COMPUTER-AIDED ESTIMATION OF ENDOSCOPIC ACTIVITY IN ULCERATIVE COLITIS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

GÖRKEM POLAT

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
MEDICAL INFORMATICS

JULY 2023

COMPUTER-AIDED ESTIMATION OF ENDOSCOPIC ACTIVITY IN ULCERATIVE COLITIS

submitted by **GÖRKEM POLAT** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Health Informatics Department, Middle East Technical University** by,

Prof. Dr. Banu Günel Kılıç
Dean, **Graduate School of Informatics**

Assoc. Prof. Dr. Yeşim Aydın Son
Head of Department, **Health Informatics**

Prof. Dr. Alptekin Temizel
Supervisor, **Modeling and Simulation, Middle East Technical University**

Examining Committee Members:

Assoc. Prof. Dr. Yeşim Aydın Son
Health Informatics, Middle East Technical University

Prof. Dr. Alptekin Temizel
Modeling and Simulation, Middle East Technical University

Assist. Prof. Dr. Aybar Can Acar
Health Informatics, Middle East Technical University

Prof. Dr. Çiğdem Gündüz Demir
Computer Engineering, Koç University

Assoc. Prof. Dr. Haluk Tarık Kani
Dept. of Gastroenterology, School of Med., Marmara University

Date: 17.07.2023

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Görkem Polat

Signature :

ABSTRACT

COMPUTER-AIDED ESTIMATION OF ENDOSCOPIC ACTIVITY IN ULCERATIVE COLITIS

Polat, Görkem
Ph.D., Department of Health Informatics
Supervisor: Prof. Dr. Alptekin Temizel

July 2023, 74 pages

Ulcerative colitis (UC) is a chronic inflammatory bowel disease that presents significant diagnostic and management challenges for clinicians. Accurate assessment of disease severity is crucial for guiding appropriate treatment strategies and improving patient outcomes. The Mayo endoscopic score (MES) is a widely used tool for evaluating UC severity; however, the assessment process relies heavily on subjective interpretation, leading to substantial intra- and inter-observer variability.

In this thesis, we present a novel loss function, termed Class Distance Weighted Cross Entropy (CDW-CE) loss, for the automated assessment of UC severity, harnessing the power of convolutional neural networks (CNN) to analyze endoscopic images of the colon. CDW-CE addresses the limitations of conventional cross-entropy loss functions in ordinal classification problems.

The proposed CDW-CE loss effectively penalizes mispredictions based on their distance from the true class, taking into account the inherent ordinal relationships among the output classes. CDW-CE has been evaluated against other loss functions and consistently outperformed them across various performance metrics and CNN architectures. Moreover, the proposed approach enables the generation of more accurate class activation maps, which can be utilized to explain model predictions —an essential aspect of translating these techniques into clinical practice. To demonstrate the generalizability of the proposed approach, it is also tested on a diabetic retinopathy dataset and got similar results, indicating that the proposed approach can be used in other applications presenting ordinal classes. The dataset created for this study, named Labeled Images for Ulcerative Colitis, is the largest publicly available labeled UC dataset to date.

Keywords: Computer-Assisted Diagnosis, Convolutional Neural Networks, Ulcerative Colitis, Ordinal Classification

ÖZ

ÜLSERATİF KOLİT ENDOSKOPİK AKTİVİTESİNİN BİLGİSAYAR YARDIMI İLE TAHMİN EDİLMESİ

Polat, Görkem
Doktora, Sağlık Bilişimi Bölümü
Tez Yöneticisi: Prof. Dr. Alptekin Temizel

Temmuz 2023, 74 sayfa

Ülseratif kolit (ÜK), klinisyenler için teşhis ve tedavi zorlukları iceren kronik bir inflamatuar barsak hastalığıdır. Hastalık şiddetinin doğru bir şekilde değerlendirilmesi, uygun tedavi stratejilerini izlemek ve hasta sonuçlarını iyileştirmek için çok önemlidir. Mayo endoskopik skoru (MES), ÜK şiddetini değerlendirmek için yaygın olarak kullanılan bir araçtır; ancak, değerlendirme süreci büyük ölçüde öznel yorumlamaya dayanır ve bu da gözlemci içi ve gözlemciler arası önemli değişkenliğe yol açar.

Bu tezde, ÜK şiddetinin otomatik değerlendirmesi için evrişimli sinir ağlarının kullanacagi Sınıf Mesafe Ağırlıklı Çapraz Entropi (SMA-ÇE) Kaybı olarak adlandırılan yeni bir kayıp fonksiyonu sunmaktayız. SMA-ÇE, sıralı sınıflandırma problemlerinde geleneksel çapraz entropi kayıp fonksiyonlarının yetersizliğine çözüm getirmektedir.

SMA-ÇE fonksiyonu, yanlış tahminleri gerçek sınıftan uzaklıklarıyla orantılı olarak etkili bir şekilde cezalandırır, böylece çıktı sınıfları arasındaki doğal sıralı ilişkileri yakalar. SMA-ÇE, diğer kayıp işlevlerine karsi değerlendirildiğinde bütün performans ölçümleri ve CNN mimarilerinde istikrarli bir sekilde onlardan daha iyi performans göstermektedir. Ayrıca, önerilen yaklaşım, model tahminlerini açıklamak için kullanılabilecek sınıf aktivasyon haritalarının daha dogru oluşturulmasını da sağlamaktadır; daha doğru açıklanabilirlik görüntüleri, geliştirilmekte olan tekniklerin klinik uygulamaya dönüştürülmesi için önemli bir özelliktir. Önerilen yaklaşımın başka veri setlerindeki performansını ölçmek icin diyabetik retinopati veri seti üzerinde de deneyler yapılmıştır ve benzer sonuçlar alınmıştır; bu durum önerilen yaklaşımın başka sıralı sınıf özelliğine sahip uygulamalarda da kullanılabileceğini göstermektedir. Bu çalışma için oluşturulan, Etiketli Ülseratif Kolit Görüntüleri adlı veri seti, bugüne kadar halka açık en büyük etiketli ÜK veri setidir.

Anahtar Kelimeler: Bilgisayar Destekli Teşhis, Evrişimsel Sinir Ağları, Ülseratif Kolit, Sıralı Sınıflandırma

To my family

ACKNOWLEDGMENTS

As I reflect upon the journey that has led me to the culmination of my Ph.D., I am grateful to the individuals and institutions who have supported, guided, and inspired me along the way.

Foremost, I express my deepest gratitude to my advisor, Prof. Alptekin Temizel. His guidance and critical thinking approach have provided me with an ideal environment to grow both academically and personally. I am particularly thankful for his efforts that enabled access to significant hardware resources and fostered key institutional partnerships, fundamental to my research. It is under his mentorship and thoughtful care that I have been able to reach this milestone.

I would like to thank METU Deep Learning and Computer Vision Lab team. It was inspiring and enlightening to listen to presentations that we held weekly and discuss the topics with the brilliant minds there. I feel lucky to be surrounded by the researchers in this lab.

I am thankful to my collaborators from Marmara University, Tarık Kani, İlkay Ergenç, Yeşim Özen Alahdab, and Özlen Atuğ, whose contributions have been invaluable to the process. Their rigorous work in annotating the LIMUC dataset formed the backbone of this thesis. Their cooperation and team spirit have made our collaborative efforts productive and enjoyable.

I would like to thank Sharib Ali, whose leadership in organizing the EndoCV challenges introduced me to new opportunities and perspectives that have significantly enriched my Ph.D. journey. His commendable efforts in facilitating these workshops and challenges have made me a much more equipped person in his field.

The financial support from TÜBİTAK and YÖK during my Ph.D. has been a significant enabler. Their generous scholarships provided me the freedom to focus solely on my research, without worrying about external distractions. I am truly thankful for their support.

I would like to thank the academic staff of the Informatics Institute at METU. I learned a lot of valuable things here and feel lucky to have done my Ph.D. in such a supportive environment.

At the heart of this journey lies the unwavering support of my family. My wife, Ece, has been my rock through the most challenging times. She has been a beacon of inspiration, support, and love. More than a spouse, she has been an unwavering collaborator, my confidant, the one I could always turn to.

Finally, I want to express my heartfelt gratitude to my parents, Ali, Gülcan, and Saliha. Their love and support have been my sanctuary. Their readiness to lend a helping hand in times of need is something I am profoundly grateful for. Their belief in me has always given me the strength to push through the toughest hurdles.

And lastly, but not least, I wish to acknowledge the joy and inspiration brought into my life by my two daughters, Mavi and Nova, who were born during this journey. Their presence has been a constant reminder of the beauty and joy of life amidst all the trials and challenges. Their laughter and smiles have been the most wonderful distraction from my work, and their innocence has provided a refreshing perspective when needed. As I watch them grow, I am reminded of the growth I have undergone during this process. I dedicate this work to them, hoping that one day, they may draw strength and inspiration from my journey.

Completing this thesis has been a remarkable journey, made possible by the combined efforts and contributions of everyone acknowledged here. I extend my sincere appreciation to all of you. My gratitude for your part in my journey will forever be a part of this work.

TABLE OF CONTENTS

ABSTRACT iv

ÖZ vi

DEDICATION viii

ACKNOWLEDGMENTS ix

TABLE OF CONTENTS xi

LIST OF TABLES xiv

LIST OF FIGURES xv

LIST OF ABBREVIATIONS xvii

CHAPTERS

1 INTRODUCTION 1

- 1.1 Research Questions 2
- 1.2 Contributions of the Study 3
- 1.3 Thesis Outline 4

2 LITERATURE REVIEW 5

- 2.1 Computer Aided Diagnosis of Ulcerative Colitis 5
- 2.2 Ordinal Classification 9

3 LABELED IMAGES FOR ULCERATIVE COLITIS DATASET 17

3.1 Ulcerative Colitis and Mayo Endoscopic Score 18

3.2 Data Collection and Processing 20

3.3 Data Labeling 22

4 CLASS DISTANCE WEIGHTED CROSS ENTROPY 25

4.1 Motivation 25

4.2 Class Distance Weighted Cross Entropy Loss Function 26

5 EXPERIMENTAL SETTINGS 29

5.1 Model Training and Evaluation 29

5.2 Explainability Analysis 34

5.3 Experimental Evaluation on an Additional Dataset: Diabetic Retinopathy 35

6 RESULTS AND DISCUSSION 39

6.1 Power Factor Analysis of CDW-CE 39

6.2 Comparison of Approaches Targeting Loss Functions 40

6.2.1 Ordinal Approaches 40

6.2.2 Non-Ordinal Approaches 43

6.3 Pushing the Boundaries of CDW-CE with Additive Margin 48

6.4 Performance on the Hard Samples 49

6.5 Explainability Analysis 53

6.6 Experiment Results on Diabetic Retinopathy Dataset 56

7 CONCLUSION AND FUTURE WORK 59

7.1 Conclusion 59

7.2 Future Work 60

REFERENCES 63

APPENDICES

A THE IMPLEMENTATION OF CDW-CE WITH PYTORCH 71

LIST OF TABLES

Table 1 Detailed overview of studies in CAD of UC domain. In studies that have multiple datasets, labeling as A and B is performed to show their performance results separately in 2 and 3. 13

Table 2 Performance results of the studies (Part 1). 14

Table 3 Performance results of the studies (Part 2). 15

Table 4 Characteristics of the study cohort. 20

Table 5 Distributions of annotation among different classes. 22

Table 6 Distribution of annotations of the third reviewer 23

Table 7 Comparison of different model predictions in terms of loss. 26

Table 8 Drawback of unimodality approach in ordinal classification. 26

Table 9 Experiment results for all Mayo scores. 42

Table 10 Experiment results for remission classification. 43

Table 11 CDW-CE vs. ArcFace. 45

Table 12 CDW-CE vs. CDW-CE with margin 48

Table 13 Number of hard samples for each class in the test set. 49

Table 14 Performance increase comparison for the hard samples vs. all samples. 50

Table 15 Comparison of loss functions on diabetic retinopathy dataset. 57

LIST OF FIGURES

Figure 1 Prediction inconsistencies in binary classification. 10

Figure 2 CE loss function and unimodal distribution. 11

Figure 3 Example patterns of affected parts of CD and UC [1]. 18

Figure 4 Example UC images and their corresponding scores according to the MES. 19

Figure 5 Sample images from the LIMUC dataset 21

Figure 6 User interface designed for the third reviewer to annotate images. 23

Figure 7 Colonoscopy operations and images data. 24

Figure 8 Mayo classes and patient images distribution. 24

Figure 9 Data splitting for the model training and evaluation. 31

Figure 10 Masking regions including text. 32

Figure 11 Training a CNN with the ArcFace loss. 33

Figure 12 User interface for CAM visualizations. 35

Figure 13 Class distributions in the DR dataset. 37

Figure 14 Sample images from the DR dataset. 37

Figure 15 Change of mean and standard deviation of QWK scores according to varying α for three models. 40

Figure 16 Mean confusion matrix of each CNN model trained with CE and CDW-CE for full Mayo score classification. 44

Figure 17 Mean confusion matrix of each CNN model trained with CE and CDW-CE for remission classification. 45

Figure 18 t-SNE embeddings of features obtained from ResNet18 model trained with CE. 46

Figure 19 t-SNE embeddings of features obtained from ResNet18 model trained with CDW-CE. 47

Figure 20 CDW-CE vs. CDW-CE with margin for varying α values. 49

Figure 21 Confusion matrices for inference on hard samples for full Mayo score estimation. 51

Figure 22 Confusion matrices for inference on hard samples for remission classification. 52

Figure 23 ResNet18 model CAM visualizations. 54

Figure 24 The assessment results of CAM visualizations for models trained with CE and CDW-CE. 55

LIST OF ABBREVIATIONS

CAD	Computer-Aided Diagnosis
CAM	Class Activation Map
CD	Crohn’s Disease
CDW-CE	Class Distance Weighted Cross Entropy
CE	Cross Entropy
CNN	Convolutional Neural Network
CORAL	Consistent Rank Logits
CORN	Conditional Ordinal Regression for Neural Network
DL	Deep Learning
DR	Diabetic Retinopathy
IBD	Inflammatory Bowel Disease
GI	Gastrointestinal
LIMUC	Labeled Images for Ulcerative Colitis
MAE	Mean Absolute Error
MES	Mayo Endoscopic Score
ML	Machine Learning
MSE	Mean Squared Error
SSL	Semi-Supervised Learning
SVM	Support Vector Machines
UC	Ulcerative Colitis
UCEIS	Ulcerative Colitis Endoscopic Index of Severity

CHAPTER 1

INTRODUCTION

Ulcerative colitis (UC) is a chronic inflammatory bowel disease that affects millions of people worldwide. It is characterized by inflammation and ulcers in the inner lining of the colon and rectum, leading to various symptoms such as abdominal pain, diarrhea, and rectal bleeding. Early and accurate diagnosis of UC is crucial for the effective treatment and management of the disease. To measure the endoscopic activity of the UC, practitioners use scoring systems like Mayo endoscopic score (MES) [2] or Ulcerative Colitis Endoscopic Index Of Severity (UCEIS) [3] to grade the severity of the disease. However, this evaluation is dependent on the experience and education of the endoscopist, which can lead to subjectivity in the assessment. Previous studies have shown that there are significant differences in the grading of endoscopic severity between observers, particularly regarding the level of experience [4, 3]. As a result, the reliability and reproducibility of grading endoscopic severity remain major concerns.

Over the past few decades, there has been an increasing interest in developing computer-aided diagnosis (CAD) systems for the severity estimation of UC. In this regard, deep learning algorithms, especially Convolutional Neural Networks (CNNs), have shown promising results in analyzing endoscopic images of the colon and predicting the severity of the disease [5, 6, 7, 8, 9]. The studies in this area have not only remained at the academic level but also many private companies have integrated them into their own products [10, 11, 12, 13].

This thesis presents an in-depth investigation into the development of a CAD system for UC using CNNs. The proposed system aims to improve the accuracy and efficiency of UC severity estimation, thereby facilitating timely and effective treatment for patients. More specifically, we investigate the ways of incorporating ordinality existing in the MES system to improve CAD performance. We have proposed a novel loss function called Class Distance Weighted Cross Entropy (CDW-CE) and did extensive experiments to show its superiority to the previous approaches. In addition, we have gathered the largest publicly available UC dataset called Labeled Images for Ulcerative Colitis (LIMUC) [14].

A central consideration in the domain of machine learning is the robustness and generalizability of an algorithm or model. Beyond the efficacy and precision in a singular, specialized dataset, an algorithm's performance on external datasets is a measure of its adaptability and resilience to varying contexts and conditions - its ability to generalize. The proposed loss function CDW-CE can be used in other domains that possess an ordinal structure in their labels. To assess the adaptability and performance of our proposed method on an external dataset, we conducted several experiments utilizing a widely recognized diabetic retinopathy dataset. The results were consistent with those from the UC experiments. This consistency demonstrates that our proposed approach is not confined to the domain of UC but extends its efficacy to other domains characterized by an ordinal relationship among their labels.

In summary, this thesis aims to advance the field of computer-assisted diagnosis of ulcerative colitis by proposing a novel loss function for training CNN models. We anticipate that our findings will contribute to improving the accuracy and reliability of UC severity estimation, ultimately leading to better patient care and outcomes. We hope that the usage of CDW-CE will not only be limited to the UC, but other domains will also benefit from that.

1.1 Research Questions

The primary goal of this research is to address the problem of accurately and reliably estimating UC severity from endoscopic images using deep learning techniques. To achieve this goal, our research will focus on the development of a novel loss function that can improve the performance of CNNs in this specific application. In this context, the central research question of this thesis is:

How can we develop a more effective and robust loss function for training CNN models that can accurately and reliably estimate UC severity from endoscopic images, taking into account inter-class distance?

To answer this research question, we will address the following sub-questions:

1. What are the limitations of the existing loss functions used for training CNN models in UC severity estimation, and how do they impact the model's performance and reliability?

2. How can we incorporate the distance between different severity levels into the loss function to enhance the performance of CNN models in UC severity estimation?

3. How does the proposed CDW-CE compare to the traditional loss functions in terms of model performance and reliability for UC severity estimation? Does it also generalizable to other domains that have an ordinal label structure?

4. To what extent can the proposed approach improve the explainability of UC severity grading from endoscopic images?

1.2 Contributions of the Study

This study makes several significant contributions to the field of CAD of UC severity estimation. These contributions are expected to advance the state of the art and provide a solid foundation for future research in this area. The main contributions of this study are as follows:

1. **LIMUC dataset:** We have collected and curated the largest publicly available dataset of labeled endoscopic images for UC, called the "Labeled Images for Ulcerative Colitis" (LIMUC) dataset [14]. This comprehensive dataset will enable researchers and practitioners to develop, test, and validate their own models, providing a common benchmark for comparing methods and fostering further advancements in the field. By making the LIMUC dataset publicly accessible, we aim to promote transparency, reproducibility, and collaboration in the research community.

2. **Class Distance Weighted Cross-Entropy (CDW-CE) Loss:** We have proposed a novel loss function, named Class Distance Weighted Cross-Entropy [15], specifically designed for ordinal classification tasks in UC severity estimation. Through extensive experimentation and comparison with existing loss functions, we have demonstrated that our proposed CDW-CE loss function consistently yields superior results in terms of prediction performance and reliability. The CDW-CE loss function addresses the limitations of traditional loss functions by effectively incorporating inter-class distance considerations, thus enhancing the performance of CNN models for UC severity estimation. By incorporating an additive margin term, we also share ways to further increase the performance of the proposed method and the results of experiments related to it. We have tested the generalizability and efficacy of CDW-CE on an out-of-domain dataset, which possesses the ordinal structure, and got consistent results with the UC experiments.

3. **Explainability and robustness analysis of CDW-CE Loss:** To ensure the clinical utility and trustworthiness of our proposed approach, we have conducted a thorough explainability and robustness analysis of the CDW-CE loss function. This analysis is crucial for the widespread adoption of computer-aided detection systems in UC diagnosis, as it provides insights into the decision-making process of the CNN models and ensures that the models are resilient to different settings. We identified challenging samples in the training set and obtained the performance of the proposed approach on these samples to measure how robust it is. By demonstrating the explainability and robustness of the CDW-CE loss, we aim to facilitate the translation of our research findings into real-world clinical practice.

The work presented in this thesis has led to the following publications:

- G. Polat, H. T. Kani, I. Ergenc, Y. Ozen 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, 11 2022.

- G. Polat, I. Ergenc, H. T. Kani, Y. O. Alahdab, O. Atug, and A. Temizel, "Class distance weighted cross-entropy loss for ulcerative colitis severity estimation," in Medical Image Understanding and Analysis: 26th Annual Conference, MIUA 2022, Cambridge, UK, July 27–29, 2022, Proceedings. Springer, 2022, pp. 157–171.

- H. T. Kani, I. Ergenc, G. Polat, Y. O. Alahdab, A. Temizel, and O. Atug, "Evaluation of ulcerative colitis endoscopic mayo score with artificial intelligence," Endoscopy, vol. 54, no. S 01, p. eP083, 2022.

- H. T. Kani, I. Ergenc, G. Polat, Y. Ozen Alahdab, A. Temizel, and O. Atug, "Evaluation of endoscopic Mayo score with an artificial intelligence algorithm," Journal of Crohn's and Colitis, vol. 15, no. Supplement_1, pp. S195–S196, 2021.

1.3 Thesis Outline

Chapter 2 provides a comprehensive literature review of studies in computer-aided diagnosis of the UC, and loss functions for the ordinal classification problems. It highlights the key advancements and results of these studies.

Chapter 3 provides all the steps in the data collection and annotation process and presents statistics related to the LIMUC dataset.

Chapter 4 analyzes the cross-entropy loss function's limitations for ordinal classification problems, emphasizing its inability to factor in the ordinal relationship among classes. This chapter reviews alternative methods, including binary sub-classification and enforcing unimodal distributions, highlighting their respective drawbacks and challenges. It details the proposed loss function CDW-CE and explains how it achieves the desired result.

Chapter 5 provides an explanation of the experimental design, underscoring the extensive comparisons made across the different CNN architectures and loss functions to ensure an objective and comprehensive evaluation. Additionally, the chapter details the strategies employed in data split, model evaluation metrics, and training strategies, enhancing the reliability of the performance results.

Chapter 6 provides the results of the experiments, comments on findings, and their discussions.

Finally, Chapter 7 outlines the main points of this study and provides guidance for future work.

