ACTIVE LEARNING BASED SYNTHETIC SAMPLE SELECTION FOR ENDOSCOPIC IMAGE CLASSIFICATION

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

ALPEREN İNCİ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF
MASTER OF SCIENCE
IN
MULTIMEDIA INFORMATICS, THE DEPARTMENT OF MODELLING AND SIMULATION

SEPTEMBER 2022

Active Learning Based Synthetic Sample Selection for Endoscopic Image Classification

submitted by **ALPEREN İNCİ** in partial fulfillment of the requirements for the degree of **Master of Science in Modelling and Simulation Department, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin _____
Dean, **Graduate School of Informatics**

Assoc. Prof. Dr. Elif Sürer _____
Head of Department, **Modelling and Simulation**

Prof. Dr. Alptekin Temizel _____
Supervisor, **Modelling and Simulation, METU**

Examining Committee Members:

Assoc. Prof. Dr. Elif Sürer
Modelling and Simulation, Middle East Technical University _____

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

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

Date: 02.09.2022

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: Alperen İnci

Signature :

ABSTRACT

ACTIVE LEARNING BASED SYNTHETIC SAMPLE SELECTION FOR ENDOSCOPIC IMAGE CLASSIFICATION

İnci, Alperen
M.S., Department of Modelling and Simulation
Supervisor: Prof. Dr. Alptekin Temizel

September 2022, 63 pages

Many people suffer from Ulcerative Colitis (UC), which is a chronic inflammatory bowel disease. UC exhibits itself as ulcers, inflammation, and sores in the colon. In order to provide proper treatment, an expert physician needs to assess the severity of the disease. However, physicians have disagreements on the severity of the disease in terms of the widely used Mayo scoring. This brings in the need for reliable and automated methods. Even though automated disease diagnosis using medical imaging has become a trending topic, labeling data in the medical field requires the agreement of a committee of experts and this is a limiting factor in obtaining a sufficient amount and variety of data for robust model training. Although generative data augmentation methods have been proven to increase data diversity and quality, directly adding synthetic samples may not contribute to a model's performance, and may even result in a performance drop. Some of these generated samples might be unrealistic, or already have been learned by the model and predicted with a high confidence score. In this thesis, we propose generation of a synthetic data pool, a subset of which is then selected and used for training. The data pool is created by generating samples with different truncation parameters. Then, active learning approaches are adopted to select informative synthetic samples among these samples in the data pool for model training. The results show that the results are favorable, and the performance is more stable compared to the baseline method and random selection of generated samples.

Keywords: active learning, dataset cleaning, image classification, synthetic sample selection

ÖZ

ENDOSKOPİK GÖRÜNTÜ SINIFLANDIRILMASI İÇİN AKTİF ÖĞRENME TABANLI SENTETİK VERİ SEÇİMİ

İnci, Alperen
Yüksek Lisans, Çokluortam Bilişimi, Modelleme ve Simülasyon Bölümü
Tez Yöneticisi: Prof. Dr. Alptekin Temizel

Eylül 2022, 63 sayfa

Birçok insan, kronik bir iltihaplı bağırsak hastalığı olan Ülseratif Kolit (ÜK) hastalığına yakalanmaktadır. Hastalık kendini kalın bağırsakta ülser, iltihap ve yara olarak gösterir. Uygun tedaviyi sağlayabilmek için uzman bir hekimin hastalığın şiddetini belirlemesi gerekir. Ancak hastalığın şiddeti konusunda sıklıkla kullanılmakta olan Mayo puanlamasını belirlerken hekimler arasında görüş ayrılığı bulunmaktadır. Bu da güvenilir ve otomatikleştirilmiş yöntemlere olan ihtiyacı beraberinde getirmektedir. Hastalığın otomatik teşhisi için tıbbi görüntülemenin kullanıldığı yöntemlerin ortaya konulmasına rağmen, tıp alanında verilerin etiketlenmesi için uzmanlardan oluşan bir komitenin mutabakata varması gerekmektedir. Bu durum gürbüz bir model eğitimi için gereken veri miktarına ve çeşitliliğine ulaşmayı zorlaştırır. Üretken veri artırma yöntemlerinin veri çeşitliliğini ve kalitesini artırdığı kanıtlanmış olsa da, doğrudan sentetik örneklerin verisetine eklenmesi model performansına katkıda bulunmayabilir ve hatta performans düşüşüne neden olabilir. Üretilen örneklerin bazıları gerçekçi olmayabilir veya model tarafından zaten öğrenilmiş ve yüksek skor ile tahmin ediliyor olabilir. Bu tezde, sentetik görüntülerden önerilen bir veri havuzunun oluşturulması ve bu havuzun bir altkümesinin seçilerek eğitim için kullanılması önerilir. Havuz farklı kesme parametrelerine sahip örneklerin üretilmesi ile oluşturulur. Daha sonra bu sentetik veri havuzundan model eğitimi için kullanılacak bilgilendirici sentetik örnekleri seçmek için aktif öğrenme yaklaşımları kullanılır. Sonuçlar, temel yöntemlere ve rastgele eklemeye göre performans artışı gözlemlendiğini ve performansın daha kararlı olduğunu göstermiştir.

