International Research Journal of Multidisciplinary Scope (IRJMS), 2025; 6(1):238-249

Original Article | ISSN (O): 2582-631X DOI: 10.47857/irjms.2025.v06i01.02299

A Hybrid CNN-Transformer Deep Learning Framework with Convolutional Block Attention Module for Enhanced Gastrointestinal Endoscopy Analysis

Pradeepan P[1]*, Gladston Raj S[2], Juby George K[3],

[1]Centre for Development of Imaging Technology (C-DIT) University of Kerala, Kerala, India, [2]Department of Computer Science, Government College Kariavattom, Trivandrum, Kerala, India, [3]Department of Computer Applications, Marian College Kuttikanam, Idukki, Kerala, India. *Corresponding Author's Email: pradeepanp@cdit.org

Abstract

Gastrointestinal (GI) endoscopy is crucial for the diagnosis of digestive diseases. It provides detailed visual information about the GI tract and helps identify abnormalities. However, the analysis of endoscopic images is challenging due to their complexity and variations caused by factors like lighting, texture and patient movement. These challenges highlight the need for advanced methodologies to enhance diagnostic accuracy and efficiency. This study introduces a novel deep learning framework integrating hybrid CNN-transformer models enhanced by a Convolutional Block Attention Module (CBAM). The framework utilizes a pre-trained Vision Transformer (ViT) to capture global image features and a convolutional neural network (CNN) to extract local features. CBAM refines the focus on relevant regions and enhances the interpretability and performance of the model. Ensemble learning was used to combine predictions from multiple models and improve the reliability and accuracy of the framework. The proposed model was evaluated on the publicly available Kvasir GI endoscopy dataset and demonstrated superior performance with an accuracy of 94.13% and a precision of 94.21%, outperforming existing methods. This framework offers a reliable and effective solution for analysing GI endoscopy images, potentially improving the accuracy and reliability of automated diagnosis. This can lead to early disease detection and improved patient outcomes.

Keywords: Convolutional Block Attention Module, Gastrointestinal Disease Detection, Hybrid CNN-Transformer, Wireless Capsule Endoscopy.

Introduction

The gastrointestinal (GI) tract is an essential part of the digestive system. It is a common site for neoplastic and non-neoplastic diseases (1). The GI tract is composed of organs that extend from the oesophagus to the anus. Colorectal cancer (CRC) and stomach cancer are among the major causes of the global health burden, underscoring the importance of precise and timely diagnosis (2). Traditionally, endoscopy has been one of the most essential tools for GI tract assessment. This minimally invasive technique uses a forward-viewing fiberscope to inspect the luminal surfaces of the organs directly. Despite the promise of new capabilities, accurate interpretation of endoscopic images remains a major challenge (3). Neoplastic lesions frequently begin with minor mucosal changes. The surface texture, crypt pattern, and micro vascularisation were all subject to change. Inexperienced endoscopists may find distinguishing between modest and regular alterations challenging. Consequently, this can lead to missed diagnosis and delayed treatment (4). The varying illumination due to light source changes, patient motion artifacts that create image instability and the complex anatomical topography of the GI tract with overlapping mucosal folds further complicate lesion diagnosis. Because of these constraints, endoscopists must review images meticulously, which increases their cognitive burden and may affect diagnostic consistency (5). These problems highlight the need for innovative solutions to enhance the diagnostic accuracy and efficiency of gastrointestinal disease detection. Recent advances in artificial intelligence (AI), notably deep learning algorithms based on multi-layered artificial neural networks, have resulted in a robust medical image interpretation toolbox (6). These algorithms, especially Convolutional Neural Networks (CNNs), are excellent for extracting

(Received 04th September 2024; Accepted 22nd January 2025; Published 31st January 2025)

complex patterns from large medical-image datasets. They can accurately automate endoscopic procedures such as image segmentation and lesion detection (7). CNNs have demonstrated significant promise in gastrointestinal disease classification, as evidenced by various studies achieving high accuracy in detecting conditions such as precancerous oesophagus and colon polyps (8). Despite progress in CNNs, they still have limitations when it comes to capturing long-range dependencies in complex images. These dependencies are important because they show subtle connections between different parts of the image. These connections can help identify early signs of disease on GI endoscopy, such as changes in mucosal texture or blood vessel patterns. To address these limitations, recent advancements in medical image analysis have increasingly focused on incorporating attention mechanisms. These mechanisms enable models to focus on the most important parts of an image, which is similar to human visual attention. In medical imaging, spatial attention highlights important regions within an image, whereas channel attention emphasizes key feature channels. Both have proven to be especially valuable for improving the analysis. This has led to the development of various sophisticated attention modules designed to capture both the spatial and channel-wise dependencies in medical images. Transformers have recently emerged as promising solutions for medical-image analysis (9). They used self-attention mechanisms that help capture long-range relationships more effectively than CNNs. This makes transformers ideal for analyzing complex medical images, such as endoscopy images, where understanding the overall image is key for accurate disease detection. Deep neural networks based on transformers have been developed to improve the diagnostic accuracy of endoscopies. These networks have shown better performance in detecting gastrointestinal issues such as lesions and colon polyps. Vision transformer models with hybrid-shifted windows were designed to capture both short- and long-range dependencies. Transformers have also been used to analyze images from endoscopic capsule videos, allowing a highly accurate diagnosis of GI tract issues, even with limited data (10). Existing research has highlighted the need for more robust models that effectively combine the strengths of CNNs and transformers, especially for tasks requiring local and global feature extraction. Combining CNNs and Transformers in hybrid models has proven to be an effective approach for taking advantage of their individual strengths (11). CNNs are good at capturing detailed local features, whereas transformers excel in understanding the broader context and long-range relationships in data (12). However, despite several attempts at hybrid models (13), the effective merging of these two architectures remains a challenge. Many models still struggle to emphasize important local features while capturing the larger context in endoscopic images. In addition, many models do not fully utilize attention mechanisms to enhance feature extraction in CNNs and transformers. This study introduced a hybrid CNN-transformer model enhanced with a Convolutional Block Attention Module (CBAM) to improve the analysis of gastrointestinal endoscopy images. The proposed framework combines the local feature-extraction capabilities of CNNs with the ability of transformers to model long-range dependencies. CBAM focuses on spatial and channel-wise details to improve feature extraction. Unlike many existing hybrid models, this method creates a balanced system that addresses the limitations of relying solely on either local or global feature extraction. This approach aims to enhance the diagnostic accuracy and robustness in detecting gastrointestinal diseases. We also used ensemble-learning techniques to improve the model's performance and generalizability. The proposed work makes the following contributions.

