Ingénierie des Systèmes d'Information
Vol. 30, No. 3, March, 2025, pp. 607-617
Journal homepage: http://iieta.org/journals/isi

Enhancing Colon Cancer Diagnosis Through Explainable Vision Transformer Architecture with Selective Kernel Learning and Multi-Objective Differential Evolution

N. Purushotham*, J. Avanija

School of Computing, Mohan Babu University, Tirupati 517102, India

Corresponding Author Email: nimmalanimmala10@gmail.com

https://doi.org/10.18280/isi.300305

Received: 4 September 2024
Revised: 13 December 2024
Accepted: 17 January 2025
Available online: 31 March 2025

Keywords:
histopathological images, Grad-CAM, SKL, MODE, EViT

ABSTRACT

Colon cancer diagnosis is critical and requires accurate deep learning models to help in clinical diagnosis. In this paper, an advanced deep learning model integrating Explainable Vision Transformer (EViT) architecture with Selective Kernel Learning (SKL) and Multi-Objective Differential Evolution (MODE) method has been proposed. The EViT model utilizes the capability of vision transformer by incorporating SKL method for dynamic selection of optimal kernels to enhance the feature extraction. This process also improves the identification of relevant patterns and structures in the tissue samples. Hyperparameter optimization is performed using MODE algorithm to improve the model's efficiency and interpretability. For preprocessing the input histopathological images of colon cancer non-local mean filtering method is used enabling precise path extraction and feature embeddings. By incorporating MODE algorithm for hyperparameter optimization balancing accuracy and reliability ensuring better efficiency of the model. The explainability of the model is facilitated by using Gradient Weighted Class Average Mapping (GRAD-CAM) and attention maps to improve decision making process. Experimentation has been carried out on histopathological images of colon cancer. EViT model performs better with an accuracy of 93.2%. The proposed EViT based model not only improves accuracy but also enhances the deeper understanding of pathology enhancing the clinical diagnosis of colon cancer.

1. INTRODUCTION

Colon cancer is observed as a prevalent cancer around the globe causing a major threat to the health globally [1]. Traditional method of such cancer diagnosis involves an invasive approach considering the histopathological analysis of biopsies. This approach is expensive and leads to inconsistencies. Methods such as colonoscopic screening used for predicting colorectal cancer leads to problems such as bleeding, colonic perforation and inability to detect all polyps [2], Early and proper diagnosis for such disease is important to improve patient outcomes and increase the survival rate. Survey reveals that the early stage diagnosis of colon cancer increases the survival rate of the patient by 90% [3]. However, for patients with cancer spread across different organs, the survival rate drops accordingly.

Deep learning models offer a better diagnosis for colon cancer using histopathological images by differentiating the normal and cancerous tissues. These models focus only on high accuracy for diagnosis, interpretability is also important to help the healthcare professionals to understand the predictions of the model to improve decision making. Recently deep learning architectures such as convolutional neural networks (CNNs) have been utilized in medical image analysis, particularly for diagnosis of critical diseases like colon cancer [4]. Nowadays, vision transformer (ViT) based deep learning architecture termed as a better alternative than CNN is used much in the Computer Vision domain [5]. The applications of deep learning methods in healthcare due to their "black box" nature lead to limitations such as lack of interpretability, limiting clinical adoption, and hindering trust in diagnosis [6]. This paper proposed a novel approach to enhance the diagnosis of colon cancer utilizing EViT architecture through features such as SKL and MODE. EViT is a modified version of the standard Vision Transformer architecture to capture long range dependencies in image data sing self-attention mechanism. SKL method utilizes learnable weights termed as gates in each kernel of the convolution layer in EViT architecture. This helps the model to suppress the channels that are less informative and improve the interpretability and efficiency. MODE is an evolutionary optimization method utilized to optimize the EViT architecture for objectives such as classification accuracy and interpretability. Also, developing deep learning models with interpretability leads to several benefits such as a better understanding of patterns and image features for improving classification accuracy. This process helps healthcare professionals generate valuable insights, facilitating improved diagnostics and better communication between patients and healthcare professionals.

Balancing factors such as accuracy and interpretability become a major challenge in medical diagnosis, especially

when using deep learning models. Performing optimization for accuracy alone can lead to black box models with a lack of transparency in predictions. Alternately, focusing on the optimization of interpretability will become a challenge for the model to handle complex patterns leading to compromise in the predictions. The proposed approach can overcome these problems by utilizing EViT architecture and MODE by formulating interpretability as an objective function. By addressing these issues, the proposed EViT based model can be used as an efficient tool for accurate diagnosis of colon cancer.

The major work of the proposed system includes the design of EViT architecture with SKL approach to improve the interpretability of colon cancer diagnosis. The optimization of EViT architecture is performed using MODE for classification accuracy and interpretability to address the multi-objective nature while performing a diagnosis of colon cancer. The proposed approach has been evaluated using colon cancer dataset with histopathological images. The results were promising with high accuracy compared to other state-of-art methods. By combining the ViT architecture with the technique to enhance interpretability and Multi objective optimization, the proposed work contributes for accurate diagnosis of colon cancer.

2. RELATED WORKS

The authors proposed deep learning models with residual networks for the detection of intestinal crypts using biopsy images. The work highlights the use of deep learning to perform tasks such as object detection within a medical context. The discussion was not clear in handling image variations [7]. The authors analyzed various images related to colon cancer diagnosis using deep learning methods [8]. A lightweight deep learning framework was introduced for colon cancer prediction in real time environment using portable endoscopy devices leading to faster diagnosis. The lightweight framework design has the possibility of compromising the accuracy of prediction [9].

A unified framework for interpreting prediction of the model has been proposed by the authors focusing on the explanation of the model's output using a single instance. The framework might not address the problem of generalizing the model for unseen data [10]. The authors demonstrated the need of transparency in machine learning for disease diagnosis applications. The implementation of these solutions in deep learning model was not discussed by authors [11]. The authors provided a comprehensive review of utilizing Explainable Artificial Intelligence (EAI) for healthcare applications. The challenge is to select the suitable method based on the applications [12].

The authors introduced Vision Transformer (ViT) architecture for recognizing images and achieved better results compared to Convolutional Neural Network based models and require only fewer computational resources [13]. The authors improved the training efficiency of transformers using mixed precision techniques. The computational cost of the ViT architecture has been addressed, but the alternate architectures efficient for image recognition were not explored [14]. The data efficient training procedure has been introduced for image transformers. The convolutional neural network with a distillation procedure has been proposed [15].

This paper provides an overview of the deep learning techniques with interpretability for medical image analysis [16]. The study [17] reviewed the explainable methods for vision transformers, the taxonomy for organizing them, and some application scenarios. This work proposes Gradient-weighted Class Activation Mapping (GRAD-CAM) for visualization of high-resolution images applied for image classification and captioning [18, 19], which provide an extensive survey of the EViT architecture and its applications in image classification, object detection and semantic segmentation [20].

