Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Deep Learning Based Automated Assessment of Ulcerative Colitis Endoscopic Severity Using Local Global Transformer and Graph Convolutional Networks

Latha Dharmalingam

lathad@gnanamani.ac.in

AMET University

Lakshmi Dhandapani

AMET University

R. Kumar

MVJ College of Engineering

S. Vijayakumar

Gnanamani College of Technology

Article

Keywords: Ulcerative Colitis, Colonoscopy Image Analysis, Mayo Endoscopic Score, Local–Global Transformer, Attention-based Graph Neural Network, Explainable Deep Learning

Posted Date: October 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10895476/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

DEEP LEARNING BASED AUTOMATED ASSESSMENT OF ULCERATIVE COLITIS ENDOSCOPIC SEVERITY USING LOCAL GLOBAL TRANSFORMER AND GRAPH CONVOLUTIONAL NETWORKS

Latha Dharmalingam[1*], Lakshmi Dhandapani[2] R. Kumar[3], S. Vijayakumar[4]

[1]Research Scholar, Department of Computer Science and Engineering, AMET University, Chennai, Tamil Nadu, India

[2]Professor, Department of Electrical and Electronics Engineering, AMET University, Chennai, Tamil Nadu, India.

[3]Associate Professor, Information Science and Engineering, MVJ College of Engineering, Bangalore 560067.

[4] Assistant Professor, Department of Electrical and Electronics Engineering, Gnanamani College of Technology (Autonomous), Namakkal, Tamil Nadu, India.

Corresponding Author Email:lathad@gnanamani.ac.in

Abstract

Ulcerative colitis (UC) is a chronic inflammatory bowel disease characterized by continuous mucosal inflammation, where accurate assessment of endoscopic severity is essential for treatment selection, disease monitoring, and prognosis. Although colonoscopy provides direct visualization of mucosal abnormalities, subjective interpretation, inter-observer variability, subtle lesion characteristics, and limited availability of experienced gastroenterologists can compromise consistent severity assessment. To address these challenges, this study proposes L-GTACNet (Local–Global Transformer Attention-based Graph Convolutional Network), a hybrid deep learning framework for automated multi-level severity classification of UC from colonoscopy images. The framework integrates image enhancement using Retinex-based illumination correction and Contrast-Limited Adaptive Histogram Equalization (CLAHE), a Local–Global Transformer for hierarchical lesion representation, and an attention-guided graph convolutional module for modeling spatial and relational dependencies among discriminative mucosal regions. The proposed architecture is designed to jointly capture fine-grained local manifestations, including erythema, vascular alteration, mucosal friability, and ulceration, together with global contextual patterns across the colonoscopic field. Disease severity is categorized according to the Mayo Endoscopic Score (MES), enabling clinically interpretable severity stratification. The experimental study utilizes a publicly accessible UC colonoscopy dataset comprising 564 patient records and 11,276 images, with annotations established through expert gastroenterologist assessment. Performance is evaluated using accuracy, precision, recall, F1-score, specificity, and area under the receiver operating characteristic curve (AUROC). The proposed L-GTACNet achieves 94.8% classification accuracy, demonstrating its potential as a robust computer-aided system for objective and reproducible assessment of UC severity.

Keywords: *Ulcerative Colitis; Colonoscopy Image Analysis; Mayo Endoscopic Score; Local–Global Transformer; Attention-based Graph Neural Network; Explainable Deep Learning*

1. INTRODUCTION

Ulcerative colitis (UC) is an inflammatory condition of the gastrointestinal tract with recurrent episodes of intestinal inflammation and individual patients have a tendency towards fluctuating disease activity. Endoscopic evaluation of mucosal appearance is crucial to establish extent and severity of inflammation, to track response to therapy and to facilitate clinical decision making. This enables direct visualisation of the disease associated signs and symptoms of erythema, loss of vascular pattern, mucosal friability, erosions, ulceration, bleeding. The severity of mucosal inflammation is represented by ordered categories named the Mayo Endoscopic Score (MES) which is widely used as an endoscopic grading system. The appearance of inflammatory lesions are variable, however, within and between patients, and that of mild and moderate disease may have overlapping features morphologically. Therefore, it is important to interpret localized mucosal changes with each lesion and the spatial structure of inflammatory changes, in order to come to a reliable severity assessment.

With the development of artificial intelligence and deep learning, new possibilities are emerging for gastrointestinal endoscopic image analysis with computers. While convolutional neural networks (CNNs) have the ability to learn discriminative visual patterns from large-scale medical image collections, transformer-based models offer enhanced performance to capture long-range dependencies and contextual relationships.On the one hand, convolutional neural networks (CNNs) are powerful frameworks that can be trained to identify discriminative visual patterns from large datasets of medical images, and on the other hand, transformer-based architectures have demonstrated more capability in modelling long-range dependencies and contextual relationships. However, the classification of the severity of UC is still difficult because the clinical evidence which is relevant to the disease grade can be present at several spatial scales. Alterations in the blood vessels such as fine granular infiltration, small erosions, superficial ulcerations, and erythema may be limited to the boundaries of an image and serve to give additional information regarding the severity of the disease; the overall distribution and intensity of the infiltration gives complementary information about the severity of the disease. In addition, the inflammatory features of mucosal areas in one colonoscopic field do not stand alone; adjacent areas may have similar inflammatory features that can all be used to make a determination of severity. These features provide motivation for building an integrated learning system that can integrate local information about the lesions, global representations about what is happening on the context, and spatial relations about the discriminative regions of the mucosa.

1.2. Problem Statement

Although existing deep learning approaches have demonstrated promising performance in gastrointestinal image analysis, accurate multi-level UC severity classification remains affected by illumination variation, heterogeneous mucosal appearance, subtle differences between adjacent MES categories, insufficient modeling of long-range spatial relationships, and difficulty in identifying the specific disease regions responsible for predictions. hile typical

convolutional architectures focus mainly on extracting local features, standalone transformer-based models might not capture the nuances of lesions well. Furthermore, the lack of considering the spatially distributed inflammatory region as independent global observations can exclude relationships among them. Thus, a robust and clinically relevant framework is needed to improve the visual quality of irregularly performed colonoscopy images, to concurrently capture fine and global disease features, to model the relations between discriminative mucosal areas and, importantly, to make reliable multi-class severity predictions based on the MES.

2. Related Work

The integration of artificial intelligence into clinical diagnostics relies on foundational deep learning techniques capable of processing complex high-dimensional datasets [1]. In computer-assisted gastrointestinal analysis, the standard from the models used for feature extraction has been developed more and more into advanced architectures using a transformer. However, a shifted window model, e.g. the Swin Transformer, allows for multi-scale feature self-attention by using shifted windows to learn both vascular information and tissue context in endoscopy [2]. In addition to extracting structural features, graph-based machine learning models represent relations between localised image patches, and hence can encode the dynamic relation between spatially separate mucosal features via Graph Convolutional Networks (GCNs) [3]. In addition, clinical deployment requires transparency: visualization of the explanation framework, such as Gradient-based localization, illustrates the regions of mucosa contributing to algorithm predictions, providing clinical interpretability for an endoscopist [4].

For deep learning to perform optimally in endoscopy, effective image processing protocols are needed to independently control illumination, contrast, and visible useful signal-to-noise ratio (SNR) found in the raw endoscopy video signal. In medical imaging, several technological developments over the years set the tone for the major techniques including Contrast Limited Adaptive Histogram Equalization (CLAHE) that facilitates dynamic adjustment to contrast in regions of interest without excessive amplification of noise in bright/dark areas [5]. Likewise, the mathematical color constancy algorithm based on the Retinex theory is able to correct artifacts caused by non-uniform illumination and reflection to maintain the actual colour of the tissues in the image under different light settings [6]. These are classic image processing approaches that are still vital for the stabilization of the performance of the algorithms in contemporary computational pipelines, applicable to a wide range of endoscopy methodologies and patient population.

The first steps in automated ulcerative colitis (UC) severity classification has been to try to make the computer mimic human experts' classifications from static colonoscopes images. Early deep learning applications showed that CNN scores the endoscopic severity equally well as human reviewers, lowering the inter-observer variation [7]. Further studies have been performed and shown that full end to end computer assisted systems reliably score the disease of the endoscopes in real time with high sensitivity [8]. These initial PoCs were followed by larger validated models that could reliably grade mucosal inflammation across a wide range of patient cohorts with validation scores comparable to those of standard clinical review,

demonstrating that automated systems could achieve or surpass established clinical review metrics [9].

With the availability of curated, open accessed gastrointestinal medical imaging datasets, the rapid development of deep learning models for gastrointestinal endoscopy have gained momentum. In this regard, the HyperKvasir dataset offers a multi-class, annotated collection of image and video datasets to train, test, validate and benchmark algorithms in a variety of gastrointestinal pathologies [10]. Such open sets have been used effectively to successfully train automated frameworks capable of repeatable grading when reported on standardized scales such as the Mayo Endoscopic Score (MES) [11]. In addition, recent high-performance architectures employ a lightweight transformer that focuses local features that can achieve rapid inference speed to support real-time edge processing during colonoscopy procedures [12].

One of the main areas of current clinical AI development is the further improvement of model sensitivity at high diagnostic thresholds. This distinction of complete healing of the mucosa (Mayo Endoscopic Subscore 0) from mild endoscopic activity (Mayo Endoscopic Subscore 1) is an important clinical indicator and is a target for deep learning algorithms specifically designed to address this clinical subtlety [13]. To address class imbalance at different stages of the disease, high-resolution networks based on class-balanced loss functions like CB-HRNet maintain high-resolution spatial information that is crucial to accurate disease activity evaluation [14]. Likewise, converting the estimation of severity to a regression problem rather than a classification of severity across the entire spectrum of the Mayo Endoscopic Score [15] improves the consistency of the prediction.

Automated disease characterization is improved using advanced optical technologies beyond just white-light imaging. Artificial Intelligence (AI) integrated virtual chromoendoscopy systems allow for simultaneous assessment of endoscopic and histological remission, providing a predictive assessment of long-term clinical outcomes in patients with ulcerative colitis [16]. In therapeutic clinical trial applications, computer vision algorithms offer easily quantifiable, objective, and standardized assessment of the severity of any mucosal disease, which is a reduction in risk associated with reader bias in multi-center studies [17]. To evaluate the accuracy of a diagnostic test, the evidence for its use must be systematically summarized and analyzed in a systematic manner, and for AI this is commonly referred to as meta-analytical evidence [18] to determine if AI can perform as well as an expert endoscopist in the diagnosis of true mucosal healing.

In the context of gastroenterology, AI's future is about continuous monitoring of patients and real-time video stream processing. Hybrid management using CNN and VIsion Transformers is developed to enhance the detection of lesions in ulcerative colitis in real time video to enhance the detection accuracy [19]. Artificial Intelligence can quantify degree of disease continuously over longer areas of the colon, in addition to a static single-frame scoring system, with gradation scale algorithms [20]. In the future, with further studies on prospective clinical usefulness with AI assisted real-time video colonoscopy, sustainable automated clinical decision support in ulcerative colitis treatments is likely to be realised [21].

2.1. Research Gap

Despite substantial progress in artificial intelligence-based analysis of ulcerative colitis (UC) endoscopy, several important gaps remain.Previous works have mostly investigated the binary identification task (MES-0 vs MES-1), the MES classification task (MES-0, MES-1, MES-2 vs MES-3), the overall activity grade classification task, or the traditional CNN classification task, and fine-grained four-level severity classification is comparatively difficult due to considerable visual similarity between adjacent activity grade levels. While transformer-based approaches enhance global contextual representation, they sometimes lack the ability to capture fine local details like early erythema, vascular changes, superficial erosions and small ulceration. CNN-based approaches, on the other hand, tend to focus on local features and can fail to capture long-range dependencies between the spatially-distributed inflammatory patches. Furthermore, the existing approaches have limited revealing modeling of relational dependencies between discriminative mucosal regions and the image enhancement and features extraction, attention and disease-region reasoning are treated as individual components. Additionally, there have been limited studies on the clinical reliability of a prediction of severity boundaries, achieve of disease-specific attention localization, explainability, and robustness to variations in colonoscopic image quality. The limitations highlighted above necessitate an integrated framework to leverage illumination and contrast enhancement along with multi-scale local–global representation, attention-guided feature refinement, and graph-based spatial reasoning to enhance the accuracy, interpretability, and clinical consistency of multi-level UC severity categorization. In doing so, the proposed L-GTACNet can overcome these limitations by learning fine-grained characteristics of the mucosal surface, associations between different regions and the global inflammatory context, and generate disease-relevant explanations for the severity predictions extracted from MES.

