scientific reports

Check for updates

OPEN

An efficient colorectal cancer detection network using atrous convolution with coordinate attention transformer and histopathological images

Majdi Khalid[1], Sugitha Deivasigamani[2], Sathiya V[3] & Surendran Rajendran[4]✉

The second most common type of malignant tumor worldwide is colorectal cancer. Histopathology image analysis offers crucial data for the clinical diagnosis of colorectal cancer. Currently, deep learning techniques are applied to enhance cancer classification and tumor localization in histopathological image analysis. Moreover, traditional deep learning techniques might loss integrated information in the image while evaluating thousands of patches recovered from whole slide images (WSIs). This research proposes a novel colorectal cancer detection network (CCDNet) that combines coordinate attention transformer with atrous convolution. CCDNet first denoises the input histopathological image using a Wiener based Midpoint weighted non-local means filter (WMW-NLM) for guaranteeing precise diagnoses and maintain image features. Also, a novel atrous convolution with coordinate attention transformer (AConvCAT) is introduced, which successfully combines the advantages of two networks to classify colorectal tissue at various scales by capturing local and global information. Further, coordinate attention model is integrated with a Cross-shaped window (CrSWin) transformer for capturing tiny changes in colorectal tissue from multiple angles. The proposed CCDNet achieved accuracy rates of 98.61% and 98.96%, on the colorectal histological image and NCT-CRC-HE-100 K datasets correspondingly. The comparison analysis demonstrates that the suggested framework performed better than the most advanced methods already in use. In hospitals, clinicians can use the proposed CCDNet to verify the diagnosis.

Keywords Colorectal cancer, Colorectal cancer detection network, Atrous convolution with coordinate attention transformer, Cross-shaped window transformer, Histopathology image

The digestive system consists of the colon, cecum, and rectum in addition to the large intestine. On the other hand, the colorectal is separated into ascending colon, transverse colon, and descending colon. One kind of cancer that develops in the colon of the large intestine is colorectal cancer (CLC)[1]. CLC is the second most prevalent cause of death worldwide, according to cancer statistics[2]. Radiation therapy has proven an efficient treatment for colorectal cancer in the last few years. Nevertheless, radiotherapy tolerance is a common problem among some patients, and it is the primary cause of radiotherapy failure and a poor prognosis for patients with colorectal cancer[3]. Therefore, early detection of cancer is essential to battling the disease and prolongs human life. It is a difficult task for researchers and physicians to detect colorectal cancer.

Medical imaging is an effective and important tool for early cancer detection[3]. Optical colonoscopy is a medical treatment that examines a variety of abnormalities on the colon's surface, including their position, shape, and pathological alterations for making a clinical diagnosis[4]. Histopathology image analysis (HIA) gives critical information for the clinical diagnosis of colorectal cancer. Even though accurately classifying pathological

[1]Department of Computer Science and Artificial Intelligence, College of Computing, Umm Al-Qura University, 21955 Makkah, Saudi Arabia. [2]Department of Computer Science and Engineering, University College of Engineering, A Constituent College of Anna University, Thirukkuvalai, Chennai, India. [3]Department of Computer Science and Engineering, Panimalar Engineering College, Chennai 600123, India. [4]Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai 602105, Tamil Nadu, India. ✉email: surendran.phd.it@gmail.com

images helps physicians determine the optimal course of treatment, the analysis of histopathological images takes a significant amount of time and effort, and the assessment of tissue categorization is simply influenced by numerous subjective parameters[5]. As a result, a more consistent and automated method based on computer-aided diagnosis (CAD) has recently grown in popularity and demand.

Machine learning, a subfield of artificial intelligence, is commonly utilized in biomedical applications for detecting cancer, classifying and segmenting the tumor[6–10]. Machine learning (ML) based CAD algorithms relied heavily on data, since they were mostly focused on feature descriptors. Feature descriptors was ineffective in the creation of CAD models due to its numerous flaws. To address these shortcomings, deep learning took the role of feature extraction. They can enhance the effectiveness of the CAD models[11]. In recent days, high-quality features are automatically mined from the input utilizing a deep learning model. It is an effective technique to identify numerous health problems[12–14].

Deep learning algorithms are eliminating the need of unambiguous feature specification; they analyse data and interpret high-dimensional features to produce a result. Convolutional neural networks (CNNs) show a significant part in the area of medical image processing[15,16]. Although spatial characteristics have been effectively retrieved using CNN-based algorithms, these methods have inherent drawbacks. One of these drawbacks is that their inability to accurately capture sequential features, particularly long-term interdependence. Furthermore, the constant size of the receptive fields limits their efficacy in retrieving short-range characteristics, since they would not be able to sufficiently seize precise and local fluctuations in the image.

The vision transformer (ViT) has motivated academics to approach image categorization using serial data to successfully seize long-range dependences[17]. A hierarchical Swin Transformer is suggested for the categorization of histopathology images in order to lower the computing complexity[18]. According to the literature, CNNs' constant receptive field size makes them poor at capturing long-range spatial properties, while they are excellent at retrieving local spatial information. It is commonly understood that atrous-convolution (AConv) computes feature by uniformly sampling data from the raw input. It can expand the receptive field and perform lossless feature processing[19]. Transformers have also exhibited an impressive capacity to understand the interrelationships between long-range features. Thus, merging these designs have the ability to improve feature learning through the analysis of short- and long-term dependences. The key contributions of this paper are provided as:

To improve diagnosis accuracy and preserve image features, a new Wiener based Midpoint weighted non-local means filter (WMW-NLM) is introduced.
To introduce a novel approach to colorectal tissue classification by integrating multiscale atrous convolution (MAConv) and a Cross-shaped window (CrSWin) transformer.
To catch small changes in colorectal tissues from several perspectives using coordinate attention driven patch division unit.
To test the robustness of the proposed CCDNet on the multiple class colorectal cancer tissues datasets using histopathological image

Deep learning in Colorectal Cancer (CRC) diagnosis: Colorectal cancer or CRC is one of the deadliest diseases that is being treated by deep learning through increasing accuracy and automating the process of screening and early detection. One of the best is its ability to utilize neural networks to analyze medical imagery and assimilate various forms of data, allowing for a detailed and tailored assessment of a patient's cancer diagnosis and treatment options. It minimizes false positive and negative findings; assists in forecasting; and advances medical study, which in turn enhances patient care and the delivery of healthcare services. With further development of the deep learning the role of this approach to the diagnosis of CRC or other diseases will only increase providing major breakthrough in medical diagnostics and treatment. The remaining part of the paper is organized as follows: Sect. 2 analyses the current literature on CLC detection using deep learning. Section 3 describes the proposed CCDNet in detail. Section 4 validates the proposed CCDNet through simulation. Section 5 completes the paper.

Related works

This section provides a quick summary of several recent colorectal cancer detection studies. Deep neural networks (DNNs) often require a vast quantity of data to do accurate analysis and classification. Ghosh et al.[20] proposed an Ensemble DNN for detecting tumors in colorectal histological images. To do this, three customized CNN architectures have been trained using two big datasets separately. The features were extracted using each architecture and the outcomes were concatenated to attain the final result. Khan et al.[21] proposed a CAD system for diagnosing Lymph Node Metastases in CLC based on ensemble and transfer learning network. Initially, the segmentation model (UNet) was trained for segmenting the lymph node tissue. Subsequently, Xception and ViT16, have been trained separately for the categorization of a positive and negative tissue.

Graham et al.[22] presented a multi-task learning method for segmenting and classifying the nuclei, glands, lumina and other tissue areas. This method makes use of data from several distinct data sources. Zidan et al.[23] proposed a Swin Transformer (ST) with up sampled network for segmenting histopathological images. They employed a hierarchical ST with shifted windows for extracting global contextual characteristics. The multiple scale feature extraction in a ST allowed the network to focus on various portions of the images at different sizes. To enhance feature concatenation, an encoder was used with a cascaded upsampling decoder.

Liang et al.[24] introduced a multiple scale feature fusion CNN using the shearlet transformation to detect histological images of CLC. This network extracted the shearlet coefficients at several decomposition scales. The network was also supplied with original pathological image. The input images were then processed via couples of convolution and pooling layers. After three pairs of convolution and pooling layers, the outputs of every channel were combined into the feature vector by the fully connected layers. The vector was then fed into 3 fully linked

layers to learn the features and perform classification task. Kumar et al.[25] proposed a lightweight CNN architecture for classifying the multi-class colorectal tissue histopathology images automatically. This model was known as a CLC classification CNN (CLCCN-Net). CLCCN-Net consisted of 5 convolutional layers, 5 pooling layers, 3 fully connected layers and a softmax layer. The utilization of three fully connected layers and fewer filters at each convolutional layer were the strengths of this network.

