applied sciences

Article

Ulcerative Severity Estimation Based on Advanced CNN–Transformer Hybrid Models

Boying Nie † **and Gaofeng Zhang** *,†

School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China; 320220941890@lzu.edu.cn

* Correspondence: zhanggaof@lzu.edu.cn

† These authors contributed equally to this work.

Abstract

The neural network-based classification of endoscopy images plays a key role in diagnosing gastrointestinal diseases. However, current models for estimating ulcerative colitis (UC) severity still lack high performance, highlighting the need for more advanced and accurate solutions. This study aims to apply a state-of-the-art hybrid neural network architecture—combining convolutional neural networks (CNNs) and transformer models—to classify intestinal endoscopy images, utilizing the largest publicly available annotated UC dataset. A 10-fold cross-validation is performed on the LIMUC dataset using CoAtNet models, combined with the Class Distance Weighted Cross-Entropy (CDW-CE) loss function. The best model is compared against pure CNN and transformer baselines by evaluating performance metrics, including quadratically weighted kappa (QWK) and macro *F*1, for full Mayo score classification, and kappa and *F*1 scores for remission classification. The CoAtNet models outperformed both pure CNN and transformer models. The most effective model, CoAtNet_2, improved classification accuracy by 1.76% and QWK by 1.46% over the previous state-of-the-art models on the LIMUC dataset. Other metrics, including *F*1 score, also showed clear improvements. Experiments show that the CoAtNet model, which integrates convolutional and transformer components, improves UC assessment from endoscopic images, enhancing AI's role in computer-aided diagnosis.

Keywords: ulcerative colitis (UC); CoAtNet; Mayo score; computer-aided diagnosis

check for updates

Academic Editor: Douglas O'Shaughnessy

Received: 5 June 2025
Revised: 1 July 2025
Accepted: 2 July 2025
Published: 3 July 2025

Citation: Nie, B.; Zhang, G. Ulcerative Severity Estimation Based on Advanced CNN–Transformer Hybrid Models. *Appl. Sci.* **2025**, *15*, 7484. https://doi.org/10.3390/app15137484

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Ulcerative colitis (UC), a chronic and relapsing subtype of inflammatory bowel disease (IBD), is characterized by continuous inflammation and ulceration confined to the mucosal layer of the colon [1]. This pathological process manifests clinically as prominent symptoms, with diarrhea and rectal bleeding being the most characteristic hallmarks [2]. However, the implications of UC extend beyond these acute symptoms; without proper management, the condition gradually increases the risk of colorectal cancer over time [3]. Given the evolving disease landscape of ulcerative colitis (UC) and its potential long-term consequences, early and accurate diagnosis is crucial for effective disease control. Conventionally, diagnosis relies on a comprehensive assessment that integrates clinical manifestations, endoscopic examinations, histopathological findings, and the exclusion of other infectious or noninfectious forms of colitis [4]. This limitation, however, is being addressed by the rapid progress of artificial intelligence (AI) in healthcare. Deep learning, in particular, has emerged as a

potent force in medical image analysis, effectively showcasing its substantial potential for detecting and diagnosing diverse diseases [5].

Prevalent CNN models primarily focus on local characteristics, whereas transformers emphasize global information. In addition, although CNNs and transformers have demonstrated robust performance on large-scale datasets, they face significant challenges in the medical domain, where datasets are typically much smaller. Moreover, considering the characteristics of medical images—particularly intestinal endoscope images—pure CNNs may tend to focus excessively on specific local features, which can lead to the learning of spurious correlations and ultimately result in overfitting. On the other hand, pure transformers may fail to adequately capture subtle differences in lesion areas, thereby reducing their ability to distinguish between different degrees of pathological severity. Against the backdrop of these technological advancements, this study undertakes an in-depth exploration of cutting-edge neural network architectures, specifically focusing on their application in the classification and interpretability of intestinal endoscopic images. Specifically, the CoAtNet model inherits the generalization capability of CNNs, such as ConvNets, along with the global feature awareness of traditional vision transformers (ViTs), making it particularly well suited for medical imaging tasks such as intestinal endoscopy. Moreover, unlike other hybrid CNN–transformer models, CoAtNet adopts a vertically structured design that clearly separates the CNN and transformer stages. This design leverages prior knowledge, emphasizing the convolution network's foundational role. Notably, the original study demonstrates that CoAtNet achieves strong performance even on relatively small datasets, as evidenced by its success when trained directly on ImageNet-1k. These strengths collectively motivate our choice of CoAtNet as the backbone model for our experiments [6]. By leveraging its architectural strengths in capturing both local and global features, we aim to enhance the accuracy and reliability of clinical examination and classification. Additionally, we seek to elucidate the decision-making mechanisms of the model to validate its practical usability and trustworthiness.

The main contributions of this study are summarized as follows:

- This study evaluates the classification performance of the CoAtNet model on the LIMUC dataset, benchmarking it against cutting-edge architectures, including pure convolutional and transformer-based models. Through this comparative analysis, the distinct strengths and potential advantages of CoAtNet emerge, positioning it as a notable advancement within the field.
- Through the combination of the CDW-CE loss function, we achieve further improvements in both QWK and classification accuracy.
- To enhance model interpretability and validate practical usability, we visualize the decision-making process using class activation maps (CAMs), offering intuitive insights into how predictions are made.

Related Work

CNNs in the Field of Medical Imaging: Convolutional neural networks have brought about transformative advancements in image analysis by emulating the visual processing mechanisms of the brain [7]. Researchers have extensively applied CNN-based models to extract distinctive and meaningful features from medical images, and they often surpass traditional machine learning methods and manually intensive volumetric analysis techniques [8]. For instance, Ismail et al. proposed a CNN-based regular pattern discovery model that effectively extracts disease-related knowledge from unstructured medical records [9]. Li et al. developed the multi-instance multiscale CNN (MIMS-CNN) model, which integrates MSConv layers and top-k pooling, and demonstrates superior performance across three medical-related tasks [10]. Moreover, in the area of endoscopic analysis,

numerous CNN-based models have been developed for wireless capsule endoscopy (WCE) image classification [11]. For example, Rustam et al. introduced the BIR (Bleedy Image Recognizer) model, which combines MobileNet with a custom CNN to classify WCE images showing bleeding [12]. Sharif et al. employed a novel fusion approach combining deep CNN features (VGG16 and VGG19) with geometric descriptors using a Euclidean Fisher Vector [13]. While the mentioned demonstrate effectiveness in specific tasks, they exhibit notable limitations. For instance, MIMS-CNN relies on fixed convolutional kernel designs and specific pooling strategies, which restrict its adaptability to variable and multiscale lesion features in medical images. The CNN-based approach of Ismail et al. for text extraction lacks deep integration with standard medical terminology, knowledge graphs, and structured medical knowledge, thereby limiting clinical interpretability and generalizability. The BIR model, despite combining MobileNet with a custom CNN, heavily relies on local convolution operations, resulting in reduced robustness to changes in lesion locations within images. Additionally, Sharif et al.'s introduction of geometric descriptors and deep feature fusion does not include visualization or decision-making analysis illustrating how the model integrates such information for diagnostic purposes.