- Introduced a novel hybrid model that combines a Vision Transformer (ViT) with an EfficientNet-B1 backbone enhanced by Convolutional Block Attention Modules (CBAM). This design captures detailed local features (EfficientNet-B1 with CBAM) and long-range dependencies (ViT).
- Replacing the original squeeze-and-excitation blocks in EfficientNet-B1 with CBAM enabled the model to learn both channel-wise and spatial attention. This change enhances the focus on the critical features within the CNN backbone.
- The learnable weighted ensemble method combined different backbone outputs. This

approach allowed the model to prioritize the most important features of each backbone.

- The proposed model was tested on the KVASIR wireless endoscopic dataset and showed superior performance compared to current leading models.

The remainder of this paper follows this organization:—Section 2 explains the proposed hybrid CNN-transformer model, including the integration of the Convolutional Block Attention Module and details of the dataset used. Section 3 presents the experiments, results, discussion, ablation study and a comparison with existing research. Finally, Section 4 concludes the paper.

Materials and Method

This section presents a novel deep-learning framework for analysing gastrointestinal endoscopy images. The proposed architecture uses a hierarchical feature extraction approach to capture the fine details and global image features. First, a pre-trained EfficientNetB1 model enhanced with a Convolutional Block Attention Module (CBAM) extracts basic image features, such as textures and geometric patterns. EfficientNetB1 provides a good balance between the model complexity and performance. The integrated CBAM focuses on the essential regions in the image to improve feature representation. Subsequently, the Vision Transformer (ViT) architecture extracts high-level image features from the input. ViT is excellent at capturing long-range dependencies and global contextual information, which are crucial for understanding an image. Using a cascade approach, this framework extracts detailed local information and global image features. The features extracted by EfficientNet-CBAM and ViT are combined into a single feature vector. This fusion uses an ensemble approach with learnable weights for each architecture, allowing the model to combine complementary information from both feature sets adaptively. The final feature vector is fed into a fully connected classifier for the classification task. Figure 1 illustrates the architecture of the proposed framework. The following subsections explain each processing stage in detail.

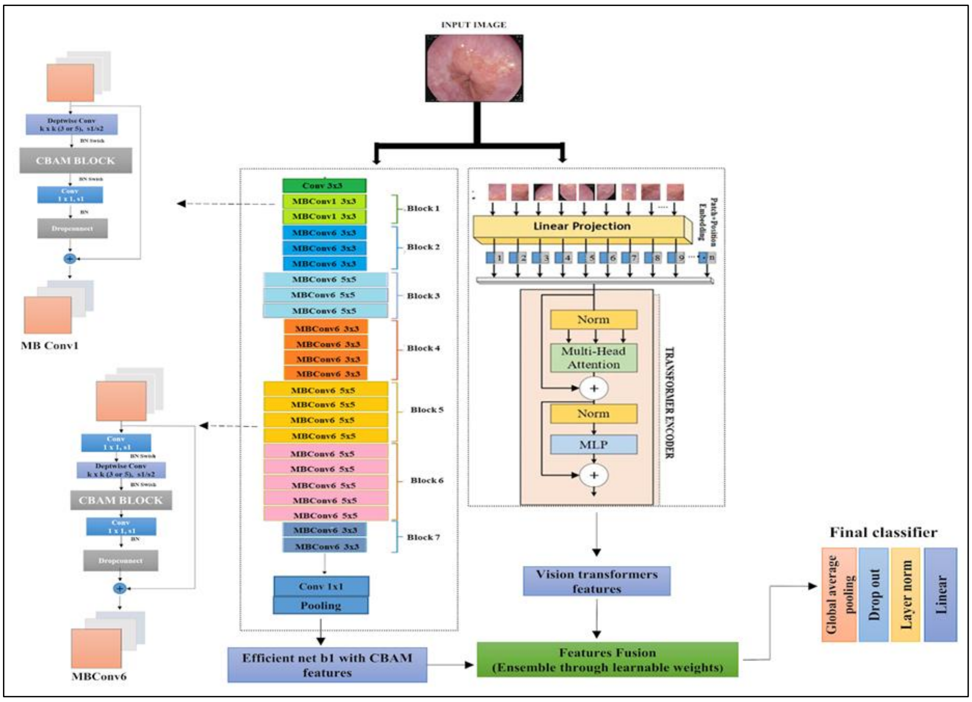


Figure 1: The Proposed Architecture

Convolutional Block Attention Module (CBAM)

This section provides a synopsis of the Convolutional Block Attention Module (CBAM), an attention mechanism introduced by Woo *et al.* (14). CBAM enhances the feature representation by incorporating sequential channels and spatial attention mechanisms. Figure 2 shows the hierarchical structure of the CBAM attention block and provides a visual representation of the steps described below. Given an input feature map (F) of dimensions (C × H × W), where C denotes the number of channels and H and W represent the height and width, respectively, the module sequentially refines this representation in two stages.

Channel Attention Module

This module dynamically modulates the significance of individual channels within the feature map. To achieve this, the process applies average and maximum pooling operations sequentially along the spatial dimensions (height and width). This process generates two separate feature maps (C × 1 × 1) that encapsulate global channel-wise statistical information. Subsequently, these feature maps were independently processed using a shared multilayer perceptron (MLP) network for dimensionality reduction and feature transformation. As shown in Figure 2B, the channel attention weights ((M_c)) were calculated using the following formula:

$$M\_c(F) = \sigma[MLP(AvgPool(F))] + \sigma[MLP(MaxPool(F))] \qquad [1]$$

Where (σ) represents the sigmoid activation function. As shown in Figure 2B, (AvgPool(F)) and (MaxPool(F)) are the average and maximum pooled feature maps, respectively. The outputs from both MLPs were then aggregated through summation and further processed using a sigmoid activation function (σ) to yield the final channel attention weights (Mc) as a feature map of the dimensions (C × 1 × 1).

Spatial Attention Module

This module emphasizes the dynamic modulation of salience across spatial locations within individual feature channels. It inputs a channel-wise refined feature map (F). Similar to the previous channel attention module, it performs the average and maximum pooling operations along the channel axis.