This investigation demonstrates that the majority of existing deep learning models are limited due to their huge sizes, several hyperparameters, long training time, and poor categorization results. Additionally, the approaches that made use of CNN models showed poor in the process of classifying histopathology images because they are limited with smaller receptive field and have incapability in seizing middle- and long-term dependences. Also, the authors Wang et al.[18] claimed that combining a CNN and a multiple scale ST architecture can improve histopathology image segmentation and classification performance. CNN retrieves local characteristics using convolutional processing, while Transformer captures global dependences via the interface of CNN-made tokens. However, typical swin transformers were unable to capture local contextual or semantic elements, resulting in ineffective utilization of spatial information. As a result, efficient fusion of local and global features in CNNs and transformers requires additional research for colorectal tissue categorization. This work proposes a cross-shaped window (CrSWin) transformer to address the limits of CNN's ability to capture middle and long-term dependencies. Furthermore, this CrSWin transformer is coupled with MAConv, which uses parallel atrous convolutions with different atrous rates to capture strong spatial characteristics with good discriminative capacities for colorectal tissues. To further improve the classification accuracy, coordinate attention is introduced for capturing small details from several perspectives.

Deep learning algorithms have shown promise in improving the accuracy and efficacy of colorectal cancer (CRC) detection through histopathology image analysis. Recent studies have explored the potential of these algorithms for clinical practice implementation, highlighting advancements in [specific examples, e.g., tumor classification, early detection]. Overcoming challenges such as [challenges, e.g., data quality, model interpretability] will be crucial for successful integration into routine clinical care[26].

Advances in deep learning have the potential to significantly enhance the accuracy and efficiency of colorectal cancer (CRC) diagnosis based on histopathology images. As demonstrated by Bousis et al. and Chlorogiannis et al.,[27] these algorithms can improve CRC detection rates by [percentage or other quantitative measure].

Proposed methodology

In this work, a new colorectal cancer detection network (CCDNet) is introduced using atrous convolution with coordinate attention transformer (AConvCAT) and histopathological images. In tissue images, the pre-processing stage is crucial for reducing different types of noise. Initially, the proposed CCDNet denoises the input histopathological images using Wiener based Midpoint weighted non-local means filter (WMW-NLM). After denoising the input, CCDNet uses data augmentation to mitigate overfitting. Then, the images are provided as input to the proposed AConvCAT for the categorization of colorectal tissues. The proposed AConvCAT framework consists of four modules, multiscale atrous convolution (MAConv), coordinate attention driven patch division unit (CA-PDU), Swin transformer, and classification. The goal is to properly acquire discriminative information from histopathology images. Figure 1 depicts the complete block of the proposed CCDNet.

Pre-processing and data augmentation stage

The gold standard for diagnosing cancer at the moment is histopathology image analysis. On the other hand, the presence of noise in histopathology creates a substantial problem and may result in misdiagnosis. In order to maintain image features and guarantee precise diagnoses, a new Wiener based Midpoint weighted non-local means filter (WMW-NLM) is presented. It can successfully eliminate noise from histological images to improve the diagnosis of colorectal cancer. Additionally, a significant amount of data is required for DL models in order to prevent over-fitting issues. The availability of more training data can improve the network generalization performance. In this work, data augmentation approaches are used to improve network's resistance to data fluctuations and ensure that our trained models can be applied to varied settings and histopathology images. This is accomplished by applying geometric transformation to the images (such as translation, scaling, and rotation). The data augmentation strategy expands the number of samples.

Wiener based midpoint weighted non-local means filter for denoising

In the non-local means (NLM) approach of image denoising, the weight of the center pixel is known as the midpoint weight (MW). The MW has a substantial influence on the effectiveness of NLM. In this study, the MW is calculated using a Wiener-like estimator. The observation model of the proposed WMW-NLM is expressed as given below:

$$I_N(m,n) = I_{NF}(m,n) + \eta(m,n) \tag{1}$$

where $I_N(m,n)$ denotes noisy image,$I_{NF}(m,n)$ represents the raw noise free image, and $\eta(m,n)$ defines the Gaussian noise with variance of σ^2. The observed data $I_N(m,n)$ is mapped into $U(m,n)$ by the spatially adaptive NLM filter in the following manner:

$$U(m,n) = \frac{\sum_{(m',n')\in R} I_{NF}\left(m',n'\right) W\left(m,n,m',n'\right)}{\sum_{(m',n')\in R} W(m,n,m',n')} \tag{2}$$

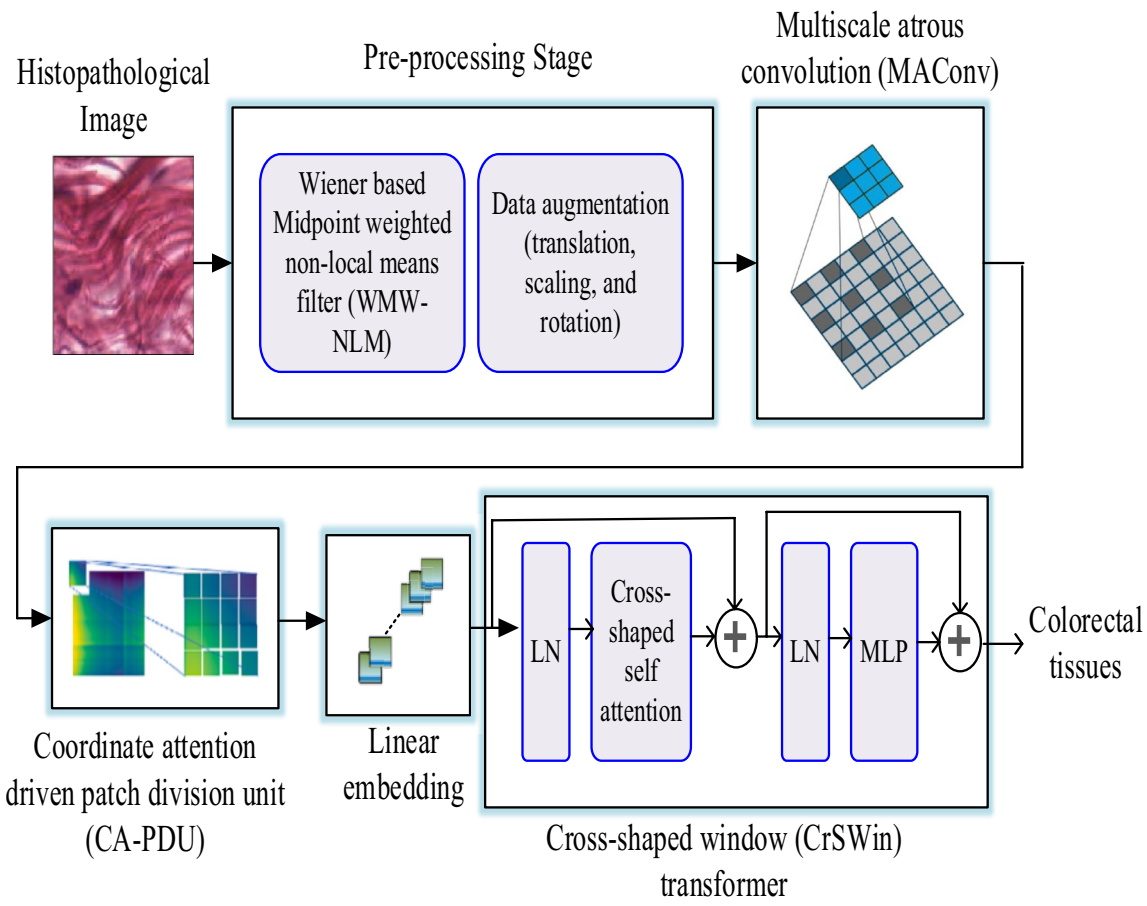

Figure 1. Proposed CCDNet architecture.

where R is a search window around (m, n) and $W\left(m, n, m', n'\right)$ is the weights. It is used for comparing the adjacent patches $\rho(m, n)$ and $\rho\left(m', n'\right)$ surrounding two pixels at (m, n) and $\left(m', n'\right)$, respectively. The midpoint weight is $W\left(m, n, m', n'\right)$ and it can be described as:

$$W\left(m, n, m', n'\right) = exp\left(-\frac{\|I_N(\rho(m, n)) - I_N\left(\rho\left(m', n'\right)\right)\|_{2,\sigma}^2}{f^2}\right) \tag{3}$$

where f represents the smoothing variable that controls the filtering level. $\|I_N(\rho(m, n)) - -I_N(\rho(m\prime, n\prime))\|_{2,\sigma}^2$ describes the weighted Euclidean distance of image patch $\rho(m, n)$ from patch $\rho(m\prime, n\prime)$ with size of $2\left(M_\rho + 1\right) \times 2\left(M_\rho + 1\right)$. It can be mathematically modelled as:

$$\|I_N(\rho(m, n)) - I_N(\rho(m\prime, n\prime))\|_{2,x}^2 = \sum_{i=-M_\rho}^{M_\rho} \sum_{j=-M_\rho}^{M_\rho} \zeta_\sigma\left(i, j\right)\left(I_N\left(m\prime + i, n\prime + j\right) - I_N\left(m + i, n + j\right)\right)^2 \tag{4}$$

where $\zeta_\sigma\left(i, j\right)$ denotes the Gaussian Kernel with variance $\sigma > 0$. Here, the (m, n) changes in the related window $R\prime$. According to Eq. (3), the center weights are always 1. In this work, the MWs are optimized by the use of a Wiener filter to achieve the excellent image processing quality. Let $m_n = (m, n), m_{n\prime} = (m\prime, n\prime), u_{m_n}^1 = U(m, n)$, $I_{m_n} = I_N(m, n), I_{m\prime_{n\prime}} = I_N(m\prime, n\prime)$ and $W_{m_n m\prime_{n\prime}} = W(m, n, m\prime, n\prime) \times \lambda_{m_n}$ is utilized for representing the MW. Consider $R_0 = R - (m, n)$ and (2) is simplified as:

$$u_{m_n}^1 - I_{m_n} = \frac{1}{\lambda_{m_n} + \sum_{m\prime_{n\prime} \in R_0} W_{m_n m\prime_{n\prime}}} \left(\sum_{m\prime_{n\prime} \in R_0} W_{m_n m\prime_{n\prime}}\left(I_{m\prime_{n\prime}} - I_{m_n}\right)\right) \tag{5}$$

