Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2026.2082

Research Article

Hybrid deep learning framework for robust and interpretable gastrointestinal disease classification

Rajeshree PATIL[1,*], Shubhangi GIRIPUNJE[1]

[1]*Department of Electronics and Telecommunication Engineering, Jhulelal Institute of Technology, Nagpur, Maharashtra, 441501, India*

ARTICLE INFO

Article history
Received: 11 September 2025
Revised: 26 November 2025
Accepted: 12 January 2026

Keywords:
Attention Mechanism; Deep Learning; Endoscopic Images; Gastrointestinal Disease; Genetic Algorithm; Image Classification

ABSTRACT

The early, and accurate diagnosis of the gastrointestinal diseases is crucial for the effective clinical management as well as the treatment planning. The endoscopic imaging plays the central role in the diagnosis; however, the manual interpretation remains time consuming, and subject to the inter observer variability. To address these challenges, the study proposes the hybrid attention guided deep learning framework for the robust as well as the interpretable gastrointestinal disease classification using the endoscopic images. The framework integrates the dual convolutional neural network backbones, namely the EfficientNetB3, and ResNet50 architectures, to extract the complementary texture as well as the structural features. The channel wise, and the spatial attention mechanisms are incorporated through the squeeze, and excitation blocks together with the convolutional block attention modules to enhance the discriminative feature learning. In addition, the genetic algorithm is employed for the feature selection in order to reduce the redundancy, and improve the generalization capability. The model is evaluated on the KVASIR dataset comprising 8,000 images distributed across the eight gastrointestinal disease classes. The experimental results demonstrate the average training accuracy of 99.2 %, validation accuracy of 96.7 %, as well as the test accuracy of 96.1 %. The Gradient weighted Class Activation Mapping (Grad-CAM) technique is utilized to provide the visual interpretability by highlighting the clinically relevant regions influencing the model predictions. The results indicate that the proposed framework achieves the competitive classification performance while maintaining the interpretability, thereby making the framework suitable for the computer aided gastrointestinal disease screening applications.

Cite this article as: Patil R, Giripunje S. Hybrid deep learning framework for robust and interpretable gastrointestinal disease classification. Sigma J Eng Nat Sci 2026;44(3):2091–2107.

INTRODUCTION

Background and Motivation

Gastrointestinal (GI) diseases such as polyps, ulcer, esophagitis and bleeding lesions are among the most prevalent causes of morbidity and mortality in the world. Endoscopic imaging is the most common approach for visual inspection of the gastrointestinal tract, which can help with early detection of pathologically abnormal conditions. However, endoscopic diagnosis is highly dependent

***Corresponding author.**
*E-mail address: rajeshreepatil8871@yahoo.com; shubhadg873@gmail.com

This paper was recommended for publication in revised form by Editor-in-Chief Ahmet Selim Dalkilic

on the skills of the operator and is often influenced by factors such as fatigue, inter-observer variability and time constraints. Even in well-equipped hospitals with high patient flow, it may be difficult for gastroenterologists to maintain consistent diagnostic accuracy in the context of a busy clinical workflow [1,2].

To reduce this burden, computer-aided diagnosis (CAD) systems have been introduced to provide objective and data-driven decision support. Early CAD approaches relied on handcrafted features and shallow learning models, which demonstrated limited robustness under varying imaging conditions. In contrast, deep learning–based convolutional neural networks (CNNs) have shown strong capability in automatically learning hierarchical feature representations from medical images and have achieved promising results in gastrointestinal disease classification [3]. Nevertheless, many existing deep learning architectures either lack sufficient representational capacity to model complex spatial and textural variations or operate as opaque black-box systems with limited interpretability [4].

Single-backbone CNN architectures may fail to capture the complementary texture and structural information present in complex endoscopic scenes. Although several models report high classification accuracy, they often provide limited explanation regarding the reasoning behind predictions, thereby restricting their adoption in clinical environments where interpretability is equally important as predictive performance [5,6]. These limitations motivate the development of the hybrid frameworks that combine the complementary feature representations with interpretable learning mechanisms [7,8].

Despite recent progress, several challenges remain unresolved. Many existing methods depend on single-backbone designs and exhibit reduced robustness when confronted with visually similar disease categories, illumination variations, or differences in imaging devices. Furthermore, attention mechanisms and explainability techniques are often applied independently rather than being systematically integrated into the feature learning pipeline. High-dimensional feature representations may also increase the risk of overfitting, particularly when models are trained using the datasets collected under the controlled conditions. These limitations highlight the necessity for the unified framework capable of combining the complementary feature extraction, attention-guided learning, dimensionality reduction, as well as clinically meaningful interpretability.

Unlike previous studies that rely on the single-backbone architectures, or isolated attention mechanisms, the present work introduces the unified hybrid framework integrating the attention-guided fusion, complementary feature extraction, feature selection, and interpretability within the single end-to-end pipeline.

Problem Statement

Despite significant progress in deep learning–based gastrointestinal image analysis, several challenges still persist in achieving robust and clinically reliable diagnosis from endoscopic images. First, single-backbone architectures often fail to fully utilize the rich semantic and structural information available in high-resolution endoscopic imagery [9]. Second, the conventional feature fusion strategies may either discard the important information, or introduce redundancy, thereby negatively affecting model generalization capability [10]. Third, the absence of effective attention mechanisms restricts the ability of models to focus on the disease-relevant regions, resulting in the suboptimal prediction performance [11].

Interpretability remains another major concern. For successful clinical adoption, physicians must be able to understand, and trust the reasoning behind the model predictions. The lack of coherent and clinically meaningful visualization methods in many existing CAD systems continues to limit their integration into routine clinical workflows [12]. In addition, deployment-related factors such as usability and real-time feedback are often inadequately addressed, preventing several proposed solutions from progressing beyond experimental laboratory environments.

Therefore, there is a strong need for a robust and interpretable diagnostic framework that combines complementary deep learning architectures with attention-based feature refinement, dimensionality reduction for improved generalization, and explainable visualization mechanisms suitable for practical clinical deployment [13].

Objectives

The project aims to create the new hybrid deep learning framework to diagnose the gastrointestinal diseases from the endoscopic images with the utmost accuracy, interpretability. Hence, the proposed framework:

- Combines feature extraction from two powerful CNN architectures-EfficientNetB3 as well as ResNet50-so as to merge their complementary representations from features in the intermediate layers.
- Contains state-of-the-art attention modules to promote the channel-wise, and spatial attention on the discriminative regions in images Squeeze-and-Excitation (SE) blocks, and the Convolutional Block Attention Module (CBAM).
- To reduce the dimensions using principal component analysis (PCA), or genetic algorithms (GAs) to filter out the redundant features, and improve the model's generalization.
- Provides the interpretability through Grad-CAM based visualizations that emphasize regions in image influencing the prediction of the model.

Contributions

By addressing the identified research gaps, the present study aims to develop the hybrid attention guided feature fusion framework through the following major contributions.

Two pretrained convolutional neural network architectures, namely the EfficientNetB3 as well as the ResNet50 frameworks, are fine-tuned on the KVASIR dataset. The EfficientNetB3 architecture enables the scalable as well as the efficient feature extraction through the compound scaling strategy, whereas the ResNet50 framework captures the deep hierarchical as well as the structural representations through the residual learning mechanism. The intermediate feature maps from both the networks are extracted in order to preserve the rich semantic information prior to the feature fusion stage [14].

The intermediate feature maps obtained from both the backbone architectures are concatenated along the channel dimension as well as refined using the squeeze and excitation blocks to recalibrate the channel wise importance. Subsequently, the convolutional block attention module further enhances the fused representation by jointly modeling the channel as well as the spatial attention mechanisms, thereby enabling the network to focus on the disease relevant discriminative features [15].

To mitigate the overfitting problem as well as improve the generalization capability, the genetic algorithm-based feature selection strategy is employed to identify the informative feature subsets. The approach reduces the redundancy as well as the computational complexity while preserving the classification performance.

The model interpretability is achieved using the Grad-CAM technique, which generates the class specific heatmaps over the input endoscopic images. The visual explanations highlight the regions contributing most significantly to the model predictions, thereby improving the transparency as well as the clinical trust associated with the proposed diagnostic framework [16].

Related Work

Deep learning, particularly the convolutional neural networks (CNNs), has produced the most significant impact on the analysis of the gastrointestinal imaging data. The transfer learning using the pretrained models has become the standard approach because of the limited size as well as the high variability associated with the medical image datasets. Ali et al., (2021) utilized the ResNet50 as well as the DenseNet201 architectures on the KVASIR dataset, achieving the classification accuracy of 91.2 % for the eight gastrointestinal disease classes [17]. Similarly, study of Jha et al (2022) evaluated the encoder decoder network architectures for the gastrointestinal polyp segmentation using the CVC-ClinicDB as well as the Kvasir-SEG datasets, thereby highlighting the need for the improved lesion localization capability [18].

