Multimedia Tools and Applications (2024) 83:78183–78206
https://doi.org/10.1007/s11042-024-18607-z

DeepCPD: deep learning with vision transformer for colorectal polyp detection

Raseena T.P[1] · **Jitendra Kumar**[1] · **S.R. Balasundaram**[1]

Received: 3 October 2023 / Revised: 12 January 2024 / Accepted: 9 February 2024 /
Published online: 28 February 2024

Abstract

One of the most severe cancers worldwide is Colorectal Cancer (CRC), which has the third-highest incidence of cancer cases and the second-highest rate of cancer mortality. Early diagnosis and treatment are receiving much attention globally due to the increasing incidence and death rates. Colonoscopy is acknowledged as the gold standard for screening CRC. Despite early screening, doctors miss approximately 25% of polyps during a colonoscopy examination because the diagnosis varies from expert to expert. After a few years, this missing polyp may develop into cancer. This study is focused on addressing such diagnostic challenges, aiming to minimize the risk of misdiagnosis and enhance the overall accuracy of diagnostic procedures. Thus, we propose an efficient deep learning method, DeepCPD, combining transformer architecture and Linear Multihead Self-Attention (LMSA) mechanism with data augmentation to classify colonoscopy images into two categories: polyp versus non-polyp and hyperplastic versus adenoma based on the dataset. The experiments are conducted on four benchmark datasets: PolypsSet, CP-CHILD-A, CP-CHILD-B, and Kvasir V2. The proposed model demonstrated superior performance compared to the existing state-of-the-art methods with an accuracy above 98.05%, precision above 97.71%, and recall above 98.10%. Notably, the model exhibited a training time improvement of over 1.2x across all datasets. The strong performance of the recall metric shows the ability of DeepCPD to detect polyps by minimizing the false negative rate. These results indicate that this model can be used effectively to create a diagnostic tool with computer assistance that can be highly helpful to clinicians during the diagnosing process.

Keywords Classification · CRC · Colonoscopy · Deep learning · Vision transformer

Jitendra Kumar and S.R. Balasundaram are contributed equally to this work.

✉ Raseena T.P
rasi.tp@gmail.com

Jitendra Kumar
jitendra@nitt.edu

S.R. Balasundaram
blsundar@nitt.edu

[1] Department of Computer Applications, National Institute of Technology Tiruchirappalli, Tiruchirappalli, Tamilnadu 620015, India

1 Introduction

In this world of modern medical science, cancer is still a nightmare, and Colorectal Cancer (CRC) is one of them. According to the World Health Organization's (WHO) Global Cancer Observatory studies, in 2020, there were 1.9 million CRC diagnoses worldwide [1]. There are 1.06 million men and 0.86 million women among them. The mortality rate is 48.41% in men and women, 48.37% in men, and 48.46% in women [2]. Colorectal polyps are anomalous tissues that begin as a growth in the innermost lining of the colon or rectum. This benign growth is known as a polyp, which may progress into CRC over the years [3]. The colon's healthy inner lining cells are susceptible to DNA alterations that cause them to spread out of control. In most cases, CRC was identified at later ages (particularly in people older than 50). Nevertheless, it can be seen at any age. After 45 years of age, a screening test is required for the diagnosis of CRC [4]. It is necessary to give careful attention to diagnosing the disease. As a result, it will open a hopeful door for the patient to avail of the best possible treatment if the condition is detected early; in turn, the patient's survival chance may increase [5–7]. CRC can be prevented by identifying suspicious tissues through early diagnosis using standard methods and regular screening [8]. There are five stages of CRC. The structure and characteristics of polyps will differ in shape, size, spread area, and appearance in each stage [9]. As the stages progress from 0 to 4, cancer spreads to other body parts; hence, the survival rate decreases [10].

Colonoscopy (Fig. 1 illustrates a visual explanation of colonoscopy) is the primarily used diagnostic tool to detect colorectal polyp [11, 12]. A colonoscopy (real-time video examination) examines the large bowel and rectum. A significant chance of survival exists when malignant growths are found early and removed, which can lower the mortality rate [13, 14]. Colonoscopy is a procedure in which a flexible tube comprising a light and camera at one end is inserted into the below part of the patient's body and moved to the colon. The physician will monitor the procedure using a screen, and any suspicious polyps will be removed if found [15]. However, in colonoscopic examinations, approximately one of every four polyps may not be correctly identified, which can be influenced by factors such as the physician's experience level [16]. It has been reported that approximately 22-25% is the miss rate of polyps during colonoscopy examination due to human error [11, 17]. Examining the polyps manually is an extensive process, even if the polyp is diagnosed correctly. This subjectivity may lead to inter-observer variability and false negatives, causing missed or delayed diagnoses. Therefore, a computer-assisted diagnostic system is necessary to diagnose the condition quickly, effectively, and accurately to support physicians in shortening the diagnostic procedure [18, 19]. In response to these challenges, we propose DeepCPD, a transformer-based LMSA mechanism designed to offer a robust solution for accurately distinguishing between normal and pathological conditions in a more time-efficient manner. Adding an Acceleration layer within the Multihead Self-Attention (MSA) block enables the model to extract global features efficiently in linear time. This enhancement substantially improves the model's capability to effectively capture intricate image features, achieving an $O(n)$ time complexity [20]. The primary objective is to elevate the sensitivity of the diagnostic process with less training time, facilitating the identification of abnormalities at a stage when interventions can have the most significant impact. This work is dedicated to mitigating diagnostic challenges, thereby reducing the risk of misdiagnosis and improving overall diagnostic accuracy.

The primary contributions of this research paper are:

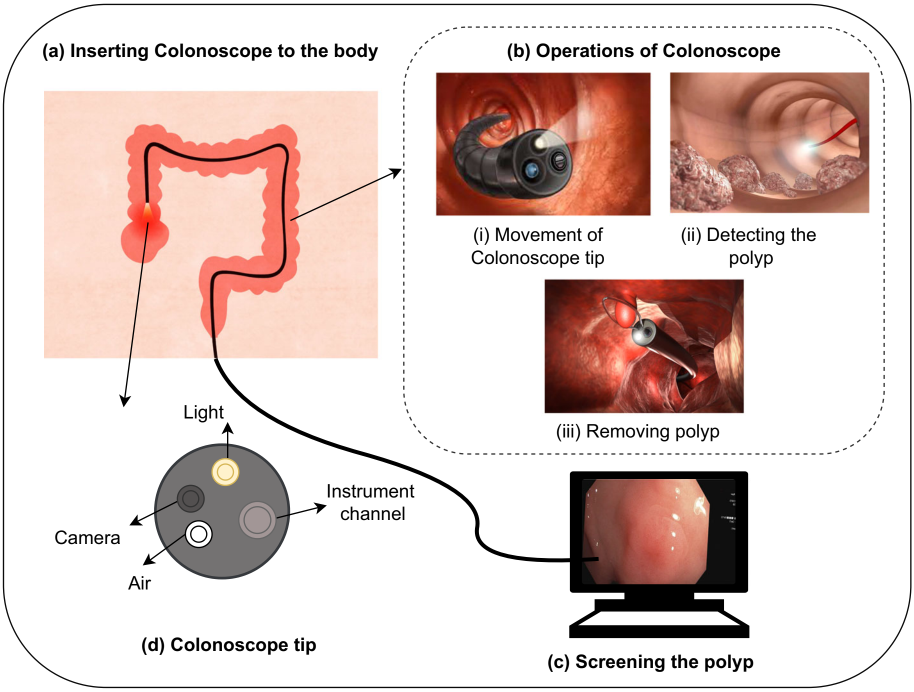


Fig. 1 Illustration of colonoscopy test. (a) to (c) demonstrate the different stages of colonoscopy examination, and (d) is an example of a colonoscope device tip

- DeepCPD effectively incorporates a Linear Multihead Self-Attention mechanism to perform binary classification on the colonoscopy images as polyp versus non-polyp and hyperplastic versus adenoma, depending on the specific dataset. This approach facilitates comprehensive global feature extraction from the image in a timely and efficient manner, encompassing features at various levels, thereby enhancing the model's performance.
- The conventional Non-Linear Multihead Self-Attention layers in the base ViT are replaced with Linear Multihead Self-Attention layers by adding two extra Acceleration layers to simplify model training, facilitating the efficient computation of contextual mappings for attention scores of image patches with linear time complexity.
- The combination of transformer design and attention mechanism has demonstrated its ability to effectively handle imbalanced datasets, as confirmed through detailed experimental validation.
- The generalizability of the DeepCPD model is validated by an extensive experimental study using four different benchmark datasets such as PolypsSet (a combination of MICCAI 2017, CVC-colonDB, GLRC and KUMC datasets), CP-CHILD-A, CP-CHILD-B and Kvasir V2.
- Finally, A series of comprehensive experiments were conducted to assess the DeepCPD model's efficacy compared to the state-of-the-art CNN models, including DenseNet201, ResNet-50, ResNet-152, and VGG19.

The subsequent sections of this paper are structured in the following manner: The prior studies in this area utilising different methodologies and datasets are described in Section 2.

The model utilised in this study is described in detail in Section 3. Section 4 portrays the architecture and workings of the proposed model. The details about the experimental setup and detailed analysis of the research's findings are provided in Section 5. Lastly, the paper concludes in Section 6.

2 Related works

The examination of medical images holds immense importance in the realm of modern medicine. Even today, healthcare professionals face challenges in effectively interpreting and comprehending diagnoses from these images. The importance of deep learning gained the researcher's attention here. Deep learning-driven Computer-Aided Diagnosis (CAD) tools can assist medical professionals during the diagnosis phase, resulting in faster diagnoses. Consequently, deep learning has gained significant popularity in medical imaging ʒ There are numerous studies conducted on this topic. This section summarises many such studies previously carried out on various datasets using feature-based and deep learning techniques.