Anahtar Kelimeler: aktif öğrenme, veriseti temizleme, görüntü sınıflandırma, sentetik örnek seçimi

To my father

ACKNOWLEDGMENTS

First of all, I would like to express my deepest gratitude to my supervisor Prof. Dr. Alptekin Temizel for his invaluable guidance, patience, immerse knowledge, important suggestions and feedbacks throughout my thesis study. His endless support, motivation and encouragement helped me in this process.

I would also like to thank dear committee members Prof. Dr. Çiğdem Gündüz Demir and Assoc. Prof. Dr. Elif Sürer for their time, support, suggestions, helpful comments and discussions that contributed to the development of this research. Thanks to their valuable feedback and suggestions, I had the opportunity to improve my work and to make this study better.

I want to express my deepest and greatest thanks to my family who have always supported me throughout my life. Even though we were far away, they provided their unflagging support, continuous encouragement, moral, and unconditional love. Thanks to the opportunities and support they have offered me, I have now come to these points.

I would like to thank Kuartis family very much for the opportunities, guidance and support it has provided me. They always motivated me throughout my thesis. I would like to thank The Scientific and Technological Research Council of Turkey (TÜBİTAK), which always values and supports scientists, for their M.S. studies scholarship, BIDEB 2210.

I am also grateful to Oğuz Hanoğlu, Mert Çağlar and Görkem Polat, who did not spare their help throughout the thesis, devoted their precious time to me and supported me. I would also like to thank my dear friends Cem Gültekin, Mehmetcan Söylemez and İsmail Başarır for their valuable friendship, motivation, and encouragement. During this process, they were always by my side and provided full support. I would like to express my sincere thanks to my friend Deniz Şen, with whom I have been together since the beginning of the master's degree, for helping me whenever I needed and befriending me in the courses we took together and in the thesis process.

Last but not least, my deepest gratitude goes to my girlfriend. Her love and support provided me strength and motivation to achieve this goal. I express my sincere thanks to her for always being there for me whenever I need.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Al	Active Learning
ADA	Adaptive Discriminator Augmentation
BADGE	Batch Active learning by Diverse Gradient Embeddings
BvSB	Best versus Second Best
CGAN	Conditional Generative Adversarial Networks
DAAL	Dual Adversarial Network
EMS	Endoscopic Mayo Score
GAAL	Generative Adversarial Active Learning
GAN	Generative Adversarial Network
GMM	Gaussian Mixture Models
IBD	Inflammatory Bowel Disease
LIMUC	Labeled Images for Ulcerative Colitis
LL	Learning Loss
MAE	Mean Absolute Error
MNIST	Modified National Institute of Standards and Technology
PCA	Principal Component Analysis
QWK	Quadratic Weighted Kappa
RSS	Random Subset Selection
SDA	Stochastic Discriminator Augmentation
T-SNE	t-distributed Stochastic Neighbor Embedding
UC	Ulcerative Colitis
VAAL	Variational Adversarial Active Learning

VAE	Variational AutoEncoder
WANDB	Weight and Biases
WGAN	Wasserstein Generative Adversarial Networks

CHAPTER 1

INTRODUCTION

The impressive development of deep learning in the field of computer vision has been the solution to many problems. Deep learning offers various solutions to problems with its many sub-branches, such as supervised, unsupervised, semi-supervised, and reinforcement learning.

Supervised learning, one of the most common deep learning applications, provides very successful results in many areas, such as object detection and classification. Although deep learning methods can vary according to the difficulty of the task, they usually require a large amount of data to achieve satisfactory performance and converge the models, so they are known as greedy to data [1]. Therefore, supervised learning owes its success to a large amount of high-quality data that can be easily found and annotated in many fields.

The widespread of supervised learning has brought along the effort to find solutions to the problems in the medical field. Reliable and robust techniques are required to employ these algorithms in the field of medical imaging and to assist physicians. However, the fact that the models need a large number of data makes this situation even more difficult in the medical field. This problem arises since data labeling requires expert judgment, even unanimous decisions of experts, and data in the medical area is sensitive. This emphasizes the necessity of unanimous decision in this field. Ulcerative Colitis (UC), which may cause ulcers in the colon, which is the subject of this thesis, can be given as an example. The degree of disease is determined by the Endoscopic Mayo Score (EMS) [2], widely used in the literature. On the EMS score, while Mayo 0 means no disease, Mayo 1, Mayo 2, and Mayo 3 indicate an increasing degree of disease. Cohen's Kappa value considers the possibility of an agreement by chance. Vashist et al. [3] found the Cohen's Kappa value $k = 0.45$ for non-experts and $k = 0.71$ for experts, indicating moderate and substantial reliability, respectively.

