A Hybrid Approach of Vision Transformers and CNNs for Detection of Ulcerative Colitis

Syed Abdullah Shah[1], Imran Taj[2], Syed Muhammad Usman[1], Syed Nehal Hassan Shah[1], Ali Shariq Imran[3,*], and Shehzad Khalid[4]

[1]Department of Creative Technologies, Faculty of Computing and Artificial Intelligence, Air University, Islamabad, 44000, Pakistan
[2]College of Interdisciplinary Studies, Zayed University, Abu Dhabi P.O. Box 144534, United Arab Emirates
[3]Department of Computer Science Norwegian University of Science and Technology Gjøvik, Norway [4]Department of Computer Engineering, Bahria University, Islamabad, 44000, Pakistan
[*]ali.imran@ntnu.no

ABSTRACT

Ulcerative Colitis is an Inflammatory Bowel disease caused by a variety of factors that lead to a serious impact on the quality of life of the patients if left untreated. Due to complexities in the identification procedures of this disease, the treatment timeline and quality can be severely affected, leading to further consequences for the sufferer. The difficulties in identification stem from inter-observer disagreement on the disease status of progression, and due to the uneven distribution of resources and professionals across the globe, dealing with this malady is a serious concern for medical professionals. Machine Learning research can contribute to this gap in professional human resources through preparing architectures by training, testing, and validation due to these ML models being easily scalable, along with constant refinements allowing them to enhance their identification efficiency as well. We contribute to this research by utilizing the TMC-UCM and LIMUC[11] datasets as inputs for a 3-fold Vision Transformer from which we extract features and combine them through averaging fusion for application as input along with our dataset on a CNN model to ensure that the model delivers better results than the ViT and therefore exhibits greater classification abilities. Through the evaluation metric of AUC-ROC, our 3-fold ViT and feature-fused CNN achieve accuracies as high as 85% and 90%, respectively, with the corresponding AUC scores being 0.91, 0.81, 0.94, and 0.94 for the endoscopic classes of Mayo 0, Mayo 1, Mayo 2, and Mayo 3 respectively.

Introduction

Ulcerative Colitis is an Inflammatory Bowel Disease (IBD) occurring in individuals due to a combination of genetic and environmental factors along with abnormal reactions from the immune system. It is divided into multiple stages of progression, which are further grouped into scoring systems like MES and UCEIS: benign, mild, moderate, and severe. At the later stage of progression, the disease severely impacts the quality of life to the extent that full-time medical care may be necessitated for treatment. The detection of this disease is done through a variety of tests like imaging, biopsy, blood tests, and stool tests. Many forms of detection are not the optimal choices for this purpose, e.g., CT-Scan's non-invasiveness limits its capabilities for Ulcerative Colitis and might even promote tumor growth and cause irritation on the lesions while the results might be obscured by any presence of higher gas contents in the colon[2]. Endoscopy is generally considered the best possible way as it provides a visual input regarding the patient's current situation and even histological testing can be carried out through the samples extracted in this process[3].

The treatment of this disease is complicated by difficulties in the correct identification of its symptoms in affected regions, and therefore, inter-observer disagreements are common, which can lead to delays in the actual treatment[4]. Researchers proposed none invasive method by draw out disease activity from the samples from genes to detect inherited capability of disease[5] or by using a combination of symp bio-scopic solutions estimating neutrophils detection to uncover the activity of the disease through the Histogrammical test[67] by gene testing, pooling from research driven sampling to identify important genomes with high coherent relations to the disease's existence by deriving and predicting results from clinical testing and variables related to patient habits to detect activity and severity[8], and Endoscopic studies for disease detection through image classification on the basis of the MES or UCEIS classification system for UC to detect progression of disease and relapsing through standard monitoring instruments[910111241314] for innovation such as the camera capsule technology[15].

On Observation, it is understood that various symptoms of Ulcerative Colitis seem to share similar symptoms with CD (Crohn's disease) with the observed difference among the diseases being the area of affliction, as declared by

Davidson Principal and Practice of Medicine[3], which is the colon or large intestine for UC and smaller intestine for CD. Several studies have been applied involving histological findings[1617] and genes sample for biomarker identification[18]. Similarly, our search led us to the classification of all 23 classes of IBD disease using endoscopic images by Machine learning models, which not only classified the activity of UC but all the other labels of IBD in the HyperKavsir dataset[19]

The Aim of our study is to shift our attention to the application of Endoscopic images with high performing Machine Learning models to promote efficient classification of disease state of progress within an afflicted individual to assist professionals in the domain of Gastroenterology, so professionals can attain sufficient tool to help divert their attention to driving a better course of recover for those who are afflicted with Ulcerative Colitis. The objectives of this research include:

- To mitigate the imbalanced dataset problem, strategies will be designed to better deal with class imbalance and biases in severity classification.
- Exploration of pre-processing and feature extraction models for deriving efficient features.
- To test the developed model on the applied available dataset for better generalization of the model for more practical use purposes.
- To develop a state-of-the-art model for accurate classification of severity growth with satisfactory sensitivity, specificity and high precision.

Figure 1. Preprocessing methods proposed by researchers in recent years

Related Work

Experienced surgeons and gastroenterologists are required to identify different stages of Ulcerative Colitis using various methods such as biopsy, colonoscopy, and sigmoidoscopy. However, such experienced and qualified medical professionals are not available across all geographic regions of the world. Therefore, researchers have been trying to implement machine learning methods for automated UC detection and identification so that an easily scalable and resource-intensive solution can be designed to deal with this problem. Different research has been conducted on the basis of varying methodologies proposing unique ways of tackling these difficulties, and we, too, propose a different methodology that serves to widen the available strategies that can be utilized for further experimentation.

