Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

An Improved ResNet-Based Deep Learning Model with Modified Mountain Gazelle Optimization for Gastrointestinal Disease Classification

Harshitha E.S.[1], Narmadha G.[2*], Sakthivel B.[3], Muneeswari B.[4]

[1]Research Scholar, [2]Professor, Department of Electrical and Electronics Engineering, Sethu Institute of Technology, Pulloor, Kariapatti, India.
[3]Professor, Department of Electronics and Communication Engineering, Pandian Saraswathi Yadav Engineering College, Sivagangai, India.
[4]Professor, Department of Electronics and Communication Engineering, Velammal College of Engineering and Technology, Madurai, India.

E-mail: [1]harshitha19phd@gmail.com, [2*]gnarmadha@sethu.ac.in, [3]sakthivelece@psyec.edu.in, [4]bes@vcet.ac.in

Orcid ID: [1]0009-0003-3242-8853, [2*]0000-0001-6515-202X, [3]0000-0003-0032-6775, [4]0009-0004-2147-267X

Abstract

Gastrointestinal (GI) tract diseases occur due to abnormalities affecting different regions of the digestive system. Diagnosing GI tract diseases typically involves methods such as endoscopy, imaging examinations, biopsy analysis, and clinical assessments. The severity and stage of the disease play a crucial role in determining the appropriate treatment plan. Medical imaging techniques are used to identify abnormal regions and assess the progression of diseases. The categorization of GI tract diseases, including ulcerative lesions, polyps, bleeding regions, inflammatory conditions, and normal tissues, is crucial for accurate diagnosis and effective treatment planning. Recently, Artificial Intelligence (AI)-based Deep Learning (DL) models have received greater attention due to their accuracy and flexibility. In this work, a threefold DL model is suggested to classify GI tract disease types. An improved ResNet, called Channel-Spatial Context Gating ResNet (CSCG-ResNet), is proposed for efficient feature extraction. Then, a modified Mountain Gazelle Optimizer (MGO) is proposed for feature optimization. Finally, the Quantum-inspired TabNet is proposed for multi-class categorization. The performance of the model is validated using the Kvasir dataset. The model achieves an overall accuracy of 96.75% when compared to previously proposed models.

Keywords: Gastrointestinal Tract Diseases, Deep Learning, Improved ResNet, Channel-Spatial Context Gating, Mountain Gazelle Optimizer.

1. Introduction

Gastrointestinal (GI) tract diseases refer to disorders that affect different parts of the digestive system, including the esophagus, stomach, small intestine, and colon [1]. These diseases are among the most prevalent health problems globally. The likelihood of developing GI tract diseases increases due to factors such as age, lifestyle, dietary habits, and genetic

*** Corresponding Author**

DOI: https://doi.org/10.36548/jiip.2026.3.024
Received: 27.05.2026, received in revised form: 03.07.2026, accepted: 21.07.2026, published: 31.07.2026

conditions, which can lead to significant health complications if not diagnosed at an early stage. Based on various surveys, GI tract diseases contribute to a major burden of healthcare challenges worldwide [2].

Doctors follow different methods to diagnose GI tract diseases, Includes physical examination, endoscopy, imaging, biopsy, and laboratory-based testing. Laboratory testing identifies specific biological changes and disease-related abnormalities for further treatment planning. Imaging methods include endoscopic imaging, ultrasound, and computed tomography scans for better visualization of affected regions. Early diagnosis of GI tract diseases supports appropriate treatment and improves the quality of life. Compared to other approaches, imaging-based detection achieves better accuracy in abnormality diagnosis. Endoscopic imaging techniques examine the GI tract and provide detailed visualization of internal structures [3]. These high-resolution imaging technologies are used to detect potential lesions, polyps, bleeding regions, and inflammatory changes present in the digestive tract.

In GI tract disease imaging, AI techniques have enhanced the precision of segmentation and classification tasks. Machine Learning (ML) [4] and Deep Learning (DL) [5] models are applied to endoscopic images for disease identification. These models automate the detection and classification of GI tract abnormalities, reducing the need for manual interpretation and increasing the speed of diagnosis. However, despite advancements in AI-driven GI disease detection, existing methods have failed to address challenges in optimizing feature extraction and model interpretability. Many existing models do not fully capture the spatial and contextual information present in endoscopic images, which limits their overall diagnostic performance.

Attention-based models like SE-ResNet (Squeeze-and-Excitation ResNet) and CBAM-ResNet (Convolutional Block Attention Module ResNet) improve feature weighting by highlighting important channels and spatial regions. However, they process attention information separately and may not fully exploit the interaction between spatial location and contextual features. Transformer-based models, such as Vision Transformer (ViT) and Swin Transformer, achieve global feature learning through self-attention strategies. These models require large-scale training data and higher computational resources, which restricts their practical application in medical image classification.

To address these gaps, a threefold DL model for GI tract disease classification is proposed using endoscopic images. The model integrates an improved ResNet architecture for efficient feature extraction. A Modified Mountain Gazelle Optimizer (MGO) is suggested for feature optimization, and the TabNet model is used for accurate multi-class classification of different GI tract disease categories. The major contributions of this work are summarized as follows:

- A CSCG-ResNet model is developed to enhance spatial and contextual feature extraction from endoscopic images.
- A Modified Mountain Gazelle Optimizer (M-MGO) is proposed with improved exploration–exploitation capability for selecting optimal discriminative features.
- A Quantum-Inspired Adaptive TabNet (QI-TabNet) classifier is suggested to improve feature selection and multi-class GI disease classification performance.

The novelty of this work lies not in the simple combination of existing models, but in the development of task-specific modifications introduced at each stage of the proposed system.

The ResNet architecture is modified by incorporating the Channel-Spatial Context Gating (CSCG) module joint channel and spatial contextual feature refinement, improving the representation of complex GI abnormalities. The conventional Mountain Gazelle Optimizer is modified into M-MGO by introducing adaptive learning rate control, momentum-assisted search, exploration adjustment, and random-walk mechanisms to improve feature optimization capability. Furthermore, the standard TabNet model is enhanced as Quantum-Inspired Adaptive TabNet (QI-TabNet) by incorporating quantum-inspired probability-based feature representation for dynamic feature selection. The novelty of this work lies in the development of task-specific architectural modifications, particularly the CSCG module that jointly models channel importance, spatial localization, and contextual information for GI endoscopic image classification.

2. Related Works

Several studies have developed DL-based models for GI tract disease classification. It involves CNNs, hybrid networks, and attention mechanisms. Öztürk, Ş. et al. [6] proposed a GI tract classification model based on the integration of a CNN with a Long Short-Term Memory (LSTM) network. The proposed model achieves 97.90% accuracy compared with existing state-of-the-art approaches. In their study, Muhammad Ramzan et al. [7] proposed a computer-aided diagnosis system using texture and deep learning feature fusion for GI disease classification. The optimized features were selected through Principal Component Analysis (PCA) and entropy.

Zubair Rahman et al. [8] suggested an EfficientNetB5-based DL model with advanced augmentation and regularization techniques for GI disease classification. The proposed approach improved generalization ability and achieved a test accuracy of 98.89%. Likewise, Haile, M. B. et al. [9] developed a concatenated neural network model for GI disease diagnosis. It combines VGGNet and InceptionNet features for feature learning. The extracted features were classified using the SVM model.

Rubab, S. et al. [10] suggested a CAD model based on modified DL models. Optimization-based feature selection is applied for GI disease classification. In their work, Nouman Noor, M., et al. [11] proposed an automated GI disease localization and classification system using encoder–decoder segmentation. XceptionNet is used for classification with explainable AI techniques.

A deep CNN-based model combined with Empirical Wavelet Transform is proposed by Subhashree Mohapatra et al. [12] for GI disease classification. Similarly, Md. Faysal Ahamed et al. [13] proposed a lightweight Pearson Correlation Coefficient-based feature selection and Ensemble Extreme Learning Machine (PD-CNN-PCC-EELM) model for multiclass GI disease categorisation. The model achieves higher classification accuracy with reduced computational complexity. Şener, A. et al. [14] developed GISegNet for GI abnormality segmentation. The features extracted from GISegNet are classified using a transformer-based hybrid classification model. Likewise, Sivari, E. et al [15] proposed a bi-level stacking ensemble approach for GI disease classification. It combines CNN predictions with ML classifications.

Tanwar, V. et al. [16] suggested an Efficient Vision Transformer model architecture for GI disease classification. It involves EfficientNetB0 and Vision Transformer architectures for accurate classification. Sultan Daud Khan et al. [17] proposed an endoscopy image classification using Local-Global CNN and Endoscopy-Lesion Attention Module. Arefin

Ittesafun Abian et al. [18] developed an Atrous Spatial Pyramid Pooling (ASPP)-Swin Transformer-based architecture for GI video categorisation. In addition, a new imbalance handling approach is proposed with Grad-CAM visualization. Hussein, A. F. et al. [19] recommended a hybrid DL model for GI disease diagnosis. It integrates ResNet18 with an ML model of SVM for classification.