In a recent study, Farah Younas et al. [15] introduced an ensemble deep learning classification model for colorectal polyp classification based on colonoscopy images. The initial step involved pretraining base classifiers, including GoogLeNet, Xception, and Resnet50, on the ImageNet database. Subsequently, a suitable combination of weights was determined through a grid search, and these weights were assigned to individual base classifier models to construct a weighted-average ensemble model. Alqudah et al. [21] presented an approach for CRC classification employing various machine learning algorithms, including support vector machine (SVM), artificial neural network, K-nearest neighbor (KNN), quadratic discriminant analysis, and classification decision tree (DT). This method utilizes features extracted from 3D Gray Level Cooccurrence Matrix matrices within three distinct color spaces: RGB, HSV, and L*A*B color spaces. A VGG16-based computer-aided diagnosis system was recently created by Ying-Chun Jheng et al. [22] to detect different colon polyps. The authors acquired and made use of a private dataset from a hospital. They used a set of approaches for data augmentation on the training set of data to improve the model's performance. Saraswati Koppad et al. [23] put forward an approach to analyze CRC gene characteristics of healthy and cancer patients. These genes can be taken as a measurement for CRC. This method employs three publicly accessible gene expression datasets from the GEO database and six distinct machine learning methods, including naive Bayes classifier, Adaboost, random forest, ExtraTrees, logistic regression, and XGBoost. Random forest outperformed other algorithms among these.

Ying Su et al. [24] proposed a model that utilized gene expression profiling data to diagnose colon cancer and determine its staging. Initially, gene modules were chosen, and characteristic genes were extracted using the least absolute shrinkage and selection operator algorithm and then integrated to distinguish between colon cancer and healthy controls using random forest, SVM, and DT. Furthermore, colon cancer staging was determined by leveraging differentially expressed genes associated with each stage. Subsequently, a survival analysis was also conducted. Several methods have been used to screen CRC in patients, such as fecal occult blood test, flexible sigmoidoscopy, and colonoscopy. Each technique has its drawbacks. Recent studies show that using microbiome analysis to detect CRC is a better option than using the current approaches for CRC screening. M Mulenga et al. [25] developed a deep learning model with stacking and chaining techniques as an alternative for augmentation and data normalization of microbiome data from stool samples of the patients. This suggested deep learning model is paired with rank transformation and feature selection to enhance the CRC's prediction task performance. Here, the experiment uses three microbiome datasets that are openly accessible. Devi Sarwinda et al. [26] conducted a study using ResNet models like ResNet-18 and ResNet-50 on colon gland images to categorize colon glands as benign or malignant. ResNet-50 can perform better than ResNet-18, according to the authors. They used the Warwick-QU public dataset. The researchers performed a contrast-limited adaptive histogram Equalization image preprocessing task to get more accurate images of this dataset. The classification of CRC using a deep neural network model from the gut microbiota in stool samples was another project that M Mulenga et al. [27] proposed. It would include feature extension and data augmentation. The authors employed two publicly accessible CRC-based microbiome databases. The method transfer learning is used by CP Tang [28] in a study to identify colon polyps and presented a method for computer-aided polyp detection. The authors used Faster-CNN, R-FCN, and Single Short Detector along with three other network structures, such as ResNet-50, ResNet-101, and Inception V2. R-FCN with ResNet-101 had the best outcome out of the group. Recently, deep learning produced excellent results in the domain of medicine for effectively diagnosing various diseases.

A study was conducted by CM Hsu et al. [29] to identify colorectal polyps and classify CRC. The authors experimented with alternative input data in place of RGB images. Grayscale images were created by the authors using RGB images. CNN model is used to detect and classify polyps. The data was collected from the public dataset CVC-ClinicDB and one private hospital. The researchers discovered that when the size of the polyp image is less than 1600 pixels, the accuracy in classifying and detecting polyps diminishes. A novel technique for classifying and localizing CRC in whole slide images (WSIs) simply with global labels was developed by Changjiang Zhou et al. [30], e.g., malignant or normal, by leveraging different models for deep learning. The experiment showed that ResNet performs well. The researchers also presented a new approach to classifying normal and cancerous tissues utilizing three frameworks: image-level, cell-level, and a combination of two. The experiment was carried out on two histopathology image datasets: WSI images of CRC from three different hospitals, along with histopathology images of CRC sourced from the cancer genome atlas. Paladini et al. [31] conducted a study on CRC tissue phenotyping to identify CRC using WSI pictures in the past. The authors completed this study by introducing Mean-Ensemble-CNN and NN-Ensemble-CNN, two ensemble techniques. The four pre-trained models ResNet-101, ResNet-50, InceptionV3, and DenseNet-161 are combined in these methods. The Kather-CRC-2016 database and the CRP-TP database are the two public datasets used in this research. This study was a multiclass tissue classification. In a prior study, Liew et al. [32] proposed a novel technique to classify colonic polyps using modified ResNet-50 architecture as a feature extractor and then applied Adaboost ensemble learning as the classifier. Using three publicly accessible datasets-Kvasir, ETIS-LaribpolypDB, and CVC-ClinicDB, the authors performed binary classification to distinguish between polyps and non-polyps. These three datasets are combined after selecting two classes from the Kvasir dataset.

Earlier, Mesejo et al. [18] developed a framework for doing virtual biopsies of lesions that classify lesions into three categories: hyperplastic, serrated adenomas, and adenomas lesions. The authors extracted 3D images of the lesions using the SfM algorithm and then employed white light and narrow-band imaging to enhance the features of the lesions as needed for SfM. In a study by Ruikai Zhang et al. [33], CNN was used to create a fully automated system for diagnosing and classifying colorectal polyps. This system learned basic CNN properties from two public datasets unrelated to medicine. The authors made use of the available PWH database. In a previous study, Xiaoda Liu et al. [34] introduced a deep CNN model called faster-rcnn-inception-resnet-v2 for classifying polyps and adenomatous lesions. The authors used a private dataset for this work. In a recent study, Nisha J.S et al. [35] developed a Dual-path CNN model to detect polyps through an image enhancement technique on three datasets, CVC-clinicDB, CVC-colonDB, and Etis-Larib, and the authors achieved high performance on CVC-colonDB. Krushi Patel et al. [36] researched various deep learning models to classify the polyps into hyperplastic and adenomatous. The authors used six CNN models and five different colonoscopy datasets to carry out their study. The VGG19 model outperformed the other models in that group.

In recent work, Chung Ming Lo et al. [37] developed the FEViT model, an ensemble classifier of ViT and KNN, and equipped it as a classifier in ViT rather than a multilayer perception layer to perform binary classification. Some of the clinically based numerical features are integrated and then served to the classifier with the image features obtained by ViT. Similarly, Mohamed et al. [38] suggested a vision transformer-based multiclassifier for CRC histological images. Before being processed by the transformer, the authors underwent various preprocessing procedures on the dataset. A publicly available benchmark dataset named CRC-5000 has been utilized to assess the model. Wang et al. [39] proposed an innovative self-supervised learning approach termed semantically-relevant contrastive learning. Their methodology involves a hybrid model integrating a CNN and a multi-scale Swin Transformer architecture to classify histopathology images. Hussein et al. [40] suggested a DeepPoly method that uses DoubleU-Net for polyp segmentation and a vision transformer for binary classification. Using a fine-tuned vision transformer, the DeepPoly approach classifies the segmented polyps as hyperplastic or adenoma. Both private and public datasets are utilized for this study after annotation.

From the literature studies, it has been understood that most of the previous studies were conducted using different CNN algorithms in combination with other techniques; most of them are complex in nature and require high training time. The excessive down-sampling task performed on images by some of the methods results in a loss of information. As a result, the models miss several tiny polyps. Hence, a new approach is necessary to detect colon polyps from the images. Comparatively, fewer studies have been conducted on the classification task utilizing transformer-based models, especially within CRC. Furthermore, the majority of these studies have centered around CRC histological images rather than directly addressing colonoscopy images. Colonoscopy images hold a pivotal role in the initial diagnosis of colorectal conditions. Conversely, CRC histological images are typically employed for in-depth analysis of the condition, offering detailed insights into tissue structures and abnormalities. Empirical evidence derived from these studies has consistently demonstrated the superiority of vision transformers over traditional CNN models in classification performance. However, a significant gap in research focused on enhancing CRC detection using colonoscopy images still needs to be addressed. Additional investigations and studies are warranted to harness the potential of transformer-based models in improving the accuracy and effectiveness of CRC detection from colonoscopy images.

3 Vision transformer (ViT)

Typically, in computer vision, the attention mechanism is either employed directly on CNN or with some changes to the design of the CNN. An elementary ViT can attain a performance level of 4.6 times more resilient than the top-performing transfer learning models based on CNN [41]. ViT is a transformer architecture working based on the attention mechanism [42, 43]. The three components that make up the transformer encoder are Layer Normalization (LN), Multi-Head Self Attention Layer (MSA), and Multi-Layer Perceptrons Layer (MLP).

The self-attention mechanisms used in natural language processing inspired using ViT in computer vision tasks, notably in image classification [44, 45]. The attention mechanism helps to capture features of the image at the lowest level of abstraction. ViT has undergone pretraining on ImageNet-21k, which contains 14 million images in 224 × 224 resolution and 21,000 classes. ViT divides the images into visual tokens by splitting the image into patches of the same size and feeds the series of their linear embeddings into a Transformer encoder along with positional encoding. The sequence of linear embeddings also has a particular learnable token prepended to it. Following that, the transformer uses the attention method to produce a series of output tokens. The output value of the learnable token is fed into a classification head connected to the encoder's output and is implemented by MLP, which produces a classification label depending on the state.