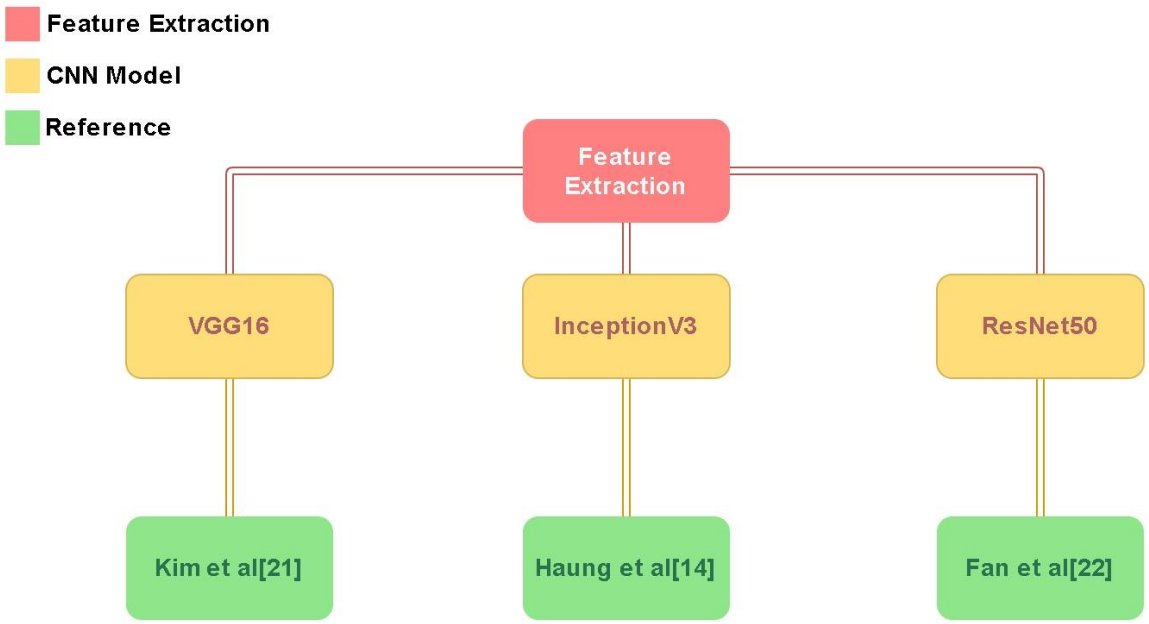


Figure 2. Feature extraction methods proposed by researchers in recent years

Figure 1 shows multiple techniques of preprocessing proposed by researchers for automated detection of Ulcerative Colitis. Cropping was utilized by Wang[4], Becker[11], J. E. Kim[12] and Ozawa[20] while resizing was utilized in Wang[4], J. E. Kim[12] Stidham[13], Huang[14], Ozawa[20] and J. H. Kim[21]. Other than Fan[22], all of the rest applied augmentation, while Kim et al[12] used normalization. Filtering was employed in Gottlieb[9] while manual frame selection was used in Becker[11], Ozawa[20] and J. H. Kim[21]. J. E. Kim[12] also chose grayscale conversion and binarization, while Becker[11] used text removal and field of view masking for their preprocessing stage. J. H. Kim[21] employed RGB to HSV conversion, Ozawa[20] used annotation, Huang[14] used pixel value rescaling, and J. E. Kim[12] used gray scaling. Gottlieb[9] used abnormality extraction and sequential processing as well at this stage of experimentation. Becker[11] and J. H. Kim[21] also used color section extraction. Finally rotation was employed by Becker[11], J. E. Kim[12], Stidham[13] and Ozawa[20].

Figure 2 presents methods used for feature extraction which include VGG16 encoder proposed by Kim et al[12], Huang[14] choosing InceptionV3, and Fan[22] employing a ResNet50 for this purpose. In contrast to the scarce use of feature extraction, a great variety of machine learning architectures and accompanying methods were employed by these researchers to contribute to the latest developments in the UC imaging domain. Huang[14] utilized the k-nearest Neighbours, Support Vector Machine, and a Deep Neural Network along with the ensemble learning for their classification process while J. H. Kim[21] employed MobileNetV2, MobileNetV3Large and Xception.

Figure 3 shows multiple methods for classification proposed by researchers in recent years. HRnet was used by Wang[4] along with ResNeXt, while Quality Control and Ulcerative Colitis Scoring models were utilized by Becker[11]. EfficientNet was used by Wang[4] and J. H. Kim[21], while DenseNet was chosen by Wang[4], and J. H. Kim[21]. ResNet family models were employed by Wang[4], Becker[11],Kim et al[21], though Gottlieb[9] utilized RNN. J. E. Kim[12] and J.

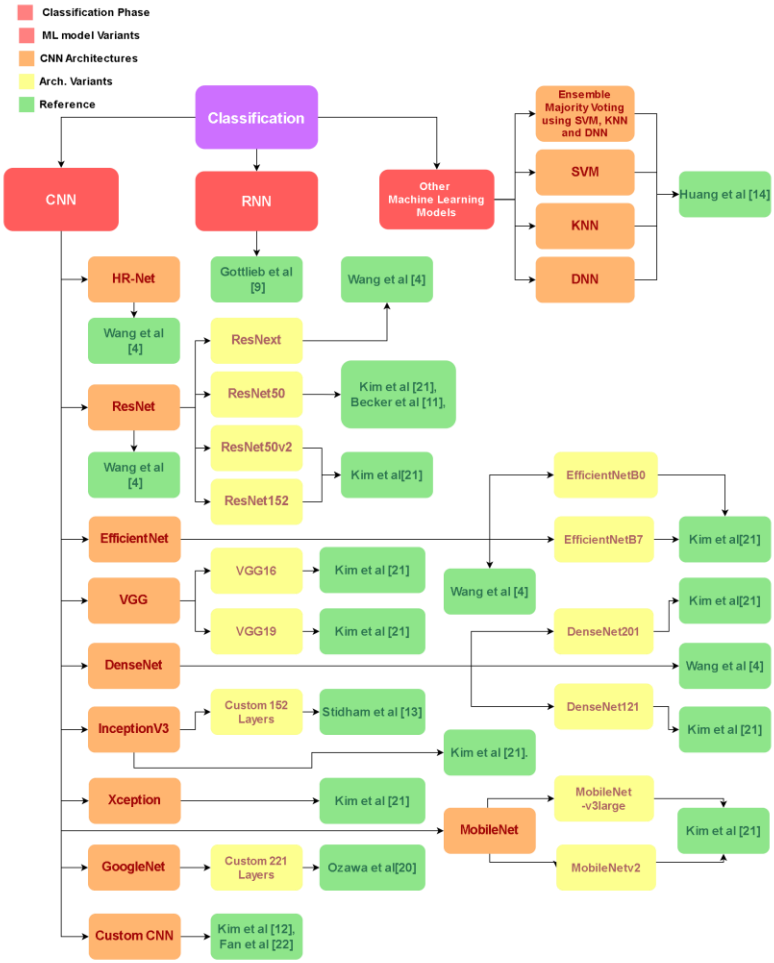


Figure 3. Classification methods proposed by researchers in recent years

H. Kim[21] used VGG16, while Fan[22] used a custom CNN for classification. GoogleNet was used by Ozawa[20] while VGG19 was employed by Kim et al[21]. Finally InceptionV3 was utilized by Stidham[13] and J. H. Kim[21].

After compilation and summarizing of the attained research material, It has been observed that in the scope of this study, several authors that have worked in this domain had undergone some common challenges that still hinder the progress of attainment of piratical resources for the actual development of useful Computer-Aided Diagnosis tools. Due to the existence of the following research gaps in this study as:

- Lack of effective pre-processing techniques in current studies.
- Failure to address class imbalance issue, impacting prediction method performance.
- Defining a suitable evaluation metric for classification purposes of ulcerative colitis, considering the compromise between accuracy, precision, and mean Average Precision.
- Need for external validation on datasets for generalization and bias mitigation.

Proposed Methodology:

The proposed method aims to address a number of problems, including class imbalance, with increased performance measures. Figure 4 shows the flow diagram of the proposed method. It consists of data preprocessing, followed by feature extraction of a customized Vision Transformer (ViT) and then fusion of features and classification.