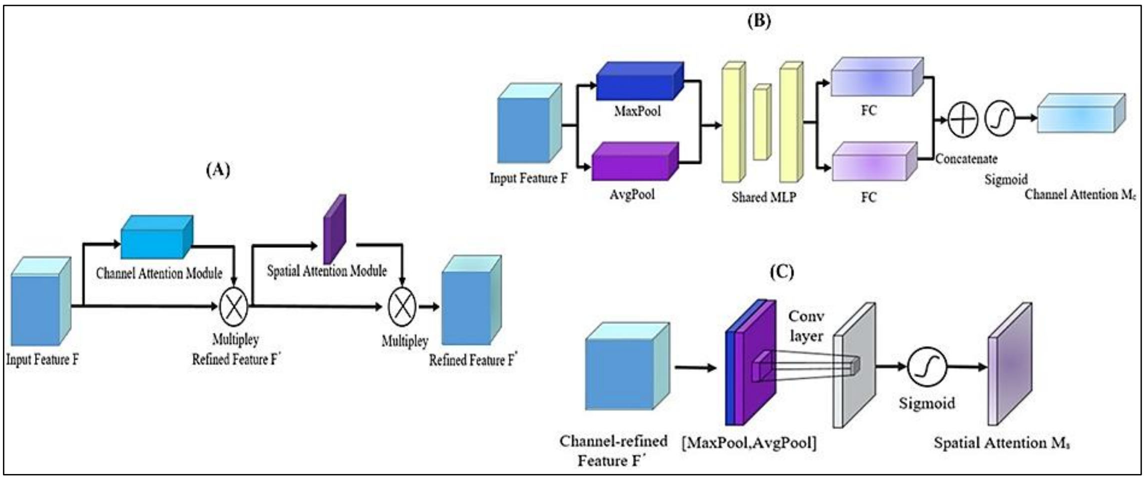


Figure 2: Illustrates the hierarchical structure of the CBAM attention block: (A) CBAM Module, Architecture, (B) Channel Attention Module Architecture, and (C) Spatial Attention Module Architecture

This yields two feature maps of dimensionality (1 × H × W) that encapsulate localized and global spatial statistics. The module concatenates these maps to facilitate the extraction of multi-scale spatial information. A 7 × 7 convolution operation subsequently reduces the dimensionality of the concatenated feature map. The spatial attention weights ((M_s)), as shown in Figure 2C, are then calculated as:

$$M\_s(F') = \sigma\{f\_7x7[Concat(AvgPool(M\_c), MaxPool(M\_c))]\} \qquad [2]$$

Finally, a sigmoid activation function (σ) generated spatial attention weights (M_s), resulting in a feature map of dimensions (1 × H × W). Where f_ 7 × 7 denotes the convolution operation with a kernel size of 7 × 7, and the final refined feature map (F'') is obtained by element-wise multiplication of the input feature map (F) with the channel attention weights (M_c) and spatial attention weights (M_s), as follows:

$$F'' = M\_c(F) \otimes F \otimes M\_s(F') \qquad [3]$$

This refined feature map (F'') incorporates both channel-wise and spatial attention, potentially improving feature representation and model performance.

Efficient Net

Deep Convolutional Neural Networks (CNNs) often face a trade-off between model capacity, training efficiency, and accuracy. This trade-off is influenced by three key dimensions: the network depth (number of layers), the number of channels (feature maps per layer), and input image resolution. EfficientNet addresses this challenge by using a composite scaling method (15). This method optimizes these three dimensions with a fixed scaling coefficient, allowing a balanced increase in model capacity while maintaining computational efficiency. EfficientNet's architecture uses a stack of Mobile Inverted Bottleneck Convolution (MBConv) blocks.

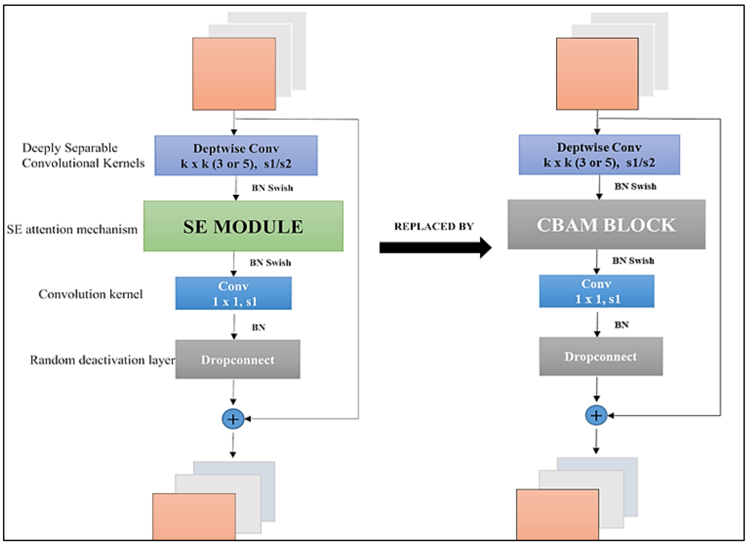


Figure 3: Shows the Replacement of the Squeeze-and-Excitation (SE) Module in MBConv1 with a CBAM Block

Each MBConv block includes a squeeze-and-excitation (SE) attention mechanism (16). The SE module compresses channel-wise features into a 1D representation of input features. A fully connected layer estimates the weights for each channel and multiplies them element-wise by the original feature maps. This dynamic weighting scheme focuses on informative channels within the feature map and enhances feature representation. Replacing the SE mechanism with a Convolutional Block Attention Module (CBAM) can enhance model performance. Figures 3 and 4 show the substitutions in MBConv1 and MBConv6, respectively. CBAM integrates channel and spatial attention and provides a more comprehensive focus on critical features. This dual-attention mechanism helps to capture more detailed and contextually relevant information and improves the model’s ability to interpret complex medical images.

