Mayo Endoscopic Score Classification

Engin Deniz Erkan
Department of Data Informatics
Middle East Technical University
Ankara, TURKEY
engin.erkan@metu.edu.tr

***Abstract*— Ulcerative colitis (UC) is a prevalent disease within the adult population. The Mayo Endoscopic Score (MES) is a significant diagnostic metric that is used in clinical studies to evaluate disease activity and see the response to treatment. Using MES for automated diagnosis is an essential subject since it may reduce doctors' workloads and improve the accuracy of diagnosis. Transformers can be used to analyze endoscopic pictures efficiently and increase the precision of MES categorization because of their well-known ability to capture long-range relationships and contextual information. In this study, the primary purpose is to develop a transformer model to categorize endoscopic images into four different MES groups, and the results will be compared with ResNet18 and a naive model, which are selected as the baseline models. This study highlights the potential for revolutionary technologies to improve healthcare delivery systems and advance the area of automated medical diagnostics.**

Keywords— Ulcerative colitis, Mayo Endoscopic Score, automated diagnosis, Transformers, ResNet18, medical diagnostics

I. Introduction

Natural language processing (NLP) is where transformers first emerged. They are deep neural networks that primarily use the self-attention mechanism to extract meaningful characteristics from textual material. Researchers have looked for ways to expand the use of transformers to computer vision applications because they see the opportunity for adaptation of this concept into different domains. But switching from text to images presents its own set of difficulties, such as more extensive data sets, noise, and other modalities, which makes straight transformer adoption extremely difficult. However, the incorporation of transformers has been a noteworthy advance in computer vision in recent times.

Transformers have shown impressive adaptability and performance when processing visual input. Vision Transformer (ViT), one of the first studies in this field, revolutionized image classification by presenting a revolutionary paradigm that perceives images as collections of patches, allowing transformers to interpret visual inputs directly. According to ViT's design, the input image is divided into smaller patches, which are subsequently linearly embedded and subjected to transformer block processing. With this method, global contextual information can be captured more efficiently, and transformers may make use of their natural ability to capture long-range relationships in images.

There are several benefits to use transformer architectures for diagnosing illnesses through classification, especially when there is a collection of labeled datasets. These benefits emerge from the architecture's fundamental characteristics. Transformers are distinguished by their attention mechanism, which gives them the unique capacity to recognize complex relationships and connections in data that is received. This corresponds to a strong ability to identify minor patterns and characteristics of different diseases in the context of medical diagnostics, especially in complicated and varied datasets. Transformers are superior to conventional convolutional neural networks (CNNs) in finding long-range relationships, allowing them to combine data from many input images efficiently. This feature is very useful in medical diagnosis, as diseases can show up as small variations over a variety of spatial scales. Furthermore, the transformers' self-attention mechanism makes adaptive feature extraction easier, dynamically prioritizing relevant data while suppressing noise and extraneous features. This flexibility is essential for managing the natural noise and unpredictability seen in medical imaging datasets, which improves the model's resilience and capacity for generalization.

A good measure of the severity of the disease is the Mayo endoscopic score, which is a direct reflection of the mucosal involvement seen during endoscopic examination. This system of ratings, which ranges from 0 to 3, classifies the severity of ulcerative colitis according to certain parameters. Higher scores imply more severe activity, whereas lower values suggest less activity related to the condition. When it comes to managing ulcerative colitis, the Mayo endoscopic score is essential for assessing disease activity, directing therapy choices, and tracking therapeutic response.

The dataset used in this study is the Labeled Images for Ulcerative Colitis (LIMUC) dataset, which consists of 1,043 colonoscopy procedures and 11,276 endoscopic images collected from 564 patients between December 2011 and July 2019 at the Marmara University School of Medicine's Department of Gastroenterology. For the purpose of treating their ulcerative colitis, these patients had colonoscopies. Two skilled gastroenterologists carefully examined the pictures and categorized them using the Mayo Endoscopic Score (MES). When the two reviewers couldn't agree, a third, additional expert reviewer evaluated and classified the images on their own without knowing the preceding classifications. A majority vote method was employed to select the final MES for images with conflicting labels. All images in the LIMUC dataset have a standardized size of 352x288 pixels. The distribution of MES categories within the dataset is as follows: Mayo 0 (54.14%), Mayo 1 (27.70%), Mayo 2 (11.12%), and Mayo 3 (7.67%) [1].