Transformers in the Field of Medical Imaging: Recently, with transformer models extending to computer vision fields [14], transformer-based architectures have demonstrated significant potential across various medical imaging tasks [15]. For robust medical image classification, the self-ensembling vision transformer (SEViT) introduces a self-distillation framework that enhances the resilience of early layer feature representations in vision transformers (ViTs) against adversarial perturbations [16]. In the context of 3D medical image segmentation, the UNEt transformer (UNETR) architecture leverages a transformer-based encoder to effectively model global multiscale contextual information while maintaining the well-established U-shaped design to integrate both high-level semantic features and low-level spatial details [17]. Furthermore, in the domain of gastroscopy and colonoscopy, a vision transformer model incorporating hybrid shifted windows was proposed to capture both short-range and long-range dependencies for digestive tract image classification [18]. However, despite the strong performance of vision transformer-based models, several challenges remain. For example, although SEViT employs a self-distillation mechanism to address the issue of limited training data, it is still prone to overfitting, particularly in cases with extremely small sample sizes or imbalanced class distributions. Additionally, models such as UNETR, which integrate transformer encoders with U-Net decoders, incur considerable computational and memory overhead, posing difficulties for deployment on resource-constrained medical devices, such as endoscopic image analysis systems.

Hybrid Models: From the above analysis, it is clear that both CNN and transformer models are often task-specific, with limited generalizability and cross-modal adaptability, making them insufficient to meet the diverse and evolving demands of real-world medical applications. To overcome these limitations, hybrid CNN–transformer models have been proposed, aiming to leverage the complementary strengths of both architectures. By combining the strong local feature extraction capabilities of CNNs with the global contextual understanding of transformers, these models improve the ability to detect both fine-grained lesion features and broader spatial patterns, thereby enhancing diagnostic performance.

By incorporating locality principles into transformer frameworks, some hybrid architectures achieve superior performance [19]. MedViT integrates efficient convolution blocks (ECBs) to reduce the quadratic complexity of self-attention and enable joint attention across diverse representational subspaces [20]. Turan et al. proposed UC-NfNet, a hybrid CNN–transformer model with an image generation module, which demonstrated high effectiveness on a newly curated ulcerative colitis (UC) dataset of 673 images [21]. While these previously discussed models have demonstrated efficacy in their respective domains,

they are not well suited for our research objectives due to their highly specialized design. MedViT, for instance, predominantly emphasizes the extraction of global structural features such as shape and overall appearance but overlooks fine-grained texture details critical for our dataset. Meanwhile, UC-NfNet's integration with an image production technique renders it incompatible with our research focus.

In contrast, our study aims to identify a versatile hybrid model that seamlessly combines the strengths of CNNs and transformer architectures. Given the distinct structural patterns present in our images, CoAtNet emerges as the optimal choice. By effectively integrating local feature extraction capabilities of CNNs with the global attention mechanism of transformers, CoAtNet strikes a balance between capturing fine-grained details and understanding overall structural contexts, thereby fulfilling the specific requirements of our research.

Research Status of Intestinal Endoscopic Image Classification: To the best of our knowledge, intestinal endoscopic image classification still largely depends on traditional deep CNNs, with most studies conducted on relatively small datasets. In contrast, the LIMUC dataset, released in 2022, is the largest publicly available collection of intestinal endoscopic images to date, and its high quality has been demonstrated in previously published studies, as discussed later in this paper. However, related research is scarce, with only two published studies. Polat et al. introduced the dataset and applied regression-based deep learning to establish baseline performance, highlighting its potential for model development and external validation [22]. Another study enhances classification accuracy and quadratically weighted kappa (QWK) by proposing a Class Distance Weighted Cross-Entropy (CDW-CE) loss function based on conventional CNN architectures [23].

Despite its scale and significance, research on the LIMUC dataset has seen little progress over the past two years. This highlights the need for further exploration of this dataset and the development of more effective models tailored to its characteristics.

A positive development is that the field of computer vision has seen the emergence of advanced models such as OverLoCK—the first pure ConvNet-based backbone integrating a top-down attention mechanism [24], and MaxViT, a hierarchical vision backbone featuring an efficient and scalable attention mechanism known as the multi-axis attention model [25]. These state-of-the-art models have been applied to the LIMUC dataset to explore improved results, as detailed in the experimental section.

2. Materials and Methods

2.1. LIMUC Dataset

The dataset utilized in this study is the Labeled Images for Ulcerative Colitis (LIMUC) dataset [26]. It comprises colonoscopic imaging data acquired between December 2011 and July 2019 for the diagnosis of ulcerative colitis. The LIMUC dataset includes 11,276 images from 564 patients across 1043 colonoscopy procedures. (The Mayo Endoscopic Score is the most widely used endoscopic assessment tool in clinical trials [27].) The dataset was split into a training/validation set consisting of 9590 images from 479 patients (85%), and a test set with 1686 images from 85 patients (15%).

Ulcerative colitis (UC) is typically categorized into mild, moderate, and severe stages based on clinical and laboratory findings [28]. The Mayo scoring system ranges from 0 (normal mucosa) to 3 (severe disease), and the LIMUC dataset reflects this with four classes, namely Mayo 0 (54.14%), Mayo 1 (27.70%), Mayo 2 (11.12%), and Mayo 3 (7.67%), with the symptoms as follows:

Mayo 0: normal or mild inflammation: visible vascular patterns without bleeding;

Mayo 1: mild inflammation: vascular pattern loss, mild erythema, edema, or minimal bleeding;

Mayo 2: moderate inflammation: noticeable erythema, edema, spontaneous bleeding, or erosion;

Mayo 3: severe inflammation: ulcers, spontaneous bleeding, or pronounced colonic wall swelling [29].

2.2. CoAtNet Models

CoAtNet is a CNN–transformer hybrid model introduced by the Google team. It combines the strong generalization ability of convolutional networks—particularly effective under limited data—with the scalability and representation capacity of transformer models when ample data are available. It also exhibits faster convergence and enhanced efficiency.