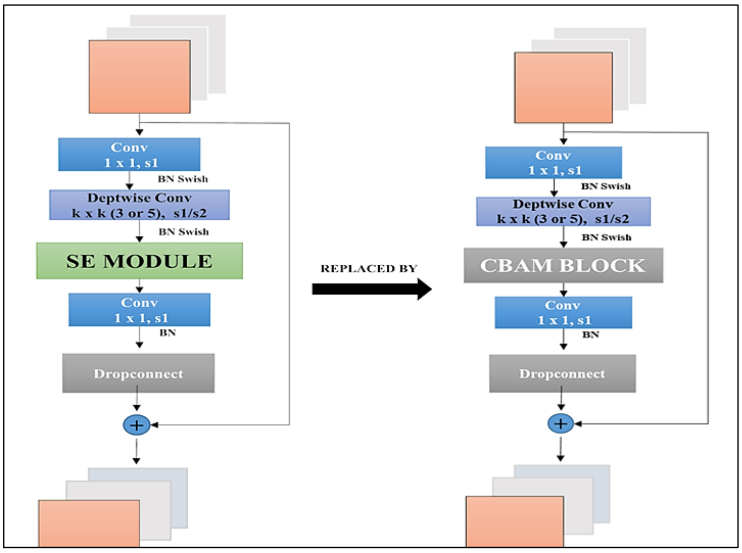


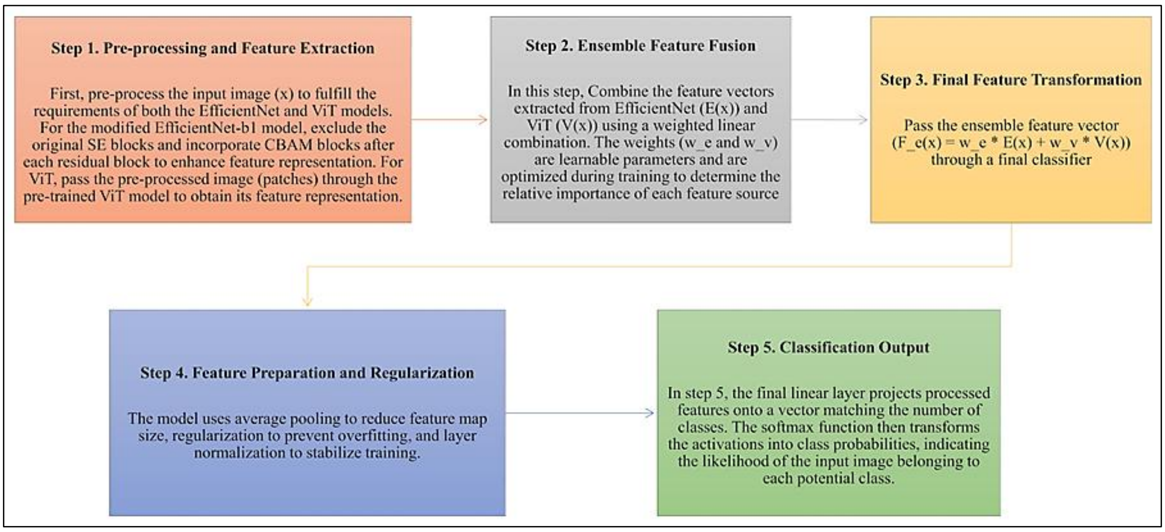
Figure 4: Illustrates the Replacement of the SE Module in MBConv6 with a CBAM Block

Ensemble Feature Fusion with Learnable Weights

This section describes the ensemble feature-fusion strategy used in the proposed hybrid model. Ensemble learning is a well-established machine-learning technique. It combines the strengths of multiple models to achieve improved performance compared with a single model (17). In this implementation, combined features were extracted using two pre-trained deep learning models: EfficientNet with CBAM integration and a Vision Transformer (ViT). For a given input image x, E (x) represents the feature vector extracted by the EfficientNet-CBAM model. V (x) represents the feature vector extracted using the ViT model for the same input image. The ensemble feature vector F_e(x) results from a weighted linear combination of these feature vectors.

$$F\_e(x) = w\_e * E(x) + w\_v * V(x) \qquad [4]$$

Where w_e and w_v represent the learnable weights for the EfficientNet and ViT features, respectively. We initialized both weights to 0.5 at the beginning of the training process. The model implements these as trainable parameters and optimises them during the training process using a dedicated learning rate (ensemble_lr=.0001) specified in the proposed model. Figure 5 shows the flowchart of the proposed architecture.



Step 1. Pre-processing and Feature Extraction

First, pre-process the input image (x) to fulfill the requirements of both the EfficientNet and ViT models. For the modified EfficientNet-b1 model, exclude the original SE blocks and incorporate CBAM blocks after each residual block to enhance feature representation. For ViT, pass the pre-processed image (patches) through the pre-trained ViT model to obtain its feature representation.

Step 2. Ensemble Feature Fusion

In this step, Combine the feature vectors extracted from EfficientNet (E(x)) and ViT (V(x)) using a weighted linear combination. The weights (w_e and w_v) are learnable parameters and are optimized during training to determine the relative importance of each feature source

Step 3. Final Feature Transformation

Pass the ensemble feature vector (F_e(x) = w_e * E(x) + w_v * V(x)) through a final classifier

Step 4. Feature Preparation and Regularization

The model uses average pooling to reduce feature map size, regularization to prevent overfitting, and layer normalization to stabilize training.

Step 5. Classification Output

In step 5, the final linear layer projects processed features onto a vector matching the number of classes. The softmax function then transforms the activations into class probabilities, indicating the likelihood of the input image belonging to each potential class.

Figure 5: The Flowchart of the Proposed Architecture

Materials

Kvasir Endoscopy Dataset: The publicly available Kvasir dataset (18) is a valuable resource for developing and evaluating deep learning algorithms for endoscopic image analysis. This dataset comprised 8,000 high-resolution images (with dimensions ranging from 720 × 579 to 1920 × 1070 pixels). These images were acquired during upper and lower gastrointestinal endoscopy procedures targeting anatomical regions such as the esophagus, stomach, and colon. Each image was annotated by experts to indicate the presence or absence of various gastrointestinal pathologies, including polyps, ulcers, and inflammatory conditions. The analysis of the Kvasir dataset identified eight diagnostic groups: the z-line, cecum, pylorus, esophagitis, polyps, ulcerative colitis, dyed resection margins, and dyed-lifted polyps. The division of the dataset into training, validation, and test sets followed an 8:1:1 split, ensuring a robust distribution for model development and performance assessment. The training set contains 6,400 images, while the validation and test sets contain 800 images. Figure 6 presents the representative images from the dataset.