Şaban Öztürk et al. [20] suggested a residual LSTM-based CNN for GI image classification. This method improves feature learning by combining CNN-extracted features with LSTM classification. Gunasekaran, H. et al. [21] developed a model called GIT-NET for GI disease classification. This weighted ensemble approach combines DenseNet201, InceptionV3, and ResNet50 for classification. Majid, A. et al. [22] recommend a lightweight attention-based polyp segmentation model based on the EfficientNet-b0 encoder. It uses channel-spatial attention and dynamic gating modules to handle complex endoscopic image patterns. Similarly, Samra Siddiqui et al. [23] developed a robust DL model called SNet for complex GI disease classification. The proposed model uses CNN-based feature extraction and Minimum Redundancy Maximum Relevance (mRMR) based feature selection to reduce dimensionality. However, the existing approaches still face limitations in handling multi-scale contextual information and achieving robust generalisation. This highlights the need for an efficient and adaptive GI disease classification model. The overall summary of limitations of existing models is given in Table 1.

Table 1. Comparative analysis of existing methods and limitations

Method Category	Advantages	Limitations
CNN-based models	Effective local feature extraction	Limited contextual dependency learning
SE/CBAM attention models	Improve important feature selection	Separate channel/spatial refinement
Transformer models	Capture global relationships	High computational cost and data dependency
Optimization-based methods	Reduce irrelevant features	May not consider feature compactness and stability
Proposed CSCG-ResNet + M-MGO	Contextual attention with optimized feature selection	Addresses above limitations

3. CSCG-ResNet Model

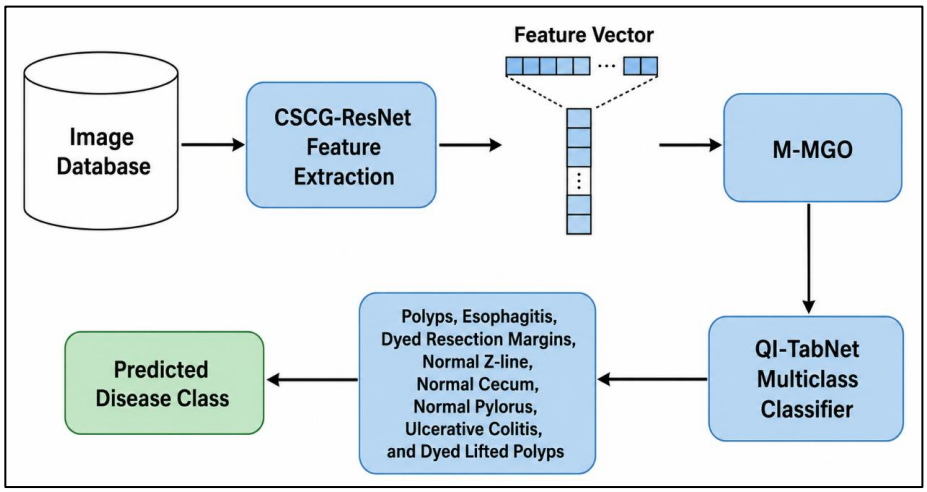

Figure 1. Workflow of the proposed classification

In this work, a three-stage system is recommended for the categorisation of GI tract disease types. Initially, features from the endoscopic images are extracted using Channel-Spatial Context Gating ResNet (CSCG-ResNet). The CSCG-ResNet applies dynamic recalibration of features at both the spatial and channel levels with its unique gating block. To further increase performance, the Modified Mountain Gazelle Optimizer (MGO) is applied for feature optimization. For classification, the quantum-inspired TabNet is applied to classify different GI tract disease categories. The complete workflow of the three-stage model is visualised in Figure 1.

3.1 CSCG-ResNet

Residual Networks (ResNet) were established by He et al. in 2015. ResNet uses residual connections for stable gradient flow during backpropagation. The main concept of ResNet is to learn a residual mapping instead of directly learning the original mapping. The residual function is denoted as:

$$Y = \sigma(\mathcal{F}(X, W) + X) \tag{1}$$

where X characterises the input feature map, $\mathcal{F}(X, W)$ denotes the residual transformation function with learnable parameters W, and σ denotes the activation function. However, ResNet does not account for dynamic recalibration of important features across spatial and channel dimensions. This is a critical requirement for medical image classification, where precise feature localization and contextual feature understanding are needed. To overcome this limitation, a novel architecture called Channel-Spatial Context Gating ResNet (CSCG-ResNet) is proposed. The proposed architecture uses dynamic recalibration through a CSCG block. It incorporates a gating mechanism which works at both the channel and spatial levels. This modification is used to improve feature extraction and contextual understanding. The overall output of CSCG-ResNet can be represented as:

$$F_{out} = \sigma(F_{CSCG} + F_{res}) \tag{2}$$

where F_{CSCG} represents the output of the Channel-Spatial Context Gating block and F_{res}represents the residual feature information. The architecture of CSCG-ResNet is shown in Figure 2. The key components of the CSCG-ResNet architecture include:

- **CSCG Block:** A novel block that replaces the standard ResNet block. It incorporates dual gating mechanisms for channel-wise and spatial recalibration of features.
- **Multi-Scale Context Extraction:** Uses dilated convolutions and global context pooling to capture fine-grained and global contextual information.
- **Residual Connections:** Ensure stable training by adding the input to the output of each block, just like in standard ResNet.

3.1.1 Initial Feature Extraction

Initially, the convolution layer is used to extract low-level features such as edges, corners, and textures from the input image. It applies a convolution operation to transform the input image into feature representations. The convolution operation is mathematically expressed as follows:

$$F_1(i, j, k) = \sum_m \sum_n \sum_c X\,(i, +, mj + nc)W(m, n, c, k) + b_k \quad (3)$$

where X represents the input image, Wrepresents the convolution kernel weights, $b_k$represents the bias term, and F_1represents the extracted feature map. After convolution, Batch Normalization (BN) is applied to stabilize the training process with the ReLU activation function:

$$F_{conv1} = ReLU\big(BN(F_1)\big) \quad (4)$$

where $BN(.)$normalizes the feature distribution and ReLU introduces non-linearity. The output of this layer represents the initial low-level feature representation.

3.1.2 Standard Convolution Path

The standard convolution path consists of two consecutive convolutional operations. Each operation extracts deeper feature representations from the previous feature map. The first convolution transformation is expressed as:

$$F_{std1} = ReLU\big(BN(W_2 * F_{conv1} + b_2)\big) \quad (5)$$

The second convolution operation generates the transformed feature representation:

$$F_{std2} = BN(W_3 * F_{std1} + b_3) \quad (6)$$

where W_2 and W_3 represent the learnable convolution weights. These operations extract high-level feature information from the input image. The residual connection combines the original input information with the transformed features:

$$F_{res} = ReLU\big(F_{std2} + P(X)\big) \quad (7)$$

where $P(X)$ represents the projection function used to match the feature dimensions. The residual connection enables the network to retain important features from earlier layers and improve gradient propagation.

3.1.3 Dilated Convolution and Global Context

To capture a wider context, the model uses dilated convolutions. These convolutions expand the receptive field without increasing the number of parameters. The dilated convolution operation is applied to the output of the previous standard convolution layer to extract features that span a larger region of the input image. The dilated convolution operation is represented as:

$$F_{dilated}(i, j) = \sum_m \sum_n W\,(m, n)F_{std}(i, +r.mj + r.n) \quad (8)$$

where r represents the dilation rate. The dilated convolution enables the network to capture multi-scale spatial information from GI tract images. In addition, the global context is obtained using Global Average Pooling (GAP). The global feature representation is calculated as:

$$F_{global}(c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_{std}\,(i, j, c) \quad (9)$$

where H and Wrepresent the spatial dimensions of the feature map and $c$represents the channel index. The GAP operation summarizes the global information of each feature channel.

3.1.4 Dual Gating Mechanism

The dual-gating mechanism dynamically recalibrates the features enabling the network to focus on the most important channels and spatial regions. It consists of two components: channel gating and spatial gating.

3.1.5 Channel Gating

Channel gating learns to emphasize important feature channels by analyzing the global context information. The channel attention weights are generated using the following methods:

$$A_c = \sigma\left(W_{c2}\delta\left(W_{c1}F_{global}\right)\right) \tag{10}$$

where $W_{c1}$and W_{c2}represent fully connected transformation matrices, δrepresents the ReLU activation function, and σrepresents the sigmoid activation function. The obtained channel attention weights are applied to the feature map:

$$F_c = A_c \otimes F_{dilated} \tag{11}$$

where $\otimes$ signifies element-wise multiplication. This operation increases the importance of significant feature channels and suppresses irrelevant information.

3.1.6 Spatial Gating

Spatial gating recalibrates the spatial regions by applying learned weights to the dilated feature maps. The spatial attention weights are obtained using the following method:

$$A_s = \sigma(W_s * F_{dilated} + b_s) \tag{12}$$

where W_s represents the spatial gating filter weights. The spatially refined feature representation is calculated as:

$$F_s = A_s \otimes F_{dilated} \tag{13}$$

The spatial gating mechanism enables the network to focus on important disease-related regions in the input image.