The model adopts a multi-stage architecture that mimics traditional ConvNets by gradually performing pooling operations across stages. Specifically, the network is composed of five stages (S0 to S4), where each stage reduces the spatial resolution by a factor of two while increasing the number of channels. In Stages S1 and S2, the network stacks MBConv blocks to capture spatial interactions, leveraging prior layers, given that convolution is more effective at extracting local features, particularly in early layers. In the later stages, transformer blocks are introduced to expand the global receptive field and model long-range dependencies and the model employs a pre-normalized relative attention variant in the transformer blocks, which serves as a core component of the architecture. Both the Feed-Forward Network (FFN) in the transformer and the MBConv block utilize the inverted bottleneck design to balance model capacity and generalization [6]. The overall structural pattern is C–C–T–T, where C denotes convolutional modules, and T denotes transformer modules. The main flowchart and structure are shown in Figure 1.

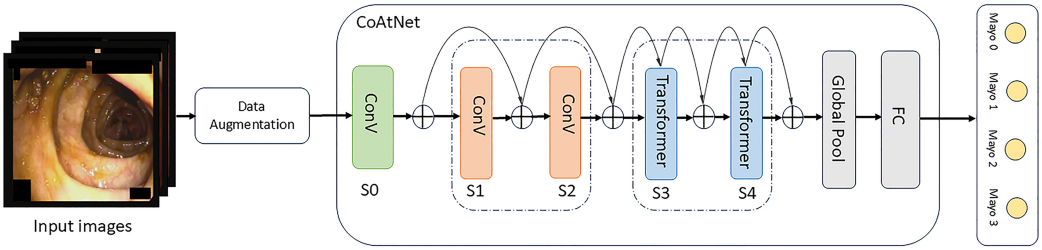

Figure 1. The main flowchart and structure of our experiment and model.

It is worth noting that a key innovation of CoAtNet lies in its combination of a global static convolution kernel with an adaptive attention matrix, as defined in Equations (1) and (2):

$$Y_i = \sum_{j \in G} \frac{exp(x_i^{\mathrm{T}} x_j + w_{i-j})}{A_{ij}} x_j, \tag{1}$$

$$A_{ij} = \sum_{k \in G} exp(x_i^{\mathrm{T}} x_k + w_{i-k}), \tag{2}$$

where $x_i, Y_i \in \mathbb{R}^D$ denote the input and output at position i, respectively. For example, x_i and x_j represent two patches from the same image. Each patch is compared with all other patches in the image using their dot product, i.e., $x_i^T x_j$. G refers to the global spatial space, and A_{ij} represents a modified attention weight incorporating an additional local component. For any position pair (i, j), the corresponding convolutional weight w_{i-j} depends solely on their relative displacement.

The above formulas are based on the standard attention weight computation

$$Attention(x_i, x_j) = \sum_{j \in G} \frac{exp(x_i^{\mathrm{T}} x_j)}{\sum_{k \in G} exp(x_i^T x_k)}, \quad (3)$$

which relies solely on the similarity between input features. An additional term, w_{i-j}, is introduced to represent a convolution-style spatial bias. This term depends only on the relative positional difference $(i - j)$ and not on the input features themselves. This means that for each patch, while computing its similarity with other patches, the model also takes into account their relative spatial distance. This approach effectively integrates the translational invariance and local structural modeling capabilities of CNNs with the global contextual modeling strength of transformers.

2.3. Loss Function

Experiments revealed that using the CoAtNet model can improve performance on the LIMUC dataset; therefore, to achieve better results, we further adopted the Class Distance Weighted Cross-Entropy (CDW-CE) loss function, which is specifically designed for datasets involving ordinal labels, such as the Mayo scoring system. Unlike the traditional cross-entropy loss function that treats all misclassifications equally, the CDW-CE computes the logarithmic loss for each misclassification based on a category distance matrix; the term $|i - i_{true}|$ computes the distance between each predicted Mayo score and the true score, capturing the ordinal nature of the Mayo score classification. The associated penalty is scaled by a weighting coefficient α. This coefficient modulates the loss according to the severity of inter-category differences [23]. This design is formally captured in the loss function calculated using Equation (4)

$$Loss = -\sum_{i=0}^{N-1} log(1 - y_i) \times |i - i_{true}|^{\alpha}, \quad (4)$$

where i represents the category index in the output layer, ranging from 0 to 3; y_i denotes the predicted probability distribution across categories for a given image; and i_{true} signifies the index of the true category.

The methodology enhances the model's ability to discern between different levels of misclassification severity, aligning more closely with the nuanced demands of our dataset's categorical structure. This adjustment aims to improve the decision-making process of the model by incorporating the informational value of misclassifications based on their proximity to the correct category, thus refining model training outcomes.

2.4. Evaluation Indicators

Quadratic weighted kappa (QWK) is commonly used to assess agreement on ordinal scales and is particularly well suited for addressing class imbalance issues in this task. In remission classification, metrics like Cohen's kappa and $F1$ score offer more reliable evaluation than sensitivity, precision, or recall, which are often overly influenced by false positives or false negatives. Given the characteristics of the LIMUC dataset, QWK and Cohen's kappa were adopted as the primary evaluation metrics, with macro $F1$ score serving as an additional measure for assessing performance across all Mayo subscores.

Cohen's kappa: The kappa statistic is widely utilized to assess interrater reliability metric, offering a method that accounts for the possibility of agreement occurring by chance, particularly when raters face uncertainty and may guess on some variables [30]. Cohen's kappa specifically evaluates the degree of agreement between two raters independently classifying N items into m mutually exclusive categories. The formal definition of the kappa coefficient(κ) is given as Equation (5) [31]

$$\kappa \equiv \frac{P_o - P_e}{1 - P_e} = \frac{1 - P_o}{1 - Pe}, \tag{5}$$

where P_o denotes the observed agreement, defined as the proportion of instances where two observers or classifiers assign the same label to a given sample. Conversely, P_e represents the expected agreement by chance, indicating the probability that two observers or classifiers coincide purely due to random assignment.

Quadratic Weighted Kappa(QWK): The quadratic weighted kappa (QWK) enables differential weighting of disagreements between raters [32]. It is particularly well suited for ordered classification tasks, as it effectively captures the degree of deviation between predicted outcomes and true labels [33]. The corresponding mathematical formulation is provided in Equation (6)

$$QWK = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}}, \tag{6}$$

where $O_{i,j}$ denotes the element of the observed confusion matrix, representing the number of instances whose true class is i but are predicted as class j. $E_{i,j}$ refers to the corresponding element in the expected confusion matrix, indicating the expected number of samples classified as j when the true category is i under random assignment (the equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability). $W_{i,j}$ is an element of the weight matrix, typically defined as $w_{i,j} = \frac{(i-j)^2}{(N-1)^2}$, where N denotes the total number of classes.