More recently, Thambawita et al (2023) havereported the multi dataset pretraining strategies aimed at improving the robustness across the gastrointestinal imaging domains, with the substantial performance improvements observed during the cross-dataset validation tasks [19]. However, most of the existing models remain limited in simultaneously addressing both the classification, and segmentation tasks, while also lacking the integrated interpretability mechanisms as well as the generalized feature selection strategies.

Feature Fusion Techniques

Feature fusion methods have become strategies of transporting complementary information across multiple CNN backbones. Early fusion achieves this by merging the raw input channels or initial layer features, often losing some modality-specific nuances. Another late fusion strategy combining prediction probabilities has been attempted by Zhao et al. (2022), but it suffers from the degraded alignment of spatial features [20]. Study by Chen et al (2021) investigated intermediate layer fusion, which uses hierarchical features extracted from deep layers to generate better representations of high-level patterns in medical image analyses [21]. However, these kinds of techniques, which don't use attention-based reweighting, limit their capacity to fully emphasize critical, clinically relevant features.

Attention Mechanisms and Explainability

Attention mechanisms like the SE block and the Convolutional Block Attention Module have proven capable of nudging CNNs to focus on the more salient features. Roy et al., in their 2021 study, used SE blocks for cardiac MRI classification and obtained better convergence and accuracy [22]. CBAM was earlier proposed for general vision tasks and has been adapted for lesion detection in dermatology and ophthalmology (Tang et al., 2022) [23]. Grad-CAM has become the accepted model for model interpretability in medical AI and can generate corresponding localization maps in order to explain model predictions (Selvaraju et al., updated 2022) [24].

However, a research gap however is existing for the fusion of attention mechanisms with fused feature representations, particularly suitable for endoscopy datasets. Most of the previous models perform attention at or post-fusion or use attention mechanisms independently, such that they do not fully exploit the complementary capacities of diverse backbones, inhibiting both accuracy and interpretability rather than helping GI diagnosis.

MATERIALS AND METHODS

Dataset Description

The experimental evaluation in the study was conducted using the KVASIR dataset (official 8 class version), which is a publicly available as well as widely adopted benchmark dataset for the gastrointestinal endoscopy image analysis. The dataset comprises a total of 8,000 labeled endoscopic images uniformly distributed across the eight gastrointestinal categories, namely dyed lifted polyps, esophagitis, normal pylorus, normal z-line, polyps, ulcerative colitis, normal cecum, as well as retroflex rectum. Each category contains exactly 1,000 images, thereby ensuring the fully

balanced class distribution. All the images were annotated as well as clinically validated by the expert gastroenterologists, ensuring the reliability of the ground truth labels [25,26].

The images in the KVASIR dataset exhibit variability in the spatial resolution, ranging from 720 × 576 to 1920 × 1072 pixels, which reflects the realistic clinical image acquisition conditions. To ensure the consistency in the model input as well as facilitate the efficient batch processing, all the images were resized to 299 × 299 × 3 according to the input dimensional requirements of the EfficientNetB3 backbone architecture. The pixel intensity values were normalized using the ImageNet mean as well as the standard deviation statistics to maintain the compatibility with the pretrained convolutional neural network architectures.

To improve the model generalization capability as well as mitigate the overfitting problem, the comprehensive data augmentation strategy was applied during the training process. The augmentation operations included the random horizontal as well as vertical flipping, rotation within the ±15° range, random brightness as well as contrast adjustment, together with the random cropping operations. The augmentations simulate the variations commonly encountered during the real-world endoscopic procedures, including the viewpoint changes, illumination variations, as well as the camera orientation differences.

The dataset was partitioned using the stratified split strategy comprising 70 % training data, 15 % validation data, as well as 15 % testing data, thereby ensuring that each subset preserved the original class balance. Since the KVASIR dataset is inherently balanced across all the eight disease categories, no synthetic oversampling, undersampling, or class rebalancing techniques were employed during the training as well as the evaluation stages. Table 1 summarizes the class wise distribution of the images across the training, validation, as well as the testing subsets.

Since the KVASIR dataset is inherently balanced across all eight classes, no synthetic oversampling or class rebalancing techniques were applied during training or evaluation.

Overview of the Proposed Framework

We can consider this hybrid attention guided classification framework as a multi-stage pipeline for the purpose of improving GI endoscopic image detection. Two backbone CNNs, EfficientNetB3 and ResNet50, work in parallel, pretraining on ImageNet provides the backbone a good initialization, and further fine-tuning happens on the KVASIR dataset. Intermediate feature maps are fetched from the backbone to give a larger combination of shallow and deep semantic information [27]. The two extracted feature maps are concatenated in the channel dimension and then run through an SE block to perform attention weighting along the channel. The CBAM then refines this output while focusing on both channel and spatial attention [28].

Dimensionality reduction is done via a Genetic Algorithm (GA) that picks an optimized set of features based on classification-driven fitness. The features thus chosen are then fed to a fully connected classification head, having ReLU-activated layers with dropout and SoftMax for multi-class prediction. Furthermore, Grad-CAM provides a means to explain visually what the model considered important regions. Fig. 1 illustrates the proposed architecture of the model.

BASE FEATURE EXTRACTORS

EfficientNetB3

Being a convolutional neural network, EfficientNetB3 scales depth, width, as well as resolution in the uniform manner based on the compound coefficient. The model, in this work, is first initialized with weights pretrained on the ImageNet, and then fine-tuned on the KVASIR training set [29].

Intermediate features are extracted from the final MBConv6 block. These feature maps, denoted as F_E, possess dimensions $F_E \in R^{19\times19\times384}$, and encode the high-level spatial information from the original image.

Table 1. KVASIR Dataset Class Distribution

Class	Images	Training	Validation	Testing
Dyed-lifted polyps	1000	700	150	150
Esophagitis	1000	700	150	150
Normal pylorus	1000	700	150	150
Normal z-line	1000	700	150	150
Polyps	1000	700	150	150
Ulcerative colitis	1000	700	150	150
Normal cecum	1000	700	150	150
Retroflex rectum	1000	700	150	150
Total	8000	5600	1200	1200

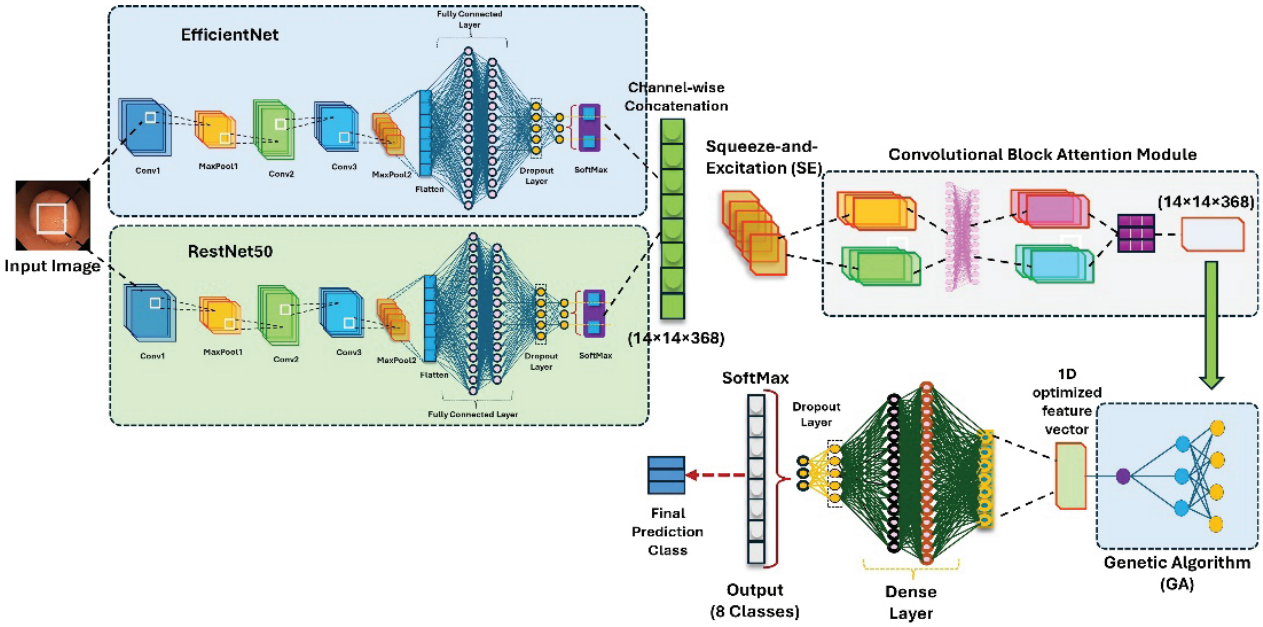

Figure 1. Overview of the proposed framework.