3.1.7 Feature Fusion

After applying gating mechanisms, the channel- and spatial-gated features are combined with the dilated features to preserve both fine-grained details and global context. The gated feature representation is calculated as:

$$F_{gated} = F_{dilated} \otimes A_c \otimes A_s \tag{14}$$

The final feature fusion operation is expressed as:

$$F_{final} = W_f * \left[F_{gated} \oplus F_{dilated}\right] + b_f \tag{15}$$

where $\otimes$ denotes concatenation and W_f represents the fusion weights. The fusion operation combines spatial details and contextual information, which improves the feature representation for classification.

3.1.8 Residual Connection and Final Output

As with traditional ResNet, the residual connection is added to the final output to stabilise the training. It enables the network to learn from both the original input and the transformed features. The final output is obtained as:

$$F_{output} = ReLU\left(F_{final} + P(X)\right) \tag{16}$$

where $P(X)$ is a projection function that adjusts the dimensions of the input feature map. The final extracted feature representation is expressed as follows:

$$F_{feature} = F_{output} \tag{17}$$

The obtained feature representation consists of optimized channel information and spatial dependencies. It also contains contextual features for better classification.

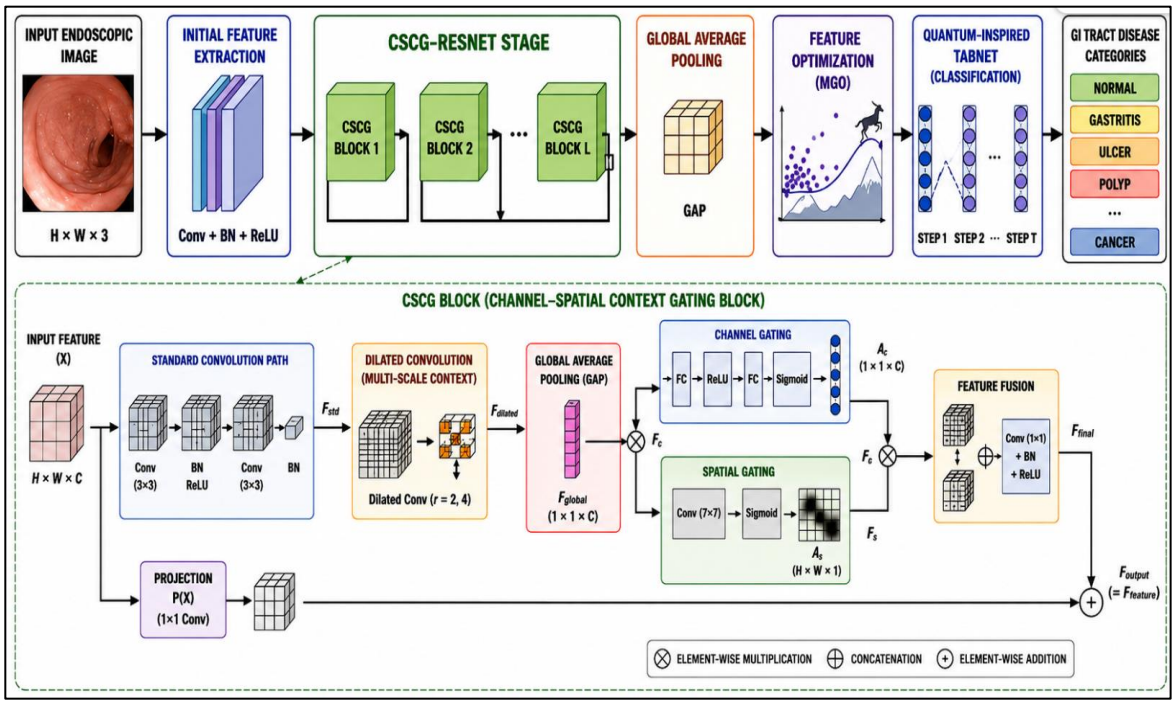

Figure 2. Architecture of CSCG-ResNet

3.2 Mountain Gazelle Optimizer (MGO)

The MGO is a nature-inspired metaheuristic algorithm. The MGO algorithm mimics the interactions and roles of different gazelles in a herd. It uses their natural behaviour to find optimal solutions in a complex search space. The optimization process is influenced by the roles of four types of gazelles: Territorial Solitary Males (TSM), Maternity Herds (MH), Bachelor Male Herds (BMH), and Migration to Search for Food (MSF). Each group performs specific tasks to balance the search between exploration and exploitation. However, MGO suffers from various drawbacks, including limited exploration in complex landscapes,

premature convergence, and parameter sensitivity. To mitigate these issues, in this work, the Modified MGO is suggested for feature optimization. Modified MGO uses dynamic learning rate adjustments and momentum-based refinement for problem solving.

3.2.1 Territorial Solitary Males (TSM) with Adaptive Learning Rate

In the proposed M-MGO, this phase is enhanced by incorporating an adaptive learning rate. This is used for the optimizer to refine the best solutions effectively without overshooting or getting stuck in a local optimum. A dynamic learning rate helps improve convergence by making the search more stable as the optimizer zeroes in on the optimal solution. It can be modelled as follows:

$$TSM = \text{malegazelle} - | (r_{i1} * BH - r_{i2} * X(t)) * F | * Cofr + \text{LearningRateAdjustment}(Iter) \quad (18)$$

Where, $\eta_{\max}$ is the maximum learning rate, $\alpha$is the rate of decay, $Iter$ represents the current iteration, LearningRateAdjustment(Iter) is calculated as:

$$\text{LearningRateAdjustment}(Iter) = \frac{\eta_{\max}}{1+\alpha \cdot Iter} \quad (19)$$

The adaptive learning rate provides a larger search step to improve global exploration and gradually reduces the step size during later iterations to achieve precise local exploitation. This mechanism avoids unstable updates and improves convergence reliability.

3.2.2 Maternity Herds (MH) with Momentum

The MH phase focuses on balancing exploration and exploitation. In the M-MGO, momentum is added to this phase to avoid getting stuck in local minima. The momentum term is used for the search process to "carry forward" solutions. This momentum provides additional push to the search process, which allows it to explore better solutions and avoid being trapped in less optimal areas. It can be modelled as follows:

$$MH = (BH + Cof1) + r_{i3} * \text{malegazelle} - r_{i4} * X_{\text{rand}} * Cof1 + \text{MomentumFactor}(Iter) \quad (20)$$

The momentum factor is used to preserve the previous movement direction of each search agent. It is calculated as:

$$MomentumFactor(Iter) = \beta M_{t-1} + (1-\beta)(X_t - X_{t-1}) \quad (21)$$

where M_t represents the current momentum value, M_{t-1} represents the previous momentum value, and β is the momentum coefficient. A higher momentum contribution maintains the previous search direction. A lower value is used for more adaptive position updates. This improves convergence stability and reduces sudden oscillations during optimization.

3.2.3 Bachelor Male Herds (BMH) with Exploration Adjustment

The BMH phase in MGO is responsible for global exploration. In the M-MGO, the BMH phase is modified with an ExplorationAdjustment(Iter) function. This function increases the exploration intensity as the algorithm nears convergence, which prevents the optimizer from prematurely focusing too much on exploitation. It can be modelled as follows:

$$BMH = (X(t) - D) + r_{i5} * \text{malegazelle} - r_{i6} * BH * Cofr + \text{ExplorationAdjustment}(Iter) \quad (22)$$

Where, ExplorationAdjustment(Iter) is a function that gradually increases the exploration component. The exploration adjustment factor dynamically controls the search intensity by maintaining higher exploration in early iterations and increasing exploitation capability in later iterations. It is expressed as:

$$ExplorationAdjustment(Iter) = E_{max} - (E_{max} - E_{min}) \frac{Iter}{T} \quad (23)$$

where E_{max} and E_{min} represent the maximum and minimum exploration limits, respectively, and Trepresents the maximum number of iterations.

3.2.4 Migration to Search for Food (MSF) with Random Walk Factor

In the MSF phase, the original MGO uses randomness to escape local optima. The M-MGO further enhances this phase by adding a RandomWalkFactor(Iter). It can be modelled as follows:

$$MSF = (ub - lb) * r7 + lb + \text{RandomWalkFactor}(Iter) \quad (24)$$

Where, RandomWalkFactor(Iter) uses a degree of randomness to ensure broad exploration. The random-walk factor introduces controlled stochastic variation into the position update process to maintain population diversity. It is defined as:

$$RandomWalkFactor(Iter) = \gamma \times rand(0, 1) \times (X_{max} - X_{min}) \quad (25)$$

where γ represents the random-walk coefficient, X_{max} and X_{min} represent the search boundaries, and $X_t$represents the current position of the search agent.