CHAPTER 2

LITERATURE REVIEW

The literature review in this thesis aims to provide an overview of the existing research in two main areas: (1) computer-aided diagnosis of UC and (2) loss functions for ordinal classification/regression tasks. These two domains are crucial for understanding the context and motivation behind our research, as our study lies at the intersection of ulcerative colitis diagnosis, computer-aided diagnosis systems, and ordinal loss functions. By reviewing the literature in these fields, we will establish a solid foundation for our proposed approach, identify the gaps in current knowledge, and highlight the potential of our study to make significant contributions to the state of the art in computer-assisted ulcerative colitis severity estimation. The following subsections will present a comprehensive review of the relevant literature in each domain, discussing the key advancements, challenges, and opportunities for future research.

2.1 Computer Aided Diagnosis of Ulcerative Colitis

In recent years, there has been a significant surge in research focusing on the automated detection and assessment of UC using CAD. The application of machine learning techniques, particularly deep learning models such as CNNs, has led to the development of numerous methods for the severity assessment of UC from endoscopic images and videos. This literature review aims to provide a comprehensive overview of the current state-of-the-art approaches in this rapidly evolving field. We will summarize various studies that have employed different CNN architectures, preprocessing steps, and datasets to tackle the challenges associated with UC diagnosis.

Alammari et al. [16] used a simple 9-layers (4 convolution layers, 4 pooling layers, and 1 fully connected layer) CNN architecture to classify UC severity for four severity levels (normal, mild, moderate, severe). They sampled two frames per second from the 328 colonoscopy videos, which resulted in 92614 frames. The authors applied preprocessing steps to distinguish informative frames from non-informative ones, which include blurring, specularity, and other artifacts. After discarding the non-informative frames, 17534 images from 65 videos were used to train the model and performance measurement was performed using 10004 images from 29 videos. They reported an accuracy of 67.9%. Furthermore, individual frame scores were averaged to obtain a final score for the video, and a Pearson correlation coefficient of 0.68 is obtained. The authors reported that the proposed system can classify a 128×128

image in 25 milliseconds, which makes the system usable in real-time. Although the authors employed a naive CNN, this is the first study that employed deep learning in the context of UC activity estimation from endoscopic images. Tejaswani et al. [17] advanced the previous study using the same dataset with a more rigorous preprocessing step, refinement of UC severity classes, and a more advanced model, AlexNet [18]. The authors discarded frames that contain large amounts of water and bubbles, have excessive specular reflection, and have uneven illumination. They subdivided each class of UC severity, and in total, 14 classes were generated to train the CNN architecture. These classes, then, were mapped to normal, mild, moderate, and severe for the performance evaluations. They reported an accuracy of 60.6%, and the Pearson correlation coefficient for the video-level scores was 0.94.

Maeda et al. [19] used endocytoscopy (EC) data to predict histologic inflammation. EC images were labeled using the biopsy samples' histologic activity, which was obtained after the EC procedure. 12900 EC images were used to train the model, and 9935 images were used for the validation. In total, 312 features extracted from images were used to train a support vector machine (SVM), which has two diagnostic classes, namely active and healing. Overall diagnostic sensitivity, specificity, and accuracy were reported as 74%, 97%, and 91%, respectively. The reported performance values indicate that the proposed CAD system is capable of estimating histologic inflammation.

Ozawa et al. [20] used GoogLeNet [21] to classify endoscopic images into three classes, namely Mayo 0, Mayo 1, and Mayo 2-3. 26304 images from 444 unique patients were used to train the model, and 3981 images from 114 patients were used as a validation set. The authors reported a classification accuracy of 70.4%, an AUROC value of 0.86 for differentiating Mayo 0 from Mayo 1-3 and 0.98 for differentiating Mayo 0-1 from Mayo 2-3. Although promising results were reported, having relatively small patients that have severe inflammation (13 patients and 1 patient for the training and validation sets, respectively) requires more experimentation on larger datasets for reliable performance results.

Stidham et al. [22] employed Inception-v3 [23] model for distinguishing both remission state from moderate-to-severe disease and exact Mayo subscore. 16514 images from 3082 unique patients were used in this study, where 90% was used for the model development, and 10% was used as a test set. A weighted Kappa score of 0.84 was obtained for the agreement between the CAD model and human reviewers. For distinguishing endoscopic remission from the moderate-to-severe disease, CNN obtained an AUROC of 0.966, a sensitivity of 83.0% and a specificity of 96.0%. Moreover, images of colonoscopy videos were aggregated to predict the video-level score by applying threshold rules. The authors reported that 25 out of 30 videos were correctly classified.

Takenaka et al. [24] developed a deep neural network for endoscopic images of UC, DNUC, which is based on Inception-v3 [23] model, to predict endoscopic remission (yes/no), histologic remission (yes/no), and Ulcerative Colitis Endoscopic Index of Severity (UCEIS) score. The DNUC model was developed using 40758 images and 6885 biopsy results from 2012 patients. The authors validated the study in a prospective study with 4187 images and 4104 biopsy specimens from 875 patients. A sensitivity of 93.3%, a specificity of 87.8%, an accuracy of 90.1%, and a kappa score of 0.798 were obtained for the endoscopic remission classification. Predicting histologic remission demonstrated similar performance. The authors reported an interclass correlation coefficient (ICC) of 0.917 between DNUC and endoscopists' scores. Regarding the subscores of UCEIS, 0.868, 0.796, and 0.851 ICCs were obtained for the vascular pattern, bleeding, and erosions, respectively.

Bhambhvani et al. [25] used UC images in publicly available HyperKvasir dataset [26] to train a ResNeXt-101 model [27]. The CNN model was trained with a total of 777 labeled Mayo 1, Mayo 2, and Mayo 3 images. The authors reported that the overall accuracy of the model was 77.2%, sensitivity was 72.4%, and specificity was 85.7%.

Gottlieb et al. [28] estimated MES and UCEIS directly for the full-length colonoscopy videos. They used the dataset resulting from the mirikizumab clinical trial (ClinicalTrials.gov ID: NCT02589665), which consists of 795 full-length colonoscopy videos from 249 patients from 14 countries. They filtered the video frames to extract visually clear images using specialized CNNs, which are responsible for out-of-colon filtering, fuzzy frame filtering, bad prep filtering, and abnormality feature extraction. The resulting frames were fed into 2-dimensional RNN [29] that consists of long-short-term-memory cells [30]. The model's overall performance of QWK was 0.844 for MES and 0.855 for UCEIS. This study showed that rather than labeling the still-frames individually, full-length colonoscopy videos can be utilized to predict endoscopic activity. Using the proposed end-to-end design, a significant workload of manual annotation of individual frames can be eliminated.

Yao et al. [31] developed a fully automated video analysis system to grade endoscopic activity. First, they trained a detector based on the Inception-v3 model [23] to classify informative and non-informative frames. The authors quantitatively showed that using informative image classification to extract clear images improved the overall MES classification performance. MES for each informative frame was obtained using a pretrained model, which is detailed in their previous work [22]. For each video, the ratios of MES were extracted, and thresholds for each MES class were used to assign an overall video score. The proposed system correctly predicted 40 of 51 (78%) internal videos with a QWK of 0.840 and 151 of 264 (57.1%) clinical trial videos (clinical trial ID: NCT02762500) with a QWK of 0.59. When clinical trial videos went through a dual review process, and only 169 of them were included, accuracy increased to 82.8% (140 of 169) with a QWK of 0.78. This study clearly demonstrates that when performance evaluations are performed on samples from different distributions, much lower values are obtained.

Huang et al. [32] employed a pretrained Inception-v3 model [23] on ImageNet dataset [33] to extract features and feed them into three different classifiers: deep neural network (DNN), SVM and k-nearest neighbor (k-NN). 856 colonoscopy images from 54 patients were used to train the different classifiers. The main outcome measures were to differentiate Mayo 0-1 from Mayo 2-3 and Mayo 0 from Mayo 1. The ensemble of the three classifiers resulted in 94.5% accuracy with a sensitivity of 89.2% and a specificity of 96.3%. The proposed system differentiated Mayo 0 and Mayo 1 with an accuracy of 89.1%, sensitivity of 82.3%, and specificity of 92.2%.

Becker et al. [34] proposed an end-to-end fully automated system to train binary classifiers to discriminate if a MES of an entire colon section is above or below a certain grade (MES >= 1, MES >=2, MES >=3). First, they trained a quality control model to differentiate readable and non-readable images. Then, they utilized a weak label approach by assigning the Mayo score of each colon section to readable still images. The final MES for the entire colon section was determined by averaging the scores of all frames. 1672 videos from 1105 patients from 28 countries were used to train a ResNet50 [35] model for the classification of the MES. The proposed system achieved an AUROC of 0.84 for MES >=1, an AUROC of 0.85 for MES >=2, and an AUROC of 0.85 for MES >=3.

Schwab et al. [36] employed a similar weak label approach for the entire colonoscopy video as Becker et al. [34] did for the colon sections. Moreover, they incorporated different ordinal regression frameworks to increase the overall performance. They used a UNIFI clinical trial of Ustekinumab [37], which consists of 1881 endoscopic videos from 726 subjects. This is the first study that utilized ordinal regression approaches for the estimation of endoscopic activity of UC. The proposed approach obtained a QWK of 0.68 for the video-level MES estimation and a QWK of 0.66 for the frame-level MES estimation.

Harada et al. [38] proposed a semi-supervised learning (SSL) method that utilizes location and temporal ordering information of the colonoscopy images to classify UC images as positive or negative (normal). 7183 images were used as a training set, and only 10% of them were used by the proposed SSL approach. The authors reported an accuracy of 84.5% and an F1 score of 75.3%. Although the reported performance results are lower than the supervised learning approach (88.5% accuracy and 82.6% F1 score), considering that only 10% of the labels of the training set are used, this was a promising result in the context of incorporating SSL techniques in this domain.

Maeda et al. [39] has evaluated the real-time use of AI for predicting the clinical relapse of UC. The employed AI system is based on their previous work [19], which is trained to classify images into two categories ("Active" or "Healing"). 135 patients were evaluated using AI; 74 patients were diagnosed as the *AI-Active* group, and 61 patients were diagnosed as the *AI-Healing* group. During the 12-month follow-up, relapse occurred in 28.4% of the AI-Active group and 4.9% of the AI-Healing group. The authors stated that real-time use of AI has the potential to help clinicians in their decision-making regarding treatment.

Unlike previous studies in the automatic estimation of UC severity, Luo et al. [40] proposed a new method called UC-DenseNet. On top of the DenseNet architecture, they utilized an attention mechanism and RNN in two different branches, then the outputs of the two branches were concatenated. They compared the proposed approach with existing methods on both the internal dataset consisting of 14306 images and the Kvasir [16] dataset consisting of 1000 UC images. The authors reported that the proposed CNN architecture has superior performance compared to the previously used methods.

Sutton et al. [41] compared several commonly used CNN models on publicly available HyperKvasir [26] dataset. The authors performed two binary classification tasks: 1) distinguishing UC from non-UC pathology on endoscopic frames, and 2) distinguishing inactive/mild (Mayo 0 and 1) from moderate/severe activity (Mayo 2 and 3). The authors reported that when comparing UC with non-UC pathologies, all models achieved high predictive performances; when comparing remission to moderate-to-severe activity, DenseNet121 [42] and Inception-v3 [23] models got the highest results, which is AUROC of 0.90.

Kadota et al. [43] proposed a cost-effective approach in terms of labeling. Instead of labeling individual Mayo scores of still-frames (*absolute-labels*), their approach utilizes labels that represent ranking between different image pairs (*relative-labels*). Since comparing two images with each other is much faster than assigning a score for a given image, the total annotation process for the whole dataset takes much less time (overall, 10% of the conventional method). Their proposed system utilizes RankNet [44] approach using relative-labels and a small set of absolute-labels in multi-task learning settings. The authors reported that the proposed approach even performs better than the conventional methods (0.578 vs. 0.559 for F1 score).

Xu et al. [45] used additive angular margin loss (ArcFace) [46] to train a DNN for the classification of UC severity. The ArcFace loss uses feature embeddings and weights in the last fully connected layer to improve the discriminative power of DNNs. The authors trained ResNet-152 [35], DenseNet-161 [42], and EfficientNet [47] with different scaling (b0, b1, b3, b4, b7) with ArcFace function and demonstrated improved performance compared to the cross-entropy loss. The authors used HyperKvasir [26] dataset in their experiments.

2.2 Ordinal Classification

Ordinal classification is an active field of research in machine learning due to the prevalence of ordinal categories in numerous real-world problems, particularly within the healthcare domain. The intrinsic nature of ordinal categories, where the order of the labels carries valuable information, necessitates specialized methodologies and techniques for optimal prediction performance. As a result, many studies have been conducted to develop innovative approaches to address these challenges, including loss functions for CNNs. These studies show that the proposed methods for ordinal classification give much better results than the classical methods. In this section, we will summarize the recent advancements targeting ordinal loss functions for CNNs.

Niu et al. [48] proposed an end-to-end deep learning approach for ordinal regression problems, specifically for age estimation from face images. They transformed ordinal regression into a series of binary classification sub-problems and employed a multiple-output CNN to solve these tasks jointly. Output nodes are responsible for if

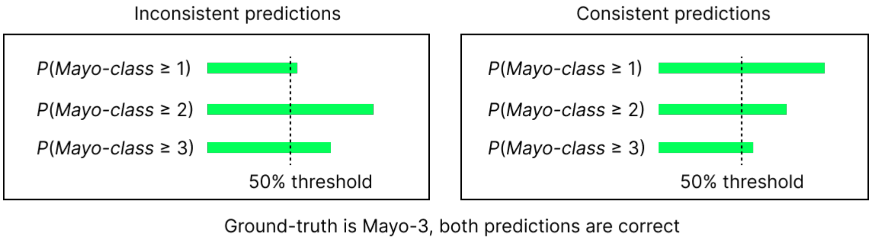


Figure 1: In the binary classification approach, output nodes are independent of each other; therefore, inconsistencies in their individual predictions may occur. Although both predictions are correct for the case in the figure, the prediction on the right-hand side is more ideal.

they are greater than a certain rank or not. For N classes, $N-1$ output nodes are sufficient for the output layer, where the label extension should be applied to the target class for this approach. The authors observed improved performance in comparison to other ordinal regression techniques, such as metric learning and the prevalent cross-entropy loss function. Despite the enhanced results achieved by the proposed method, inconsistencies in the ranking of the outputs were present in the output classification subtasks. Additionally, they published the Asian Face Age Dataset (AFAD), containing over 160000 facial images with precise age labels, making it the largest public age dataset at the time.

Cao et al. [49] proposed a consistent rank logits (CORAL) framework for rank-inconsistencies. The authors addressed the issue of classifier inconsistencies in neural network-based implementations of extended binary classification approaches. Normally, the confidence of the model predictions should follow a non-increasing order when the rank increases; however, in Niu et al.'s work, this is not guaranteed (see Figure 1). The main difference between this framework compared to Niu et al.'s work is that during the training, weight sharing (except the bias term) is applied in the penultimate layer. They showed that the CORAL framework offers theoretical guarantees for classifier consistency. By implementing CORAL in common CNN architectures like ResNet, they demonstrated improved predictive performance in age estimation tasks compared to the previous approaches.

Shi et al. [50] introduced the Conditional Ordinal Regression for Neural Network (CORN) framework, which aims to enhance the capacity of neural networks by loosening the constraint on the penultimate layer of the CORAL framework and incorporating conditional probabilities. The authors conducted experiments on various datasets such as MORPH-2 [51], AFAD [48], AES [52], and FIREMAN [53] and reported that the performance of the CORN method surpassed earlier approaches.

A significant drawback of techniques similar to CORN is the necessity to modify both the model architecture (output layer) and the labeling structure. Another line of approach for the ordinal classification problems is integrating unimodality distribution on the model's output predictions. This method enforces unimodality by penalizing inconsistencies in the posterior probability distribution between neighboring labels.

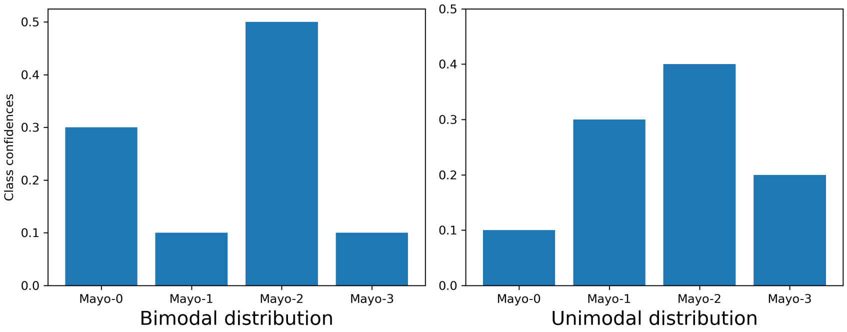


Figure 2: Assuming Mayo-2 is the true class, the CE loss function provides the same loss for both cases. The unimodal distribution is more intuitive and provides more realistic results.

Typically, the penalizing term is incorporated alongside the primary loss function, with cross-entropy being the most commonly used.

Belharbi et al. [54] proposed a non-parametric ordinal loss for neural networks that aims at promoting output probabilities to follow a unimodal distribution (see Figure 2). Basically, they evaluated the neighborhood couples in the output probabilities and applied a punishment if their probabilities are not compatible with the unimodal distribution. They have validated their methods on different problems such as breast cancer grading [55], predicting the decade of a photograph taken [56], and age estimation [48]. Albuquerque et al. [57] applied the same technique by adding unimodality losses onto cross-entropy (referred to as CO2 in Equation 1) and entropy losses (referred to as HO2 in Equation 3) for the final loss function. The Herlev dataset [58], consisting of 917 images of individual cervical cells in different stages of the disease, is used in their experiments, along with a range of CNN architectures. Both papers reported superior performance when compared to Niu et al.'s work [48] and CE loss.

CO2 and HO2 losses are frequently used for comparison in this study:

$$\mathrm{CO2}\left(y_n, \hat{y}_n\right) = \mathrm{CE}\left(y_n, \hat{y}_n\right) + \lambda \sum_{k=0}^{K-1} \mathbf{1}(k \geq k_n^*)\mathrm{RELU}(\delta + \hat{y}_{n(k+1)} - \hat{y}_{n(k)}) + \tag{1}$$

$$\lambda \sum_{k=0}^{K-1} \mathbf{1}(k \leq k_n^*)\mathrm{RELU}(\delta + \hat{y}_{n(k)} - \hat{y}_{n(k+1)}) \tag{2}$$

$$\mathrm{HO2}\left(y_n, \hat{y}_n\right) = \mathrm{H}\left(y_n, \hat{y}_n\right) + \lambda \sum_{k=0}^{K-1} \mathbf{1}(k \geq k_n^*)\mathrm{RELU}(\delta + \hat{y}_{n(k+1)} - \hat{y}_{n(k)}) + \tag{3}$$

$$\lambda \sum_{k=0}^{K-1} \mathbf{1}(k \leq k_n^*)\mathrm{RELU}(\delta + \hat{y}_{n(k)} - \hat{y}_{n(k+1)}) \tag{4}$$

where CE and H refer to Cross entropy and Entropy losses, respectively, y_n is ground truth, $\hat{y}_n$ is model predictions, K is the total number of classes, λ is a parameter that determines the strength of the unimodal loss, and δ is the margin term.

An alternative group of methods involves utilizing regression to estimate a single continuous value at the output or applying a sigmoid activation function on top of it to constrain the prediction within the range of [0, 1]. Subsequently, thresholds or probability distributions are employed to transform the output into discrete levels. Beckham et al. [59] proposed a method that adds another layer consisting of a single node on top of the final layer and employs squared-error loss. They reported that if the final output is processed through a sigmoid function followed by multiplication of K (Number of classes) -1, it gives better results. For the inference, they simply rounded the predicted value to the nearest integer. They have compared their method on a diabetic retinopathy dataset [60] and reported better performance compared to the standard CE loss. However, regression-based approaches have been shown to be inferior to other methods in many studies [57, 54].

Table 1: Detailed overview of studies in CAD of UC domain. In studies that have multiple datasets, labeling as A and B is performed to show their performance results separately in 2 and 3.

Study (year)	Dataset	Model Development Set	Test Set	Model
Alammari et al. (2017) [16]	Private, single center	17534 images from 65 videos (A) 32753 images from 116 videos (B)	10004 images from 29 videos	9-layers CNN
Tejaswani et al. (2019) [17]	Private, single center	29841 images from 254 videos	14925 images from 62 videos	AlexNet
Ozawa et al. (2019) [20]	Private, single center	26304 images from 444 patients	3981 images from 114 patients	GoogLeNet
Stidham et al. (2019) [22]	Private, single center	14862 images from 2778 patients[3]	1652 images from 304 patients 11432 images from 30 videos	Inception-v3
Maeda et al.(2019) [19] [1]	Private, single center	12900 images from 87 patients	9935 images from 100 patients	SVM
Takenaka et al. (2020) [24]	Private, single center	40758 images from 2012	4187 images from 875 patients	Inception-v3
Bhambhvani et al. (2021) [25]	Public (HyperKvasir), single Center	90% of 777 images from 777 patients	10% of 777 images from 777 patients	ResNext-101
Gottlieb et al. (2021) [28] [2]	Private, multi center	80% of 795 videos from 249 patients[3]	20% of 795 videos from 249 patients[3]	Proprietary algorithm, RNN
Yao et al. (2021) [31] [4]	Private, multi center	16000 images from 3000 patients [3]	51 videos (A) 264 videos from 157 patients (B)	Inception-v3
Huang et al. (2021) [32]	Private, single center	70% of 856 images from 54 patients	30% of 856 images from 54 patients	Inception-v3, SVM, k-NN
Becker et al. (2021) [34]	Private, multi center	80% of 1672 videos from 1105 patients[3,4]	20% of 1672 videos from 1105 patients[3,4]	ResNet50
Schwab et al. (2021) [36]	Private, multi center	80% 1881 videos from 726 patients[3,4]	20% 1881 videos from 726 patients[3,4]	ResNet34
Harada et al. (2021) [38]	Private, single center	7183 images for training 2052 images for validation	1027 images	-
Maeda et al.(2021) [39] [5]	Private, single center	44097 images	135 patients	SVM
Luo et al. (2022) [40]	Private, single center	80% of 9928 images (A) 80% of 4378 images (B) A+B: 1317 patients	20% of 9928 images (A) 20% of 4378 images (B) A+B: 1317 patients	UC_DenseNet
Sutton et al. (2022) [41]	Public (HyperKvasir), single center	80% of 2642 images (A)3 80% of 840 images (B)3	20% of 2642 images (A)3 20% of 840 images (B)3	DenseNet121
Kadota et al. (2022) [43]	Private,single center	80% of 10265 images[3]	20% of 10265 images[3]	DenseNet169 % RankNet
Polat et al. (2022) [61]	Public (LIMUC), single center	9590 images from 462 patients[3]	1686 images from 85 patients	DenseNet121
Polat et al. [15]	Public (LIMUC), single center	9590 images from 462 patients[3]	1686 images from 85 patients	Inception-v3

[1] This work is based on Endocytoscopy data.