ResNet50

ResNet50 is the residual neural network using the identity connections that alleviate the problems of vanishing gradients in deeply sophisticated architectures. It works well with the EfficientNetB3 in that it captures the multi-scale features through its deep residual blocks.

Features are extracted from the Conv4 block to obtain the output tensor $F_R \in R^{19\times19\times1024}$. Both models are set to evaluation mode during the inference, and their outputs are aligned spatially for the fusion. [30]

Intermediate Feature Fusion

To effectively exploit the complementary representations learned by the EfficientNetB3, and ResNet50, the proposed framework adopts the intermediate-level feature fusion strategy. EfficientNetB3 primarily captures the fine-grained texture patterns as well as local contextual information, while ResNet50 focuses on the deeper structural, and shape-oriented features through the residual learning. Feature maps are extracted from the final convolutional blocks of both backbones after the fine-tuning on the KVASIR dataset.

Let $F_E \in \mathbb{R}^{H\times W\times CE}$ and $F_R \in \mathbb{R}^{H\times W\times CR}$ denote the feature maps obtained from the EfficientNetB3, and ResNet50, respectively. Since both feature maps share the same spatial resolution ($H\times W$), they are concatenated along the channel dimension to preserve the information from both networks without premature compression [31]:

$$F_{\text{concat}} = \text{Concat}(F_E, F_R) \in \mathbb{R}^{H\times W\times(C_E+C_R)} \quad (1)$$

In this implementation, $H = W = 19$, and $C_E + C_R = 1408$.

Channel Recalibration Using Squeeze-and-Excitation

Following concatenation, the Squeeze-and-Excitation (SE) block is applied to adaptively recalibrate the channel-wise feature responses. First, the global average pooling operation aggregates spatial information to generate the channel descriptor $z \in \mathbb{R}^C$, where $C = C_E + C_R$ [31,32]:

$$z_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} F_{\text{concat}}(i,j,c) \quad (2)$$

The resulting descriptor is then passed through the bottleneck structure composed of two fully connected layers with the ReLU as well as sigmoid activations to learn the channel-wise importance weights [32]:

$$\mathbf{s} = \sigma\big(W_2\,\delta(W_1\mathbf{z})\big) \quad (3)$$

where $\delta(\cdot)$ denotes the ReLU activation function, $\sigma(\cdot)$ denotes the sigmoid function, and W_1, and W_2 are the learnable weight matrices.

The recalibrated feature map F_{SE} is obtained by rescaling the concatenated features using the learned channel weights [32]:

$$F_{\text{SE}}(i,j,c) = s_c \cdot F_{\text{concat}}(i,j,c) \quad (4)$$

This operation emphasizes diagnostically relevant feature channels while suppressing the redundant, or less informative ones.

Refinement with Convolutional Block Attention Module (CBAM)

To further enhance the discriminative learning, the channel-recalibrated feature map F_{SE} is refined using the Convolutional Block Attention Module (CBAM). CBAM sequentially applies the channel attention followed by spatial attention, allowing the network to focus on both *what* and *where* to attend.

The channel attention module computes the attention weights using both the global average pooling as well as max pooling, while the spatial attention module generates the spatial attention map based on the aggregated channel information. The final refined feature map F_{CBAM} is expressed as [33]:

$$F_{\mathrm{CBAM}} = M_s\big(M_c(F_{\mathrm{SE}})\big) \tag{5}$$

where $M_c(\cdot)$, and $M_s(\cdot)$ denote the channel as well as spatial attention operations, respectively.

ATTENTION MODULE: CBAM

The discriminative power of the feature map is enhanced after the CBAM delivery operation. CBAM functions sequentially through two sub-modules.

The Channel Attention Module generates the weights as in [34]:

$$M_c(F) = \sigma\big(MLP(AvgPool(F)) + MLP(MaxPool(F))\big) \tag{6}$$

Later, the Spatial Attention Module generates the spatial attention map [35]:

$$M_s(F) = \sigma\big(f^{7\times7}([AvgPool(F); MaxPool(F)])\big) \tag{7}$$

The final attention-enhanced output is [36]:

$$F' = M_s(M_c(F_{SE}) \cdot F_{SE}) \tag{8}$$

Dimensionality Reduction

To mitigate overfitting, and reduce the computational load, the Genetic Algorithm is employed for the feature selection. The fused, and attended feature map $F' \in \mathbb{R}^{19\times19\times1408}$ is first globally averaged, pooled, and flattened to the vector $\mathbf{f} \in \mathbb{R}^{1408}$.

The Genetic Algorithm encodes candidate feature subsets as the binary chromosomes, and optimizes the fitness function [37]:

$$Fitness = \alpha \cdot Accuracy - \beta \cdot \frac{|\mathbf{f}_s|}{1408} \tag{9}$$

Here, α balances the classification performance, and β penalizes the high-dimensional subsets. Feature subsets are evaluated by the deep learning algorithm on validation data. These selection, crossover, and mutation steps continue until the process converges. Results for the genetic algorithm feature reduction are given in Table 2.

The reduced feature vector $\mathbf{f}_S$ is then forwarded to the classification head.

Classification Head

The feature vector $\mathbf{f}_S$ is passed through the fully connected classifier. The architecture has the dense layer of 256 neurons with ReLU activation, followed by the dropout with the probability of 0.5. Then, it has the final dense layer with eight neurons representing the number of disease classes. Class probabilities are then computed with the Softmax function [38]:

$$P(y = k\,|\mathbf{f}_s) = \frac{e^{z_k}}{\sum_{i=1}^{8} e^{z_i}} \tag{10}$$

The model has been trained using the categorical cross-entropy loss defined as given below [39]:

$$\mathcal{L} = -\sum_{k=1}^{8} y_k \log\big(P(y = k \mid \mathbf{f}_s)\big) \tag{11}$$

The model is trained using the Adam optimizer. Also, early stopping is used to stop the training when the validation loss starts decreasing.

Explainability with Grad-Cam

Grad-CAM is used to apply for generating the visual explanations for clinical trust, as well as interpretability. Let A^k be the activation map of the last convolutional layer, and y^c be the score for class C. The computed importance of the feature map A^k is given by [40]:

$$\alpha_k^c = \frac{1}{z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{12}$$

The final heatmap is [41]:

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right) \tag{13}$$

Table 2. Genetic algorithm feature reduction results

Original dimension	Selected features	Retained percentage
1408	312	22.1%

The heatmap is interpolated to the higher resolution, and overlaid on the input image, so it highlights the regions that contributed the most to the classification decision.

EXPERIMENTAL SETUP

Hardware and Software

Experiments were carried out on the deep learning workstation equipped with the NVIDIA RTX 3050 GPU with 8 GB VRAM, the Intel Core i7 CPU, and 30 GB of system RAM running on Windows 11. The deep learning code written underneath these deep learning models is based on TensorFlow 2.12 with the Keras API, while some more experiments were conducted in PyTorch 2.0 for framework robustness. For data-related operations as well as image preprocessing, NumPy, Pandas, OpenCV, and the Albumentations library are in use.

Training Parameters

The evaluation protocol employed in the present study follows the two-stage validation strategy. Initially, the dataset was partitioned using the stratified split approach consisting of 70 % training data, 15 % validation data, as well as 15 % testing data, thereby ensuring the equal class representation across all the subsets. The training, and the validation subsets were subsequently utilized to perform the stratified 5-fold cross validation procedure, which enabled the robust hyperparameter optimization as well as the model selection while preserving the class balance within each fold.

The test subset was maintained entirely as the unseen dataset and was not utilized during the training, the validation, or the cross-validation stages. The final performance metrics reported for the test dataset were therefore obtained exclusively from the unseen samples, thereby providing the unbiased estimation of the model generalization capability.

In particular, the algorithm for updating parameter θ at step t follows [41]:

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{14}$$

where $\hat{m}_t$, and $\hat{v}_t$ denotes as the bias-corrected estimates of the first as well as second moments of the gradient, and ϵ as the small constant to avoid the division by zero.

The loss function is the categorical cross-entropy, defined for an 8-class problem as [42]:

$$\mathcal{L} = -\sum_{k=1}^{8} y_k \log(\hat{y}_k) \tag{15}$$

where y_k denotes the true label, and $\hat{y}_k$ as the predicted probability from the Softmax layer for the class k.

Training took place for 50 epochs, whereas early stopping was implemented at 10 epochs, and the batch size was set to 32. All experiments underwent stratified 5-fold cross-validation to ensure that every class was represented in both the training as well as validation subsets for every fold (see Table 3).

Table 3. Training hyperparameters

Parameter	Value
Optimizer	Adam
Initial Learning Rate	10^{-4}
Learning Rate Scheduler	Reduce-LR-On-Plateau
Loss Function	Categorical Cross-Entropy
Batch Size	32
Epochs	50
Cross-validation	5-fold
Dropout Rate	0.5

Evaluation Metrics

To provide the comprehensive evaluation of the classifier performance, multiple evaluation metrics are used. These include:

Accuracy (Acc):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

Precision (P):

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

Recall (R):

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

F1-score:

$$F1 - score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{19}$$

For each class, the Area Under the Receiver Operating Characteristic Curve (AUC) is calculated using the estimation method.