3.3 Feature Optimization Using M-MGO in ResNet

In classification, feature optimization plays a fundamental role in increasing the accuracy of DL models. Here, the Modified Mountain Gazelle Optimizer (M-MGO) is applied to refine the features extracted features from CSCG-ResNet. In the proposed M-MGO-based feature optimization process, the continuous position values are generated by the gazelle agents. These are then converted into binary feature-selection vectors to identify the most informative features. Since the extracted CSCG-ResNet features are represented in continuous space, a sigmoid transfer function is applied to map the position values into a probability range between 0 and 1. The transfer function is expressed as:

$$S(X_i) = \frac{1}{1+e^{-X_i}} \quad (26)$$

where X_i represents the continuous position of the i^{th}feature and $S(X_i)$ denotes the probability of selecting that feature. The binary conversion rule is defined using a threshold value as follows:

$$B_i = \begin{cases} 1, & S(X_i) \geq 0.5 \\ 0, & S(X_i) < 0.5 \end{cases} \quad (27)$$

where $B_i = 1$ indicates the corresponding feature is selected, $B_i = 0$ indicates the feature is discarded. The threshold value of 0.5 provides an equal-probability boundary for feature selection and avoids bias toward feature inclusion or exclusion. Then, the generated binary feature vector is evaluated using the fitness function.

The M-MGO and backpropagation perform different optimization tasks in the proposed work. Backpropagation is used to update the trainable weights of CSCG-ResNet and QI-TabNet to reduce classification loss. Here, the M-MGO acts as an external metaheuristic optimizer to select the optimal feature subset from the extracted features. Thus, M-MGO does not modify network weights but improves classification performance through feature optimization.

The fitness or loss function for feature optimization using M-MGO can be expressed as the combination of the cross-entropy loss and a feature optimization penalty. The cross-entropy loss minimizes the classification error.

$$L_{total} = L_{CE} + \lambda_{feature} \cdot F_{opt} \tag{28}$$

Where, $L_{CE}$is the cross-entropy loss. It determines how well the model predicts the true labels. The fitness function of M-MGO is formulated to directly evaluate the quality of the selected feature subset. The optimization objective considers classification error and feature sparsity. The classification loss ensures that the selected features maintain discriminative capability. The sparsity term reduces redundant features and generates a compact feature representation. The fitness equation is expressed as follows:

$$F_{fitness} = L_{CE} + \lambda \frac{N_{selected}}{N_{total}} \tag{29}$$

where L_{CE}represents the classification error, $N_{selected}$ represents the number of selected features, N_{total}represents the total extracted features, and λ controls the contribution of the sparsity constraint. The Pseudocode for feature selection is given below:

Algorithm 1: Deep feature selection using modified mountain gazelle optimizer (M-MGO)

Input:
Deep feature matrix F_{CNN}extracted from CSCG-ResNet
Population size N
Maximum iteration number Max_{iter}
Lower bound lb, Upper bound ub
Regularization coefficient $\lambda_{feature}$
Output:
Optimal feature subset F_{opt}
Begin
1. Initialize N gazelle agents Xi (i = 1,2,...,N)
   randomly within the feature search space.
2. Initialize M-MGO parameters:
   Learning rate ηmax
   Decay factor α
   Momentum coefficient β
   Exploration limits (Emax, Emin)
   Random walk coefficient γ

3. Convert continuous gazelle positions into binary feature vectors:
   Si = 1/(1+exp(-Xi))
   If Si ≥ 0.5
      Feature = 1 (Selected)
   Else
      Feature = 0 (Discarded)
   End If
4. Evaluate fitness of each gazelle:
   Fitness = LCE + λfeature × Fopt
   where LCE represents classification error and
   Fopt represents feature optimization penalty.
5. Select the best gazelle position as BH.
6. For Iter = 1 to Max_iter do
   6.1 Update adaptive learning rate:
      LR = ηmax/(1 + α × Iter)
   6.2 Update Territorial Solitary Male (TSM):
      Apply adaptive learning rate to move
      current gazelle position towards BH.
   6.3 Update Maternity Herd (MH):
      Calculate momentum factor:
      M(t)=βM(t-1)+(1-β)(X(t)-X(t-1))
      Update gazelle position using momentum
      enhanced exploitation.
   6.4 Update Bachelor Male Herd (BMH):
      Calculate exploration adjustment:
      E(t)=Emax-(Emax-Emin)(Iter/Max_iter)
      Increase exploration capability to avoid
      premature convergence.
   6.5 Update Migration Search for Food (MSF):
      Apply random walk factor:
      RW(t)=γ×rand(0,1)×(BH-X(t))
      Improve population diversity and escape
      local optimum.
   6.6 Convert updated continuous positions into
      binary feature selection vectors using
      sigmoid transfer function and threshold 0.5.
   6.7 Calculate new fitness value of each
      feature subset.
   6.8 If new fitness < previous fitness:
      Replace previous solution.
      End If
   6.9 Update best gazelle position BH.
   End For
7. Select features corresponding to the final best
   gazelle position.
8. Generate deep contextual feature representation from RGB endoscopic images:
   Fopt = FCSCG(BH)
9. Return Fopt.
End

Initially, the M-MGO population is randomly generated. The fitness of each solution is calculated using a combined objective function consisting of classification loss and a feature optimization penalty. The optimization process aims to minimize classification error and maintain a compact feature representation. During optimization, the TSM phase improves promising feature subsets using an adaptive learning rate. The MH phase uses momentum-based updating to maintain previous search information. Finally, the MSF phase uses random exploration to avoid local optima and improve population diversity. After completing all iterations, the best gazelle position represents the optimal feature selection vector. The corresponding features are selected from the CSCG-ResNet feature set and forwarded to the TabNet classifier.

3.4 Classification

In this work, a Quantum-Inspired Adaptive TabNet (QI-TabNet) is proposed for classification with quantum-inspired feature refinement. QI-TabNet is a modified version of the TabNet model that incorporates quantum-inspired probability representation into the feature selection process. The architecture of QI-TabNet is based on a combination of neural decision blocks and quantum-inspired attention mechanisms, where the model learns to identify the most relevant features at each decision point through probability amplitude-based feature selection. Unlike conventional TabNet, which uses sparse attention masks for feature selection, QI-TabNet utilizes quantum-inspired feature states to represent the importance of each feature dynamically. A sequence of quantum-guided decision steps is applied. It computes a set of decisions by selecting important features based on their probability contribution and uses quantum-inspired attention to concentrate on the most significant information required for accurate prediction.

Initially, the input data is forwarded through a feature encoder to convert the extracted features into a latent feature space. The quantum-inspired attention mechanism is the core component of QI-TabNet. The input features are converted into a quantum-inspired feature state using probability amplitudes. These amplitudes indicate the contribution of individual features and help the model select highly discriminative features during classification. The decision steps consist of learnable blocks that integrate quantum-inspired feature selection with non-linear transformations. These blocks regulate the information flow and determine the important features that are forwarded to the next decision step. Finally, the output from each quantum-guided decision step is passed through a multi-layer perceptron (MLP) layer to generate the final classification output. The quantum-inspired feature representation is expressed as:

$$| \psi\rangle = \sum_{i=1}^{n} a_i \, | x_i\rangle \tag{30}$$

where, $| \psi\rangle$ represents the quantum-inspired feature state, a_i represents the probability amplitude of the i^{th}feature, and $| x_i\rangle$represents the corresponding feature state. The probability amplitude is calculated as:

$$a_i = \frac{x_i^2}{\sum_{j=1}^{n} x_j^2} \tag{31}$$

where, x_i represents the input feature value and $a_i$represents the important contribution of each feature in the quantum feature space. The quantum-inspired attention mechanism generates the feature selection mask at each decision step as:

$$M_t^Q = \frac{|W_t \psi_t|^2}{\sum_{i=1}^{n} |W_t \psi_t|^2} \tag{32}$$

where, M_t^Qrepresents the quantum-inspired attention mask, $W_t$represents the learnable attention parameters, and $\psi_t$represents the quantum feature state at the t^{th}decision step. The selected feature representation is obtained as:

$$X_t = M_t^Q \odot X \tag{33}$$

where, $\odot X$ represents element-wise multiplication and $X_t$represents the quantum-guided selected feature representation. The decision-making process at each quantum-guided step is represented as:

$$z_t = MLP_t \left(M_t^Q (x_{t-1}) \right) \tag{34}$$

where, x_{t-1}represents the input features from the previous decision step, M_t^Qrepresents the quantum-inspired attention mechanism that selects the relevant features, MLP_trepresents the multi-layer perceptron applied to the selected features, and $z_t$represents the decision output at step t. The final classification output is produced by combining all quantum-guided decision step outputs as follows:

$$\hat{y} = \sum_{t=1}^{T} \gamma_t \, z_t \tag{35}$$

where, $\gamma_t$represents the quantum-inspired adaptive weights. It determines the contribution of each decision step to the final prediction, and Trepresents the total number of decision steps. The proposed QI-TabNet improves classification performance by incorporating quantum-inspired feature probability estimation and adaptive attention-based feature selection. This strategy allows the model to concentrate on the most discriminative features during the prediction process.