[2] In this study, the DL model is trained with video-level labels.

[3] Cross-validation is applied for model performance assessment.

[4] MES estimations were performed for the whole video, not still frames.

[5] Contact microscopy.

Table 2: Performance results of the studies (Part 1).

Study (year)	Outcome Measures	Performance result (MES based)	Remission Estimation (Mayo 0-1 vs. Mayo 2-3)	Histologic Remission	Frame Scoring
Alammari et al. (2017) [16]	MES	Macro accuracy (A): 0.676 Macro accuracy (B): 0.436	-	-	Frame, Video
Tejaswani et al. (2019) [17]	MES	Macro accuracy: 0.606	-	-	Frame, Video
Ozawa et al. (2019) [20]	MES (Mayo 0, Mayo 1, and Mayo 2-3)	Accuracy: 0.704*	AUROC: 0.980 Accuracy: 0.946*	-	Frame
Stidham et al. (2019) [22]	MES	Kappa: 0.840 Accuracy: 0.778*	AUROC: 0.970 Accuracy: 0.917* Sensitivity: 0.830 Specificity: 0.960	-	Frame, Video
Maeda et al.1 (2019) [19]	Histologic inflammation estimation (active vs. healing)	-	-	Accuracy: 0.910 Sensitivity: 0.740 Specificity: 0.970	Frame
Takenaka et al. (2020) (2020) [24]	Histologic remission Endoscopic remission UCEIS	-	Kappa: 0.798 Accuracy: 0.901 Sensitivity: 0.933 Specificity: 0.878	Kappa: 0.859 Accuracy: 0.929 Sensitivity: 0.924 Specificity: 0.935	Frame
Bhambhvani et al. (2021) [25]	MES estimation (Mayo 1, Mayo 2, and Mayo 3)	Accuracy: 0.772 Sensitivity: 0.724 Specificity: 0.857	-	-	Frame
Gottlieb et al.2 (2021) [28]	MES UCEIS	QWK: 0.844 Accuracy: 0.702* Sensitivity: 0.716** Specificity: 0.901*	Accuracy: 0.866*	-	Video
Yao et al. (2021) [31]	MES	QWK (A): 0.840 Accuracy (A): 0.780 QWK (B): 0.590 F1 (B): 0.571	Accuracy (B): 0.837	-	Video
Huang et al. (2021) [32]	MES 0-1 vs. MES 2-3 MES 0 vs. MES 1	-	Accuracy: 0.945 Sensitivity: 0.892 Specificity: 0.963	-	Frame

* Performance metrics marked * are calculated using the reported numerical values in the study.

Table 3: Performance results of the studies (Part 2).

Study (year)	Outcome Measures	Performance result (MES based)	Remission Estimation (Mayo 0-1 vs. Mayo 2-3)	Histologic Remission	Frame Scoring
Becker et al. (2021) [34]	MES 0 vs. MES 1-2-3 MES 0-1 vs. MES 2-3 MES 0-1-2 vs. MES 3	-	AUROC: 0.850 Precision: 0.850 Recall: 0.810	-	Video
Schwab et al. (2021) [36]	MES	QWK: 0.680 (video-level) QWK: 0.660 (frame-level)	-	-	Frame, Video
Harada et al. (2021) [38]	UC vs. normal	Accuracy: 0.845 F1: 0.753 Specificity: 0.899	-	-	Frame
Maeda et al. (2021) [39]	Healing group vs. Active group	-	-	-	Frame, Video
Luo et al. (2022) [40]	MES	Accuracy (A): 0.906 F1 (A): 0.868 Accuracy (B): 0.916 F1 (B): 0.858	Accuracy (A): 0.976 F1 (A): 0.976 AUROC (A): 0.975 Accuracy (B): 0.989 F1 (B): 0.989 AUROC (B): 0.988	-	Frame, Video
Sutton et al. (2022) [41]	MES 0 vs MES 1-2-3 (A) MES 0-1 vs MES 2-3 (B)	-	F1: 0.913 Accuracy: 0.875 Sensitivity: 0.790 Specificity: 0.910	-	Frame
Kadota et al. (2022) [43]	MES	QWK: 0.578 Accuracy: 0.720	-	-	Frame
Polat et al. (2022) [61]	MES	QWK: 0.854 F1: 0.697 Accuracy: 0.772 Sensitivity: 0.693 Specificity: 0.911	Kappa: 0.827 F1: 0.858 Accuracy: 0.957 Sensitivity: 0.974 Specificity: 0.876	-	Frame
Polat et al. (2022) [15]	MES, Remission	QWK: 0.8678 F1: 0.7261 Accuracy: 0.7880	Kappa: 0.8598 F1: 0.8847 Accuracy: 0.9590	-	Frame

* Performance metrics marked * are calculated using the reported numerical values in the study.

CHAPTER 3

LABELED IMAGES FOR ULCERATIVE COLITIS DATASET

In the field of machine learning research, datasets play a fundamental role as they are the core part of developing, training, and evaluating various algorithms and models. The availability of high-quality, diverse, and representative datasets is essential for driving advancements in the field and fostering a deeper understanding of complex problems. The process of annotating a medical dataset is laborious and demands the involvement of numerous experts, with a strong emphasis on attention and accuracy. As detailed in Section 2, the majority of studies conducted on the computer-aided diagnosis (CAD) of UC have relied on private datasets. To our knowledge, the HyperKvasir dataset [26] is the sole publicly available resource containing labeled UC images. However, its limited size, comprising only 851 images, has hindered its widespread adoption in related studies. All in all, the practice of using private datasets has several drawbacks that hamper the advancement of research in this field:

- **Reproducibility.** It becomes difficult for other researchers to reproduce the same result, leading to a lack of validation of the proposed work. As a result, the study becomes questionable.

- **Comparison.** It prevents transparent comparisons between different studies and methodologies, which creates uncertainty about which method works well and which works poorly, and complicates the transition of these studies to clinical use.

- **Advancement of technology.** It is difficult for researchers who cannot reach the necessary datasets but have sufficient technical knowledge to advance in this field. So, research and development in this field becomes less democratic.

In light of these, it was necessary to form a labeled UC dataset. As a result, we have created the largest publicly available UC dataset, LIMUC, which has been utilized extensively in our research. LIMUC aims to foster more effective and accurate machine-learning models for the automated diagnosis and treatment of UC. In this section, details on the data collection and labeling process will be shared.

3.1 Ulcerative Colitis and Mayo Endoscopic Score

Inflammatory bowel disease (IBD) encompasses a group of disorders characterized by chronic inflammation of the gastrointestinal (GI) tract. The two primary types of IBD are UC and Crohn's disease (CD). These chronic, lifelong conditions can be managed through treatment but are currently incurable. UC, in particular, is characterized by continuous inflammation and ulcers in the lining of the colon and rectum, causing symptoms such as abdominal pain, diarrhea, and rectal bleeding. The disease's severity and extent can vary among individuals, necessitating personalized approaches to treatment and ongoing monitoring. UC typically starts in the rectum and lower colon before progressively extending throughout the entire colon, as illustrated in Figure 3.

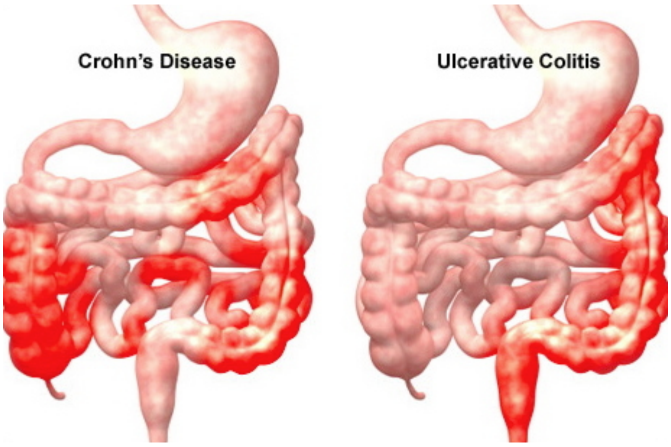


Figure 3: Example patterns of affected parts of CD and UC [1].

Endoscopic procedures, such as colonoscopy and sigmoidoscopy, are vital in evaluating the endoscopic activity of both UC and CD. These procedures enable medical professionals to visualize the entire colon using a slender, flexible, lighted tube equipped with an attached camera. During the examination, the physician may also collect small tissue samples for histopathological analysis. The assessment of IBD does not rely on a single gold standard method; instead, experts consider various patient data sources, including blood tests, endoscopic evaluations, genetic factors, and histopathological examination results, to determine an activity score. Among these indicators, endoscopic assessment plays a pivotal role in the overall evaluation of IBD.

There are several scoring systems for the assessment of UC from endoscopic images. Among these, MES is the most widely used scoring system to assess the disease severity of the UC [62]. Moreover, it is also one of the most reliable scoring systems in terms of low intra- and interobserver variability [4]; therefore, it is chosen as the main scoring system for this study. MES system evaluates the stage of UC based only on endoscopic examination, and each Mayo grade is given according to the frequency of certain symptoms and patterns on the tissue as seen in Figure 4.

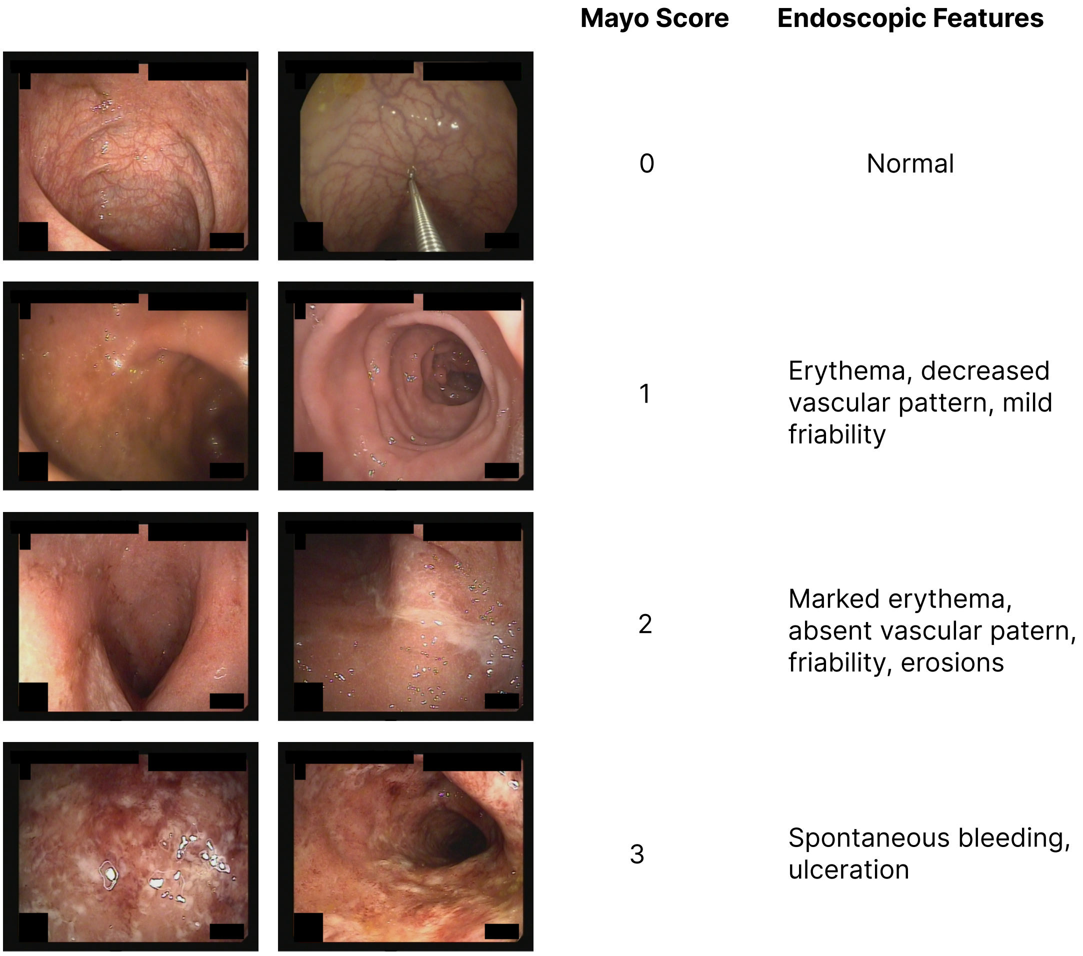

Mayo Score
Endoscopic Features
0
Normal
1
Erythema, decreased vascular pattern, mild friability
2
Marked erythema, absent vascular patern, friability, erosions
3
Spontaneous bleeding, ulceration

Figure 4: Example UC images from the LIMUC dataset and their corresponding scores according to the MES grading systems.

When experts perform an assessment, they mainly try to answer the questions for the following three descriptors:

- **Vascular Pattern:** How much obliteration exists in the vascular pattern? Are capillaries clearly defined, or is there complete obliteration? Is there an erythema in the tissue?
- **Bleeding:** Is there a bleeding? If yes, how prevalent and spontaneous is that?
- **Erosions and Ulcers:** Is the mucosa normal? If there are erosions, what are their sizes? Are the ulcers superficial or deep?

Table 4: Characteristics of the cohort (n=561, three patients have no information in the hospital's database). Values are in mean ± standard deviation format.

Male/female	**317 (56.5%) / 244 (43.5%)**
Age (year)	43.3 ± 13.7
Male	44.3 ± 13.7
Female	42.1 ± 13.7
Colonoscopies per patient	2.0 ± 1.2
Male	2.0 ± 1.3
Female	1.9 ± 1.2

3.2 Data Collection and Processing

A total of 19537 endoscopic images were collected from 1043 colonoscopy procedures involving 572 UC patients who underwent colonoscopy at Marmara University Institute of Gastroenterology between December 2011 and July 2019. During this time interval, the gastroenterologists applying the colonoscopy procedures captured some frames, and these frames were recorded in the hospital's database. All images were captured using a Pentax EPK-i video processor and Pentax EC-380LKp video colonoscope (Pentax, Tokyo, Japan) and were resized to a resolution of 352×288 when added to the database. The images to be captured were chosen by the operator at various moments throughout the colonoscopy procedure: as a result, there is no spatial connection among the images belonging to the same colonoscopy procedure. This leads to greater heterogeneity, resulting in a more diverse dataset. The research design and all data obtained from electronic health records received approval, prior to this study, from the Ethical Review Board of Marmara University School of Medicine (Study Protocol No: 09.2020.627, Approval date: 12.06.2020). Figure 5 shows a sample image for each Mayo class from the dataset. The original version of the images contained certain information such as patient-id and date/time. Before the images were made public, these regions were covered with black pixels. Moreover, the captions of the images were changed to make them anonymized.

Along with the images, patient age and sex information is drawn from the database of the hospital. Statistics related to the study cohort are given in Table 4. The mean age for the cohort is 43.3 ± 13.7 years, with males having a slightly higher mean age of 44.3 ± 13.7 years, while females have a mean age of 42.1 ± 13.7 years. On average, each patient underwent 2.0 ± 1.2 colonoscopies, with both male and female patients exhibiting similar numbers of colonoscopies per patient (2.0 ± 1.3 for males and 1.9 ± 1.2 for females).

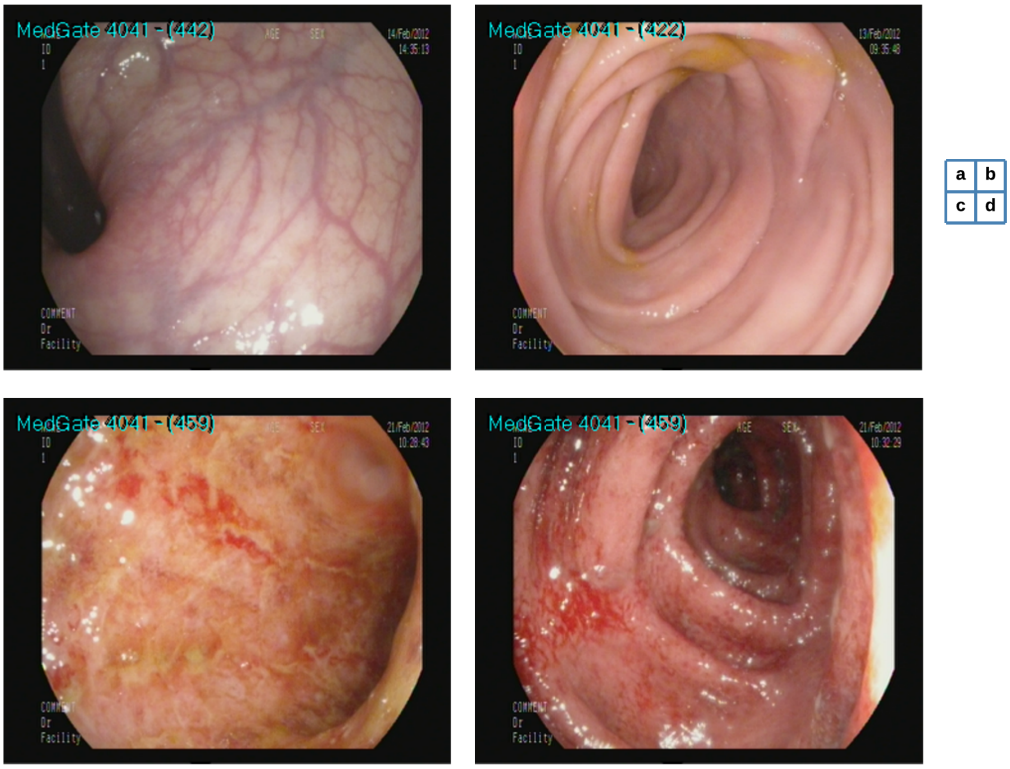


Figure 5: Sample images from the LIMUC dataset (**a**: Mayo-0, **b**: Mayo-1, **c**: Mayo-2, **d**: Mayo-3).

3.3 Data Labeling

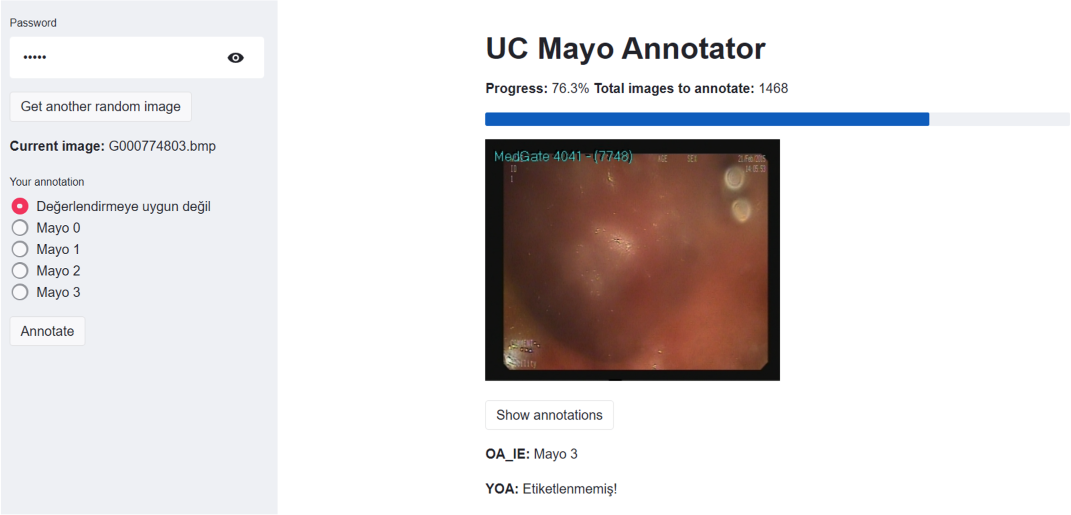
The whole dataset was initially given to two experienced gastroenterologists, who specialized in IBD, to be annotated. The annotators were asked to label images into five different classes, namely, ‘not suitable for evaluation’, ‘Mayo-0’, ‘Mayo-1’, ‘Mayo-2’, and ‘Mayo-3’. MES was chosen as the scoring system as it is the most common and reliable grading system that is used to assess the disease severity of UC [62, 4]. The annotators simply put the currently evaluated image into a folder named with the target class when performing the annotation. Results of the review process were presented in Table 5. The total number of images to evaluate was 19537 for both reviewers. Reviewer-1 considered 7621 images as not suitable for evaluation, while Reviewer-2 deemed 9207 images as not suitable. There were 11916 and 10330 images evaluated as one of the Mayo scores by Reviewer-1 and Reviewer-2, respectively, with 5720 images being commonly assessed by both reviewers. The interreader reliability for the annotation was measured with quadratic weighted kappa (QWK) score and obtained as 0.781. 7652 images, which are annotated differently by two reviewers and at least one reviewer has annotated as one of the Mayo classes, were reviewed by a third experienced gastroenterologist. The third reviewer independently annotated these differently labeled images as one of the five classes without observing the previous annotations. A web-based user interface for labeling the contradictory samples was prepared for the third reviewer, as seen in Figure 6. Final scores were determined using majority voting. Table 6 shows the breakdown of annotations of the third reviewer. Out of 7652 images, 1895 images were assessed as not suitable to annotate, and 5757 images were evaluated into one of the Mayo scores by the third reviewer. The third reviewer had differing Mayo annotations from the previous two reviewers for 201 images, in which both the first two reviewers also assigned a Mayo score. The remaining 5556 images were in agreement with one of the initial reviewers. Combining these images with the 5720 images from the first and second reviewers’ common annotations, the final dataset consisted of 11276 images. The **Final** column in Table 6 corresponds to the resulting dataset, which is published as the LIMUC dataset. The LIMUC dataset is shared on the Zenodo platform[1] under the Creative Commons Attribution 4.0 license.