II. Literature Review

TransMed is an innovative multi-modal CNN and transformer architecture-based medical image classification method. Compared to conventional CNN-based techniques, introducing transformers improves performance by capturing long-range dependencies between image sequences. The experimental findings demonstrate TransMed's superiority over current models, underscoring its potential to further the multi-modal medical picture classification field. The paper proposes a hybrid model that efficiently utilizes transformers to capture global characteristics and CNN for low-level feature extraction, leading to improved classification accuracy. When everything is considered, the incorporation of transformer architecture into TransMed shows encouraging outcomes and represents a noteworthy development in the field of medical picture classification. In multi-modal medical picture classification, the suggested approach has demonstrated a remarkable overall improvement of 10.1% when compared to earlier state-of-the-art models. This is a significant improvement over conventional CNN-based methods, which can be attributed to the efficient use of transformers to capture long-range relationships in the images. [2].

The paper explores the application of transformer architectures, notably the Vision Transformer (ViT), in image classification tasks. While resolving issues like training instability with increasing model depth, it demonstrates ViT's effectiveness in the natural language processing (NLP) and computer vision domains. In order to improve model performance, a unique approach is presented that prioritizes crucial heads in each layer. The research highlights the promise of ViT and related transformers for image classification, but it also notes that current issues must be resolved. The main goal is to compare the efficacy of the suggested model with the most advanced methods for classifying diseases, using the Cassava Leaf Disease Dataset as a reference. Findings show that the proposed model outperforms the top-performing model in earlier research by 3%, with an impressive F1-score of 96.80%. This demonstrates how well the suggested model correctly categorizes diseases seen in the Cassava Leaf Disease Dataset [3].

The study's main goal was to use deep learning algorithms to estimate Ulcerative Colitis (UC) endoscopic activity from static image data. Convolutional neural network (CNN) models were used in the studies, and the task was presented as a regression problem. The research aimed to improve classification accuracy and handle the ordinal development of UC severity by including domain information in the model design. The study results showed that using deep learning techniques made it easier to estimate the Mayo endoscopic score accurately and produced high-performance measures such as quadratic weighted kappa. DenseNet121 was the best-performing CNN model for assessing endoscopic activity in ulcerative colitis among those that were tested. It performed better on most measures of evaluation; however, the Inception-v3 model outperformed it in terms of macro F1 score. Notably, DenseNet121 proved to be the most successful model in the investigation by exhibiting the highest skill level in categorizing all Mayo endoscopic scores [4].

III. Exploratory Data Analysis

A. *Statistics about the datasets*

There are totally two datasets to be used for this study. One dataset is for training and validation, and the other is used for testing purposes. Tables 1 and 2 represent some of the statistics of the datasets.

TABLE I. Training/Validation Dataset Statistics

Sub-file	Number of Images	Image Size	Image Format	Total Size
Mayo 0	5180	(288, 352)	.bmp	1502.67 MB
Mayo 1	2588	(288, 352)	.bmp	750.75 MB
Mayo 2	1077	(288, 352)	.bmp	312.43 MB
Mayo 3	745	(288, 352)	.bmp	216.12 MB

TABLE II. Test Dataset Statistics

Sub-file	Number of Images	Image Size	Image Format	Total Size
Mayo 0	925	(288, 352)	.bmp	268.33 MB
Mayo 1	464	(288, 352)	.bmp	134.60 MB
Mayo 2	177	(288, 352)	.bmp	51.35 MB
Mayo 3	120	(288, 352)	.bmp	34.81 MB

B. *Visualizion of sample images from each class*

Figure 1 shows sample images that correlate to various MES levels.

Fig. 1. Actual endoscopic images that represent various MES levels

C. Pixel Intensity Distribution

The concept of Maximum Pixel Intensity (MPI) entails identifying the average pixel intensity value that exhibits the highest frequency across images within a specific class. Frequency describes how often the detected pixel intensity occurs in the class's image data. The results show that the same maximum pixel intensity value of 14 is seen in all classes. Nonetheless, there are tiny variations in the corresponding frequencies, indicating slight differences in the pixel intensity distribution between the classes.

IV. Experiments

A. Naïve Model

The study's naive baseline model makes predictions based on the dataset's most prevalent class, operating on a basic assumption. This method just uses a simple heuristic and does not require any training or learning procedures. The model labels all predictions with the class that is most common in the dataset because it believes that this class is reflective of most occurrences. When looking at the training data set, the label most used was Mayo 0, so all images in the test data were classified this way.