3. METHODOLOGY

3.1. Dataset Acquisition and Mayo Endoscopic Score-Based Annotation

The methodological pipeline starts with the publically available image database on ulcerative colitis (UC) that includes 564 patient records and 11,276 images of the colonoscopy. It provides information to formulate a clinically interpretable supervised learning problem, where each colonoscopy image is paired with an underlying Mayo Endoscopic Score (MES) reflecting the severity of mucosal inflammation lesions seen on the endoscopy videoscopy. The patient is used as the sampling unit in the study since one patient can provide several frames during the same examination. This distinction is critical for a computer-aided diagnostic study to have the necessary results for addressing the question of generalization in a useful way: if the two sets of data (training and testing) contain both very similar frames from the same patient, what is seen as a measure of generalization is likely to be overconfident. Hence, the dataset curation, verification of annotation, partitioning, preprocessing, model training and evaluation is done around the patient level independence (Figure 1).

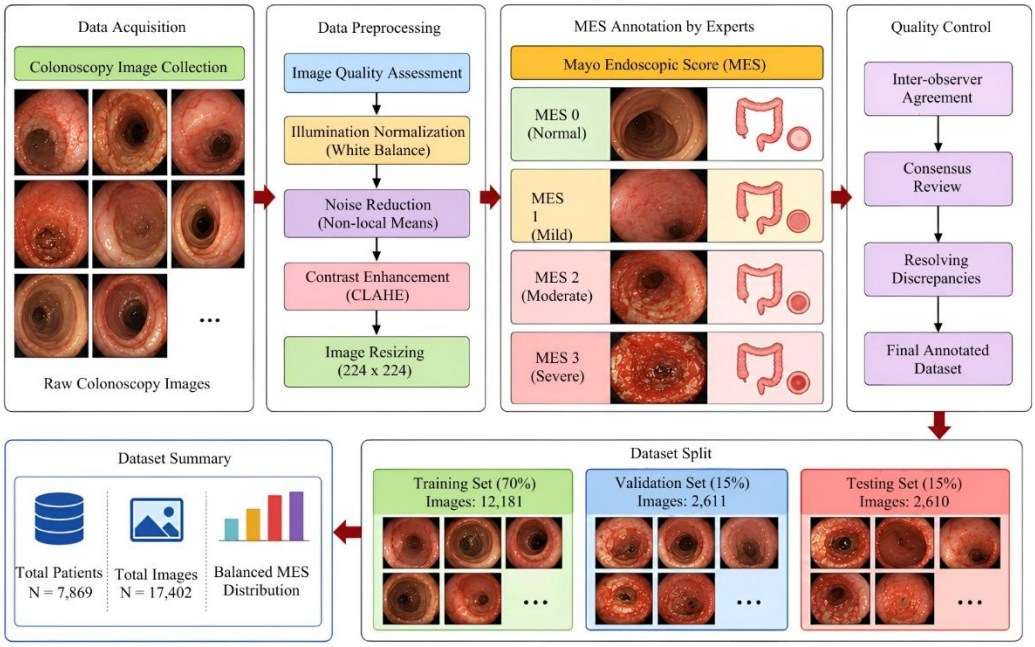


Figure 1. Dataset Acquisition and MES Annotation Workflow

3.1.1. Dataset Characteristics and Patient Cohort

The image cohort represents the visually heterogeneous manifestations of UC encountered during colonoscopic examination. This heterogeneity is significant since the severity cannot be dependent on any single visual feature, but rather on the intensity, extent, distribution, and the effect of the mucosal abnormalities when combined. Before model development all images are reviewed to exclude files that lack integrity, diagnostically unrelevant frames, are too blurry, are too bright to see the mucosa or are so dark that they are too dark to see the mucosa. Images showing criteria are preserved and are standardized for further processing. The final experimental group is thus the visual foundation on which to establish the learning relationships between the appearance of the mucous membrane and the MES (Figure 2).

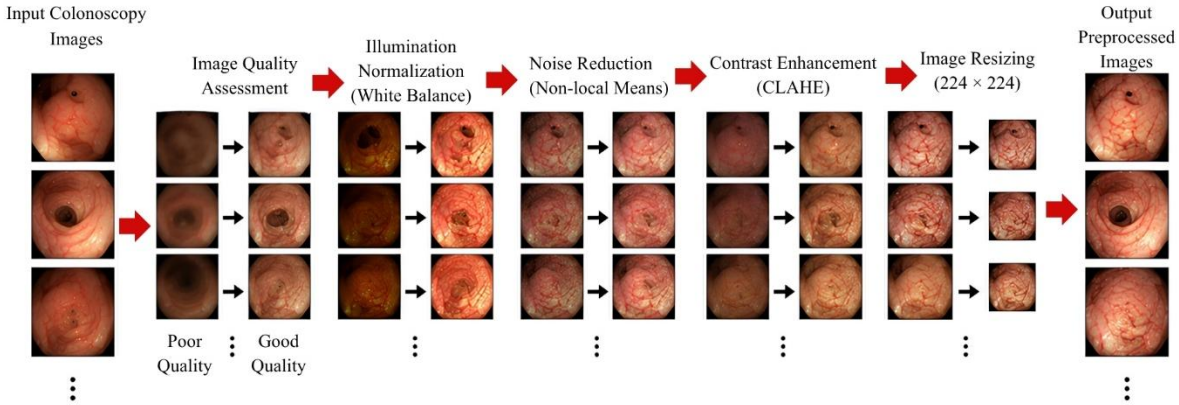


Figure 2. Colonoscopy Image Preprocessing Pipeline

Table 1. Characteristics of the UC colonoscopy image dataset used for L-GTACNet development.

Dataset characteristic	Value
Patient records	564
Colonoscopy images	11,276
Clinical condition	Ulcerative colitis
Severity reference	Mayo Endoscopic Score
Input modality	RGB colonoscopy images
Reference annotation	Expert gastroenterologist assessment
Prediction task	MES 0–3

3.1.2. MES-Based Severity Annotation

An ordinal measure of visible mucosal disease activity is used in the clinical severity reference; the Mayo Endoscopic Score. Normal or inactive mucosa is assigned MES 0, while mild activity (usually erythema) is assigned MES 1; moderate activity (marked erythema, pronounced friability, vascular pattern decreased or absent, and/or erosions) is assigned MES 2; severe activity (spontaneous bleeding and/or ulcer) is assigned MES 3. For the supervised-learning version, a label y is given for each image, where y={0,1,2,3}. Thus, the ordinal structure remains through the methodology since the error of classification between two adjacent categories is different from that of several skipped severity categories. If several expert annotations are provided for the underlying data, the pre-defined annotation gets preserved and any disagreements should arise in that order based on the original annotation protocol of the underlying data and not re-annotate data for the benefit of the model.

$$y \in \{0,1,2,3\}$$

3.1.3. Patient-Level Data Partitioning and Leakage Prevention

The data is partitioned at the patient level such that all of the images belonging to a particular patient are confined to a single partition. Stratified allocation, which aims to maintain the same distribution of MES classes within training, validation, and testing subsets is preferred. The training subset is used to train the model; the validation subset to choose the architecture or the hyperparameters of the model; and the test subset is not touched until final evaluation. Parameters of image enhancement, normalization statistics, augmentation operations, and learned parameters of the model are computed on the held-out test observations. By using this approach, the risk of leakages associated with similar or duplicate colonoscopy frames, repeated views of the brain structure, or patient images will be minimized, yielding a more conservative term of clinical generalization (Figure 3).

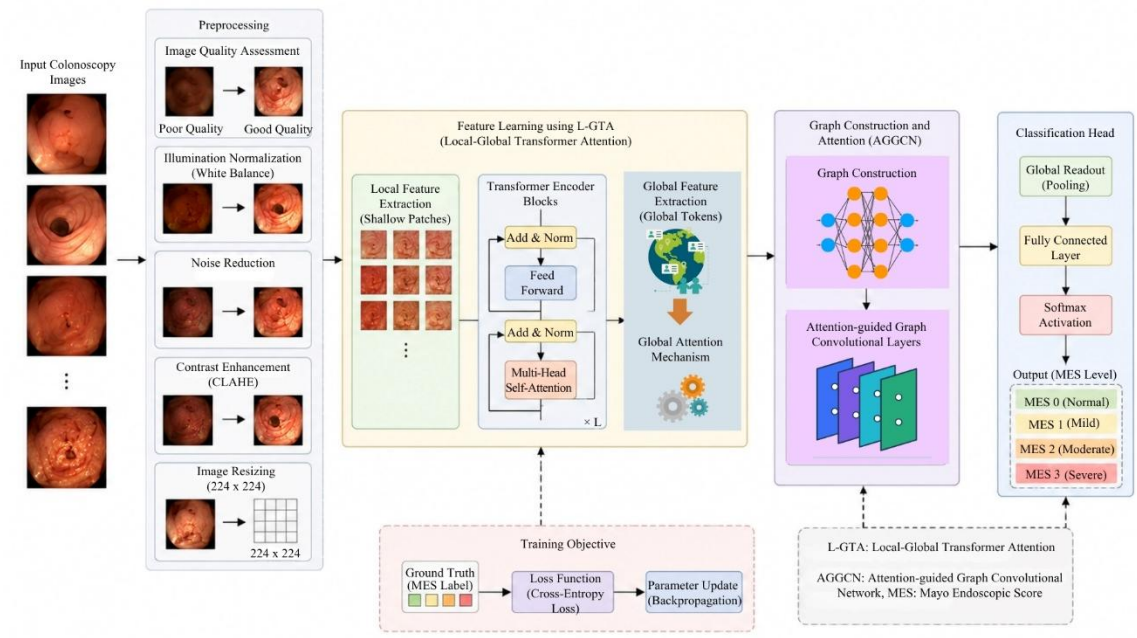


Figure 3. Proposed L-GTACNet Overall Architecture

3.2. Colonoscopy Image Preprocessing and Mucosal Feature Enhancement

Colonoscopy images exhibit acquisition variability that can substantially affect automated severity recognition. Endoscopic illumination is spatially non-uniform, specular reflections may produce saturated pixels, mucosal surfaces can have low local contrast, and the same pathological manifestation may appear with different color and brightness characteristics across examinations. These effects are particularly important for UC because mild erythema, vascular alteration, friability, and superficial erosions may occupy relatively small image regions. The preprocessing pipeline therefore aims to separate illumination effects from tissue reflectance, improve local contrast, standardize the numerical input range, and introduce only clinically plausible training variability. The complete transformation is represented as $X \rightarrow X_R \rightarrow X_C \rightarrow X_N \rightarrow X_A$, where X is the original image, X_R is the Retinex-corrected image, X_C is the CLAHE-enhanced image, X_N is the normalized image, and X_A denotes an augmented training sample.