Table 5: Distributions of annotation among different classes.

	Reviewer-1	**Reviewer-2**	**Common**
Total images to evaluate	19537	19537	-
Evaluated as not suitable for evaluation	7621	9207	-
Evaluated as one of the Mayo scores	11916	10330	5720
Mayo-0	7398	4503	3472
Mayo-1	2473	3796	1210
Mayo-2	1190	1014	470
Mayo-3	855	1017	568

---

[1] `https://doi.org/10.5281/zenodo.5827695`



Password

•••••

Get another random image

Current image: G000774803.bmp

Your annotation

● Değerlendirmeye uygun değil
○ Mayo 0
○ Mayo 1
○ Mayo 2
○ Mayo 3

Annotate

UC Mayo Annotator

Progress: 76.3% **Total images to annotate:** 1468

Show annotations

OA_IE: Mayo 3

YOA: Etiketlenmemiş!

Figure 6: User interface designed for the third reviewer to annotate images.

Table 6: Distribution of annotations of the third reviewer

	Reviewer-3	**From Reviewer 1&2**	**Final**
Total images to evaluate	7652	-	-
Evaluated as not suitable to annotate	1895	-	-
Evaluated into one of the Mayo scores	5757	-	-
Differently Annotated from the other reviewers	201	-	-
To join dataset (agreement with one of the observers)	5556	5720	11276
Mayo-0	2633	3472	6105
Mayo-1	1842	1210	3052
Mayo-2	784	470	1254
Mayo-3	297	568	865

Statistics related to both the original dataset and annotated dataset are presented in Figures 7 and 8. On average, each patient has 1.8 colonoscopy procedures in the data collection time period, and for each colonoscopy, 18.8 images exist on average for the dataset (19537 samples); as a result, there were nearly 34 images on average per patient before the annotation process. The final dataset has an imbalanced structure in terms of class sizes, where Mayo-0 is 54.14%, Mayo-1 is 27.07%, Mayo-2 is 11.12%, and Mayo-3 is 5.67% of all annotated images. Since some of the images were excluded during the annotation process, the average number of images per patient has fallen to 20.

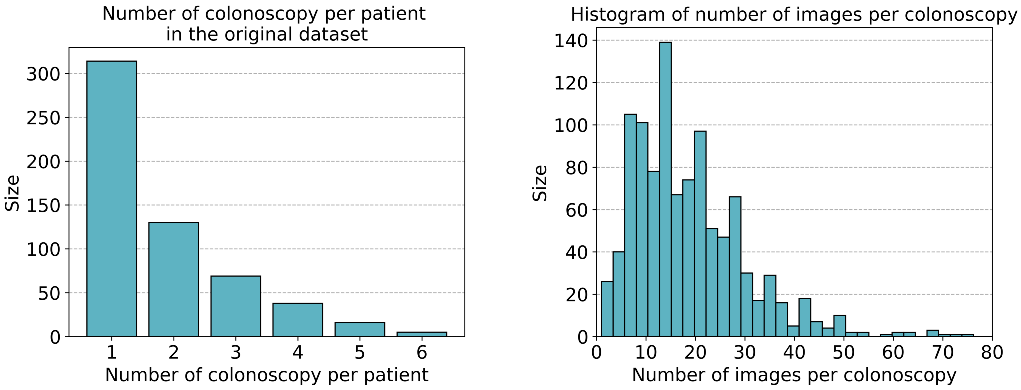


Figure 7: Left: Number of colonoscopy operations per patient in the original dataset (mean: 1.8). Right: Histogram of number of images per colonoscopy (mean: 18.8).

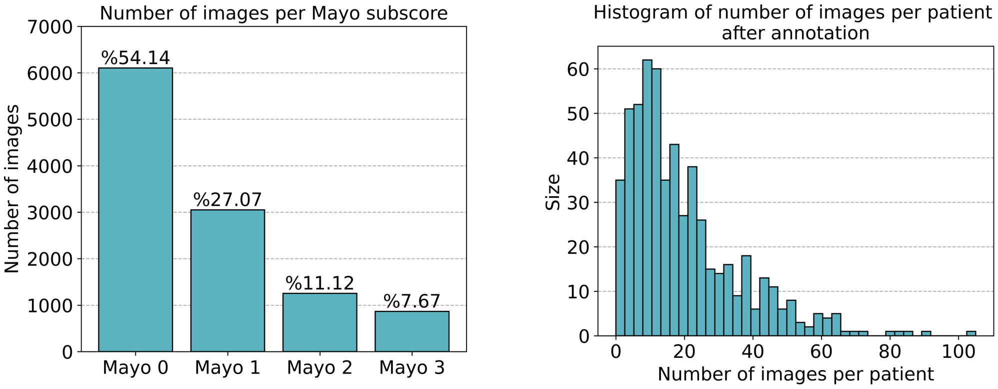


Figure 8: Left: Distributions of images among Mayo classes. Right: Histogram of the number of images per patient for the final dataset (mean: 20).

CHAPTER 4

CLASS DISTANCE WEIGHTED CROSS ENTROPY

4.1 Motivation

The Cross-entropy (CE) loss function (Equation 5) is one of the most commonly used loss functions in classification problems. When using CE loss, the output layer of the DL model contains as many nodes as the number of classes, where each node corresponds to a different class. Then, the model's output predictions are converted to a scalar loss value using the ground-truth values.

$$\text{CE} = -\sum_{i=0}^{N-1} y_i \times \log \hat{y}_i = -\log \hat{y}_c \tag{5}$$

In Equation 5, i refers to the index of the class in the output layer, c is the index of the ground-truth class, y is the ground-truth label, and $\hat{y}$ refers to the prediction. This widely adopted approach for tackling classification tasks misses very critical information that exists among the ordinal classes, which is the ranking of the classes. Given that one-hot encoding is employed for ground-truth labels at the output layer, y_i becomes 0 for $\forall i \neq c$. Consequently, the CE loss only evaluates the predicted confidence of the true class. However, in scenarios where an ordinal relationship exists among output classes, whole probability distributions should be taken into account. For instance, within an ordinal class structure ranging from 0 to 9, a prediction of 0 for class 9 is considerably more detrimental than predicting 8. However, CE loss results in exactly the same value for those two different predictions (given that their output confidences are the same). An improved loss function should assess the ranking and impose a greater penalty on predictions that deviate further from the true class (see Table 7). Due to its inability to penalize predictions at a greater distance from the correct classes more heavily than those closer, the CE loss function is suboptimal for ordinal classification problems.

In Section 2.2, several approaches addressing this issue were presented. However, some of these approaches bring different problems within themselves. For example, approaches that transform the multiclassification problem into binary sub-classification tasks, such as CORAL [49] and CORN [50], require a change in model architecture and labeling structure, so they cannot be used out-of-box. Although the approaches that use enforcing unimodal distributions [54, 57] can be used for the existing ar-

Table 7: Consider the following three sample cases with the same cross-entropy loss, where Class 0 is the ground truth, and the classes exhibit an ordinal relationship. An ideal loss function for ordinal classification should assign the lowest cost to Case 1, indicating the most favorable outcome, while allocating the highest cost to Case 3, signifying the least favorable outcome.

Classes	Case 1	Case 2	Case 3
0	0.6	0.6	0.6
1	0.3	0.1	0
2	0.1	0.3	0.1
3	0	0	0.3

Table 8: Approaches enforcing unimodality distribution result in the same loss value for the two different cases, where Class 0 is the ground truth, and the classes exhibit an ordinal relationship. For both cases, the loss value is calculated in transitions from 0 to 0.19 and from 0.19 to 0.21.

Classes	Case 1	Case 2
0	0.6	0.6
1	0.0	0.0
2	0.19	0.0
3	0.21	0.19
4	0.0	0.21

chitectures, they only evaluate the neighborhood couples ordering, which makes it insufficient in assessing the distance of the predictions to the real classes (see Table 8). Moreover, regression-based approaches have been shown to be inferior to other ordinal losses in many studies [48, 49, 50, 54, 57].

4.2 Class Distance Weighted Cross Entropy Loss Function

We introduce a novel, non-parametric loss function, denoted as CDW-CE (6), which evaluates the confidences of non-true classes as opposed to focusing on the true class confidence. Our proposed methodology incorporates two primary components. First, we impose a penalty that reflects the degree of deviation of each misprediction from the actual value, employing log loss as the metric. Given that one-hot encoding is the standard encoding technique for class labels in multi-class classification problems, the predicted confidences for non-true classes are expected to be zero. Thus, as the mispredictions move away from the zero value, the loss increases. Secondly, we incorporate a coefficient into the loss calculation for each class. This coefficient leverages the distance to the ground-truth class and exhibits an increasing trend with respect to the distance. This added component ensures that our proposed loss function effectively accounts for the ordinal nature of the classes and penalizes mispredictions more heavily as they deviate further from the true class. We use a power term α in the loss coefficient that determines its strength. The power term α in the loss coefficient plays a crucial role in our proposed loss function. As a hyperparameter, alpha governs the degree of penalization with respect to the distance between predicted and ground-truth classes. As alpha increases, the influence of the distance grows, leading to stronger penalization for mispredictions that deviate further from the true class.

$$\text{CDW-CE} = -\sum_{i=0}^{N-1} \log(1-\hat{y}_i) \times |i-c|^{\alpha} \tag{6}$$

where the notation is the same as in Eqn. 5. CDW-CE loss is differentiable; therefore, it can be directy used in backpropogation calculation:

$$\frac{d(L)}{d(\hat{y}_i)} = -\sum_{i=0}^{N-1} \frac{d}{d(\hat{y}_i)}(\log(1-\hat{y}_i) \times |i-c|^{\alpha})$$

Notice that only the term where the index i is in the summation will have a nonzero derivative. For all other terms, the derivative will be zero. So, we can rewrite the expression as:

$$\frac{d(L)}{d(\hat{y}_i)} = -\frac{d}{d(\hat{y}_i)}(\log(1-\hat{y}_i) \times |i-c|^{\alpha})$$

$$\frac{d(L)}{d(\hat{y}_i)} = -|i-c|^{\alpha} \times \frac{d}{d(\hat{y}_i)}(\log(1-\hat{y}_i))$$

$$\frac{d(L)}{d(\hat{y}_i)} = -|i-c|^{\alpha} \times \left(-\frac{1}{1-\hat{y}_i}\right)$$

$$\frac{d(L)}{d(\hat{y}_i)} = \frac{|i-c|^{\alpha}}{1-\hat{y}_i} \tag{7}$$

As seen in Equation 7, when the prediction value deviates further away from its target value of zero, the derivative term increases.

The implementation of CDW-CE with the PyTorch framework is presented in Appendix A.

CHAPTER 5

EXPERIMENTAL SETTINGS

In this section, details of the experimental environment are given as it is foundational for the comparison of the results. An important deficiency in both the CAD of UC and ordinal classification literature is that in most of the studies, the proposed methods are not compared with each other sufficiently. For example, in section 2.1, in a large portion of all studies, there is no comparison with other CNN architectures. Likewise, in section 2.2, proposed ordinal losses were only evaluated against naive methods such as CE or regression-based losses. In this work, we perform a comprehensive evaluation by training different CNN architectures over different loss functions, and in this chapter, we give the details of the experimental settings.

5.1 Model Training and Evaluation

In our study, we allocated 15% of the images (comprising 1686 images from 85 patients) as the test set, ensuring that the class (i.e., MES) ratios remained consistent with the overall study group. We performed the splitting at the patient level, assigning all images from a single patient to either the test set or the model development set. The remaining 85% of images (which included 9590 images from 479 patients) were utilized for 10-fold cross-validation.

For each fold, the training and validation split was conducted at the patient level, with random selection while maintaining class ratios, mirroring the approach used for the test set. CNN architectures, trained using different cross-validation folds, were evaluated on a separate, held-out test set (refer to Figure 9). We report performance metrics as the mean values derived from the 10-fold cross-validation results.

We assessed model performance based on two baselines: full 4-level Mayo score classification and remission state classification, which consists of two levels (remission: Mayo 0 or 1, and non-remission: Mayo 2 or 3). While the CNNs were trained exclusively for full Mayo scores, we converted model predictions to remission states for remission classification purposes. We compared model predictions with ground-truth labels provided by human experts to measure performance.

Given the presence of class imbalances and ordinal relationships among them, we identified the Quadratic Weighted Kappa (QWK) score as the primary performance metric but also reported other related metrics such as macro F1 score, accuracy, and mean-absolute score (MAE) for the classification of all Mayo scores. For remission classification, we report Cohen's kappa, accuracy, and F1 scores. The QWK is a commonly used statistic for evaluating agreement on an ordinal scale and serves as one of the most suitable singular performance metrics for this problem, considering class imbalances.

The QWK is calculated based on a confusion matrix of the model's predictions. It involves the following steps:

- **Confusion Matrix:** First, a confusion matrix is created that outlines the actual versus the predicted classifications.

- **Weights Matrix:** Next, a weights matrix is created as in Equation 8. This matrix is the same shape as the confusion matrix, but it's filled with the squared difference between the actual and predicted ratings.

$$w_{i,j} = \frac{(i-j)^2}{(N-1)^2} \tag{8}$$

- **Expected Matrix:** This is a matrix showing what we would expect the confusion matrix to look like if the predictions were random. It's calculated as the outer product of the row and column totals of the confusion matrix.

- **Normalization:** Both the confusion matrix and the expected matrix are normalized by dividing each by the total number of observations.

Finally, QWK is calculated as in Equation 9.

$$QWK = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}} \tag{9}$$

where O is the correlation matrix and E is the expected matrix.

We employed three widely recognized CNN architectures, namely ResNet18 [35], Inception-v3 [23], and MobileNet-v3-large [63], for training with various loss functions. Both ResNet and Inception model families have been extensively utilized for estimating UC severity in the literature [22, 20, 24, 25, 34, 36, 31]. MobileNet-v3-large, a more recent model, distinguishes itself through its exceptional speed and performance, rendering it an ideal candidate for real-time UC severity estimation derived from video frames. Therefore, measuring performance on these three models is compatible with the literature and the domain.

For data augmentation, we incorporated random rotation ($0^\circ - 360^\circ$) and horizontal flipping techniques and initialized the weights using pretrained models from the ImageNet dataset [33]. Due to class imbalance among the dataset, when forming the batch for each forward-pass operation, an equal number of samples from each class is ensured to overcome class imbalance problems. We employed the Adam optimizer

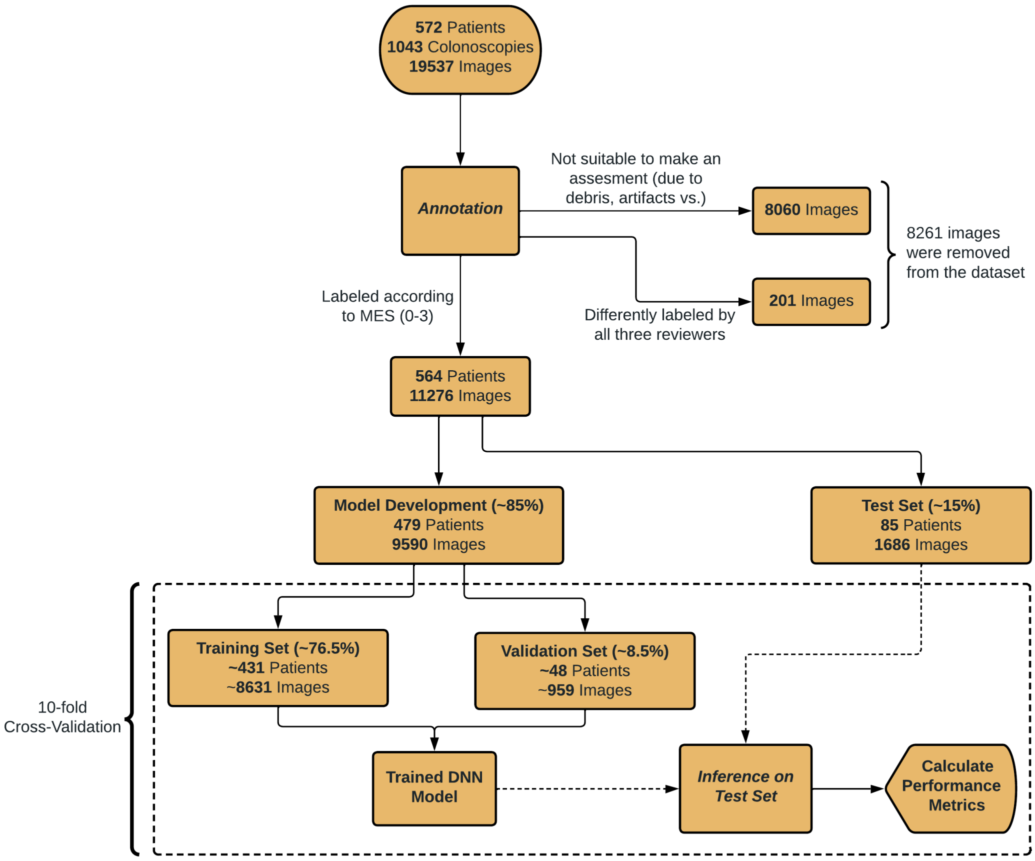


Figure 9: Data splitting for the model training and evaluation. Image is reprinted from the author's publication in [61].

[64], featuring a learning rate of $2e4$, and implemented learning rate scheduling with a scaling factor of 0.2 when no improvement in validation set accuracy was observed over the last 10 epochs. Early stopping was utilized to terminate training if performance did not enhance during the previous 25 epochs. The optimal model checkpoint on the validation set for each fold was employed to measure test set performance. The training and testing code is implemented using PyTorch [65] and CNN architectures obtained from the Torchvision library [66].

The initial version of the images incorporated specific details, including software ID and date/time information. Before the images are used for the training, a black-out operation is applied for these regions to prevent any bias in the model (Figure 10).

In the interest of promoting transparency and ensuring the replicability of the experiments presented in this thesis, all source code associated with the research has been made publicly available. This initiative aligns with the growing trend in the scientific community to encourage open research practices, which facilitate the validation and extension of findings by other researchers.

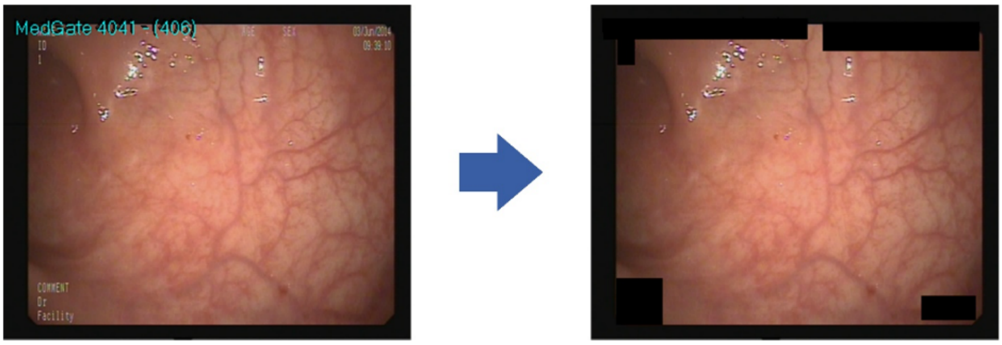


Figure 10: Masking regions including text.

The source code can be accessed through a dedicated repository hosted on the GitHub platform [1], with appropriate documentation provided to enable ease of use and understanding. By releasing the source code, this research aims to contribute to the ongoing advancement of knowledge in the field and foster collaborative efforts among researchers. It is encouraged that any researchers interested in building upon or validating the findings presented in this thesis make use of this resource.

We assessed the proposed model in comparison with four distinct methodologies specifically tailored for ordinal regression tasks: the CORN framework [50], CO2 [57], HO2 [57], using regression with MSE loss, and the CE loss function, which serves as the primary baseline. For the CORN approach, three output nodes are used at the output layer, where each node is responsible for an independent binary task. In particular, node-0 predicts if the ground truth is greater than Mayo-0 or not, node-1 predicts if the ground truth is greater than Mayo-1 or not, and node-2 predicts if the ground truth is greater than Mayo-2 or not. Then, the final loss value is calculated as the sum of binary cross-entropy losses obtained from each node. Label extension is applied for the CORN loss. For the CO2 (Equation 1) and HO2 (Equation 3) losses, the main loss function (either cross-entropy or entropy loss) is scaled with a λ coefficient as in the original paper implementation [67]. Although the formulation in the paper is different, where the ordinal loss term is scaled (see Equation 1 and 3), mathematically, both formulations work in the same way, which basically adjusts the ratio between the main loss term and ordinal loss term. Tuning of the hyperparameter λ was carried out by exploring values from the set $\{0.1, 0.01, 0.001\}$, utilizing 10-fold cross-validation to evaluate the performance of each value. For the regression approach, we followed the approach of Polat et al. [61]. A single node at the output layer without any activation function is employed, and MSE is used to calculate the loss. As a baseline method, we employed standard cross entropy loss as shown in Equation 5.

Apart from the ordinal losses, other loss function-specific approaches were also evaluated in this work to see the state of the proposed approach with respect to other uti-

---

[1] `https://github.com/GorkemP/labeled-images-for-ulcerative-colitis`

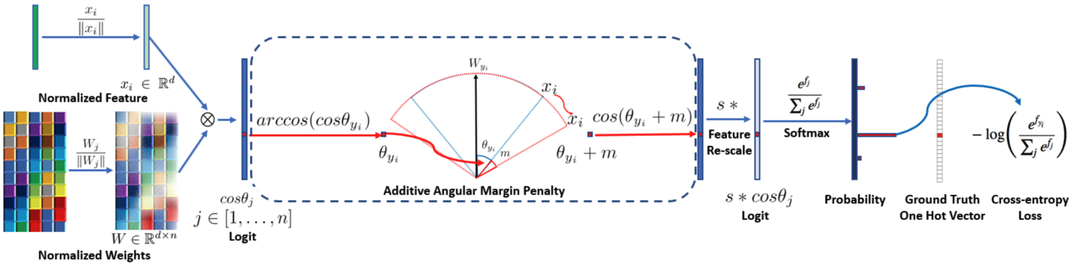


Figure 11: Training a CNN with the ArcFace loss. The margin penalty m and feature scale s are hyperparameters that need to be tuned [46].

lized losses in this domain. Recently, Xu et al. [45] employed additive angular margin loss (ArcFace) [46], which is one of the state-of-the-art techniques in face recognition domain, for training CNNs in the task of UC severity classification. ArcFace loss increases the discriminative capacity of DNNs by leveraging feature embeddings and weights in the final fully connected layer. As seen in Figure 11, a margin m angle is added to the angle between the feature vector and the weight vector. Then, the resulting logits are scaled with s. Scaled logits are processed through the softmax function and then evaluated in CE to obtain a loss value. In our study, we replicated their experiment to see its result on the LIMUC dataset and compare it with the CDW-CE.