4. Results and Discussion

The performance of the CSCG-ResNet model is assessed using the combined Kvasir v1 and Kvasir v2 datasets [24]. The combined dataset consists of eight GI tract categories;Polyps, Esophagitis, Dyed resection margins, Normal z-line, Normal cecum, Normal pylorus, Ulcerative colitis, and Dyed lifted polyps. Sample images from the dataset are shown in Figure 3. The dataset contains variations in image quality and illumination to analyze the robustness of the model. Initially, the input endoscopic images are resized to 224 × 224 pixels. Then, a median filter with a 3 × 3 kernel size is applied to remove impulse noise and smooth intensity variations. The pixels are normalized by a scaling process, which converts the intensity values into the range of 0 and 1. The parameter settings of the model are given in Table 2. The dataset is divided into training and testing subsets using a 70:30 ratio. Duplicate image verification is performed before training to remove repeated samples and avoid biased evaluation. The training subset contains 6720 images and the testing subset contains 2880 images. Each GI category contains 840 training images and 360 testing images. In the proposed model, the modules are integrated through a sequential hybrid learning strategy rather than independent training. The CSCG-ResNet and QI-TabNet parameters are optimized using gradient-based backpropagation. The M-MGO performs metaheuristic optimization independently for feature

selection. The optimized feature subset generated by M-MGO is used as the input for QI-TabNet classification. Therefore, the complete model tracks a sequential optimization process.

Table 2. Parameter settings of proposed model

Component	Parameter	Value	Component	Parameter	Value
Image preprocessing	Input image size	224 × 224	QI-TabNet	Number of decision steps	5
	Image format	RGB		Feature transformer dimension	128
	Noise removal method	Median filter		Attention dimension	64
	Median filter kernel	3 × 3		Batch size	32
	Pixel normalization	0–1		Training epochs	100
CSCG-ResNet	Initial convolution kernel	7 × 7		Optimizer	Adam
	Standard convolution kernel	3 × 3		Learning rate	0.001
	Dilated convolution rate	2		Dropout rate	0.2
	Activation function	ReLU		Loss function	Categorical cross-entropy
	Extracted feature dimension	2048			
M-MGO	Population size	30	Output layer	Number of classes	8
	Maximum iterations	100		Activation	Softmax
	Maximum learning rate	0.1			
	Learning rate decay	0.01			
	Momentum factor	0.9			
	Random walk coefficient	0.5			

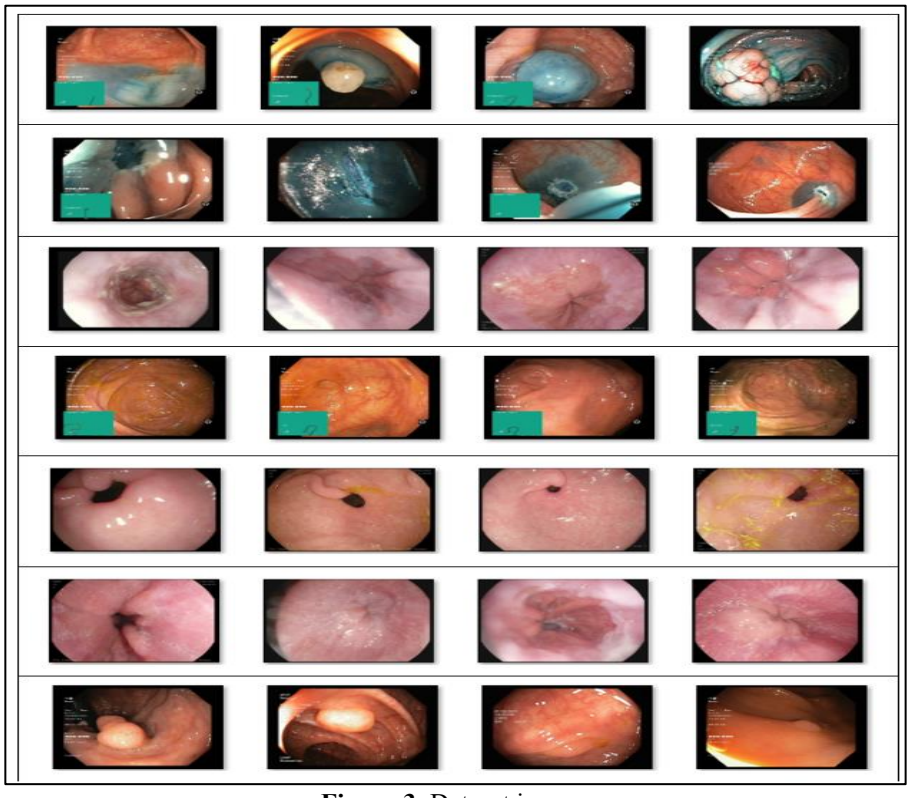

Figure 3. Dataset images

The overall CSCG-ResNet training and loss plot is visualised in Figure 4. Both the loss and accuracy curves show consistent improvement. The validation curves closely align with the training curves without significant divergence. The model is not overfitting and will perform reliably on unseen data.

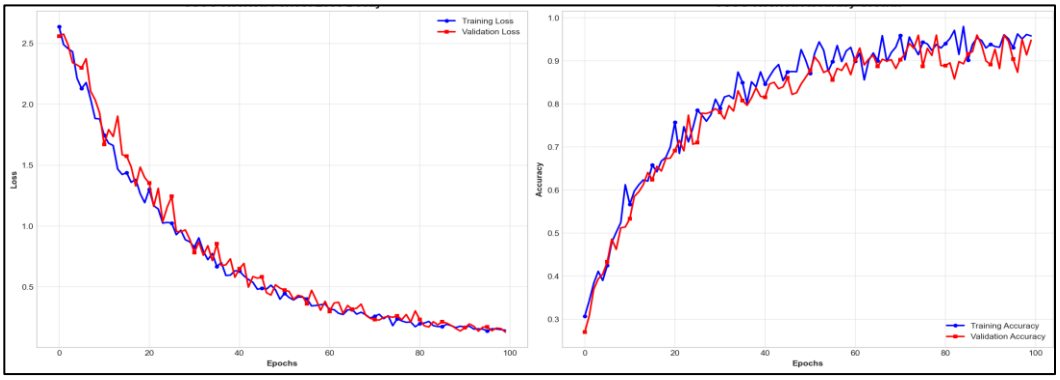

Figure 4. Model training and validation curves

Figure 5 shows the Receiver Operating Characteristic (ROC) curves for eight classes. Each subplot represents the ROC curve for a specific class. This curve illustrate the trade-off between the true positive rate (TPR) and the false positive rate (FPR). Overall, the AUC values for all eight classes indicate that the model is performing well with all values above 0.95. It is observed that the classifier has strong discriminatory power for differentiating between the different types of GI diseases.

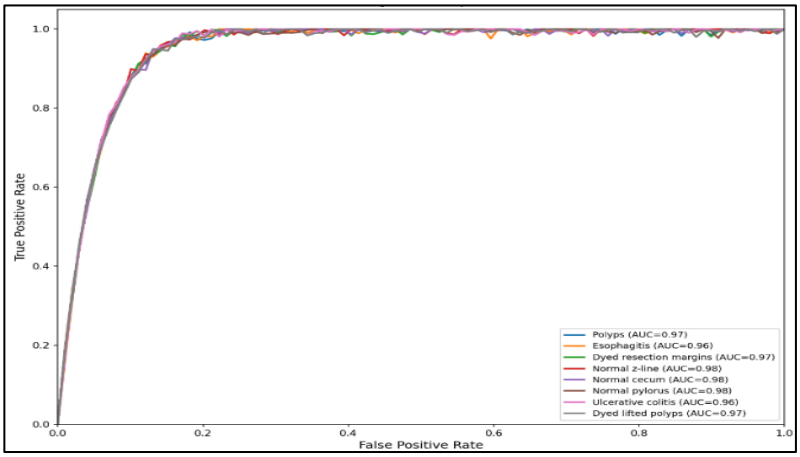

Figure 5. ROC plot of the model

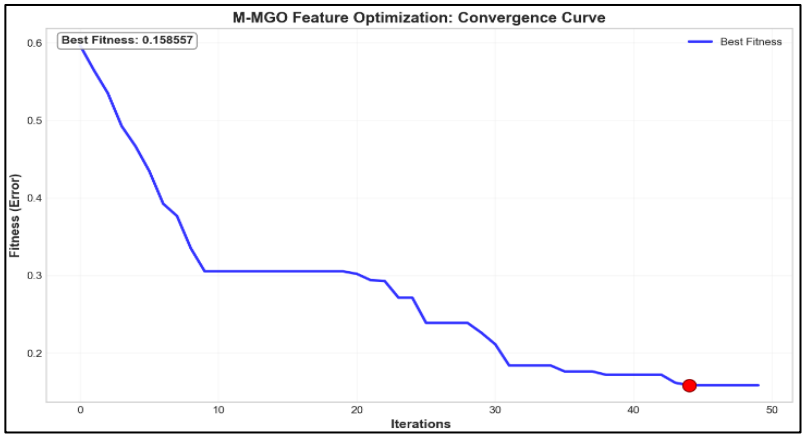

Figure 6. M-MGO optimization curve

The convergence curve for MGO-based feature optimization is shown in Figure 6. The plot tracks the progression of the fitness (error) value over 50 iterations. The best fitness value achieved by the optimizer is 0.158557. This indicates the lowest error the algorithm reached during the optimization process. This convergence curve suggests that MGO successfully minimized the error across iterations. The rapid reduction in fitness early on and the stabilization towards the end prove the algorithm's efficiency in converging to a solution.