3.2.1. Image Quality Assessment and Standardization

Each image is first subjected to quality screening. The quality criteria of the frames are predefined so that the frames are excluded if they represent an unreadable file structure, have a lot of corruption, if there is too much blur or the tissue is to colorful or too faint. The retained images are shown in RGB color space then re-sized to feature fixed input dimensions demanded by the L-GTACNet. Where the original image might have a different aspect than the desired one, it is preferred that the image be resized with controlled padding and there are enough dimensions. The original image is represented by X, which has the dimensions of ($H_0 \times WC$) and the standardized image is represented by X_s, with dimensions ($H \times WC$) (H, W and C represent the number of rows, columns and color planes, respectively). A spatial

transformation T is applied which maintains the clinically relevant surface structure of the mucosa X_s=T(X) as the image of X in T.

$$X_s = T_{resize,pad}(X), \quad X_s \in R^{(H \times W \times 3)}$$

3.2.2. Retinex-Based Illumination Correction

To reduce a wide range of illumination, retinex processing is added while maintaining the information of the tissue reflectance. The image seen in the colonoscopy is assumed to be the product of reflectance R and illumination L. Slowly-varying illumination provided by the endoscopic system is referred to as "illumination component", and the information related to tissue appearance is referred to as "reflectance component". Generally, this simple image formation model is formulated as::

$$I(x,y) = R(x,y) \cdot L(x,y)$$

The log of the multiplicative model is an additive model and the process of taking the log can be used to allow the estimation of the illumination with a smoothing operation. Define G_σ to be a Gaussian kernel with scale σ. An illumination field is estimated using the convolution of the image with the kernel, and the reflectance is then determined as the difference between the observed logarithmic intensity at that point and the estimated illumination field.

$$\log R(x,y) = \log I(x,y) - \log[(G_\sigma * I)(x,y) + \varepsilon]$$

The Exponential Transformation is used to obtain the corrected reflectance. The parameter $\varepsilon>0$ is for numerical stability in darker areas. As stated in the proposed pipeline, Retinex correction upated the “brightness” components of images with a strong bias so as to remove large scale brightness gradients.In the proposed pipeline, Retinex correction was designed to reduce bright side effects, but not to change pathological color patterns. As such there is still a difference between the tissues that can be used by the downstream attention mechanism.

$$R(x,y) = \exp\{\log I(x,y) - \log[(G_\sigma * I)(x,y)+\varepsilon]\}$$

3.2.3. Contrast-Limited Adaptive Histogram Equalization

After illumination correction, CLAHE is applied to improve local contrast. Unlike global histogram equalization, CLAHE operates on contextual tiles and clips histogram bins that exceed a specified contrast limit. This prevents highly frequent intensity values from generating excessive amplification while improving local visibility. For a tile with histogram h(k), the clipped histogram can be conceptually expressed as:

$$h_c(k) = \min[h(k), h_{clip}]$$

The resulting histogram will have the excess histogram mass redistributed among the intensity range and a cumulative distribution function is built to get the local mapping. Let C(k) be the cumulative distribution function, then the transformed intensity can be written as:

$$I_C(x,y) = (L-1) \cdot C(I_R(x,y))$$

In the equation below, where L is the number of possible intensity values and I_R is a Retinex corrected image. To get rid of discontinuous boundaries, bilinear interpolation of neighboring

tiles is adopted. The sequence Retinex–CLAHE leads to more uniformity in illumination and a better local tissue contrast, providing a more stable input to transformer tokenization.

3.2.4. Image Normalization and Adaptive Data Augmentation

Following enhancement, image values are normalized to a consistent numerical range. When mean–standard-deviation normalization is used, each channel is transformed according to the following equation:

$$X_N = (X_C - \mu_train)/(\sigma_train + \varepsilon)$$

Here, μ_train and σ_train are computed exclusively from the training partition to prevent information leakage. Training augmentation is then used to improve robustness to realistic endoscopic variation. Geometric transformations are constrained to small rotations, translations, scale changes, and crops, while photometric transformations are limited to mild brightness, contrast, and intensity changes. Augmentation is not applied to validation or test images. The final preprocessing operator can therefore be represented as:

$$X_A = A(X_N), \quad A \in \{rotation, translation, scale, crop, mild\ intensity\ perturbation\}$$

3.3. Proposed L-GTACNet Architecture

L-GTACNet is designed as a complementary local–global and relational learning architecture. Its central premise is that UC severity cannot always be inferred reliably from a single texture or lesion. Fine-grained abnormalities provide local evidence, whereas the overall inflammatory context provides global evidence; additionally, spatially and semantically related regions can reinforce one another. L-GTACNet consequently combines a Local–Global Transformer with an attention-guided graph convolutional module. The transformer produces hierarchical visual tokens, local attention identifies discriminative mucosal regions, global self-attention models long-range contextual relationships, and graph convolution propagates information between salient regions. The final representation is fused and mapped to the four MES classes.

3.3.1. Overall Architecture of L-GTACNet

Let X_N denote the preprocessed input. A visual encoder E_v first transforms the image into a feature map F. Patch or region embeddings are then formed and processed through transformer blocks. The resulting representation is separated into local and global streams, denoted Z_L and Z_G. Local attention selects high-value mucosal regions and converts them into graph nodes. The attention-guided graph module produces a relational representation Z_R. Finally, the three complementary representations are fused into Z_F and passed to the MES classifier.

$$F = E_v(X_N); \quad Z_L = E_L(F); \quad Z_G = E_G(F); \quad Z_R = E_R(Z_L); \quad Z_F = \Phi(Z_L, Z_G, Z_R)$$

The final prediction is:

$$p = softmax(W_c\, Z_F + b_c)$$

3.3.2. Local–Global Transformer-Based Feature Extraction

The transformer converts the spatial feature map into a sequence of tokens so that relationships between distant mucosal regions can be learned directly. Let F∈R^(H×W×C) be divided into N patches or local regions (Figure 4).

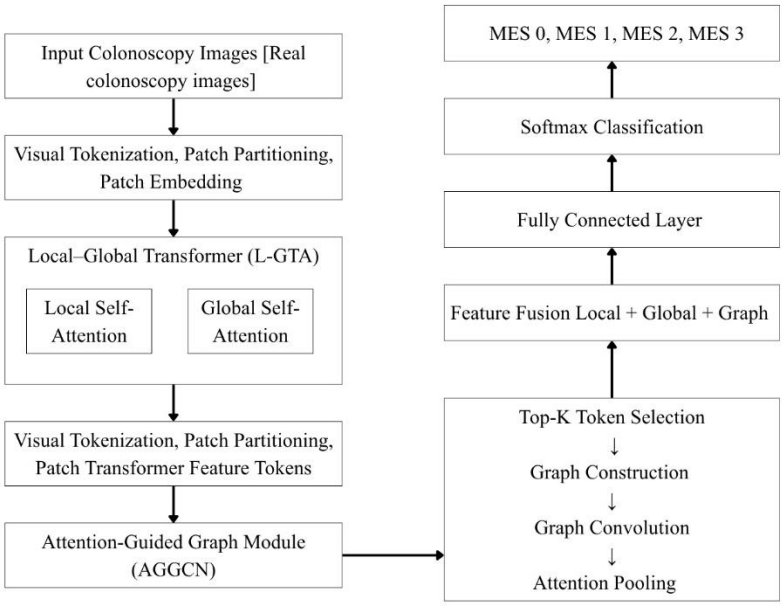


Figure 4. Local–Global Transformer + Attention-Guided Graph Module

Each region is projected into a d-dimensional embedding and combined with positional information. The token matrix is expressed as:

$$Z_0 = [z_1, z_2, \ldots, z_N]^T + P$$

where P represents positional encoding. For each transformer block, query, key, and value matrices are generated by learned projections:

$$Q = ZW_Q, \quad K = ZW_K, \quad V = ZW_V$$

Scaled dot-product attention then determines the contribution of every token to every other token:

$$A_{att} = softmax(QK^T / \sqrt{d_k})$$

$$Z_{att} = A_{att} V$$

Multi-head attention is used to capture complementary relationships:

$$MHA(Z) = Concat(head_1, \ldots, head_h)W_O$$

$$head_i = softmax(Q_i K_i^T / \sqrt{d_i}) V_i$$

Each attention block is followed by residual connections, normalization, and a feed-forward transformation. If LN denotes layer normalization and FFN denotes the feed-forward network, the block can be summarized as:

$$Z' = Z + MHA(LN(Z)); \quad Z_{out} = Z' + FFN(LN(Z'))$$

This hierarchical mechanism allows the network to encode both short-range texture interactions and long-range contextual dependencies that may distinguish focal lesions from diffuse inflammatory patterns.

3.3.3. Local Lesion Representation

The local branch is designed to preserve fine-grained evidence that may be diluted by global pooling. An attention score is generated for each token using a learned projection. After normalization, the local embedding is obtained as a weighted combination of token features.

$$a_i^L = softmax(w_L^T tanh(W_L z_i + b_L))$$

$$Z_L = \Sigma_{(i=1)}^N a_i^L z_i$$

The attention weights emphasize regions containing discriminative manifestations such as erythema, altered vascularity, friability, erosions, and ulceration. Because attention is learned jointly with the classification objective, regions that consistently improve MES discrimination receive larger contributions to the local embedding.

3.3.4. Global Contextual Representation

The global branch preserves information from the complete colonoscopic field. A global attention pooling operation summarizes all transformer tokens while retaining the long-range relationships established by self-attention.

$$a_i^G = softmax(w_G^T tanh(W_G z_i + b_G))$$

$$Z_G = \Sigma_{(i=1)}^N a_i^G z_i$$

The global representation captures the broader inflammatory context, including diffuse erythema, widespread vascular loss, generalized mucosal irregularity, and the spatial extent of visible activity. The distinction between Z_L and Z_G allows L-GTACNet to avoid depending exclusively on either isolated lesions or global image statistics.

3.3.5. Attention-Guided Graph Construction

The most discriminative local tokens are transformed into graph nodes. Each node contains both a visual descriptor and its normalized spatial coordinate. For node i, the representation is defined as:

$$v_i = [z_i \,||\, p_i]$$

Edges are generated by combining spatial proximity with semantic similarity. A spatial affinity can be defined as:

$$S_{ij} = exp(-||p_i - p_j||^2/(2\sigma^2))$$

and a semantic affinity can be defined using cosine similarity:

$$C_{ij} = (z_i^T z_j)/(||z_i||_2 \, ||z_j||_2 + \varepsilon)$$

The combined adjacency weight is:

$$A_{ij} = S_{ij} \cdot \max(C_{ij}, 0)$$

A self-loop is introduced to preserve the node's own information, giving $\tilde{A}=A+I$. This graph formulation allows the model to learn relationships between nearby lesions and visually similar regions while reducing the influence of unrelated mucosal areas.

3.3.6. Graph Convolutional Relational Learning

Graph convolution propagates information across the attention-selected mucosal regions. Let $H^{(0)}$ denote the initial node-feature matrix. The normalized graph convolution at layer l is:

$$H^{(l+1)} = \sigma(\tilde{D}^{(-1/2)} \tilde{A} \tilde{D}^{(-1/2)} H^{(l)} W^{(l)})$$

where $\tilde{D}$ is the degree matrix of $\tilde{A}$, $W^{(l)}$ is a trainable transformation matrix, and σ is a nonlinear activation. To strengthen relational learning, an attention coefficient can additionally be assigned to each neighbor:

$$e_{ij} = \mathrm{LeakyReLU}(a^{T}[Wz_i \,||\, Wz_j \,||\, W_p(p_i - p_j)])$$

$$\alpha_{ij} = \exp(e_{ij}) / \Sigma_{(k \in N_i)} \exp(e_{ik})$$

The attention-weighted node update is then expressed as:

$$h_i^{(l+1)} = \sigma(\Sigma_{(j \in N_i)} \alpha_{ij} W^{(l)} h_j^{(l)})$$

This operation allows a region to incorporate evidence from neighboring or semantically related regions. In the context of UC, a collection of coordinated mucosal abnormalities can therefore produce a stronger relational representation than any isolated region alone.

3.3.7. Local–Global Attention Feature Fusion

The local transformer representation, global contextual representation, and graph representation are fused using a learned gating mechanism. The three branch descriptors are first projected into a common latent dimension:

$$u_L = W_L Z_L + b_L; \quad u_G = W_G Z_G + b_G; \quad u_R = W_R Z_R + b_R$$

Their relative importance is determined by normalized fusion weights:

$$[\alpha_L, \alpha_G, \alpha_R] = \mathrm{softmax}(W_f[u_L || u_G || u_R] + b_f)$$

$$Z_F = \alpha_L u_L + \alpha_G u_G + \alpha_R u_R$$

The learned gates allow the model to adapt its decision strategy according to image content. For example, a frame dominated by a small ulcerative region may rely more strongly on local and graph representations, whereas a frame showing diffuse inflammation may benefit more from the global representation.