Another line of experiments was also performed by adding a margin to the CDW-CE loss function as in Equation 10. As shown in ArcFace and several other works (CosFace [68], SphereFace [69], NormFace [70]), adding a margin to the loss function enforces intra-class compactness and increases inter-class distances. This idea is incorporated into CDW-CE loss by adding a margin to the class probabilities. Since CDW-CE penalizes the non-ground truth classes (logits, whose value must be zero), we need to incorporate the margin in an additive way. Moreover, the value of the new logit is limited to a value of one for numerical stability. Therefore, the addition of margin onto predicted confidence should be clipped at the value of one.

$$\textbf{CDW-CE with Margin} = -\sum_{i=0}^{N-1} \log(1 - max(1, \hat{y}_i + m)) \times |i - c|^{\alpha} \tag{10}$$

Some images, which were labeled differently by the first two reviewers, had to be labeled by the third reviewer. If the third reviewer agreed with one of the first two reviewers, a 2 versus 1 condition occurred, and these images were entered into the final data set. We have marked these images as *hard samples* because at least one reviewer made a different assessment on these images, which means these are challenging (or confusing) images for experts. In order to provide a more fine granular analysis, we have also provided performance results and analysis for the hard samples.

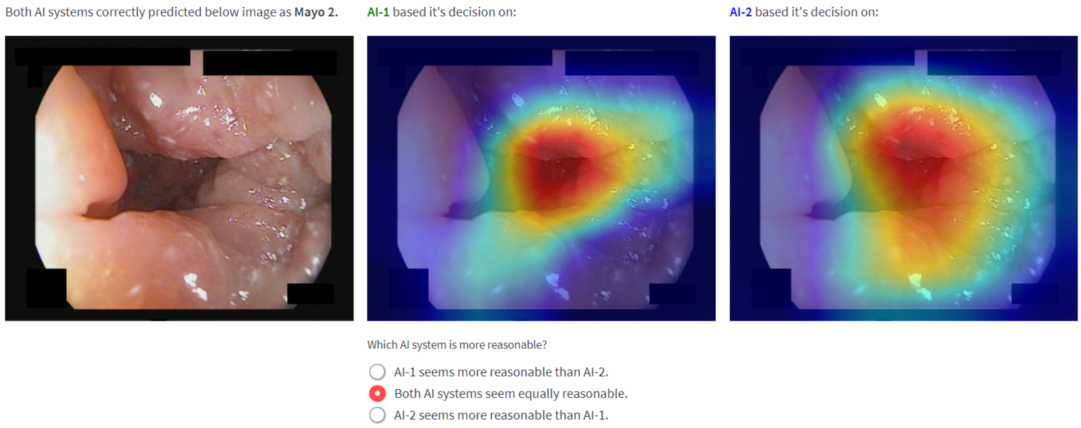
5.2 Explainability Analysis

In order to enhance the transparency and interpretability of CNN models, a variety of visualization techniques have been proposed in recent literature, including Class Activation Mapping (CAM) [71] and Gradient-weighted Class Activation Mapping (Grad-CAM) [72]. The ability to visualize the most salient regions that a model utilizes for making predictions is particularly crucial within the medical domain, where the alignment of a model's decision-making process with that of domain experts is of paramount importance. As a result, models that demonstrate similar focus areas as experts are more likely to gain trust and adoption among end-users.

The employment of CAM visualizations as a comparative criterion enables the assessment and selection of models based on their interpretability, especially when their performance metrics are comparable. In such cases, models with more reasonable and justifiable activation maps could be favored over others, despite having similar performance outcomes. Furthermore, these visualizations offer developers a valuable tool for debugging their approach, identifying any potential biases or issues in the model's prediction process, and ultimately refining the model's performance [72].

In this study, we have generated CAM visualizations using the methodology outlined by Zhou et al. [71]. It is important to note that CAMs are specifically generated for each class, highlighting only the class-specific discriminative regions corresponding to the target class. In this context, we have compared CDW-CE's explainability characteristics with the widely used CE loss function. This comparison aims to elucidate further the advantages and potential improvements offered by our proposed loss function in the realm of deep neural network interpretability.

To conduct a quantitative and objective evaluation of the CAMs generated by the models trained with CDW-CE and CE loss functions, two ResNet18 models trained with CE and CDW-CE losses are used to generate CAMs for different images in the test set. Only the predictions, which were correctly predicted by the two models, were included in the comparison. Then, the CAMs of these images were evaluated by three IBD experts independently. These experts were asked to assess the compatibility of the CAMs with symptomatic areas in the tissue, i.e., to determine which CAM is more closely aligned with the regions they would consider in their own decision-making process. Additionally, the experts were provided with the option to indicate that both CAMs were equally reasonable if they were unable to discern a clear distinction between the two visualizations. We have implemented a user interface for IBD experts to make their decisions easily (see Figure 12). In the user interface, we have shown the original image, indicating its class, accompanied by CAMs. The CAM images generated by the models for each image were randomly labeled as AI-1 (Artificial Intelligence 1) and AI-2 to ensure anonymity. The clinicians were then asked to select between the three available options: 1) AI-1 seems more reasonable than AI-2, 2) AI-2 seems more reasonable than AI-1, and 3) Both AI systems seem equally reasonable. In total, the IBD experts were presented with 240 images, including 60 images from each class. This evaluation process was completed independently


Both AI systems correctly predicted below image as **Mayo 2.**

AI-1 based it's decision on:

AI-2 based it's decision on:

Which AI system is more reasonable?

○ AI-1 seems more reasonable than AI-2.
● Both AI systems seem equally reasonable.
○ AI-2 seems more reasonable than AI-1.

Figure 12: User interface provided to the experts displays the CAM visualizations alongside the image. Experts are asked to evaluate the spots used in the decision-making process of CE and CDW-CE and choose the one which they think is more reasonable to them (i.e., more aligned with their decision-making).

by each IBD expert to ensure an unbiased comparison of the CAMs produced by the CE and CDW-CE loss function-trained models.

5.3 Experimental Evaluation on an Additional Dataset: Diabetic Retinopathy

In order to further validate the generalizability and applicability of the proposed CDW-CE loss function, it is essential to examine its performance on other datasets that exhibit ordinality in their annotations. Assessing the proposed method on various datasets is critical for establishing its robustness, as well as highlighting its potential for adaptation and usage across a diverse range of medical imaging applications. Conducting such evaluations will not only provide valuable insights into the transferability of the CDW-CE loss function but also ensure that the research findings are comprehensive, thereby strengthening the overall conclusions drawn from the study.

One such suitable problem for the evaluation of the CDW-CE loss function is the diabetic retinopathy (or diabetic eye disease) assessment. Diabetic retinopathy (DR) is a common complication of diabetes that affects the blood vessels in the retina and is a leading cause of blindness worldwide [73]. The severity of DR is often assessed using ordinal grading systems [74], which makes the DR dataset an appropriate choice for assessing the performance of the proposed CDW-CE loss function in the context of ordinal classification tasks.

By conducting an extensive experimental evaluation of the CDW-CE loss function on the DR dataset, this study aims to provide a thorough and rigorous assessment of the proposed method's ability to handle ordinal classification tasks in various medical imaging domains. The results obtained from this evaluation will not only serve to reinforce the validity of the CDW-CE loss function in the context of UC severity estimation but also demonstrate its potential applicability and effectiveness in addressing other ordinal classification problems in the medical imaging field.

We utilize the dataset used in Diabetic Retinopathy Challenge [75] hosted on the Kaggle platform, which is aimed at developing automated methods for detecting DR from digital retinal images. The dataset provided for this challenge offers an excellent resource for evaluating the performance of the proposed CDW-CE loss function on ordinal classification tasks in the context of DR due to the following reasons:

- The retinal images in the dataset are labeled by clinicians using a grading system that assigns an ordinal severity level to each image. The severity levels range from 0 (no diabetic retinopathy) to 4 (proliferative diabetic retinopathy), with each level representing increasing severity of the disease (DR-0: No DR, DR-1: Mild, DR-2: Moderate, DR-3: Severe, DR-4: Proliferative DR).

- Training and test sets are large, which contain 35126 and 53576 images, respectively. This ensures a fair assessment of the model's generalization capabilities and helps to prevent overfitting.

- Images are provided in high resolutions, where the average image width is 4009 and height is 2712 in pixels.

Figure 13 shows the distribution of the samples across different DR classes. The majority of the samples belong to the DR-0 class, which has 25810 samples. This indicates that a large portion of the dataset consists of images with no signs of diabetic retinopathy. The number of samples in the other classes (DR-1 to DR-4) is significantly lower than that in the DR-0 class. The difference is particularly pronounced for the DR-1, DR-3, and DR-4 classes, which have 2443, 873, and 708 samples, respectively. This suggests that the dataset is heavily skewed toward non-diabetic retinopathy cases.

Figure 14 shows two DR images belonging to DR-0 and DR-4 images. When the severity of DR increases, hard exudates appear (yellowish spots), blood vessels grow abnormally, and hemorrhages occur on the retina wall.

The 30% of the training set is separated as the validation set, and images are downscaled into 256×256 before feeding into the model. The same augmentation, learning rate scheduling, and early stopping procedures were applied as we did in UC training. QWK score is the main performance metric to compare different approaches, which is also employed in the challenge as the target performance metric.

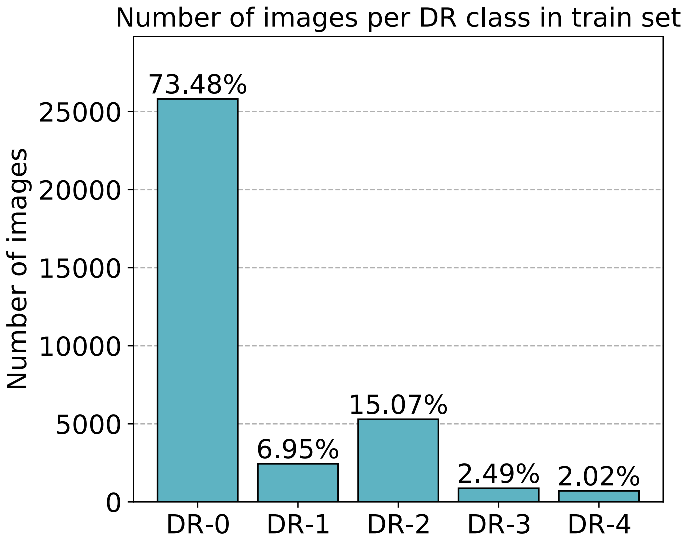


Figure 13: Class distributions of the training set of the DR dataset. There is a very high class imbalance where nearly 7 out of 10 images belong to DR-0 class, and DR-0 to DR-3 and DR-4 ratio is 32:1. The test set also follows the same pattern.

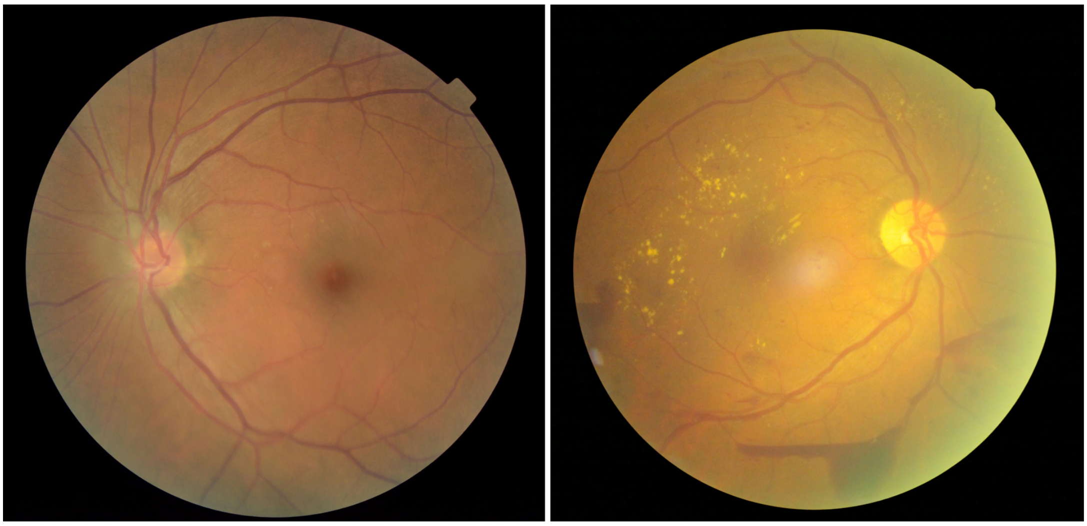


Figure 14: Two sample images from the DR dataset. **Left:** DR-0, healthy retina. **Right:** DR-4, proliferative retina.

CHAPTER 6

RESULTS AND DISCUSSION

In this chapter, we present the results obtained from the implementation of the proposed CDW-CE loss function for estimating UC severity in endoscopic images and discuss the findings in detail. The assessment of our proposed approach is conducted in comparison with several existing state-of-the-art methods, including the CORN framework, CO2, HO2, regression, and the CE loss functions. Furthermore, we explore additional experiments involving the integration of margins with CDW-CE, as well as the comparison with the ArcFace approach in UC severity estimation. Moreover, we share the results of the experimental evaluation on a diabetic retinopathy dataset. Our evaluation extends beyond model performance metrics, encompassing an in-depth analysis of explainability through CAM visualizations. The insights gained from the comparison of CAMs generated by models trained with CDW-CE and CE loss functions are subjected to expert evaluation, providing a valuable perspective on the potential improvements offered by our proposed loss function in terms of deep neural network interpretability. Through this extensive analysis, we aim to demonstrate the superiority of our novel CDW-CE loss function in facilitating more accurate, reliable, and interpretable ulcerative colitis severity estimation using CNN models.

6.1 Power Factor Analysis of CDW-CE

The power term α present in the CDW-CE loss serves as a mechanism to regularize the extent of penalization of the distant classes. With an increase in the value of α, the penalty imposed on more distant classes intensifies. However, the degree of penalization is subject to variation due to external factors, including the dataset, the number of labels, and the specific CNN model employed. Therefore, before performing a comparison of different approaches, we need to determine the optimal values of α for the CDW-CE loss function. In order to ascertain the most suitable α value for each CNN model, we conducted an extensive analysis of various α values. The outcomes reported in this chapter for CDW-CE are obtained from models trained with experimentally determined optimal α values. For each CNN model, we present the mean and standard deviation of the QWK scores corresponding to different α values in Figure 15, offering a comprehensive insight into the impact of the power term on model performance. When choosing the optimal α value, the QWK score is chosen as the target performance metric.

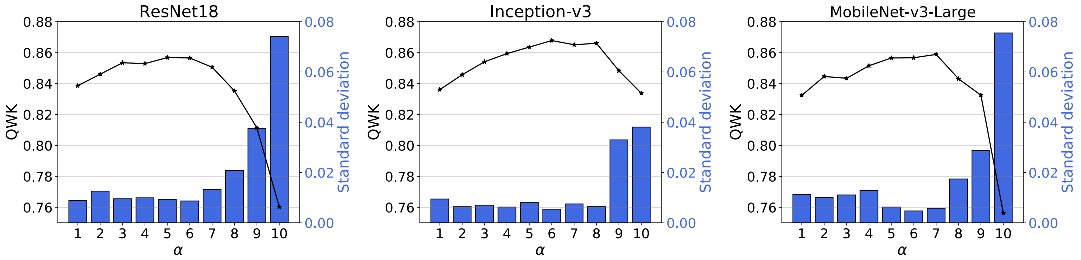


Figure 15: Change of mean and standard deviation of QWK scores according to varying α for three models.

Figure 15 demonstrates that different models may exhibit varying optimal α parameters. As the value of α increases and approaches the optimum, the model's performance increases, too; however, once the value surpasses the optimum, the accuracy of the model declines. Given that α is an exponential term, elevating it beyond the optimal value results in a substantially heightened cost coefficient, which in turn, destabilizes the training process and leads to an increase in the standard deviation of cross-validation results, as illustrated in Figure 15. The power analysis indicates that, counterintuitively, imposing a relatively high penalty on distant classes facilitates more effective optimization of model training (e.g., for $\alpha = 5$, a 2-level neighborhood coefficient of $2^5 = 32$ and a 3-level neighborhood coefficient of $3^5 = 243$, which is a very high cost coefficient compared to the 1-level neighborhood). It is worth noting that α is not an exceedingly sensitive parameter with respect to performance, as Figure 15 reveals that even training with non-optimal α values can surpass the baseline and other ordinal approaches in terms of performance. As Figure 15 shows, optimum α values for ResNet18, Inception-v3, and MobileNet-v3-large are 5, 6, and 7, respectively.

6.2 Comparison of Approaches Targeting Loss Functions

6.2.1 Ordinal Approaches

Table 9 and Table 10 show the comparison of approaches, which utilizes the ordinality information of the problem, for the full Mayo score estimation and remission state estimation, respectively.

For the full Mayo score estimation, the CDW-CE loss function consistently outperforms the other methods, achieving the highest QWK, F1 score, and accuracy, as well as the lowest MAE. In contrast, the CE loss function demonstrates the weakest performance, which indicates that this widely used loss function is not optimal and other approaches that incorporate the ordinality information are preferable. The remaining loss functions - MSE, CORN, CO2, and HO2 - exhibit intermediate performance levels, with some variations across the different CNN models. However, none of them manage to surpass the CDW-CE loss function in any of the evaluation metrics. Unimodality approaches (CO2 and HO2) compare favorably to the CORN framework for the ResNet18 and Inception-v3 models while only falling slightly behind for the MobileNet-v3-Large model, with an insignificant margin. Among the unimodality approaches, the HO2 results are mostly better than those of CO2, which aligns with the findings reported in the literature [57]. An important result presented by this comparison is that the naive MSE loss gives very good results compared to other methods except for the CDW-CE.

Due to IBD experts' interest in the binary classification of the UC disease (remission vs. non-remission), similar performance measurements were performed for the remission case as seen in Table 10. We also see a similar trend here. For each metric and across all three CNN models, the CDW-CE loss function once again demonstrates superior performance, achieving the highest Kappa, F1-score, and accuracy values consistently. The CE loss function generally performs the worst among the tested methods, with the lowest scores. However, it should be noted that it does not consistently yield the lowest scores in this case, as CORN and CO2 results are slightly lower in some metrics and CNN models. The remaining loss functions (MSE, CORN, CO2, and HO2) exhibit varying intermediate performance levels, with the MSE loss function typically attaining better results than the rest. The CO2 and CORN frameworks demonstrate closely comparable performances across the experiments. However, the HO2 method consistently surpasses them for all models, suggesting that HO2 is more effective at concentrating estimations around the true class.

Along with the summary performance metrics as shown in Table 10 and Table 9, we also present the confusion matrices for CE and CDW-CE loss to examine and compare the individual class performances in Figure 16 and Figure 17. To obtain an average confusion matrix that represents all cross-validation results, each individual confusion matrice is summed, and numbers are normalized with the total true predictions for each class (sum of rows); therefore, each diagonal cell corresponds to the recall (i.e., sensitivity or true positive rate) value for that class. Figure 16 reveals that the CDW-CE loss substantially diminishes mispredictions with a two-class distance or greater from the true class (for example, see the results of predictions for Mayo-2 and Mayo-3, which were in fact Mayo-0 or predictions for Mayo-0 and Mayo-1 which were in fact Mayo-3). CDW-CE predominantly centers incorrect estimates around classes with a one-neighborhood distance which results in higher performance. This behavior is more apparent in remission classification (see Figure 17) because, in this case, each class (remission/non-remission) corresponds to two Mayo classes, where 1-level neighborhood mistakes become unimportant (except for Mayo-2 predicted as Mayo-3 and Mayo-3 predicted as Mayo-2 because the decision boundary is between Mayo-2 and Mayo-3).

High-quality embeddings are crucial as they facilitate better representation of the dataset and class distinction, ultimately leading to improved classification performance. To compare the embeddings of a ResNet18 model trained with the CE and CDW-CE, we transformed the feature of the last fully connected layer of each image in the test set into 2D points using t-Distributed Stochastic Neighbor Embedding (t-SNE) dimensionality reduction technique [76]. Figure 18 and 19 show the t-SNE embeddings of features obtained from the ResNet18 model trained with CE

Table 9: Experiment results for all Mayo scores.

	Loss Function	**ResNet18**	**Inception-v3**	**MobileNet-v3-Large**
QWK	Cross-Entropy	0.8296 ± 0.014	0.8360 ± 0.011	0.8302 ± 0.015
	MSE	0.8540 ± 0.007	0.8517 ± 0.007	0.8467 ± 0.005
	CORN	0.8366 ± 0.007	0.8431 ± 0.009	0.8412 ± 0.010
	CO2	0.8394 ± 0.009	0.8482 ± 0.009	0.8354 ± 0.009
	HO2	0.8446 ± 0.007	0.8458 ± 0.010	0.8378 ± 0.007
	CDW-CE	**0.8568 ± 0.010**	**0.8678 ± 0.006**	**0.8588 ± 0.006**
F1	Cross-Entropy	0.6720 ± 0.026	0.6829 ± 0.023	0.6668 ± 0.028
	MSE	0.6925 ± 0.015	0.6881 ± 0.013	0.6946 ± 0.011
	CORN	0.6809 ± 0.014	0.6832 ± 0.013	0.6847 ± 0.020
	CO2	0.6782 ± 0.014	0.6846 ± 0.016	0.6793 ± 0.012
	HO2	0.6856 ± 0.016	0.6901 ± 0.008	0.6741 ± 0.030
	CDW-CE	**0.7055 ± 0.021**	**0.7261 ± 0.015**	**0.7254 ± 0.010**
Accuracy	Cross-Entropy	0.7566 ± 0.015	0.7600 ± 0.012	0.7564 ± 0.011
	MSE	0.7702 ± 0.009	0.7690 ± 0.008	0.7677 ± 0.009
	CORN	0.7591 ± 0.009	0.7600 ± 0.008	0.7613 ± 0.012
	CO2	0.7601 ± 0.008	0.7654 ± 0.008	0.7572 ± 0.009
	HO2	0.7625 ± 0.011	0.766 ± 0.010	0.7583 ± 0.005
	CDW-CE	**0.7740 ± 0.011**	**0.7880 ± 0.011**	**0.7759 ± 0.010**
MAE	Cross-Entropy	0.2581 ± 0.018	0.2526 ± 0.013	0.2563 ± 0.012
	MSE	0.2346 ± 0.009	0.2359 ± 0.009	0.2383 ± 0.009
	CORN	0.2517 ± 0.012	0.2497 ± 0.010	0.2480 ± 0.012
	CO2	0.2497 ± 0.011	0.2404 ± 0.008	0.2524 ± 0.010
	HO2	0.2460 ± 0.011	0.2424 ± 0.011	0.2487 ± 0.005
	CDW-CE	**0.2300 ± 0.011**	**0.2147 ± 0.010**	**0.2272 ± 0.011**

and CDW-CE, respectively. As seen from both figures, edge classes (Mayo-0 and Mayo-3) are well separated from each other, while intermediate classes (Mayo-1 and Mayo-2) are intertwined with their neighbor classes. Nevertheless, when these two figures are compared with each other qualitatively, it is hard to say whose representation of the dataset is better.