Equation (5) indicates that the noise $u_{m_n}^1 - I_{m_n}$ is eliminated during the NLM filter operation. The noise that has been eliminated is calculated by multiplying two terms: $1/\left(\lambda_{m_n} + \sum_{m\prime_{n\prime} \in R_0} W_{m_n m\prime_{n\prime}}\right)$ and $\sum_{m\prime_{n\prime} \in R_0} W_{m_n m\prime_{n\prime}}\left(I_{m\prime_{n\prime}} - I_{m_n}\right)$. Here, λ_{m_n} is crucial for the denoising process. The amount of noise eliminated increases with decreasing λ_{m_n}. Hence, over smoothness is easily caused. Certain visual features are likely to be degraded or eliminated. For instance, an oversmoothed result will be produced by NLM with a center weight of 0. The eliminated noise decreases as λ_{m_n} increases. It shows that the denoising is not enough. For instance,

NLM yields the desired outcome when its midpoint weight is 1. The denoised image has a lot of residual noise. Thus, the midpoint weight need to appropriately fluctuate between 0 and 1. It is the initial rule of choosing the midpoint weight. Furthermore, smaller midpoint weights are selected to better remove noise while preserving more structures in the denoised image. Higher midpoint weights are assigned in the flat domain in order to remove a reasonable amount of noise without over smoothing. It is the second rule to determine the midpoint weight. According to these points, the proposed midpoint weight is expressed as:

$$\lambda_{m_n} = \frac{\tau f^2}{V_{m_n} + \tau f^2} \tag{6}$$

It is equivalent to Wiener filter. In contrast to previous Wiener filters in the gradient and wavelet domains, the suggested V_{m_n} is calculated as

$$V_{m_n} = max\left(q_{m_n} - \tau f^2, 0\right) \tag{7}$$

Here, the feature descriptor q_{m_n} is used to compute the features of image, and it is provided as follows:

$$q_{m_n} = \frac{1}{|R_0|} \sum_{(m',n') \in R_0} \|I_N(\rho(m,n)) - I_N\left(\rho\left(m',n'\right)\right)\|_{2,\sigma}^2 \tag{8}$$

and the constant τ is utilized for tuning the denoising limit in local Wiener filter. Equation (6) indicates that λ_{m_n} should be large for the region with smaller V_{m_n}. Hence (6) satisfies second rule. As a result, WMW-NLM can produce images with higher quality than those produced by NLM filters with different midpoint weights.

Atrous convolution with coordinate attention transformer

The proposed CCDNet uses AConvCAT for classifying the colorectal tissues. The four modules that make up the proposed AConvCAT framework are MAConv, CA-PDU, cross-shaped window (CrSWin) transformer and classification. The pre-processed histopathological image is expressed as $Z \in R^{h \times w}$ where $h \times w$ stands for the spatial dimension. For every pixel in the image, the class probability is represented as $X \in R^{1 \times c}$, where c is the number of tissue classifications. Initially, 2D patches are taken from the pre-processed histopathological image Z and fed into the network. The target pixel and its neighboring pixels make up each patch denoted as $P \in R^{u \times v}$ which allows for the extraction of pixel-level features, where $u \times v$ standing for window size.

However, the edge pixels cannot be caught while extracting a patch encircling a single pixel because of the limits of the patch extraction procedure. Consequently, a padding operation is performed on these pixels, and the recovered patches are then fed into the MAConv module. The MAConv module efficiently absorbs the deep spatial characteristics through the enlargement of the receptive arenas and preservation of the precise data. The discriminative representation is created by fusing features from several branches of MAConv. The features that are obtained after fusion are altered and put into an atrous convolution layer. The atrous convolution layer generates output feature maps and are separated into patch tokens utilizing CA-PDU. These tokens are converted into linear embeddings. Here, the coordinate attention (CA) block is used to produce a patch-by-patch embedding. The suggested CA-PDU may effectively enhance information by embedding location information into channel attention. It can also create long-range dependencies and thereby the features of blurred edges of histopathological images are detected precisely. The resulting embeddings are then fed to CrSWin transformer blocks for capturing the local and global characteristics, representing long-range dependences in an image. The acquired features from different CSWin transformer blocks are down sampled and transformed into a feature vector using a fully linked layer and global average pooling. At last the created feature vector is used to forecast the class labels for each pixel using a linear Softmax function. The architecture of the proposed AConvCAT based classifier is illustrated in Fig. 2.

Multiscale atrous convolution (MAConv)

Atrous convolution (AConv) produces convolutions through the injection of hovels or nil values into the CNN filters. As a result, it produces a bigger receptive field that resembles the result of combining pooling and convolution layers. Smaller kernels are frequently used in current convolutional networks to minimize parameters and optimize computation time. On the other hand, the dimension of kernel can be efficiently increased without raising the hyper-parameters by using AConv with a degree R. The suggested MAConv module uses different atrous degrees in each branch for having bigger and smaller views from the spatial space. This has been shown to be useful in producing vigorous and differentiable features. The objective of this multi-branch technique is to improve the feature fusion process and learn features at various scales. The spatial information is effectively included into the model through the use of parallel AConv with varied kernel sizes. After AConv, the worth of every neuron is represented as G_{ij}^{xy} ,where (x, y) is the neuron's location in two dimensions and j is the specific feature map on the i th layer, which is calculated as follows:

$$G_{ij}^{xy} = ReLU\left(\beta_{i,j} + \sum_{k} \sum_{l=0}^{A_i-1} \sum_{m=0}^{B_i-1} \omega_{ijk}^{lm} G_{(i-1)k}^{(x+lr)(y+mR)}\right) \tag{9}$$

Here, ReLU denotes the activation function, $\beta_{i,j}$ represents the bias, and k is the feature map positioned on $(i-1)$ layers and ω_{ijk}^{lm} denotes the weight parameters at the location (l, m). The definition of the atrous degree is R.