Deep learning researchers have conducted various studies to solve the problem of data deficiency. The classical data augmentation is the most widely used method, which is now added to every trained scheme as default. These methods aim to increase model performance, generalization capacity, and robustness by applying various transformations to the image. Generative Adversarial Networks, which can be defined as a more advanced data augmentation, can increase data diversity and extract extra information from the existing data. The drawback of Generative Adversarial Networks is the need

for a large amount of data for training. Karras et al. [4] proposed StyleGAN2-ADA, which is capable of training GANs with fewer data.

Çağlar applied the StyleGAN2-ADA approach to the Labeled Images for Ulcerative Colitis (LIMUC) Dataset [5] to generate synthetic samples in his thesis. Generated samples are used to extend the training set for image classification. The results show that it is possible to improve classification performance by training StyleGAN2-ADA with a very less amount of data, which is only 50. The effectiveness of the applied strategy is proven on subsets with a different number of samples on the LIMUC dataset. Even though training GANs contributed to an improvement in the classification performance, fluctuation and instability in the performance still remain an open question. The generated synthetic samples are directly added to the training set, and no selection mechanism is applied to the samples. The samples may be predicted by the model with very high confidence or may be unrealistic. These samples may have a destructive effect on the model training instead of contributing.

Active learning approaches aim to maximize model performance in a cost-efficient way. Selecting informative samples provides better performance with reduced labeling costs. These methods can be applied to choose synthetic samples to enlarge the training set. However, active learning strategies are prone to select outliers. Karamcheti et al. [6] show that random selection performs better than eight active learning strategies in 4 different datasets. This emphasizes the importance of the synthetic sample selection strategy since GANs frequently produce outliers.

This thesis takes Çağlar's work [7] as a baseline and tries to select synthetic examples wisely to guarantee performance improvement.

1.1 Research Questions

Generative data augmentation methods are capable of generating realistic samples with small datasets. The effectiveness of using generative data augmentation to improve image classification performance in small datasets is also proven. Can the performance be further increased by wisely selecting samples using active learning?

1.2 Motivation

Active learning strategies in this thesis can be applied to select more informative synthetic samples to improve image classification performance. Extending the training dataset with wisely selected samples instead of randomly selected yields better image classification performance.

The recognition of the severity of the disease is crucial for physicians to administer appropriate treatment for UC disease. This diagnosis and treatment protect patients from undesired short-long term potentially fatal effects of the disease. Improving the performance of the model decreases the need for additional labeled data and may contribute to guiding the non-experts physicians, saving expert physicians time. In addition to this, stable and robust results provide more reliable systems, which make technology automatized and expand usage areas faster.

In this thesis, we aim to increase the image classification performance by active learning strategies on a small size dataset to save physicians time reserved for annotations.

1.3 Contributions of the Study

Our contributions in this thesis are as follows:

Firstly, we create a large pool that consists of synthetic samples generated with different truncation parameters to have images with different properties.

Secondly, we apply active learning strategies in the literature to select synthetic samples in a more informed fashion.

Thirdly, a different selection method, namely Weighted Margin is proposed.

Fourthly, dataset cleaning is applied to remove the samples that may negatively affect the model training.

Finally, a diversity-based approach is adopted to further increase performance by solving the overlapping problem.

1.4 Organization of the Thesis

The thesis is organized as follows; firstly, in Chapter 2, the necessary background about active learning, generative data augmentation and the medical context of the problem is provided. In Chapter 3, implementation details of the algorithms and evaluation metrics are introduced. Then, in Chapter 4, the dataset used for the problem, pre-processing steps, active learning strategies are explained. Later, in Chapter 5, experimental results are shared and discussed. Finally, in Chapter 6, conclusion is given, and proposed future works are shared for later studies.

CHAPTER 2

RELATED WORK

In this thesis, Active Learning (AL) methods are used to choose synthetic examples generated by a Generative Data Augmentation algorithm with an aim to improve the Endoscopic Mayo Score (EMS) classification performance of Ulcerative Colitis (UC) disease.

This chapter provides a background on the problem and related methods. Firstly, the literature on active learning strategies, which aims to find the most informative samples, is introduced. Then, Generative Data Augmentation techniques used in the medical image domain are mentioned, and methods in which Active Learning and Generative Data Augmentation methods are used together to both reduce the need for labeling and increase model performance are described. This is followed by a brief introduction on classical data augmentation techniques and StyleGAN2-ADA [4] which is the Generative Data Augmentation used in [7] that forms the basis of this thesis. Finally, the Ulcerative Colitis (UC) disease dataset used throughout this thesis is explained.

2.1 Active Learning

Deep learning methods are greedy for data, and conventional supervised methods need large amounts of data to converge and optimize many parameters [1]. However, in many cases, it may be difficult to obtain more that, and it may not be feasible to annotate the whole available data, especially if the amount of data keeps growing or the annotation requires special expertise such as specific medical knowledge. While the use of medical imaging for medical diagnosis is becoming widespread, there is an increasing need for reliable, automated methods, especially considering the limited number of specialized radiologists who could interpret them [8]. Generating ground truth annotations requires a unanimous decision of a committee of expert annotators, especially in safety-critical domains, increasing the annotation cost and time. Therefore, it is desired to have automated ways to select informative samples for labeling.