4 Proposed method

The system workflow of the proposed model is given in Fig. 2. There are three major phases in the proposed model. The first phase is to increase the data size which is necessary to avoid the challenges associated with less data during training. The second phase is preprocessing the data, which includes resizing and normalizing procedures. In the third phase, a thorough exploration of the proposed model and its architecture is provided, offering detailed insights into the process of classifying colonoscopy images into binary categories. These three phases are described in more detail in subsequent subsections.

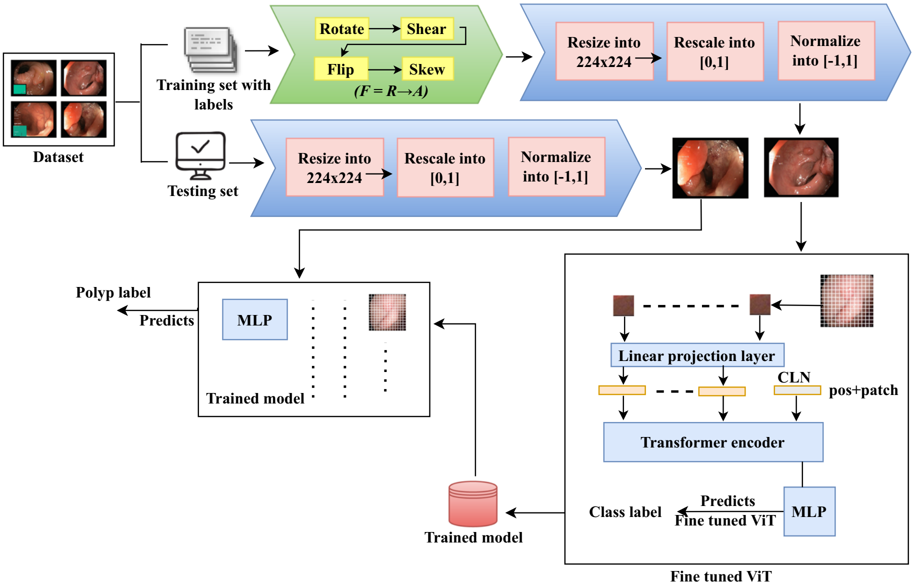


Fig. 2 System workflow of the proposed method for colorectal polyps detection

4.1 Data augmentation

The quality and size of the training data have a significant influence on the deep learning model's performance. However, one of the most prevalent problems that deep learning faces is inadequate data. A deep learning model performs better if the dataset is large and sufficient [46]. Data augmentation is applied to address this issue, which generates new data samples from the existing ones by combining multiple mathematical operations, which is stated in (1).

$$F(x) : R \rightarrow A; \qquad x \in \{rotation, shearing, flipping, skewing\} \tag{1}$$

Where $F(x)$ is the transformation function that performs augmentation, x is the transformation set, R is the training set of the original dataset, and A is the augmented set of R. Thus, the artificially generated training set is shown in (2).

$$R' = R \cup A \tag{2}$$

Where R' represents the entire training set, including both the original training set and the augmented set, which is created through a sequence of distinct geometric transformations, including rotation, shearing, flipping, and skewing, these four geometric transformations were selected with consideration for the characteristics of the dataset. The rotation transformation is applied within an angle range spanning from -90 to +90 degrees. For the shear transformation, the parameters are configured to shear the images by a maximum angle of 20 degrees to the left along the x-axis and a maximum angle of 20 degrees to the right along the x-axis. The flipping operation is employed to flip the colon images horizontally or vertically. Lastly, a skew transformation distorts the images towards random corners. Each of these four geometric transformations is executed with a probability of 0.5 on the datasets.

Data augmentation is performed only on the CP-CHILD-B and Kvasir V2 datasets since these have less number of samples. Performing different augmentation techniques with x on the training set of the datasets resulted in the generation of 10000 images overall (7352 for polyp class and 2648 for non-polyp class) in CP-CHILD-B and 10000 images overall (4964 for polyp class and 5036 for non-polyp class) in Kvasir datasets.

4.2 Preprocessing

Data preprocessing is a crucial step in medical imaging tasks and in many other fields of image analysis. Medical image data often come in various forms and may exhibit characteristics such as varying dimensions, noise, and inconsistencies. Therefore, standardizing and preparing the colorectal datasets before feeding into the proposed model is essential to ensure that the model can effectively learn and make accurate predictions.

The preprocessing pipeline applied to the proposed model encompasses resizing, rescaling, and normalization procedures. The dataset comprises colonoscopy images of varying dimensions, necessitating the transformation of colon images into the correct input format for training with the proposed model. Initially, every image in the dataset undergoes resizing to achieve a uniform size of 224 × 224 pixels, aligning with the standard size required for generating patches from the input images. Subsequently, rescaling is executed, which maps the pixel values from their original range of [0, 255] to the normalized range [0, 1]. This rescaling operation is accomplished by dividing the pixel values by a rescaling factor of

1/255. Finally, the normalization step is applied to the RGB channels of the images using the specified mean (0.5, 0.5, 0.5) and standard deviation (0.5, 0.5, 0.5) using (3). Consequently, the resulting pixel range is standardized to [-1, 1].

$$I_n = \frac{(I_c - M_c)}{SD_c} \tag{3}$$

Where I_n is the normalized image, I_c is the pixel value of the input's RGB channel, M_c is the mean value of the RGB channel, and SD_c is the standard deviation of the RGB channel.

4.3 DeepCPD

The proposed model uses ViT as the foundational architecture for image classification, integrating two additional Acceleration layers to enhance training efficiency, subsequently leading to fine-tuning with colonoscopy datasets. This fine-tuning process enhances the model's ability to distinguish between colonoscopy images. The initial weights of DeepCPD are set using pre-trained ViT weights, and the classification layer is modified with specific hyperparameters optimized for the characteristics of colonoscopy datasets. Figure 3 visually depicts the proposed model architecture in detail. The model processes the input image by dividing it into non-overlapping fixed-size patches of 16 × 16. Subsequently, these patches are vectorized, resulting in a 1D patch representation after flattening, as described in (4), (5), (6), (7) and (8).

$$I_s = h \times w \times c \qquad ie,\, I_s = (224 \times 224 \times 3) \tag{4}$$

$$P_s = P_h \times P_w \qquad ie,\, P_s = (16 \times 16) \tag{5}$$

$$N_p = \frac{(h \times w)}{Ph \times P_w} \qquad ie,\, N_p = 196 \tag{6}$$

$$T_i = (N_p, P_h \times P_w \times c) \qquad ie,\, T_i = (196, 16 \times 16 \times 3) \tag{7}$$

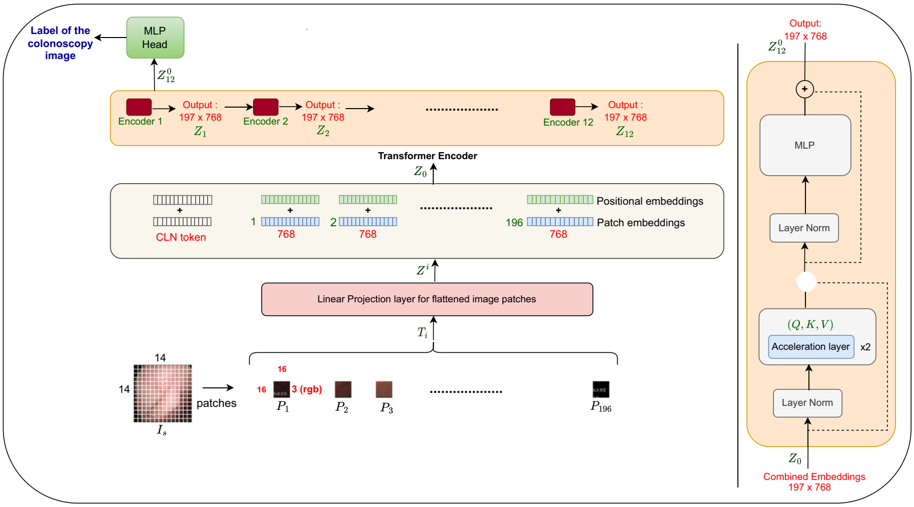


Fig. 3 System architecture of the proposed method for colorectal polyps detection (on the left) and a detailed view of the design of the modified encoder block in the transformer (on the right)

$$I_p^i \in \mathbb{R}^{P_h \times P_w \times c} \rightarrow I_p^i \in \mathbb{R}^{1 \times P_h \times P_w \times c} \qquad ie, I_p^i[1 \times 16 \times 16 \times 3] \tag{8}$$

Where I_s stands for the size of the image, h stands for the image's height, w stands for the Image's width, c stands for the number of RGB channel, P_s stands for the patch's size, P_h stands for the Patch's height, P_w stands for the patch's width, N_p stands for the number of Patches, T_i denotes the transformed input, while I_p^i represents the reshaped 1D vector of each patch. Following the conversion of the input image into a 1D patch representation, a linear projection layer is used to map each of these patches to D (1×768) dimensional patch embeddings. The entirety of patch representations undergoes this linear transformation, producing the vector Z. This vector Z functions as the feature vector for the image patches and is formally represented in (9).

$$I_p \in \mathbb{R}^{P_h \times P_w \times c} \rightarrow Z \in \mathbb{R}^{N \times D} \tag{9}$$

Equations (10) and (11) are the mathematical operations involved in the linear projection layer:

$$I_p^i \in \mathbb{R}^{1 \times P_h \times P_w \times c} \cdot W \in \mathbb{R}^{D \times P_h \times P_w \times c} = I_p^i W = Z^i \in \mathbb{R}^{1 \times D} \qquad ie, Z^i[1 \times 768] \tag{10}$$

In this step, the reshaped 1D patch is multiplied with a weight matrix W to produce Z^i, which is the patch embedding of each patch from the linear projection layer, where $i = 1, 2....196$. Such 196 patch embeddings are obtained. This process is done for all patches as given below,

$$I_p \in \mathbb{R}^{N \times P_h \times P_w \times c} \cdot W \in \mathbb{R}^{P_h \times P_w \times c \times D} = I_p W = Z \in \mathbb{R}^{N \times D} \qquad ie, Z[196 \times 768] \tag{11}$$