Single objective optimization is a time-consuming process and does not provide an optimal solution, The authors provided a comprehensive survey on the multi objective optimization in deep learning for parameter optimization with several case studies [21, 22]. Deep learning models plays important role in extracting features from images. Multi objective optimization algorithm helps in finetuning the hyperparameters to increase the accuracy of the model [23]. Colon cancer prediction can be performed using deep learning architecture such as ResNet, EfficientNet and Convolutional Neural Network [24-29].

Recently Vision Transformer models have been used in medical image analysis considering its ability to outperform traditional convolutional models in various applications of medical image analysis [30]. ViT models used in breast cancer detection from mammography images, the model showed efficient performance in detecting malignant lesions [31]. The authors proposed a hybrid approach combining ViT and Convolutional neural network for better feature extraction. The authors applied ViT for skin lesion classification in order to handle complex dermatological images effectively [32]. Besides these applications the lack of interpretability remains a challenge in clinical adoption. Recent studies reveal that the methods such as GRAD-CAM and self-attention mechanism can be used to enhance the transparency of prediction in ViT models [33]. Considering these aspects, the proposed work utilizes EViT architecture with SKL and Multi Objective Differential Evolution for effective feature extraction leading to efficient diagnosis of colon cancer.

3. METHODOLOGY

This section outlines the stages used in developing a framework for colon cancer diagnosis, from data preparation to evaluation of the model.

3.1 Image preprocessing

Image preprocessing techniques are used to remove noise from images for better accuracy while performing classification tasks. Traditional filtering methods such as blurring and sharpening are not sufficient to perform classification for histopathological images. Considering this fact advanced filtering method termed Non-local Means filtering with patch similarity measure has been used to remove noise in histopathology image samples of colon cancer.

Input: Noisy Image (N), Patch Size(s)
Output: Denoised Image (M)
Algorithm 1:

1. Patch Extraction:
    - Image N divided into overlapping patches with

size (s×s)
- Each patch represented as $S_{j,k}$, j,k specifies pixel location

2. For each patch $S_{j,k}$
   - Search the image for similar patch (P) considering average intensity within a patch
3. Weight Calculation:
   - Assign a weight w to the similar patch (P) based on average intensity difference between $S_{j,k}$ and P. Higher similarity means higher weight
4. Weighted Averaging:
   - Calculate denoised pixel value $v_{j,k}$ by averaging the intensity of similar patch P, with weight w using formula specified in Eq. (1)

$$v_{j,k} = (\mathrm{w} * \mathrm{P}) + [(1 - \mathrm{w}) * I_{j,k}] \quad (1)$$

where, $I_{j,k}$ represents the intensity of noisy pixel at (j,k)

5. Image Reconstruction:
   - Repeat steps 2-4 for all pixels in the image

Combine denoised pixels $v_{j,k}$ to form denoised image (M)

The algorithm for image preprocessing using Non-local Means filtering is specified in Algorithm 1. This method performs a search over the entire image to identify similarity patches using parameters such as intensity, local features and texture and utilizes a weighted averaging approach to reconstruct the denoised image. This approach removes noise by preserving vital information such as tissue boundaries and improves the efficiency of colon cancer classification.

Histogram equalization is also performed to enhance the contrast to increase the visibility of the tissue features for the deep learning model during colon cancer classification. This process is performed by stretching the distribution of pixel intensities across the entire range of grayscale images. This process also addresses stain variations by normalizing the intensity distributions.

3.2 Colon cancer classification using EViT architecture

The proposed Enhanced Vision Transformer (EViT) architecture specified in Figure 1 addresses the challenge of traditional deep learning model being black box in nature lacking transparency. This helps in healthcare applications, particularly in cancer diagnosis to enhance interpretability. EViT architecture is the modified version of vision transformer utilized for the diagnosis of colon cancer through its capacity for hierarchical feature representation. This architecture uses self-attention mechanism to model long range dependencies across histopathological images. This enhancement reduces the computational complexity by utilizing sparse attention mechanism allowing effective processing of high dimensional data. This feature is introduced in token interactions in order to improve speed and memory efficiency during attention computation. Also, the hierarchical feature extraction process across various scales enables efficient localization of cancerours regions.

EViT architecture uses attention maps to address black box problem by providing explanations for their classification. These maps highlight the image regions that the model focuses allowing the healthcare professionals to understand the models decision in identifying whether cancerous or non-cancerous. Also, this architecture provides early detection insights for researchers by analyzing the image features that the EViT architecture prioritizes. This process helps to identify the visual patterns associated with colon cancer leading to better diagnosis.

The shortcomings of the model can also identify which helps to identify the areas of improvement or to address the biases in training data. Also, histopathological images are complex in nature to signify cancerous and non-cancerous samples which can be overcome by EViT architecture focusing on specific regions required in such scenarios. The tasks such as processing histopathological images and performing feature extraction are carried out by the Vision Transformer architecture. Metaformer block is used in the architecture to enhance the performance of transformer encoder. The method such as SKL has been utilized to enhance the feature extraction process concentrating on the informative region of the image.

The initial component specified in the EViT architecture is used to preprocess the input images using denoising method mentioned in Figure 1. Then normalization is applied to scale the images to a standard range to ensure the quality of image input and to improve convergence during training process. The denoised image of size h×w×c (height×width×channel), where channel is 3 for RGB, is considered for patch extraction and embeddings, discussed in detail in the sub-section (3.3).

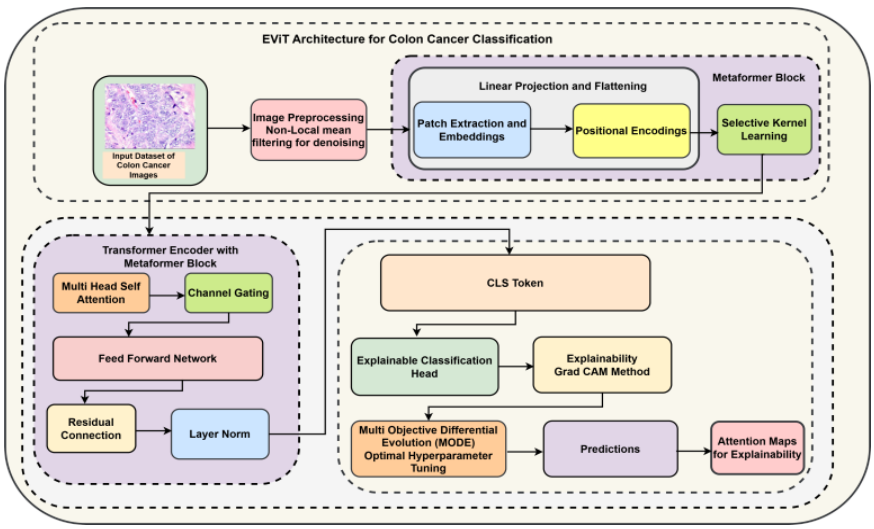

Figure 1. EViT architecture for colon cancer diagnosis

During this process patch size is defined for example, P×P, then a grid of non-overlapping images will be performed. Each patch is flattened using a vector of size P^2×c. During the process of embedding flattened patches into a format through linear projection that the transformer model can process. The CLS token is then added to the patch embeddings and then positional encoding is added to provide positional information. SKL is applied to the sequence of embeddings to select relevant kernels. This process selects the most relevant features dynamically based on the input data. This method is

The components of transformer encoder with metaformer block are specified in Figure 2. The patch embedding component divides the histopathological images into patches and converts them into lower dimensional vectors suitable for processing by transformer encoder. This process is performed through an embedding function preferably linear projection that divides image into patches. SKL uses pre-trained kernels to emphasize the image features related to colon cancer, particularly textures and shapes.