To quantitatively evaluate the feature-selection capability of M-MGO, the number of retained features after optimization is analyzed (Table 3). Initially, the CSCG-ResNet extracts 2048-dimensional deep features. After feature selection, 512 features are retained. It achieves a 75% feature reduction rate. The reduction of redundant features improves computational efficiency and preserves the discriminative information required for GI disease classification. In addition, the stability of the selected feature subset is evaluated through five independent M-MGO optimization runs. The selected subsets achieved an average stability score of 94.6%. This denotes that M-MGO provides consistent and reliable feature selection.

Table 3. Feature selection performance of M-MGO

Method	Original Features	Selected Features	Reduction Rate (%)	Stability (%)
CSCG-ResNet + M-MGO	2048	512	75	94.6

The confusion matrix of the CSCG-ResNet is shown in Figure 7. The model correctly identified 347 instances as Polyps. For Esophagitis, 346 instances were correctly identified. Dyed resection margins were correctly predicted 352 times. The Normal z-line class was accurately identified 353 times. Normal cecum was correctly classified 354 times. Normal pylorus was correctly predicted 335 times. Ulcerative colitis was accurately identified 348 times. The dyed-lifted polyps were correctly classified 340 times. Overall, the confusion matrix suggests that the model works well, with a high number of accurate categorisations for each type. The performance of the CSCG-ResNet for all classes is given in Table 4.

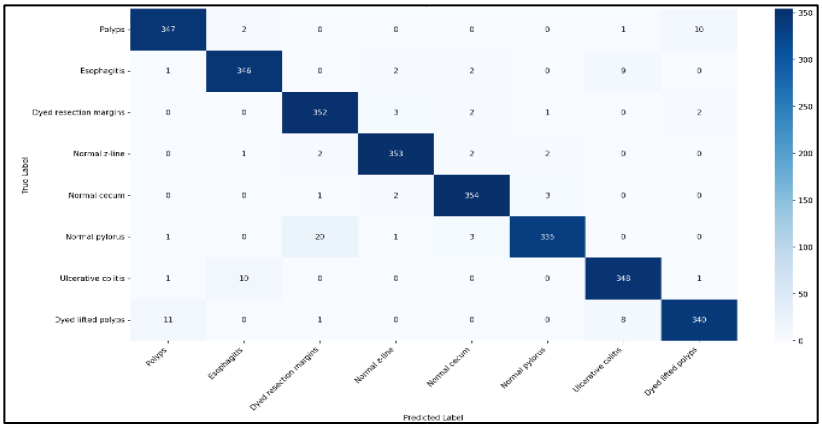

Figure 7. Confusion matrix

Table 4. Class-wise performance

Class	Precision (Pr)	Recall (Re)	F1-Score(F1)
Polyps	0.9612	0.9639	0.9626
Esophagitis	0.9638	0.9611	0.9624
Dyed resection margins	0.9362	0.9778	0.9565
Normal z-line	0.9778	0.9806	0.9792
Normal cecum	0.9752	0.9833	0.9793
Normal pylorus	0.9824	0.9306	0.9558
Ulcerative colitis	0.9508	0.9667	0.9587
Dyed lifted polyps	0.9632	0.9444	0.9537

The model performance comparison with previous approaches is given in Table 5. All models are trained using the same parameter settings with validation splits. The CNN + LSTM model shows an accuracy of 91.82%. The texture and deep feature fusion approach shows 92.46% accuracy. The EfficientNetB5 model shows improved classification performance with an accuracy of 93.72%. Recent attention-based models like SE-ResNet and CBAM-ResNet show accuracies of 94.62% and 95.08%, respectively. The hybrid VGGNet + InceptionNet + SVM model shows an accuracy of 90.94%. Similarly, the XceptionNet with Explainable AI shows 94.15% accuracy by enhancing feature representation. The Deep CNN combined with Empirical Wavelet Transform and the PD-CNN-PCC-EELM methods show accuracies of 91.63% and 93.26%, respectively. Furthermore, GISegNet integrated with Transformer shows 94.72% accuracy. The Bi-level Stacking Ensemble and Efficient Vision Transformer models show accuracies of 95.10% and 95.34%, respectively. The Weighted Ensemble CNN Model and Attention EfficientNet-B0 show accuracies of 95.02% and 95.18%. The ASPP-Swin Transformer model shows a higher accuracy of 95.62% by effectively capturing multi-scale and global contextual features. The ViT-3 [25] model attains an accuracy of 95.16%. It uses three transformer encoder blocks for feature learning. However, the CSCG-ResNet shows the highest performance with an accuracy of 96.75%, along with precision, recall, and F1 values of 96.68%, 96.76%, and 96.85%, respectively. This superior performance proves the efficiency of the proposed model in extracting discriminative features and increasing classification accuracy compared with existing state-of-the-art methods.

Table 5. Performance comparison with other models

Method	Ac (%)	Pr (%)	Re (%)	F1 (%)
CNN + LSTM	91.82	91.75	91.63	91.69
Texture and Deep Feature Fusion	92.46	92.31	92.18	92.24
EfficientNetB5	93.72	93.65	93.58	93.61
SE-ResNet	94.62	94.55	94.58	94.56
CBAM-ResNet	95.08	95.01	95.04	95.02
VGGNet + InceptionNet + SVM	90.94	91.02	90.86	90.92
XceptionNet with Explainable AI	94.15	94.10	94.05	94.07
Deep CNN + Empirical Wavelet Transform	91.63	91.57	91.48	91.52
PD-CNN-PCC-EELM	93.26	93.18	93.22	93.20
GISegNet + Transformer	94.72	94.65	94.68	94.66
Bi-level Stacking Ensemble	95.10	95.05	95.02	95.03
Efficient Vision Transformer	95.34	95.28	95.30	95.29
Local-Global CNN + Attention Module	94.86	94.80	94.83	94.81
CNN + mRMR Feature Selection	95.46	95.41	95.38	95.39
ResNet18 + SVM	92.74	92.69	92.72	92.70
Residual LSTM-CNN	94.38	94.31	94.35	94.33
Weighted Ensemble CNN Model	95.02	94.96	95.00	94.98
Vision Transformer (ViT-3)	95.16	95.10	95.13	95.11
Attention EfficientNet-B0	95.18	95.12	95.15	95.13
ASPP-Swin Transformer	95.62	95.55	95.59	95.57
CSCG-ResNet	96.75	96.68	96.76	96.85

The five-fold cross-validation results validate the stability and reliability of the CSCG-ResNet (Table 6). The model achieves consistent performance across all folds with accuracy values ranging from 96.42% to 97.02%. The low standard deviation of 0.24% indicates minimal variation between training and testing subsets. The average F1-score of 96.85% confirms that the proposed model maintains a balanced classification performance across different GI disease categories. These results prove that the CSCG-ResNet is robust and not dependent on a specific data split.

Table 6. Cross-validation performance

Fold	Acy (%)	Pr (%)	Re (%)	F1 (%)
Fold 1	96.42	96.35	96.41	96.38
Fold 2	96.81	96.72	96.85	96.78
Fold 3	96.58	96.61	96.55	96.57
Fold 4	97.02	96.94	97.05	96.99
Fold 5	96.92	96.78	96.94	96.86
Average	96.75	96.68	96.76	96.85
Standard Deviation	0.24	0.22	0.26	0.23

To verify the effectiveness of the proposed method, a statistical significance analysis was performed against the three best-performing existing models (Table 7). Each model was trained and evaluated five times using different random initializations, Identical experimental settings were used. In addition, an independent statistical significance test was conducted in comparison with the existing methods. The obtained p-values (<0.05) indicate that the improvement achieved by the proposed model is statistically significant and is not due to random variations. The proposed method achieves an improvement of 1.13%, 1.29%, and 1.41% over the three comparative methods, respectively. The obtained p-values are less than 0.05 in comparison. This proves that the performance improvement of the proposed method is statistically significant rather than occurring due to random variations in the dataset.