Since Z^i is represented as a 1×768 long vector, the patch embedding matrix Z is 196×768 in size. Transformers lack a built-in mechanism that considers the "order" of patch embeddings. The order of the image patches in the image can significantly change its meaning, so a method for letting the model guess the order of the image patches is necessary. By using the positional embedding method, add a unique position to the linear projection of each patch in the form of vectors, which is mentioned in (12). The proposed model is thus aware of the patch sequence's order throughout training. At this stage, before feeding the colorectal image patches to the encoders, the proposed model introduces a learnable class token CLN to the patch embeddings. The final feature vector corresponding to the CLN token is used by the MLP head for classification.

$$Z_0 = [X_{CLN}; Z^i \in \mathbb{R}^{1 \times D}] + E_{pos}, \ E_{pos} \in \mathbb{R}^{(N+1) \times D} \qquad ie, Z_0[197 \times 768] \tag{12}$$

The patch embeddings and positional embeddings are merged into vector space during training. The embeddings exhibit significant similarity to their nearby position embeddings, especially those sharing the same column and row. The patch embeddings are now 197×768 in size. Subsequently, the series of embedded patches Z_0 is fed into the transformer's architecture comprising 12 identical encoders. The LMSA and MLP blocks comprise most of the transformer encoder's structure, which is depicted on the right side of Fig. 3. To simplify model training, the input undergoes normalization using layer normalization, followed by two Acceleration layers inside the LMSA block. Furthermore, a residual connection is also employed after each block. The Acceleration layers incorporate two matrices, A_l^K and A_l^V, to simplify the calculation of attention scores for patch embeddings, reducing the complexity from $O(n^2)$ to $O(n)$ [20]. Equations (13) and (14) describe how the transformer encoder's entire encoding process works.

$$Z_l' = LMSA(LN(Z_{l-1})) + Z_{l-1} \qquad ie, l = 1......12, \quad Z_{l-1} = Z_0.....Z_{11} \tag{13}$$

$$Z_l = MLP(LN(Z'_l)) + Z'_l \qquad ie, l = 1......12 \tag{14}$$

Here, l represents the encoder number. Z'_l and Z_l are the feature vectors generated by LMSA and MLP layers in each encoder, respectively. The process outlined in (13) is subdivided and defined independently, with detailed explanations using (15), (16), (17) and (18). Three parameters, query (Q), key (K), and value (V) are used by the LMSA layer of the transformer's attention mechanism to find dependencies among several patches of the input image.

$$K^A = A_l^K \cdot Z_{l-1}; \qquad V^A = A_l^V \cdot Z_{l-1} \qquad ie, A_l^K, A_l^V \in \mathbb{R}^{M \times N} \tag{15}$$

The K^A and V^A pairs are initially calculated using the matrices A_l^K and A_l^V generated by the two Acceleration layers introduced inside the LMSA block. These layers transform the original $(N \times D)$ dimensional K and V matrices into $(M \times D)$ dimensional matrices, facilitating the efficient computation of contextual mappings for attention scores with linear time and memory complexity by opting for a smaller projected dimension M, where $M < N$. The value of M is fixed at 90 for all the encoders. Subsequently, the QKV scores are computed by multiplying Z_{l-1}, K^A, and V^A with the learnable weight matrices W_l^Q, W_l^K, and W_l^V, respectively. The calculation of attention scores for patch embeddings by the LMSA block is elaborated in (15), (16) and (17).

$$Q = Z_{l-1} \cdot W_l^Q; \qquad K = K^A \cdot W_l^K; \qquad V = V^A \cdot W_l^V \quad ie, Q \in \mathbb{R}^{N \times D} and K, V \in \mathbb{R}^{M \times D} \tag{16}$$

After calculating the dot product of Q with the transpose of K, the dimension is scaled with the square root to avoid the vanishing gradient problem. The attention score for each patch of the image is obtained by applying the softmax function to the, which is then multiplied with V, which is given in (17).

$$A^j_{QKV}(S) = softmax\left(\frac{QK^T}{\sqrt{D}}\right)V \qquad ie, A^j_{QKV}(S) \in \mathbb{R}^{N \times D} \tag{17}$$

The attention score of the image generated by one head is $A^j_{QKV}(S), \quad j = 1, 2, 3, ...12;$ LMSA has such 12 heads. The attention scores of all heads are then concatenated together and projected through a dense layer with a learnable weight matrix W to make the output into the desired dimension, as mentioned in (18).

$$LMSA(Z_l) = concat(A_1(Z_l), A_2(Z_l), ...A_j(Z_l))W^0, W^0 \in \mathbb{R}^{(D \times h), D} \tag{18}$$

A residual connection is integrated with $LMSA(Z_l)$ to obtain Z'_l, representing the final attention scores for the LMSA layer in the transformer encoder as described in (13), which is produced by processing the scores of 12 self-attention heads simultaneously, each of which can concentrate on distinct relationships between the image patches. Z'_l is then fed into the MLP block of the encoder architecture, which starts with a Layer normalization followed by an MLP layer and a residual connection as described in (14). Two fully connected layers with a GELU at the end make up the MLP layer. Z^0_{12} represents the feature vector extracted from the input image by CLN token generated from the 12^{th} encoder of the transformer, which is used for classification purposes. The other 196 tokens are not considered since the CLN token encapsulates a feature vector that effectively summarizes information from the entire set of image patches. The resultant feature vector Z^0_{12} is the final representation of the CLN token from the transformer encoder, which is 1 × 768 in size, which is further normalized by a linear layer as described in (19),

$$Y = LN(Z_{12}^{0}) \tag{19}$$

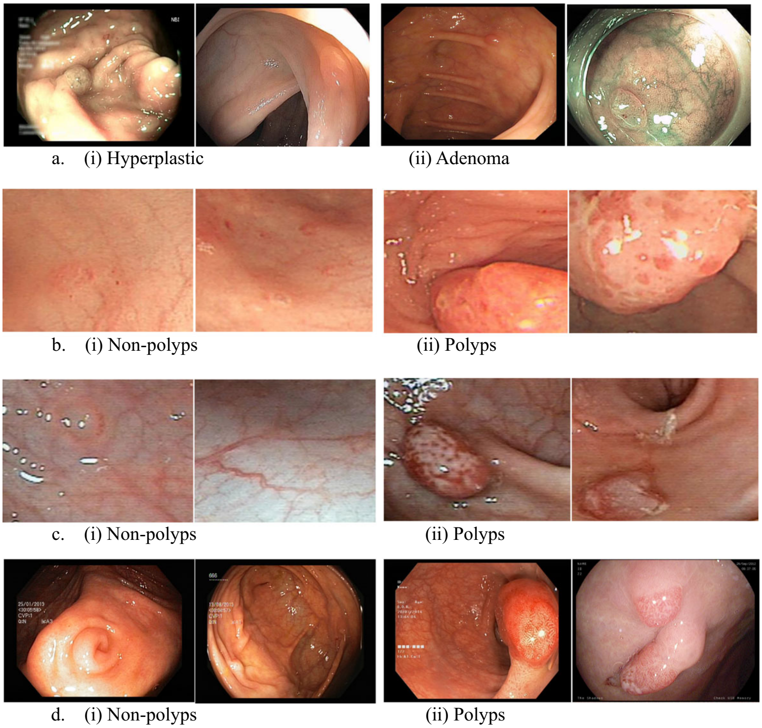
Y is the final feature vector of the colorectal image with attention score after normalization, which is processed by the classification head in the final step to predict the class label of the colorectal image. The activation function Softmax is applied to this output to produce classification labels, $Y_{pred} = softmax(Y)$. Y_{pred} can be labeled as a polyp versus non-polyp or hyperplastic versus adenoma. The Algorithm to classify colorectal images is elucidated in Algorithm 1.

Algorithm 1 Algorithm for colorectal image classification using DeepCPD.

```
Initialize $epoch, M$;
Apply data augmentation on images ($I$);
Apply rescaling and normalization on $I$;
Divide $I$ into patches, $I_p \leftarrow I$ (Refer (4), (5) and (6));
Reshape patches into 1D vectors,
**while** $i \leq epoch$ **do**
    $I_p W = Z \in \mathbb{R}^{N \times D}$ (Refer (11));
    Add positional encoding and CLN token (Non-trainable parameters) to $Z$ to produce $Z_0$;
    Apply normalization on $Z_l$, where $l = 1, 2, ...12$;
    Generate $Q = Z_{l-1} \cdot W_l^Q$, $K = K^A \cdot W_l^K$ and $V = V^A \cdot W_l^V$ (Refer (15) and (16));
    Compute attention scores by one head $A_{QKV}^{j}(S) = softmax\left(\frac{QK^T}{\sqrt{D}}\right)V$;
    Concatenate attention scores of 12 heads, $LMSA(Z_l) = concat(A_1(Z_l), A_2(Z_l), .....A_j(Z_l))W^0$, where $j = 1, 2, ...12$;
    Compute $Z_l' = LMSA(Z_l) + Z_{l-1}$;
    Apply normalization on $Z_l'$;
    Compute $Z_l = MLP(Z_l') + Z_l'$;
    Apply Normalization on $Z_{12}^0$, $Y = Z_{12}^0$. Here, $Z_{12}^0$ represents the entire feature representation of $I$ from $12^{th}$ encoder;
    Apply Softmax on $Y$;
**end**
Predict the label
```

5 Experiment results and analysis

This section begins by addressing the datasets employed in the study, detailing the experimental setup for DeepCPD, and outlining the various evaluation metrics used to assess the performance of the proposed method alongside other comparative studies. Then, a comprehensive analysis of DeepCPD's performance follows, including comparisons with benchmark methods and other pretrained CNN models. The discussion section concludes by providing detailed insights into the significance of this study and acknowledging its limitations.


a. (i) Hyperplastic (ii) Adenoma

b. (i) Non-polyps (ii) Polyps

c. (i) Non-polyps (ii) Polyps

d. (i) Non-polyps (ii) Polyps

Fig. 4 Different examples of the images in the datasets. (a) is the example of hyperplastic and adenoma from PolypsSet, (b) is the example of non-polyps and polyps from CP-CHILD-A, (c) is the example of non-polyps and polyps from CP-CHILD-B, and (d) is the example of non-polyps and polyps from Kvasir V2

5.1 Datasets

The four benchmark datasets of colonoscopy images used in this study are PolypsSet (combinations of four distinct datasets), CP-CHILD-A, CP-CHILD-B, and Kvasir V2 datasets. For examples of diverse images in various datasets, Fig. 4 depicts examples of images from each dataset. The details of each dataset are described in the Table 1.