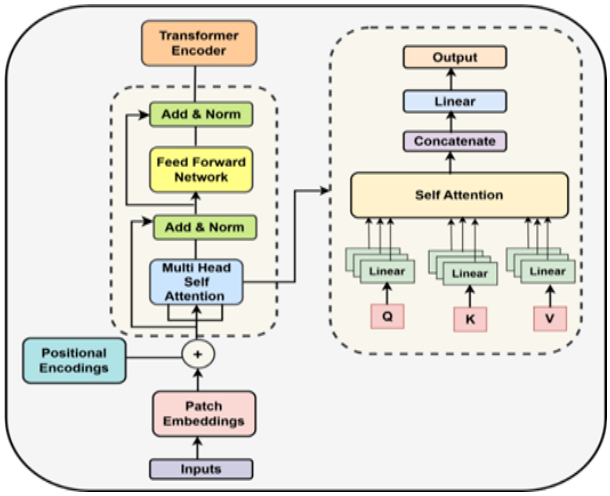

Figure 2. Transformer encoder architecture with metaformer block

Multi head self attention mechanism is used to capture the dependencies between tokens in the enhanced feature map. Also, it analyzes the image patches enhanced by the meta former block. During this process the patch embeddings transformed into query (Q), key (K) and value (V) matrices and then attention score is calculated as specified in Eqs. (2) and (3).

$$Q = Xw_q, K = Xw_k, V = Xw_v \tag{2}$$

$$Attention(Q,K,V) = softmax\left(\frac{Qk^T}{\sqrt{d_k}}\right)V \tag{3}$$

where, w_q, w_k, w_v forms learnable weight matrices for query, key and value, d_k represents the dimensionality of key vectors. Concatenation of attention heads is performed and linear transformation is applied as in Eq. (4).

$$MHA(Q,K.V) = concat(h_1, h_2 \dots h_h)w_y \tag{4}$$

where, h_i=*Attention*(Q_i,K_i,V_i) and w_y forms learnable output weight matrix. Channel gating mechanism is used to activate channels in the feature map. This mechanism is applied as in Eq. (5) to enhance the feature maps.

$$z_{gt} = \sigma\left(w_{gt}\mathrm{Z} + b_{gt}\right) \tag{5}$$

where, σ forms sigmoid function, w_{gt} and b_{gt} are learnable parameters and z_{gt} forms gating weights.

Feed forward network is used to apply non-linear transformations to the features as in Eq. (6). Residual connections and layer normalization are applied to stabilize training and improve gradient flow as in Eq. (7).

$$FFN(x) = ReLU(Xw_1 + b_1)w_2 + b_2 \tag{6}$$

where, w_1 and w_2 forms weight vectors and b_1 and b_2 form bias vectors of the feed forward network.

$$Y = \left(\frac{X - mean(X)}{stdv(X) + \epsilon}\right) * \gamma + \beta \tag{7}$$

where, *mean(X)* forms mean value of the input, *std(X)* forms the standard deviation of the input *X*, ϵ forms a small constant to avoid divide by zero error, γ, β forms learnable scaling and shifting parameters. The special token termed as class (CLS) token representing the entire sequence is used for classification task. Explainable classification head maps the class (CLS) token to output classes and then provides interpretability through linear transformations applied to the CLS token. Softmax function is used to produce probability distribution over classes using the technique Grad-CAM (Gradient weighted Class Activation Mapping). This method is a visualization technique used by the proposed EViT based model to make classification decisions. During training, the EViT architecture learns to extract features from images to differentiate from cancerous and healthy tissues.

Grad-CAM is then applied to a specific image to understand the models decision by focusing on the final layer of the model containing softmax function. During forward pass the image is divided into patches and positional encodings are added. Then SKL is applied to select appropriate features. The transformed embeddings passed through the transformer encoder with metaformer block using multi-head self attention, channel gating and feed forward network. The feature maps from SKL are selected to compute Grad-CAM through predicted class score with respect to these feature maps using Eq. (8).

$$\frac{dy^c}{dm^f} \tag{8}$$

where, y^c forms the class score for colon cancer and m^f specifies the feature map f for the selected convolution layer.

Then average the gradients to get importance of weights for each feature map f as specified in Eq. (9).

$$D_f^c = \frac{1}{z_i}\sum_i \frac{dy^c}{dm_{i,j}^f} \tag{1}$$

where, m_f^c specifies importance weight for feature map f and class c. z represents number of pixels in the feature map and $m_{i,j}^f$ represents element at position (i,j) in feature map m^f. Compute the weighted sum of the feature weight using importance weight as in Eq. (10), where G^c provides heatmap of class c, ReLU represents rectified linear unit activation.

$$G^c = ReLU(\textstyle\sum_k D_f^c m^f) \tag{10}$$

Upsample the heatmap to the original image size as in Eq. (11). This output reveals the important region in the image highlighting the abnormal tissues which helps in generating meaningful visual representation helping in colon cancer diagnosis.

$$GU^c = upsample(G^c) \tag{11}$$

3.3 Optimal kernel selection using SKL

SKL introduces the adaptive kernel selection mechanism into the model. This feature ensures that the model captures features across different spatial resolutions which are important for detecting patterns in colon cancer images. This method utilizes parallel convolutional pathways with varying sizes of kernel and integrates their outputs through attention mechanism in order to prioritize the most relevant scale. This mechanism focuses on features of interest considering the biological vision systems. Also, SKL enhances the ability of the model to detect fine-grained details and global patterns through feature aggregation process. SKL is used to dynamically combine different convolutional kernels to enhance feature representation. A kernel refers to a convolutional filter applied to a feature map to extract relevant features improving the learning process. Kernels are trained to detect features such as edges and textures. Multiple kernels are used to capture diverse features. The algorithm for SKL is specified in Algorithm 2.