Table 7. Statistical significance analysis of the CSCG-ResNet

Model / Comparison	Mean Accuracy (%)	Standard Deviation (%)	95% Confidence Interval (%)	Mean Difference (%)	t-value	p-value	Statistical Significance
Efficient Vision Transformer	95.34	0.18	95.18 – 95.50	1.41	5.74	0.001	Significant
ASPP-Swin Transformer	95.62	0.16	95.48 – 95.76	1.13	4.82	0.003	Significant
Attention EfficientNet-B0	95.18	0.20	95.00 – 95.36	1.57	–	–	–
CNN + mRMR Feature Selection	95.46	0.17	95.31 – 95.61	1.29	5.16	0.002	Significant
Proposed CSCG-ResNet + M-MGO + QI-TabNet	96.75	0.12	96.64 – 96.86	–	–	–	Reference Model

To further validate the effectiveness and generalization capability of the CSCG-ResNet, an additional experiment is conducted using the Gastrointestinal Bleeding WCE Images Dataset. The dataset consists of gastrointestinal bleeding and normal images. The results are given in Table 8. The CSCG-ResNet model achieved an accuracy of 95.82%, with precision, recall, and F1-score values of 95.76%, 95.91%, and 95.83%, respectively. The consistent performance on an independent dataset proves that the proposed model is not overfitted to the primary dataset and can maintain reliable classification performance under different image acquisition conditions. To further strengthen the clinical applicability and robustness analysis, the GastroVision dataset is considered as an additional multi-centre validation resource. GastroVision is an open-access gastrointestinal endoscopy image dataset. It consists of 8,000 images with different gastrointestinal abnormalities. The dataset is collected from Baerum Hospital, Norway, and Karolinska University Hospital, Sweden. The model achieved an accuracy of 94.96%, with precision, recall, and F1-score values of 94.72%, 95.08%, and 94.89%, respectively. The obtained results demonstrate that CSCG-ResNet maintains stable classification performance across different institutions and acquisition devices. External

validation is performed to assess the generalization capability of the proposed model on different dataset distributions. Since the external datasets contain different class numbers, direct class-wise comparison with the eight-class Kvasir dataset was not considered. The evaluation was conducted based on the available categories of each dataset to verify the robustness of the learned feature representation under different imaging conditions.

Table 8. External dataset validation

Dataset	Classes	Ac (%)	Pr (%)	Re (%)	F1 (%)
Gastrointestinal Bleeding WCE Images Dataset	Bleeding / Normal	95.82	95.76	95.91	95.83
GastroVision Dataset	27 GI classes	94.96	94.72	95.08	94.89

The ablation study was conducted to analyse the contribution of each component in the CSCG-ResNet model. The outcome of the CSCG-ResNet is given in Table 9. The baseline ResNet model achieves lower performance due to limited feature extraction capability. The CSCG-ResNet model accuracy increases because the cascaded spatial convolution and residual learning improve feature representation. Also, the model with M-MGO further improves the performance by selecting optimal features and eliminating redundant information. Finally, the addition of QI-TabNet achieves additional improvement by enhancing classification capability. The complete framework achieves the highest accuracy of 96.75%.

Table 9. Ablation Study of CSCG-ResNet Model

Model Variant	Ac (%)	Pr (%)	Re (%)	F1 (%)
ResNet18 + TabNet	92.74	92.69	92.72	92.70
ResNet18 + QI-TabNet	94.21	94.15	94.18	94.16
CSCG-ResNet + TabNet	95.08	95.02	95.05	95.04
CSCG-ResNet + QI-TabNet	95.86	95.79	95.83	95.81
ResNet18 + M-MGO + TabNet	95.71	95.64	95.68	95.66
CSCG-ResNet + M-MGO + TabNet	96.24	96.18	96.21	96.20
CSCG-ResNet + M-MGO + QI-TabNet (Proposed)	96.75	96.68	96.76	96.85

The classification performance of the CSCG-ResNet model is compared with other classifiers. The extracted features from the CSCG-ResNet are classified using different classifiers. The outcomes are given in Table 10. The QI-TabNet achieves the highest classification accuracy of 96.75%. The QI-TabNet improves the accuracy by 1.73% compared with Standard TabNet and provides better classification performance than conventional classifiers.

Table 10. Comparison of proposed QI-TabNet with other classifiers

Classifier	Ac (%)	Pr (%)	Re (%)	F1(%)
Support Vector Machine (SVM)	91.86	91.72	91.80	91.75
k-Nearest Neighbor (KNN)	90.94	90.82	90.91	90.86
Random Forest (RF)	92.74	92.65	92.71	92.68
Extreme Learning Machine (ELM)	93.28	93.16	93.25	93.20
Multilayer Perceptron (MLP)	94.12	94.05	94.18	94.11
Standard TabNet	95.02	94.96	95.04	95.00
Proposed QI-TabNet	96.75	96.68	96.76	96.85

The performance comparison of the M-MGO optimizer with other existing optimization algorithms is given in Table 11. All optimization algorithms show similar computational complexity of O(T×N×D). The proposed modifications do not add an additional computational burden. However, the proposed M-MGO achieves faster convergence with only 60 iterations compared with GA, PSO, GWO, and MPO. Compared with the standard MPO algorithm, the proposed M-MGO provides an improvement of 1.67%. This confirms the advantage of the modified optimization strategy in achieving better convergence and feature selection capability.

Table 11. Comparison of different optimization algorithms

Optimization Algorithm	Time Complexity	Convergence Iterations	Ac (%)	Pr (%)	Re (%)	F1 (%)
Genetic Algorithm (GA)	O(T×N×D)	96	93.84	93.72	93.80	93.75
Particle Swarm Optimization (PSO)	O(T×N×D)	88	94.16	94.05	94.12	94.08
Grey Wolf Optimizer (GWO)	O(T×N×D)	82	94.52	94.41	94.48	94.45
Marine Predator Optimization (MPO)	O(T×N×D)	74	95.08	95.02	95.11	95.06
Modified Mountain Gazelle Optimizer (M-MGO)	O(T×N×D)	60	96.75	96.68	96.76	96.85

The sensitivity analysis of important hyperparameters is given in Table 10. To avoid test-set optimization bias, the hyperparameter selection process is performed only using the training data with internal validation during model development. The testing subset was not involved in parameter tuning or selection of optimal values. The values reported in Table 12 represent the parameters that provided the best validation performance during the training phase. After fixing the optimal parameters, the final model evaluation is performed using the independent testing subset. The reported results reflect the generalization capability of the proposed model rather than optimization toward the test data. From the results, it was observed that the M-MGO population size affects the balance between optimization diversity and computational cost. A learning rate of 0.001 achieves optimal convergence by balancing training stability and parameter updating speed. For QI-TabNet, five decision steps provide sufficient feature refinement without increasing model complexity. The selected dropout rate and CSCG channel reduction ratio improve generalization and feature representation.

Table 12. Hyperparameter Sensitivity Analysis

Parameter	Tested Values	Optimal Value	Accuracy (%)
M-MGO Population Size	10, 20, 30, 40, 50	30	96.75
Maximum Iterations of M-MGO	25, 50, 75, 100, 125	100	96.75
Learning Rate	0.01, 0.005, 0.001, 0.0005	0.001	96.75
QI-TabNet Decision Steps	3, 4, 5, 6, 7	5	96.75
Feature Selection Threshold	0.3, 0.4, 0.5, 0.6, 0.7	0.5	96.75
Dropout Rate	0.1, 0.2, 0.3, 0.4	0.2	96.75
CSCG Channel Reduction Ratio	4, 8, 16, 32	16	96.75

The computational analysis of the model is given in Table 13. Although transformer-based models like EfficientViT and ASPP-Swin Transformer provide competitive accuracy, they require more parameters. The proposed model achieves improved classification accuracy with moderate computational complexity. The inference time of 24.6 ms per image and memory requirement of 126.3 MB indicate that the proposed model provides a suitable balance between diagnostic performance and computational efficiency.

Table 13. Model complexity and computational analysis

Model	Parameters (Million)	FLOPs (GFLOPs)	Inference Time (ms/image)	Memory Usage (MB)	Accuracy (%)
ResNet18 + TabNet	12.45	1.82	18.6	96.4	92.74
EfficientNet-B0	5.29	0.39	15.8	82.6	95.18
Efficient Vision Transformer	86.50	17.60	42.5	312.8	95.34
ASPP-Swin Transformer	88.20	18.40	45.2	328.5	95.62
CSCG-ResNet + M-MGO + TabNet	14.82	2.15	22.4	118.7	96.24
CSCG-ResNet + M-MGO + QI-TabNet	16.35	2.48	24.6	126.3	96.75

To enhance the interpretability of the proposed CSCG-ResNet-MGO, Grad-CAM visualisation is used to highlight the discriminative regions for classification decisions, as shown in Figure 8. The generated heatmaps denote that the model effectively focuses on clinically significant regions like bleeding areas and abnormal mucosal patterns. The high-activation regions correspond well with the pathological regions observed in the original endoscopic images. The proposed model learns meaningful disease-specific features and improves transparency.