Table 3 shows the results of evaluation metrics for the test dataset. Figure 2 represents the confusion matrix for the test dataset.

TABLE III. Naïve Model Evaluation Metrics

Metric	Test Set
MAE	0.698695
Quadratic Weighted Kappa	0
F1 Score (Macro)	0.177135
Accuracy (Macro)	0.548636

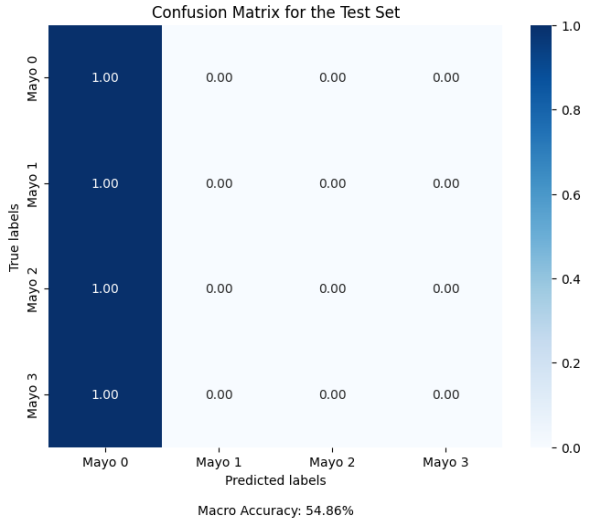

Fig. 2. Confusion matrix for the test dataset using naïve model

B. ResNet18

Residual Networks, or ResNets, are a significant development in deep learning architecture. Introducing residual connections to address the vanishing gradient problem can simplify the training of deep neural networks, which is one of its essential characteristics. The network can learn residual mappings due to these connections, which helps with optimization and prevents accuracy from decreasing as network depth increases. ResNet's skip connections allow for direct information flow between layers, aiding in the efficient learning of both shallow and deep features. Nevertheless, the insertion of these skip connections faces a more significant computational cost when the model parameters are increased. Moreover, in scenarios when the dataset may be larger, the depth of ResNet may result in overfitting. ResNet has demonstrated remarkable success in various computer vision tasks, particularly image classification, and remains a foundational architecture in the deep learning landscape.

Significant preprocessing operations called data transformations are performed on input images before supplying them to the neural network model. In this study, a series of transformations are encapsulated. To enhance the dataset, the images are resized to a uniform 224x224 pixel size. Additionally, the images are randomly rotated by up to 15 degrees to introduce variability, flipped horizontally to augment the dataset, and adjusted for color jitter, modifying brightness, contrast, saturation, and hue. Then, they are converted to tensors to make computation more manageable, and finally, the pixel values are normalized using predetermined mean and standard deviation values. Normalization ensures that the pixel values have a mean of 0 and a standard deviation of 1, which helps stabilize and expedite the training process of the neural network.

This study puts 20% of the dataset aside for validation, leaving the remaining 80% as training data. This split guarantees that the model is trained on an adequate dataset and provides a distinct subset for assessing its performance on untested data. Such partitioning is essential for preventing overfitting and ensuring the generalizability of the trained model to new, unseen instances.

Figure 3 shows the training and validation data loss curves during the training process. As both losses decrease with increasing epochs, no overfitting is observed during the training process. Table 4 shows the results of evaluation metrics for both the validation and test datasets. The primary evaluation metric that will be used to compare different models is Quadratic Weighted Kappa. This metric is calculated as 0.62 for the test data. Figure 4 represents the confusion matrix for the test dataset.

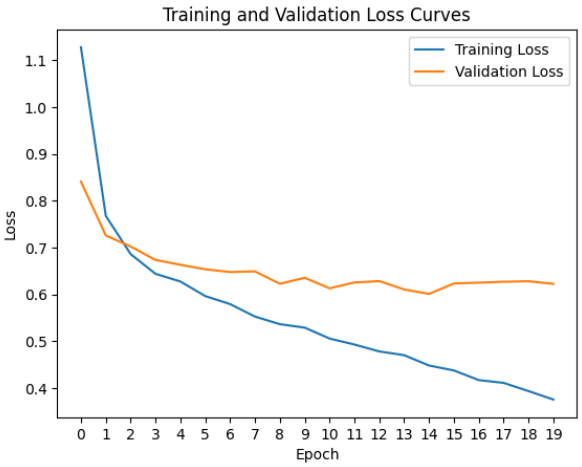

Fig. 3. Training and validation loss curves

TABLE IV. RESNET18 EVALUATION METRICS

Metric	Validation Set	Test Set
MAE	0.315954	0.33274
Quadratic Weighted Kappa	0.664719	0.616622
F1 Score (Macro)	0.686931	0.655201
Accuracy (Macro)	0.749218	0.732503