Input: Feature Map (X) (h×w×c), Convolutional kernels $\{k_1,k_2\ldots,k_n\}$

Output: Feature Map (M)

Algorithm 2:

1. **Apply Convolutional Kernels**
   - Apply convolutional kernels k_i to feature map X to produce set of features X_i as mentioned in Eq. (12).

$$X_i = Conv(X, k_i)\ \forall i \in \{1,2,\ldots,n\} \tag{12}$$

2. **Global Information Aggregation**
   - Apply Global Average Pooling to each feature map X_i to produce global descriptor d_i as in Eq. (13).

$$d_i = GAP(X_i) = \frac{1}{hxw}\sum_{he=1}^{h}\sum_{we=1}^{w} X_i(h,w,c) \quad \forall c \in \{1,2,\ldots,n\} \tag{13}$$

3. **Concatenate Global Descriptors**
   - Concatenate the global descriptors from all feature map to form a single vector as in Eq. (14).

$$d = \text{concat}(d_1, d_2, \ldots d_k) \tag{14}$$

4. **Compute Importance Weights**
   - Pass the concatenated weight d through fully connected layers to derive importance weights w as in Eq. (15).

$$w = FC_2\left(ReLU\left(FC_1(d)\right)\right) \tag{15}$$

5. **Softmax Normalization**
   - Normalize the importance weights using softmax function to ensure they sum to 1 as in Eq. (16).

$$w_{soft,i} = \frac{\exp(w_i)}{\sum_{j=1}^{k}\exp(w_j)} \tag{16}$$

6. **Re-weight Feature Maps**
   - Multiply feature map by its normalized importance weights as in Eq. (17).

$$M_i = w_{soft,i} * X_i \tag{17}$$

7. **Combine Re-weighted Feature Maps**
   - Add the re-weighted feature maps to produce final output feature map as in Eq. (18). The dimension of output feature map will be h×w×c.

$$M = \sum_{i=1}^{k} M_i \tag{18}$$

3.4 Hyperparameter optimization using MODE

MODE method is utilized for the optimization of hyperparameters balancing multiple objectives like computational efficiency, accuracy and generalization. MODE method explores pareto-optimal solution space to ensure that the model achieves balance between competing performance metrics. This method uses differential evolution method by incorporating multi objective differential evolution in order to maintain diversity in solution space ensuring better optimization. MODE is used to optimize the hyperparameter and feature selection to balance multiple objectives like accuracy and interpretability. This method uses the approach of evolutionary algorithm to improve the candidate solutions. Also, this method maximizes the accuracy and minimizes the complexity of the model by balancing the multiple objectives and identifying the best tradeoffs. The hyperparameters of model building such as learning rate, patch size, number of layers can be optimized using this method to enhance the efficiency of EViT model. The histopathological images of colon cancer dataset used to train the EViT architecture with different hyperparameter configurations.

The hyperparameters to be optimized includes: patch size, transformer encoder layers, self-attention heads, learning rate and drop-out rate. This method is well suited for optimizing problems with multiple objective functions. This process involves generating and iterating candidate solutions by updating the population based on pareto dominance. The algorithm for MODE to optimize hyperparameters is specified in Algorithm 3.

Input:

Objective Functions: $O_1(x)$, $O_2(x)$,…$O_m(x)$

Population Size: N

Scaling Factor: S

Crossover Probability: CP

Stopping Criteria: Maximum number of generations $\mathbf{GE_{max}}$

Output: Pareto front,optimal parameter set for pareto front solutions

Algorithm 3:

Initialization

- Generate an initial population of size N with random parameter vectors as specified in Eq. (19).

$$P = (y_1, y_2, \dots, y_N) \quad (19)$$

Evaluation

- Evaluate each individual in the population for all objective functions as in Eq. (20).

$$f(y_i) = f_1(y_i), f_2(y_i), \dots f_m(y_i) \quad (20)$$

While stopping criteria not met

a. Mutation

- For each individual y_i generate mutant vector m_i as in Eq. (21).

$$m_i = y_{r1} + F.(y_{r2} - y_{r3}) \quad (21)$$

where r_1, r_2, r_3 form distinct random indices.

b. Crossover

- Create a trial vector v_i by combining m_i and y_i as in Eq. (22).

$$v_i = \begin{cases} m_{ij}, if\ rand_j \le Cp \\ y_{ij}, \quad Otherwise \end{cases} \quad (22)$$

c. Selection

- Evaluate the trial vector m_i as in Eq. (23).

$$f(m_i) = f_1(m_i), f_2(m_i), \dots f_m(m_i) \quad (23)$$

- Update population based on pareto dominance as in Eq. (24).

$$(v_i)^{c+1} = \begin{cases} m_i, if\ m_i\ dominates\ y_i \\ y_i, \quad Otherwise \end{cases} \quad (24)$$

d. Update Pareto Front

- Maintain a set of non dominated solutions representing current pareto front.

e. Increment generation counter as in Eq. (25).

$$c = c + 1 \quad (25)$$

Return

- The pareto front, set of non-dominated solutions

Optimal parameter set y for the solutions in the pareto front.

The proposed system has been evaluated using metrics like accuracy, precision, recall and F1-score. The accuracy measure is used to find the ratio of correctly classified samples indicating the overall correctness of the model as specified as in Eq. (26).

$$Accuracy = \frac{TRP + TRN}{TRP + TRN + FAP + FAN} \quad (26)$$

where, *TRP*, *TRN* represents true positives and true negatives, *FAP* and *FAN* represents false positives and false negatives. The metric precision representing the ratio of true positive prediction over all positive predictions measures how many of the predicted positive cases are correct as in Eq. (27).

$$Precision = \frac{TRP}{TRP + FAP} \quad (27)$$

The evaluation metric recall also termed as sensitivity provides the ratio of true positive predictions to all actual positive cases reflecting the model's ability to identify positive cases as in Eq. (28).

$$Recall = \frac{TRP}{TRP + FAN} \quad (28)$$

The metric F1-Score represents mean of precision and recall used to evaluate the performance when the dataset is imbalanced as in Eq. (29).

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (29)$$

4. RESULTS AND DISCUSSION

The performance of the proposed work has been evaluated using histopathological image dataset of colon cancer containing high resolution images of cancerous and non-cancerous tissues. The input image samples consist of histopathological images with two classes such as Colon-ACA and Colon-N. Colon-ACA termed as colon adenocarcinoma represents the cancerous tissues in colon lining. Colon-N represents the normal tissues without abnormalities. These images can be used to distinguish the normal and cancerous tissues. The dataset consists of 500 images in total with two classes of 250 images of Colon-N tissues and 250 images of Colon-ACA and then augmented to 10,000 images [29]. Figure 3 specifies the sample images of these two classes.

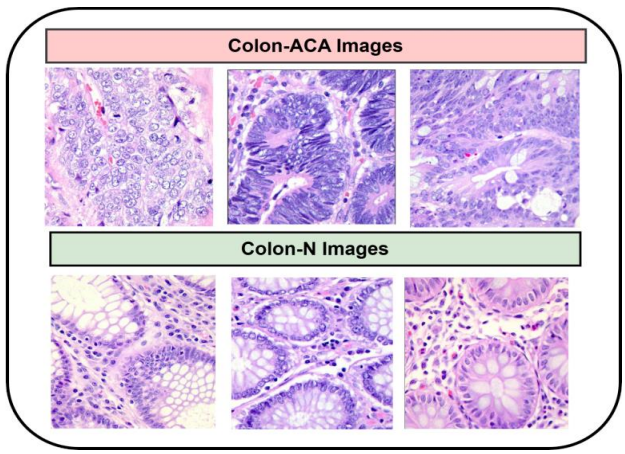

Figure 3. Sample histopathological images of colon cancer dataset

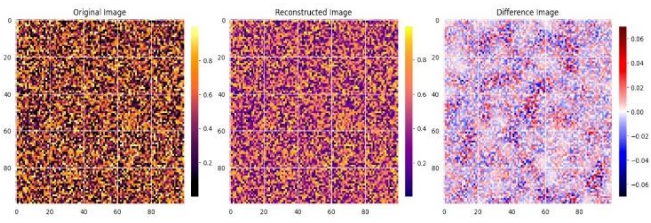

Figure 4. Reconstructed image after noise removal using non-local mean filtering

The input images were preprocessed with non-local mean filtering method to remove noise in the input image. This process enhances the performance of the feature extraction and classification process. The original image with noise and

reconstructed image using non-local mean filtering along with difference image highlighting the removal of noise during this process is specified in Figure 4.