Active learning is a set of strategies that attempts to maximize model performance while annotating as few samples as possible, resulting in reduced labeling cost and better performance with fewer samples. As such, active learning aims to fill this gap [1, 9]

Joshi et al. [10] developed a simple yet efficient method for multi-class classification problems based on the difference between the best and the second-best prediction probabilities named as *Margin* or *Best versus Second Best (BvSB)* to estimate uncertainty. They compared their results with random sampling and entropy sampling, which uses Shannon entropy [11] to select uncertain samples. They emphasized why entropy measurement is not suitable for a multi-class classification problem under the assumption that the initial training set is large enough to make a good estimation. For a ten-class classification problem, they showed that the entropy score of a sample that is highly confused between two classes might be lower than a sample for which the model is relatively more confident about its class because of the probabilities of other classes, which are relatively unimportant. Therefore, they suggested the margin method that outperforms entropy and random for multi-class classification problems.

Sener et al. [12] introduced a method that chooses a small subset of the dataset instead of using the whole dataset. The proposed method chooses diverse samples by finding k-centers, which have the maximum distance from other data points to achieve competitive performance with the whole dataset.

Thapa et al. [13] applied active learning to select discriminative and diverse samples for semantic segmentation of polyps and depth estimation tasks. Utilization of a pre-trained model rather than PCA [14] to extract features allows for selecting task-aware samples. The combination of the Coreset [12] approach and margin uncertainty [10] provided promising results in terms of the mean intersection of union for the segmentation task and the root mean square error for the depth estimation task.

Label noise is a common problem in machine learning, leading to poor model performance. While relabeling the dataset samples may solve the problem, relabeling the entire dataset may not be possible in practice, especially in cases where there is a need for expert labelers such as a specialist medical doctor. Since relabelling budget is usually limited, it is preferable to choose re-annotation samples wisely. Bernhardt et al. [15] benefited from active learning for label cleaning. Cross-entropy loss was used to estimate the noisy samples. The ambiguity of the samples measured with entropy is used to penalize the loss to select over-confident mislabeled samples.

Gal et al. [16] used a regularization technique, Dropout [17], to estimate uncertainties. A model was trained with dropout layers placed before each weight layer. Mutual information was computed after multi-forward pass with activated dropout layers at test time. Beluch et al. [18] gathered a committee by training five classification networks with the same architecture but different initialization instead of using dropout layers for the committee. State-of-the-art results were reported emphasizing the power of ensemble knowledge. However, the ensemble of multiple models is computationally inefficient since many models need to be trained from scratch and tested. In addition, multiple forward passes are required to get multiple inferences. Choi et al. [19] estimated uncertainty by using Gaussian Mixture Models. They designed the classification and the localization heads of their object detection network to give probabilistic outputs instead of deterministic outputs as in the conventional object detection networks [20, 21, 22]. They predicted the weights, variances and means of mixtures which are used to estimate uncertainty and model output parameters.

Neural network models are likely to generate a high loss on uncertain samples and wrong predictions. Since this knowledge can be used to estimate informative samples, Yoo et al. [23] proposed adding a Learning Loss (LL) module to the network, which learns to predict the loss of unlabeled samples. Directly enlarging the dataset with top uncertain samples leads to a collection of similar samples, so the proposed method firstly selects a random subset. Then, the module can suggest that the model is likely to make a wrong decision and highly informative sample by predicting a target loss for unlabeled samples using features of several layers. Although uncertainty-based methods succeed in determining where the model has difficulty in deciding, they are prone to choose nearly identical samples and suffer from overlapping problem. Random subset selection (RSS) used in [23] is simple but not the best method since it does not consider the similarity between samples. Therefore, the overlapping problem still may occur. Wang et al. [24] proved the effectiveness of random subset selection by removing RSS from LL. Instead of RSS, they propose a combination of score network, Variational AutoEncoder (VAE) and 2 different discriminators, which construct a dual adversarial network (DAAL). The score network takes the features generated by the task model and generates informativeness scores to assign weight to features. Discriminators separate out the samples with high uncertainty representative using encoded and reconstructed weighted features via VAE.

Sinha et al. [25] trained a Variational AutoEncoder (VAE) to select unlabeled samples which are sufficiently different from the labeled samples in the latent space. VAE discriminates the labeled and unlabeled data to find the most representative samples from the unlabeled data pool. Even though data-distribution-based methods may yield good performance, under-exploiting the general structure of the task may harm the model by choosing points with little information. On the other hand, uncertainty-based methods such as the ones proposed in [26, 10, 23] disregard the data distribution. The method proposed in [25] is further improved in [27] by integration of a learning-loss module [23] combined with the "ranker" concept in RankCGAN [28] which focus on relative loss ranking more than the exact value of the loss into Variational Adversarial Active Learning (VAAL) to reach state-of-the-art results.