2.5. Data Augmentation

An automated data augmentation strategy was employed for the training set to enhance data diversity, incorporating supplementary techniques such as image color dithering, image differencing, and random erasure. For the validation set, preprocessing included image differencing, center cropping, and normalization. In contrast, the test set underwent only image resizing to 224 × 224 pixels followed by normalization. The selected data augmentation strategy was determined through extensive experiments as the most effective combination to minimize bias and overfitting. The same preprocessing procedures were applied to all models for a fair comparison.

2.6. Training Parameters and Processes

In this study, pre-trained weights from the ImageNet-1K dataset were adopted as the initial parameters for the CoAtNet model. Freezing early layers of the model often helps reduce training costs, improve stability, and prevent overfitting by retaining general features, and this approach is well suited for small-sample and transfer learning tasks. Based on multiple experimental evaluations, we decided to freeze the early stages of the network to achieve a better balance between performance and computational efficiency. A uniform 10-fold cross-validation strategy was employed throughout all experiments, with the model trained for 200 epochs per fold. For each fold, the optimal model checkpoint—determined by performance on the validation set—was used to evaluate test set performance. After extensive experimental tuning, the Adam optimizer was employed, with a learning rate decay factor of 0.1 and a weight decay coefficient of 0.05. All experiments were conducted using the PyTorch 2.0.0 framework, and computations were accelerated with a VGPU-32G and RTX 4090 graphics processing unit. The average training time for each fold was approximately 2 h.

2.7. Selection of CDW-CE Penalty Factors

Since the selection of the α parameter in the CDW-CE loss function significantly affects model performance across different architectures, this study systematically investigated its impact. To determine the optimal α values for various models, we plotted the test curves of CoAtNet (Levels 0 to 3), DeiT, MaxViT, and OverLock architectures—with the latter three being base-level models—across a range of α values (from 1 to 10), as shown in Figure 2, which presents the corresponding accuracy, quadratic weighted kappa (QWK), and standard deviation values. All subsequent comparative experiments were conducted using the best-performing α values identified for each model based on this analysis. According to the experimental results, the optimal α values for CoAtNet_0, CoAtNet_1, CoAtNet_2, CoAtNet_3, DeiT, MaxViT, and OverLock were 6.0, 6.0, 5.0, 6.0, 5.0, 6.0, and 5.0, respectively.

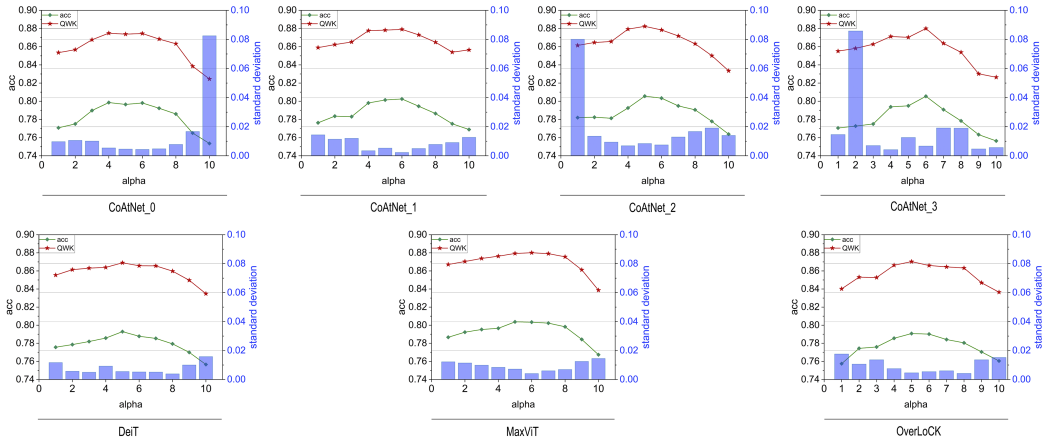

Figure 2. Test results of various models across different α parameter settings.

3. Results

3.1. Key Findings

In this study, the quadratic weighted kappa coefficient (QWK), accuracy, and *F*1 score were employed as the primary metrics for model performance evaluation. Model performance was evaluated under two assessment schemes: a full four-grade Mayo score classification and a binary remission classification. In the binary classification setting, remission is defined as a Mayo score of 0 or 1, while active disease corresponds to a score of 2 or 3 [22]. Notably, the model was exclusively trained on the full Mayo score classification task, and its predictions were subsequently mapped to the remission categories for binary evaluation.

CoAtNet provides four structural variants, defined by the number of blocks at each stage, assuming an input resolution of 224 × 224. In this study, we utilized the CoAtNet_2 architecture, which consisted of block configurations of (2, 2, 6, 14, 2) and corresponding hidden dimensions of (192, 192, 256, 512, 1014) across Stages S0 to S4, respectively. A kernel size of 3 was uniformly employed for all convolutional (Conv) and mobile inverted bottleneck convolutional (MBConv) blocks. The detailed parameter settings for other CoAtNet variants are available in the original publication [6].

Under identical training configurations, the complete four-level Mayo score classification results and the remission classification outcomes for the four models with different specifications are summarized in Table 1. The results indicate that, for the LIMUC dataset, the CoAtNet-2 model achieved the best performance. Meanwhile, the performance of all models significantly exceeded the current best benchmark on LIMUC. To further highlight the inherent advantages of the model architecture, we select the CoAtNet_2 model as a

representative and compare it with CNN and Transformer models of varying structures. The results are presented in Tables 2 and 3, respectively.

It can be observed from the results that although models such as MaxVit and Overlock incorporate a balance between local and global attention, their performance still shows a noticeable gap compared to the best CoAtNet model. To further compare the classification results of different models (CoAtNet_2, Deit, MaxVit, and Overlock) for each category, we provide the final confusion matrix of the model for the classification of different grades of ulcerative colitis. This matrix, representing the aggregate results from 10-fold cross-validation, visually illustrates the model's classification performance, as shown in Figure 3. It is worth noting that the confusion matrices for Inception-v3, MobileNet_large_v3, and ResNet18 are omitted, as they were presented in the research paper on the LIMUC dataset [23].

Table 1. Classification results for the complete 4-level Mayo scores and remission tasks across four CoAtNet model specifications.