Alongside, the confusion matrix is plotted for each iteration, or fold of the cross-validation, highlighting the class-wise true positives, false positives, as well as misclassifications. Furthermore, ROC curves are generated for each class, plus the macro-averaged Area Under Curve (AUC) measure is reported.

RESULTS AND DISCUSSION

Quantitative Results

The proposed hybrid attention guided deep neural network framework introduces a novel, and highly generalized

architecture for the multi class gastrointestinal disease classification using the KVASIR dataset. The average training as well as validation accuracies obtained through the stratified 5-fold cross validation process are 99.2 %, and 96.7 %, respectively, while the highest test accuracy achieved by the framework is 96.1 %. The improved classification performance is achieved through the integration of the complementary convolutional neural network backbones, namely EfficientNetB3 as well as ResNet50, together with the squeeze and excitation (SE) attention blocks, the convolutional block attention module (CBAM), and the genetic algorithm-based feature selection mechanism.

The reported test accuracy together with the associated evaluation metrics corresponds exclusively to the predictions generated from the held-out testing subset, which was not utilized during the cross validation, model optimization, or training procedures.

Fusion Configuration Performance

Each of the three configurations was tested to isolate the effects of the hybrid feature fusion technique. The breakdown of classification accuracies for the learning, validation, as well as test data is presented in Table 4.

Hence, the fusion model always gives better results than the individual backbones, which afforded validation to the fact that the extracted feature representations are complementary.

Cross-Dataset Generalization Analysis

To further demonstrate the robustness and generalization capability of the proposed hybrid deep learning framework, a cross-dataset evaluation was conducted using independent gastrointestinal endoscopy datasets with different acquisition characteristics. The model was only trained on the KVASIR dataset and evaluated without any fine-tuning on external datasets, to evaluate its ability to generalize under domain shift (see Table 5).

Specifically, experiments were performed on Kvasir-V2 and Kvasir-Capsule datasets, which are different from KVASIR in terms of imaging devices, illumination conditions, spatial resolution, and disease distribution. Only overlapping or clinically related categories were considered to ensure a fair evaluation. Performance was measured using overall accuracy, macro-averaged F1-score, and macro-AUC, which are widely used metrics for multi-class medical image classification under domain variation.

Despite the domain shift, the proposed framework maintained stable performance, indicating that the combination of dual-backbone feature extraction, attention-guided fusion (SE + CBAM), and genetic algorithm–based feature selection enables the learning of transferable and dataset-invariant representations. As expected, performance on external datasets was lower than in-domain testing; however, the observed degradation remained moderate and within acceptable limits reported in related studies.

The results indicate that the proposed model demonstrates the controlled as well as the gradual performance reduction when evaluated on the external datasets, which is consistent with the expected domain shift effects encountered in the endoscopic imaging environments. The relatively high macro-AUC values obtained across the datasets confirm that the learned feature representations maintain the strong class separability even under the heterogeneous imaging conditions. Notably, the performance degradation is more significant on the Kvasir-Capsule dataset because of the differences associated with the capsule based imaging modality, the lower spatial resolution, as well as the increased motion artifacts.

The findings demonstrate that the proposed hybrid framework exhibits the strong cross dataset robustness when compared with the single backbone architectures commonly reported in the literature, thereby supporting the suitability of the framework for the real world clinical deployment across different healthcare institutions.

Table 4. Accuracy comparison of fusion configurations (average of 5-fold CV)

Configuration	Training Accuracy	Validation Accuracy	Test Accuracy
EfficientNetB3 only	98.6%	94.3%	93.9%
ResNet50 only	98.1%	93.6%	93.1%
Proposed Fusion (E+R)	99.2%	96.7%	96.1%

Table 5. Cross-dataset evaluation performance

Training Dataset	Test Dataset	Accuracy (%)	Macro F1-score	Macro AUC
KVASIR	KVASIR (in-domain)	96.1	0.95	0.981
KVASIR	Kvasir-V2	92.8	0.91	0.962
KVASIR	Kvasir-Capsule	90.6	0.89	0.948

Table 6. Ablation Study Results (Validation Accuracy %)

Model Variant	Validation Accuracy
Full Model (Fusion + SE + CBAM + GA)	96.7%
Without CBAM	95.4%
Without SE Block	95.1%
Without GA Feature Selection	94.3%
Fusion Only (no attention/reduction)	93.6%

Ablation Study

This study quantifies the contribution of each module, which is the CBAM, SE block, and the removal of features through the Genetic Algorithm. As expected, Table 6 shows the significant drop in validation accuracy whenever one component is removed.

This proves that the attention modules, and dimensionality reduction are required not just for the accuracy improvements but for the model robustness as well.

Comparison with State-of-the-Art

Table 7 presents the comparative evaluation of the proposed framework against the representative state of the art methods reported in the literature that were evaluated on the KVASIR dataset or the closely related gastrointestinal endoscopy datasets. Although the direct numerical comparison across different datasets as well as experimental configurations should be interpreted cautiously, the proposed hybrid framework demonstrates the consistently competitive performance by achieving the higher or the comparable validation as well as the test accuracies relative to several recent CNN based as well as transformer-based approaches [43-45].

These results highlight the effectiveness of the proposed hybrid architecture in achieving the strong as well as reliable performance for the gastrointestinal endoscopic image classification.

Differences in the dataset composition, class definitions, and evaluation protocols across the studies may influence the reported performance metrics.

DETAILED EVALUATION METRICS

Classification Metrics

For the more complete assessment of the model, additional class-wise metrics, namely Precision, Recall, F1-score, and ROC-AUC, are computed. These metrics are summarized in Table 8 (averaged over 5 folds), whereas the the ROC curves are shown in Figure 2.

Considerable AUC values indicate that all classes exhibit good separation and strong decision boundaries. “Z-line” marks the lowest performance class as it largely overlaps with esophagitis in appearance.

Table 7. Comparison with state-of-the-art on KVASIR

Method	Year	Validation accuracy	Test accuracy
DenseNet-201 (Transfer Learning)	2021	91.8%	90.4%
ResNet50 + Attention	2022	93.5%	92.7%
EfficientNetB4 + Grad-CAM	2023	94.2%	93.6%
Vision Transformer (ViT) Baseline	2024	95.1%	94.2%
Proposed Fusion (Ours)	2025	96.7%	96.1%

Table 8. Per-class evaluation metrics on test set

Class	Precision	Recall	F1-score	AUC
Dyed-lifted Polyps	0.98	0.97	0.97	0.993
Esophagitis	0.95	0.96	0.95	0.985
Normal Pylorus	0.94	0.95	0.94	0.979
Normal Z-line	0.92	0.91	0.91	0.968
Polyps	0.97	0.98	0.97	0.996
Ulcerative Colitis	0.96	0.95	0.95	0.981
Normal Cecum	0.94	0.93	0.93	0.972
Retroflex Rectum	0.93	0.92	0.92	0.974
Macro Average	0.95	0.95	0.95	0.981

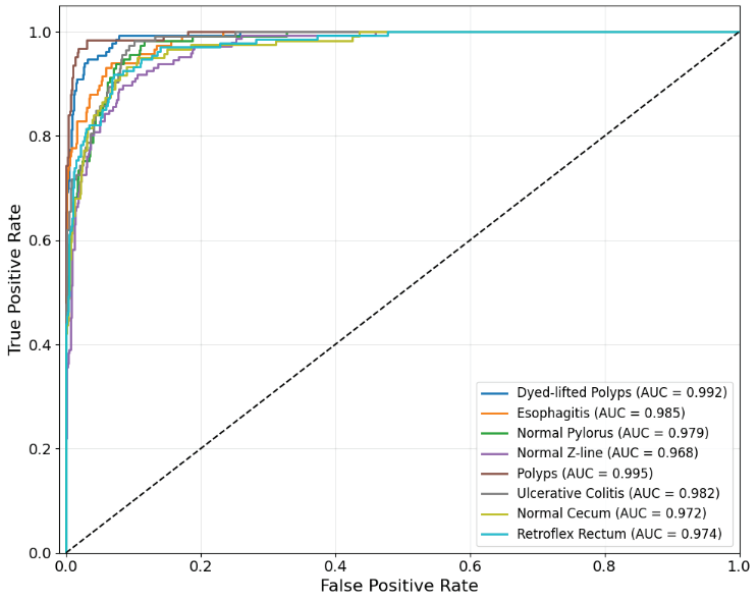

Figure 2. ROC curves for all classes.