Table 1 Information on the sample size of each dataset used for the proposed study

Dataset	Class wise number of samples	
	Original dataset	Augmented dataset
PolypsSet	Adenoma: 19240 Hyperplastic: 16741	–
CP-CHILD-A	Non-polyp: 7000 Polyp: 1000	–
CP-CHILD-B	Non-polyp: 1100 Polyp: 400	Non-polyp: 7652 Polyp: 2748
Kvasir V2	Non-polyp: 1000 Polyp: 1000	Non-polyp: 5236 Polyp: 5164

5.1.1 PolypsSet dataset

This dataset includes the hyperplastic and adenomatous images from the KUMC dataset, the CVC-colon DB dataset, the GLRC dataset, and the MICCAI 2017 dataset. All images of this dataset are extracted from different frames of colonoscopy videos and labelled by Li et al. [47].

MICCAI 2017 dataset: This dataset was released at the 2017 MICCAI GIANA Endoscopic Vision Challenge [48]. The training and testing set data are extracted from 18 short colonoscopy videos and 20 short colonoscopy videos, respectively.

CVC-colon DB dataset: The dataset includes different frames from 15 colonoscopy video sequences [49].

GLRC dataset: This dataset contains images from 76 colonoscopy video sequences [18].

KUMC dataset: The KUMC dataset contains images taken from 80 short colonoscopy video sequences [47].

The polypsSet dataset has 35981 images of polyps with various dimensions, divided into hyperplastic and adenomatous classes. The split between training and testing for the dataset is 80% and 20%.

5.1.2 CP-CHILD-A and CP-CHILD-B datasets

Images from 1600 children's colonoscopies, ranging in age from 0 to 18 years, are included in this dataset [50]. The images are all annotated and split into CP-CHILD-A and CP-CHILD-B datasets. The CP-CHILD-A dataset has 8000 colonoscopy images, of which 7000 are non-polyp, and 1000 are polyp images. The dataset includes a training set of 6200 non-polyp and 800 polyp images and a testing set of 800 non-polyp and 200 polyp images. The dataset CP-CHILD-B contains 1500 RGB colonoscopy images, of which 1100 are non-polyp and 400 are polyp images. The dataset comprises 800 non-polyp images and 300 polyp images for training, as well as 300 non-polyp images and 100 polyp images for testing.

5.1.3 Kvasir V2 dataset

This dataset is prepared from the Kvasir dataset [51]. The Kvasir dataset includes eight classes of endoscopically acquired colour images of the gastrointestinal tract that medical professionals have annotated and confirmed. However, this study requires only two classes, polyps and non-polyps (normal cecum), related to colorectal disease; the remaining classes are unrelated to this study. The dataset contains images of various resolutions, ranging from 720 × 576 to 1920 × 1072 pixels. The dataset contains two versions, Kvasir version 1 and Kvasir version 2. The proposed study made use of the latest version. The dataset's images were gathered from four hospitals under the Vestre Viken Health Trust in Norway, which consists of 2000 images of polyps and non-polyps, 1000 in each class. The split between training and testing for the dataset is 80% and 20%.

5.2 Implementation details

The proposed model is implemented on the PARAM PORUL supercomputing unit provided by the National Supercomputing Mission with the Centre for Development in Advanced Computing support. A GPU node comprises two Intel Xenon Gold-6248 processors and two Nvidia V100 GPU cards. The DeepCPD is implemented using the Pytorch framework. The model performed well with the hyperparameters such as the Adam optimizer with a batch size of 32, learning rate of 2e-5, momentum values of beta1 of 0.9 and beta2 of 0.999, and epsilon of 1e-8. The model exhibited superior performance with a learning rate of 2e-5 compared to 2e-6 and 2e-7 and a batch size of 32 compared with 64, while the remaining parameters were selected randomly. The proposed model was trained five times on each dataset, and the average of the experiments is reported in this work.

5.3 Evaluation metrics

The efficacy of the proposed method is measured using various metrics. In view of the imbalanced datasets used in this research, recall and precision are the primary evaluation matrices of the suggested model. Equation (20) can be used to compute recall (Rec) and precision ($Prec$).

$$\begin{aligned} Rec &= \frac{TP}{TP + FN} \\ Prec &= \frac{TP}{TP + FP} \end{aligned} \tag{20}$$

In addition, the F1-score ($F1$), accuracy (Acc), specificity ($Spec$), and Matthews Correlation Coefficient (MCC) metrics are computed as given in (21), (22), (23) and (24).

$$F1 = 2 * \frac{Prec * Rec}{Prec + Rec} \tag{21}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{22}$$

$$Spec = \frac{TN}{TN + FP} \tag{23}$$

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \tag{24}$$

TP, TN, FP, and FN represent True Positive, True Negative, False Positive, and False Negative, respectively. These terms can be derived from the confusion matrix, a performance metric used in classification problems where the output is categorized into two or more classes.

In addition, to determine the correctness of the prediction of the classes, the AUC-ROC curve is plotted to validate the performance of the model visually.

5.4 Performance analysis of the proposed method

The results and analysis of the proposed method for classifying colonoscopy images into polyp vs. non-polyp or hyperplastic vs. adenoma are discussed in this section. This study involved extensive experiments aimed at assessing the DeepCPD's performance. The DeepCPD model demonstrated remarkable performance across various metrics, including accuracy, recall, precision, F1-score, specificity, and MCC, achieving results such as 99.90% accuracy, 99.87% recall, 99.94% precision, 99.90% F1-score, 99.94% specificity, and a 99.81% MCC on PolypsSet, 99.60% accuracy, 98.10% recall, 99.89% precision, 98.98% F1-score, 99.97% specificity, and a 98.74% MCC on CP-CHILD-A, 99.45% accuracy, 98.20%

Table 2 Classification report of the proposed DeepCPD method on four different benchmark datasets

Dataset	Acc (%)	Rec (%)	Prec (%)	F1 (%)	Spec (%)	MCC (%)	Training time
PolypsSet	99.90	99.87	99.94	99.90	99.94	99.81	1.2x
CP-CHILD-A	99.60	98.10	99.89	98.98	99.97	98.74	1.4x
CP-CHILD-B	99.45	98.20	99.59	98.88	99.86	98.53	1.3x
Kvasir V2	98.05	98.40	97.71	97.94	97.70	95.81	1.3x

recall, 99.59% precision, 98.88% F1-score, 99.86% specificity, and a 98.53% MCC on CP-CHILD-B, and 98.05% accuracy, 98.40% recall, 97.71% precision, 97.94% F1-score, 97.70% specificity, and a 95.81% MCC on Kvasir V2 datasets. Detailed performance analyses of the proposed method across the four datasets are provided in Table 2, representing the average results obtained from five experimental trials. The analysis of training time underscores the computational efficiency of DeepCPD in contrast to the foundational ViT architecture. The observations clearly indicate that DeepCPD, incorporating two extra acceleration layers in the LMSA block, can compute attention scores for patch embeddings in a shorter training time compared to the base ViT. Notably, DeepCPD's training is accelerated by over 1.2 times compared to the base ViT. These performance metrics highlight DeepCPD's potential to efficiently distinguish colonoscopy images, particularly its strong performance in the recall metric, which is considered the most valuable metric in this study. This metric is crucial because it accounts for false negative classifications, ensuring that individuals with suspected CRC are not incorrectly categorized as non-polyp or hyperplastic cases, representing normal colon conditions. Simultaneously, further investigations are necessary for polyp and adenoma cases. Thus, minimizing the false negative rate is of utmost importance.