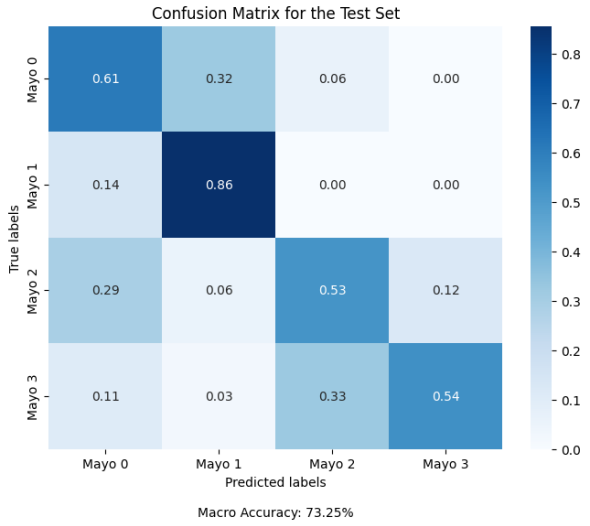

Fig. 4. Confusion matrix for the test dataset using ResNet18 model

C. VIT Model

Motivated by the current developments in deep learning architectures, this work presented a new approach for image classification of ulcerative colitis disease phases using a Vision Transformer (ViT) model. This strategy substitutes the classification head with the pre-trained DeiT model, making it more specific to a task. Data preprocessing techniques used in the ResNet18 model were applied similarly. During the examination, the ViT model achieved competitive metrics, indicating its exceptional ability to differentiate between the phases of ulcerative colitis. Using a methodical grid search, hyperparameters like learning rate and batch size were meticulously selected to provide the best possible model convergence and generalization. The Weights and Biases platform was utilized to explore hyperparameters systematically using a grid search approach. Batch sizes of 16 and 32 were used, and the learning rate was carefully tuned throughout a range of values, including 1e-5, 1e-4, and 1e-3. Notably, the model performed at its peak with a learning rate of 1e-5 and a batch size of 32. These results highlight hyperparameter adjustment's role in optimizing model performance, improving its capacity for convergence and generalization. This method offers insights into illness development with implications for clinical diagnosis and treatment, marking a promising medical image analysis direction. The evaluation metrics results for the validation and test datasets are displayed in Table 5. Quadratic Weighted Kappa is the primary evaluation statistic that will be utilized to evaluate various models. Regarding the test data, this measure is computed as 0.73. The test dataset's confusion matrix is shown in Figure 5.

TABLE V. VIT MODEL EVALUATION METRICS

Metric	Validation Set	Test Set
MAE	0.27268	0.263938
Quadratic Weighted Kappa	0.738572	0.725913
F1 Score (Macro)	0.730975	0.710441
Accuracy (Macro)	0.773201	0.77758

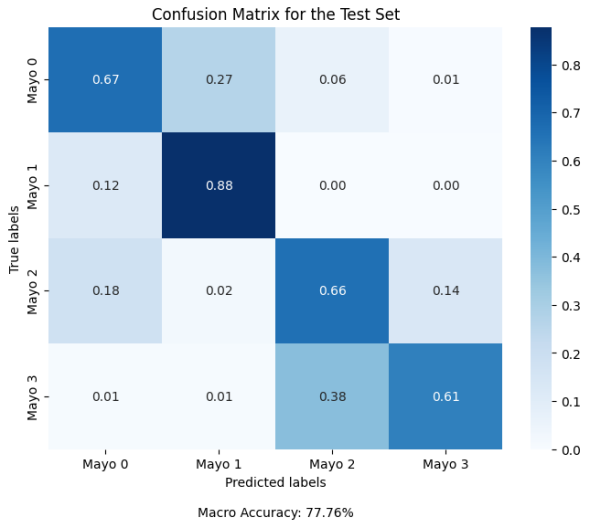

Fig. 5. Confusion matrix for the test dataset using VIT model

V. Discussions

Three models were compared in terms of performance for the classification of ulcerative colitis disease stages: the ViT model, the ResNet18 model, and the naïve baseline model. The objective of the naïve baseline model was to forecast the dataset's most common class, in this example, Mayo 0. With no training required, this straightforward method performed poorly, yielding a Quadratic Weighted Kappa measure of 0 on the test set. This finding emphasizes the requirement for more advanced methods by underscoring the model's incapacity to distinguish between the different phases of ulcerative colitis.