Citovsky et al. [29] suggested the Cluster-Margin algorithm to address the lack of information in case of using uncertainty and diversity-based algorithms separately. Firstly, a subset of unlabelled samples is selected according to the margin uncertainty score, which is the difference between the largest two class probabilities. Then, a smaller subset is selected, promoting diversity from the chosen subset according to the uncertainty scores. However, this may create batches with similar examples, resulting in overlapping problems. Besides, diversity-promoting algorithms might fail with small batch sizes while succeeding with large batch sizes. This causes variable results according to design parameters in practical applications. Ash et al. [30] used gradient embeddings of the unlabelled pool, which involve information both hidden representation and uncertainty to form a diverse informative set. The method named Batch Active learning by Diverse Gradient Embeddings (BADGE) firstly computes the gradients of the last feature layer whose magnitude indicates the uncertainty using a hypothetical label. Then, comprise a diverse set of samples k-MEANS++ [31] using the distance of gradient embeddings.

Active learning strategies are generally evaluated on balanced datasets such as [32, 33] even though in real-world scenarios data distributions are generally imbalanced. Using active learning strategies directly may promote class imbalance and result in sub-optimal classifiers. Bengar et al. [34] propose a method that considers class-balancing for both diversity and uncertainty approaches.

The efficiency of active learning is proven in many fields, such as image classification and object detection. Haussmann et al. [35] showed that using the whole dataset may lead to worse results, supporting the importance of selection of informative samples, in addition to reduced labeling efforts. They even propound that reusing informative samples rather than including all of the samples of unlabeled pool empowers the models. However, it is risky to completely rely on active learning strategies to work well since they are prone to collect outliers. Karamcheti et al. [6] exposed that random selection beats out eight different active learning strategies on a visual question answering task with extensive experiments on five models and four datasets. A remarkable boost in the performance of active learning was also reported as the number of outliers decreases. These observations highlight the importance of choosing the synthetic data produced with GANs wisely, considering outliers can frequently be generated by GANs.

2.2 Generative Data Augmentation

In the field of medical image analysis, the limited number of data, the need for experts for labeling, and the difficulty of engaging experts for labeling due to their busy schedules have forced researchers to find various methods to train robust models. Generative Data Augmentation is one of the hot topics that helps increase the limited data size with the help of synthetically generated data.

Frid-Adar et al. [36] used classical data augmentation methods to improve the performance and reported the effects of using synthetic data over baseline obtained with classical data augmentation methods. Results were supported by visualizations obtained by the t-distributed stochastic neighbor embedding (t-SNE) algorithm, which enables embedding high-dimensional data to lower-dimensional space that is two-dimensional in this case.

Çağlar [7] used Generative Adversarial Network (GAN) in their work to extend limited data size in Ulcerative Colitis (UC) disease problem and reported improvement in the performance over baseline classification models by combining classical data augmentation methods and synthetic samples. The synthetic samples were generated by Adaptive Discriminator Augmentation (ADA) [4] which is specialized for limited data. Extensive experiments using subsets of the dataset comprised of the different number of samples and design parameters are examined to enhance the dataset quality and model performance.

2.3 Combination of Generative Data Augmentation and Active Learning

Since the generative models combining two neural nets contesting each other to generate realistic images introduced by Goodfellow et al. [37], they have found several uses in a vast array of applications. As mentioned in Section 2.1, active learning enables models to be trained with less data more effectively and reduces the need for labeling efforts. Based on this idea, methods that use both active learning and generative adversarial networks have been proposed to generate synthetic samples, and select the most informative ones among them have been proposed recently.

Active learning loop includes training a model with labeled samples, inference on unlabeled pool to select informative samples, and sending them for annotation. However, the unlabeled pool may contain similar or uninformative samples, and the model developer may not have any control over the unlabelled pool. Zhu et al. [38] developed a method called Generative Adversarial Active Learning (GAAL) to generate informative synthetic samples to replace unlabeled real samples. Instead of using samples from an unlabelled pool, they used synthetic samples that are generated to be labeled by an oracle until the labeling budget is exceeded. Even though their performance does not always outperform other active learning strategies, they stated that the method has the power to compete or even beat fully supervised schemes in some cases. Mayer et al. [39] stated GAN based AL methods are usually developed and tested in simple datasets, and random sample selection does better than GAAL. Their method uses generated uncertain samples to select the most informative samples from an unlabeled pool by searching for the most similar one rather than enlarging the dataset with synthetics.

Huijser et al. [40] presented a new annotation method. The method generates synthetic samples along a 1-dimensional line using latent space representation obtained by pre-trained GAN. Annotators are asked to select the boundary point where samples change class. Even though the method outperforms uncertainty-based sampling methods, it is not easily applicable to data that cannot be visualized. According to this paper, the method is highly dependent on the quality of GANs to reach competitive or better performance with the literature and may fail otherwise.