To quantitatively compare the embeddings generated by models trained with CDW-CE and CE, we employ the Silhouette score metric [77], which evaluates intra-class compactness and inter-class separation. A higher Silhouette coefficient score is indicative of a model with more distinct and well-defined clusters. The Silhouette coefficient is a metric that is individually computed for each sample, incorporating two distance measurements:

- d_s**:** The average distance between a sample and all other data points within the same class.
- d_n**:** The average distance between a sample and all other data points in the nearest adjacent cluster.

Subsequently, the Silhouette coefficient (s) for a singular sample is calculated using the following formula:

Table 10: Experiment results for remission classification.

	Loss Function	**ResNet18**	**Inception-v3**	**MobileNet-v3-Large**
Kappa	Cross-Entropy	0.8077 ± 0.023	0.8074 ± 0.021	0.8122 ± 0.018
	MSE	0.8406 ± 0.013	0.8404 ± 0.017	0.8339 ± 0.012
	CORN	0.8191 ± 0.021	0.8077 ± 0.022	0.8203 ± 0.016
	CO2	0.8185 ± 0.020	0.8243 ± 0.011	0.8067 ± 0.020
	HO2	0.8318 ± 0.015	0.8251 ± 0.015	0.8283 ± 0.018
	CDW-CE	**0.8521 ± 0.016**	**0.8598 ± 0.012**	**0.8592 ± 0.012**
F1	Cross-Entropy	0.8419 ± 0.018	0.8420 ± 0.017	0.8451 ± 0.016
	MSE	0.8691 ± 0.011	0.8686 ± 0.014	0.8634 ± 0.010
	CORN	0.8511 ± 0.016	0.8425 ± 0.018	0.8523 ± 0.013
	CO2	0.8513 ± 0.015	0.8561 ± 0.009	0.8404 ± 0.017
	HO2	0.8618 ± 0.012	0.8565 ± 0.011	0.8583 ± 0.015
	CDW-CE	**0.8785 ± 0.013**	**0.8847 ± 0.010**	**0.8842 ± 0.010**
Accuracy	Cross-Entropy	0.9436 ± 0.009	0.9432 ± 0.007	0.9456 ± 0.005
	MSE	0.9531 ± 0.004	0.9536 ± 0.006	0.9514 ± 0.005
	CORN	0.9473 ± 0.007	0.9429 ± 0.008	0.9473 ± 0.006
	CO2	0.9461 ± 0.008	0.9479 ± 0.004	0.9444 ± 0.006
	HO2	0.9507 ± 0.005	0.9485 ± 0.005	0.9504 ± 0.005
	CDW-CE	**0.9566 ± 0.005**	**0.9590 ± 0.003**	**0.9588 ± 0.005**

$$\mathbf{s} = \frac{d_n - d_s}{\max(d_n, d_s)} \tag{11}$$

For a collection of samples, the overall Silhouette coefficient is determined by computing the mean Silhouette coefficient (s) of each individual sample within the set. The overall Silhouette coefficient is bounded between -1 and 1, and the score is higher when the clusters are compact and distinctly separated from each other. When we compare the representations of embeddings of two models trained with CE and CDW-CE, they get a value of 0.121 and 0.222, respectively. As a result, the models trained with CDW-CE learn better representations of the dataset compared to the standard CE loss.

6.2.2 Non-Ordinal Approaches

In this study, to provide a comprehensive analysis, we not only compared ordinal loss functions but also conducted comparisons with non-ordinal loss functions that have been studied in the context of CAD of UC. As mentioned previously, ArcFace loss is compared against the proposed CDW-CE loss in this regard. The experiments are performed on the ResNet18 model following the same cross-validation settings as before. As indicated in the ArcFace paper [46], m and s parameters are sensitive to the dataset; therefore, they need to be tuned accordingly. Different combinations from the set $\{1, 2, 4, 8, 16, 32\}$ and $\{0, 0.2, 0.4, 0.5\}$ for the feature scale parameter s and margin m, respectively, are experimented with. ArcFace loss gave the best results when the scale parameter was around 4, and the margin parameter was around 0.4.

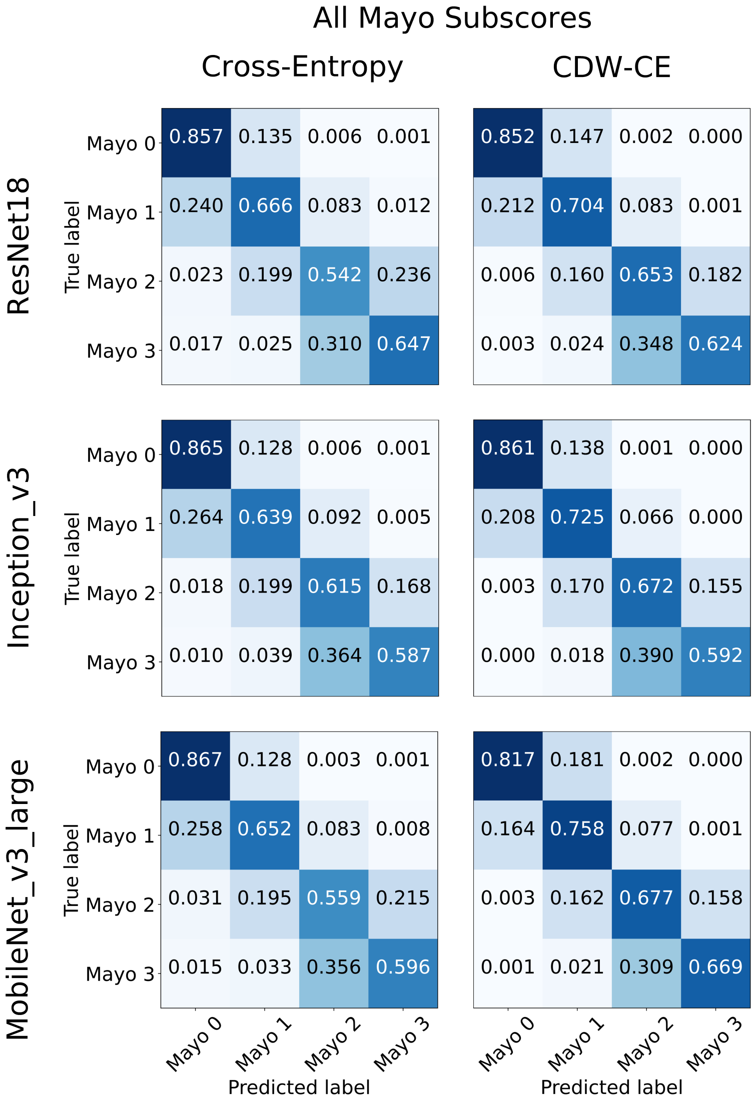


Figure 16: Mean confusion matrix of each CNN model trained with CE and CDW-CE for full Mayo score classification.

When the scale was above 8, the performance was mostly worse than the CE loss. One important point we draw from these experiments is that these two hyperparameters are highly dependent on each other; therefore, they need to be tuned together very carefully. In Table 11, we have shared the result of the top three performing combinations, where a margin of 0.5 and scaling of 1 gives the best performance. Although they are better than the baseline loss function CE, they consistently underperformed compared to CDW-CE.

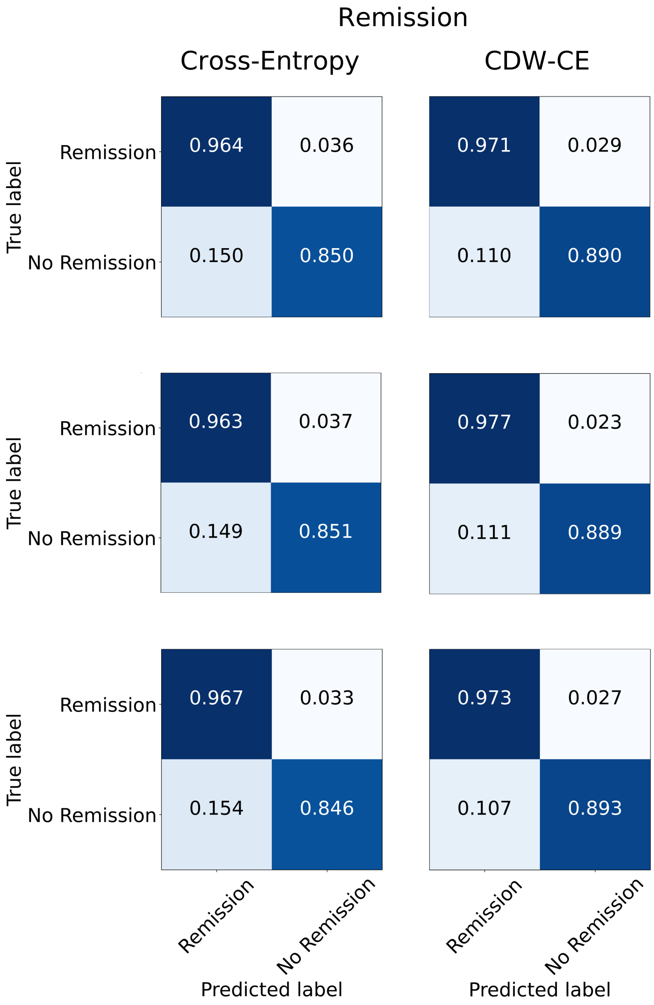


Figure 17: Mean confusion matrix of each CNN model trained with CE and CDW-CE for remission classification.

Table 11: CDW-CE vs. ArcFace. Only the top three performing results were shared for the ArcFace.

Approach	**s**	**m**	**QWK**
CE	-	-	0.8296
CDW-CE	-	-	0.**8568**
ArcFace	1	0	0.8385
ArcFace	1	0.5	0.8399
ArcFace	4	0.4	0.8371

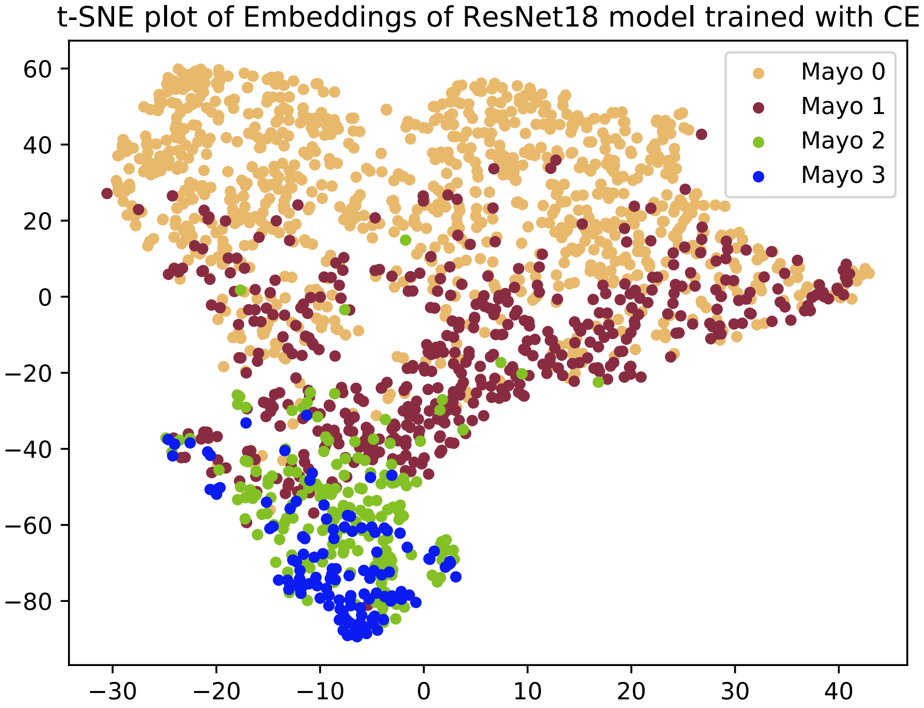


Figure 18: t-SNE embeddings of features obtained from ResNet18 model trained with CE.

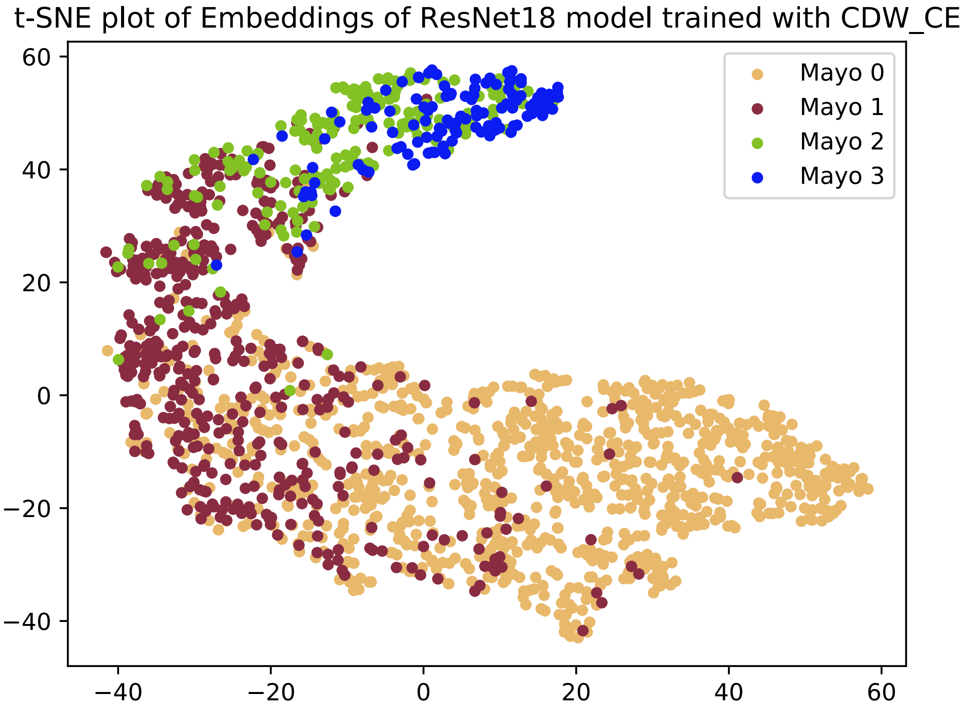


Figure 19: t-SNE embeddings of features obtained from ResNet18 model trained with CDW-CE.

6.3 Pushing the Boundaries of CDW-CE with Additive Margin

By integrating the concept of margin, which has been successful in other applications, the CDW-CE with margin (Equation 10) loss function demonstrates better performance compared to the original CDW-CE. First, in order to observe the consistency, we have obtained the result for varying α values for the ResNet18 model. Figure 20 shows the performances of CDW-CE and CDW-CE with margin. We used a margin value of 0.5, and it outperformed the base CDW-CE result until the value of 5 for the α parameter. When experimenting with $\alpha = 6$, it underperformed compared to the baseline; therefore, we have changed the margin value to obtain an improvement. Finally, a margin of 0.015 surpassed the baseline result. Eventually, Figure 20 shows that with a properly fine-tuned margin value, CDW-CE with margin can provide better results. For Inception-v3 and MobileNet-v3-large architectures, we directly tried to improve the results with the optimum α values, where they perform the best (6 for the Inception-v3 and 7 for the MobileNet-v3-large). After experimenting with different margin values, better-performing results were obtained. All in all, for all three models, the best results obtained with CDW-CE were exceeded by the additive margin strategy. Although the additive margin strategy outperforms the proposed approach, it comes with the cost of tuning an additional hyperparameter. Moreover, obtained performance increase is not that significant (0.37%, 0.34%, 0.48% for ResNet18, MobileNet-v3-large, and Inception-v3, respectively); therefore, it creates a trade-off between performance and experimental-effort.

Table 12: CDW-CE vs. CDW-CE with margin (m refers to additive margin value).

Loss function	**ResNet18**	**Inception-v3**	**MobileNet-v3-large**
CDW-CE	0.8568	0.8678	0.8588
CDW-CE w/ margin	0.8600 (m=0.05)	0.8719 (m=0.025)	0.8617 (m=0.0025)

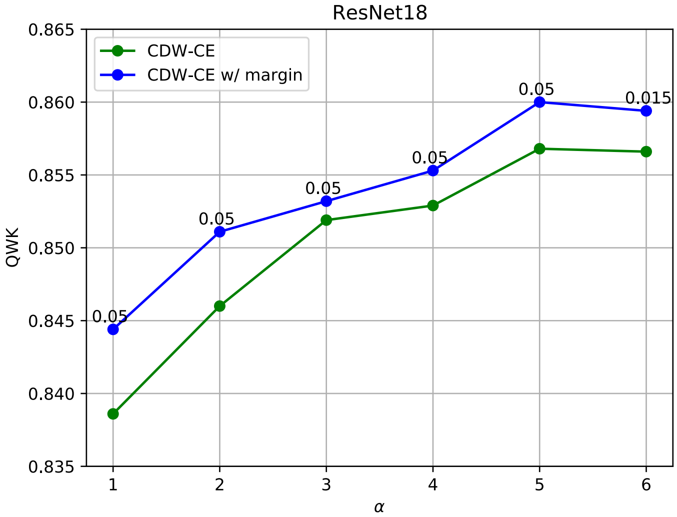


Figure 20: CDW-CE vs. CDW-CE with margin for varying α values. Numbers above the blue markers show the margin value.

6.4 Performance on the Hard Samples

A total of 863 samples are identified as hard samples (explained in section 5) in the test set (breakdown with respect to classes are given in Table 13). In Table 14, we present the results and performance increase ratios for all samples and hard samples. Performance results obtained only for the hard samples in the test set are much lower compared to all samples, which is very intuitive because these images are challenging for the human experts and, naturally, for the CNN model, too. For example, the QWK value for CE loss drops from 0.8296 to 0.7673 for ResNet18, from 0.8360 to 0.7688 for the Inception-v3, and from 0.8302 to 0.7774 for the MobileNet-v3-large model. When we obtain how much performance increase is obtained by CDW-CE compared to the CE, we observe a much more performance gain compared to when it is tested on all test samples. This situation indicates that CDW-CE loss is more effective when the samples are more challenging. In real-life scenarios, during the colonoscopy procedure, frames are taken in very different and challenging environments. So, an approach that performs much better in challenging conditions is more favorable in the clinical setting.

Table 13: Number of hard samples for each class in the test set.

Class	**Number of samples**
Mayo 0	404
Mayo 1	314
Mayo 2	105
Mayo 3	40
Total	863

Table 14: Performance increase comparison for the hard samples vs. all samples (full test set). QWK refers to full Mayo score estimation, and Kappa refers to remission estimation. The ratio column indicates how much percentage gain is obtained with CDW-CE compared to CE loss.

		ResNet18	**ratio**	**Inception-v3**	**ratio**	**MobileNet-v3-large**	**ratio**
				All samples			
QWK	Cross-Entropy	0.8296 ± 0.014		0.8360 ± 0.011		0.8302 ± 0.015	
	CDW-CE	0.8568 ± 0.010	3.40%	0.8678 ± 0.006	3.80%	0.8588 ± 0.006	3.50%
Kappa	Cross-Entropy	0.8077 ± 0.023		0.8074 ± 0.021		0.8122 ± 0.018	
	CDW-CE	0.8521 ± 0.016	5.50%	0.8598 ± 0.012	6.50%	0.8592 ± 0.012	5.80%
				Hard samples			
QWK	Cross-Entropy	0.7673 + 0.019		0.7688 + 0.015		0.7774 + 0.021	
	CDW-CE	0.8046 + 0.016	4.90%	0.8124 + 0.013	5.70%	0.8137 + 0.006	4.60%
Kappa	Cross-Entropy	0.7576 + 0.026		0.7593 + 0.030		0.7696 + 0.039	
	CDW-CE	0.8211 + 0.017	8.40%	0.8299 + 0.015	9.30%	0.8323 + 0.016	8.20%

Figure 21 shows the inference results of models trained with CE and CDW-CE on hard samples. Individual class performances are mostly lower than when the inference was performed on the full test set, which is expected. However, general trends and improvements are very similar to the full test set (see Figure 16). For all models, there are not any three-level neighborhood mistakes (Mayo-0 predicted as Mayo-3 or vice versa); moreover, 2-level neighborhood mistakes (Mayo-0 predicted as Mayo-2 or Mayo-1 predicted as Mayo-3) are very low in CDW-CE when compared to the CE loss. Naturally, this result is also reflected in confusion matrices for the remission classification (see Figure 22)

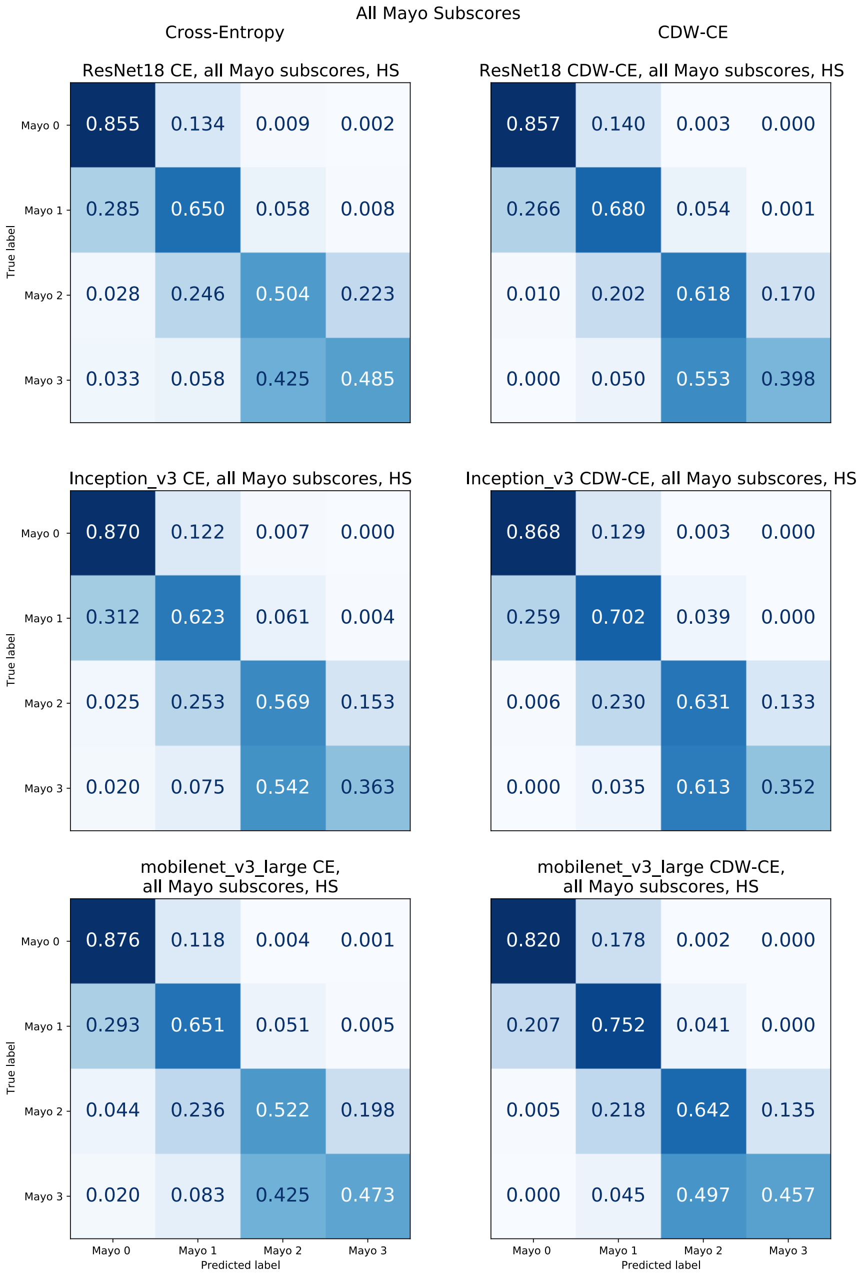


Figure 21: Confusion matrices for inference on hard samples for full Mayo score estimation.