	4-Score			Remission		
Model	**Accuracy**	***QWK***	***F1-Score***	**Accuracy**	***kappa***	***F1-Score***
CoAtNet_0	0.7981 ± 0.0040	0.8745 ± 0.0044	0.7381 ± 0.0085	0.9610 ± 0.0026	0.8634 ± 0.0097	0.8869 ± 0.0082
CoAtNet_1	0.8025 ± 0.0040	0.8791 ± 0.0075	0.7410 ± 0.0075	0.9613 ± 0.0018	0.8654 ± 0.0072	0.8889 ± 0.0062
CoAtNet_2	0.8056 ± 0.0093	0.8824 ± 0.0084	0.7428 ± 0.0128	0.9630 ± 0.0042	0.8737 ± 0.0012	0.8962 ± 0.0093
CoAtNet_3	0.8054 ± 0.0072	0.8791 ± 0.0067	0.7426 ± 0.0075	0.9620 ± 0.0030	0.8701 ± 0.0084	0.8932 ± 0.0066

Table 2. Performance comparison between the CoAtNet-2 model and CNN models based on the complete 4-score Mayo score and remission classification.

	4-Score			Remission		
Model	**Accuracy**	***QWK***	***F1-Score***	**Accuracy**	***kappa***	***F1-Score***
CoAtNet_2	0.8056 ± 0.0093	0.8824 ± 0.0084	0.7428 ± 0.0128	0.9630 ± 0.0042	0.8737 ± 0.0012	0.8962 ± 0.0093
Inception_v3 *	0.7880 ± 0.0110	0.8678 ± 0.0060	0.7261 ± 0.0150	0.9590 ± 0.0030	0.8598 ± 0.0120	0.8847 ± 0.0100
ResNet18 *	0.7740 ± 0.0110	0.8568 ± 0.0100	0.7055 ± 0.0210	0.9566 ± 0.0050	0.8521 ± 0.0160	0.8785 ± 0.0130
MobileNet *	0.7759 ± 0.0100	0.8588 ± 0.0060	0.7254 ± 0.0100	0.9588 ± 0.0050	0.8592 ± 0.0120	0.8842 ± 0.0100
Overlock_b	0.7908 ± 0.0042	0.8670 ± 0.0040	0.7225 ± 0.0101	0.9574 ± 0.0026	0.8514 ± 0.0088	0.8771 ± 0.0074

* Note: The results for Inception_v3, ResNet18, and MobileNet_V3_Large (referred to as MobileNet in this table) are cited directly from their respective original publications [23].

Table 3. Performance comparison between the Coatnet_2 model and pure transformer models based on the complete 4-level Mayo score and remission classification.

	4-Score			Remission		
Model	**Accuracy**	***QWK***	***F1-Score***	**Accuracy**	***kappa***	***F1-Score***
CoAtNet_2	0.8056 ± 0.0093	0.8824 ± 0.0084	0.7428 ± 0.0128	0.9630 ± 0.0042	0.8737 ± 0.0012	0.8962 ± 0.0093
DeiT	0.7929 ± 0.0069	0.8690 ± 0.0055	0.7309 ± 0.0121	0.9590 ± 0.0038	0.8573 ± 0.0139	0.8822 ± 0.0117
MaxViT	0.8034 ± 0.0060	0.8801 ± 0.0041	0.7455 ± 0.0105	0.9612 ± 0.0017	0.8668 ± 0.0067	0.8904 ± 0.0059

Based on the confusion matrix results, it is evident that all models achieved the highest classification performance on Mayo 0, followed by Mayo 3 or Mayo 1, while Mayo 2 was associated with the worst scores. This trend may be attributed to the imbalanced data distribution. Comparative analysis also shows that our models exhibited better classification ability on class Mayo 3 compared to traditional CNN models, and the CoaTNet model improved classification performance across all categories compared to all the other models, rather than being optimized for a specific class.

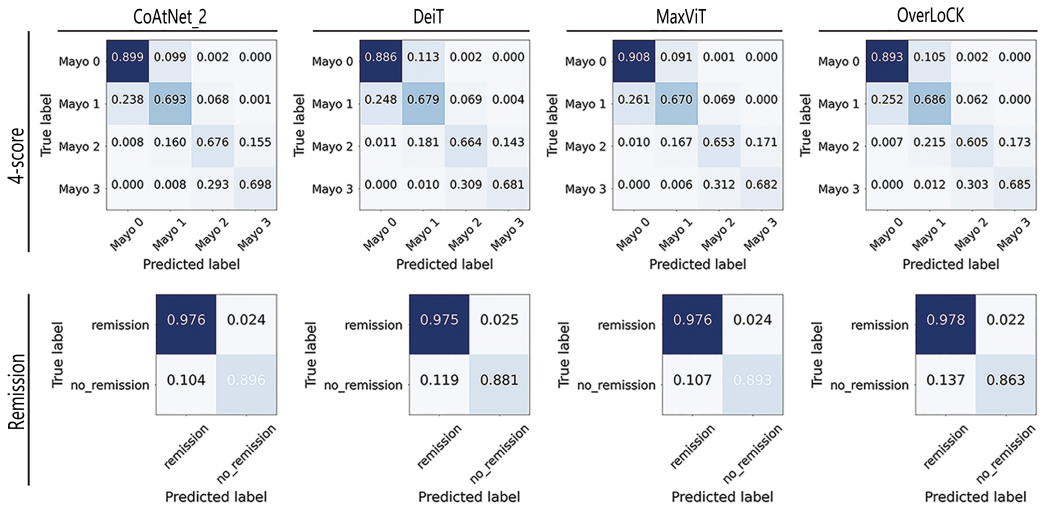

Figure 3. The confusion matrix of the 10-fold test results of different models.

Moreover, we conducted ROC and AUC analyses to provide a more intuitive comparison of the classification capabilities of different models across each category from a new viewpoint. As shown in Figure 4, the performance of each model in the four-score plus remission classification task is illustrated using ROCs based on the one-vs-rest strategy for each class. The corresponding macro-averaged AUC values were also calculated. These observations imply that both the individual class AUCs and the macro-average AUCs were close to or exceeded 0.90, demonstrating the models' strong and consistent classification performance across all categories. In terms of the remission classification, the AUC results were 0.99, which revealed high classification performance. An interesting phenomenon is that the Mayo 3 category, despite having the lowest classification accuracy, achieved the highest AUC score. This is because AUC and accuracy evaluate model performance from different perspectives. AUC measures the model's overall discriminative ability and is threshold-independent, whereas accuracy reflects the model's correctness at a specific decision threshold. The high AUC for the Mayo 3 category indicates that the model generally assigns higher prediction probabilities to Mayo 3 samples compared to non-Mayo 3 samples. However, its low accuracy suggests that Mayo 3 is inherently more challenging to classify than other categories.