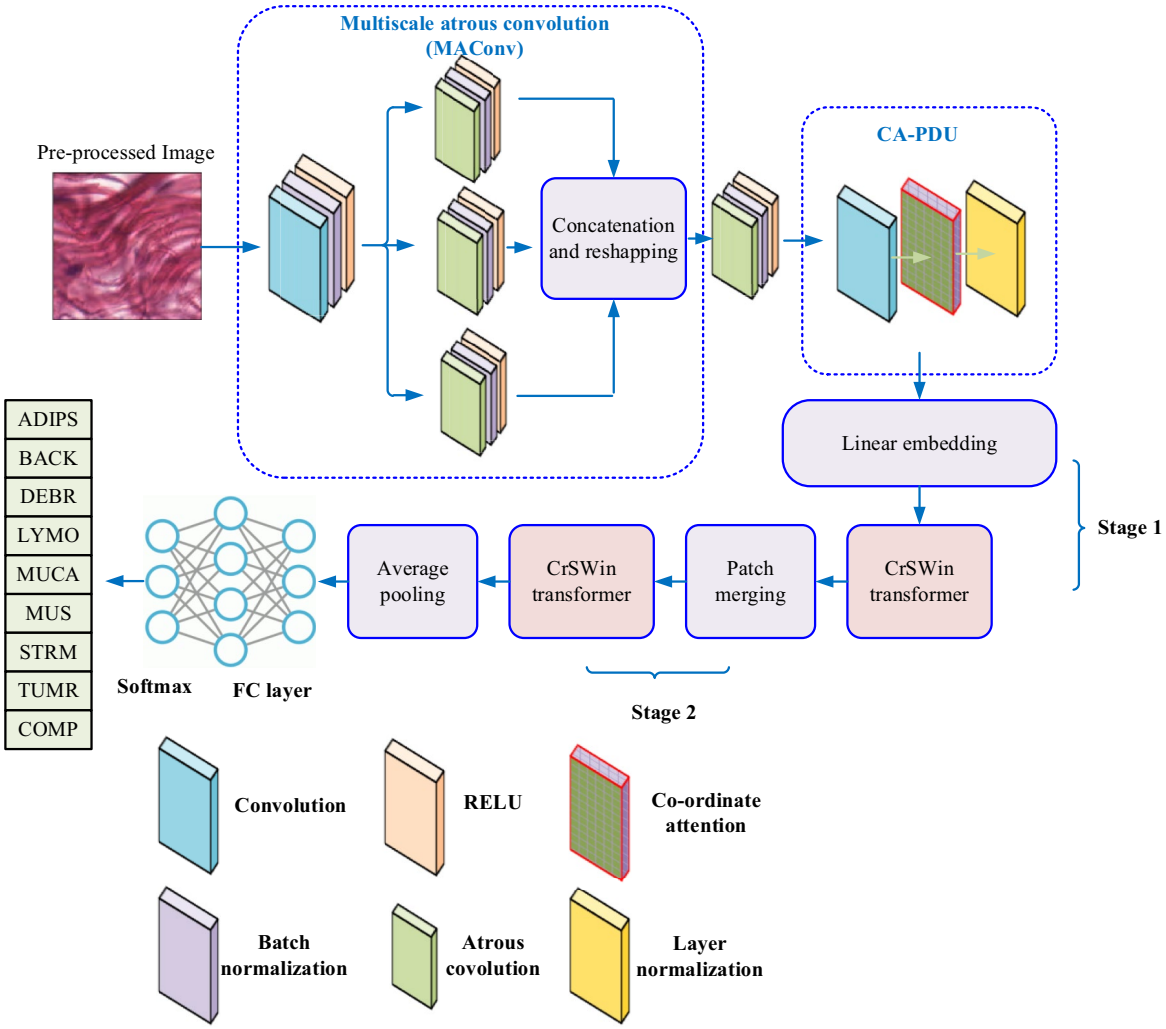

Figure 2. Proposed AConvCAT architecture.

Coordinate attention driven patch division unit (CA-PDU)

In AConvCAT framework, the output of the MAConv module is used as the input to the patch division unit (PDU) for obtaining patch tokens for transformer model. In this work, a new CA-PDU is introduced for the extraction of location information that is overlooked by all previous techniques. It embeds the position information in the histopathological image into the channel attention for enhancing the capability of the network to extract features. Additionally, this technique is used to solve the potential problem of information loss between patches. A 4×4 convolutional layer in CA-PDU divides the input feature map into non-overlapping, equal-sized patches. As seen in Fig. 3, the CA block averages pooling operations on the input features in the width and height directions, respectively.

As a result, the output of the c-th channel at height $h\prime$ and the width $w\prime$ can be expressed as

$$\widetilde{Z}(h\prime) = \frac{1}{w\prime} \sum_{0 \le i < w\prime} Z_c(h\prime, i) \tag{10}$$

$$\widetilde{Z}(w\prime) = \frac{1}{h\prime} \sum_{0 \le j < h\prime} Z_c(j, w\prime) \tag{11}$$

In this manner, the feature maps in both directions are obtained by aggregating the input features from the height and width directions. Also, a 1×1 convolution is utilized for compressing the number of channels and obtaining the feature map by performing the concatenation of spatial dimension for two multi-channel vectors. Batch normalization is then used to encode spatial information in both vertical and horizontal directions. The two directional eigenvectors, $\widetilde{Z}(h\prime)$ and $\widetilde{Z}(w\prime)$ can be reshaped to have the same number of channels as the input by using a 1×1 convolution. After that, the sigmoid function generates the outputs and it is ultimately weighted with the initial input data in both directions. The feature map is then converted into linear embeddings using

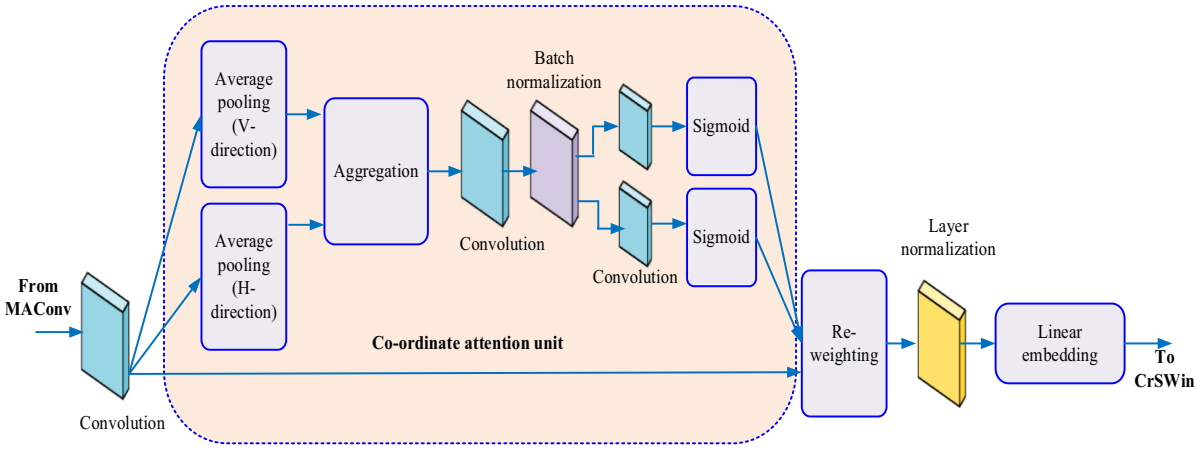

Figure 3. Architecture of Coordinate attention driven patch division unit.

patch partition. The CrSWin transformer blocks are then used to capture local and global features using the obtained embeddings as input.

Cross-shaped window (CrSWin) transformer

After linear embeddings, the proposed AConvCAT framework utilizes various stages of CrSWin Transform to generate a hierarchical feature representation. The Swin Transformer and the CrSWin Transform operate similarly. Unlike Swin Transformer, CrSWin Transformer uses a convolutional stem in place of the patchily stem to improve training efficiency while maintaining overall stability. The global attention is computationally more expensive than local attention but it confines the interface between distinct tokens and needs more computer blocks to obtain global attention. As a result, in contrast to the vanilla transformer, the CrSWin Transformer uses CrSWin self-attention to more successfully capture global attention as seen in Fig. 4.

To be more precise, CrSWin self-attention calculates self-attention in parallel for the horizontal and vertical stripes after dividing the input data vertically and horizontally according to a predetermined size. Initially, it projecting the input data Z'' to J heads linearly. They are then split evenly into segments I and II. Every component has $J/2$ heads. The self-attention computation is then carried out in parallel in the horizontal and vertical stripes after segments I and II have been divided vertically and horizontally. Let use segment I as an illustration. The horizontal stripes $[Z''_1, \ldots, Z''_D]$ are obtained by dividing segment I with width ω_s in the vertical direction. Here, $D = H/\omega_s$ and size of every strip Z''_i is $\omega_s \times W \times C$, where, C represents the number of channels. Next, multihead attention is calculated for every stripe Z''_i, where j th head attention α_j^i is specified as.

$$\alpha_j^i = Attention\left(Z_i\omega_J^Q, Z_i\omega_j^J, Z_i\omega_j^v\right) \tag{12}$$

where $\omega_j^Q \in R^{C\times d_j}, \omega_j^J \in R^{C\times d_j}, \omega_j^V \in R^{C\times d_j}$ and $d_j = C/J$.

For the j th head, the total horizontal stripes self-attention is described as

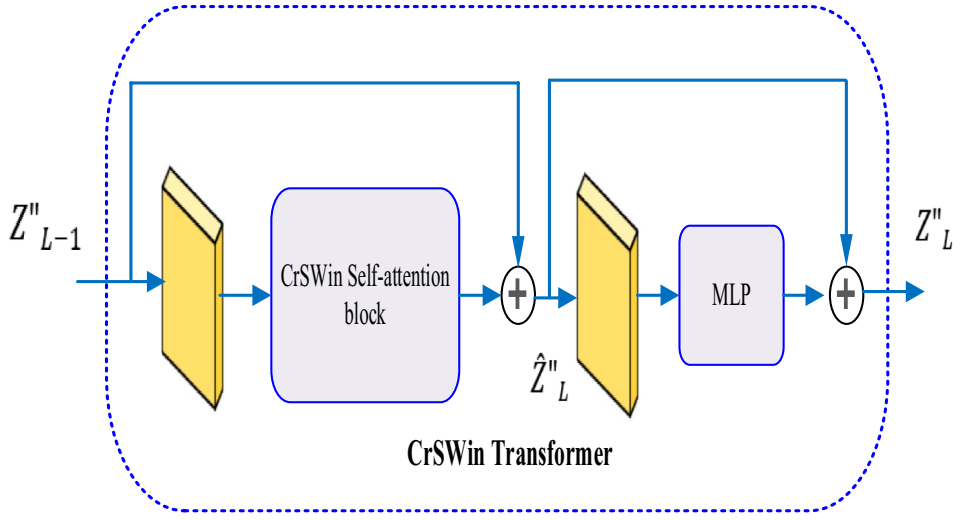

Figure 4. CrSWin Transformer.

$$Attention_H(Z") = \left[\alpha_j^1, \alpha_j^2, \ldots, \alpha_j^D\right] \tag{13}$$