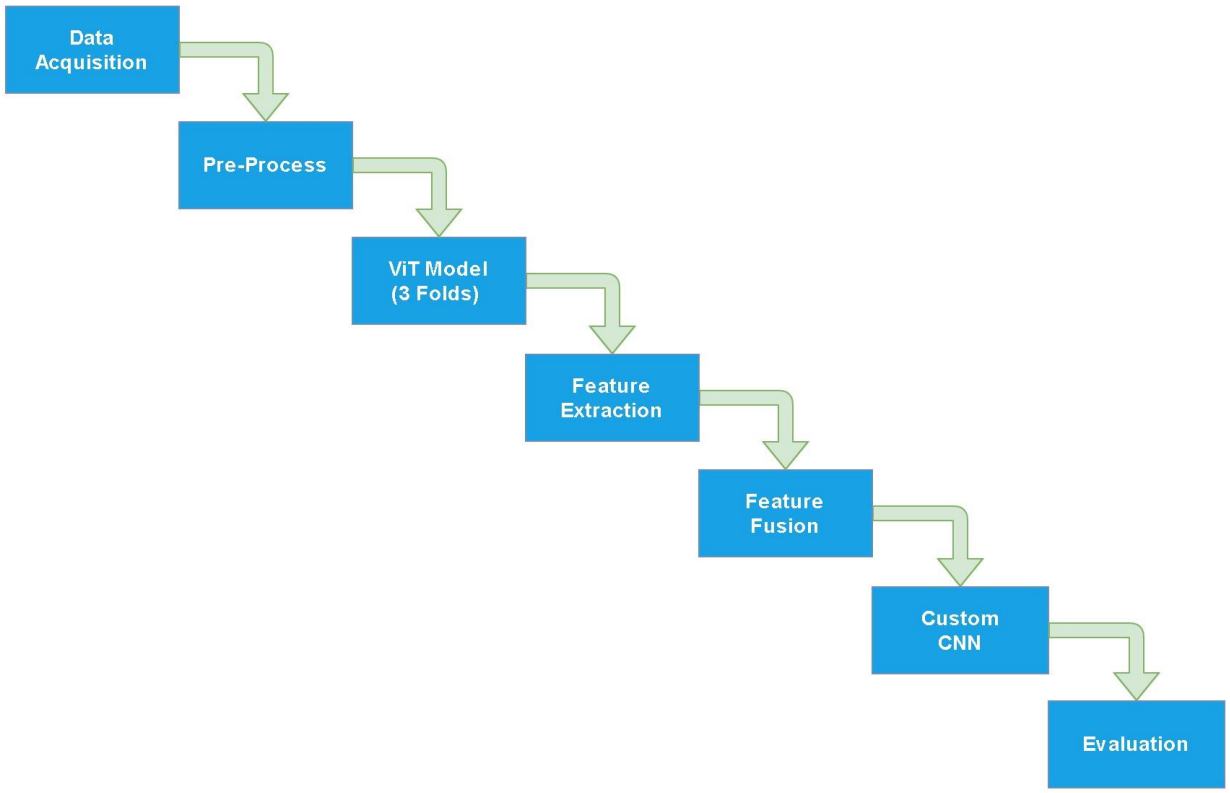


Figure 4. Figure illustrates overflow of the entire applied pipeline of the methodology.

Pre-Processing

Data Standardization was applied so that the differences in scales or units among the different datasets being employed were eliminated to implement uniformity of a single standard. In this way the input data is optimized for further usage so that the increased model performance is ensured. Figure 5 shows the proposed preprocessing steps. While the original format of the datasets was already our desired JPEG, the resolution of images for the two datasets, LIMUC[1] and TMC-UCM, were 352x288 and 300x300 which were standardized to 300x300. Afterwards, normalization was applied to mitigate the pixel value fluctuation from 0 and 255 to 0 and 1. Then to deal with the problem of class imbalance in the input data, we avoided employing oversampling and simple augmentation due to the former tending to over-represent the minority classes across the entire data space leading to poor generalization and overfitting while the latter tends to apply transformations on the basis of perspective, orientation, color, scale, brightness, contrast, etc. across the entire data as well without considering the presence of certain areas in data space where the minority class might be more or less concentrated. Both of these techniques do not adequately address the class imbalance problems, and therefore, we utilize the High-Frequency Balancing and Augmentation technique, which specifically targets areas where minority classes are present in higher numbers. By compensating for class imbalance only in those areas where required, this technique not only deals with the problem of class imbalance. To reduce over-fitting to also introduce a dropout layer. simultaneously so that important patterns and features of the minority classes are also accounted for by the model.

Feature Extraction and Classification

We propose a 3-fold Vision Transformer where the usage of 3 folds allows us to prepare three iterations of the ViT with the data subsets being rearranged each time. Figure 6 shows the architecture of the proposed ViT. This approach ensures a holistic

analysis of the different features in different regions of the dataset. The Vision Transformer architecture consists of a Patch extractor, Token Embedder, transformer encoder, Self-attention mechanism, and MLP head[23]. The Patch Extractor divides images into patches, which are then transformed into a lower-dimensional

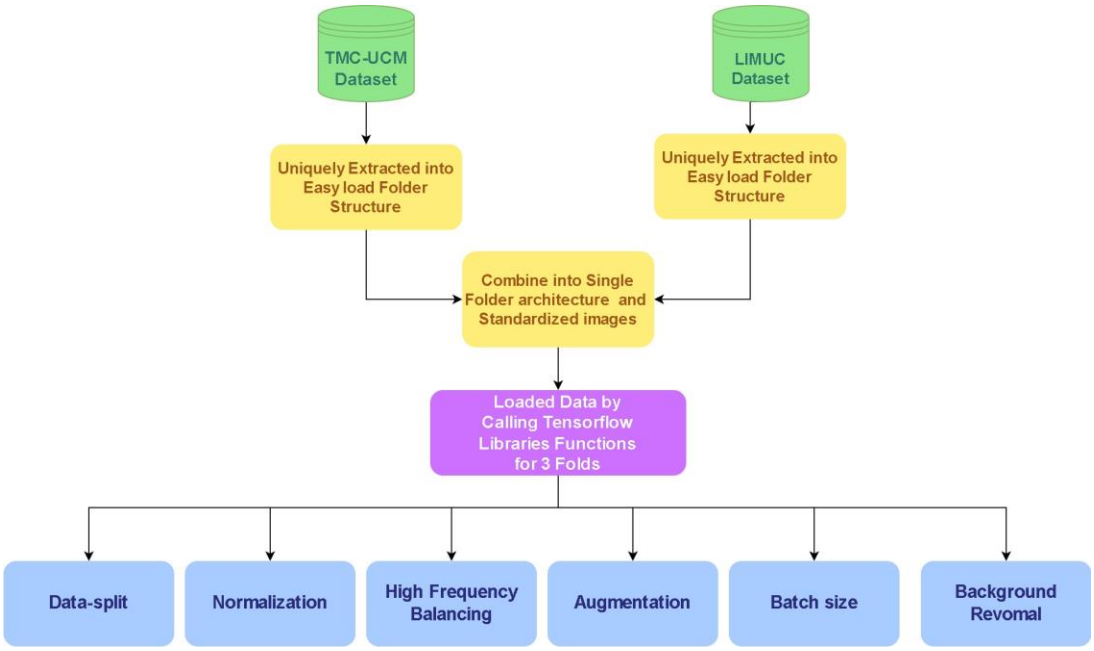


Figure 5. Figure illustrates the involved the general structuring and Pre-processing applied.