After preprocessing the images were divided as smaller patches and embedded into lower dimensional space suitable for vision transformer. This process is to ensure that the local features are captured effectively. The patch extraction process is specified in Figure 5.

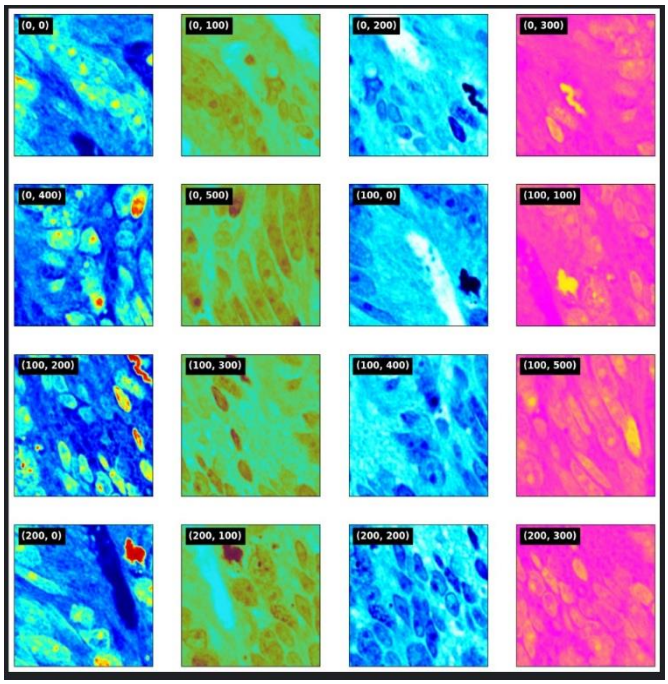

Figure 5. Patch extractions of colon cancer histopathological images

Table 1. Layered architecture of EViT for colon cancer diagnosis

No.	Layer (Type)	Output Shape	Param #
0	Conv2d-1	[-1, 768, 14, 14]	590592
1	Dropout-2	[-1, 197, 768]	0
2	MultiheadAttention-3	[[-1, 197, 768], [-1, 2, 2]]	0
3	LayerNorm-4	[-1, 197, 768]	1536
4	Linear-5	[-1, 197, 3072]	2362368
5	GELU-6	[-1, 197, 3072]	0
6	Linear-7	[-1, 197, 768]	2360064
7	Dropout-8	[-1, 197, 768]	0
8	LayerNorm-9	[-1, 197, 768]	1536
9	EViTLayer-10	[-1, 197, 768]	0
10	MultiheadAttention-11	[[-1, 197, 768], [-1, 2, 2]]	0
11	LayerNorm-12	[-1, 197, 768]	1536
12	Linear-13	[-1, 197, 3072]	2362368
13	GELU-14	[-1, 197, 3072]	0
14	Linear-15	[-1, 197, 768]	2360064
15	Dropout-16	[-1, 197, 768]	0
16	LayerNorm-17	[-1, 197, 768]	1536
17	EViTLayer-18	[-1, 197, 768]	0
18	MultiheadAttention-19	[[-1, 197, 768], [-1, 2, 2]]	0
19	LayerNorm-20	[-1, 197, 768]	1536
20	Linear-21	[-1, 197, 3072]	2362368
21	GELU-22	[-1, 197, 3072]	0
22	Linear-23	[-1, 197, 768]	2360064
23	Dropout-24	[-1, 197, 768]	0

The layered architecture of the proposed EViT is specified in Table 1. The type of layer include convolution layer for feature selection of initial input images and the drop-out layer is used for regularization. The multihead self attention mechanism is used to process the inputs and it is considered as a key component of the transformer. The layernorm component is used to normalize the activation between the layers. Fully connected layer is specified as linear and GELU (Gaussian error linear unit) activation function is utilized for non-linearity. EViT layers are the custom layers incorporating in the attention and feedforward mechanism. The numbers after each layer, e.g. ([-1, 197, 3072]) represents the shape of the output tensor for layer operation.

Gradient Weighted Class Activation Mapping (GRAD-CAM MAP) is utilized in deep learning to enhance the interpretability of convolutional neural networks and transformer architectures. It highlights the influential regions of the image which helps in decision making process of the model. This method computes the gradient score of the output class considering the feature maps of the convolutional layer. The proposed system uses Grad-CAM map to highlight the critical regions of the input images that contribute much in the colon cancer diagnosis by the model.

The Grad-CAM heatmap generated by the proposed system is specified in Figure 6. By highlighting the affected tissues that help in diagnosis, the output of the model becomes interpretable to healthcare professionals.

Including SKL with EViT architecture in the proposed system demonstrated improvement in the classification of colon cancer. This process is performed through the generation SKL feature map specified in Figure 7. This map provides insights about the utilization of the critical regions by the model to make predictions accurately.

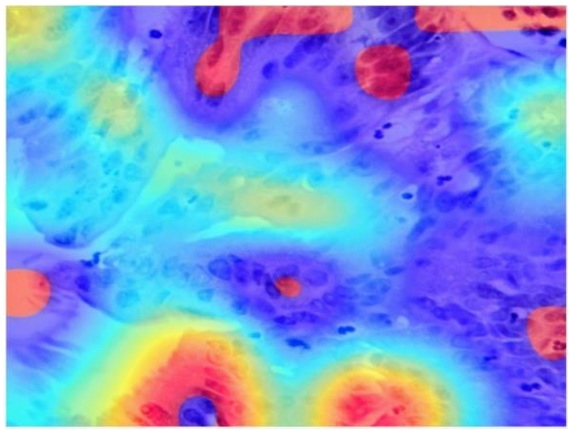

Figure 6. Grad-CAM map visualization for histopathological images of colon cancer

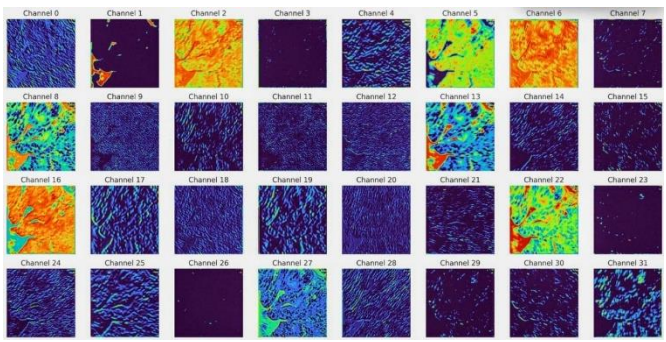

Figure 7. SKL feature map generated by the mode

The explainability component of the proposed system is included through the generation of attention map to highlight the critical regions to improve the accuracy of predictions. The attention map generated by the model specified in Figure 8 highlights the significant features such as anomalies of the cells indicating cancer. This component also contributes to the transparency in predicting the colon cancer.