The total self-attention for the vertical stripe $Attention_V(Z")$ is computed based on horizontal stripes self-attention. Following the self-attention's concatenation into horizontal and vertical stripes, the outcome is as follows:

$$CrSWin - Att \times \times ention(Z") = concat\left(Hd_1, Hd_2, \ldots, Hd_J\right)\omega^O \tag{14}$$

where $\omega^O \in R^{C\times C}$ and $Hd_J = \begin{cases} Attention_H(Z")j = 1,2,\ldots,J/2 \\ Attention_V(Z")j = J/2+1,\ldots,J \end{cases}$. $CrSWin - Attention(Z")$ is used to enlarge the attention region of every token within one transformer block. Furthermore, the strip width ω_s should be increased to increase the associating regions of CrSWin Transformer as the stage increases. Lastly, the CrSWin Transformer block processing can be properly stated as

$$\widehat{Z}"_L = CrSWin - Attention(LN(Z"_{L-1})) + Z"_{L-1} \tag{15}$$

$$Z"_L = MLP\left(LN\left(\widehat{Z}"_L\right)\right) + \widehat{Z}"_L \tag{16}$$

where the output of the L th CrSWin Transformer block is denoted by $Z"_L$. The last component of the suggested AConvCAT architecture is a global average pooling layer. The output feature of the backbone network's final stage is fed as input to pooling layer. To get the final classification results, a fully connected layer is then used as a softmax layer. The softmax function is employed to the input vector $(\widehat{Z}"_i)$ for computing the probability that an input is a part of a particular category i as:

$$SM\left(\widehat{Z}"_i\right) = \frac{exp\left(\widehat{Z}"_i\right)}{\sum_{j=1}^{n} exp\left(\widehat{Z}"_j\right)} \tag{17}$$

where $\widehat{Z}"_i$ denotes the i th component of the input $\widehat{Z}"$ and n is the total classes. The usual exponential functions employed for the the input and outputs are denoted as $exp\left(\widehat{Z}"_i\right)$ and $exp\left(\widehat{Z}"_j\right)$ correspondingly. The loss function is an essential tool to evaluate the difference of the projected results from the true labels. This loss function is used for guiding the learning process through the minimization of the error among the true and forecasted labels. Here, the clear-cut Cross-Entropy (CCE) loss function is utilized and it is described as:

$$Loss_{CCE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} v_{ij} log\left(\widehat{v}_{ij}\right) \tag{18}$$

where v_{ij} denotes the pointer function and its value is considered as 1 in cases where pixel i's ground truth label is j and 0 in all other cases. In the meantime, the estimated probability that pixel i belongs to class j is denoted as $,\widehat{v}_{ij}$.

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Results and discussion

The performance of the proposed CCDNet is validated by simulating it using python programming language. This section gives the simulation study of diverse datasets to classify CLC histopathological images. Similarly, the effectiveness of the proposed model is compared with some pre-trained models. Lastly, the efficiency of the proposed model is compared with the other state-of-the-art methods for the categorization of CLC tissue histopathology images.

Dataset description

This work proposed CCDNet to classify the multiple-class colorectal tissue by utilizing the publicly accessible datasets, colorectal histological images and NCT-CRC-HE-100 K.

Dataset A: Colorectal histological images

The dataset is accessible from https://zenodo.org/record/53169#Yp86-nZBzIX . The medical science library at the University Medical Center Mannheim, Germany given ten private colorectal cancer tissue slides stained with hematoxylin and eosin. The lower grade and higher grade tumours are included in this category. The colorectal histology dataset consists of five thousand 150×150 pixels, which are split into eight non-overlapping classes: MUCOSA (MUCA), EMPTY (EMPT), LYMPHO (LYMO), TUMOR (TUMR), COMPLEX (COMP), ADIPOSE (ADIPS), STROMA (STRM), and DEBRIS (DEBR). The image patches for every class are shown in Fig. 5.

Dataset B: NCT-CRC-HE-100 K

This dataset is accessible from https://zenodo.org/record/1214456#Yp9BY3ZBzIU . This dataset consists of 100,000 haematoxylin and eosinophilic images with size of 224×224 pixels. To enhance diversity, benign areas from gastric surgery specimens were added to normal tissue groups. There are nine classes in this dataset. DEBR,

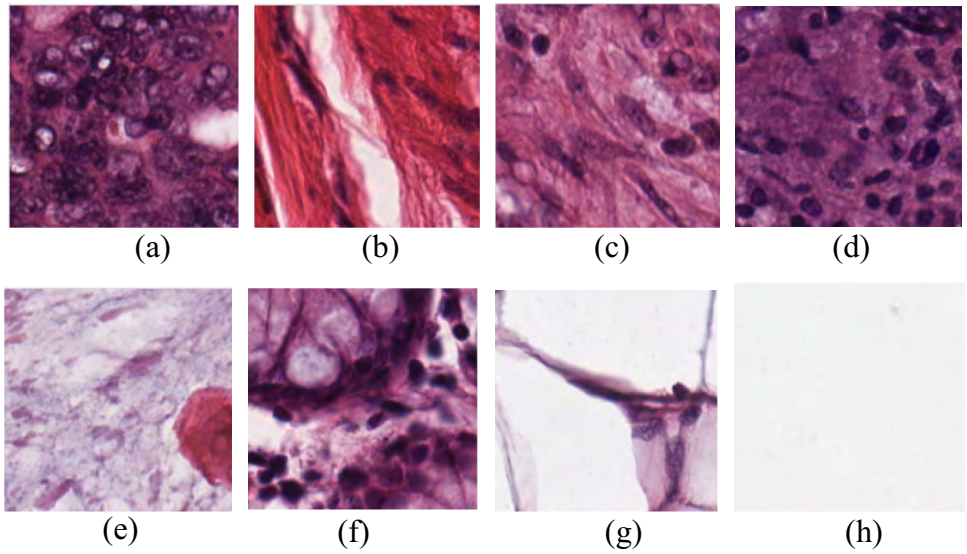

Figure 5. Sample images from Dataset A (**a**) TUMR (**b**) STRM (**c**) COMP (**d**) LYMO (**e**) DEBR (**f**) MUCA (**g**) ADIPS (h) EMPT.

background (BACK), MUCA, smooth muscle (MUS), normal colon mucosa (NORM-CM), ADIPS, LYMO, and TUMR are the designated groups. Every class denoting a region that has tumors. The image patches for every class are shown in Fig. 6.

Simulation settings

The simulation is performed on a PC with an Intel Core i7 processor. Also, PyTorch 1.10 and Python 3.8 were used to implement the proposed CCDNet. One important factor that has a big impact on the network's performance in the experiment is the partition of the dataset. In this work, the training, validating, and testing division approach is used for evaluating the network. For testing, validation, and training, the split ratios are 80%, 10%, and 10%, respectively. To have the best possible model performance, hyperparameters must be chosen and adjusted. A random search technique is used to set the important parameters for the proposed network, including batch size, optimization algorithm, epochs, and learning rate. Table 1 provides the proposed CCDNet's hyperparameter settings.

The efficiency of the proposed CCDNet was validated using accuracy, sensitivity, precision, specificity, false positive rate (FPR), False negative rate (FNR), F1-score, and AUC.

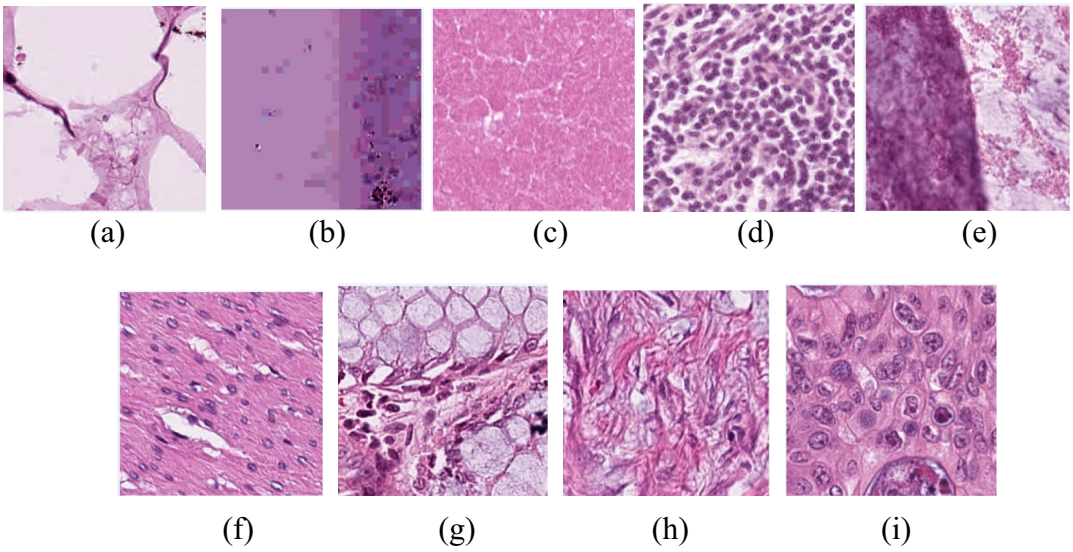

Figure 6. Sample images from Dataset B (a) ADIPS (b) BACK (c) DEBR (d) LYMO (e) MUCA (f) MUS (g) NORM-CM (h) STRM (i) TUMR.