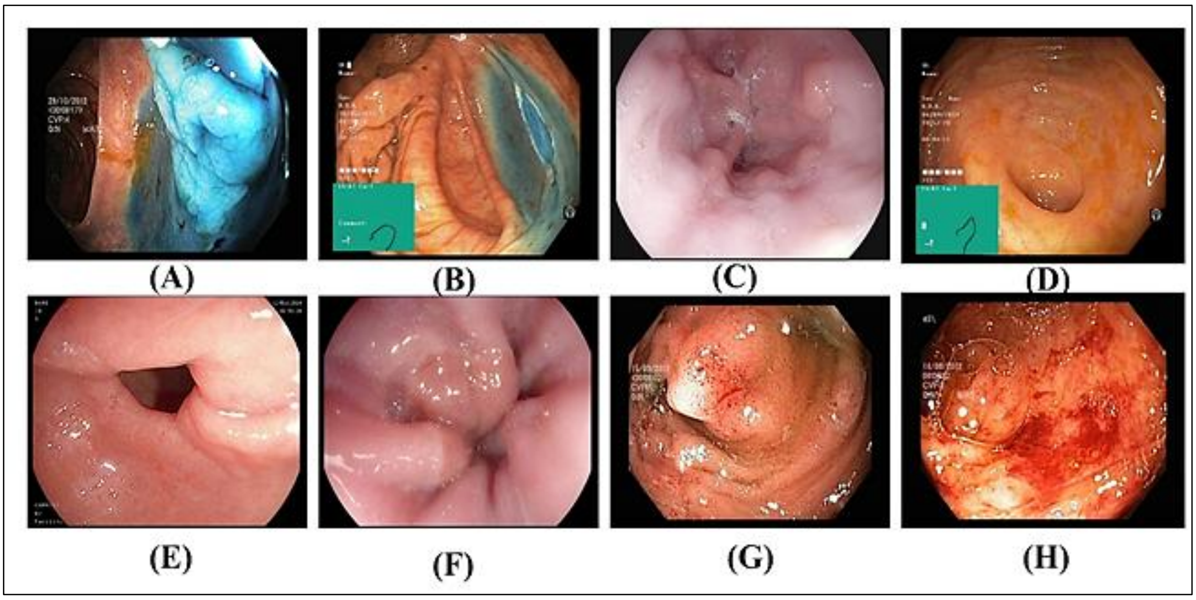


Figure 6: Representative Images from Each Class: (A) Dyed and Lifted Polyps (B) Dyed Dissection Margins (C) Esophagitis (D) Normal Cecum (E) Normal Pylorus (F) Normal Z-line (G) Polyps and (H) Ulcerative Colitis

Results

The proposed method was implemented and tested on Google Colab Pro, a cloud-based platform that provides access to high-performance computing resources. The hardware configuration included a 16-GB GPU, likely an NVIDIA Tesla series unit. The software environment utilized CUDA version 12.2 and CUDNN version 8.9. CUDA, a parallel computing platform developed by NVIDIA, enables the efficient use of GPUs for computationally intensive tasks such as deep learning. CUDNN, a library designed by NVIDIA, accelerates deep neural network computations on GPUs. The operating system used was Windows 10, with Python 3.8.8 as the foundation for the PyTorch deep learning framework. Effective image pre-processing is crucial for optimally preparing data in deep learning models. This involves techniques to improve data quality, reduce noise, and enhance relevant features by applying various transformations to the existing images. These transformations can be categorized into geometric and color variations. Geometric transformations include resizing images to a standard size of 448 × 448 pixels, applying random horizontal flips to account for object orientation variations, and introducing slight rotations to help the model handle rotations in real-world scenarios. Color jittering introduces controlled variations in brightness and contrast, mimics real-world lighting conditions, and prevents the model from overfitting specific color distributions in the training data. Additionally, advanced techniques like CLAHE (contrast-limited adaptive histogram equalization (CLAHE) can enhance image contrast, potentially leading to better feature extraction and model performance. Incorporating these diverse data augmentation strategies effectively diversifies the training dataset, promoting the ability of the model to learn more robust and generalizable features, ultimately reducing overfitting and enhancing the performance of unseen data. The training used RGB endoscopic images for 50 epochs with the Adam optimizer and applied a low learning rate of 0.0001 to prevent overfitting. The model employed categorical cross-entropy as the loss function, which is well-suited for multiclass classification tasks in gastrointestinal disease identification. To mitigate overfitting, the training process incorporated several regularization techniques. First, an early stopping mechanism halted training when the validation loss plateaued. Additionally, the learning rate dynamically adjusted by a factor of 0.01 upon encountering a plateaued validation loss.