vector space using a linear transformation[24]. This process can be mathematically expressed as:

$$\mathbf{x}_p = \mathbf{W}_p \cdot \mathbf{x} + \mathbf{b}_p \tag{1}$$

where, $\mathbf{x}_p$ is the vector representation of the image patch, $\mathbf{W}_p$ is the weight matrix for the linear transformation, $\mathbf{x}$ is the original image patch and $\mathbf{b}_p$ is the bias term. The Token Embedder embeds a token within the sequence of patches to represent the entire image. This embedding process can be described by[25]:

$$\mathbf{z}_0 = [\mathbf{x}_{class}; \mathbf{x}_{p1}; \mathbf{x}_{p2}; \ldots; \mathbf{x}_{pN}] + \mathbf{E}_{pos} \tag{2}$$

where, $\mathbf{z}_0$ is the initial embedded sequence, $\mathbf{x}_{class}$ is the class token, $\mathbf{x}_p{}^i$ are the patch vectors and $\mathbf{E}_{pos}$ is the positional encoding. The Transformer Encoder consists of the Self-Attention Mechanism and the Feed Forward Neural Network. The Self-Attention Mechanism points out the dependencies among patches and annotates their significance:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{3}$$

where, $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are the query, key, and value matrices and d_k is the dimension of the key vectors[26]. The Feed Forward Neural Network captures complex relations within each patch:

$$\text{FFN}(\mathbf{z}) = \max(0, \mathbf{z}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2 \tag{4}$$

where, $\mathbf{z}$ is the input from the self-attention layer, $\mathbf{W}_1$ and $\mathbf{W}_2$ are weight matrices, $\mathbf{b}_1$ and $\mathbf{b}_2$ are bias terms. Finally, the MLP head undertakes classification and object detection:

$$\hat{y} = \mathrm{MLP}(\mathbf{z}_L) \tag{5}$$

where, $\hat{y}$ is the output prediction and $\mathbf{z}_L$ is the output of the final transformer encoder layer. The MLP processes the patches through a series of connected layers containing non-linear activation functions like ReLU to perform higher classification and feature extraction[2728].

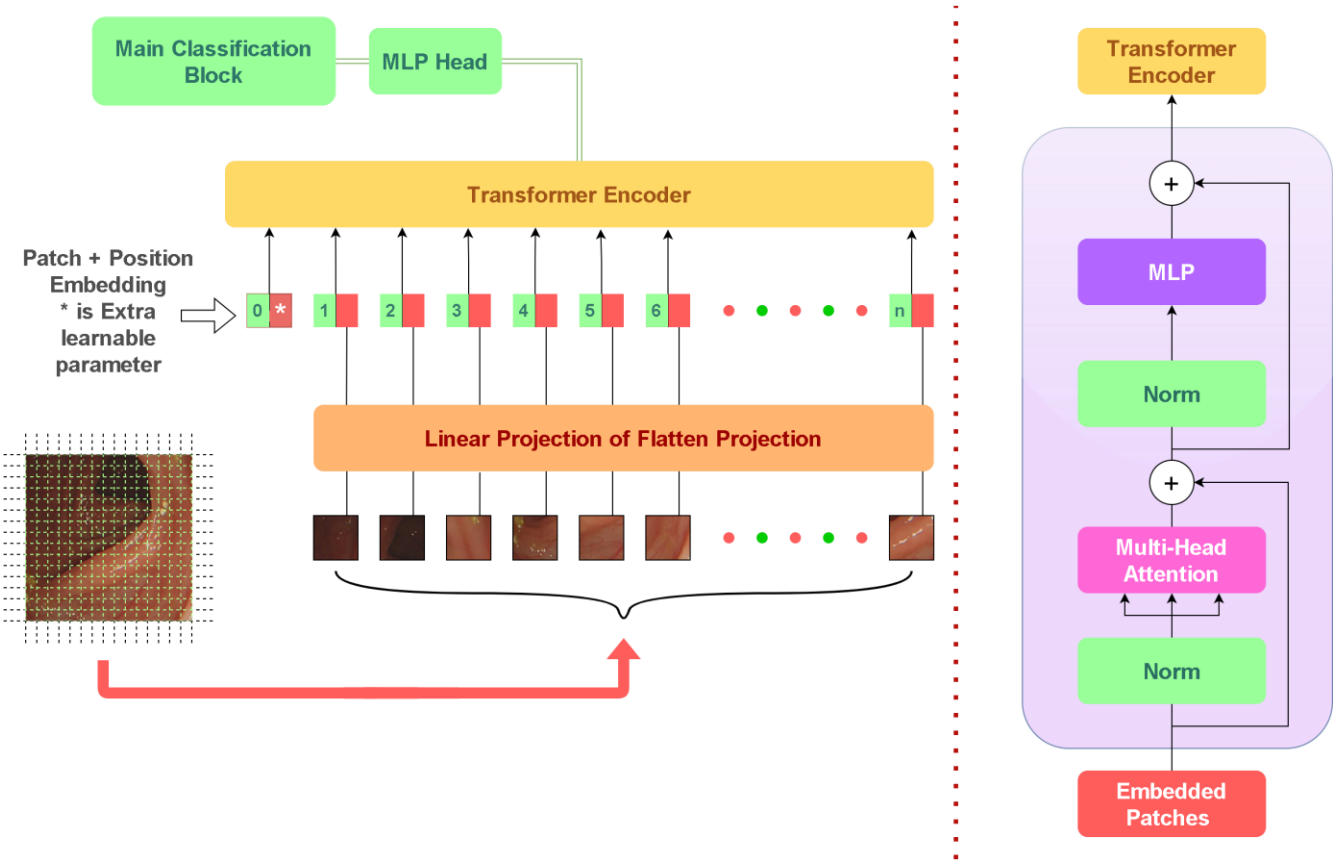


Figure 6. Figure Illustrates the overall flow of the applied ViT Architecture on UC Dataset.

With results from the 3-fold ViT acquired, we extract the penultimate layer of each ViT fold to acquire features[2925]. These 3 different feature vectors are then fused through the usage of the averaging fusion technique because fusion allows the unification of multiple representations into a single feature map, along with images of the combined datasets are fed to improve generalization[3031], Custom CNN is applied to get enhanced and more refined feature vector. This features vector has enhanced capability to detect images that were used for the training[3130]. *CNN Model Architecture and Methods*

After having prepared the input dataset along with the fused feature map from the 3-fold Vision Transformer, we will utilize these as input for a CNN model composed of a number of layers with specific functions to be performed.

The first layer is a Conv2D layer where kernels convolve across the dataset to create feature maps. This layer focuses on the detection of local patterns and spatial relationships while employing the activation function of ReLU to introduce non-linearity and diminish the impact of irrelevant features and the vanishing gradient, so that the model can learn higher-level features. Mathematically, the convolution operation can be expressed as:

$$\mathrm{Conv2D}(\mathbf{x}) = \mathrm{ReLU}(\mathbf{W}_c * \mathbf{x} + \mathbf{b}_c) \tag{6}$$

where, $\mathbf{W}_c$ are the convolutional kernels, $\mathbf{x}$ is the input feature map, $\mathbf{b}_c$ is the bias term, $*$ denotes the convolution operation.

The Batch Normalization technique optimizes the stability and speed of the network so that deeper layers can be utilized. It works by increasing the convergence by reducing the mean and variance of each layer while utilizing parameters like shift and scale that allow adaptation to specific data features. This can be expressed as:

$$\mathrm{BN}(\mathbf{x}) = \gamma \frac{\mathbf{x}-\mu}{\sqrt{\sigma^2+\epsilon}} + \beta \tag{7}$$

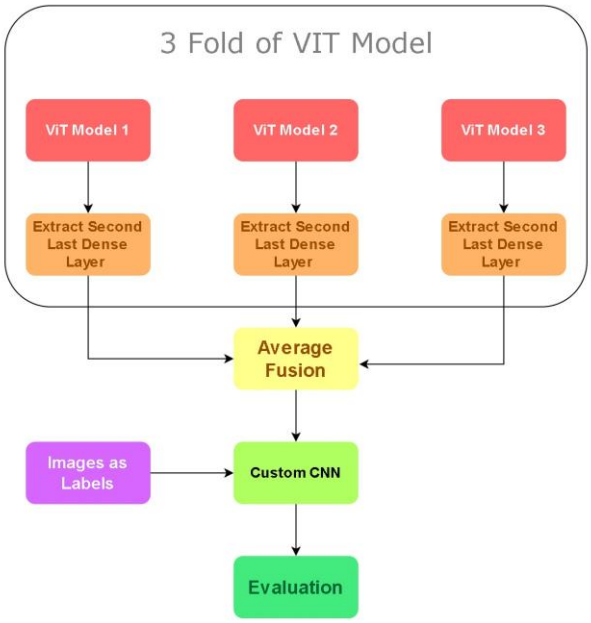


Figure 7. Averaging Fusion on the ViT folds.

where, μ and σ^2 are the mean and variance of the input, γ and β are the learnable scale and shift parameters and ϵ is a small constant for numerical stability.