On the other hand, a well-known convolutional neural network model called ResNet18 showed significant improvements. This model obtained a Quadratic Weighted Kappa measure of 0.62 on the test data, suggesting moderate acceptance of the actual labels. The ResNet18 performed better than the naïve baseline, perhaps because of its deep design, which records hierarchical characteristics.

With the greatest QWK score of 0.72 on the test data, the Vision Transformer model beat the ResNet18 model as well as the naïve baseline. Compared to conventional convolutional models, the ViT model's architecture, which makes use of the self-attention mechanism, enables it to better capture complex patterns and global context in visual information. The model's performance was also much enhanced by the hyperparameter tuning carried out using the Weights and Biases platform; the best outcomes were attained with a batch size of 32 and a learning rate of 1e-5.

These results imply that transformer-based models, like the ViT, have a lot of potential for classification tasks involving medical images. The ViT model's better performance highlights its ability to correctly categorize ulcerative colitis into several phases, which might be crucial for clinical assessment and treatment planning.

From the confusion matrix results, the ViT model consistently performs better than the ResNet18 model in all classes. The Mayo 2 and Mayo 3 stage classifications, which are essential for proper disease staging and treatment planning, are significantly improved by the ViT model. The improved accuracy for these difficult classes demonstrates how well the ViT model can identify minor discrepancies in the medical images, highlighting why it's suitable for this kind of application.

VI. Conclusions

This study investigated the use of several machine learning models to categorize ulcerative colitis phases using data from medical images. Three models were compared: a ResNet18 model, a ViT model, and a naive baseline model. With a Quadratic Weighted Kappa measure of 0.72, the ViT model performed the best, indicating its higher capacity to identify complex patterns in the data.

The outcomes emphasize how crucial model selection and hyperparameter tuning are to achieving peak performance. The success of the ViT model indicates that transformer-based designs have potential in the field of medical image categorization. Further studies may enhance these models and investigate their suitability for additional medical diseases, which might lead to better treatment planning and increased diagnostic precision.

VII. Future Works

In order to better enhance the performance of the model, future studies might include investigating a more extensive variety of hyperparameters. Furthermore, particular focus might be focused on pictures with annotations supplied by medical professionals, including circles and arrows. A couple of images have that kind of annotation in the dataset used. Creating strategies to direct the model's attention toward these labeled regions might improve classification accuracy by utilizing the medical experts' domain-specific knowledge. This method may be beneficial for enhancing the identification and categorization of minor characteristics that point to distinct ulcerative colitis phases.

References

[1] Gorkem Polat, Haluk Tarik Kani, Ilkay Ergenc, Yesim Ozen Alahdab, Alptekin Temizelve Ozlen Atug, “Labeled Images for Ulcerative Colitis (LIMUC) Dataset”, *Inflammatory Bowel Diseases*, c. 2022, sy 11. Zenodo, Mar. 14, 2022. doi: 10.5281/zenodo.5827695.

[2] Dai, Y.; Gao, Y.; Liu, F. TransMed: Transformers Advance Multi-Modal Medical Image Classification. Diagnostics 2021, 11, 1384. https://doi.org/10.3390/diagnostics11081384

[3] H.-T. Thai, K.-H. Le, and N. L.-T. Nguyen, "FormerLeaf: An efficient vision transformer for Cassava Leaf Disease detection," *Computers and Electronics in Agriculture*, vol. 204, p. 107518, 2023. [Online]. Available: https://doi.org/10.1016/j.compag.2022.107518

[4] G. Polat, H. T. Kani, I. Ergenc, Y. O. Alahdab, A. Temizel, and O. Atug, "Improving the Computer-Aided Estimation of Ulcerative Colitis Severity According to Mayo Endoscopic Score by Using Regression-Based Deep Learning," *Inflammatory Bowel Diseases*, Nov. 2022.