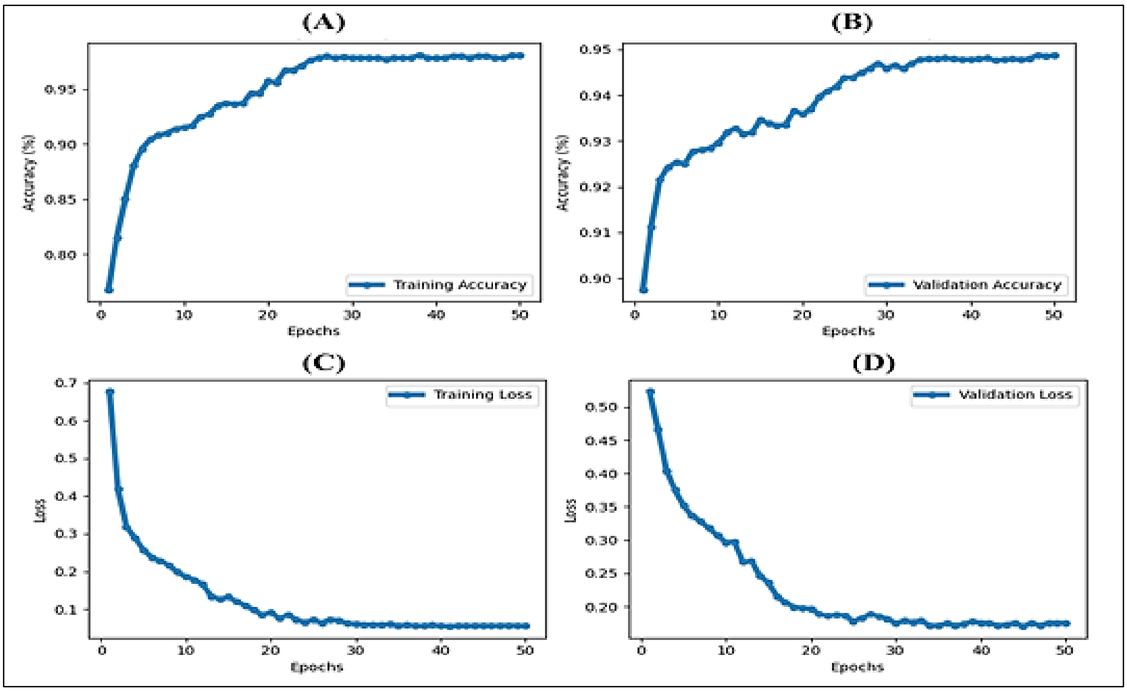


Figure 7: Training and Validation Results — (A) Training Accuracy, (B) Validation Accuracy, (C) Training Loss, (D) Validation Loss

Weight decay with a coefficient of 0.001 penalized large weights in the network. This approach helps the model learn more features and reduces the over fitting of the training data. These combined strategies help the model to converge to a better minimum and avoid over fitting. Figure 7 shows the accuracy and loss curves during the training.

Discussion

We conducted a rigorous experimental analysis to assess the effectiveness of the proposed hybrid model in classifying endoscopic images. This analysis compares the model's performance against established benchmarks on a well-known endoscopic image dataset. The dataset includes diverse gastrointestinal (GI) pathologies, allowing the model to learn and generalize across various disease presentations. Standard pre-processing techniques ensured data quality and consistency, enhancing the model's performance. A comprehensive suite of metrics, commonly used in multiclass classification tasks, assessed the model's ability to differentiate accurately between classes. Precision measures the proportion of correctly identified positive cases, whereas recall focuses on the model's ability to capture all true positives. Accuracy provides a general overview of the overall accuracy of the model. The F1 score offers a balanced view by combining precision and recall. The Matthews Correlation Coefficient (MCC) incorporates true negatives and false positives, providing a more robust evaluation of imbalanced datasets. Figure 8 shows the confusion matrix, which visually shows the model's performance by displaying the distribution of correct and incorrect classifications across all classes.

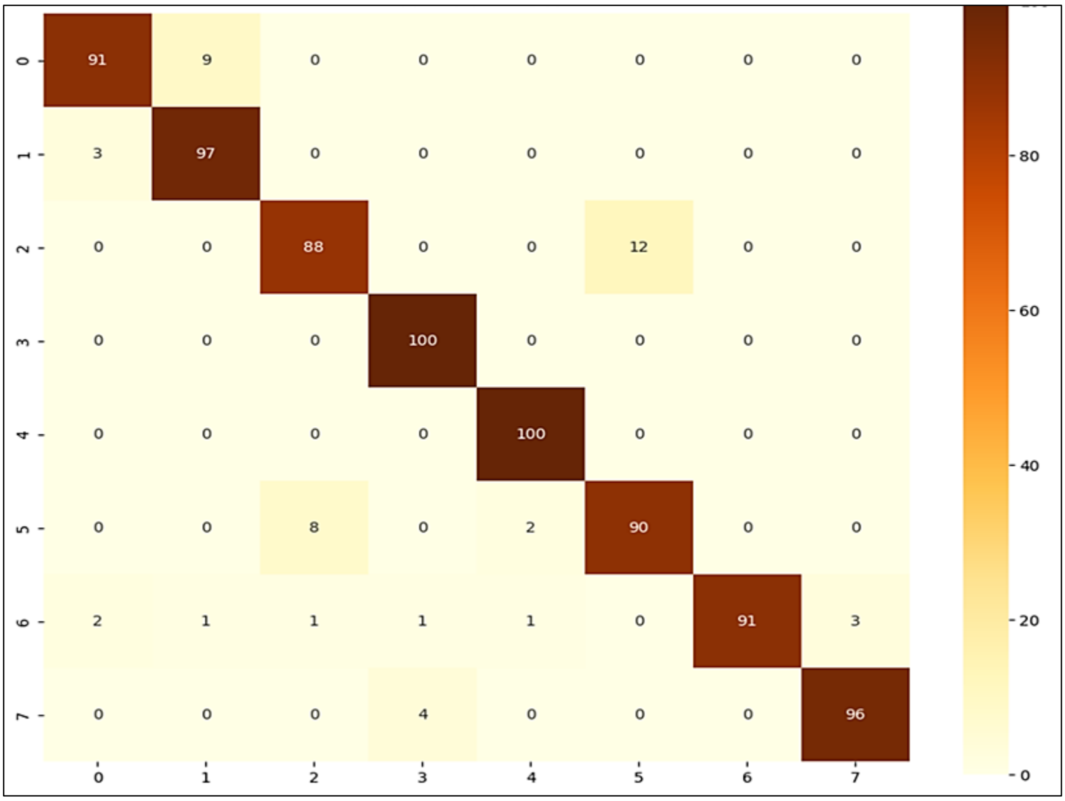


Figure 8: This Figure Depicts the Confusion Matrix Generated by the Model

The confusion matrix visualizes the performance of the model for each class. The labels (0-7) correspond to the following gastrointestinal conditions: (0) Dyed and Lifted Polyps, (1) Dyed Dissection Margins, (2) Esophagitis, (3) Normal Cecum, (4) Normal Pylorus, (5) Normal Z-Line, (6) Polyps and (7) Ulcerative Colitis. The receiver operating characteristic (ROC) curve and area under the curve (AUC) assess the model's ability to discriminate between positive and negative instances. The precision-recall curve visualizes the trade-off between precision and recall for different classification thresholds. Figure 9 displays the performance evaluation of the proposed model on both ROC curves and AUC scores (figure 9A), as well as precision-recall curves and AUC scores (figure 9B). This comprehensive evaluation strategy allows a thorough understanding of the strengths and weaknesses of the model in the context of endoscopic images for GI disease identification.