Finally, to interpret the decision-making process of the model, we employed Gradient-Weighted Class Activation Mapping (Grad-CAM) to visualize the output feature maps at different stages of the network. This includes the feature maps from the last convolutional layers in the convolutional modules (Stage 1–2), as well as the output of the final normalization layers in the last transformer modules (Stage 3–4), which reflect the attention-enhanced feature representations. These visualizations effectively highlight the regions of interest attended to by the model. The detailed visualization and workflow are presented in Figure 5.

From the figure, it can be observed that after the initial two convolutional stages, the model primarily attends to scattered local regions, with limited focus and minimal differentiation in attention across areas. In contrast, following the final two transformer stages, the model exhibits a more concentrated and coherent focus on broader regions associated with lesion presence. This suggests that the model's decision-making is aligned with clinically relevant patterns rather than random or arbitrary attention.

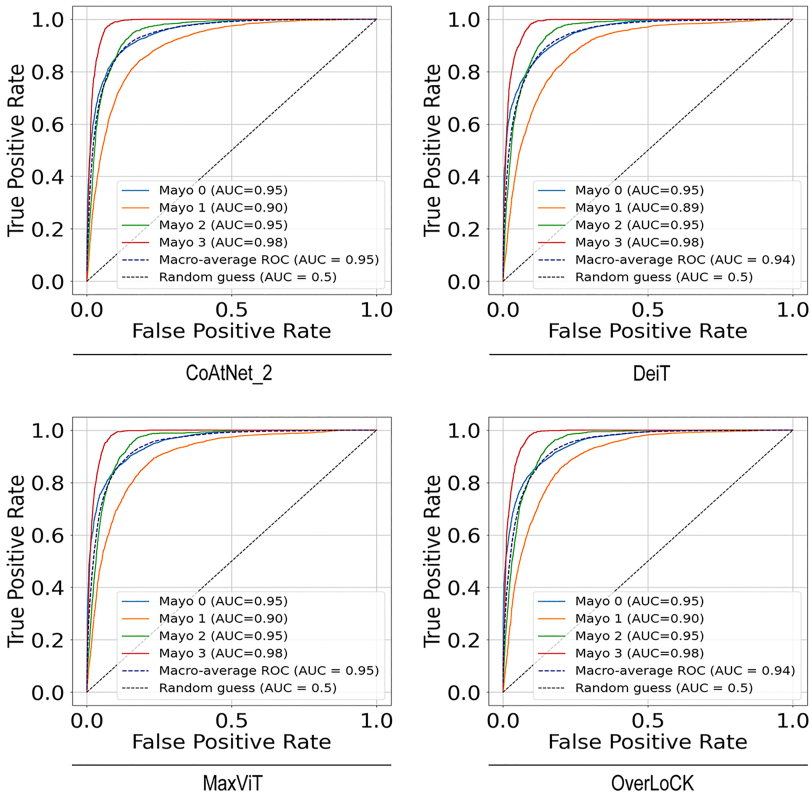

Figure 4. ROCs (receiver operating characteristic curves) for 4-level and remission classification across different models.

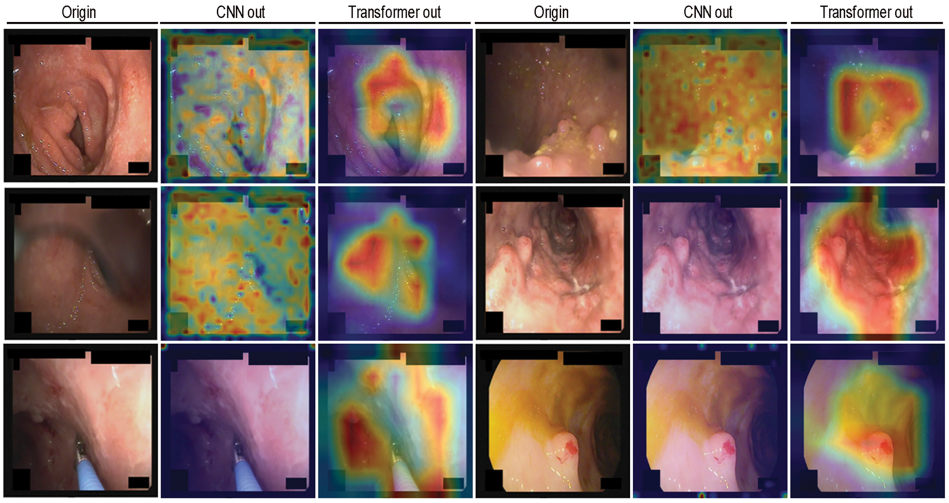

Figure 5. Gradient-Weighted Class Activation Mapping (Grad-CAM) visualizations of the CoaTNet model on the LIMUC dataset.

3.2. Statistical Result Analysis

In this part of the study, we conducted a Wilcoxon signed-rank test [34] on the 10-fold results of accuracy, quadratic weighted kappa (QWK), and *F*1 score to perform pairwise comparisons between the CoAtNet_2 model and other baseline models, including pure convolutional neural networks (CNNs) and transformer-based architectures. The statistical test results are summarized in Table 4.

The test results indicate that CoAtNet_2 model showed statistically significant performance differences compared to most baseline models across the evaluated metrics. Among all pairwise comparisons, only two *p*-values exceeded the 0.05 significance threshold, suggesting that the superiority of CoAtNet is generally consistent and statistically robust.

These findings highlight the effectiveness of CoAtNet in capturing both local and global features, contributing to improved classification performance on the LIMUC dataset.

Table 4. *p*-values from the Wilcoxon signed-rank test comparing the performance of the CoAtNet_2 model with pure convolutional neural network(CNN) and transformer models.

Models	Inception *	ResNet18	MobileNet *	OverLock_b	DeiT	MaxViT
accuracy	0.0039	0.0020	0.0020	0.0195	0.0195	0.3750
QWK	0.0059	0.0020	0.0020	0.0059	0.0137	0.0020
*F*1-score	0.0371	0.0020	0.0098	0.0039	0.1309	0.0020

* Note: In this table, 'Inception' refers to Inception_v3, and 'MobileNet' refers to MobileNet_V3_Large.

3.3. Comparison and Ablation Experiments

Although previous experiments demonstrated the superior overall performance of CoAtNet combined with the CDE-CE loss on the LIMUC dataset, the individual contributions of the loss function and the CoAtNet architecture were not clearly distinguished. To further investigate these aspects, additional experiments were conducted to evaluate the impact of different loss functions on model performance. These experiments also aimed to address the residual class imbalance issue that may still influence the results. The detailed outcomes of these evaluations are presented in Table 5.