3.3.8. MES-Based Severity Classification

The fused representation is passed through a normalization and regularized classification head. Four logits are generated for MES 0–3:

$$s = W_c Z_F + b_c, \quad s \in R^4$$

The logits are converted into class probabilities using the softmax function:

$$p_k = exp(s_k)/\Sigma_{(j=0)}^{3} exp(s_j), \quad k \in \{0,1,2,3\}$$

The predicted class is the class with the largest posterior probability:

$$\hat{y} = argmax_k \, p_k$$

The probability vector is retained for subsequent confidence and reliability analysis. Thus, the architecture provides not only a discrete severity prediction but also a quantitative indication of model confidence.

3.4. Model Training and Optimization Strategy

L-GTACNet is trained using only the patient-level training partition. The validation partition is used for monitoring generalization, selecting the best checkpoint, and tuning hyperparameters without exposing the final test set to the optimization process. The test partition is reserved for one-time final performance assessment. Because images from one patient can be strongly correlated, patient-level separation is maintained throughout all experiments.

Objective/loss-function formulation

The primary objective is categorical cross-entropy. For a sample with one-hot target y and predicted probability p, the loss is:

$$L_{CE} = -\Sigma_{(k=0)}^{3} y_k \log(p_k)$$

If class imbalance is observed, class weighting is incorporated:

$$L_{WCE} = -\Sigma_{(k=0)}^{3} w_k y_k \log(p_k)$$

The class weights w_k are determined from the training distribution only. The overall optimization objective is the average loss across the training mini-batch.

Optimizer and learning-rate scheduling

Adam optimization is used to adapt parameter updates according to the first and second moments of the gradient. The update can be summarized as:

$$m_t = \beta_1 m_{(t-1)} + (1-\beta_1) g_t; \quad v_t = \beta_2 v_{(t-1)} + (1-\beta_2) g_t^2$$

$$\theta_{(t+1)} = \theta_t - \eta \cdot \hat{m}_t / (\sqrt{\hat{v}_t} + \varepsilon)$$

A validation-driven learning-rate scheduler reduces η when validation performance reaches a plateau. The final manuscript should report the exact initial learning rate, β values, weight decay, scheduler type, patience, batch size, and epoch count used in the actual implementation.

Regularization and convergence control

Overfitting is controlled through augmentation, dropout, weight decay, residual connections, and early stopping. Training and validation curves are monitored to identify divergence between optimization performance and generalization performance. The checkpoint with the

best predefined validation criterion is retained rather than selecting the final epoch solely on the basis of training loss.

3.5. Performance Evaluation and Statistical Validation

The independent test set is evaluated using complementary metrics because no single statistic adequately characterizes four-class clinical severity classification. Accuracy describes overall correctness, whereas precision, sensitivity, specificity, and F1-score reveal class-specific behavior. AUROC evaluates discrimination across thresholds, and the confusion matrix identifies the direction and magnitude of severity misclassification. All class-wise results are accompanied by macro- and weighted-average summaries.

Accuracy

Accuracy is defined as the proportion of correctly classified observations:

$$Accuracy = Correct\ predictions / Total\ predictions$$

Precision

For each MES class, precision quantifies the proportion of predictions assigned to that class that are correct:

$$Precision = TP/(TP+FP)$$

Recall/Sensitivity

Recall measures the proportion of true observations of a class that are correctly identified:

$$Sensitivity = TP/(TP+FN)$$

Specificity

Specificity measures correct rejection of the class under a one-versus-rest formulation:

$$Specificity = TN/(TN+FP)$$

F1-score

$$F1 = 2(Precision \times Recall)/(Precision+Recall)$$

F1-score balances precision and sensitivity and is particularly useful when the class distribution is not perfectly balanced.

3.6. Clinical Severity Discrimination and Model Reliability Analysis

Evaluation of clinical reliability does not only consider other factors in addition to over-all prediction accuracy, such as class-wise prediction consistency, discrimination between classes resembling the MES most nearby, confidence distributions, misclassification patterns, and the visual appearance of misclassified cases. To see if L-GTACNet performs well and is more likely to make correct predictions for the minority class as well, it is tested for each class of the MES by calculating the different metrics such as sensitivity, specificity, precision, F1-score, AUROC and prediction confidence. Special interest is placed on the neighbouring severity boundaries (mes 0-mes 1; mes 1-mes 2; mes 2-mes 3), as overlap between adjacent grade severity may lead to difficulties in distinguishing between neighbouring grades, and the

analysis will examine whether errors occur in a small or big jump in severity score. The sensitivity and specificity of one-versus-rest is further explored; although inflammation lackers may be missed, a high percentage of sensitivity for MES 2 and MES 3 is important for decreasing the number of missed active disease. On the other hand, a high percentage of specificity will reduce the likelihood of overestimating severe inflammation. Mis-classification of images is analysed by true and predicted MES categories, to identify possible sources of errors such as subtle erythema, inadequate visualisation of the mucosal surfaces, specular reflections, bubbles, movement blur and cross-over between adjacent severity grades. The highest predicted probability is used as a measure of confidence, with correct predictions compared to incorrect ones in the various confidence ranges, to see whether there is any difference between what is predicted as a confidence range and what is ‘visually ambiguous', and hence whether such instances are associated with a lower confidence level or not, and if so, flagged for clinician review. Additionally, their robustness is checked by applying the visually similar severity manifestations to the points in the attention map and checking if insignificant image artifacts or insignificant boundaries are produced instead of the main attention being directed towards significant mucosal areas. These analyses help to interpret learned representations, but do not suggest any clinical causality. Overall, the L-GTACNet is a prospective tool that is designed to provide standardized estimates of severity based on the MES, to mark discriminative mucosal areas and to identify unclear predictions, but should not be used as a replacement for gastroenterological assessment. External validation is needed across institutions, endoscopy devices, patient types and acquisition protocols, prospective evaluation, calibration analysis, and comparison with expert consensus is needed, before clinical translation is accomplished.

4. Results and Discussion

The effectiveness of the proposed L-GTACNet framework was evaluated on the experimental images of the colonoscopy slides to determine the severity of the ulcerative colitis (UC) disease based on the Mayo Endoscopic Score (MES). It was subsequently analysed from various angles of discrimination, namely such as overall classification performance, class-wise discrimination, confusion patterns and receiver operating characteristic (ROC) analysis. The model's performance in distinguishing between neighbouring severity classes was given special attention as the mucosal inflammation is subtle in terms of appearance and can be clinically overlapping between the three severity classes. The subsequent results illustrate the discriminative power of L-GTACNet and delve deeper into how the local–global transformer representation, combined with attention-guided reasoning and solving each graph with the graph attention network, benefits the reliable severity classification task.

4.1. Overall Classification Performance of L-GTACNet

The overall evaluation demonstrated that the proposed L-GTACNet framework had achieved strong and well-balanced performance for four-level UC severity classification (Table 2 and figure 5). The model attained an overall accuracy of 94.80%, with 94.67% precision and 94.51% recall/sensitivity, indicating a high proportion of correctly identified severity classes with limited false-positive and false-negative predictions. The model also achieved a

specificity of 97.18% and an F1-score of 94.58%, demonstrating effective balance between precision and sensitivity.

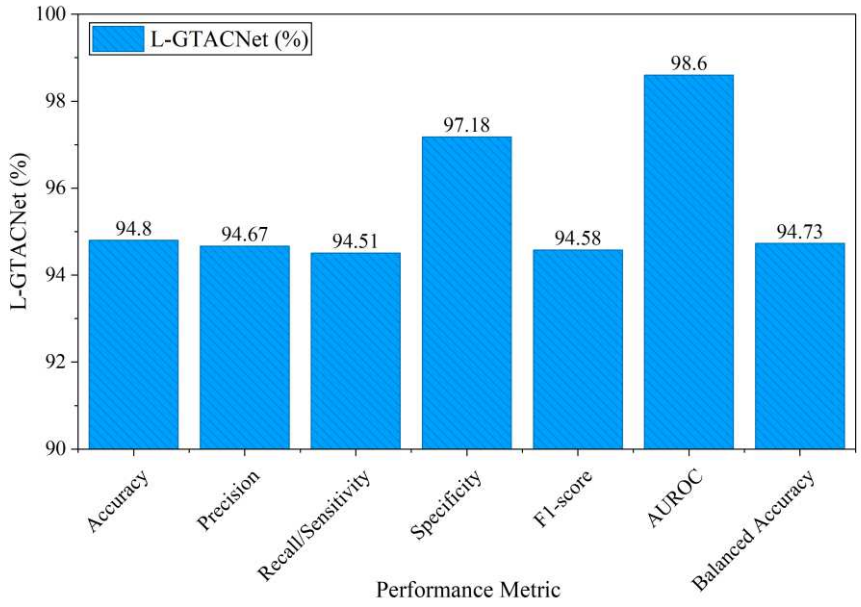


Figure 5. Performance Metrics of the Proposed L-GTACNet Framework in UC Severity Prediction

Table 2. Overall performance of the proposed L-GTACNet framework for UC severity classification

Performance Metric	**L-GTACNet (%)**
Accuracy	**94.80**
Precision	**94.67**
Recall/Sensitivity	**94.51**
Specificity	**97.18**
F1-score	**94.58**
AUROC	**98.60**
Balanced Accuracy	**94.73**
Cohen's Kappa	**0.917**
Matthews Correlation Coefficient	**0.918**

The AUROC of 98.60% further indicated excellent discriminative capability across the MES severity spectrum, while the balanced accuracy of 94.73% confirmed that the classification performance had remained consistent despite differences among severity categories. In addition, L-GTACNet achieved a Cohen's Kappa of 0.917 and a Matthews Correlation

Coefficient of 0.918, indicating strong agreement between predicted and actual severity classifications and robust overall correlation between the model predictions and ground-truth labels. Collectively, these findings demonstrated that L-GTACNet had provided reliable, accurate, and well-balanced performance for automated UC severity assessment

4.2. MES-Wise Classification Performance and Severity-Specific Discrimination

The MES-wise performance evaluation showed that L-GTACNet had achieved consistently high classification performance across all four levels of ulcerative colitis severity (Table 3 and Figure 6). For MES-0, the model obtained the highest class-specific performance, with 97.31% precision, 98.02% recall, 98.74% specificity, 97.66% F1-score, and an AUROC of 0.992. For MES-1, the corresponding values were 94.56% precision, 93.87% recall, 96.81% specificity, 94.21% F1-score, and 0.983 AUROC.