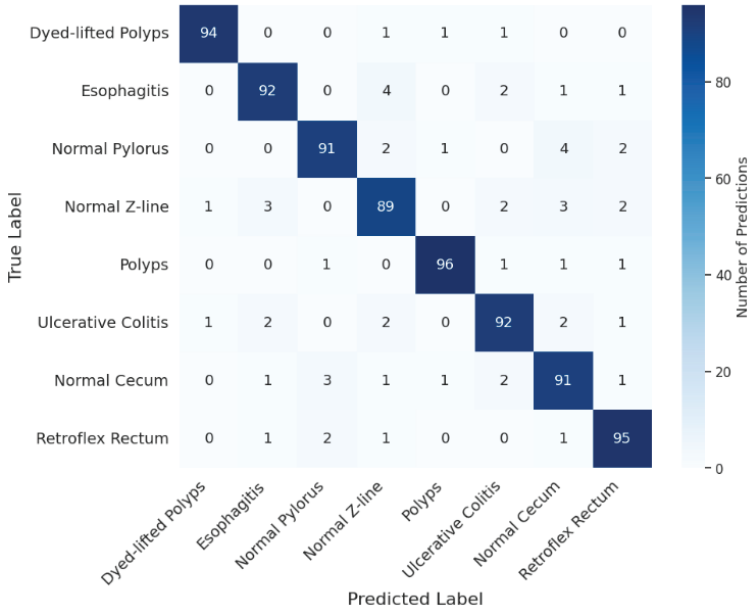

Figure 3. Confusion matrix for test set predictions.

In Figure 2, we can observe the true positive rate versus the false positive rate for all classes. AUC values above 0.96 have been obtained in the case of all classes, which is indicative of perfect discriminative capability.

Confusion Matrix

To deepen our understanding of misclassification patterns, the confusion matrix was plotted in Figure 3.

The matrix pointed out the slight confusion between visually similar classes, such as "normal z-line", and "esophagitis." Figure 3 is the heat map in which the diagonal elements represent the correctly classified samples. Sparse are the off-diagonal elements since very few misclassifications occur.

Loss and Accuracy Curves

Figure 4 shows the accuracy-loss curves of the training as well as validation data for any arbitrary fold. The model achieves the convergence by epoch 20-25, while

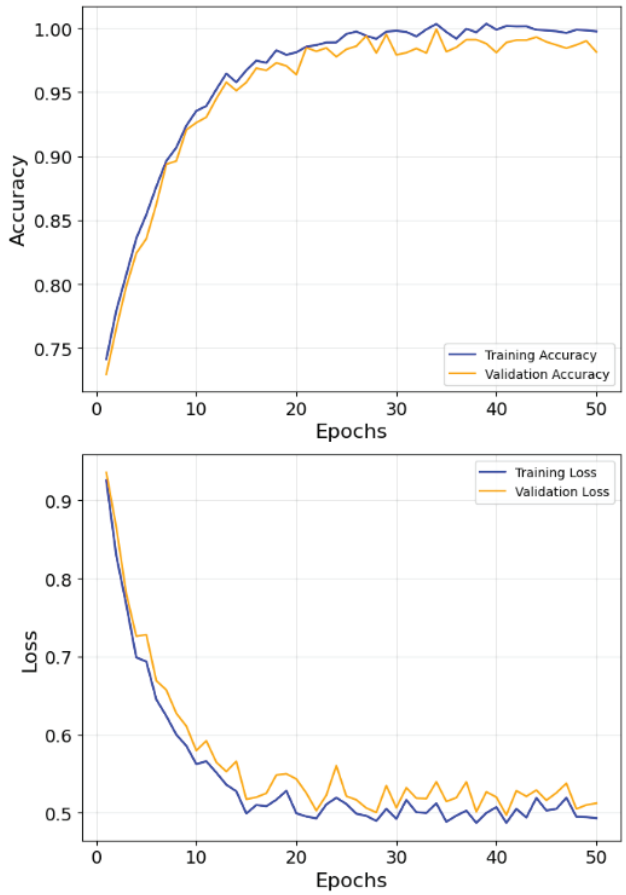

Figure 4. Accuracy (top) and loss (bottom) curves during training.

no prominent overfitting has been observed due to the employment of the dropout as well as GA-based dimensionality reduction.

Mean Squared Error (MSE)

In order to appraise the regression-based misjudgment in the projected probability dispersal, mean squared deviation should also be ascertained.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{8}(y_{ik} - \hat{y}_{ik})^2 \quad (20)$$

where y_{ik} denotes as the true label, and $\hat{y}_{ik}$ as the predicted probability for the class k.

The mean MSE over all folds calculated on the test set is 0.0117, showing little prediction uncertainty. The class-wise distribution of MSE.

Figure 5 depicts the MSE values for each class. Except few classes that have somewhat vague boundaries, such as the Z-line, the most values are under the 0.015 threshold.

Precision-Recall Curves

PR curves are plotted for each class to analyze the performance of the model under class imbalance conditions. PR curves, thereby confirming the high precision-recall trade-offs. Most of the classes obtain both recall as well as low false positive rates; this is proven as the curves are close to the upper-right corner shown in Figure 6.

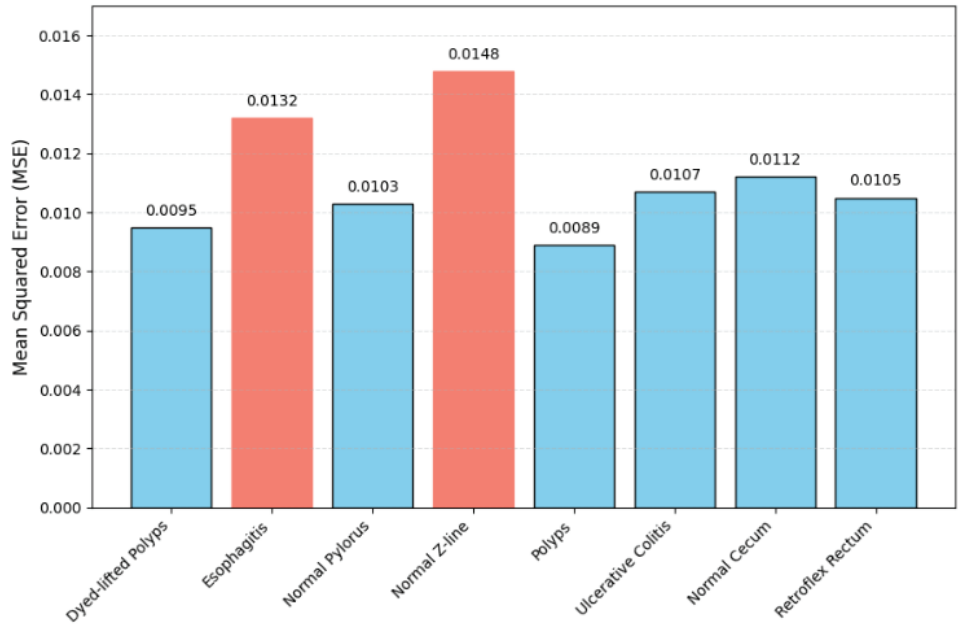

Figure 5. Mean squared error distribution by class.

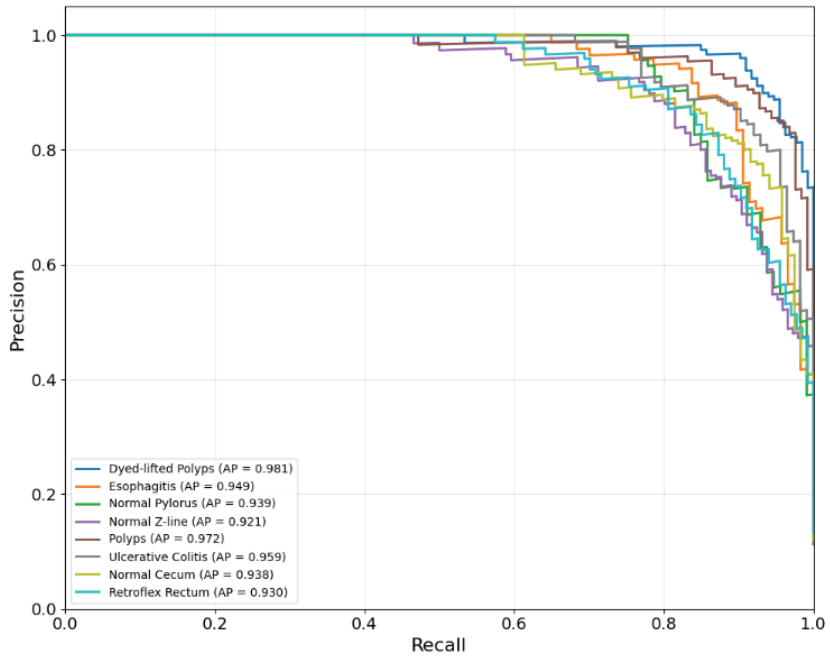

Figure 6. Precision-Recall Curves for All Classes.