The Flatten layer reduces the dimensionality of the input to linearity through concatenation, which allows focusing of all dimensions onto a single axis. This is typically used when transitioning from convolutional layers to fully connected layers. It can be described as:

$$\mathrm{Flatten}(\mathbf{x}) = \mathrm{reshape}(\mathbf{x}, [-1]) \tag{8}$$

where the input $\mathbf{x}$ is reshaped into a single-dimensional vector.

Figure 7 shows the proposed method for fusion of ViTs, whereas, Figure 8 shows the proposed feature fusion of CNNs. The fully connected layer, referred to as the Dense Layer, performs a linear operation followed by a non-linear activation function such as ReLU, GeLU, tanh, etc. This layer performs higher-level feature extraction and classification by utilizing all the information from previous layers to make predictions. The operation can be described as:

$$\mathrm{Dense}(\mathbf{x}) = \sigma(\mathbf{W}_d\mathbf{x}+\mathbf{b}_d) \tag{9}$$

where, $\mathbf{W}_d$ is the weight matrix, $\mathbf{x}$ is the input vector, $\mathbf{b}_d$ is the bias term and σ is the activation function.

Finally, a regularization technique referred to as the Dropout Layer is utilized to prevent overfitting. This layer introduces redundancy into the network through random deactivation of a small fraction of the total neurons in a layer during each fold. This allows the model to be better generalized by reducing its dependency on certain neurons for analysis, thus forcing the model to avoid memorizing the data too closely. The number of neurons to be deactivated each time is completely determinable by the user. The dropout operation can be represented as:

$$\mathrm{Dropout}(\mathbf{x}) = \mathbf{x} \odot \mathbf{r} \tag{10}$$

whee, $\odot$ denotes element-wise multiplication and $\mathbf{r}$ is a binary mask vector with a certain probability of zero entries.

Results

Before we move on to the performance evaluation of the Vision Transformer and the CNN employed in this research, we will elaborate on the datasets that were acquired for utilization in this experiment.

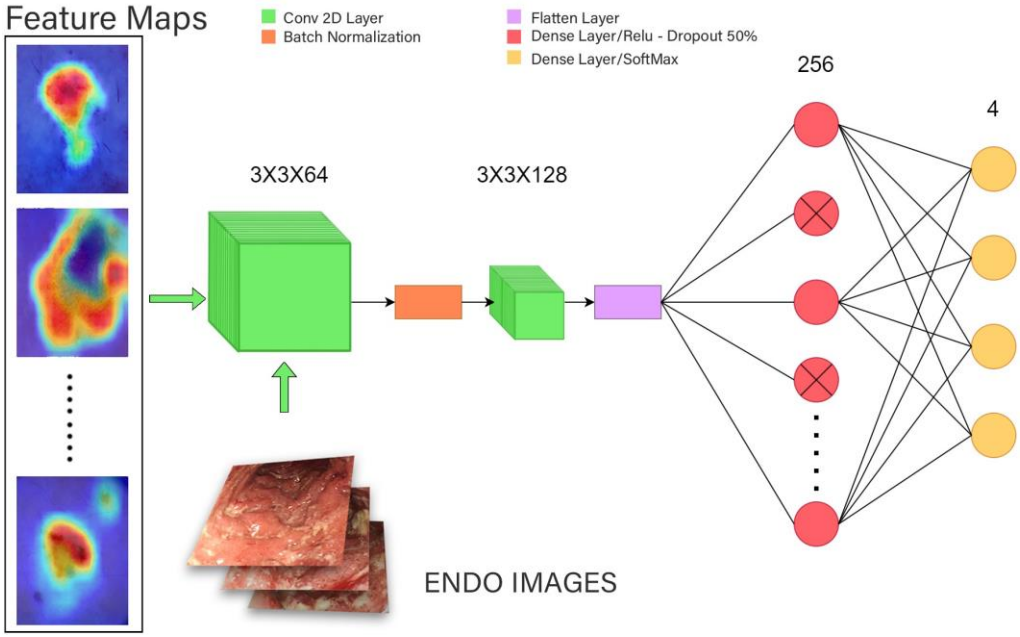


Figure 8. Architecture of Custom CNN with Feature Fusion.

Datasets

The two datasets LIMUC[1] and TMC-UCM have been used in this research.

LIMUC Dataset

Labelled Images for Ulcerative Colitis[1] (LIMUC)dataset was compiled by the Marmara University School of Medicine's Department of Gastroenterology using 11,276 Ulcerative Colitis images of size 352x288. Having acquired these diagnoses from 564 patients through 1043 colonoscopy exams carried out from December 20111 to July 2019, the dataset was assessed by two gastroenterology experts who classified using the Mayo Endoscopic Scoring system. Then, a third-party professional independently assessed the data and assigned Mayo scores without any knowledge of the prior assessment, with the final Mayo score being assigned through a majority vote system.

TMC-UCM Dataset

Originating from the Tongji Medical College of Huazhong University of Science and Technology in Wuhan, China, this publicly available dataset is derived from the tests performed on 308 patients with Ulcerative Colitis[4], who underwent colonoscopy based on IBD guideline between January 2014 and December 2021. It contains 12,163 images obtained through the Olympus colonoscope, which, when filtered to exclude images with stool, blur, or halation, is reduced to 7,978 images. This dataset was reviewed by at least three expert gastroenterologists with sufficient experience in tackling IBDs, who resolved any judgmental differences through discussion, and even then, their final classification decision on the basis of MES was further reviewed by another IBD specialist.

Associated with the diagnosis of Ulcerative Colitis are the different diagnosis systems used to identify the stage of progression of the disease, such as UCEIS and MES. In our research, we utilize the Mayo Endoscopic Scoring system, which assigns four categories, ranging from 0 to 3, to the different disease progression stages. Mayo 0 indicates no inflammation with the surface being smooth and healthy with a normal mucosa. Mayo 1 represents a mild stage of inflammation where the mucosa might be irregular or granular, though no erosions or ulcers can be found. Further signs can be mild erythema and friability. Mayo 2 symbolizes moderate inflammation, where limited ulceration is identifiable, along with more pronounced signs like erosion, friability, and marked erythema. The final Mayo 3 stage is potentially fatal due to severe inflammation marked by extensive ulcerations and spontaneous bleeding. Furthermore, the presence of a pseudo-

polypoid is also possible. Table 1 presents the detailed description of datasets used in this study, whereas, Figure 9 shows the sample images of the dataset.