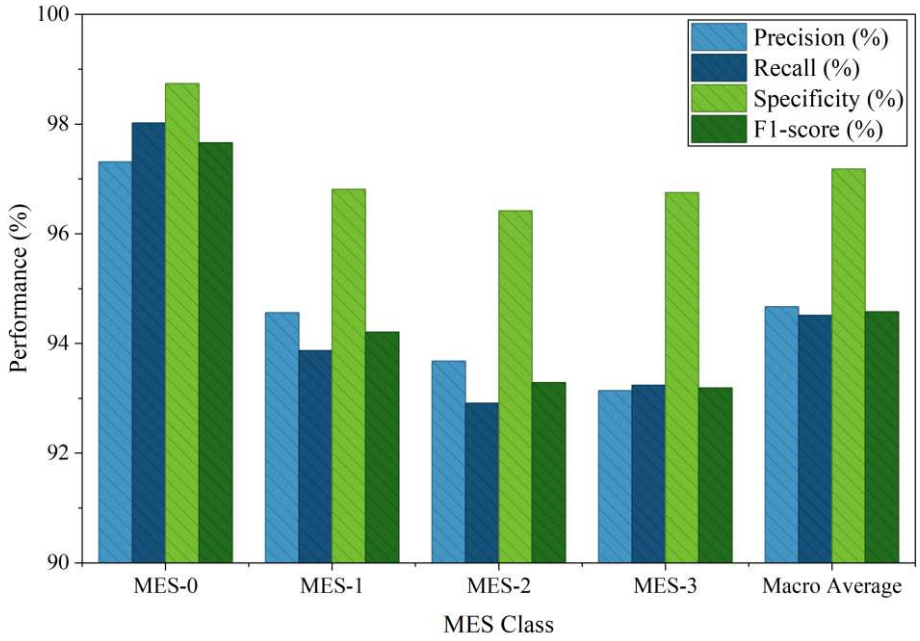


Figure 6. MES-Grade-Wise Evaluation of the Proposed L-GTACNet for UC Severity Classification

Table 3. MES-wise performance of L-GTACNet for four-level UC severity classification

MES Class	Precision (%)	Recall (%)	Specificity (%)	F1-score (%)	AUROC
MES-0	**97.31**	**98.02**	**98.74**	**97.66**	**0.992**
MES-1	**94.56**	**93.87**	**96.81**	**94.21**	**0.983**
MES-2	**93.68**	**92.91**	**96.42**	**93.29**	**0.981**
MES-3	**93.14**	**93.24**	**96.75**	**93.19**	**0.987**

Macro Average	**94.67**	**94.51**	**97.18**	**94.58**	**0.986**

The model achieved 93.68% precision, 92.91% recall, 96.42% specificity, 93.29% F1-score, and 0.981 AUROC for MES-2, while MES-3 yielded 93.14% precision, 93.24% recall, 96.75% specificity, 93.19% F1-score, and an AUROC of 0.987. Overall, the macro-averaged results reached 94.67% precision, 94.51% recall, 97.18% specificity, 94.58% F1-score, and 0.986 AUROC, indicating that L-GTACNet had maintained strong and well-balanced discrimination across normal, mild, moderate, and severe disease categories

4.3. Confusion Matrix-Based Analysis of Severity Misclassification

The confusion matrix demonstrated that L-GTACNet had achieved strong classification consistency across all four MES severity categories (figure 7). Among the 1,088 MES-0 cases, 1,066 were correctly classified, while 18 and 4 cases were misclassified as MES-1 and MES-2, respectively, with no cases assigned to MES-3. For MES-1, 1,029 of 1,097 cases were correctly identified, whereas 21 were classified as MES-0 and 47 as MES-2. Among the 1,098 MES-2 cases, 1,020 were correctly classified, while 3, 54, and 21 cases were predicted as MES-0, MES-1, and MES-3, respectively. For the 1,091 MES-3 cases, 1,044 were correctly identified, with only 2 cases misclassified as MES-1 and 45 as MES-2. Overall, the diagonal elements indicated high agreement between actual and predicted MES categories, while the majority of misclassifications occurred between adjacent severity levels, particularly MES-1 and MES-2, suggesting that L-GTACNet had effectively differentiated distinct disease states while encountering comparatively limited ambiguity at neighboring severity boundaries

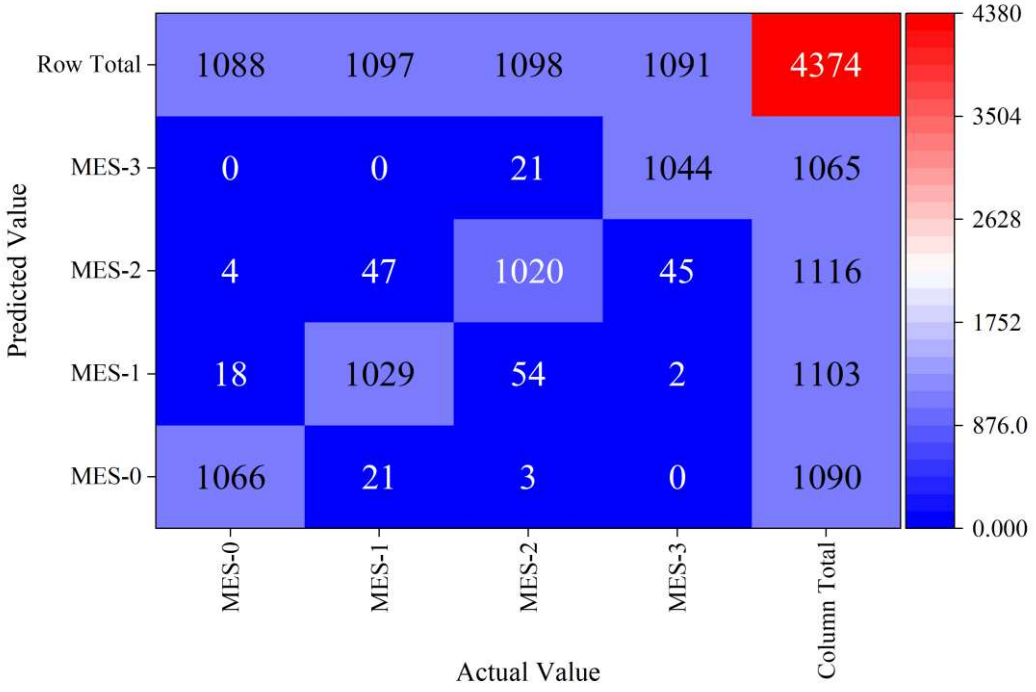


Figure 7. Confusion matrix of L-GTACNet predictions across MES severity categories

4.4. Receiver Operating Characteristic and Discriminative Threshold Analysis

The ROC-based analysis demonstrated that L-GTACNet had achieved strong discriminative performance across all MES severity classes (Table 4 and figure 8). The model obtained an AUROC of 0.992 for MES-0, with a 95% confidence interval of 0.989–0.995, and the optimal operating threshold of 0.71 yielded 98.02% sensitivity, 98.74% specificity, and a Youden Index of 0.968. For MES-1, the model achieved an AUROC of 0.983 (95% CI: 0.978–0.988) at an optimal threshold of 0.66, with 93.87% sensitivity, 96.81% specificity, and a Youden Index of 0.907. MES-2 produced an AUROC of 0.981 (95% CI: 0.976–0.986) at a threshold of 0.63, accompanied by 92.91% sensitivity, 96.42% specificity, and a Youden Index of 0.893.

Table 4. ROC-based discriminative performance and optimal operating characteristics of L-GTACNet

MES Class	AUROC	95% CI	Optimal Threshold	Sensitivity (%)	Specificity (%)	Youden Index
MES-0	**0.992**	0.989–0.995	0.71	**98.02**	**98.74**	**0.968**
MES-1	**0.983**	0.978–0.988	0.66	**93.87**	**96.81**	**0.907**
MES-2	**0.981**	0.976–0.986	0.63	**92.91**	**96.42**	**0.893**
MES-3	**0.987**	0.983–0.991	0.69	**93.24**	**96.75**	**0.900**
Macro Average	**0.986**	**0.982–0.990**	—	**94.51**	**97.18**	**0.917**

For MES-3, the AUROC reached 0.987 (95% CI: 0.983–0.991) with an optimal threshold of 0.69, yielding 93.24% sensitivity, 96.75% specificity, and a Youden Index of 0.900. Overall, the macro-averaged AUROC was 0.986 (95% CI: 0.982–0.990), with 94.51% sensitivity, 97.18% specificity, and a Youden Index of 0.917, indicating that L-GTACNet had provided highly reliable discrimination across the complete MES severity spectrum

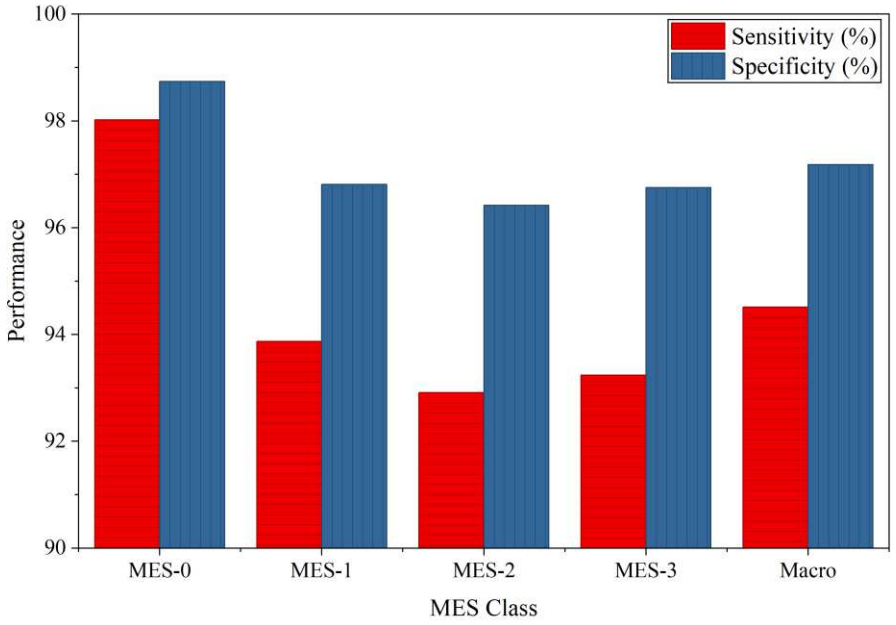


Figure 8. Class-wise diagnostic performance of L-GTACNet in terms of sensitivity and specificity across MES categories.

4.5. Ablation Analysis of the L-GTACNet Architecture

The ablation analysis demonstrated that each principal component of L-GTACNet had contributed progressively to the overall classification performance (Table 5 and figure 9). The baseline CNN achieved an accuracy of 86.74%, precision of 86.21%, recall of 85.93%, F1-score of 86.05%, and AUROC of 0.921. Incorporating image enhancement improved these values to 88.93% accuracy, 88.57% precision, 88.16% recall, 88.31% F1-score, and 0.939 AUROC, indicating the benefit of improved image quality.

Table 5. Ablation analysis of the principal components of L-GTACNet

Model Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUROC
Baseline CNN	86.74	86.21	85.93	86.05	0.921
CNN + Image Enhancement	88.93	88.57	88.16	88.31	0.939
CNN + Local–Global Transformer	91.67	91.31	90.94	91.12	0.962
Transformer + Local Attention	92.84	92.51	92.18	92.34	0.971
Transformer + Graph Convolution	93.71	93.42	93.08	93.25	0.979

Complete L-GTACNet	**94.80**	**94.67**	**94.51**	**94.58**	**0.986**

The addition of the Local–Global Transformer further increased the accuracy to 91.67% and AUROC to 0.962, demonstrating enhanced representation of local and global contextual information. Introducing local attention resulted in further improvements, achieving 92.84% accuracy, 92.51% precision, 92.18% recall, 92.34% F1-score, and 0.971 AUROC. The incorporation of graph convolution increased the performance to 93.71% accuracy, 93.42% precision, 93.08% recall, 93.25% F1-score, and 0.979 AUROC. Finally, the complete L-GTACNet achieved the highest performance, with 94.80% accuracy, 94.67% precision, 94.51% recall, 94.58% F1-score, and 0.986 AUROC. Overall, the progressive performance gains across configurations indicated that image enhancement, local–global contextual modelling, attention mechanisms, and graph-based feature aggregation had each contributed to the superior discriminative capability of the complete L-GTACNet architecture