Figure 5 illustrates the confusion matrix for each dataset, providing insights into the number of images correctly and incorrectly classified into their respective categories. For example, (a) in Fig. 5, it is evident that 3749 adenoma images out of 3751 were correctly classified into the adenoma class, while 3442 hyperplastic images out of 3446 were correctly classified into the hyperplastic class. Figure 6 graphically visualizes the AUC-ROC curves for all datasets. These curves demonstrate that the DeepCPD distinguishes colon images into the correct classes, consistently achieving accurate classifications for all four datasets. The best results from five experiments are reported in both the confusion matrix and AUC-ROC curves. Based on these findings, the DeepCPD model has demonstrated state-of-the-art

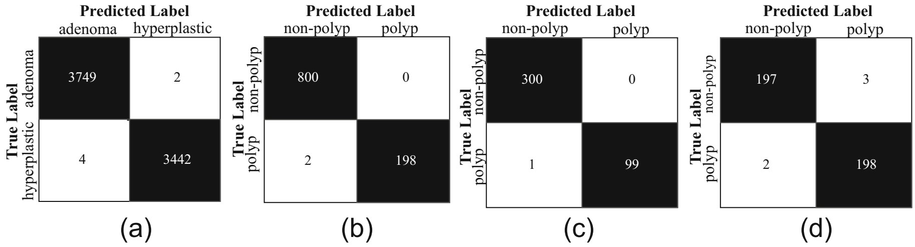


Fig. 5 Confusion matrix of the proposed DeepCPD method for all datasets. (a) confusion matrix for PolypsSet, (b) confusion matrix for CP-CHILD-A, (c) confusion matrix for CP-CHILD-B, and (d) confusion matrix for Kvasir V2

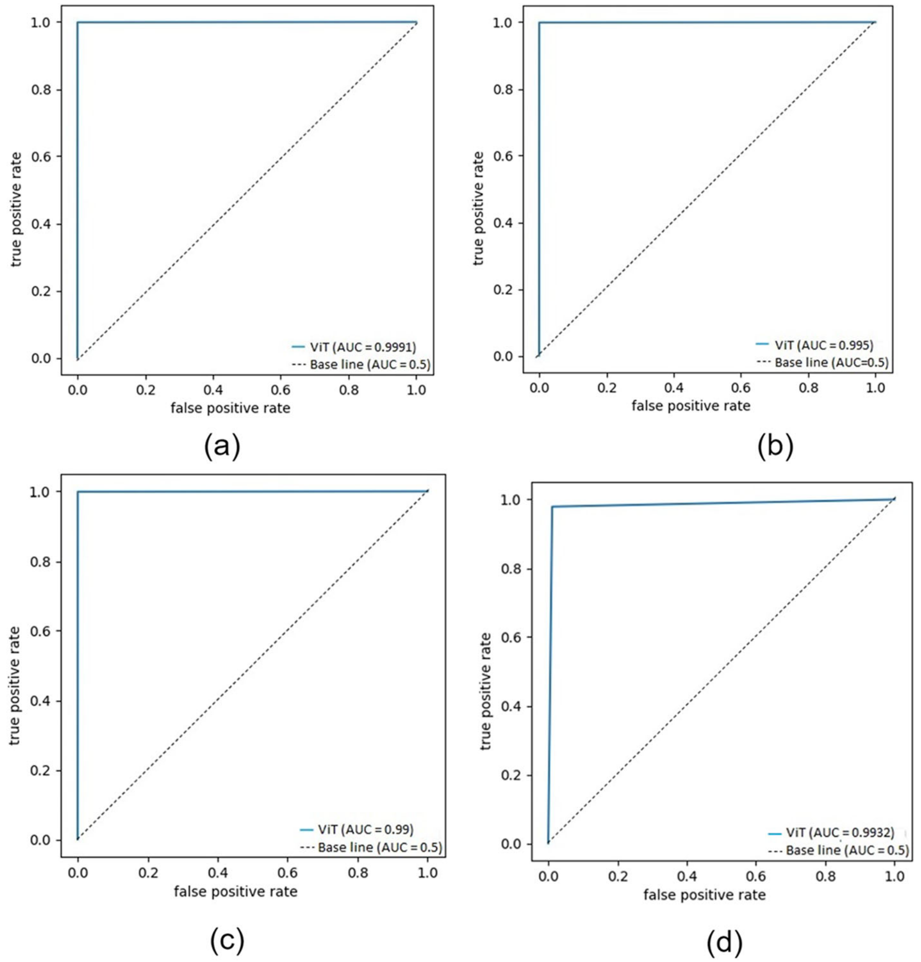


Fig. 6 AUC curve of the proposed DeepCPD method for all the datasets. (a) AUC curve for PolypSet, (b) AUC curve for CP-CHILD-A, (c) AUC curve for CP-CHILD-B, and (d) AUC curve for Kvasir V2 datasets

performance by leveraging a novel approach for the image classification of colonoscopy, showcasing its potential to advance the field.

5.5 Comparison with the benchmark methods

A comparative analysis is conducted to compare and evaluate the performance of the proposed DeepCPD method with state-of-the-art polyp classification methods using benchmark datasets, namely PolypsSet, CP-CHILD-A, CP-CHILD-B, and Kvasir V2. This analysis aims to showcase the effectiveness of the proposed method in comparison to state-of-the-art polyp classification methods. In a related study, Patel et al. used a basic VGG19 model to train it end-to-end on the PolypsSet dataset to classify the images into hyperplastic and adenoma following preprocessing steps that involved cropping the images. The authors obtained an accuracy of 79.78%, recall of 78.64%, precision of 78.71%, and F1-score of 78.67%. The pro-

Table 3 Comparison of the proposed DeepCPD method with the existing state-of-the-art polyp classification methods on four different benchmark datasets

Dataset	Method	Underlying architecture	Acc (%)	Rec (%)	Prec (%)	F1 (%)
PolypsSet	DeepCPD	ViT	**99.93**	**99.89**	**99.97**	**99.92**
	Patel et al. [36]	VGG-19	79.78	78.64	78.71	78.67
CP-CHILD-A	DeepCPD	ViT	**99.80**	**99.00**	**100**	**99.49**
	Wei Wang et al. [50]	ResNet152-GAP	99.29	97.55	–	–
CP-CHILD-B	DeepCPD	ViT	**99.75**	**99.00**	**100**	**99.49**
	Wei Wang et al. [50]	ResNet152-GAP	99.35	97.70	–	–
Kvasir V2	DeepCPD	ViT	**98.75**	**99.00**	98.50	**98.01**
	Liew et al. [32]	Modified ResNet-50 with AdaBoost	97.91	96.45	**99.35**	97.90

vided recall, precision, and F1-score metrics were calculated using a macro-average approach across the two classes, considering that the authors specified individual recall, precision, and F1-score values for each class in their study. In another study, Wei Wang et al. [50] introduced a ResNet152-GAP model based on the ResNet architecture and incorporates Global Average Pooling in place of a fully connected layer. After applying preprocessing operations, the model was evaluated on two separate datasets, CP-CHILD-A and CP-CHILD-B. How-

Table 4 Comparison of the proposed DeepCPD method with the state-of-the-art CNN methods on four different benchmark datasets

Dataset	Method	Acc(%)	Rec (%)	Prec (%)	F1 (%)	Spec (%)	MCC (%)
PolypsSet	DeepCPD	99.93	99.89	99.97	99.92	99.97	99.86
	DenseNet201	65.07	69.73	65.51	67.55	59.99	29.87
	ResNet-50	75.19	78.04	75.24	76.62	72.01	50.18
	Resnet-152	73.18	74.63	74.12	74.38	71.60	46.25
	VGG19	77.59	78.18	78.70	78.44	76.95	55.11
CP-CHILD-A	DeepCPD	99.80	99.00	100	99.49	100	99.37
	DenseNet201	96.60	85.00	97.70	90.90	99.50	89.16
	ResNet-50	99.00	96.50	98.46	97.47	99.63	96.86
	Resnet-152	99.20	97.00	98.97	97.97	99.75	97.49
	VGG19	99.40	98.00	98.98	98.49	99.75	98.12
CP-CHILD-B	DeepCPD	99.75	99.00	100	99.49	100	99.33
	DenseNet	92.50	88.00	83.01	85.43	94.00	80.45
	ResNet-50	99.00	97.00	98.97	97.97	99.67	97.32
	Resnet-152	99.25	99.00	98.01	98.50	99.33	98.01
	VGG19	99.25	99.00	98.01	98.50	99.33	98.01
Kvasir V2	DeepCPD	98.75	99.00	98.50	98.01	98.50	97.50
	DenseNet201	89.00	93.00	86.11	89.42	85.00	78.25
	ResNet-50	94.25	93.50	94.92	94.20	95.00	88.51
	Resnet-152	96.00	96.00	96.00	96.00	96.00	92.00
	VGG19	97.00	97.50	96.53	97.01	96.50	94.00

ever, augmentation techniques were only used for CP-CHILD-A. ResNet152-GAP achieved an accuracy of 99.29% and a recall of 97.55% on CP-CHILD-A, while on CP-CHILD-B, it attained an accuracy of 99.35% and a recall of 97.70%. The authors did not provide metrics such as precision and F1-score for ResNet152-GAP. Liew et al. [32] proposed a modified ResNet50 in combination with the PCA, AdaBoost and other preprocessing techniques, including a median filter for noise removal of images for a colonic polyp classification system utilizing Kvasir V2. The proposed model achieved an accuracy of 97.91%, recall of 96.45%, and precision of 99.35%. The F1-score, although not directly provided by the authors, was computed as the mean value of precision and recall, resulting in an F1-score of 97.90%. On the benchmark datasets, DeepCPD outperformed the performance of all three discussed state-of-the-art methods, except the precision value of Kvasir V2 is precisely 0.85% less than the benchmark result. Table 3 presents a comparative analysis of DeepCPD's performance with existing state-of-the-art methods across all four benchmark datasets. This analysis uses the optimal result obtained from the five experiments conducted on each dataset.