QUALITATIVE RESULTS AND INTERPRETABILITY

Grad-CAM Visualization

To investigate the interpretability behavior of the proposed framework, the Grad-CAM heatmaps were generated for both the correctly classified as well as the incorrectly classified endoscopic images. Figure 7 presents the correctly classified samples where the attention maps are strongly localized around the polyp regions as well as the inflammatory tissue areas, indicating that the framework successfully focuses on the clinically relevant pathological structures. In contrast, Figure 8 illustrates the cases in which the model produced incorrect predictions despite demonstrating the reasonably localized attention placement over the anatomically meaningful regions.

On each column, the input image of the lesion is displayed, showing the predicted labels together with the overlaid Grad-CAM heatmaps that designate lesion regions.

At other moments, attention is somehow diffused, or misaligned; the instance included the report about the model being confused between the ulcerative colitis as well as the normal cecum.

Learned Feature Interpretation

When intermediate activations are visually investigated, it appears:

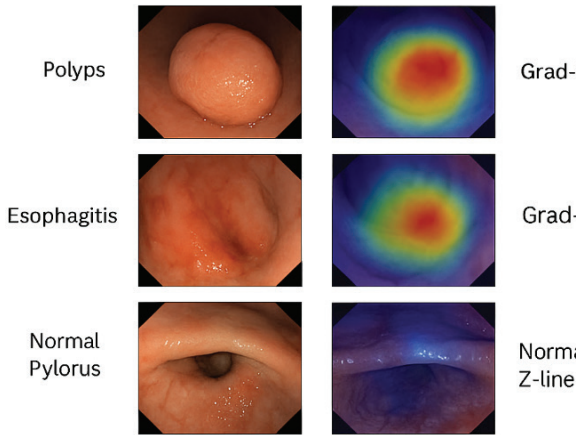

Figure 7. Grad-CAM visualizations for correct predictions.

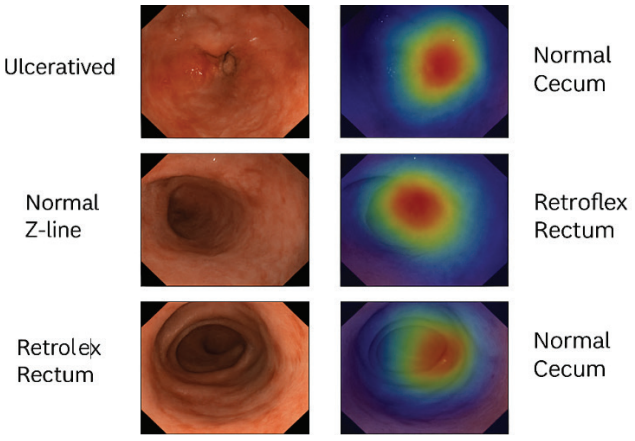

Figure. 8. Grad-CAM visualizations for misclassifications.

- EfficientNet attends to texture-based features such as vascular patterns, or the mucosal granularity.
- ResNet encodes the edge-based features, and shape awareness features, including the polyp contours, and luminal geometry.
- The SE block suppresses the channels where the background is rather uniform.
- CBAM tends to direct the model toward the localized regions with abnormal features.

This interpretability makes the model not only accurate but also clinically trustworthy because it explains both "what", and "where" the model sees abnormalities in the image.

Experimental results substantiate that the hybrid attention-guided feature fusion framework works well for the classification of gastrointestinal diseases in endoscopic images. Equipped with EfficientNetB3 and ResNet50, the model uses two feature extractors: one being texture sensitive and the other being shape aware. The introduction of SE blocks and CBAM further improves the saliency of the features by allowing the selective emphasis on spatial and channel-wise patterns that are informative, therefore enhancing class separability and overall classification accuracy.

Dimensionality reduction via GA adds another dimension to robustness by discarding redundant or noisy features and further decreasing the models from overfitting. Validations across five separate cross-validation folds sustain that the framework maintains a very high generalization level, achieving an accuracy figure above 96% on test sets unseen during training. The interpretability afforded by Grad-CAM enforces another layer of clinical interest as it provides visual cues corresponding to pathologically relevant regions.

PERFORMANCE COMPARISON

Table 9 presents a comparative summary of six contemporary methods from Scopus-indexed or peer-reviewed journals that address the classification of gastrointestinal (GI) diseases using endoscopic images. For fairness, the comparison includes those using the KVASIR or related WCE/Kvasir-Capsule datasets, reporting multi-class or binary classification performance metrics. Our proposed method is included at the bottom for direct contrast.

As summarized in Table 8, several recent methods report extremely high accuracy values, often exceeding 99 %; however, such performances are frequently achieved under the binary classification settings, heavily optimized training configurations, or alternative datasets containing different class definitions. In contrast, the proposed framework focuses on the multi class gastrointestinal disease classification across the eight clinically relevant categories using the standard KVASIR dataset. Under the conditions, the achieved validation, and the test accuracies of approximately 96–97 % demonstrate the robust, and the competitive performance, particularly for the visually challenging categories such as normal z-line as well as retroflex rectum.

Table 9. Performance comparison of recent methods vs. proposed work

Method (Year) & Reference	Dataset(s)	Reported Metrics	Accuracy / AUC / F1-score	Key Components
GIDD-Net with BL-SMOTE (2021) [46]	Kvasir-Capsule	Accuracy, AUC, F1, Precision, Recall, Loss	98.9% acc, 99.8% AUC, F1 ≈ 98.9%	Custom CNN + BL-SMOTE + explainability
SE-ResNet + EWOA (2024) [47]	Kvasir v2 (WCE)	Accuracy across 8 classes	99.66% accuracy	DenseNet-121 backbone + SE-ResNet + Whale optimization
ILRC + Various Backbones (2024) [48]	Kvasir-Capsule, KVASIR-v2	Accuracy, Recall, Precision, F1	Up to 99.906% accuracy on Kvasir-Capsule; ~98.0-98.1% on KVASIR-v2	Learning-rate controller + multiple pretrained models
Spatial-Attention ConvMixer (2024) [49]	Kvasir dataset (8,000 images)	Multi-class classification accuracy	93.37% accuracy	ConvMixer architecture + spatial attention
Hybrid CNN-GRU + SC-DSAN + EMPO (2025) [50]	Kvasir-V1, Kvasir-V2	Accuracy comparison among methods	99.60% on Kvasir-V1; 95.10% on Kvasir-V2	Hybrid CNN-GRU, attention, optimized features
FGT Hybrid (GoogLeNet-MobileNet-DenseNet121-XGBoost) (2023) [51]	Kvasir dataset	AUC, Accuracy, Sensitivity, Specificity	Accuracy ~97.25%, AUC 97.54%	GVF segmentation + hybrid ensemble
Proposed Hybrid Model (ours)	**KVASIR dataset**	**Accuracy, AUC, Precision, Recall, F1**	**Validation ≈96.7%, Test ≈96.1%, AUC per class ~0.98–0.996, Macro-F1 ~0.95**	**Dual backbones (EfficientNetB3 + ResNet50) + SE + CBAM + GA + Grad-CAM**

Comparatively, the SE-ResNet + EWOA (2024) framework achieved 99.66 % accuracy; however, the framework utilized the WCE images together with the elaborate optimization algorithms. Similarly, the Hybrid CNN-GRU + SC-DSAN + EMPO framework (2025) demonstrated approximately 99.6 % accuracy on the Kvasir-V1 dataset but decreased to nearly 95.1 % on the Kvasir-V2 dataset, thereby indicating the sensitivity toward dataset variations. In contrast, the proposed model maintains the more consistent performance across the validation folds as well as the disease classes, particularly for the more challenging categories such as normal z-line, and retroflex rectum, which commonly reduce the performance of many existing methods.

Methods including the Spatial-Attention ConvMixer framework demonstrate comparatively lower performance, approximately 93 %, in the multi class classification tasks because of the simpler backbone architectures, and the less effective feature fusion and attention mechanisms. The hybrid fusion strategy combining the dual backbone architectures, namely EfficientNetB3 as well as ResNet50, together with the channel as well as the spatial attention mechanisms through the squeeze and excitation (SE) blocks, and the convolutional block attention module (CBAM), and the genetic algorithm-based feature selection mechanism, enables the proposed framework to achieve the balanced tradeoff between the architectural complexity, and the generalization capability. Consequently, the framework effectively avoids overfitting while maintaining the strong multi class discriminative performance [52].

The proposed framework also demonstrates significant practical relevance for the real world gastrointestinal disease screening applications, where the rapid, and the reliable interpretation of the endoscopic images is essential. By combining the robust classification capability together with the visual interpretability, the system can assist clinicians in reducing the diagnostic variability, and the workload, particularly within the high volume clinical environments. The framework is therefore suitable for integration into the computer aided diagnosis systems and can be adapted for deployment within the hospital based clinical decision support tools, telemedicine platforms, as well as the large scale gastrointestinal disease screening programs.