Parameter	Value
Batch size	64
learning rate	0.001
Epochs	100
spatial patch sizes	[9 × 9], [11 × 11], [13 × 13], [15 × 15], [17 × 17]
Optimizer	ADAM

Table 1. Hyperparameters setting.

Performance analysis

This section shows the efficiency of the proposed CCDNet on publicly accessible datasets. Figure 7 shows the training approach for the proposed CCDNet and different pre-trained models, Xception, InceptionResNetV2, DenseNet121, and VGG16, on the Dataset A. The multiplicity of instances enables designs to develop a steady learning curve and validation graphs with relatively less disturbances. This collection consists of large-scale histopathology images. Existing pre-trained models are unable to represent large-scale spatial structures. In this dataset, the proposed CCDNet outperforms the rest of the models. This demonstrates that our CrSWin transformer transformer can successfully utilize appropriate data, seize feature relations, and accurately represent long-term dependences.

Figure 8 depicts how the individual models were trained over 100 epochs. The Colorectal Histology (5000 images) collection has fewer images than the NCT-CRC-HE100K dataset. Sparse datasets produce much worse performance than datasets with a large number of images. However, the suggested strategy outperforms all other

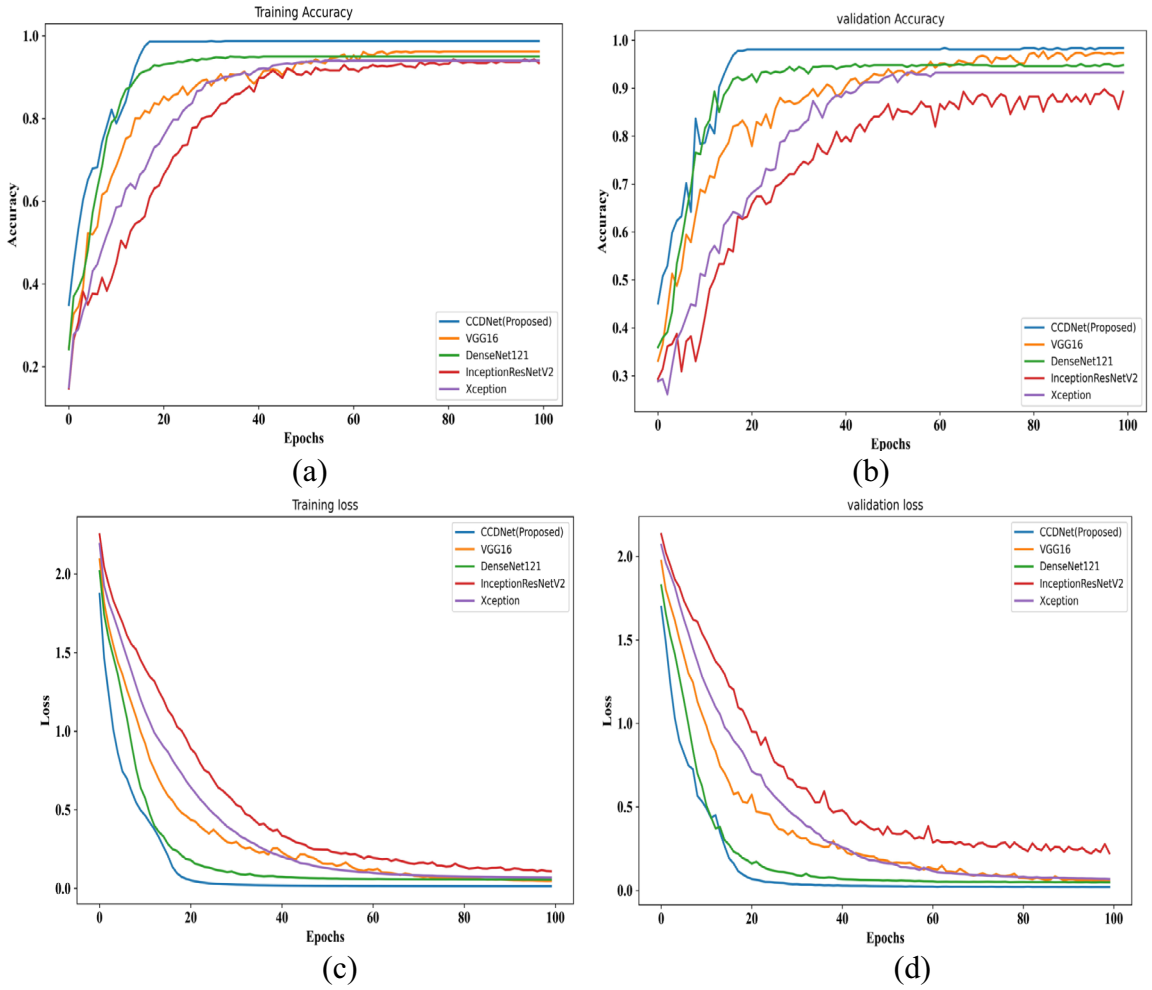

Figure 7. Accuracy and loss plot on Dataset A (**a**) training accuracy (**b**) validation accuracy (c) training loss (d) validation loss.

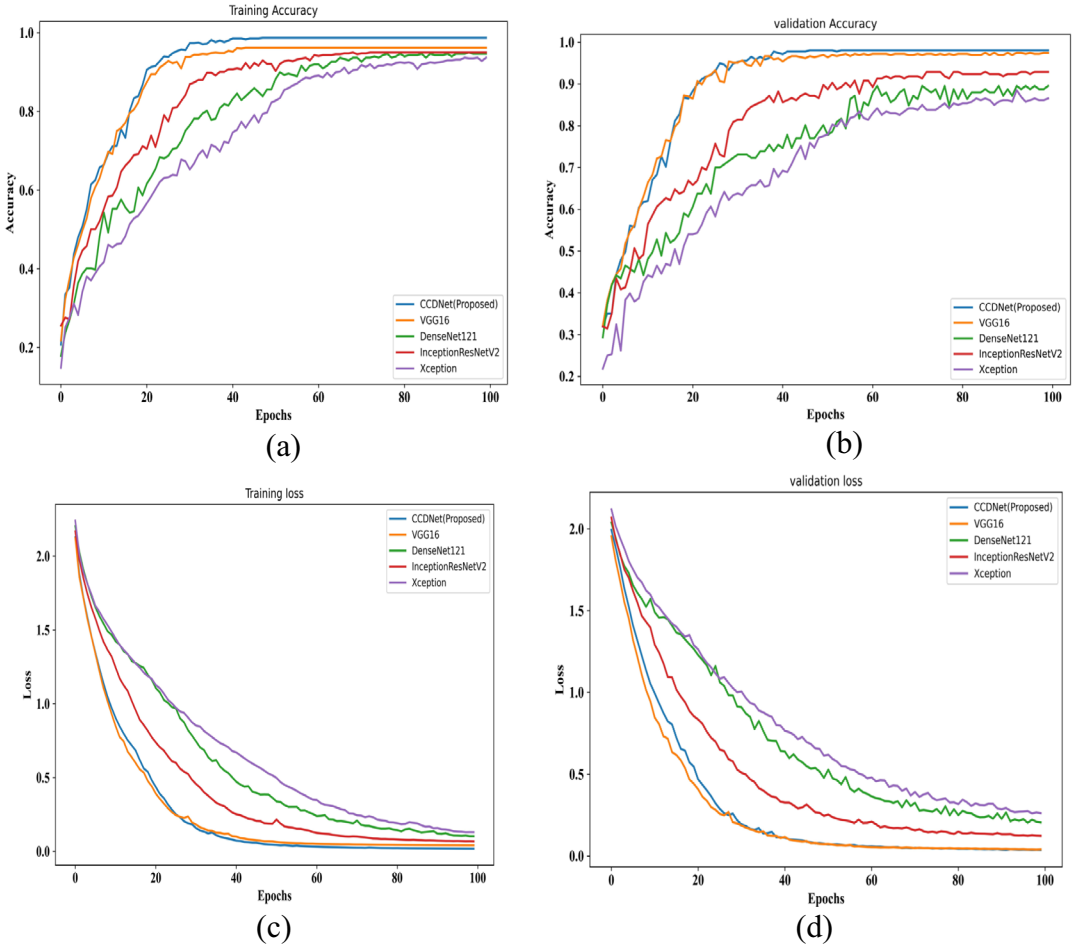

Figure 8. Accuracy and loss plot on Dataset B (**a**) training accuracy (**b**) validation accuracy (**c**) training loss (**d**) validation loss.

methods and achieves nearly 100% accuracy on short datasets because of its capacity for learning discriminative spatial information.