The proposed EViT based model utilizes Multiobjective Differential Evolution (MODE) method to optimize the hyperparameters to enhance the models performance. MODE method effectively searches the hyperparameter space to identify optimal values to improve the accuracy of predictions. Table 2 specifies the range of hyperparameter and optimal values selected by MODE method.

The performance of the proposed system is measured using different metrics specified in Table 3 and the graph is specified in Figure 9.

The proposed EViT based model for colon cancer diagnosis achieves an accuracy of 93.2% which is better than the CNN, ResNet, EfficientNet, DenseNet and Vision Transformer models as specified in Table 3. The precision score of the proposed model is 92.4% which is higher than other existing architectures specified in Table 3. This represents that the proposed model reduces false positive rates efficiently which is important for medical diagnosis. The recall score of the proposed model is 91% which is better than other existing models specified in Table 3 indicating that the proposed model is effective in detecting colon cancer cases reducing the risk of incorrect diagnosis. The F1-Score of the proposed EViT model is also better highlighting the overall effectiveness of the proposed system in handling both positive and negative samples. The results specified in Table 3 reveal that the EViT based model with SKL and MODE algorithm performs well with an accuracy of 93.2% compared to existing state-of-art methods [24-28].

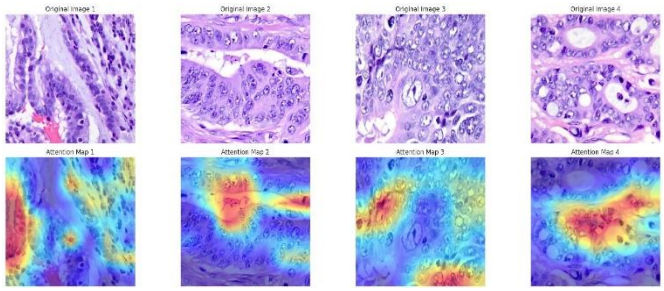

Figure 8. Attention map generated by the proposed model

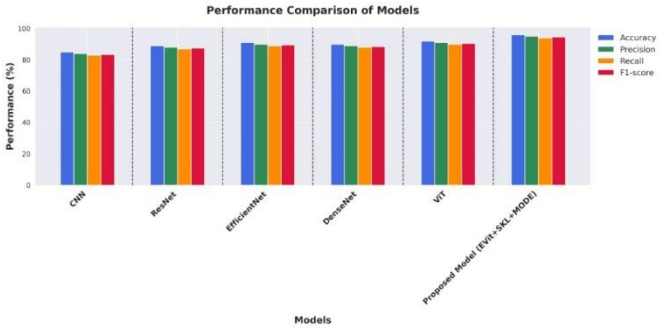

Figure 9. Performance metrics of the proposed model compared with existing methods

The confusion matrix of the proposed model revealing the accuracy of classifying cancerous and non-cancerous samples is specified in Figure 10. The enhanced accuracy and interpretability of the proposed model help medical practitioners for accurate diagnosis of colon cancer.

Table 2. Hyperparameters, possible and optimized values using MODE

Hyperparameter	Description	Possible Values / Range	Optimized Value
	EViT		
Batch Size Patch	Size of each image	{8, 16, 32}	16
Embedding Dimension	Dimension of the embedding space	{128, 256, 512, 768, 1024}	768
Number of Layers	Number of transformer layers	{6, 8, 10, 12, 14}	10
Number of Heads	Number of attention heads	{4, 8, 12, 16}	12
Feed-Forward Network Dimension	Dimension of the feed-forward network	{512, 1024, 2048, 3072, 4096}	3072
Dropout Rate	Dropout rate	{0.1, 0.2, 0.3, 0.4, 0.5}	0.1
Learning Rate	Learning rate for the optimizer	{0.0001, 0.0005, 0.001, 0.005, 0.01}	0.001
Batch Size	Number of samples per gradient update	{16, 32, 64, 128, 256}	64
Weight Decay	Regularization parameter	{0.00001, 0.0001, 0.001, 0.01, 0.1}	0.0001
	SKL		
Kernel Sizes	Sizes of the kernels used	{3, 5, 7, 9}	[3, 5, 7]
Number of Kernels	Number of different kernel sizes used	{1, 2, 3, 4}	3
	MODE		
Population Size	Number of individuals in the population	{20, 30, 40, 50, 60}	50
Crossover Probability	Probability of crossover	{0.6, 0.7, 0.8, 0.9, 1.0}	0.9
Differential Weight	Scaling factor for the differential mutation	{0.5, 0.6, 0.7, 0.8, 0.9, 1.0}	0.8
Number of Generations	Number of generations	{50, 75, 100, 125, 150}	100

Table 3. Performance of the proposed EViT+SKL+MODE model

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN	89	87	85	86
ResNet	91	90	88	89
EfficientNet	91.5	90.5	90	90.3
DenseNet	92	91	90.4	90.7
Vision Transformer- (ViT)	92.3	91	90.6	90.8
Proposed System (EViT-SKL-MODE)	93.2	92.4	91	91.7

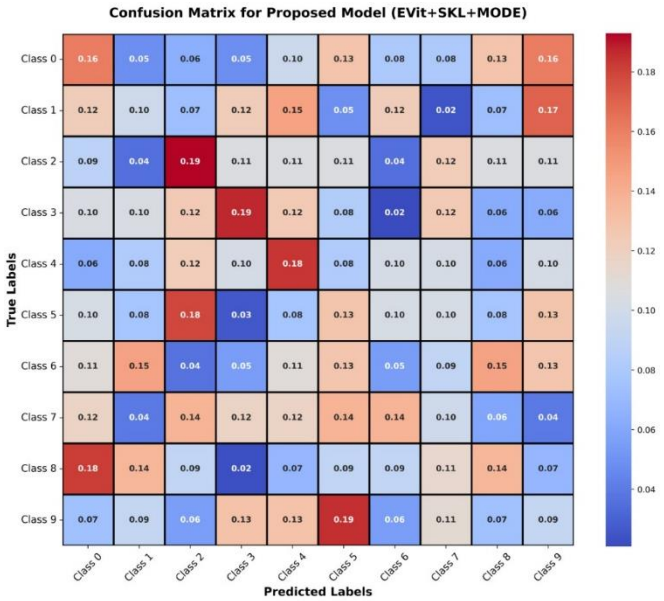

Figure 10. Confusion matrix for the proposed model

The proposed EViT model for colon cancer detection was evaluated through training and validation accuracy based on the number of epochs is specified in Figure 11. The graph reveals that the accuracy increases over the number of epochs. The consistent performance shows the effectiveness of the model.