Moving on to the methods of evaluating the performance of our models, we decided to utilize the AUC (Area Under the Curve) score due to this specific metric being a suitable way to display the varied discriminating thresholds regarding the rarity of different stages of diseases in the medical field by using the graphical plot of the Receiver

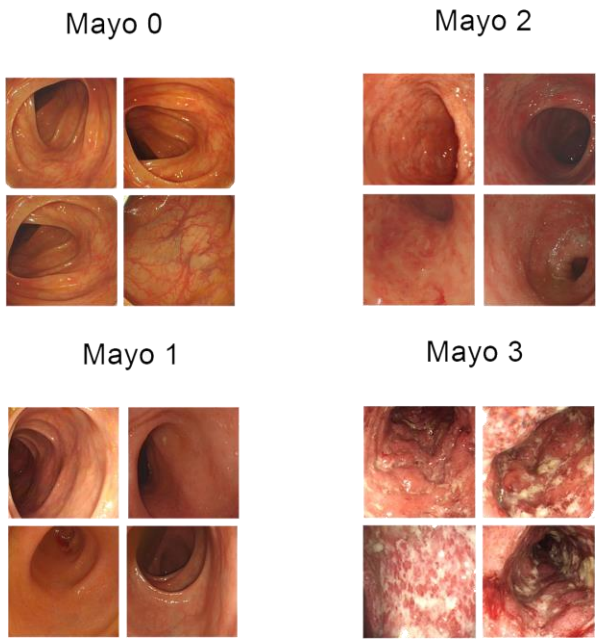


Figure 9. Figure illustrates the visual Aspects of the applied Images.

Table 1. Quantitative attributes of each dataset with respect to the details and dimension of the involved labels

Dataset Name	Total_images	Mayo0	Mayo1	Mayo2	Mayo3	Dimensions
LIMUC	11,276	6105	3052	1254	865	352*288
TMC-UCM	7,978	3031	1835	1675	1423	300*300

Operating Characteristic (ROC). These variations in discrimination thresholds are simply not possible with simple accuracy, and therefore, we forego the simple accuracy metric in favor of the more suitable AUC-ROC. A detailed description of our attained results in terms of AUC-ROC, precision, recall, and f1-score is given below.

Performance Measures

Moving on to the methods of evaluating the performance of our models, we decided to utilize the AUC (Area Under the Curve) score due to this specific metric being a suitable way to display the varied discriminating thresholds regarding the rarity of different stages of diseases in the medical field by using the graphical plot of the Receiver Operating Characteristic (ROC). These variations in discrimination thresholds are simply not possible with simple accuracy, and therefore, we forego the simple accuracy metric in favor of the more suitable AUC-ROC. A detailed description of our attained results is given below.

ROC refers to the Receiver Operating Characteristic, which is a graphical plot drawn on the basis of the discriminating thresholds being varied so that the trade-offs between sensitivity and precision can be displayed. It is derived through the calculation of specificity and sensitivity.

$$ROC = \{(Specificity_i, Precision_i)\}^{n}_{i=0} \quad (11)$$

Specificity measures the proportion of true negatives that are correctly identified by the neural network architecture. It is calculated through the following formula:

$$\text{Specificity} = \frac{TN}{FP+TN} \quad (12)$$

Where:

- FP = False Positives
- TN = True Negatives

Precision refers to the proportion of true positives that are correctly identified by the neural network. Its calculation is carried out through the following process:

$$\text{Sensitivity} = \frac{TP}{FP+TP} \quad (13)$$

Where:

- TP = True Positives
- FP = False Positives

AUC refers to Area Under the ROC Curve, which is a singular value between 0 and 1 that represents the overall efficiency of the classifier.

$$AUC = \int_0^{Z_n} \text{Sensitivity}(\text{Specificity}) d\text{Specificity} \quad (14)$$

There are a multitude of reasons for not using simple accuracy. Firstly, medical datasets can suffer from class imbalance due to some conditions being rarer than others, thus generating a difference in the weight of different severity classes. Therefore, a model achieving high accuracy could simply be doing so through bias towards the majority class. Furthermore, simple accuracy is unable to account for the different types of errors i.e., false positives and false negatives. Simple accuracy is also limited to a simple fixed threshold of 0.5, which is not optimal for the multiple severity levels inherent in disease detection. Accuracy is thus not a comprehensively suitable metric for our research domain.

In contrast, AUC-ROC addresses the class imbalance by considering all possible thresholds of classification, which leads to it being less affected by any imbalance if present, and therefore provides for a more robust measurement of effectiveness. It also captures two critical aspects of diagnosis through a single measure: true positives and true negatives or sensitivity and precision. This means that AUC-ROC can declare both a diseased state when there is a disease and a non-diseased state when there is none. In the medical field, where false positives and false negatives can constitute serious consequences, AUC-ROC is a better measure due to balancing these considerations of clinical requirements or the costs associated with diagnostic errors through adjustment to the thresholds of classification probability. The ROC curve provides a comprehensive overview regarding the trade-offs involved in the binary classification methods by plotting the True Positive rate against the False Positive rate.

Recall is a metric of measure which involves the identification of correct samples are identified from the overall dataset by the performing model under study it is described by the following formula:

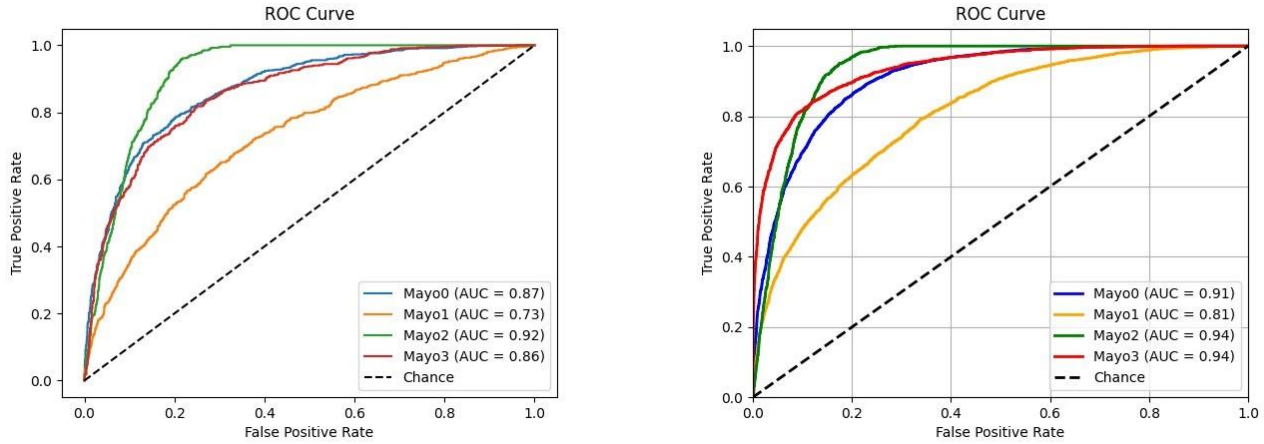
$$R = \frac{TP}{TP+FN} \quad (15)$$

F-1 score another metric of measure used in machine learning It serves as a hormonanic mean between precision and recall and is described by the following formula:.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall} \quad (16)$$

Experimental Results

As is noticeable from the ROC curve graph given above, we can see that the AUC scores of the different classes of Ulcerative Colitis attained by the best fold of the 3-fold ViT model are 0.87 (Mayo 0), 0.73 (Mayo 1), 0.92 (Mayo 2), and 0.86 (Mayo 3). In contrast to these results are those of the feature-fused CNN, which provided scores of 0.91, 0.81, 0.94, and 0.94 for the Mayo 0, 1, 2, and 3 categories, respectively. Then we utilize the following formula to calculate the average AUC-ROC score:



(a) Figure illustrates the attained ROC curve of each class label by Custom CNN.