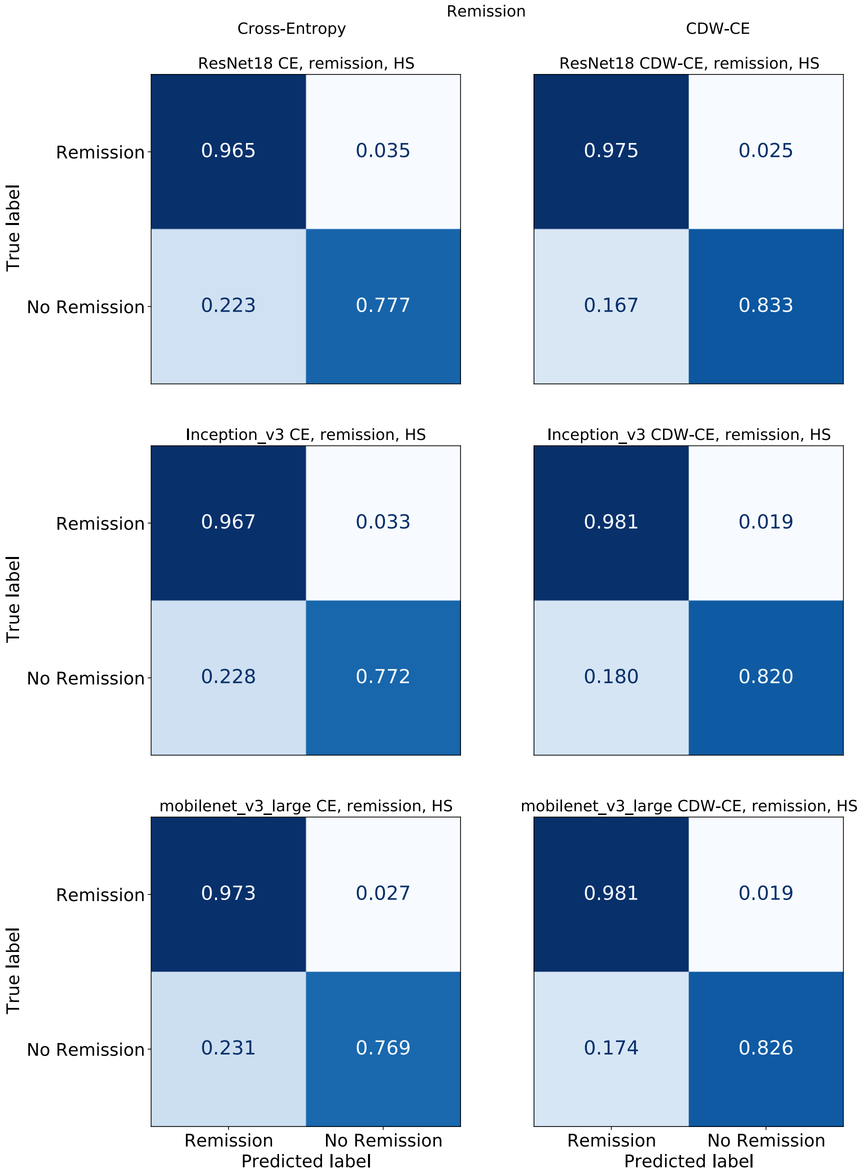


Figure 22: Confusion matrices for inference on hard samples for remission classification.

6.5 Explainability Analysis

The training of CNNs using the proposed CDW-CE loss function not only enhances their overall performance but also improves the explainability of these networks through CAM visualizations. By utilizing the CDW-CE loss function, the trained models are able to highlight more relevant and discriminative regions for all Mayo scores compared to models trained with the classic CE loss function.

A closer examination of sample CAM visualizations, as depicted in Figure 23, reveals that the CDW-CE loss function enables the model to extract features that are more compatible with disease symptoms, ultimately leading to superior performance. The CAM regions extracted by CDW-CE appear to be more extensive, and these expansions predominantly occur in relevant areas rather than unrelated ones. For example, in Figure 23, the most active regions for CAM generated with CDW-CE for the Mayo-0 class represents the underlying health tissue (where capillaries are clearly seen) better. For the Mayo-1 case, CDW-CE CAM successfully highlights the region around the lumen, and for Mayo-2 and Mayo-3 samples, the most active regions are better overlapped with the ulcers. These observations suggest that CDW-CE has effectively captured semantically meaningful features.

In order to further assess the quality of CAM visualizations, we sought the opinions of three experts who compared the visualizations produced by models trained with CDW-CE and CE loss functions. As illustrated in Figure 24, the experts unanimously found the CAM visualizations of the CDW-CE-trained model to be more reasonable across all Mayo classes. On average, the experts found that nearly half of the images were equally reasonable (47.4%), while the rate of selecting CDW-CE was twice as high as that of CE (35.0% vs. 17.6%). CDW-CE CAMs perform much better, especially in severe classes (37.8% vs. 4.4% for Mayo-2 and 24.4% vs. 12.8% for Mayo-3).

The improved interpretability provided by CDW-CE, in conjunction with its superior estimation performance, enhances the credibility and trustworthiness of CAD systems for clinical use. As the CDW-CE loss function bolsters interpretability, the transition of such systems into clinical practice is expected to be expedited.

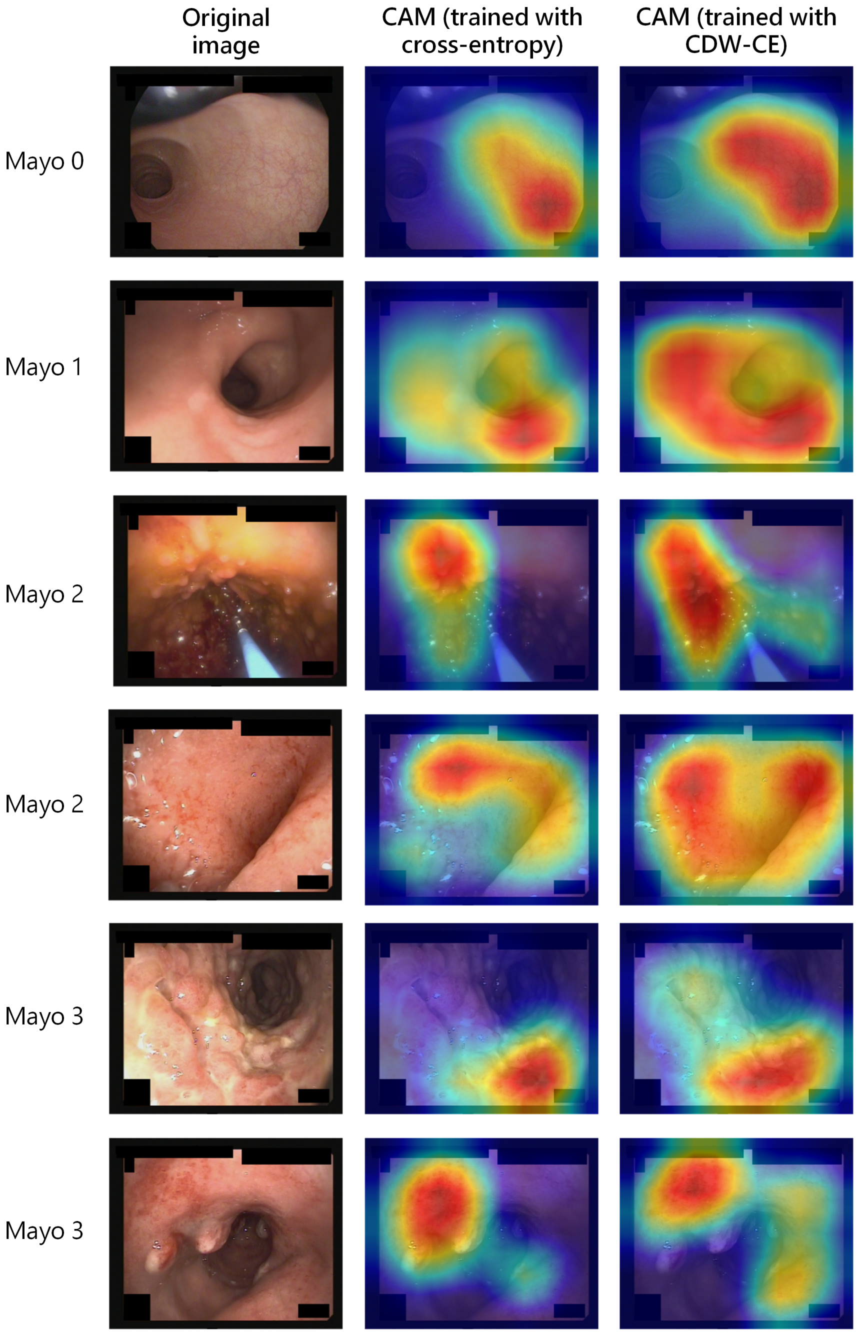


Figure 23: Original images and CAM visualizations of the ResNet18 model trained with Cross-Entropy and CDW-CE. The model trained with CDW-CE highlights broader and more relevant areas related to the disease.

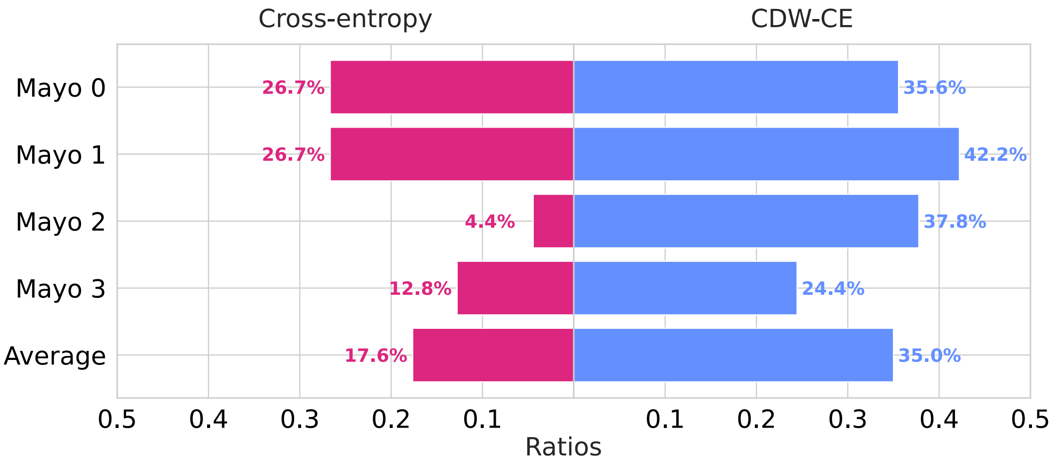


Figure 24: The assessment results of CAM visualizations for models trained with CE and CDW-CE were evaluated by experts. The percentage values representing the instances where experts found both visualizations to be equally reasonable were as follows: 37.7%, 31.1%, 57.8%, 62.8%, and 47.4%, respectively. Image is reprinted from the author's publication in [15].

6.6 Experiment Results on Diabetic Retinopathy Dataset

Table 15 shows the results of the performance comparisons with respect to the QWK score across three different CNN architectures. The CDW-CE loss shows competitive results across all three architectures. Notably, it consistently outperforms the baseline CE and other loss functions except the MSE across all models. The best-performing loss function varies with the choice of CNN architecture: MSE loss yields the highest performance for ResNet18 and Inception-v3, while for MobileNet-v3-large, the proposed CDW-CE demonstrates the superior result. However, the differences between MSE and CDW-CE are very small. The other loss functions, namely CORN, CO2, and HO2, show varying results across different architectures, but generally, they are better than the baseline CE loss, however, they can't surpass the proposed CDW-CE or the MSE loss functions. The performance of loss functions seems to vary with different CNN architectures. This observation suggests that the choice of model architecture can significantly impact the performance of the loss function, and thus, the overall performance of the model.

The reason behind this result can be attributed to the fundamental principle that both these methods, MSE and CDW-CE, essentially operate on. At their core, these two functions punish mispredictions based on distance; therefore, they may result in very similar performances. Meanwhile, CDW-CE loss has advantages, particularly in two critical aspects:

1. **Interpretability and Decision Confidence:** The CDW-CE loss function utilizes discrete output nodes, which provide class-specific confidence levels. This distinct characteristic facilitates a more nuanced and comprehensible interpretation of the model's output, thereby aiding in decision-making. On the contrary, the implementation of MSE in a CNN architecture typically incorporates a single output node, which provides a continuous numerical output. Consequently, the interpretation of the MSE output becomes challenging. For instance, deciphering the confidence level or certainty associated with a continuous output like 7.3 poses a significant interpretative challenge. Thus, the CDW-CE loss function offers an inherent advantage in providing interpretable and confidence-associated outputs.

2. **Compatibility with Class Activation Mapping (CAM) Methods:** Class Activation Mapping is an important technique in the interpretability and understanding of CNN models, as it highlights the regions in the input image that are instrumental in the model's decision-making. These mapping techniques are designed to work with individual output nodes corresponding to each class, which is the approach adopted by the CDW-CE loss function. Conversely, these methods are incapable of generating meaningful maps for architectures with a single output node, which is responsible for all classes, as is the case with MSE. Therefore, the use of CDW-CE loss further enhances the model's interpretability by enabling compatibility with Class Activation Mapping methods.

Table 15: Comparison of loss functions on diabetic retinopathy dataset.

	ResNet18	**Inception-v3**	**MobileNet-v3-large**
CE	0.6490	0.6440	0.6121
CORN	0.6608	0.6682	0.6325
CO2	0.6729	0.6753	0.6325
HO2	0.6614	0.6456	0.6099
MSE	**0.6818**	**0.6894**	0.6522
CDW-CE	0.6754	0.6867	**0.6560**

In summary, while the proposed CDW-CE loss function is not the absolute best across all architectures, it performs consistently well and shows competitive performance compared to other loss functions. This observation reinforces the efficacy of the proposed CDW-CE loss function in ordinal classification tasks across various domains. Furthermore, the results emphasize the importance of considering both the loss function and the model architecture in the design of effective deep-learning models. Despite the seemingly comparable performance of the CDW-CE and MSE loss functions, the CDW-CE provides superior interpretability and compatibility with CAM techniques, thereby offering a more holistic and practical solution in the realm of ordinal classification tasks.

CHAPTER 7

CONCLUSION AND FUTURE WORK

7.1 Conclusion

This thesis has presented a comprehensive study on improving the performance of deep learning models on UC datasets by taking into account relationships between classes. The primary objective of this work was to develop a robust loss function for the computer-aided diagnosis of UC from endoscopic images.

In the dataset creation phase, we collected the largest publicly available annotated UC dataset, called LIMUC, which is one of the main contributions of this study. Due to the lack of a comprehensive dataset for UC research and the drawbacks associated with using private datasets, we created LIMUC to foster more effective and accurate machine-learning models for the automated diagnosis and treatment of UC. The dataset contains 19537 endoscopic images collected from 1043 colonoscopy procedures involving 572 UC patients. The images were labeled using the widely used MES system by experienced gastroenterologists, and the final annotated dataset consists of 11276 images from 564 patients. We have provided details on the data collection, processing, and labeling procedures, along with the challenges faced during the annotation process. Finally, we shared the detailed statistics, distribution of annotations and the final LIMUC dataset, which is publicly available under the Creative Commons Attribution 4.0 license.

Class Distance Weighted Cross Entropy loss was introduced as a novel, non-parametric method to address the limitations of the standard Cross-Entropy loss function for ordinal classification tasks. The CDW-CE loss function not only evaluates the confidence of non-true classes but also considers the ordinal relationship between classes by incorporating a distance-based coefficient. This coefficient penalizes mispredictions more heavily as they deviate further from the true class, effectively accounting for the ordinal nature of the classes. Compared to other approaches such as CORAL, CORN, unimodal constraints, and regression-based methods, the CDW-CE loss function is more suitable for ordinal classification problems as it provides an efficient and effective way to consider the ordinal relationship between classes without requiring any changes to the model architecture or labeling structure. The implementation of CDW-CE demonstrates its usability and ease of integration into existing deep-learning architectures.

One of the key highlights of our study was the successful application of the proposed method to an external dataset - a widely acknowledged diabetic retinopathy dataset. The consistency of the results obtained from these external experiments with those from the UC experiments serves as a strong signal to the robustness and adaptability of our proposed method. The ability of our approach to effectively handle and yield high performance on this distinct dataset provides clear evidence of its generalization capacity.

The experimental results demonstrated that the proposed CDW-CE loss function consistently outperformed the other methods in all performance metrics and CNN models, showcasing its ability to handle complex classification tasks effectively. Moreover, the CDW-CE loss function was observed to provide better explainability of the model through class activation maps, which is an important criterion in determining its use in clinical settings.

7.2 Future Work

The work in this thesis was focused on evaluating individual endoscopic frames and assigning a score independent of each other. A natural extension of this study would be to perform an automated diagnosis for entire endoscopic video sequences rather than individual frames. Currently, the assessment of UC severity is conducted using the Mayo endoscopic score, which is assigned to the whole video captured during the endoscopic examination. By considering the complete video sequence, the model would be able to account for the temporal dynamics and contextual information present in the endoscopic videos, potentially leading to a more accurate and holistic diagnosis of UC severity. Developing an efficient and effective approach for processing and analyzing such video sequences would not only enhance the model's clinical applicability but also contribute to a more comprehensive understanding of the progression and manifestation of ulcerative colitis.

Looking ahead, there are a multitude of avenues for further exploration and expansion of the present work. An especially promising prospect lies in the potential application of our proposed CDW-CE loss function to more fine-grained scoring systems, such as the Ulcerative Colitis Endoscopic Index of Severity (UCEIS). The UCEIS represents a more intricate and detailed system for assessing UC severity, where the assessment is performed for the three distinct major indicators, namely, vascular pattern, bleeding, erosions and ulcers. Each group has an ordinal grading in itself, and the total score of the UCEIS is determined as the sum of the sub-indicators. This granularity poses an exciting challenge and opportunity for the application of the CDW-CE. Given its robust performance in both the UC and diabetic retinopathy datasets, we believe that the CDW-CE could provide a valuable tool for enhancing the accuracy and efficacy of these finer-scale severity indices.

Another promising direction for future work involves addressing the challenges associated with obtaining accurately labeled ulcerative colitis images. The presence of high intra- and inter-observer variability, coupled with the limited availability of publicly shared datasets, creates a bottleneck in the development of robust computer-aided diagnosis models. To overcome these limitations, it would be worthwhile to investigate the integration of self-supervised or semi-supervised learning techniques into the current framework. These alternative learning strategies could enable the effective utilization of limited labeled data by leveraging the inherent structure and patterns present in the vast amounts of unlabeled ulcerative colitis images.

REFERENCES

[1] "Ulcerative colitis - ibd," https://www.crohns-disease-probiotics.com/ulcerativecolitis/, accessed: 2023-05-01.

[2] K. W. Schroeder, W. J. Tremaine, and D. M. Ilstrup, "Coated oral 5-aminosalicylic acid therapy for mildly to moderately active ulcerative colitis," *New England Journal of Medicine*, vol. 317, no. 26, pp. 1625–1629, 1987.

[3] S. P. Travis, D. Schnell, P. Krzeski, M. T. Abreu, D. G. Altman, J.-F. Colombel, B. G. Feagan, S. B. Hanauer, G. R. Lichtenstein, P. R. Marteau *et al.*, "Reliability and initial validation of the ulcerative colitis endoscopic index of severity," *Gastroenterology*, vol. 145, no. 5, pp. 987–995, 2013.

[4] T. Osada, T. Ohkusa, T. Yokoyama, T. Shibuya, N. Sakamoto, K. Beppu, A. Nagahara, M. Otaka, T. Ogihara, and S. Watanabe, "Comparison of several activity indices for the evaluation of endoscopic activity in uc: inter-and intraobserver consistency," *Inflammatory bowel diseases*, vol. 16, no. 2, pp. 192–197, 2010.