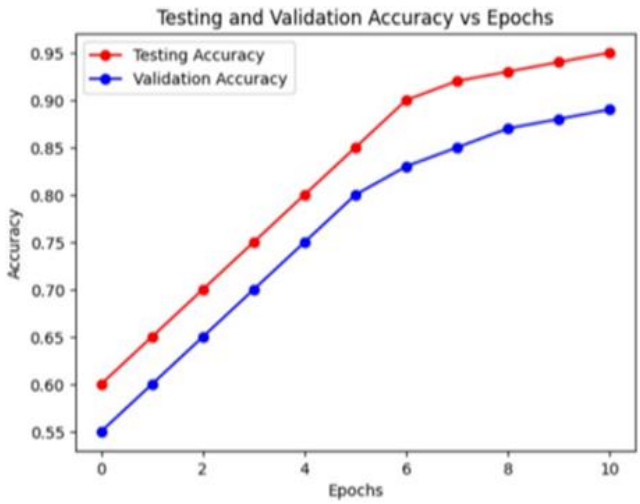

Figure 11. Testing and validation accuracy over number of epochs for the proposed system

The validation loss over the number of epochs specified in Figure 12 reveals that there was a gradual decrease in loss over epochs revealing the effective learning of the model. The loss curve not only reveals that the model is learning well but also generalizes effectively for new data. The performance of the proposed EViT based model is compared with existing methods for predicting colon cancer as specified in Figure 12. The results reveal that the proposed EViT+SKL+MODE method is better compared to the existing methods such as CNN,ResNet,DenseNet, EfficientNet and Vision transformer based models.

The Receiver Operating Characteristic (ROC) curve of the proposed system specified in Figure 13 evaluates the ability of the model to differentiate between cancerours and non-cancerous tissues. The curve plots the true positive rate and false positive rate across thresholds. The steeper curve indicates that the model performs well.

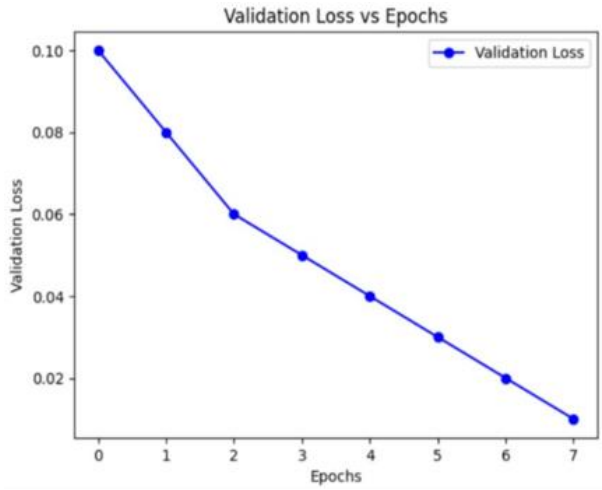

Figure 12. Validation loss of the proposed model over number of epochs

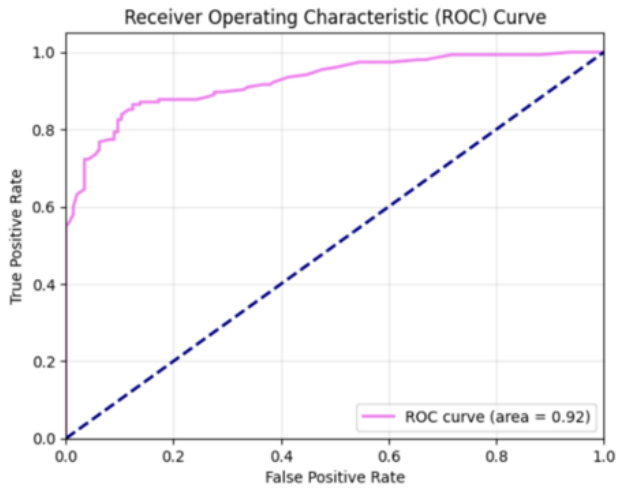

Figure 13. ROC curve for the proposed model

5. CONCLUSION AND FUTURE WORK

In this paper, the advanced model for prediction of colon cancer has been proposed by combining EViT architecture with SKL and MODE. The proposed methods show accuracy of accuracy of 93.2% outperforming the state-of-art methods. The utilization of SKL helps the model select optimal kernels for feature extraction, enhancing the capability of the model to identify important features. The MODE algorithm optimizes the hyperparameters specified in Table 2 ensuring the performance of the model by balancing between accuracy and interpretability. Also, the use of GRAD-CAM and Attention map generated by the proposed model specified in Figure 9 provides valuable insights making the model transparent and interpretable for clinical applications. Further, the utilization of advanced image pre-processing techniques like non-local mean filtering with patch similarity measure enhanced the image quality and feature extraction contributing to the better performance of the model. The model's explainability feature helps medical practitioners to identify the critical regions that contribute much to the occurrence of colon cancer. This

process facilitates better decision-making in clinical environment. Also, the proposed work has an impact on patient outcomes through early and accurate diagnosis of colon cancer. The significance of the proposed work extends beyond the technical advancements on clinical practices. The proposed model provides a reliable tool to healthcare providers for making informed diagnoses to improve patient outcomes. The innovation of the proposed work lies in the explainability features through GRAD-CAM and attention maps. These methods offer transparency by highlighting the regions of interest which helps medical practitioners to gain deeper insights into the diagnosis of colon cancer. The adoption of preprocessing methods such as non-local mean filtering with patch similarity enhances feature extraction and image clarity improving the performance of the proposed model.

Future work will focus on including more diverse samples and further larger datasets across different population considering histopathological image variations. There is also a potential to explore other types of cancers with the same architecture considering different image modalities and datasets. Implementing continuous learning mechanism will help the model to update and improve with new data, maintaining better performance. There will be a challenge deploying the model in real-time clinical settings considering the optimization of speed and efficiency without compromising the accuracy of the model. Integrating multimodal data such as histopathological images with genetic information has the possibility of enhancing the accuracy of diagnosis providing deeper insights about the disease.

REFERENCES

[1] Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 71(3): 209-249. https://doi.org/10.3322/caac.21660

[2] Navarro, M., Nicolas, A., Ferrandez, A., Lanas, A. (2017). Colorectal cancer population screening programs worldwide in 2016: An update. World Journal of Gastroenterology, 23(20): 3632. https://doi.org/10.3748/wjg.v23.i20.3632

[3] American Cancer Society.Colorectal Cancer Stages. https://www.cancer.org/cancer/types/colon-rectal-cancer/detection-diagnosis-staging.html.

[4] Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.A. (2016). Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE Transactions on Medical Imaging, 36(4): 994-1004. https://doi.org/10.1109/TMI.2016.2642839

[5] 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 16×16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929

[6] Lipton, Z.C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, 16(3): 31-57.