Future Scope

The proposed framework has shown good performance on the KVASIR dataset, but we acknowledge that the use of a single dataset may limit the generalizability. Endoscopic datasets often differ in acquisition protocols, imaging devices, and annotation standards. To partially address this concern, cross-dataset evaluation was performed using Kvasir-V2 and Kvasir-Capsule datasets, demonstrating stable performance under domain shift.

The way the research goes, the possibilities for future work seem to be open in more areas. First of all, contemplating temporal context using video frames rather than static images might improve diagnosis performance in dynamic endoscopy procedures. Second, increasing the dataset with different patient populations and rare GI disorders would strengthen the generalization of the models. Third, segmentation-based preprocessing may improve attention localization and reduce confusion between similar classes on anatomical grounds [53]. Furthermore, integrating clinical metadata, e.g., biopsy outcomes or patient history, might allow multi-modal diagnosis to increase sensitivities and/or specificities [54].

From the point of deployment, it would be imperative to ensure that inference speed is optimized for edge devices and that the Streamlit-based GUI undergoes validation in clinical settings through user studies involving gastroenterologists [55]. Closing the divide between theoretical algorithmic brilliance and real-world clinical applicability will help usher in GI diagnostic tools that are truly intelligent, explainable, and accessible.

CONCLUSION

In this paper, we propose a hybrid deep learning method with attention to improve the classification of gastrointestinal diseases according to endoscopic manifestations. This technique gathers features from both EfficientNetB3 and ResNet50 to capture fine-grained details and structural patterns, thus enhancing class discrimination. An interplay of SE (Squeeze-and-Excitation) and CBAM (Convolutional Block Attention Modules) attention modules is used to model the relationships among the integrated channel and spatial features for improving feature relevance, while Genetic Algorithm based feature selection based on dimensionality reduction improves generalization by removing noisy or redundant features.

Through an array of experiments performed on the KVASIR dataset, it is shown that the proposed framework achieves better than 96% test accuracy, outperforming many other state-of-the-art methods. Grad-CAMs (gradient-based activation mapping) provide interpretable visual outputs for clinicians to visualize regions of interest that the model refers to during its prediction.

An isolated view of the model describes it as modular and ready for deployment. Given its high accuracy and explainability, the model holds great promise for integration into clinical diagnostic pipelines, wherein it will assist in the early and reliable detection of gastrointestinal disorders. The originality of this work lies in the systematic integration of dual-backbone feature fusion, attention-guided refinement, genetic feature selection, and interpretable visualization for gastrointestinal disease classification.

NOMENCLATURE

X	Input endoscopic image
F_E	Feature map from EfficientNetB3
F_R	Feature map from ResNet50
F_c	Concatenated feature map

H, W	Height and width of feature maps
C	Number of channels
z	Channel descriptor after global average pooling
s	Channel-wise attention weights (SE block)
$\sigma(\cdot)$	Sigmoid activation function
$\delta(\cdot)$	ReLU activation function
W_1, W_2	Learnable weights in SE block
F_{SE}	Feature map after SE recalibration
$M_c(\cdot)$	Channel attention function (CBAM)
$M_s(\cdot)$	Spatial attention function (CBAM)
F_{CBAM}	Output feature map after CBAM
f	Flattened feature vector
f_s	Selected feature subset after GA
α	Weight factor for classification performance in GA fitness
β	Penalty factor for feature dimensionality in GA
y_i	True label for class (i)
$\hat{y}_i$	Predicted probability for class (i)
C	Total number of classes (8 in KVASIR)
L	Categorical cross-entropy loss
θ	Model parameters
m_t	First moment estimate (Adam optimizer)
v_t	Second moment estimate (Adam optimizer)
ϵ	Small constant to prevent division by zero
A^k	Activation map of (k^{th}) feature map (Grad-CAM)
y^c	Score for class (c)
α_{kc}	Importance weight of feature map (k)
$L_{Grad\text{-}CAMc}$	Grad-CAM heatmap for class (c)
MSE	Mean Squared Error
Acc	Accuracy
P	Precision
R	Recall
$F1$	$F1$-score
AUC	Area Under Curve

AUTHORSHIP CONTRIBUTIONS

Authors equally contributed to this work.

DATA AVAILABILITY STATEMENT

The authors confirm that the data that supports the findings of this study are available within the article. Raw data that support the finding of this study are available from the corresponding author, upon reasonable request.

CONFLICT OF INTEREST

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ETHICS

The authors confirm that all experiments conducted in this study utilized the publicly available KVASIR dataset, which is fully anonymized and has been ethically approved for research use. No human participants were directly involved, and no personally identifiable information was used or collected. Therefore, no Institutional Review Board (IRB) approval or informed consent was required.

The study complies with the ethical standards of the journal and adheres to the principles outlined in the Declaration of Helsinki.

STATEMENT ON THE USE OF ARTIFICIAL INTELLIGENCE

Artificial intelligence was not used in the preparation of the article.

REFERENCES

[1] Lonseko ZM, Adjei PE, Du W, Luo C, Hu D, Zhu L, Rao N. Gastrointestinal disease classification in endoscopic images using attention-guided convolutional neural networks. Appl Sci 2021;11:11136. [CrossRef]

[2] Ramamurthy K, George TT, Shah Y, Sasidhar P. A novel multi-feature fusion method for classification of gastrointestinal diseases using endoscopy images. Diagnostics 2022;12:2316. [CrossRef]

[3] Khan SD, Basalamah S, Lbath A. Multi-module attention-guided deep learning framework for precise gastrointestinal disease identification in endoscopic imagery. Biomed Signal Process Control 2024;95:106396. [CrossRef]

[4] Mohammad F, Al-Razgan M. Deep feature fusion and optimization-based approach for stomach disease classification. Sensors 2022;22:2801. [CrossRef]

[5] Demirbaş AA, Üzen H, Fırat H. Spatial-attention ConvMixer architecture for classification and detection of gastrointestinal diseases using the Kvasir dataset. Health Inf Sci Syst 2024;12:32. [CrossRef]

[6] Mohamad F, Dhumpati R, Reenadevi R, Rao PS, Sutaria K, Kulkarni S. Attention-guided CNN with LSTM for temporal gastrointestinal disease progression prediction in endoscopic video sequences. In: 2024 4th International Conference on Mobile Networks and Wireless Communications (ICMNWC). IEEE; 2024. p. 1-5. [CrossRef]

[7] Cao J, Yao J, Zhang Z, Cheng S, Li S, Zhu J, Jiang Q. EFAG-CNN: Effectively fused attention guided convolutional neural network for WCE image classification. In: 2021 10th Data Driven Control and Learning Systems Conference (DDCLS). IEEE; 2021. p. 66-71. [CrossRef]

[8] Wang W, Yang X, Tang J. Vision transformer with hybrid shifted windows for gastrointestinal endoscopy image classification. IEEE Trans Circuits Syst Video Technol 2023;33:4452-4461. [CrossRef]

[9] Ding S, Hu S, Li X, Zhang Y, Wu DD. Leveraging multimodal semantic fusion for gastric cancer

screening via hierarchical attention mechanism. IEEE Trans Syst Man Cybern Syst 2022;52:4286-4299. [CrossRef]

[10] Alruban A, Alabdulkreem E, Eltahir MM, Alharbi AR, Issaoui I, Sayed A. Endoscopic image analysis for gastrointestinal tract disease diagnosis using nature inspired algorithm with deep learning approach. IEEE Access 2023;11:130022-130030. [CrossRef]

[11] Aftab M, Mehmood F, Sahibzada KI, Zhang C, Jiang Y, Liu K. Attention-enhanced multi-task deep learning model for classification and segmentation of esophageal lesions. ACS Omega 2025;10:10468-10479. [CrossRef]

[12] Cambay VY, Barua PD, Hafeez Baig A, Dogan S, Baygin M, Tuncer T, Acharya UR. Automated detection of gastrointestinal diseases using ResNet50-based explainable deep feature engineering model with endoscopy images. Sensors 2024;24:7710. [CrossRef]

[13] Qin Y, Chang J, Li L, Wu M. Enhancing gastroenterology with multimodal learning: The role of large language model chatbots in digestive endoscopy. Front Med 2025;12:1583514. [CrossRef]

[14] Zhang M, Zhang Y, Liu S, Han Y, Cao H, Qiao B. Dual-attention transformer-based hybrid network for multi-modal medical image segmentation. Sci Rep 2024;14:25704. [CrossRef]

[15] Muruganantham P, Balakrishnan SM. Attention aware deep learning model for wireless capsule endoscopy lesion classification and localization. J Med Biol Eng 2022;42:157-168. [CrossRef]

[16] Wang KN, Zhuang S, Ran QY, Zhou P, Hua J, Zhou GQ, He X. DLGNet: A dual-branch lesion-aware network with the supervised Gaussian mixture model for colon lesions classification in colonoscopy images. Med Image Anal 2023;87:102832. [CrossRef]

[17] Ali S, Ismail JD, Rahim HA. Transfer learning for gastrointestinal disease classification with CNNs. J Med Imaging Health Inform 2021;11:410-417.