The shortcoming of the methods above is that they are still dependent on human annotators. On the other hand, Conditional Generative Adversarial Networks (CGAN) [41] allow generating samples alongside with their corresponding labels. CGANs generate synthetic data with given auxiliary information. Dwarikanath et al. [42] used CGANs to generate chest X-ray images with different class characteristics for both classification and segmentation tasks in the first step of their work. In the second step, they utilized a Bayesian neural network to select informative samples using uncertainty scores. Results indicate the same performance with 35% of the original dataset, which corresponds to less annotation effort and time. Kong et al. [43] state that most of the generated samples in CGANs fit well to class distributions. However, discriminative features lie on the class boundaries. The proposed method uses an external reward system that hinges on uncertainty measurement based on the smallest margin [10], and entropy [11] to promote informativeness.

In order to synthesize informative anomaly samples, Duran [44] emphasized the lack of enough anomaly data to train an accurate model in the literature and adapted [45] which is an improved version of Wasserstein Generative Adversarial Networks (WGAN) [46]. In another work, Liu et al. [47] applied generative adversarial active learning strategy to outlier detection and extended their work with multiple generators to solve the mode-collapse problem.

2.4 Data Augmentation

The performance of image classification models is highly dependent on the quality and quantity of data. However, it may not be possible to obtain sufficient data to satisfactorily train a deep model in many cases, such as in medical imaging [48]. With the lack of training data, a model cannot generalize well to the unseen data, and it might overfit [49]. A straightforward way to deal with this problem is using Data Augmentation. Data augmentation is a set of techniques used to enhance data quality and increase the amount of data by generating new data points from existing data. This process can be examined in two groups, Classical Data Augmentation and Generative Data Augmentation.

2.4.1 Classical Data Augmentations

Classical data augmentation is a simple yet efficient set of techniques that generates slightly different samples by applying various methods to enhance data variance and improve the model stability, generalization ability and performance. While there is a myriad of methods since the usage of classical augmentation methods has rapidly increased with the widespread use of deep learning methods, some commonly used data augmentation techniques are as follows:

- Saturation, Brightness, Contrast
- Translation, Scaling, Rotation, Flipping, Cropping
- Adding noise, Color jittering
- CutOut [50], Mixup [51], Cutmix [52]

Different augmentation methods yield variable results on different datasets. A popular library, Albumentations developed by Buslaev et al. [53] includes more than 70 augmentation methods. The same procedure used in [7], which uses random rotation and random horizontal flip for consistency, is followed in this work. It has to be noted that finding the best classical data augmentation methods by experimenting with an extensive collection of augmentations and finding a subset that works best with them is not the scope of this thesis.

2.4.2 Generative Data Augmentation and StyleGAN2-ADA

Access to sufficient amounts of labeled data is a crucial problem, especially in medical imaging. Obtaining and labeling medical images require domain experts' time and effort. Generative Adversarial Networks allow a convenient way to increase data diversity and expose additional information from labeled set [54]. However, generative models usually need a large amount of data to converge to generate realistic synthetic images. Karras et al. [4] developed StyleGAN2-ADA to allow GANs to be trained with a small amount of data. Çağlar [7] proved this StyleGAN2-ADA architecture is suitable for medical imaging cases and how the additional information from GANs contributes to the model performance. This thesis will use GAN models with the best Fréchet Inception Distance (FID) score obtained by Çağlar [7] to generate synthetic samples. Here StyleGAN2-ADA approach will be explained briefly.

In Generative Adversarial Networks, the Generator learns to produce data while the Discriminator learns to classify whether data comes from the Generator or a real set of training data. The main problem of GANs trained with small amounts of data is the Discriminator to prone to diverge, resulting in overfitting of training. Data augmentation methods such as rotation and color transformations are the most common ways of avoiding overfitting in classification. Data augmentation can be applied to the output of the Generator, however, it is not aware whether the augmentations come from the natural distribution or not. Therefore, the Generator will start to produce augmented data besides the realistic data, named as leaking augmentations [55]. This is an undesirable side-effect as it will also generate samples from the augmented distribution.

Karras et al.[4] proposed Stochastic Discriminator Augmentation (SDA), where the Discriminator only sees the augmented images obtained from the Generator and original training images to direct the Generator to generate only clean images (i.e., from the original distribution). They also proposed Adaptive Discriminator Augmentation (ADA), the dynamic control scheme for the magnitude of the probability of augmentation. They defined a set of augmentation methods initialized with equal probability. Proposed two novel heuristics to measure overfitting value allows them to control augmentation strength. The augmentation strength is adjusted by decrementing or incrementing, considering the overfitting measurement. This novel method allowed to train GANs with small size datasets to reach acceptable FID [56] scores.