[7] Wei, J.W., Wei, J.W., Jackson, C.R., Ren, B., Suriawinata, A.A., Hassanpour, S. (2019). Automated detection of celiac disease on duodenal biopsy slides: A deep learning approach. Journal of Pathology Informatics, 10(1): 7. https://doi.org/10.4103/jpi.jpi_87_18

[8] Hamida, A.B., Devanne, M., Weber, J., Truntzer, C., et al. (2021). Deep learning for colon cancer histopathological images analysis. Computers in Biology and Medicine, 136: 104730. https://doi.org/10.1016/j.compbiomed.2021.104730

[9] Huang, Z., Wang, Z., Chen, J., Zhu, Z., Li, J. (2020). Real-time colonoscopy image segmentation based on ensemble knowledge distillation. In 2020 5th International Conference on Advanced Robotics and Mechatronics (ICARM), Shenzhen, China, pp. 454-459. https://doi.org/10.1109/ICARM49381.2020.9195281

[10] Lundberg, S., Lee, S.I. (2017). A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874. https://doi.org/10.48550/arXiv.1705.07874

[11] Hull, R., Francies, F.Z., Oyomno, M., Dlamini, Z. (2020). Colorectal cancer genetics, incidence and risk factors: In search for targeted therapies. Cancer Management and Research, 12: 9869-9882. https://doi.org/10.2147/CMAR.S251223

[12] Nazar, M., Alam, M.M., Yafi, E., Su'ud, M.M. (2021). A systematic review of human-computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques. IEEE Access, 9: 153316-153348. https://doi.org/10.1109/ACCESS.2021.3127881

[13] Meng, L., Li, H., Chen, B.C., Lan, S., Wu, Z., Jiang, Y. G., Lim, S.N. (2022). Adavit: Adaptive vision transformers for efficient image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12309-12318.

[14] Zhuang, B., Liu, J., Pan, Z., He, H., Weng, Y., Shen, C. (2023). A survey on efficient training of transformers. arXiv preprint arXiv:2302.01107. https://doi.org/10.48550/arXiv.2302.01107

[15] Liu, P., Ji, L., Ye, F., Fu, B. (2023). GraphLSurv: A scalable survival prediction network with adaptive and sparse structure learning for histopathological whole-slide images. Computer Methods and Programs in Biomedicine, 231: 107433. https://doi.org/10.1016/j.cmpb.2023.107433

[16] Singha, A., Thakur, R.S., Patel, T. (2021). Deep learning applications in medical image analysis. In Biomedical Data Mining for Information Retrieval: Methodologies, Techniques and Applications, pp. 293-350. https://doi.org/10.1002/9781119711278.ch11

[17] Stassin, S., Corduant, V., Mahmoudi, S.A., Siebert, X. (2023). Explainability and evaluation of vision transformers: An in-depth experimental study. Electronics, 13(1): 175. https://doi.org/10.3390/electronics13010175

[18] 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. In 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, pp. 618-626. https://doi.org/10.1109/ICCV.2017.74

[19] Montavon, G., Samek, W., Müller, K.R. (2018). Methods

for interpreting and understanding deep neural networks. Digital Signal Processing, 73: 1-15. https://doi.org/10.1016/j.dsp.2017.10.011
[20] Patrício, C., Neves, J.C., Teixeira, L.F. (2023). Explainable deep learning methods in medical image classification: A survey. ACM Computing Surveys, 56(4): 85. https://doi.org/10.1145/3625287
[21] Karl, F., Pielok, T., Moosbauer, J., Pfisterer, F., et al. (2023). Multi-objective hyperparameter optimization in machine learning—An overview. ACM Transactions on Evolutionary Learning and Optimization, 3(4): 16. https://doi.org/10.1145/3610536
[22] Huang, W., Zhang, Y., Li, L. (2019). Survey on multi-objective evolutionary algorithms. Journal of Physics: Conference Series, 1288(1): 012057. https://doi.org/10.1088/1742-6596/1288/1/012057
[23] Sun, T., Jiao, L., Liu, F., Wang, S., Feng, J. (2013). Selective multiple kernel learning for classification with ensemble strategy. Pattern Recognition, 46(11): 3081-3090.
[24] Ranjan, A., Srivastva, P., Prabadevi, B., Sivakumar, R., Soangra, R., Subramaniam, S.K. (2024). Classification of colorectal cancer using ResNet and EfficientNet models. The Open Biomedical Engineering Journal, 18: e18741207280703. https://doi.org/10.2174/0118741207280703240111075752
[25] Alawi, A.E.B., Bozkurt, F. (2023). CNN-based colon cancer recognition model. In 2023 3rd International Conference on Emerging Smart Technologies and Applications (eSmarTA), Taiz, Yemen, pp. 1-5. https://doi.org/10.1109/eSmarTA59349.2023.10293562
[26] Yadav, K., Tiwari, S., Jain, A., Alshudukhi, J. (2022). Convolution neural network based model to classify colon cancerous tissue. Multimedia Tools and Applications, 22: 37461-37476. https://doi.org/10.1007/s11042-022-13504-9
[27] Ragab, M., Eljaaly, K., Sabir, M.F.S., Ashary, E.B., Abo-Dahab, S.M., Khalil, E.M. (2022). Optimized deep learning model for colorectal cancer detection and classification model. CMC-Computers Materials & Continua, 71(3): 5751-5764. https://doi.org/10.32604/cmc.2022.024658
[28] Chang, X., Wang, J., Zhang, G., Yang, M., et al. (2023). Predicting colorectal cancer microsatellite instability with a self-attention-enabled convolutional neural network. Cell Reports Medicine, 4(2): 100914. https://doi.org/10.1016/j.xcrm.2022.100914
[29] Borkowski, A.A., Bui, M.M., Thomas, L.B., Wilson, C.P., DeLand, L.A., Mastorides, S.M. (2019). Lung and colon cancer histopathological image dataset (lc25000). arXiv preprint arXiv:1912.12142. https://doi.org/10.48550/arXiv.1912.12142
[30] Takahashi, S., Sakaguchi, Y., Kouno, N., Takasawa, K., et al. (2024). Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review. Journal of Medical Systems, 48(1): 84. https://doi.org/10.1007/s10916-024-02105-8
[31] Umamaheswari, T., Babu, Y.M.M. (2024). ViT-MAENB7: An innovative breast cancer diagnosis model from 3D mammograms using advanced segmentation and classification process. Computer Methods and Programs in Biomedicine, 257: 108373. https://doi.org/10.1016/j.cmpb.2024.108373
[32] Saha, D.K., Joy, A.M., Majumder, A. (2024). YoTransViT: A transformer and CNN method for predicting and classifying skin diseases using segmentation techniques. Informatics in Medicine Unlocked, 47: 101495. https://doi.org/10.1016/j.imu.2024.101495
[33] Leem, S., Seo, H. (2024). Attention guided CAM: Visual explanations of vision transformer guided by self-attention. Proceedings of the AAAI Conference on Artificial Intelligence, 38(4): 2956-2964. https://doi.org/10.1609/aaai.v38i4.28077

NOMENCLATURE

P	Patch
w	Weight
Q, K, V	Query, Key, Value
X	Mean
G	Heatmap of Class c
GU	Upsample of Heatmap
GAP	Globa Average Pooling
d	Global Descriptor
FC	Fully Connected
P	Population
Z	Gating Weights
W,b	Learnable Parameter
FFN	Feed Forward Network
ReLU	Residual Connection
TRP,TRN	True Positive, True Negative
FAP,FAN	False Positive, False Negative

Greek symbols

σ	Sigmoid function
γ, β	Learnable scaling and shifting parameters
ϵ	Constant