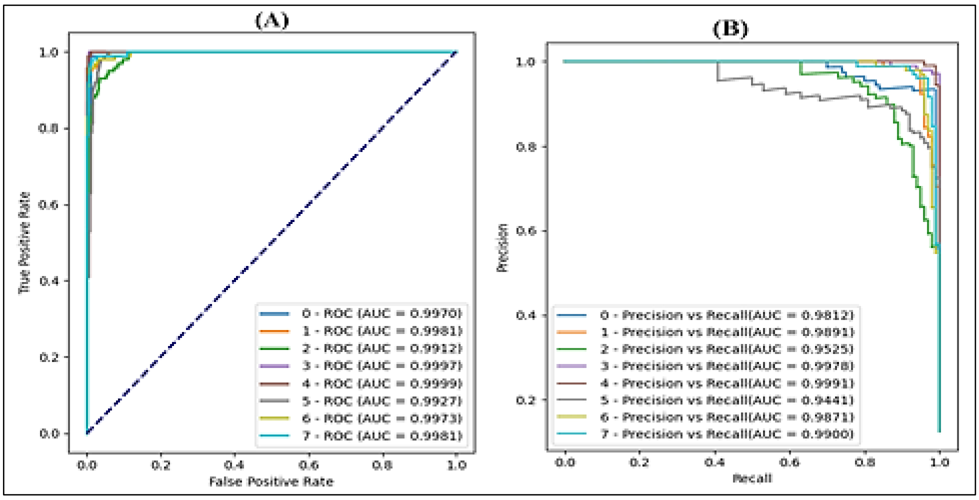


Figure 9: Performance Evaluation of the Proposed Model on 8 GI Disease Classes using ROC Curves and AUC Scores (A), as well as Precision-Recall Curves and AUC scores (B)

The labels (0-7) correspond to the following gastrointestinal conditions: (0) Dyed and Lifted Polyps, (1) Dyed Dissection Margins, (2) Esophagitis, (3) Normal Cecum, (4) Normal Pylorus, (5) Normal Z-Line, (6) Polyps, and (7) Ulcerative Colitis. We conducted an ablation study to evaluate the contributions of individual components within the proposed architecture. This involved systematically removing and reintegrating crucial elements, specifically, the Convolutional Block Attention Module (CBAM) and Vision Transformer (ViT) modules. Table 1 summarizes the results of the ablation studies. The model incorporating CBAM and ViT achieved the highest performance metrics: precision, recall, accuracy, and an F1 score of approximately 94%. This combined representation benefits from the feature concatenation of the CBAM-augmented CNN and ViT modules, suggesting that the attention mechanism significantly enhances the model's performance. Models using the CBAM block or the ViT module alone performed worse than the combined model but still showed respectable metrics. Excluding the ViT module resulted in a noticeable decline in performance, implying that relying solely on CNN features without an attention mechanism reduced the model's effectiveness. The absence of the CBAM attention mechanism and ViT module led to the poorest results. As shown in Table 1, the ablation study results prove that integrating the CBAM attention mechanism and ViT model improves the model's ability to distill essential image features from wireless-capsule endoscopic data. This enhanced representational capacity likely drives the superior performance metrics obtained using the full model configuration.

Table 1: Summarizes the Ablation Study Investigating the Impact of Individual Components within the Proposed Model

Efficient Net	CBAM	VIT	Accuracy	Precision	Recall	F1 score	MCC
✔	x	x	91.55	91.58	91.55	91.56	90.03
✔	✔	x	92.24	92.38	92.24	92.25	91.02
✔	x	✔	92.25	92.47	92.25	92.23	90.89
✔	✔	✔	94.13	94.21	94.13	94.11	93.30

Table 2 provides a detailed comparative evaluation of the performance of the proposed hybrid model and the established state-of-the-art methodologies for endoscopic image classification. This benchmark comparison highlights the model's effectiveness relative to other approaches that use the Kvasir dataset. The proposed method was designed to demonstrate superior performance through key metrics, such as accuracy, precision, recall, and F1-score. As shown in Table 2, the proposed method achieved an accuracy of 94.13%. It performs better than other state-of-the-art methods: ResNet-152 with Grad-CAM at 93.46%, Multi-model Classification at 90.20%, custom CNN for two-stage classification at 88.00%, FocalConvNet at 63.7% and MobileNetv2 at 79.15%. This strong performance was due to the combined effects of the hybrid architecture. The EfficientNet-B1 backbone excels in extracting fine-grained local features, which are crucial for identifying subtle textural changes indicative of disease. Simultaneously, the Vision Transformer effectively captured long-range dependencies, allowing the model to understand the context and relationships between different regions within the endoscopic image. CBAM further improved this by guiding the model to the most important spatial and channel-wise features. This allows it to capture critical details that other architectures often miss. This thorough benchmarking process is essential for gaining valuable insights into the effectiveness of the proposed model in classifying gastrointestinal (GI) diseases.

Table 2: Benchmarking Performance: Proposed Method vs. State-of-the-Art Approaches

Author	Methods	Accuracy
(Srivastava *et al.,* 2022) (19)	FocalConvNet	63.7%
(Sandler *et al.,* 2018) (20)	MobileNetv2	79.15 %
(Pozdeev *et al.,* 2019) (21)	Custom CNN for two-stage classification	88.00%

(Fonolla *et al.,* 2019) (22)	Multi-model Classification	90.20%
(Mukhtorov et al., 2023) (23)	ResNet-152 combined with Grad–CAM	93.46 %
The proposed method	CBAM integrated hybrid efficient net b1-VIT	94.13%

The high accuracy of our model has promising implications in clinical use. Imagine this model integrated into endoscopy systems to assist clinicians in real-time. This could lead to earlier and more accurate detection of precancerous lesions, such as subtle polyps, which is crucial for preventing colorectal cancer. By identifying suspicious areas, the model could lower the chances of missed diagnoses. This would be especially helpful for less experienced endoscopists and could improve the overall quality of the endoscopic procedures. The objective nature of AI-driven analysis can also reduce the variability among practitioners. This would lead to more consistent diagnostic results across different clinicians and healthcare centres. The CBAM module plays a key role in boosting performance. However, further research on the interpretability of this model would be beneficial. Tools such as Grad-CAM can help visualize the areas the model focuses on during predictions. This offers valuable insights into the features that the model uses to distinguish between GI conditions. It could also reveal new diagnostic markers and deepen clinicians' understanding of the disease patterns.