[18] Jha D, Ali S, Tomar NK, Johansen D, Johansen D, Riegler P, et al. Comprehensive study on colonoscopic polyp segmentation with CNNs. Med Image Anal 2022;76:102321.

[19] Thambawita V, Salehi P, Halvorsen P, Riegler M. Domain-adaptive pretraining for gastrointestinal endoscopy. IEEE J Biomed Health Inform 2023;27:1245-1257.

[20] Zhao Y, Liu H, Wang K. Late fusion-based ensemble model for GI disease diagnosis. Comput Biol Med 2022;141:105135.

[21] Chen L, Sun M, Zhang Y. Intermediate-layer fusion with deep supervision for medical image analysis. Neural Comput Appl 2021;33:10529-10540.

[22] Roy S, Banerjee A, Mitra P. SE block-enhanced CNNs for cardiac disease detection. Biomed Signal Process Control 2021;69:102814. [CrossRef]

[23] Tang Y, Li F, Guo L. CBAM-driven lesion attention for medical diagnosis. IEEE Access 2022;10:33291-33302.

[24] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. Int J Comput Vis 2020;128:3365-3387. [CrossRef]

[25] Dhivya P, Vanithamani S, Malini M, Kanimozhi T. Adaptive edge-guided segmentation with BifTransNet for gastrointestinal tract image segmentation. Indian J Sci Technol 2024;17:4524-4535. [CrossRef]

[26] Obayya M, Al-Wesabi FN, Maashi M, Mohamed A, Hamza MA, Drar S, Alsaid MI. Modified salp swarm algorithm with deep learning based gastrointestinal tract disease classification on endoscopic images. IEEE Access 2023;11:25959-25967. [CrossRef]

[27] Zhu S, Li Y, Dai X, Mao T, Wei L, Yan Y. A multi-resolution hybrid CNN-transformer network with scale-guided attention for medical image segmentation. IEEE J Biomed Health Inform 2025;29:1012-1024. [CrossRef]

[28] Zhu M, Chen Z, Yuan Y. DSI-Net: Deep synergistic interaction network for joint classification and segmentation with endoscope images. IEEE Trans Med Imaging 2021;40:3315-3325. [CrossRef]

[29] Ali A, Iqbal A, Khan S, Ahmad N, Shah S. A two-phase transfer learning framework for gastrointestinal diseases classification. PeerJ Comput Sci 2024;10:e2587. [CrossRef]

[30] Garbaz A, Lafraxo S, Charfi S, El Ansari M, Koutti L, Salihoun M. Combined deep convolutional neural networks for abnormality classification in wireless capsule endoscopy images. Multimed Tools Appl 2025;84:11201-11229. [CrossRef]

[31] Gunasekaran H, Ramalakshmi K, Swaminathan DK, Mazzara M. GIT-Net: An ensemble deep learning-based GI tract classification of endoscopic images. Bioengineering 2023;10:809. [CrossRef]

[32] Garbaz A, Oukdach Y, Charfi S, El Ansari M, Koutti L, Salihoun M, Lafraxo S. MLFE-UNet: Multi-level feature extraction transformer-based UNet for gastrointestinal disease segmentation. Int J Imaging Syst Technol 2025;35:e70030. [CrossRef]

[33] Zhong Y, Piao Y, Zhang G. Dilated and soft attention-guided convolutional neural network for breast cancer histology images classification. Microsc Res Tech 2022;85:1248-1257. [CrossRef]

[34] Pan S, Liu X, Xie N, Chong Y. EG-TransUNet: A transformer-based U-Net with enhanced and guided models for biomedical image segmentation. BMC Bioinformatics 2023;24:85. [CrossRef]

[35] Xue H, Yonggang L, Min L, Lin L. A lighter hybrid feature fusion framework for polyp segmentation. Sci Rep 2024;14:23179. [CrossRef]

[36] Vrindavanam J, Kumar P, Kamath G, Bharadwaj AS. Transfer learning in endoscopic imaging: A

machine vision approach to GIT disease identification. In: 2024 1st International Conference on Communications and Computer Science (InCCCS). IEEE; 2024. p. 1-6. [CrossRef]

[37] Yang L, Zhai C, Liu Y, Yu H. CFHA-Net: A polyp segmentation method with cross-scale fusion strategy and hybrid attention. Comput Biol Med 2023;164:107301. [CrossRef]

[38] Huang S, Wang L, Liao J, Liu L. Multi-attentional causal intervention networks for medical image diagnosis. Knowl Based Syst 2024;299:111993. [CrossRef]

[39] Lu S, Liu J, Wang X, Zhou Y. Collaborative multi-metadata fusion to improve the classification of lumbar disc herniation. IEEE Trans Med Imaging 2023;42:3590-3601. [CrossRef]

[40] Yue G, Wei P, Liu Y, Luo Y, Du J, Wang T. Automated endoscopic image classification via deep neural network with class imbalance loss. IEEE Trans Instrum Meas 2023;72:1-11. [CrossRef]

[41] Yue G, Li S, Cong R, Zhou T, Lei B, Wang T. Attention-guided pyramid context network for polyp segmentation in colonoscopy images. IEEE Trans Instrum Meas 2023;72:1-13. [CrossRef]

[42] Liu XX, Xu M, Wei Y, Qin H, Song Q, Fong S, Siu S. Enhancing diagnostic precision in gastric bleeding through automated lesion segmentation: A deep DuS-KFCM approach. arXiv Preprint 2024;arXiv:2411.14385.

[43] Zhang K, Hu D, Li X, Wang X, Hu X, Wang C, Rao N. BFE-Net: Bilateral fusion enhanced network for gastrointestinal polyp segmentation. Biomed Opt Express 2024;15:2977-2999. [CrossRef]

[44] Rajasekar D, Theja G, Prusty MR, Chinara S. Efficient colorectal polyp segmentation using wavelet transformation and AdaptUNet: A hybrid U-Net. Heliyon 2024;10:e33412. [CrossRef]

[45] Li X, Li M, Yan P, Li G, Jiang Y, Luo H, Yin S. Deep learning attention mechanism in medical image analysis: Basics and beyonds. Int J Netw Dyn Intell 2023;2:93-116. [CrossRef]

[46] Mahmood S, Fareed MMS, Ahmed G, Dawood F, Zikria S, Mostafa A, et al. A robust deep model for classification of peptic ulcer and other digestive tract disorders using endoscopic images. Biomedicines 2022;10:2195. [CrossRef]

[47] Padmavathi P, Harikiran J. SE-ResNet: A novel method for gastrointestinal diseases classification from wireless capsule endoscopy images. Traitement Signal 2023;40:1341-1353. [CrossRef]

[48] El-Ghany SA, Mahmood MA, Abd El-Aziz AA. An accurate deep learning-based computer-aided diagnosis system for gastrointestinal disease detection using wireless capsule endoscopy image analysis. Appl Sci 2024;14:10243. [CrossRef]

[49] Kuduz H, Kaçar F. Human lower-extremity movement classification based on biomechanical sensor data: Machine learning approach. Sigma J Eng Nat Sci 2024;42:1378-1390. [CrossRef]

[50] Demirbaş AA, Üzen H, Fırat H. Spatial-attention ConvMixer architecture for classification and detection of gastrointestinal diseases using the Kvasir dataset. Health Inf Sci Syst 2024;12:32. [CrossRef]

[51] Khan MA, Shafiq U, Hamza A, et al. A novel network-level fused deep learning architecture with shallow neural network classifier for gastrointestinal cancer classification from wireless capsule endoscopy images. BMC Med Inform Decis Mak 2025;25:150. [CrossRef]

[52] Malathy N, Navin G, Sriram S, Bala Subramaniyan S. A secure framework for multimedia transmission in medical images using DNA cryptography. Sigma J Eng Nat Sci 2025;43:222-233. [CrossRef]

[53] Bulut C, Sönal T, Kolca D, Tarlak F. Evaluation of health care systems using supervised and unsupervised machine learning approaches. Sigma J Eng Nat Sci 2025;43:2174-2185. [CrossRef]

[54] Ahmed IA, Senan EM, Shatnawi HSA. Hybrid models for endoscopy image analysis for early detection of gastrointestinal diseases based on fused features. Diagnostics 2023;13:1758. [CrossRef]

[55] Mathew J, Dileepkumar SR, Ardra BS. A hybrid bit-efficientNet framework for automated detection of skin cancer and monkeypox using dermatoscopic images. Sigma J Eng Nat Sci 2025;43:102. [CrossRef]