From the above results, it is evident that the CoAtNet_2 model exhibits inherent advantages when applied to the LIMUC dataset, and the incorporation of the CDW-CE loss function further enhances its performance.

Table 5. Comparison of results based on different loss functions in the CoAtNet_2 model.

Loss Function	Cross-Entropy	CDW-CE	Focal Loss	LMF *
accuracy	0.8029 ± 0.0042	0.8056 ± 0.0093	0.7697 ± 0.0207	0.7692 ± 0.0169
QWK	0.8785 ± 0.0048	0.8824 ± 0.0084	0.8537 ± 0.0153	0.8512 ± 0.0115
*F*1-score	0.7413 ± 0.0101	0.7428 ± 0.0128	0.6864 ± 0.0287	0.6827 ± 0.0214

* Note: In this table, 'LMF' refers to large-margin-aware focal loss.

Focal loss is a commonly used loss function designed to address class imbalance. By introducing a modulation factor to the standard cross-entropy loss, it down-weights well-classified (easy) examples and up-weights hard examples, encouraging the model to focus on more informative, difficult samples during training. LMF (large-margin-aware focal loss) extends this idea by linearly combining focal loss with LDAM (label-distribution-aware margin loss) [35], thereby simultaneously leveraging class-distribution-aware margins and emphasizing hard-to-classify samples [36]. In our experiments, the focal loss was configured with $\gamma = 1.5$, and the values of α were set according to the sample distribution of each class. For LMF, we adopted $\alpha = 1.5$ and $\beta = 1.0$.

However, neither focal loss nor LMF led to improved model performance on the LIMUC dataset; instead, both resulted in noticeable performance degradation. This outcome may be attributed to suboptimal hyperparameter settings or other dataset-specific factors, which warrant further investigation in our future work.

4. Discussion

This study delves deeper into the integration of artificial intelligence within intestinal endoscopy, employing advanced deep neural networks for Mayo grade assessment in ulcerative colitis. Specifically, it explores the applicability of the CoatNet architecture, coupled with the Class Distance Weighted Cross-Entropy (CDW-CE) loss function and corresponding training methodologies tailored to the LIMUC dataset. This approach enhances key metrics such as accuracy, quadratically weighted kappa (QWK), and other

performance indicators. However, the study has certain limitations, including the need for further investigation into whether all hybrid CNN–transformer networks offer superior performance, and whether they achieve similarly strong results in other medical domains, which remains to be explored.

In addition, from a practical usability perspective, the model can assist clinicians in identifying lesions more accurately and efficiently, thereby enhancing diagnostic precision. However, before clinical deployment, its performance in terms of accuracy, robustness, and safety must be thoroughly validated through prospective clinical trials and real-world data. Upon successful validation, the model may be considered for integration into hospital information systems as a clinical decision support tool.

5. Conclusions

In this study, we applied the state-of-the-art CoAtNet model, in conjunction with the Class Distance Weighted Cross-Entropy (CDW-CE) loss function, to the largest publicly available ulcerative colitis dataset—LIMUC. This work aimed to validate the effectiveness of the CNN–transformer hybrid architecture in the context of intestinal endoscopy image classification. Comparative experiments were conducted against recent advanced convolutional neural network (CNN) and pure transformer-based models, further demonstrating the superior performance of the proposed approach. Our method achieved a classification accuracy improvement from 78.80% to 80.56% and an increase in the quadratically weighted kappa (QWK) score from 86.78% to 88.24%, along with consistent improvements across other evaluation metrics. To enhance the interpretability of the model—an essential aspect in medical image analysis—we visualized the decision-making process using class activation maps (CAMs), providing insights into the regions influencing the model's predictions. Looking forward, we plan to further optimize the CoAtNet architecture by integrating more targeted strategies, refining its structural components, and exploring its compatibility with other model paradigms. These efforts aim at further improving its robustness and applicability in clinical settings.

To promote transparency and reproducibility, all source code is made publicly available at https://github.com/Nby8/limuc-nets.git (accessed on 20 May 2025).

Author Contributions: Conceptualization, B.N.; investigation, G.Z.; methodology, B.N.; project administration, G.Z.; resources, G.Z.; software, B.N.; supervision, G.Z.; validation, B.N.; writing—original draft preparation, B.N.; writing—review and editing, G.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the public availability of the dataset.

Informed Consent Statement: Patient consent was waived due to the public availability of the dataset.

Data Availability Statement: The data presented in this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.5827695, reference number 26. (accessed on 15 February 2025).

Acknowledgments: The authors gratefully acknowledge the use of the Zenodo repository for data sharing.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kanojia, N.; Thapa, K.; Verma, N.; Rani, L.; Sood, P.; Kaur, G.; Dua, K.; Kumar, J. Update on mucoadhesive approaches to target drug delivery in colorectal cancer. *J. Drug Deliv. Sci. Technol.* **2023**, *87*, 104831. [CrossRef]
2. Turner, N. D.; Lupton, J. R. Dietary fiber. *Adv. Nutr.* **2011**, *2*, 151–152. [CrossRef]