Conclusion

This paper presents a novel deep-learning framework for endoscopy image analysis that combines a Vision Transformer (ViT) with a CNN enhanced by a Convolutional Block Attention Module (CBAM). This hybrid architecture leverages ViT's global feature extraction capacity and CNN's local feature extraction ability, with CBAM further enhancing the attentional focus. We assessed the efficacy of the proposed model using the publicly available Kvasir wireless endoscopy dataset, which achieved an accuracy of 94.13%, surpassing the performance metrics reported for contemporary benchmark methods. These results demonstrate that the hybrid CNN-transformer framework enhanced with CBAM effectively captures local and global features in endoscopic images. The integration of the CBAM improved the model's attentional focus and contributed to its robust performance. The high accuracy achieved by the model suggests its potential for more precise and reliable detection of gastrointestinal disease, which is crucial for early diagnosis and treatment. Despite its promising performance, the model has limitations in accurately classifying certain cases, particularly those with subtle anomalies. To evaluate the clinical utility of the model, thorough validation using a wide range of endoscopic image datasets, including data from different clinical sites and various types of equipment, is necessary.

Future Scope

Future research will prioritize expanded testing and explore the potential advantages of ensemble learning techniques for improved decision-making. In addition, we will focus on reducing the computational complexity of the model to achieve faster and more efficient processing. This ongoing study aims to make the hybrid model more efficient and effective for clinical use, potentially revolutionizing endoscopic image analysis and improving patient outcomes through timely and accurate diagnosis.

Abbreviation

Nil.

Acknowledgement

Nil.

Authors Contribution

Pradeepan P. and Dr. Gladston Raj S. conducted the experiments and wrote the manuscript. Dr. Gladston Raj S. and Dr. Juby George K. reviewed and edited the manuscript. All authors have read and agreed to the final version of the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

Ethics Approval

Not applicable.

Funding

Nil.

References

1. Jemal A, Siegel R, Xu J, Ward E. Cancer statistics, 2010. CA: a cancer journal for clinicians. 2010; 60(5):277-300.
2. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality

worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2018; 68(6):394-424.

3. Rex DK, Johnson DA, Anderson JC, Schoenfeld PS, Burke CA, Inadomi JM. American College of Gastroenterology guidelines for colorectal cancer screening 2008. Official journal of the American College of Gastroenterology| ACG. 2009; 104(3):739-50.
4. Repici A, Badalamenti M, Maselli R, Correale L, Radaelli F, Rondonotti E, Ferrara E, Spadaccini M, Alkandari A, Fugazza A, Anderloni A. Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. Gastroenterology. 2020; 159(2):512-20.
5. Ali S, Zhou F, Bailey A, Braden B, East JE, Lu X, Rittscher J. A deep learning framework for quality assessment and restoration in video endoscopy. Medical image analysis. 2021; 68:101900.
6. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017; 542(7639):115-8.
7. Litjens G, Kooi T, Bejnordi BE, Setio AA, Ciompi F, Ghafoorian M, Van Der Laak JA, Van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Medical image analysis. 2017; 42:60-88.
8. Yu L, Chen H, Dou Q, Qin J, Heng PA. Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE transactions on medical imaging. 2016; 36(4):994-1004.
9. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, Tang Y, Xiao A, Xu C, Xu Y, Yang Z. A survey on vision transformer. IEEE transactions on pattern analysis and machine intelligence. 2022; 45(1):87-110.
10. Tang S, Yu X, Cheang CF, Liang Y, Zhao P, Yu HH, Choi IC. Transformer-based multi-task learning for classification and segmentation of gastrointestinal tract endoscopic images. Computers in Biology and Medicine. 2023; 157:106723.
11. Wu X, Feng Y, Xu H, Lin Z, Chen T, Li S, Qiu S, Liu Q, Ma Y, Zhang S. CTransCNN: Combining transformer and CNN in multilabel medical image classification. Knowledge-Based Systems. 2023; 281:111030.
12. Pan X, Ge C, Lu R, Song S, Chen G, Huang Z, Huang G. On the integration of self-attention and convolution. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2022; p. 815-825. https://doi.org/10.48550/arXiv.2111.14556
13. Wang Y, Guo J, Yang Y, Kang Y, Xia Y, Li Z, Duan Y, Wang K. CWC-transformer: a visual transformer approach for compressed whole slide image classification. Neural Computing and Applications. 2023; 1-13. https://doi.org/10.1007/s00521-022-07857-3
14. Woo S, Park J, Lee JY, Kweon IS. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV) 2018; p. 3-19. https://doi.org/10.48550/arXiv.1807.06521
15. Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. International conference on machine learning 2019 May 24; p. 6105-6114. PMLR. https://proceedings.mlr.press/v97/tan19a.html
16. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018; 7132-7141. https://doi.org/10.1109/CVPR.2018.00745
17. Dietterich TG. Ensemble methods in machine learning. In International workshop on multiple classifier systems 2000 Jun 21; p. 1-15. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45014-9_1
18. Pogorelov K, Randel KR, Griwodz C, Eskeland SL, de Lange T, Johansen D, Spampinato C, Dang-Nguyen DT, Lux M, Schmidt PT, Riegler M. Kvasir: A multi-class image dataset for computer-aided gastrointestinal disease detection. In Proceedings of the 8th ACM on Multimedia Systems Conference 2017 Jun 20; p. 164-169. https://doi.org/10.1145/3193289
19. Srivastava A, Tomar NK, Bagci U, Jha D. Video capsule endoscopy classification using focal modulation guided convolutional neural network. In 2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS) 2022 Jul 21; p. 323-328. IEEE. https://www.doi.org/10.1109/CBMS55023.2022.00064
20. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018; p. 4510-4520. https://doi.org/10.48550/arXiv.1801.04381
21. Pozdeev AA, Obukhova NA, Motyko AA. Automatic analysis of endoscopic images for polyps detection and segmentation. In2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus) 2019 Jan 28; p. 1216-1220. IEEE. https://www.doi.org/10.1109/EIConRus.2019.8657018
22. Fonolla R, van der Sommen F, Schreuder RM, Schoon EJ, de With PH. Multi-modal classification of polyp malignancy using CNN features with balanced class augmentation. In2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019) 2019 Apr 8; p. 74-78. IEEE. https://www.doi.org/10.1109/ISBI.2019.8759320
23. Mukhtorov D, Rakhmonova M, Muksimova S, Cho YI. Endoscopic image classification based on explainable deep learning. Sensors. 2023; 23(6):3176.