(b) Figure illustrates the attained ROC curve of each class label by ViT.

Figure 10. Compariosn of ROC curves for multiple classes using ViT and Custom CNN

Figure 10 presents the ROC curves of ViT and Custom CNN and it shows that the scores for different classes of Ulcerative Colitis are attained by the best fold of the 3-fold ViT model are 0.87 (Mayo 0), 0.73 (Mayo 1), 0.92 (Mayo 2), and 0.86 (Mayo 3). In contrast to these results are those of the feature-fused CNN, which provided scores of 0.91, 0.81, 0.94, and 0.94 for the Mayo 0, 1, 2, and 3 categories, respectively. It is noticeable that the performance of the CNN increases the correction rate in all the Mayo classes significantly. Furthermore, the overall performance of the CNN model displays an increase from the ViT's 85 to the CNN's 90. Table 2 compares the results acheived on different CNNs models with ViT and custom CNN, whereas, Table 3 shows the comparison of accuracy and ROC values for multiple variants of EfficientNet.

Table 2. Compariosn of results achieved using different variants of CNNs and ViT

Model	AUC-ROC	Precision				F1-Score				Recall			
		M0	M1	M2	M3	M0	M1	M2	M3	M0	M1	M2	M3
VGG-16	0.593	0.67	0.50	0.43	0.46	0.54	0.25	0.42	0.58	0.47	0.23	0.39	0.70
VGG-22	0.599	0.69	0.59	0.72	0.49	0.54	0.25	0.50	0.60	0.47	0.23	0.48	0.71
Res-Net 50	0.5282	0.58	0.43	0.46	0.46	0.55	0.36	0.45	0.46	0.43	0.33	0.45	0.46
ResNet 101	0.53	0.59	0.42	0.48	0.46	0.56	0.35	0.46	0.46	0.44	0.32	0.46	0.46
ViT	0.85	0.87	0.73	0.91	0.85	0.87	0.73	0.91	0.85	0.85	0.84	0.87	0.86
Custom CNN	0.90	0.87	0.73	0.91	0.85	0.85	0.84	0.87	0.86	0.87	0.73	0.91	0.85

Table 3. Comaprison of AUC-ROC and accuracy of different variants of EfficientNet

Performance Metric	Efficient Net b-2	Efficient Net b-3	Efficient Net b-4	Efficient Net b-5	Efficient Net b-6	Efficient Net b-7
Accuracy	0.52	0.52	0.51	0.52	0.54	0.53
AUC-roc	0.56	0.56	0.54	0.56	0.60	0.67

It is noticeable that the performance of the CNN increases the correction rate in all the Mayo classes significantly. Furthermore, the overall performance of the CNN model displays an increase from the ViT's 85% to the CNN's 90%. Figure 11 presents the confusion matrix to show the performance of the Custom CNN model attained by averaging fusion and illustrates the True Positives, False Positives, True Negatives, and False Negative prediction of labels for each label class in the study.

The advent of Vision Transformers (ViTs) has introduced numerous advantages over Convolutional Neural Networks (CNNs), especially regarding size limitations and the ability to grasp overall features for accurate

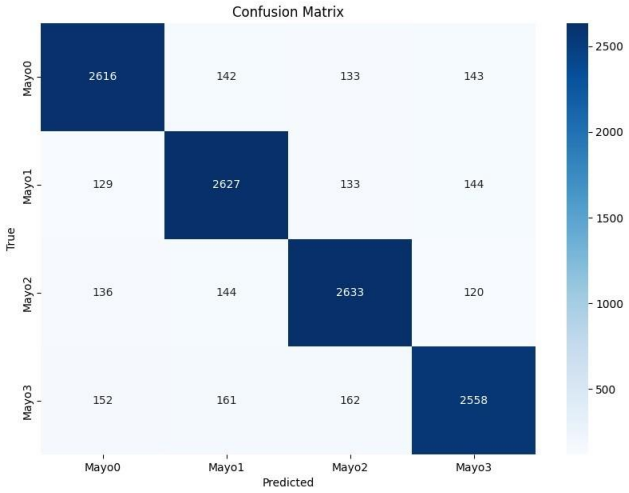


Figure 11. Figure Illustrates the confusion Matrix attained by Custom CNN model.

classification[29]. ViTs utilize self-attention mechanisms to directly capture global context, allowing them to uncover and link relationships between distant patches more effectively. This global perspective enables ViTs to generalize more efficiently on smaller datasets compared to CNNs[27], which depend on local feature extraction and often require extensive data augmentation and large datasets for similar performance. Furthermore, ViTs can leverage pre-training on large, diverse datasets and fine-tuning on smaller, task-specific datasets, enhancing their adaptability and efficiency in data-constrained scenarios.

Through our experimentation, we uncovered that even if we utilized strategies that had been established by previous research, such as attaining optimized overall accuracy and AUC, the CNN models still consistently displayed poor results. Not only the overall accuracy but also the AUC metric also remained disappointingly low for the various architectures, leading to not-so-satisfactory results being obtained through their performance. Amidst these impediments that threatened to detail the course of experimentation, our research attempts with the Vision Transformer showed promising results through its 85% accuracy and a high AUC-ROC score. Though the exact factors that obstructed the performance of the failed CNN models require further investigation, we can discern that the unique characteristics of the ViT architecture must have played a crucial role in bypassing these impediment factors. Therefore, for greater accuracy, reliability, and robustness, ViT must be utilized in the future due to it possibly being the key to uncovering crucial insights into UC that could contribute greatly to speedy diagnosis and planning targeted recovery plans.

As is noticeable from the results, our work represents an advancement regarding our objective of delivering a practical, realistic, and robust tool to medical practitioners for the correct multistage classification of Ulcerative Colitis severity on the basis of the Mayo scoring system, from 0 to 3, through the usage of endoscopic data.

Using state-of-the-art architectures like Vision Transformer (ViT) classifier and custom Convolutional Neural Network (CNN) feature fusion, promising results have been acquired, as evidenced by the high accuracy in efficient categorization of disease severity.

For practicality, as uncovered from Becker Paper implication of noisy labels helps optimize the performance of the model for practical situations, however a key limitation of the study was the limitation of the bias of the introduced labels as of implication of automatics means but with the availability of LIMUC dataset upon inspection several clearly identified labels had an amount of noise inside the images due presence of assisting diagnosis tools in the images using these images could help the performance of the model where the clearer images in quality of disease could highlight light common key features to capture from images and promote for more practical classification capability or model robustness. However, the shortcoming of our research is the lack of external validity, which is a crucial component to prepare for the deployment of

machine learning models into real-world scenarios. External Validation works by testing the model on hitherto unknown data so as to gain an idea of the generalization capabilities of the model to assess its robustness.

Conclusion

Our methodology of utilizing a 3-fold Vision Transformer (ViT) for feature extraction and then using averaging fusion to combine the three different results into a single feature map of considerable generalization that is afterward used as the input along with the acquired dataset into the CNN yields improved results in terms of classification efficiency. The ViT model by itself provides an accuracy of 85% with the corresponding scores being Mayo0 (0.8828), Mayo1 (0.7431), Mayo2 (0.9327), and Mayo3 (0.8847), but the usage of the feature map along with the dataset for the CNN provides 90% accuracy with the scores of the four categories being Mayo0 (0.91), Mayo1 (0.81), Mayo2 (0.94), and Mayo3 (0.94). which have been applied to the attained open source dataset of LIMUC and TMC-UCM by the implementation of optimal data balance and pre-processing. In the future, a multimodal approach could also prove productive for a superior understanding of UC and may help discover novel biomarkers that could contribute to a better therapeutic response. We can also deploy our developed model in real-world medical settings so that validation studies under the supervision of gastroenterologists and endoscopists can be carried out.