Furthermore, the DeepCPD surpasses other state-of-the-art deep learning methods on all the datasets. Detailed results are presented in Table 4, while a visual representation of the findings is illustrated in Fig. 7. A comparative analysis of DeepCPD against leading CNN models has been conducted, such as Densenet201, RsNet-50, ResNet-152, and VGG19, all of which were trained using identical methods, which involved data augmentation and preprocessing techniques as well as utilizing transfer learning, following the same approach used for DeepCPD. For all four CNN models, all layers except the last two layers were set to a frozen state. Subsequently, a global average pooling layer was introduced to foster

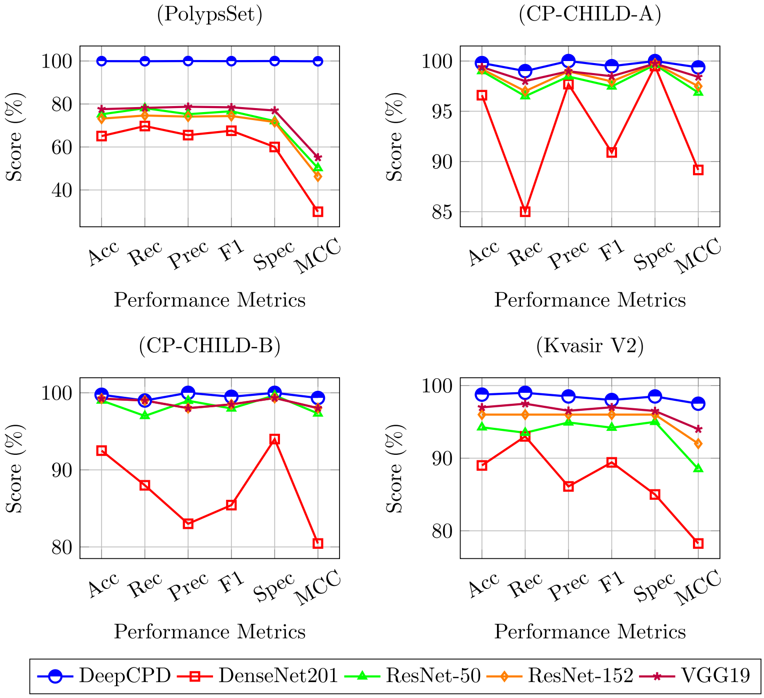


Fig. 7 Visual analysis of DeepCPD performance against standard deep learning models using various performance metrics

computational efficiency and mitigate overfitting, followed by adding two fully connected layers to improve the model's performance. The performance of different models exhibited minimal variations across all experiments. For instance, VGG19 achieved a higher accuracy of 77.59%, with a Recall of 78.18%, a Precision of 78.70%, a F1-score of 78.44%, a specificity of 76.95%, and a MCC of 55.11% on the PolypsSet dataset. On CP-CHILD-A, VGG19 achieved a higher accuracy of 99.40%, with a Recall of 98.00%, a Precision of 98.98%, a F1-score of 98.49%, a Spec of 99.75, and a MCC of 98.12%. On CP-CHILD-B, VGG19 achieved an accuracy of 99.25%, with a Recall of 99.00%, a Precision of 98.01%, a F1-score of 98.50%, a specificity of 99.33%, and a MCC of 98.01%. On Kvasir V2, VGG19 achieved an accuracy of 97.00%, Recall of 97.50%, Precision of 96.53%, a F1-score of 97.01%, a specificity of 96.50%, and a MCC of 94.00%. On the other hand, all other models showed relatively lower performance. In contrast, the DeepCPD model achieved an accuracy of 99.93%, 99.80%, 99.75%, and 98.75% on the PolypsSet, CP-CHILD-A, CP-CHILD-B, and Kvasir V2 datasets, respectively.

5.6 Discussion

CRC is a significant global health concern, ranking as one of the most prevalent cancers and a leading cause of cancer-related mortality. CRC often remains asymptomatic in its early stages, leading to late-stage diagnoses when treatment options may be less effective since making early detection and removal of these polyps crucial for preventing the progression of cancer. Diagnosis through traditional methods, such as visual examination during colonoscopy, is subjective and depends heavily on the physician's expertise. This subjectivity can lead to inter-observer variability and increase the risk of missed or delayed diagnoses. The inherent variability in colorectal images further complicates accurate diagnosis. Conventional diagnostic approaches are often time-consuming, requiring careful examination of images and data. More efficient and precise diagnostic methods are critical for timely intervention. These challenges make it imperative to explore advanced technologies for more accurate detection.

This study designed a colorectal polyp detection model, DeepCPD, for colorectal image classification that adopts ViT as its underlying architecture, wherein a linear multihead self-attention mechanism replaces the initial multihead self-attention mechanism of ViT. The model undergoes various data augmentation and preprocessing techniques to enhance the dataset. DeepCPD's capacity for global feature extraction through LMSA enables the capture of extensive dependencies and inherent relationships among pixels in images. This capability makes it particularly adept at handling the intricate patterns of colorectal polyps. Furthermore, the integrated LMSA mechanism efficiently mitigates computational complexity, reducing it from $O(n^2)$ to $O(n)$ [20]. The model demonstrated strong performance when utilizing specific hyperparameters, including the Adam optimizer with a batch size of 32, a learning rate of 2e-5, momentum values of beta1 (0.9), beta2 (0.999), and an epsilon of 1e-8. Notably, the model showcased superior performance when trained with a learning rate of 2e-5 in comparison to 2e-6 and 2e-7. Additionally, employing a batch size of 32 yielded better results compared to a batch size of 64.

DeepCPD exhibited remarkable performance, achieving an accuracy exceeding 98.05%, recall exceeding 98.10%, and precision exceeding 97.71%, while achieving a training time improvement of over 1.2x across all datasets. The model showcased superior diagnostic accuracy compared to other CNN-based approaches, as evidenced by its exceptional results on the four datasets. Although the experimentally established model has good results for the diagnosis of colon cancer, there are still some limitations. The interpretability of DeepCPD, especially the complex architecture of ViT, needs careful consideration. Understanding the decision-making process is essential to gain trust in clinical settings. Another challenge faced in this study and other medical research involves accessing labeled images for medical images. Future research should emphasize enhancing the interpretability of ViT by developing methods that enable clinicians to understand and trust the decision-making process of the models. To overcome the scarcity of labeled medical images, generating new and diverse datasets is essential, employing techniques like Few-shot fine-grained action [52, 53]. Besides, this study needs further improvement, such as to perform multiclassification for predicting various stages of polyps. Collectively, these aspects contribute to certain challenges in diagnosing and treating CRC. To address the mentioned limitations, we will continue to investigate them in subsequent studies.

6 Conclusion

In this study, a novel deep learning colorectal polyp detection model, DeepCPD, is designed using a transformer-based linear multihead self-attention mechanism combined with data augmentation to classify the colonoscopy images into polyp vs. non-polyp or hyperplastic vs. adenoma classes based on the dataset. The linear multihead self-attention mechanism efficiently extracts pertinent features from the input image, facilitating the model to represent intricate relationships and patterns within the data. It enables the model to assign attention weights to individual image patches, capturing extensive dependencies between pixels and offering insights into the image patches that play a significant role in the model's decision-making process. This is crucial for understanding the global context and relationships of different parts within the image. In contrast to non-linear multihead self-attention mechanisms, linear multihead self-attention offers a notable reduction in computational complexity, thereby elevating computational efficiency. This efficiency is notably reflected in training times, showcasing an improvement of over 1.2x in speed across all four datasets. DeepCPD was assessed across four diverse datasets-PolypsSet, CP-CHILD-A, CP-CHILD-B, and Kvasir V2-yielding exceptional recall and precision results. Specifically, it achieved a recall and precision of 99.87% and 99.94% for PolypsSet, 98.10% and 99.89% for CP-CHILD-A, 98.20% and 99.59% for CP-CHILD-B, and 98.40% and 97.71% for Kvasir V2, respectively, outperforming other approaches already in use. Similarly, The several experiments undertaken as part of this study for the comparative analysis utilizing competing state-of-the-art CNN methods consistently indicate that DeepCPD surpasses state-of-the-art deep learning techniques across all performance metrics on the four datasets. The efficiency and generalizability of the proposed DeepCPD model were substantiated through performance evaluations on four distinct datasets, considering metrics such as accuracy, recall, precision, F1-score, specificity, MCC, confusion matrix, and AUC-ROC. Notably, the experiment results underscored the efficacy of the proposed model even in scenarios involving imbalanced datasets. Automating the classification process through deep learning models, such as DeepCPD, helps alleviate inter-observer variability, ensuring consistent and reliable assessments; this is especially critical for reducing false negatives and facilitating timely interventions. Such models can support healthcare professionals in precisely identifying and characterizing polyps during routine screenings.

Acknowledgements This research was supported by the University Grants Commission of India under UGC-JRF grant 3640/(NET- JULY 2018), and the computing resources are provided by C- DAC and National Institute of Technology Tiruchirappalli.

Author Contributions All authors equally contributed to this study.

Funding This study was funded by the University Grants Commission (UGC) of India under UGC-JRF grant 3640/(NET-JULY 2018).

Availability of data and materials Public datasets are used for this study.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethics approval The authors of this study have not engaged in any research involving human subjects or animals.

Consent to participate All individual participants included in the study provided informed consent.

Consent for publication The participant has given consent for the submission of the case report to the journal.