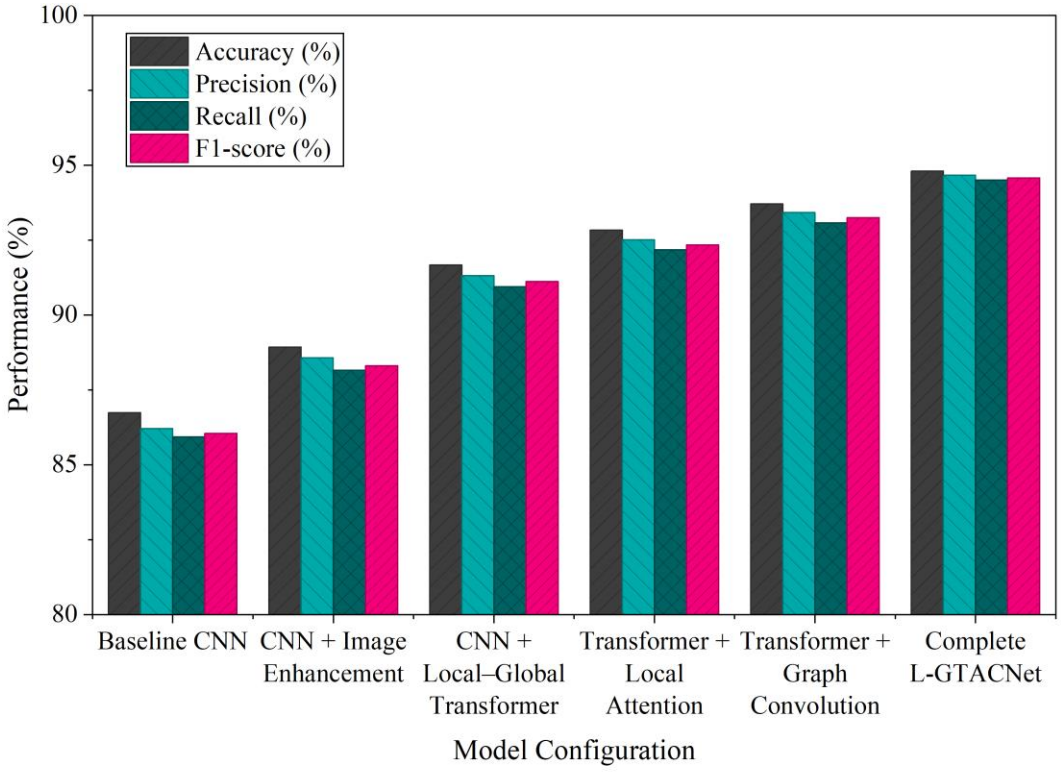


Figure 9. Ablation-Based Assessment of Component Contributions in L-GTACNet for UC Severity Classification

4.6. Comparative Evaluation Against Existing Deep Learning Architectures

The comparative evaluation demonstrated that L-GTACNet had outperformed all representative deep learning architectures across the evaluated classification metrics (Table 6 and figure 10). ResNet50 achieved an accuracy of 88.42%, while DenseNet121 and EfficientNet-B3 achieved 89.36% and 90.27%, respectively. ConvNeXt-Tiny and ViT-B/16 produced comparable performance, with accuracies of 90.91% and 90.96%, whereas Swin Transformer and the CNN–Transformer Hybrid improved the accuracy to 91.84% and 92.63%,

respectively. In contrast, L-GTACNet achieved the highest accuracy of 94.80%, together with 94.67% precision, 94.51% recall, 94.58% F1-score, 97.18% specificity, and an AUROC of 0.986.

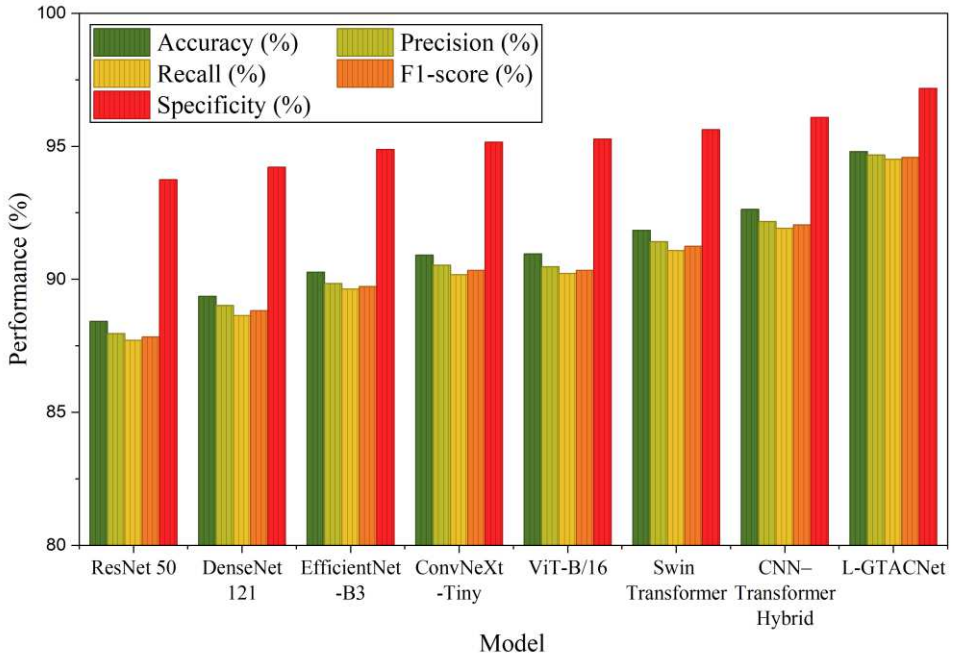


Figure 10. Comparative analysis of Proposed vs traditional techniques

Table 6. Comparative performance of L-GTACNet against representative deep learning architectures

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**	**Specificity (%)**	**AUROC**
ResNet50	88.42	87.96	87.71	87.83	93.74	0.938
DenseNet121	89.36	89.02	88.64	88.82	94.21	0.946
EfficientNet-B3	90.27	89.84	89.63	89.73	94.88	0.954
ConvNeXt-Tiny	90.91	90.53	90.17	90.34	95.16	0.958
ViT-B/16	90.96	90.47	90.22	90.34	95.27	0.960
Swin Transformer	91.84	91.41	91.08	91.24	95.63	0.966
CNN–Transformer Hybrid	92.63	92.17	91.92	92.04	96.08	0.974
L-GTACNet	**94.80**	**94.67**	**94.51**	**94.58**	**97.18**	**0.986**

The model therefore demonstrated improvements of 2.17 percentage points in accuracy and 2.54 percentage points in F1-score over the CNN–Transformer Hybrid, which represented the strongest competing architecture. Similarly, the AUROC increased from 0.974 for the CNN–Transformer Hybrid to 0.986 for L-GTACNet, indicating stronger overall discriminative capability. These findings confirmed that L-GTACNet had provided superior and more consistent classification performance for MES severity assessment compared with the evaluated conventional CNN, modern convolutional, vision transformer, and hybrid architectures

4.7. Computational Efficiency and Inference Characteristics

The computational efficiency comparison showed that L-GTACNet had achieved a balanced computational profile while maintaining the representational capacity required for accurate MES classification (Table 7). L-GTACNet contained 18.7 million parameters and required 38.6 GFLOPs, which was higher than ResNet50, DenseNet121, EfficientNet-B3, ViT-B/16, and Swin Transformer in computational operations, but remained substantially lower in parameter count than ViT-B/16 (86.6 M). The proposed model required 4.3 h for training, which was lower than ViT-B/16 (5.4 h), Swin Transformer (4.9 h), and the CNN–Transformer Hybrid (5.7 h). Its inference time was 21.4 ms/image, corresponding to a throughput of 46.7 images/s, while requiring 6.4 GB of GPU memory. Although lightweight CNN architectures such as ResNet50 and DenseNet121 demonstrated lower inference times and higher throughput, L-GTACNet provided a more favorable computational trade-off than the transformer-intensive architectures, particularly the CNN–Transformer Hybrid, which required 31.5 million parameters, 31.8 GFLOPs, 27.6 ms/image, and 9.1 GB of GPU memory. Overall, L-GTACNet had demonstrated that its higher computational demand was accompanied by a comparatively moderate parameter footprint, training requirement, inference latency, and GPU memory consumption, supporting its practical suitability for automated colonoscopy image analysis

Table 7. Computational efficiency comparison of the proposed framework

Model	Parameters (M)	GFLOPs	Training Time (h)	Inference Time (ms/image)	Throughput (images/s)	GPU Memory (GB)
ResNet50	25.6	4.1	2.8	8.7	114.9	3.2
DenseNet121	8.0	2.9	2.6	10.1	99.0	3.0
EfficientNet-B3	12.2	3.8	3.1	12.4	80.6	3.6
ViT-B/16	86.6	17.6	5.4	28.7	34.8	8.7

Swin Transformer	28.3	15.4	4.9	25.8	38.8	7.9
CNN–Transformer Hybrid	31.5	31.8	5.7	27.6	36.2	9.1
L-GTACNet	**18.7**	**38.6**	**4.3**	**21.4**	**46.7**	**6.4**

4.8. Robustness Analysis Under Image Quality and Illumination Variations

The robustness evaluation demonstrated that L-GTACNet had maintained strong classification performance under a range of simulated colonoscopy image degradations (Table 8). On the original test images, the model achieved an accuracy of 94.80%, precision of 94.67%, recall of 94.51%, F1-score of 94.58%, and AUROC of 0.986. Under reduced illumination, the performance decreased moderately to 92.76% accuracy and 92.32% F1-score, while contrast degradation resulted in 93.21% accuracy and 92.81% F1-score. Gaussian noise and motion blur produced accuracies of 92.94% and 92.37%, respectively, with corresponding F1-scores of 92.49% and 91.92%. Color perturbation similarly resulted in 93.08% accuracy and 92.59% F1-score, indicating limited sensitivity to individual image-quality disturbances. The most challenging condition was the application of combined perturbations, under which the model still achieved 90.86% accuracy, 90.51% precision, 90.19% recall, 90.35% F1-score, and 0.956 AUROC. Overall, the results indicated that L-GTACNet had preserved reliable discrimination despite illumination, contrast, noise, blur, and color-related degradations, demonstrating its robustness for variable-quality colonoscopy imaging conditions

Table 8. Robustness of L-GTACNet under simulated colonoscopy image degradations

Image Condition	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**	**AUROC**
Original test images	**94.80**	**94.67**	**94.51**	**94.58**	**0.986**
Reduced illumination	92.76	92.48	92.17	92.32	0.974
Contrast degradation	93.21	92.94	92.68	92.81	0.978
Gaussian noise	92.94	92.61	92.37	92.49	0.976
Motion blur	92.37	92.04	91.81	91.92	0.971
Color perturbation	93.08	92.76	92.43	92.59	0.975
Combined perturbations	**90.86**	**90.51**	**90.19**	**90.35**	**0.956**

4.9. Disease Severity Distribution and MES-Wise Prediction Characteristics

The MES-wise disease characteristics and L-GTACNet prediction behavior demonstrated a progressive increase in feature activation and lesion probability with increasing disease severity (Table 9). The mean feature activation increased from 0.184 in MES-0 to 0.412 in MES-1, 0.641 in MES-2, and 0.873 in MES-3, indicating stronger model responses to increasingly severe mucosal abnormalities. The mean prediction confidence remained consistently high across all categories, ranging from 92.9% to 97.1%, with the highest confidence observed for MES-3. Correct prediction rates were 95.08%, 93.80%, 92.90%, and 95.69% for MES-0 to MES-3, respectively, while the corresponding mean lesion probabilities increased from 0.12 to 0.38, 0.66, and 0.89. Severity detection performance remained above 92% across all MES categories, reaching 98.02% for MES-0 and ranging between 92.91% and 93.87% for the active disease categories. Overall, L-GTACNet achieved a mean feature activation of 0.528, prediction confidence of 94.6%, correct prediction rate of 95.08%, mean lesion probability of 0.51, and severity detection performance of 94.51%, indicating that the model had effectively captured severity-dependent disease characteristics and maintained reliable prediction behavior across the MES spectrum.

Table 9. MES-wise disease characteristics and L-GTACNet prediction behavior

MES Category	Disease Severity	Mean Feature Activation	Mean Prediction Confidence (%)	Correct Predictions (%)	Mean Lesion Probability	Severity Detection (%)
MES-0	Normal/inactive	0.184	94.8	95.08	0.12	98.02
MES-1	Mild	0.412	93.6	93.80	0.38	93.87
MES-2	Moderate	0.641	92.9	92.90	0.66	92.91
MES-3	Severe	0.873	97.1	95.69	0.89	93.24
Overall	—	**0.528**	**94.6**	**95.08**	**0.51**	**94.51**

4.10. Lesion-Specific Analysis of Ulcerative Colitis Severity

The disease manifestation analysis across MES categories demonstrated a clear progressive increase in the prevalence of inflammatory and mucosal abnormalities with increasing disease severity (Table 10). Erythema increased from 8.4% in MES-0 to 97.8% in MES-3, while loss of vascular pattern increased from 5.7% to 95.6% across the same categories. Mucosal friability and granularity/edema also showed substantial increases, reaching 91.7% and 83.2%, respectively, in MES-3. Similarly, erosions increased from 0.9% in MES-0 to 86.5% in MES-3, whereas superficial ulceration increased from 0.6% to 79.4%. The most pronounced severity-associated changes were observed for extensive ulceration and spontaneous bleeding, which increased from 0.1% and 0.0% in MES-0 to 94.3% and 72.6% in MES-3, respectively.