3. Chauhan, S.; Harwansh, R. K. Recent advances in nanocarrier systems for ulcerative colitis: A new era of targeted therapy and biomarker integration. *J. Drug Deliv. Sci. Technol.* **2024**, *93*, 105466. [CrossRef]
4. Villanacci, V.; Reggiani-Bonetti, L.; Salviato, T.; Leoncini, G.; Cadei, M.; Albarello, L.; Caputo, A.; Aquilano, M. C.; Battista, S.; Parente, P. Histopathology of IBD Colitis. A practical approach from the pathologists of the Italian Group for the study of the gastrointestinal tract (GIPAD). *Pathologica* **2021**, *113*, 39. [CrossRef]
5. Kumar, R.; Kumbharkar, P.; Vanam, S.; Sharma, S. Medical images classification using deep learning: A survey. *Multimed. Tools Appl.* **2024**, *83*, 19683–19728. [CrossRef]
6. Dai, Z.; Liu, H.; Le, Q. V.; Tan, M. Coatnet: Marrying convolution and attention for all data sizes. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 3965–3977.
7. Nie, Z.; Xu, M.; Wang, Z.; Lu, X.; Song, W. A Review of Application of Deep Learning in Endoscopic Image Processing. *J. Imaging* **2024**, *10*, 275. [CrossRef]
8. Li, M.; Jiang, Y.; Zhang, Y.; Zhu, H. Medical image analysis using deep learning algorithms. *Front. Public Health* **2023**, *11*, 1273253. [CrossRef] [PubMed]
9. Ismail, W. N.; Hassan, M. M.; Alsalamah, H. A.; Fortino, G. CNN-based health model for regular health factors analysis in internet-of-medical things environment. *IEEE Access* **2020**, *8*, 52541–52549. [CrossRef]
10. Li, S.; Liu, Y.; Sui, X.; Chen, C.; Tjio, G.; Ting, D.S.W.; Goh, R.S.M. Multi-instance multi-scale CNN for medical image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019, Proceedings of the 22nd International Conference, Shenzhen, China, 13–17 October 2019*; Proceedings, Part IV 22; Springer: Cham, Switzerland, 2019; pp. 531–539.
11. Muruganantham, P.; Balakrishnan, S. M. A survey on deep learning models for wireless capsule endoscopy image analysis. *Int. J. Cogn. Comput. Eng.* **2021**, *2*, 83–92. [CrossRef]
12. Rustam, F.; Siddique, M. A.; Siddiqui, H. U. R.; Ullah, S.; Mehmood, A.; Ashraf, I.; Choi, G. S. Wireless capsule endoscopy bleeding images classification using CNN based model. *IEEE Access* **2021**, *9*, 33675–33688. [CrossRef]
13. Sharif, M.; Attique Khan, M.; Rashid, M.; Yasmin, M.; Afza, F.; Tanik, U. J. Deep CNN and geometric features-based gastrointestinal tract diseases detection and classification from wireless capsule endoscopy images. *J. Exp. Theor. Artif. Intell.* **2021**, *33*, 577–599. [CrossRef]
14. Islam, K. Recent advances in vision transformer: A survey and outlook of recent work. *arXiv* **2022**, arXiv2203.01536.
15. Xiao, H.; Li, L.; Liu, Q.; Zhu, X.; Zhang, Q. Transformers in medical image segmentation: A review. *Biomed. Signal Process. Control* **2023**, *84*, 104791. [CrossRef]
16. Almalik, F.; Yaqub, M.; Nandakumar, K. Self-ensembling vision transformer (sevit) for robust medical image classification. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Singapore, 18–22 September 2022; pp. 376–386.
17. Hatamizadeh, A.; Tang, Y.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H.R.; Xu, D. Unetr: Transformers for 3d medical image segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2022; pp. 574–584.
18. Wang, W.; Yang, X.; Tang, J. Vision transformer with hybrid shifted windows for gastrointestinal endoscopy image classification. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 4452–4461. [CrossRef]
19. Rashid, A. B.; Kausik, A. K. AI revolutionizing industries worldwide: A comprehensive overview of its diverse applications. *Hybrid Adv.* **2024**, *7*, 100277. [CrossRef]
20. Manzari, O. N.; Ahmadabadi, H.; Kashiani, H.; Shokouhi, S. B.; Ayatollahi, A. MedViT: A robust vision transformer for generalized medical image classification. *Comput. Biol. Med.* **2023**, *157*, 106791. [CrossRef]
21. Turan, M.; Durmus, F. UC-NfNet: Deep learning-enabled assessment of ulcerative colitis from colonoscopy images. *Med. Image Anal.* **2022**, *82*, 102587. [CrossRef]
22. Polat, G.; Kani, H. T.; Ergenc, I.; Ozen Alahdab, Y.; Temizel, A.; Atug, O. Improving the computer-aided estimation of ulcerative colitis severity according to mayo endoscopic score by using regression-based deep learning. *Inflamm. Bowel Dis.* **2023**, *29*, 1431–1439. [CrossRef]
23. Polat, G.; Ergenc, I.; Kani, H. T.; Alahdab, Y. O.; Atug, O.; Temizel, A. Class distance weighted cross-entropy loss for ulcerative colitis severity estimation. In *Annual Conference on Medical Image Understanding and Analysis*; Springer: Berlin/Heidelberg, Germany, 2022; pp 157–171.
24. Lou, M.; Yu, Y. OverLoCK: An Overview-first-Look-Closely-next ConvNet with Context-Mixing Dynamic Kernels. *arXiv* **2025**, arXiv:2502.20087.
25. Tu, Z.; Talebi, H.; Zhang, H.; Yang, F.; Milanfar, P.; Bovik, A.; Li, Y. Maxvit: Multi-axis vision transformer. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 459–479.
26. Polat, G.; Kani, H. T.; Ergenc, I.; Alahdab, Y. O.; Temizel, A.; Atug, O. *Labeled Images for Ulcerative Colitis (LIMUC) Dataset (Version 1)*; Zenodo: Geneva, Switzerland 2022. [CrossRef]

27. Xie, T.; Zhang, T.; Ding, C.; Dai, X.; Li, Y.; Guo, Z.; Wei, Y.; Gong, J.; Zhu, W.; Li, J. Ulcerative Colitis Endoscopic Index of Severity (UCEIS) versus Mayo Endoscopic Score (MES) in guiding the need for colectomy in patients with acute severe colitis. *Gastroenterol. Rep.* **2018**, *6*, 38–44. [CrossRef] [PubMed]
28. Matsuoka, K.; Kobayashi, T.; Ueno, F.; Matsui, T.; Hirai, F.; Inoue, N.; Kato, J.; Kobayashi, K.; Kobayashi, K.; Koganei, K. Evidence-based clinical practice guidelines for inflammatory bowel disease. *J. Gastroenterol.* **2018**, *53*, 305–353. [CrossRef] [PubMed]
29. Vuitton, L.; Peyrin-Biroulet, L.; Colombel, J.; Pariente, B.; Pineton de Chambrun, G.; Walsh, A.; Panes, J.; Travis, S.; Mary, J.; Marteau, P. Defining endoscopic response and remission in ulcerative colitis clinical trials: An international consensus. *Aliment. Pharmacol. Ther.* **2017**, *45*, 801–813. [CrossRef]
30. McHugh, M. L. Interrater reliability: The kappa statistic. *Biochem. Medica* **2012**, *22*, 276–282. [CrossRef]
31. Banerjee, M.; Capozzoli, M.; McSweeney, L.; Sinha, D. Beyond kappa: A review of interrater agreement measures. *Can. J. Stat.* **1999**, *27*, 3–23. [CrossRef]
32. Cohen, J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychol. Bull.* **1968**, *70*, 213. [CrossRef]
33. Cohen, J. A cofficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
34. Woolson, R.F. Wilcoxon signed-rank test. *Encycl. Biostat.* **2005**, *8*.. [CrossRef]
35. Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1567–1578.
36. Sadi, A.A.; Chowdhury, L.; Jahan, N.; Rafi, M.N.S.; Chowdhury, R.; Khan, F.A.; Mohammed, N. LMFLOSS: A hybrid loss for imbalanced medical image classification. *arXiv* **2022**, arXiv:2212.12741.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.