The proposed CCDNet's class-wise sensitivity, specificity, accuracy, and AUC are validated on Datasets A and B with the radar map in Fig. 9. The suggested CCDNet framework performed the best on Dataset A, with an average accuracy of 98.61%, sensitivity of 98.33%, specificity of 99.42%, and AUC of 0.9812. Similarly, it attained the highest accuracy (98.96%), sensitivity (98.80%), specificity (99.72%), and AUC (0.9912) on Dataset

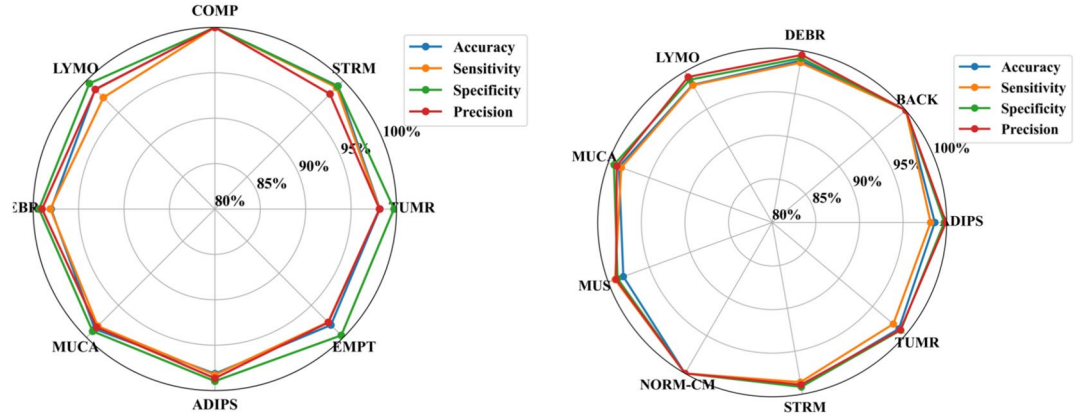

Figure 9. Radar map for class-wise performance analysis on (**a**) Dataset A (**b**) Dataset B.

B. The reason for this is that the suggested model effectively learns the discriminative spatial information of histopathology images.

Next, the proposed CCDNet performance is compared with Xception, InceptionResNetV2, DenseNet121, and VGG16. Tables 2 and 3 shows that CCDNet-Net outperforms the four pre-trained models on the test dataset. The performance of all the pre-trained models are reasonably good based on the classification results. However, pre-trained models have been found to produce subpar categorization consequences for certain classes because of their smaller sample dimension. The reason behind this is that they require more training data in order to operate at their best. It is noteworthy that the suggested AConvCAT framework delivers higher classification results on all evaluated datasets by adding spatial data with the aid of MAConv and CrSWin units. Furthermore, the accuracy of every class generated through CCDnet is relatively uniform, indicating that it is robust and successful even in classes with a smaller amount of training examples.

It is evident from the discussion above that the suggested network with histopathological images outperforms VGG16, Xception, InceptionResNetV2, DenseNet121, and others. Table 4 compares the performance of the provided network to other current techniques for classifying colorectal tissue histopathological images. By taking into account local and global representations across several layers, the proposed technique efficiently separates different tissue features, retains the best boundary areas, and achieves more realistic classification.

Class	Xception	InceptionRESNetV2	DeNSENet121	VGG16	CCDNet
TUM	88.23	89.12	87.18	90.16	98.16
STR	89.64	90.16	87.94	91.24	99.16
COM	88.34	89.67	89.14	90.37	100
LYM	90.17	91.74	88.67	90.87	98.64
DEB	87.69	88.97	87.24	91.28	98.04
MUC	88.38	89.64	88.38	90.66	98.69
ADI	89.62	90.15	87.37	92.10	98.14
EMP	86.31	87.72	87.82	91.68	98.07

Table 2. Comparative analysis of Class-wise accuracy performance with pre-trained models on Dataset A.

Class	Xception	InceptionResNetV2	DenseNet121	VGG16	CCDNet
ADI	94.18	91.89	92.64	95.61	98.64
BACK	95.42	92.14	93.47	99.12	100
DEB	94.87	91.29	92.45	96.46	98.87
LYM	95.63	92.52	93.64	94.37	98.24
MUC	95.55	90.34	91.96	95.43	98.7
MUS	96.12	91.75	92.81	96.54	98.16
NORM	94.97	92.12	93.34	94.16	100
STR	95.32	91.67	92.48	95.34	99.12
TUM	95.46	92.89	93.46	95.12	98.96

Table 3. Comparative analysis of Class-wise accuracy performance with pre-trained models on Dataset B.

Methods	Accuracy (%)	Sensitivity (%)	Precision (%)	Specificity (%)	F-1 score (%)	AUC
MFF-CNN[24]	96.00	94.05	97.42	97.50	95.94	0.9600
Ensemble DNN[20] (Dataset A)	92.83	93.11	92.83	92.54	92.83	0.9283
Ensemble DNN[20] (Dataset B)	96.16	96.15	96.17	96.17	96.16	0.9616
CRCCN-Net[25] (Dataset A)	93.50	93.62	94.12	99.06	93.86	0.9562
CRCCN-Net[25] (Dataset B)	96.26	96.34	96.44	99.52	96.38	0.9800
CCDNet (Dataset A)	98.61	98.33	98.55	99.42	98.24	0.9812
CCDNet (Dataset B)	98.96	98.80	99.37	99.72	98.64	0.9912

Table 4. Comparison with state-of-the-art methods.

Ablation study

In this section, we conducted several simulated trials to evaluate the suggested CCDNet network's performance in three viewpoints: the impact of different training samples, The effectiveness of the MAConv and CA-PDU modules and computational cost analysis. Figure 10 displays the classification performance of the suggested CCDNet model on two datasets with varying percentages of training samples. The training samples are chosen randomly 10%, 20%, 30%, 40%, and 50% on each dataset. This diverse training data distribution shows that improved classification performance is correlated with larger training data sets. The suggested CCDNet achieves exceptional categorization accuracy for all percentage of training samples on two datasets. However, the accuracy shows a slight improvement when more training data are used. The recommended model performs well in classification even when trained with a limited number of samples.

Simulations are then conducted to demonstrate the efficacy of the MAConv and CA-PDU modules with a CrSWin transformer[28]. The MAConv module is critical to the proposed model since it captures local features for compensating the transformer's deficiency of local feature modelling capacity, resulting in much improved classification. In addition, coordinate attention gathers cross-channel coordinate information, allowing the model to more correctly trace and identify the target[29,30]. To ascertain the utility of the MAConv and CA-PDU modules, experiments were conducted on each dataset both with and without these modules. Figure 11 depicts the performance attained on each dataset. The suggested module improves the model's accuracy over the traditional swin transformer, demonstrating that the MAConv and CA-PDU modules, together with a CrSWin transformer may efficiently learn discriminative local spatial features.

Next, the complexity of the proposed CCDNet is evaluated in terms of the amount of parameters, Floating Point Operations (FLOPs), memory size, training, and testing times in Table 5. According to the findings, the MAConv with CA-PDU and CrSWin transformer outperforms the MAConv with CA-PDU and conventional Swin transformer architecture. Specifically, the suggested CCDNet involves minimal training and testing time while delivering exceptional performance.

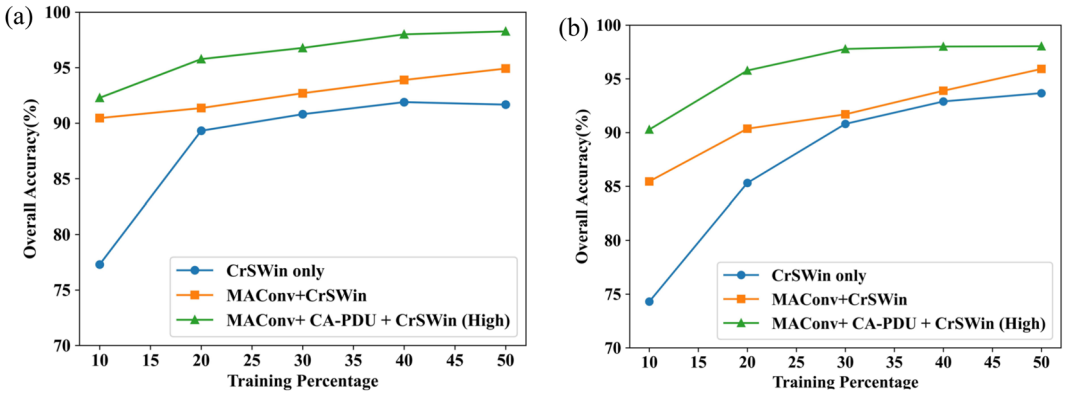

Figure 10. Evaluation of CCDNet across various training percentages (**a**) Dataset A (**b**) Dataset B.