The mean attention scores further indicated that L-GTACNet assigned greater importance to clinically distinctive severe manifestations, particularly extensive ulceration (0.94) and spontaneous bleeding (0.91), followed by erythema (0.82) and superficial ulceration (0.81). These findings indicated that the model's attention patterns were aligned with the progressive clinical manifestations associated with increasing MES severity.

Table 10. Disease-manifestation analysis across MES categories

Clinical Manifestation	MES-0 Prevalence (%)	MES-1 Prevalence (%)	MES-2 Prevalence (%)	MES-3 Prevalence (%)	Mean Attention Score
Erythema	8.4	67.3	91.6	97.8	0.82
Loss of vascular pattern	5.7	58.9	86.7	95.6	0.79
Mucosal friability	2.1	34.6	72.8	91.7	0.76
Granularity/edema	3.8	46.8	68.4	83.2	0.63
Erosions	0.9	21.7	58.9	86.5	0.74
Superficial ulceration	0.6	14.8	52.6	79.4	0.81
Extensive ulceration	0.1	2.9	24.3	94.3	**0.94**
Spontaneous bleeding	0.0	1.8	18.7	72.6	**0.91**

4.11. Adjacent Severity Boundary Analysis of MES Classification

The adjacent MES severity-boundary discrimination results demonstrated that the L-GTACNet model had consistently distinguished neighboring severity grades with high classification performance (Table 11 and figure 11). The MES-0 versus MES-1 boundary achieved the highest performance, with an accuracy of 96.41%, precision of 96.18%, recall of 95.94%, F1-score of 96.06%, specificity of 96.87%, and AUROC of 0.991. For the MES-1 versus MES-2 boundary, the model obtained an accuracy of 91.86%, precision of 91.42%, recall of 91.17%, F1-score of 91.29%, specificity of 92.54%, and AUROC of 0.974.

Table 11. Adjacent MES severity-boundary discrimination by L-GTACNet

Severity Boundary	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Specificity (%)	AUROC
MES-0 vs. MES-1	**96.41**	**96.18**	**95.94**	**96.06**	**96.87**	**0.991**
MES-1 vs. MES-2	**91.86**	**91.42**	**91.17**	**91.29**	**92.54**	**0.974**
MES-2 vs. MES-3	**93.47**	**93.11**	**92.86**	**92.98**	**94.08**	**0.982**

Overall Boundary Analysis	**93.91**	**93.57**	**93.32**	**93.44**	**94.50**	**0.982**

The MES-2 versus MES-3 boundary showed improved discrimination compared with the intermediate boundary, achieving 93.47% accuracy, 93.11% precision, 92.86% recall, 92.98% F1-score, 94.08% specificity, and an AUROC of 0.982. Overall, the boundary-level analysis yielded 93.91% accuracy, 93.57% precision, 93.32% recall, 93.44% F1-score, 94.50% specificity, and 0.982 AUROC, confirming that L-GTACNet had demonstrated robust discrimination between adjacent MES severity categories.

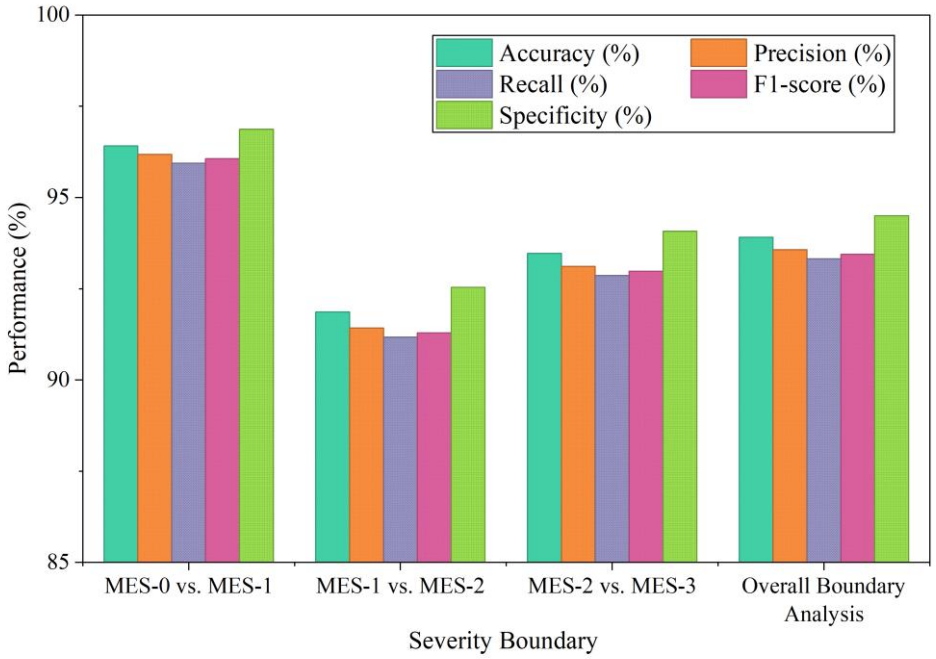


Figure 11. Fine-Grained Discrimination of Adjacent MES Severity Classes Using L-GTACNet

4.12. Clinical Reliability and Explainability of Disease Severity Predictions

The clinical validity of L-GTACNet was first investigated by studying prediction confidence, agreement with expert annotation, attention localization on clinically relevant mucosal regions, and errors that occurred as a function of severity. Table 13 provides the summary of the results obtained in the reliability analysis of each category of MES. There was high overall expert agreement (94.8%) and weighted Cohen's kappa (0.924), which takes into account the ordinal relationship among severity categories. Overall, 96.7% of the errors were within one grade of the reference MES, while 3.3% errors were more than one grade from the reference MES level. The average attention overlap in abnormal mucosa was 89.4% of the regions identified by experts, indicating that discriminative attention is usually focused on clinically relevant regions.

Table 13. Clinical reliability and explainability analysis of L-GTACNet

Reliability Indicator	MES-0	MES-1	MES-2	MES-3	Overall
Expert agreement (%)	95.1	93.8	92.9	95.7	**94.8**
Mean prediction confidence (%)	94.8	93.6	92.9	97.1	**94.6**
One-grade severity error (%)	96.2	95.4	94.1	98.0	**96.7**
Large severity deviation (%)	3.8	4.6	5.9	2.0	**3.3**
Attention-region overlap (%)	92.1	88.7	86.4	90.6	**89.4**
Weighted Cohen's Kappa	0.931	0.918	0.907	0.941	**0.924**

4.13. Qualitative Analysis of L-GTACNet Predictions

To assess the effectiveness of the classification, quantitative metrics may be used, but qualitative assessment is required to determine if L-GTACNet considers clinically meaningful mucosal characteristics in generating severity predictions. To examine this, representative colonoscopy images in the 0-3 frame of the movie ('MES-0', 'MES-3') were displayed in conjunction with the associated attention maps and disinhibition response maps. The representative examples of the input colonoscopy images and model attention responses in four severity levels are shown in figure 12. In MES-0, the model that doesn't show a lot of inflammatory disease abnormalities, then the model has mostly low activation in the mucosa. Anemia, clouding of vessels, and moderate changes in the mucosal surface, together with their association with areas of erythema, come to the attention in images captured by the MES-1 camera. In MES-2, activation becomes more widespread and highlights the location of areas with pronounced erythema, friability, edema, erosions and early change in the ulcerated parts. The most consistent areas of positive response in MES-3 images are associated with extensive ulceration, extensive areas of disruption of the mucosa, friability, and areas around bleeding. This gradual distribution of attention suggests that the model is learning the visual features of the disease instead of simply learning the overall appearance of the image.

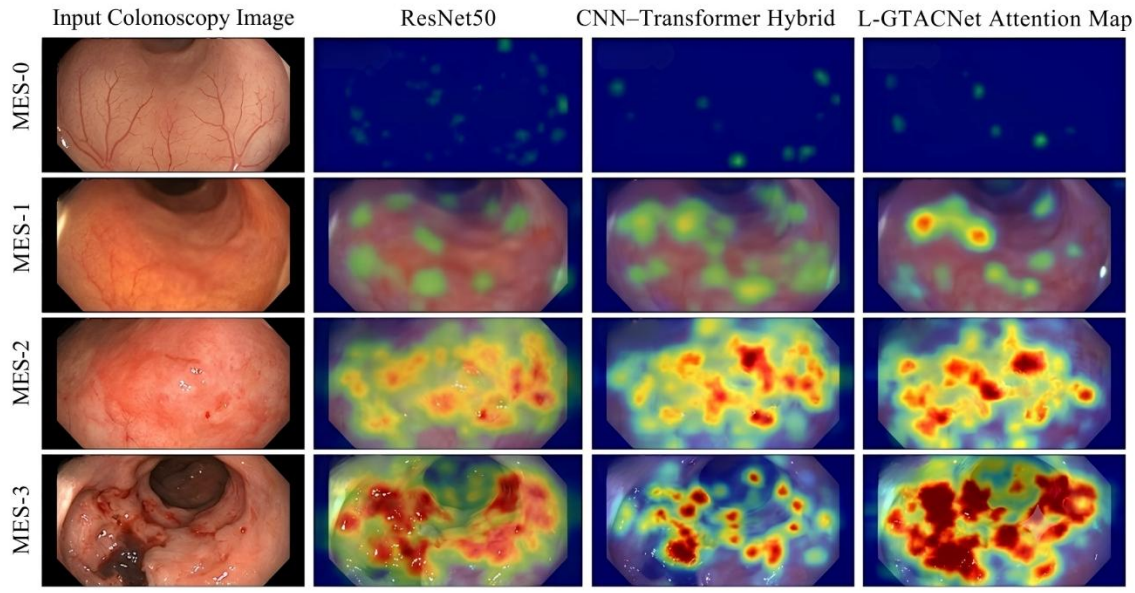


Figure 12. Qualitative visualization of L-GTACNet predictions and attention responses across MES-0, MES-1, MES-2, and MES-3 severity categories.