2.5 Active Learning Based Selection of Synthetic Images Generated with GAN to Improve Classification Performance of EMS of Colonoscopy Images

2.5.1 Ulcerative Colitis (UC)

Many people suffer from Ulcerative Colitis (UC), which is a chronic inflammatory bowel disease. UC exhibits itself as ulcers, inflammation, and sores in the colon UC, which can be seen for clinical, endoscopic and genetic reasons, affects the mucosal

structure in the colon. The diagnosis and identification of the disease level are critical to initiate appropriate treatment, which protects the patient from the effects of uncontrollable bleeding and cancer which can even be fatal [57, 3].

2.5.2 Colonoscopy (Endoscopy)

Endoscopy examines organs with bendable devices equipped with cameras and light sources. Examination of the large intestine with this method is called colonoscopy. During this non-surgical operation, a camera connected to the endoscopy device allows experts to diagnose mucosal diseases and record them for later examination.

Endoscopic imaging is an effective method to determine the level of Ulcerative Colitis disease [58], whose essential symptoms such as intestinal mucosal damage can be qualified and quantified using the findings. Furthermore, endoscopic imaging enables disease activity to be measured, which plays a critical role in the systematic measurement of disease activity for the proper treatment [59, 60]

2.5.3 Endoscopic Mayo Score

UC disease can generally be detected and evaluated by taking images by colonoscopy. There are different scores in the literature to determine the degree of the disease [61] such as Ulcerative Colitis Endoscopic Index of Severity (UCEIS) [62], Rachmilewitz Endoscopic score [63], Modified Mayo Clinic Endoscopic Subscore [64], Endoscopic Activity Index [65] and Endoscopic Mayo Score (Mayo Clinic Endoscopic Subscore) [66]. Among them, the most common and reliable one to rate disease is considered as Endoscopic Mayo Score [2].

The Colonoscopy Imaging dataset [5] in our case contains images of 4 different classes adjusted for the Endoscopic Mayo Score. EMS is defined as ordinal values, which show the disease level, 0,1,2,3 means healthy, mild disease, moderate disease and severe disease, respectively. As the severity of the disease increase, how the deterioration of the colon mucosa increases, bleeding and ulceration begins is shown in Figure 1.

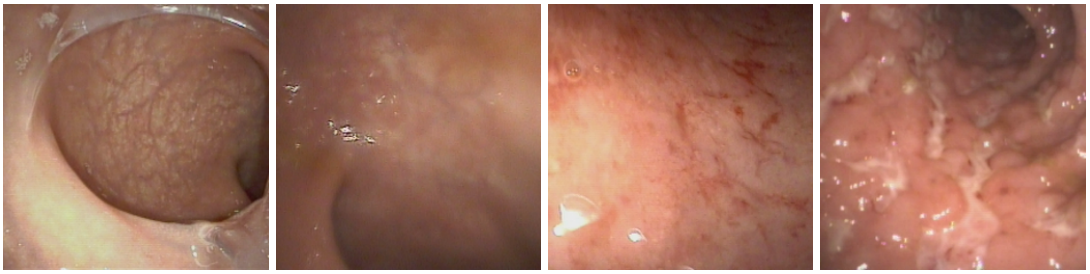


Figure 1: Example colonoscopy images with increasing EMS from 0 to 3, left to right.

2.5.4 Remission Score

Inflammatory bowel diseases are classified according to the severity of the disease and a treatment method is followed accordingly. Although there are different studies on grading the severity of the diseases in the literature, the recovery status of the disease has not been clearly defined [67]. In addition, label errors in the data, and indecision among specialist doctors, revealed a different classification that doctors often use, named Remission [7]. Therefore, Mayo-0 and Mayo-1 healthy individuals and mild cases are grouped together, while the severe cases Mayo-2 and Mayo-3 are grouped together according to the Remission score. The results considering Remission Score will also be shared throughout the thesis.

In many different applications in endoscopic imaging, computer-aided approaches are proven to be successful such as endoscopic artefact detection, polyp detection and endoscopic disease severity classification [68, 69, 70, 71, 72, 73]. The artifact and polyp detection tasks help physicians not to overlook the problematic regions and examine them in more detail. On the other hand, determining the severity of UC disease plays a critical role in predicting the possible future consequences of the disease and taking precautions. The treatment method is selected according to this severity and planned recovery. Moreover, the variability in the symptoms of UC in individuals makes it challenging to determine the correct disease severity, which causes a significant variability between the decisions of physicians and the application of the suitable treatment method [74]. In the studies conducted to measure the reliability of disease severity among the evaluators' decisions, the Cohen's Kappa value $\tilde{k}$ which also considers the possibility of an agreement by chance, was found to be $\tilde{k}$=0.45 for non-experts and $\tilde{k}$=0.71 for experts which means moderate and substantial reliability respectively [3]. Variability of decisions shows the necessity of objective evaluation for reliability and reproducibility. In this thesis, our aim is to select the most informative data to minimize the time spent by experts, obtain reliable and accurate results.