[5] S. Ali, M. Dmitrieva, N. Ghatwary, S. Bano, G. Polat, A. Temizel, A. Krenzer, A. Hekalo, Y. B. Guo, B. Matuszewski *et al.*, "Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy," *Medical image analysis*, vol. 70, p. 102002, 2021.

[6] S. Ali, N. Ghatwary, D. Jha, E. Isik-Polat, G. Polat, C. Yang, W. Li, A. Galdran, M.-Á. G. Ballester, V. Thambawita *et al.*, "Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge," *arXiv preprint arXiv:2202.12031*, 2022.

[7] W. Du, N. Rao, D. Liu, H. Jiang, C. Luo, Z. Li, T. Gan, and B. Zeng, "Review on the applications of deep learning in the analysis of gastrointestinal endoscopy images," *IEEE Access*, vol. 7, pp. 142 053–142 069, 2019.

[8] G. Polat, D. Sen, A. Inci, and A. Temizel, "Endoscopic artefact detection with ensemble of deep neural networks and false positive elimination," ser. CEUR Workshop Proceedings, vol. 2595, 2020, pp. 8–12.

[9] G. Polat, E. Isik-Polat, K. Kayabay, and A. Temizel, "Polyp detection in colonoscopy images using deep learning and bootstrap aggregation," ser. CEUR Workshop Proceedings, vol. 2886, 2021, pp. 90–100. [Online]. Available: http://ceur-ws.org/Vol-2886

[10] "Iterative health," https://iterative.health/, accessed: 2023-04-15.

[11] "Odin vision," https://odin-vision.com/, accessed: 2023-04-15.

[12] "Olympus endo-aid," https://www.olympus-europa.com/medical/en/Products-and-Solutions/Products/Product/ENDO-AID.html, accessed: 2023-04-15.

[13] "Medtronic gi genius," https://www.medtronic.com/covidien/en-gb/products/gastrointestinal-artificial-intelligence/gi-genius-intelligent-endoscopy.html, accessed: 2023-04-15.

[14] G. Polat, H. T. Kani, I. Ergenc, Y. O. Alahdab, A. Temizel, and O. Atug, "Labeled Images for Ulcerative Colitis (LIMUC) Dataset," Mar. 2022, doi:10.5281/zenodo.5827695. [Online]. Available: https://doi.org/10.5281/zenodo.5827695

[15] G. Polat, I. Ergenc, H. T. Kani, Y. O. Alahdab, O. Atug, and A. Temizel, "Class distance weighted cross-entropy loss for ulcerative colitis severity estimation," in *Medical Image Understanding and Analysis: 26th Annual Conference, MIUA 2022, Cambridge, UK, July 27–29, 2022, Proceedings*. Springer, 2022, pp. 157–171.

[16] A. Alammari, A. R. Islam, J. Oh, W. Tavanapong, J. Wong, and P. C. de Groen, "Classification of ulcerative colitis severity in colonoscopy videos using cnn," ser. ICIME 2017, 2017, p. 139–144, doi: 10.1145/3149572.3149613. [Online]. Available: https://doi.org/10.1145/3149572.3149613

[17] S. V. L. L. Tejaswini, B. Mittal, J. Oh, W. Tavanapong, J. Wong, and P. C. de Groen, "Enhanced approach for classification of ulcerative colitis severity in colonoscopy videos using cnn," in *Advances in Visual Computing*, Berlin, Heidelberg, p. 25–37, doi: 10.1007/978-3-030-33723-0_3.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25, 2012.

[19] Y. Maeda, S.-e. Kudo, Y. Mori, M. Misawa, N. Ogata, S. Sasanuma, K. Wakamura, M. Oda, K. Mori, and K. Ohtsuka, "Fully automated diagnostic system with artificial intelligence using endocytoscopy to identify the presence of histologic inflammation associated with ulcerative colitis (with video)," *Gastrointestinal endoscopy*, vol. 89, no. 2, pp. 408–415, 2019.

[20] T. Ozawa, S. Ishihara, M. Fujishiro, H. Saito, Y. Kumagai, S. Shichijo, K. Aoyama, and T. Tada, "Novel computer-assisted diagnosis system for endoscopic disease activity in patients with ulcerative colitis," *Gastrointestinal Endoscopy*, vol. 89, no. 2, pp. 416–421.e1, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0016510718331936

[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.

[22] R. W. Stidham, W. Liu, S. Bishu, M. D. Rice, P. D. R. Higgins, J. Zhu, B. K. Nallamothu, and A. K. Waljee, “Performance of a Deep Learning Model vs Human Reviewers in Grading Endoscopic Disease Severity of Patients With Ulcerative Colitis,” *JAMA Network Open*, vol. 2, no. 5, pp. e193 963–e193 963, 05 2019, doi: 10.1001/jamanetworkopen.2019.3963. [Online]. Available: https://doi.org/10.1001/jamanetworkopen.2019.3963

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[24] K. Takenaka, K. Ohtsuka, T. Fujii, M. Negi, K. Suzuki, H. Shimizu, S. Oshima, S. Akiyama, M. Motobayashi, M. Nagahori *et al.*, “Development and validation of a deep neural network for accurate evaluation of endoscopic images from patients with ulcerative colitis,” *Gastroenterology*, vol. 158, no. 8, pp. 2150–2157, 2020.

[25] H. P. Bhambhvani and A. Zamora, “Deep learning enabled classification of mayo endoscopic subscore in patients with ulcerative colitis,” *European Journal of Gastroenterology & Hepatology*, vol. 33, no. 5, pp. 645–649, 2021.

[26] H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T. D. Nguyen *et al.*, “Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy,” *Scientific data*, vol. 7, no. 1, pp. 1–14, 2020.

[27] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” *arXiv preprint arXiv:1611.05431*, 2016.

[28] K. Gottlieb, J. Requa, W. Karnes, R. C. Gudivada, J. Shen, E. Rael, V. Arora, T. Dao, A. Ninh, and J. McGill, “Central reading of ulcerative colitis clinical trial videos using neural networks,” *Gastroenterology*, vol. 160, no. 3, pp. 710–719, 2021.

[29] M. Schuster and K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997, doi: 10.1109/78.650093.

[30] F. Gers, J. Schmidhuber, and F. Cummins, “Learning to forget: continual prediction with lstm,” in *1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470)*, vol. 2, 1999, pp. 850–855 vol.2, doi: 10.1049/cp:19991218.

[31] H. Yao, K. Najarian, J. Gryak, S. Bishu, M. D. Rice, A. K. Waljee, H. J. Wilkins, and R. W. Stidham, “Fully automated endoscopic disease activity assessment in ulcerative colitis,” *Gastrointestinal Endoscopy*, vol. 93, no. 3, pp. 728–736, 2021.

[32] T.-Y. Huang, S.-Q. Zhan, P.-J. Chen, C.-W. Yang, and H. H.-S. Lu, “Accurate diagnosis of endoscopic mucosal healing in ulcerative colitis using deep learning

and machine learning," *Journal of the Chinese Medical Association*, vol. 84, no. 7, pp. 678–681, 2021.

[33] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[34] B. Gutierrez Becker, F. Arcadu, A. Thalhammer, C. Gamez Serna, O. Feehan, F. Drawnel, Y. S. Oh, and M. Prunotto, "Training and deploying a deep learning model for endoscopic severity grading in ulcerative colitis using multicenter clinical trial data," *Therapeutic advances in gastrointestinal endoscopy*, vol. 14, p. 2631774521990623, 2021.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[36] E. Schwab, G. O. Cula, K. Standish, S. S. Yip, A. Stojmirovic, L. Ghanem, and C. Chehoud, "Automatic estimation of ulcerative colitis severity from endoscopy videos using ordinal multi-instance learning," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pp. 1–9, 2021.

[37] B. E. Sands, W. J. Sandborn, R. Panaccione, C. D. O'Brien, H. Zhang, J. Johanns, O. J. Adedokun, K. Li, L. Peyrin-Biroulet, G. Van Assche, S. Danese, S. Targan, M. T. Abreu, T. Hisamatsu, P. Szapary, and C. Marano, "Ustekinumab as induction and maintenance therapy for ulcerative colitis," *New England Journal of Medicine*, vol. 381, no. 13, pp. 1201–1214, 2019, pMID: 31553833, doi:10.1056/NEJMoa1900750.

[38] S. Harada, R. Bise, H. Hayashi, K. Tanaka, and S. Uchida, "Order-guided disentangled representation learning for ulcerative colitis classification with limited labels," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 471–480.

[39] Y. Maeda, S. ei Kudo, N. Ogata, M. Misawa, M. Iacucci, M. Homma, T. Nemoto, K. Takishima, K. Mochida, H. Miyachi, T. Baba, K. Mori, K. Ohtsuka, and Y. Mori, "Evaluation in real-time use of artificial intelligence during colonoscopy to predict relapse of ulcerative colitis: a prospective study," *Gastrointestinal Endoscopy*, vol. 95, no. 4, pp. 747–756.e2, 2022, doi: 10.1016/j.gie.2021.10.019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0016510721017314

[40] X. Luo, J. Zhang, Z. Li, and R. Yang, "Diagnosis of ulcerative colitis from endoscopic images based on deep learning," *Biomedical Signal Processing and Control*, vol. 73, p. 103443, 2022, doi: 10.1016/j.bspc.2021.103443. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809421010405

[41] R. T. Sutton, O. R. Zaiane, R. Goebel, and D. C. Baumgart, "Artificial intelligence enabled automated diagnosis and grading of ulcerative colitis endoscopy images," *Scientific Reports*, vol. 12, no. 1, pp. 1–10, 2022.

[42] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[43] T. Kadota, K. Abe, R. Bise, T. Kawamura, N. Sakiyama, K. Tanaka, and S. Uchida, "Automatic estimation of ulcerative colitis severity by learning to rank with calibration," *IEEE Access*, vol. 10, pp. 25 688–25 695, 2022, doi: 10.1109/ACCESS.2022.3155769.

[44] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Proceedings of the 22nd international conference on Machine learning*, 2005, pp. 89–96.

[45] Z. Xu, S. Ali, J. East, and J. Rittscher, "Additive angular margin loss and model scaling network for optimised colitis scoring," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 2022, pp. 1–5, doi: 10.1109/ISBI52829.2022.9761437.

[46] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[47] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[48] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output cnn for age estimation," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4920–4928.

[49] W. Cao, V. Mirjalili, and S. Raschka, "Rank consistent ordinal regression for neural networks with application to age estimation," *Pattern Recognition Letters*, vol. 140, pp. 325–331, 2020.

[50] X. Shi, W. Cao, and S. Raschka, "Deep neural networks for rank-consistent ordinal regression based on conditional probabilities," *arXiv preprint arXiv:2111.08851*, 2021.

[51] K. Ricanek and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," in *7th international conference on automatic face and gesture recognition (FGR06)*. IEEE, 2006, pp. 341–345.

[52] R. Schifanella, M. Redi, and L. M. Aiello, "An image is worth more than a thousand favorites: Surfacing the hidden beauty of flickr pictures," in *Proceedings of the international AAAI conference on web and social media*, vol. 9, no. 1, 2015, pp. 397–406.

[53] “Fireman dataset,” https://github.com/gagolews/ordinal-regression-data, accessed: 2023-04-15.

[54] S. Belharbi, I. B. Ayed, L. McCaffrey, and E. Granger, “Non-parametric uni-modality constraints for deep ordinal classification,” *arXiv preprint arXiv:1911.10720*, 2019.

[55] G. Aresta, T. Araújo, S. Kwok, S. S. Chennamsetty, M. Safwan, V. Alex, B. Marami, M. Prastawa, M. Chan, M. Donovan *et al.*, “Bach: Grand challenge on breast cancer histology images,” *Medical image analysis*, vol. 56, pp. 122–139, 2019.

[56] F. Palermo, J. Hays, and A. A. Efros, “Dating historical color images,” in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*. Springer, 2012, pp. 499–512.

[57] T. Albuquerque, R. Cruz, and J. S. Cardoso, “Ordinal losses for classification of cervical cancer risk,” *PeerJ Computer Science*, vol. 7, p. e457, 2021.

[58] J. Jantzen and G. Dounias, “Analysis of pap-smear image data,” in *Proceedings of the nature-inspired smart information systems 2nd annual symposium*, vol. 10, 2006, pp. 1–11.

[59] C. Beckham and C. Pal, “A simple squared-error reformulation for ordinal classification,” *arXiv preprint arXiv:1612.00775*, 2016.

[60] “Diabetic retinopathy detection,” https://www.kaggle.com/c/diabetic-retinopathy-detection, accessed: 2023-04-17.

[61] G. Polat, H. T. Kani, I. Ergenc, Y. Ozen 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*, 11 2022, doi: 10.1093/ibd/izac226. [Online]. Available: https://doi.org/10.1093/ibd/izac226

[62] J. Satsangi, M. Silverberg, S. Vermeire, and J. Colombel, “The montreal classification of inflammatory bowel disease: controversies, consensus, and implications,” *Gut*, vol. 55, no. 6, pp. 749–753, 2006.

[63] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan *et al.*, “Searching for mobilenetv3,” in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1314–1324.

[64] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2017.

[65] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” in *Advances in Neural Information Processing Systems 32*, 2019, pp. 8024–8035.

[66] T. maintainers and contributors, “TorchVision: PyTorch’s Computer Vision library,” Nov. 2016. [Online]. Available: https://github.com/pytorch/vision

[67] “Ordinal losses,” https://github.com/tomealbuquerque/ordinal-losses, accessed: 2023-04-25.

[68] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, “Cosface: Large margin cosine loss for deep face recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5265–5274.

[69] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, “Sphereface: Deep hypersphere embedding for face recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[70] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, “Normface: L2 hypersphere embedding for face verification,” in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1041–1049.

[71] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2921–2929.

[72] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 618–626.

[73] Z. L. Teo, Y.-C. Tham, M. Yu, M. L. Chee, T. H. Rim, N. Cheung, M. M. Bikbov, Y. X. Wang, Y. Tang, Y. Lu *et al.*, “Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis,” *Ophthalmology*, vol. 128, no. 11, pp. 1580–1591, 2021.

[74] “What is diabetic retinopathy?” https://www.healthline.com/health/type-2-diabetes/retinopathy, accessed: 2023-04-17.

[75] J. W. C. Emma Dugas, Jared, “Diabetic retinopathy detection,” 2015. [Online]. Available: https://kaggle.com/competitions/diabetic-retinopathy-detection

[76] L. Van der Maaten and G. Hinton, “Visualizing data using t-sne.” *Journal of machine learning research*, vol. 9, no. 11, 2008.

[77] P. J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

APPENDIX A

THE IMPLEMENTATION OF CDW-CE WITH PYTORCH

Implementation of CDW-CE with PyTorch framework:

```python
import numpy as np
import torch
from torch import Tensor

class ClassDistanceWeightedCrossEntropyLoss(torch.nn.Module):

    def __init__(self, class_size: int, power: float = 2.0, reduction: str = "mean"):
        super(ClassDistanceWeightedCrossEntropyLoss, self).__init__()
        self.class_size = class_size
        self.power = power
        self.reduction = reduction

    def forward(self, input: Tensor, target: Tensor)  Tensor:
        input_sm = input.softmax(dim=1)

        weight_matrix = torch.zeros_like(input_sm)
        for i, target_item in enumerate(target):
            weight_matrix[i] = torch.tensor(
                [abs(k - target_item) for k in range(self.class_size)]
            )

        weight_matrix.pow_(self.power)
        reverse_probs = (1 - input_sm).clamp_(min=1e4)

        log_loss = -torch.log(reverse_probs)
        loss = log_loss * weight_matrix
        loss_sum = torch.sum(loss, dim=1)

        if self.reduction == "mean":
            loss_reduced = torch.mean(loss_sum)
        elif self.reduction == "sum":
            loss_reduced = torch.sum(loss_sum)
        else:
            raise Exception("Undefined reduction type: " + self.reduction)

        return loss_reduced
```

Gorkem Polat

Curriculum Vitae

Education

- **MSc in Electrical and Electronics Engineering**, Middle East Technical University, Turkey, 2014-2018.

- **BSc in Electrical and Electronics Engineering**, Middle East Technical University, Turkey, 2008-2013.

- **Minor Degree in Sociology**, Middle East Technical University, Turkey, 2010-2013.

- **High School in Maths-Science**, Bornova Anatolian High School, Turkey, 2004-2008.

Work Experience

- **Machine Learning Engineer**, Encord, London-UK August 2022 - Present.

- **Software Engineer**, Piri Reis Information Systems, Ankara-Turkey, August 2016 - December 2019.

- **Software Engineer**, Safe and Sound Technologies, Ankara-Turkey, June 2014 - August 2016.

- **Software Engineer**, Piri Reis Information Systems, Ankara-Turkey, September 2013 - June 2014.

Publications

1. Isik-Polat, Ece, Gorkem Polat, and Altan Kocyigit. "ARFED: Attack-Resistant Federated averaging based on outlier elimination." Future Generation Computer Systems 141 (2023): 626-650.

2. Polat, Gorkem, Haluk Tarik Kani, Ilkay Ergenc, Yesim Ozen Alahdab, Alptekin Temizel, and Ozlen Atug. ‘Improving the Computer-Aided Estimation of Ulcerative Colitis Severity According to Mayo Endoscopic Score by Using Regression-Based Deep Learning’. Inflammatory Bowel Diseases, 11 2022.

3. Kani, Haluk Tarık, I. Ergenc, Görkem Polat, Y. Ozen Alahdab, Alptekin Temizel, and O. Atug. "Evaluation of Ulcerative Colitis Endoscopic Mayo Score with Artificial Intelligence." Endoscopy 54, no. S 01 (2022): eP083.

4. Ali, Sharib, Noha Ghatwary, Debesh Jha, Ece Isik-Polat, Gorkem Polat, Chen Yang, Wuyang Li et al. "Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge." arXiv preprint arXiv:2202.12031 (2022).

5. Polat, Gorkem, Ilkay Ergenc, Haluk Tarik Kani, Yesim Ozen Alahdab, Ozlen Atug, and Alptekin Temizel. "Class distance weighted cross-entropy loss for ulcerative colitis severity estimation." In Medical Image Understanding and Analysis: 26th Annual Conference, MIUA 2022, Cambridge, UK, July 27–29, 2022, Proceedings, pp. 157-171. Cham: Springer International Publishing, 2022.

6. Isik-Polat, Ece, Gorkem Polat, Altan Kocyigit, and Alptekin Temizel. "Evaluation and analysis of different aggregation and hyperparameter selection methods for federated brain tumor segmentation." In International MICCAI Brainlesion Workshop, pp. 405-419. Cham: Springer International Publishing, 2021.

7. Polat, Gorkem, Ece Isik-Polat, Kerem Kayabay, and Alptekin Temizel. ‘Polyp Detection in Colonoscopy Images Using Deep Learning and Bootstrap Aggregation’, 2886:90–100. CEUR Workshop Proceedings, 2021. http://ceur-ws.org/Vol-2886.

8. Kani, Haluk Tarık, I. Ergenc, Görkem Polat, Y. Ozen Alahdab, Alptekin Temizel, and O. Atug. "P099 Evaluation of endoscopic mayo score with an artificial intelligence algorithm." Journal of Crohn’s and Colitis 15, no. Supplement_1 (2021): S195-S196.

9. Ali, Sharib, Mariia Dmitrieva, Noha Ghatwary, Sophia Bano, Gorkem Polat, Alptekin Temizel, Adrian Krenzer et al. "Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy." Medical image analysis 70 (2021): 102002.

10. Reinke, Annika, Minu D. Tizabi, Carole H. Sudre, Matthias Eisenmann, Tim Rädsch, Michael Baumgartner, Laura Acion et al. "Common limitations of image processing metrics: A picture story." arXiv preprint arXiv:2104.05642 (2021).

11. Polat, Gorkem, Deniz Sen, Alperen Inci, and Alptekin Temizel. "Endoscopic artefact detection with ensemble of deep neural networks and false positive elimination." In EndoCV@ ISBI, pp. 8-12. 2020.

12. Polat, Gorkem, Ugur Halici, and Yesim Serinagaoglu Dogrusoz. "False positive reduction in lung computed tomography images using convolutional neural

networks." arXiv preprint arXiv:1811.01424 (2018).

13. Polat, Görkem, Yesim Serinagaoglu Dogrusoz, and Ugur Halici. "Effect of input size on the classification of lung nodules using convolutional neural networks." In 2018 26th Signal Processing and Communications Applications Conference (SIU), pp. 1-4. IEEE, 2018.

14. Polat, Görkem. "Classification of lung nodules in CT images using convolutional neural networks." Master's thesis, Middle East Technical University, 2018.

TEZ İZİN FORMU / THESIS PERMISSION FORM

ENSTİTÜ / INSTITUTE

Fen Bilimleri Enstitüsü / Graduate School of Natural and Applied Sciences ☐

Sosyal Bilimler Enstitüsü / Graduate School of Social Sciences ☐

Uygulamalı Matematik Enstitüsü / Graduate School of Applied Mathematics ☐

Enformatik Enstitüsü / Graduate School of Informatics ☒

Deniz Bilimleri Enstitüsü / Graduate School of Marine Sciences ☐

YAZARIN / AUTHOR

Soyadı / Surname : Polat
Adı / Name : Görkem
Bölümü / Department : Sağlık Bilişimi

TEZİN ADI / TITLE OF THE THESIS (**İngilizce** / English) :
COMPUTER-AIDED ESTIMATION OF ENDOSCOPIC ACTIVITY IN ULCERATIVE COLITIS

TEZİN TÜRÜ / DEGREE: **Yüksek Lisans** / Master ☐ **Doktora** / PhD ☒

1. **Tezin tamamı dünya çapında erişime açılacaktır. /** Release the entire work immediately for access worldwide. ☒
2. **Tez iki yıl süreyle erişime kapalı olacaktır.** / Secure the entire work for patent and/or proprietary purposes for a period of **two year**. * ☐
3. **Tez altı ay süreyle erişime kapalı olacaktır.** / Secure the entire work for period of **six months**. * ☐

***** *Enstitü Yönetim Kurulu Kararının basılı kopyası tezle birlikte kütüphaneye teslim edilecektir. A copy of the Decision of the Institute Administrative Committee will be delivered to the library together with the printed thesis.*

Yazarın imzası / Signature .... **Tarih** / Date 21.07.2023