References

1. Cancer IA Cancer Today. https://gco.iarc.fr/. Accessed 15 Nov 2022
2. international W Colorectal cancer statistics. https://www.wcrf.org/cancer-trends/colorectal-cancer-statistics/. Accessed 15 Nov 2022
3. Chen H, Li C, Li X, Rahaman MM, Hu W, Li Y, Liu W, Sun C, Sun H (2022) Huang X et al Il-mcam: an interactive learning and multi-channel attention mechanism-based weakly supervised colorectal histopathology image classification approach. Computers in Biology and Medicine 143:105265
4. Society AC American Cancer Society Guideline for Colorectal Cancer Screening. https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/acs-recommendations.html. Accessed 15 Nov 2022
5. Tan J, Gao Y, Liang Z, Cao W, Pomeroy MJ, Huo Y, Li L, Barish MA, Abbasi AF, Pickhardt PJ (2019) 3D-GLCM CNN: a 3-dimensional gray-level co-occurrence matrix-based cnn model for polyp classification via ct colonography. IEEE Transactions on Medical Imaging 39(6):2013–2024
6. Nguyen H-G, Blank A, Lugli A, Zlobec I (2020) An effective deep learning architecture combination for tissue microarray spots classification of h&e stained colorectal images. In: 2020 IEEE 17th International symposium on biomedical imaging (ISBI), IEEE, pp 1271–1274
7. Pacal I, Karaboga D, Basturk A, Akay B, Nalbantoglu U (2020) A comprehensive review of deep learning in colon cancer. Computers in Biology and Medicine 126:104003
8. Solak A, Ceylan R (2023) A sensitivity analysis for polyp segmentation with u-net. Multimed Tools Appl 1–29
9. Jha D, Smedsrud PH, Johansen D, Lange T, Johansen HD, Halvorsen P, Riegler MA (2021) A comprehensive study on colorectal polyp segmentation with resunet++, conditional random field and test-time augmentation. IEEE journal of biomedical and health informatics 25(6):2029–2040
10. Tasnim Z, Chakraborty S, Shamrat F, Chowdhury AN, Nuha HA, Karim A, Zahir SB, Billah MM et al (2021) Deep learning predictive model for colon cancer patient using cnn-based classification. Int J Adv Comput Sci Appl 12
11. Lorenzovici N, Dulf E-H, Mocan T, Mocan L (2021) Artificial intelligence in colorectal cancer diagnosis using clinical data: non-invasive approach. Diagnostics 11(3):514
12. Younas F, Usman M, Yan WQ (2022) A deep ensemble learning method for colorectal polyp classification with optimized network parameters. Appl Intell 1–24
13. Xie X, Xing J, Kong N, Li C, Li J, Zhang S (2017) Improving colorectal polyp classification based on physical examination data—an ensemble learning approach. IEEE Robot Automat Lett 3(1):434–441
14. Chou Y-C, Chen C-C (2023) Improving deep learning-based polyp detection using feature extraction and data augmentation. Multimedia Tools and Applications 82(11):16817–16837
15. Younas F, Usman M, Yan WQ (2023) An ensemble framework of deep neural networks for colorectal polyp classification. Multimedia Tools and Applications 82(12):18925–18946
16. Fang Y, Zhu D, Yao J, Yuan Y, Tong K-Y (2020) Abc-net: area-boundary constraint network with dynamical feature selection for colorectal polyp segmentation. IEEE Sensors Journal 21(10):11799–11809

17. Chandan S, Mohan BP, Khan SR, Bhogal N, Ramai D, Bilal M, Aziz M, Shah AR, Mashiana HS, Jha LK et al (2021) Adenoma and polyp detection rates during insertion versus withdrawal phase of colonoscopy: a systematic review and meta-analysis of randomized controlled trials. Gastrointestinal endoscopy 93(1):68–76
18. Mesejo P, Pizarro D, Abergel A, Rouquette O, Beorchia S, Poincloux L, Bartoli A (2016) Computer-aided classification of gastrointestinal lesions in regular colonoscopy. IEEE transactions on medical imaging 35(9):2051–2063
19. Aggarwal AK (2023) Thermal imaging for cancer detection. Imaging and Radiation Research 6(1):2638
20. Wang S, Li BZ, Khabsa M, Fang H, Ma H (2020) Linformer: self-attention with linear complexity. arXiv:2006.04768
21. Alqudah AM, Alqudah A (2022) Improving machine learning recognition of colorectal cancer using 3d glcm applied to different color spaces. Multimedia Tools and Applications 81(8):10839–10860
22. Jheng Y-C, Wang Y-P, Lin H-E, Sung K-Y, Chu Y-C, Wang H-S, Jiang J-K, Hou M-C, Lee F-Y, Lu C-L (2022) A novel machine learning-based algorithm to identify and classify lesions and anatomical landmarks in colonoscopy images. Surgical Endoscopy 36(1):640–650
23. Koppad S, Basava A, Nash K, Gkoutos GV, Acharjee A (2022) Machine learning-based identification of colon cancer candidate diagnostics genes. Biology 11(3):365
24. Su Y, Tian X, Gao R, Guo W, Chen C, Chen C, Jia D, Li H, Lv X (2022) Colon cancer diagnosis and staging classification based on machine learning and bioinformatics analysis. Computers in biology and medicine 145:105409
25. Mulenga M, Kareem SA, Sabri AQM, Seera M (2021) Stacking and chaining of normalization methods in deep learning-based classification of colorectal cancer using gut microbiome data. IEEE Access 9:97296–97319
26. Sarwinda D, Paradisa RH, Bustamam A, Anggia P (2021) Deep learning in image classification using residual network (resnet) variants for detection of colorectal cancer. Procedia Computer Science 179:423–431
27. Mulenga M, Kareem SA, Sabri AQM, Seera M, Govind S, Samudi C, Mohamad SB (2021) Feature extension of gut microbiome data for deep neural network-based colorectal cancer classification. IEEE Access 9:23565–23578
28. Tang C-P, Chen K-H, Lin T-L (2021) Computer-aided colon polyp detection on high resolution colonoscopy using transfer learning techniques. Sensors 21(16):5315
29. Hsu C-M, Hsu C-C, Hsu Z-M, Shih F-Y, Chang M-L, Chen T-H (2021) Colorectal polyp image detection and classification through grayscale images and deep learning. Sensors 21(18):5995
30. Zhou C, Jin Y, Chen Y, Huang S, Huang R, Wang Y, Zhao Y, Chen Y, Guo L, Liao J (2021) Histopathology classification and localization of colorectal cancer using global labels by weakly supervised deep learning. Computerized Medical Imaging and Graphics 88:101861
31. Paladini E, Vantaggiato E, Bougourzi F, Distante C, Hadid A, Taleb-Ahmed A (2021) Two ensemble-CNN approaches for colorectal cancer tissue type classification. Journal of Imaging 7(3):51
32. Liew WS, Tang TB, Lin C-H, Lu C-K (2021) Automatic colonic polyp detection using integration of modified deep residual convolutional neural network and ensemble learning approaches. Computer Methods and Programs in Biomedicine 206:106114
33. Zhang R, Zheng Y, Mak TWC, Yu R, Wong SH, Lau JY, Poon CC (2016) Automatic detection and classification of colorectal polyps by transferring low-level cnn features from nonmedical domain. IEEE journal of biomedical and health informatics 21(1):41–47
34. Liu X, Li Y, Yao J, Chen B, Song J, Yang X (2019) Classification of polyps and adenomas using deep learning model in screening colonoscopy. In: 2019 8th International symposium on next generation electronics (ISNE), IEEE, pp 1–3
35. Nisha J, Gopi VP, Palanisamy P (2022) Automated colorectal polyp detection based on image enhancement and dual-path cnn architecture. Biomedical Signal Processing and Control 73:103465
36. Patel K, Li K, Tao K, Wang Q, Bansal A, Rastogi A, Wang G (2020) A comparative study on polyp classification using convolutional neural networks. PloS one 15(7):0236452
37. Lo C-M, Yang Y-W, Lin J-K, Lin T-C, Chen W-S, Yang S-H, Chang S-C, Wang H-S, Lan Y-T, Lin H-H et al (2023) Modeling the survival of colorectal cancer patients based on colonoscopic features in a feature ensemble vision transformer. Computer Med Imaging Graphics 107:102242
38. Mali MT, Hancer E, Samet R, Yıldırım Z, Nemati N (2022) Detection of colorectal cancer with vision transformers. In: 2022 Innovations in intelligent systems and applications conference (ASYU), IEEE, pp 1–6
39. Wang X, Yang S, Zhang J, Wang M, Zhang J, Yang W, Huang J, Han X (2022) Transformer-based unsupervised contrastive learning for histopathological image classification. Medical Image Analysis 81:102559

40. Hossain MS, Rahman MM, Syeed MM, Uddin MF, Hasan M, Hossain MA, Ksibi A, Jamjoom MM, Ullah Z, Samad MA (2023) Deeppoly: deep learning based polyps segmentation and classification for autonomous colonoscopy examination. IEEE Access
41. Zhang J (2023) Towards a high-performance object detector: insights from drone detection using vit and cnn-based deep learning models. In: 2023 IEEE International conference on sensors, electronics and computer engineering (ICSECE), IEEE, pp 141–147
42. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2021) An image is worth 16x16 words: transformers for image recognition at scale. In: International conference on learning representations
43. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inform Process Syst 30
44. Wang X, Girshick R, Gupta A, He K (2018) Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7794–7803
45. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: European conference on computer vision, Springer, pp 213–229
46. Kaur A, Chauhan APS, Aggarwal AK (2021) An automated slice sorting technique for multi-slice computed tomography liver cancer images using convolutional network. Expert Systems with Applications 186:115686
47. Li K, Fathan MI, Patel K, Zhang T, Zhong C, Bansal A, Rastogi A, Wang JS, Wang G (2021) Colonoscopy polyp detection and classification: dataset creation and comparative evaluations. Plos One 16(8):0255809
48. Bernal J, Tajkbaksh N, Sanchez FJ, Matuszewski BJ, Chen H, Yu L, Angermann Q, Romain O, Rustad B (2017) Balasingham I et al Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge. IEEE transactions on medical imaging 36(6):1231–1249
49. Bernal J, Sánchez J, Vilarino F (2012) Towards automatic polyp detection with a polyp appearance model. Pattern Recognition 45(9):3166–3182
50. Wang W, Tian J, Zhang C, Luo Y, Wang X, Li J (2020) An improved deep learning approach and its applications on colonic polyp images detection. BMC Medical Imaging 20:1–14
51. Pogorelov K, Randel KR, Griwodz C, Eskeland SL, Lange T, Johansen D, Spampinato C, Dang-Nguyen D-T, Lux M, Schmidt PT et al (2017) Kvasir: a multi-class image dataset for computer aided gastrointestinal disease detection. In: Proceedings of the 8th ACM on multimedia systems conference, pp 164–169
52. Zha Z, Tang H, Sun Y, Tang J (2023) Boosting few-shot fine-grained recognition with background suppression and foreground alignment. IEEE Trans Circuits Syst Video Technol
53. Tang H, Liu J, Yan S, Yan R, Li Z, Tang J (2023) M3net: multi-view encoding, matching, and fusion for few-shot fine-grained action recognition. In: Proceedings of the 31st ACM international conference on multimedia, pp 1719–1728

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.