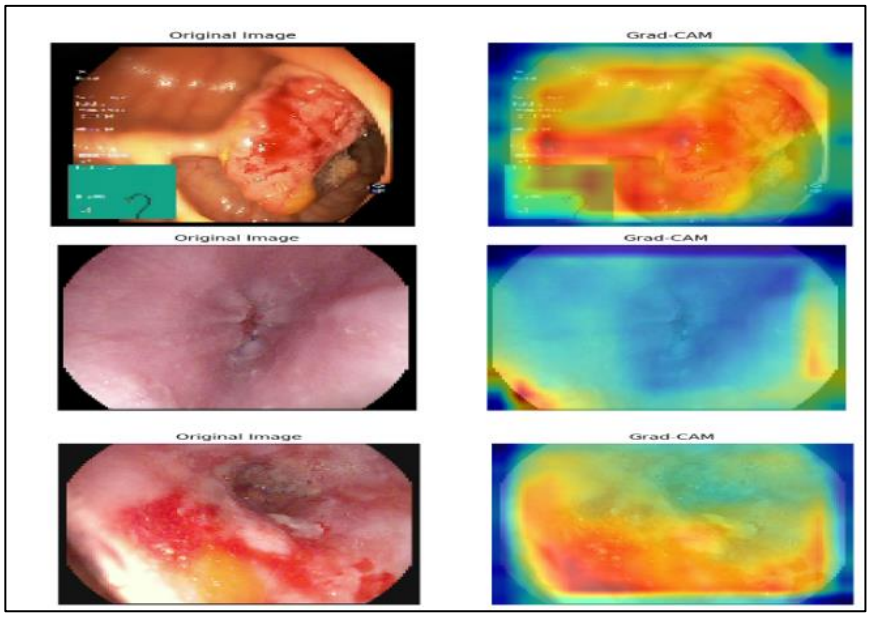

Figure 8. Grad-CAM visualisation of the model

5. Conclusion

In this work, a threefold DL model is recommended for GI tract disease categorization. It combines CSCG-ResNet and M-MGO for multi-class classification. In addition, a new variant of TabNet called QI-TabNet is proposed for classification. In validation, the proposed model outperformed existing approaches with an accuracy of 96.75%. Future work will focus on expanding the dataset using real-time gastrointestinal endoscopic images from clinical settings.

Declaration

Currently, ethical committee approval is in process under the supervision of Dr A.C. Arun, M.D., D.M. (Gastroenterology), PGI Chandigarh, Consultant Gastroenterologist, Hepatologist, and Endoscopist. Lily Hospital, Madurai, Tamil Nadu, India. It supports further validation with larger and more varied clinical datasets to improve the generalizability and real-world applicability of this research.

References

[1] Mukhtorov, Doniyorjon, Madinakhon Rakhmonova, Shakhnoza Muksimova, and Young-Im Cho. "Endoscopic Image Classification Based on Explainable Deep Learning." Sensors 23, no. 6 (2023): 3176.

[2] Li, Sheng, Jing Cao, Jiafeng Yao, Jinhui Zhu, Xiongxiong He, and Qianru Jiang. "Adaptive Aggregation with Self-Attention Network for Gastrointestinal Image Classification." IET Image Processing 16, no. 9 (2022): 2384-2397.

[3] Asif, Sohaib. "A Fuzzy Minkowski Distance-Based Fusion of Convolutional Neural Networks for Gastrointestinal Disease Detection." Applied Soft Computing 158 (2024): 111595.

[4] Khan, Muhammad Attique, Seifedine Kadry, Majed Alhaisoni, Yunyoung Nam, Yudong Zhang, Venkatesan Rajinikanth, and Muhammad Shahzad Sarfraz. "Computer-Aided Gastrointestinal Diseases Analysis from Wireless Capsule Endoscopy: A Framework of Best Features Selection." IEEE Access 8 (2020): 132850-132859.

[5] Naz, Javeria, Muhammad Sharif, Mussarat Yasmin, Mudassar Raza, and Muhammad A. Khan. "Detection and Classification of Gastrointestinal Diseases Using Machine Learning." Current Medical Imaging Reviews 17, no. 4 (2021): 479-490.

[6] Öztürk, Şaban, and Umut Özkaya. "Gastrointestinal Tract Classification Using Improved LSTM based CNN." Multimedia Tools and Applications 79, no. 39 (2020): 28825-28840.

[7] Ramzan, Muhammad, Mudassar Raza, Muhammad Sharif, Muhammad Attique Khan, and Yunyoung Nam. "Gastrointestinal Tract Infections Classification Using Deep Learning." Comput. Mater. Contin 69, no. 3 (2021): 3239-3257.

[8] Rahman, AMJ MD Zubair, R. Mythili, K. Chokkanathan, T. R. Mahesh, K. Vanitha, and Temesgen Engida Yimer. "Enhancing Image-Based Diagnosis of Gastrointestinal Tract Diseases Through Deep Learning with EfficientNet and Advanced Data Augmentation Techniques." BMC Medical Imaging 24 (2024): 1.

[9] Haile, Melaku Bitew, Ayodeji Olalekan Salau, Belay Enyew, and Abebech Jenber Belay. "Detection and Classification of Gastrointestinal Disease Using Convolutional Neural Network and SVM." Cogent Engineering 9, no. 1 (2022): 2084878.

[10] Rubab, Saddaf, Muhammad Jamshed, Muhammad Attique Khan, Nouf Abdullah Almujally, Robertas Damaševičius, Amir Hussain, Neunggyu Han, and Yunyoung Nam. "Gastrointestinal Tract Disease Classification from Wireless Capsule Endoscopy Images Based on Deep Learning Information Fusion and Newton Raphson Controlled Marine Predator Algorithm." Scientific Reports 15, no. 1 (2025): 32180.

[11] Nouman Noor, Muhammad, Muhammad Nazir, Sajid Ali Khan, Imran Ashraf, and Oh-Young Song. "Localization and Classification of Gastrointestinal Tract Disorders Using Explainable AI from Endoscopic Images." Applied Sciences 13, no. 15 (2023): 9031.

[12] Mohapatra, Subhashree, Girish Kumar Pati, Manohar Mishra, and Tripti Swarnkar. "Gastrointestinal Abnormality Detection and Classification Using Empirical Wavelet Transform and Deep Convolutional Neural Network from Endoscopic Images." Ain Shams Engineering Journal 14, no. 4 (2023): 101942.

[13] Ahamed, Md Faysal, Md Nahiduzzaman, Md Rabiul Islam, Mansura Naznine, Mohamed Arselene Ayari, Amith Khandakar, and Julfikar Haider. "Detection of Various Gastrointestinal Tract Diseases Through a Deep Learning Method with Ensemble ELM and Explainable AI." Expert Systems with Applications 256 (2024): 124908.

[14] Şener, Abdullah, and Burhan Ergen. "Automatic Detection of Gastrointestinal System Abnormalities Using Deep Learning-Based Segmentation and Classification Methods." Health Information Science and Systems 13, no. 1 (2025): 37.

[15] Sivari, Esra, Erkan Bostanci, Mehmet Serdar Guzel, Koray Acici, Tunc Asuroglu, and Tulin Ercelebi Ayyildiz. "A New Approach for Gastrointestinal Tract Findings Detection and Classification: Deep Learning-Based Hybrid Stacking Ensemble Models." Diagnostics 13, no. 4 (2023): 720.

[16] Tanwar, Vishesh, Bhisham Sharma, Dhirendra Prasad Yadav, and Abolfazl Mehbodniya. "Hybrid Deep Learning Framework Based on EfficientViT for Classification of Gastrointestinal Diseases." Scientific Reports 15, no. 1 (2025): 26982.

[17] Khan, Sultan Daud, Saleh Basalamah, and Ahmed Lbath. "Multi-Module Attention-Guided Deep Learning Framework for Precise Gastrointestinal Disease Identification in Endoscopic Imagery." Biomedical Signal Processing and Control 95 (2024): 106396.

[18] Abian, Arefin Ittesafun, Mohaimenul Azam Khan Raiaan, Mirjam Jonkman, Sheikh Mohammed Shariful Islam, and Sami Azam. "Atrous Spatial Pyramid Pooling with Swin Transformer Model for Classification of Gastrointestinal Tract Diseases from Videos with Enhanced Explainability." Engineering Applications of Artificial Intelligence 150 (2025): 110656.

[19] Hussein, Ahmed F., Auns Q. Al-Neami, and Noor Kamal Al-Qazzaz. "Transfer Learning and Hybrid Deep Convolutional Neural Networks for Detection and Classification of Gastrointestinal Diseases." International Journal of Information Technology (2025): 1-15.

[20] Öztürk, Şaban, and Umut Özkaya. "Residual LSTM Layered CNN for Classification of Gastrointestinal Tract Diseases." Journal of Biomedical Informatics 113 (2021): 103638.

[21] Gunasekaran, Hemalatha, Krishnamoorthi Ramalakshmi, Deepa Kanmani Swaminathan, and Manuel Mazzara. "GIT-Net: An Ensemble Deep Learning-Based GI Tract Classification of Endoscopic Images." Bioengineering 10, no. 7 (2023): 809.

[22] Majid, Abdul, Muhammad Sharif, and Mehwish Zafar. "A Lightweight Attention-Based Framework with Multi-Scale Deep Supervision and Boundary-Aware Learning for Gastrointestinal Tract Image Segmentation." Signal, Image and Video Processing 20, no. 2 (2026): 78.

[23] Siddiqui, Samra, Junaid A. Khan, Tallha Akram, Meshal Alharbi, Jaehyuk Cha, and Dina A. AlHammadi. "SNet: A Novel Convolutional Neural Network Architecture for Advanced Endoscopic Image Classification of Gastrointestinal Disorders." SLAS technology 33 (2025): 100304.

[24] https://datasets.simula.no/kvasir/

[25] Rauf, Rehana, Samia Ijaz, Saima Gulzar Ahmad, Ehsan Ullah Munir, and Naeem Ramzan. "Gastrointestinal Endoscopic Image Classification Using a Hybrid Modified Inception Network and a Customized Vision Transformer with Tree Growth Feature Selection." Scientific Reports (2026).