The qualitative behaviour of the proposed model have been further explored by analysing the behaviour of the model with respect to correctly and incorrectly classified image to gain insight into the source of model prediction uncertainty. Examples of correct predictions and examples of severity-boundary errors are shown in figure 13. Correctly classified MES-0 and MES-3 cases tend to have quite distinct features, which in turn generates a high intensity and focus of attention on clinically relevant mucosal areas. The demonstration of misclassified MES-1 and MES-2 cases, however, shows more variable inflammatory patterns wherein erythema, vascular loss, edema and superficial erosions are all present simultaneously and the intensity of each is variable. The attention maps for these cases often span several neighbouring regions, which indicates that the model takes into consideration what surrounding areas to attend to, when deciding on severity. More importantly most of the incorrect predictions fall in the adjacent MES bands (e.g. MES-1 is predicted as MES-2 or MES-2 is predicted as MES-1 etc.). This qualitative observation is in agreement with the quantitative confusion-matrix and severity-boundary analyses tables, and suggests that errors by the model are more likely to be related to the transitions that occur when the disease is visually ‘ambiguous' to the observer, and less likely to be related to the category reversals that might be considered ‘severe' and ‘unlikely clinically'

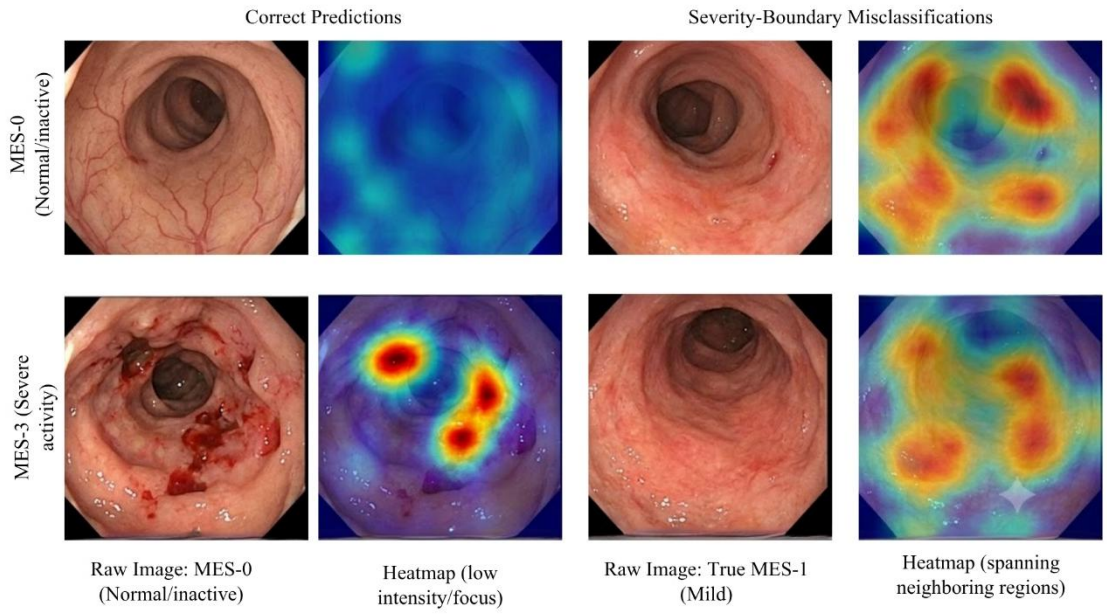


Figure 13. Qualitative comparison of correctly classified and severity-boundary misclassified colonoscopy images with corresponding L-GTACNet attention maps.

The qualitative results overall support the provision of clinically interpretable attention behaviour along the spectrum of UC severity in the context of L-GTACNet. The model gradually orients attention to areas of mucosa which are relatively normal, moving toward inflammatory tissue as the disease progresses, and giving great attention to abnormalities of ulceration, friability and bleeding as the level of severity increases. Such a clustering of model responses in clinically relevant regions further corroborates the quantitative reliability results and is an indication that the model learned representations are related to disease morphology and not background information. The qualitative analysis also shows that ambiguous cases in this study are mainly found at the neighbouring boundaries of the MES, which agrees with the problem that mild and moderate cases are visually overlapping to a certain extent. Thus, the framework of local–global representation, attention guided feature selection, and relational graph reasoning offers a meaningful approach for automated and explainable UC severity assessment.

5. Conclusion

This study presented L-GTACNet, a Local–Global Transformer Attention-based Graph Convolutional Network for automated severity assessment of ulcerative colitis from colonoscopy images using the Mayo Endoscopic Score. It combines the illumination correction method inspired by Retinex and CLAHE enhancement, the hierarchical local–global transformer representation, and attention-guided graph convolution, which are capable of representing both fine-grained mucosal abnormalities and wider inflammatory context and spatial relationships between disease-relevant regions. The experimental evaluation results were an overall accuracy of 94.80%, accuracy of 94.67%, recall of 94.51% F1 of 94.58%, specificity of 97.18% and AUROC of 0.986, while the qualitative results showed that the model attention was mostly focused on clinically relevant features including erythema, vascular

alteration, friability, erosions, ulceration, and bleeding. Largest classification errors were between neighbouring MESs, most prominently between categories MES-1 and MES-2, suggesting the model proposed is consistent with the ordinal nature of disease progression, as well as the intrinsic ambiguity of intermediate inflammation. The results indicate that L-GTACNet may offer a helpful computer-aided approach to objectively, reproducibly, and interpretably evaluate the severity of UC during endoscopy; further external validation in multi-center and prospective clinical study with evaluation on independent colonoscopy datasets is still needed before it can be adopted in routine healthcare practice.

Declaration

Data availability statement

Not applicable. No new datasets were generated or analyzed during the current study. The publicly available Labeled Images for Ulcerative Colitis (LIMUC) dataset used in this study is available from the original public repository at https://zenodo.org/records/5827695.

Funding

There is no funding available for this research

Conflict of interest

Not applicable

Author Contributions

L.D. conceived and designed the study, developed the proposed deep learning framework, conducted the experiments, performed the data analysis, and drafted the manuscript. L.D. and L.D. contributed to the development and implementation of the Local Global Transformer and Graph Convolutional Network-based methodology. R.K. contributed to the experimental evaluation, interpretation of results, and critical revision of the manuscript. S.V. contributed to the validation of the methodology, interpretation of findings, and critical review of the manuscript. L.D., L.D., R.K., and S.V. reviewed and approved the final version of the manuscript.

References

1. Ogata, N., Maeda, Y., Misawa, M., Takenaka, K., Takabayashi, K., Iacucci, M., Kuroki, T., Takishima, K., Sasabe, K., Niimura, Y., Kawashima, J., Ogawa, Y., Ichimasa, K., Nakamura, H., Matsudaira, S., Sasanuma, S., Hayashi, T., Wakamura, K., Miyachi, H., Baba, T., Mori, Y., Ohtsuka, K., Ogata, H., & Kudo, S.-E. (2025). Artificial intelligence-assisted video colonoscopy for disease monitoring of ulcerative colitis: A prospective study. *Journal of Crohn's and Colitis, 19*(1), jjae080. https://doi.org/10.1093/ecco-jcc/jjae080.
2. Shah, S. A., Taj, I., Usman, S. M., Shah, S. N. H., Imran, A. S., & Khalid, S. (2024). A hybrid approach of vision transformers and CNNs for detection of ulcerative colitis. *Scientific Reports, 14*(1), 24771. https://doi.org/10.1038/s41598-024-75901-4.

3. Takabayashi, K., Kobayashi, T., Matsuoka, K., Levesque, B. G., Kawamura, T., Tanaka, K., Kadota, T., Bise, R., Uchida, S., Kanai, T., & Ogata, H. (2024). Artificial intelligence quantifying endoscopic severity of ulcerative colitis in gradation scale. *Digestive Endoscopy, 36*(5), 582–590. https://doi.org/10.1111/den.14677.

4. Rimondi, A., Gottlieb, K., Despott, E. J., Iacucci, M., Murino, A., & Tontini, G. E. (2024). Can artificial intelligence replace endoscopists when assessing mucosal healing in ulcerative colitis? A systematic review and diagnostic test accuracy meta-analysis. *Digestive and Liver Disease, 56*(7), 1164–1172. https://doi.org/10.1016/j.dld.2023.11.005.

5. Stidham, R. W., Cai, L., Cheng, S., Rajaei, F., Hiatt, T., Wittrup, E., Rice, M. D., Bishu, S., Wehkamp, J., Schultz, W., Khan, N., Stojmirovic, A., Ghanem, L. R., & Najarian, K. (2024). Using computer vision to improve endoscopic disease quantification in therapeutic clinical trials of ulcerative colitis. *Gastroenterology, 166*(1), 155–167.e2. https://doi.org/10.1053/j.gastro.2023.09.049.

6. Kim, J. E., Choi, Y. H., Lee, Y. C., Seong, G., Song, J. H., Kim, T. J., Kim, E. R., Hong, S. N., Chang, D. K., Kim, Y.-H., & Shin, S.-Y. (2023). Deep learning model for distinguishing Mayo endoscopic subscore 0 and 1 in patients with ulcerative colitis. *Scientific Reports, 13*(1), 11351. https://doi.org/10.1038/s41598-023-38206-6.

7. Wang, G., Zhang, S., Li, J., Zhao, K., Ding, Q., Tian, D., Li, R., Zou, F., & Yu, Q. (2023). CB-HRNet: A class-balanced high-resolution network for the evaluation of endoscopic activity in patients with ulcerative colitis. *Clinical and Translational Science, 16*(8), 1421–1430. https://doi.org/10.1111/cts.13542.

8. Polat, G., Kani, H. T., Ergenc, I., Ozen Alahdab, Y., Temizel, A., & Atug, O. (2023). Improving the computer-aided estimation of ulcerative colitis severity according to Mayo Endoscopic Score by using regression-based deep learning. *Inflammatory Bowel Diseases, 29*(9), 1431–1439. https://doi.org/10.1093/ibd/izac226.

9. Iacucci, M., Cannatelli, R., Parigi, T. L., Nardone, O. M., Tontini, G. E., Labarile, N., Buda, A., Rimondi, A., Bazarova, A., Bisschops, R., del Amor, R., Meseguer, P., Naranjo, V., Ghosh, S., & Grisan, E. (2023). A virtual chromoendoscopy artificial intelligence system to detect endoscopic and histologic activity/remission and predict clinical outcomes in ulcerative colitis. *Endoscopy, 55*(4), 332–341. https://doi.org/10.1055/a-1960-3645.

10. Wu, S., Zhang, R., Yan, J., Li, C., Liu, Q., Wang, L., & Wang, H. (2023). High-speed and accurate diagnosis of gastrointestinal disease: Learning on endoscopy images using lightweight transformer with local feature attention. *Bioengineering, 10*(12), 1416. https://doi.org/10.3390/bioengineering10121416.

11. Sutton, R. T., Zaïane, O. R., Goebel, R., & Baumgart, D. C. (2022). Artificial intelligence enabled automated diagnosis and grading of ulcerative colitis endoscopy images. *Scientific Reports, 12*(1), 2748. https://doi.org/10.1038/s41598-022-06726-2.

12. Takenaka, K., Ohtsuka, K., Fujii, T., Negi, M., Suzuki, K., Shimizu, H., Oshima, S., Akiyama, S., Motobayashi, M., Nagahori, M., Saito, E., Matsuoka, K., & Watanabe, M.

(2020). Development and validation of a deep neural network for accurate evaluation of endoscopic images from patients with ulcerative colitis. *Gastroenterology, 158*(8), 2150–2157. https://doi.org/10.1053/j.gastro.2020.02.012.

13. Borgli, H., Thambawita, V., Smedsrud, P. H., Hicks, S., Jha, D., Eskeland, S. L., Randel, K. R., Pogorelov, K., Lux, M., Nguyen, D. T. D., Johansen, D., Griwodz, C., Stensland, H. K., Garcia-Ceja, E., Schmidt, P. T., Hammer, H. L., Riegler, M. A., Halvorsen, P., & de Lange, T. (2020). HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data, 7*(1), 283. https://doi.org/10.1038/s41597-020-00622-y.

14. Ozawa, T., Ishihara, S., Fujishiro, M., Saito, H., Kumagai, Y., Shichijo, S., Aoyama, K., & Tada, T. (2019). Novel computer-assisted diagnosis system for endoscopic disease activity in patients with ulcerative colitis. *Gastrointestinal Endoscopy, 89*(2), 416–421.e1. https://doi.org/10.1016/j.gie.2018.10.020.

15. Stidham, R. W., Liu, W., Bishu, S., Rice, M. D., Higgins, P. D. R., Zhu, J., Nallamothu, B. K., & Waljee, A. K. (2019). Performance of a deep learning model vs human reviewers in grading endoscopic disease severity of patients with ulcerative colitis. *JAMA Network Open, 2*(5), e193963. https://doi.org/10.1001/jamanetworkopen.2019.3963.

16. Pisano, E. D., Zong, S., Hemminger, B. M., DeLuca, M., Johnston, R. E., Muller, K., Braeuning, M. P., & Pizer, S. M. (1998). Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms. *Journal of Digital Imaging, 11*(4), 193–200. https://doi.org/10.1007/BF03178082.

17. Lecca, M., Gianini, G., & Serapioni, R. P. (2022). Mathematical insights into the original Retinex algorithm for image enhancement. *Journal of the Optical Society of America A, 39*(11), 2063–2072. https://doi.org/10.1364/JOSAA.471953.

18. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision Transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022. https://doi.org/10.1109/ICCV48922.2021.00986.

19. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*, 618–626. https://doi.org/10.1109/ICCV.2017.74.

20. Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.1609.02907.

21. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539.