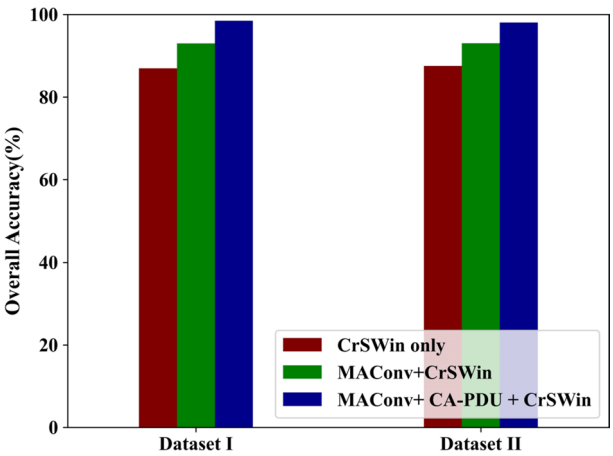

Figure 11. Performance of proposed CCDNet with and without MAConv and CA-PDU modules on two datasets.

Model	Parameters (M)	FLOPS (M)	Memory size	Training time (s)	Testing time (s)
MAConv + CA-PDU + CrSwin	1.49	169.66	6.67 MB	124.23	3.16
MAConv + CA-PDU + conventional Swin	1.82	221.18	3.14 MB	210.12	6.48

Table 5. Complexity of the proposed CCDNet.

Conclusion

This research describes a new colorectal cancer detection network for automatically classifying colorectal tissue from histopathological images. The MAConv module has performed significant role in enabling the network to learn local spatial characteristics at different resolutions and sizes by increasing its receptive field. Here, deep learning is applied for fusing the features of histopathological imaging data at several scales. It built the most discriminating high-quality feature vector, thereby overcoming the drawbacks of single feature analysis. The suggested CrSwin transformer improved feature representations and simplified feature refinement. The simulation results demonstrated that the multi-scale feature fusion technique could differentiate the colorectal tissues easily from histopathological images. As per the study results, our CCDNet has an accuracy rate of 98.61% on the colorectal histology dataset and 98.96% on the NCT-CRC-HE-100 K dataset for detecting colorectal tissue cancer. In the near future, pathologists may find the suggested CCDNet to be very helpful in aiding with histology image diagnosis, as it is both time and financially efficient.

Data availability

The datasets generated and or analyzed during the current study is publicly available of the submitted research work. Three customized CNN architectures have been trained using two big datasets separately. 1. The dataset is accessible from https://zenodo.org/record/53,169#Yp86-nZBzIX. The medical science library at the University Medical Center Mannheim, Germany given ten private colorectal cancer tissue slides stained with hematoxylin and eosin. 2.= This dataset is accessible from https://zenodo.org/record/1,214,456#Yp9BY3ZBzIU. This dataset consists of 100,000 haematoxylin and eosinophilic images with size of 224 × 224 pixels.

Received: 17 June 2024; Accepted: 13 August 2024
Published online: 17 August 2024

References

1. Sung, H. *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* https://doi.org/10.3322/caac.21660 (2021).
2. Thanarajan, T., Alotaibi, Y., Rajendran, S. & Nagappan, K. Eye-tracking based autism spectrum disorder diagnosis using chaotic butterfly optimization with deep learning model. *Comput. Mater. Continua* https://doi.org/10.32604/cmc.2023.039644 (2023).
3. Li, Y., Zhang, F. & Xing, C. Screening of pathogenic genes for colorectal cancer and deep learning in the diagnosis of colorectal cancer. *IEEE Access* **8**, 114916–114929 (2020).
4. Murakami, T. *et al.* Sessile serrated lesions: Clinicopathological characteristics, endoscopic diagnosis, and management. *Dig. Endosc.* **34**(6), 1096–1109 (2022).
5. Tsai, M.-J. & Tao, Y.-H. Deep learning techniques for the classification of colorectal cancer tissue. *Electronics* **10**(14), 1662 (2021).
6. Wang, K.-S. *et al.* Accurate diagnosis of colorectal cancer based on histopathology images using artificial intelligence. *BMC Med.* https://doi.org/10.1186/s12916-021-01942-5 (2021).
7. Rajendran, S. *et al.* Automated segmentation of brain tumor MRI images using deep learning. *IEEE Access* **11**, 64758–64768 (2023).
8. Yu, C. & Helwig, E. J. The role of AI technology in prediction, diagnosis and treatment of colorectal cancer. *Artif. Intell. Rev.* https://doi.org/10.1007/s10462-021-10034-y (2022).
9. Yin, Z., Yao, C., Zhang, L. & Qi, S. Application of artificial intelligence in diagnosis and treatment of colorectal cancer: A novel Prospect. *Front. Med.* **10**, 1128084 (2023).
10. Xu, H. *et al.* Artificial intelligence–assisted colonoscopy for colorectal cancer screening: a multicenter randomized controlled trial. *Clin. Gastroenterol. Hepatol.* https://doi.org/10.1016/j.cgh.2022.07.006 (2023).
11. Pacal, I., Karaboga, D., Basturk, A., Akay, B. & Nalbantoglu, U. A comprehensive review of deep learning in colon cancer. *Comput. Biol. Med.* **126**, 104003 (2020).
12. Jain, Astha, Manish Pandey, and Santosh Sahu. A deep learning-based feature extraction model for classification brain tumor. In *Proceedings of Data Analytics and Management: ICDAM 2021, Volume 1*, pp. 493–508. Springer Singapore, (2022).
13. Davri, A. *et al.* Deep learning on histopathological images for colorectal cancer diagnosis: A systematic review. *Diagnostics* **12**(4), 837 (2022).
14. Fadafen, K. & Masoud, and Khosro Rezaee.,. Ensemble-based multi-tissue classification approach of colorectal cancer histology images using a novel hybrid deep learning framework. *Sci Rep.* **13**, 8823 (2023).
15. Sarvamangala, D. R. & Kulkarni, R. V. Convolutional neural networks in medical image understanding: A survey. *Evol. Intel.* https://doi.org/10.1007/s12065-020-00540-3 (2022).
16. Chattopadhyay, A. & Maitra, M. MRI-based brain tumour image detection using CNN based deep learning method. *Neurosci. Inform.* **2**(4), 100060 (2022).
17. Yu, S. *et al.* Mil-vt: Multiple instance learning enhanced vision transformer for fundus image classification.". In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021* (ed. Yu, S.) (Springer International Publishing, 2021).
18. Wang, X. *et al.* Transformer-based unsupervised contrastive learning for histopathological image classification. *Med. Image Anal.* **81**, 102559 (2022).
19. Thanarajan, T., Alotaibi, Y., Rajendran, S. & Nagappan, K. Improved wolf swarm optimization with deep-learning-based movement analysis and self-regulated human activity recognition. *AIMS Math.* **8**(5), 12520–12539 (2023).
20. Ghosh, S. *et al.* Colorectal histology Tumor detection using ensemble deep neural network. *Eng.Appl. Artif. Intell.* **100**, 104202 (2021).

21. Khan, A. *et al.* Computer-assisted diagnosis of lymph node metastases in colorectal cancers using transfer learning with an ensemble model. *Modern Pathol.* **36**, 100118 (2023).
22. Graham, S. *et al.* One model is all you need: multi-task learning enables simultaneous histology image segmentation and classification. *Med. Image Anal.* **83**, 102685 (2023).
23. Zidan, U., Gaber, M. M. & Abdelsamea, M. M. SwinCup: Cascaded swin transformer for histopathological structures segmentation in colorectal cancer. *Expert Syst. Appl.* **216**, 119452 (2023).
24. Liang, M., Ren, Z., Yang, J., Feng, W. & Li, Bo. Identification of colon cancer using multi-scale feature fusion convolutional neural network based on shearlet transform. *IEEE Access* **8**, 208969–208977 (2020).
25. Kumar, A., Vishwakarma, A. & Bajaj, V. Crccn-net: Automated framework for classification of colorectal tissue using histopathological images. *Biomed. Signal Process. Control* **79**, 104172 (2023).
26. Bousis, D. *et al.* The role of deep learning in diagnosing colorectal cancer. *Prz Gastroenterol.* **18**, 266–273. https://doi.org/10.5114/pg.2023.129494 (2023).
27. Chlorogiannis, D. D. *et al.* Tissue classification and diagnosis of colorectal cancer histopathology images using deep learning algorithms. Is the time ripe for clinical practice implementation. *Prz Gastroenterol.* https://doi.org/10.5114/pg.2023.130337 (2023).
28. Ogudo, K. A., Surendran, R. & Khalaf, O. I. Optimal Artificial Intelligence Based Automated Skin Lesion Detection and Classification Model. *Comput. Syst. Sci. Eng.* https://doi.org/10.32604/csse.2023.024154 (2023).
29. Selvanarayanan, R., Rajendran, S., Algburi, S., Ibrahim Khalaf, O. & Hamam, H. Empowering coffee farming using counterfactual recommendation based RNN driven IoT integrated soil quality command system. *Sci. Rep.* **14**(1), 6269 (2024).
30. Luo, Y. *et al.* DAFNet: A dual attention-guided fuzzy network for cardiac MRI segmentation. *AIMS Math.* **9**(4), 8814–8833 (2024).

Author contributions

M.K. and S.D. wrote the main manuscript text and S.V and S.R. prepared figures and tables. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.R.

Reprints and permissions information is available at www.nature.com/reprints.

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