References

1. Polat, G. *et al.* Class distance weighted cross-entropy loss for ulcerative colitis severity estimation. In *Annual Conference on Medical Image Understanding and Analysis*, 157–171 (Springer, 2022).
2. Jiang, F., Fu, X., Kuang, K., Fan, D. *et al.* Artificial intelligence algorithm-based differential diagnosis of crohns disease and ulcerative colitis by ct image. *Comput. Math. Methods Medicine* **2022** (2022).
3. Penman, I. D., Ralston, S. H., Strachan, M. W. & Hobson, R. *Davidson's Principles and Practice of Medicine E-Book: Davidson's Principles and Practice of Medicine E-Book* (Elsevier Health Sciences, 2022).
4. Wang, G. *et al.* Cb-hrnet: A class-balanced high-resolution network for the evaluation of endoscopic activity in patients with ulcerative colitis. *Clin. Transl. Sci.* **16**, 1421–1430 (2023).
5. Carreras, J. Artificial intelligence analysis of ulcerative colitis using an autoimmune discovery transcriptomic panel. In *Healthcare*, vol. 10, 1476 (MDPI, 2022).
6. Gui, X. *et al.* Picasso histologic remission index (phri) in ulcerative colitis: development of a novel simplified histological score for monitoring mucosal healing and predicting clinical outcomes and its applicability in an artificial intelligence system. *Gut* **71**, 889–898 (2022).
7. Iacucci, M. *et al.* Artificial intelligence enabled histological prediction of remission or activity and clinical outcomes in ulcerative colitis. *Gastroenterology* **164**, 1180–1188 (2023).
8. Popa, I. V. *et al.* A new approach to predict ulcerative colitis activity through standard clinical–biological parameters using a robust neural network model. *Neural Comput. Appl.* **33**, 14133–14146 (2021).
9. Gottlieb, K. *et al.* Central reading of ulcerative colitis clinical trial videos using neural networks. *Gastroenterology* **160**, 710–719 (2021).
10. Chaddad, A., Peng, J., Xu, J. & Bouridane, A. Survey of explainable ai techniques in healthcare. *Sensors* **23**, 634 (2023).
11. Gutierrez Becker, B. *et al.* Training and deploying a deep learning model for endoscopic severity grading in ulcerative colitis using multicenter clinical trial data. *Ther. advances gastrointestinal endoscopy* **14**, 2631774521990623 (2021).
12. Kim, J. E. *et al.* Deep learning model for distinguishing mayo endoscopic subscore 0 and 1 in patients with ulcerative colitis. *Sci. reports* **13**, 11351 (2023).
13. Stidham, R. W. *et al.* Performance of a deep learning model vs human reviewers in grading endoscopic disease severity of patients with ulcerative colitis. *JAMA network open* **2**, e193963–e193963 (2019).
14. Huang, T.-Y., Zhan, S.-Q., Chen, P.-J., Yang, C.-W. & Lu, H. H.-S. Accurate diagnosis of endoscopic mucosal healing in ulcerative colitis using deep learning and machine learning. *J. Chin. Med. Assoc.* **84**, 678–681 (2021).
15. Mascarenhas, M. *et al.* Deep learning and colon capsule endoscopy: automatic detection of blood and colonic mucosal lesions using a convolutional neural network. *Endosc. Int. Open* **10**, E171–E177 (2022).

16. Ruan, G. *et al.* Development and validation of a deep neural network for accurate identification of endoscopic images from patients with ulcerative colitis and crohns disease. *Front. Medicine* **9**, 854677 (2022).

17. Chierici, M. *et al.* Automatically detecting crohn's disease and ulcerative colitis from endoscopic imaging. *BMC Med. Informatics Decis. Mak.* **22**, 300 (2022).

18. Khorasani, H. M., Usefi, H. & Pena-Castillo, L. Detecting ulcerative colitis from colon samples using efficient feature selection and machine learning. *Sci. reports* **10**, 13744 (2020).

19. Borgli, H. *et al.* Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Sci. data* **7**, 283 (2020).

20. Ozawa, T. *et al.* Novel computer-assisted diagnosis system for endoscopic disease activity in patients with ulcerative colitis. *Gastrointest. endoscopy* **89**, 416–421 (2019).

21. Kim, J.-H. *et al.* Using a deep learning model to address interobserver variability in the evaluation of ulcerative colitis (uc) severity. *J. Pers. Medicine* **13**, 1584 (2023).

22. Fan, Y. *et al.* Novel deep learning–based computer-aided diagnosis system for predicting inflammatory activity in ulcerative colitis. *Gastrointest. Endosc.* **97**, 335–346 (2023).

23. Zhang, P. *et al.* Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2998–3008 (2021).

24. Chen, H. *et al.* Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12299–12310 (2021).

25. Wang, Y., Huang, R., Song, S., Huang, Z. & Huang, G. Not all images are worth 16x16 words: Dynamic transformers for efficient image recognition. *Adv. neural information processing systems* **34**, 11960–11973 (2021).

26. Wang, W. *et al.* Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, 568–578 (2021).

27. Zhou, D. *et al.* Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886* (2021).

28. Liu, Z. *et al.* Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022 (2021).

29. Lee, S. H., Lee, S. & Song, B. C. Vision transformer for small-size datasets. *arXiv preprint arXiv:2112.13492* (2021).

30. Ali, I., Muzammil, M., Haq, I. U., Amir, M. & Abdullah, S. Deep feature selection and decision level fusion for lungs nodule classification. *IEEE Access* **9**, 18962–18973 (2021).

31. Cheng, X., Tan, L. & Ming, F. Feature fusion based on convolutional neural network for breast cancer auxiliary diagnosis. *Math. Probl. Eng.* **2021**, 1–10 (2021).

Competing Interest

Authors declare no conflicts of interest.

Author contributions

Conceptualization, S.A.S., I.T. and S.M.U.; methodology, S.A.S., I.T., S.M.U. and S.N.H.S ; software, S.A.S., S.M.U., S.K. and A.S.I. ; validation, S.M.U., S.N.H.S., S.M.U., S.K., and A.S.I.; formal analysis, I.T., S.M.U.,S.N.H.S., A.S.I. and S.K. ; investigation, S.A.S., S.M.U., S.N.H.S, S.K., and A.S.I.; resources, S.M.U., S.K. and A.S.I.; data curation, S.A.S., I.T., S.M.U., A.S.I. and S.K.; writing—original draft preparation, s.A.S.,I.T., S.M.U.,S.N.H.S., A.S.I. and
S.K.; writing—review and editing, I.T., S.M.U.,S.N.H.S., A.S.I. and S.K.; visualization, A.S.I., S.N.H.S., S.M.U and
S.K.; supervision, S.M.U. and S.K.; project administration, S.M.U and A.S.I;

Data Availability

The datasets generated and/or analysed during the current study are available in the LIMUC and TMC-UCM repository, https://zenodo.org/records/5827695#.Yi8GJ3pByUk; https://pan.baidu.com/share/init?surl=Q09eWJAQgkZf4lrvE5hkKg
& pwd=HUST