CHAPTER 3

IMPLEMENTATION DETAILS AND EVALUATION METRICS

3.1 Synthetic Dataset Generation

With the widespread use of medical imaging and the proven success of deep learning in areas such as image classification and segmentation, use of deep learning in the medical field has also been increasing. This brings the need for labeled data to train robust models. The need for specialists to label data in the medical field are factors that limit the progress of success in this field and the growth of labeled data. The fact that even experts in diseases such as UC have difficulty in making unanimous decision shows the importance of training robust models with low amount of data.

Çağlar [7] trained StyleGAN2-ADA[4] which is a Generative Adversarial Network to generate synthetic samples. Experiments were done using the truncation parameters $0.5, 0.7, 1.0, 1.2, 1.5, 2.0$, which controls the diversity of the samples. Çağlar demonstrated that there is no single best truncation value in order to get the best classification score. In each subset, a different truncation value resulted in the best F1 score. Therefore, we create a data pool with 180.000 images by generating 60.000 images with each truncation values $0.5, 1.2, 2.0$. While 0.5 value was selected to obtain samples that come from in distribution, value of 2.0 was selected to obtain diverse samples and 1.2 was a balance between them. Informative samples are selected using uncertainty score before training. Results show that samples that were generated with different truncation values are chosen by active learning approaches and contribute to the model training. Detailed explanation is given in Section 4.1.3.

3.2 Implementation details

Implementations in this thesis have been done using Python [75] language and PyTorch [76] framework. Çağlar [7] created subsets with $\mathbf{50 \times n}$ images in each class to train the generative model. A total of 10 different size subsets have been generated using $n = 1..10$. StyleGAN2-ADA [4] has been adopted as the GAN model and Resnet18 [77] as the image classification model. Experiments have been logged and tracked with the help of Weight and Biases (WANDB) framework [78].

For the evaluation of the results, two subsets having 50 and 100 images per class respectively, have been used. Since our focus is to improve the performance with limited data these subsets of having limited number of training images have been selected.

Development and training of the models have been done on a DGX1 machine with 8 V100 GPUs and a workstation with 2 RTX 3080 GPUs.

Two different training procedures, fine-tuning and training from-scratch, has been followed, details of which are given in Algorithm 1. Active learning approaches have been used to select the samples for which the model is uncertain. Since the uncertain samples are determined according to the last trained model at each step, it is better to keep the model train from that point. For the fine-tuning approach, the classification model is initialized with the weights of the baseline model in the first step. It means that all of the methods start from the same initial point. It reduces the randomness effect on the results, thus, providing a better comparison between methods. Instead of estimating boundary points from the beginning, the model is trained with relatively easy samples, which are the real ones. Then, the model is tuned with uncertain samples, which lie on the class boundaries.

In our training procedure, the total number of synthetic samples added to the training set is larger than real ones. In total, 4 times as many synthetic samples as the real samples in the training set are added. However, fine-tuning may fail if the model converges to a local minimum. Since the learning rate is decreased during fine-tuning, the model may not reach a global minimum. Therefore, the contribution of the synthetic samples might be limited. To validate this idea, we trained a baseline model from real samples. Then, the model is trained either fine-tuning or from scratch by adding new real samples. From-scratch training resulted in a better classification score. Therefore, we decided to use both training procedures to evaluate our results.

$$\begin{aligned} \mathcal{S}_{0.5} &= \{x_{0.5}^{(1)}, x_{0.5}^{(2)}, ...x_{0.5}^{(60000)}\} \\ \mathcal{S}_{1.2} &= \{x_{1.2}^{(1)}, x_{1.2}^{(2)}, ...x_{1.2}^{(60000)}\} \\ \mathcal{S}_{2.0} &= \{x_{2.0}^{(1)}, x_{2.0}^{(2)}, ...x_{2.0}^{(60000)}\} \\ \mathcal{S}^{*} &= \mathcal{S}_{0.5} \cup \mathcal{S}_{1.2} \cup \mathcal{S}_{2.0} \end{aligned} \tag{1}$$

Creation of the synthetic data pool is explained in Equation 1, where x are the samples and $\mathcal{S}_{\alpha}$ are the generated synthetic sample sets with truncation value α. α is selected as 0.5, 1.2 and 2.0. The synthetic data pool $\mathcal{S}^{*}$ is the union of those subsets. Since the same number of synthetic samples are created for each class, $\mathcal{S}^{*}$ includes 45000 images per class with a total of 180000 images.

$$\mathcal{R}_N = \{x^1, x^2, ..., x^{N \times C}\} \; where \; N = 50, 100 \tag{2}$$

Subsets of real images $\mathcal{R}_N$ are defined as in Equation 2, where C is the number of classes and N corresponds to the number of samples per class. C is selected as 4 according to Endoscopic Mayo Score in the LIMUC dataset [5].

Algorithm 1 Model training

Definitions :
N : Number of samples in a class
C : Number of classes
$\mathcal{L}$: Training set
$\mathcal{S}^*$: Synthetic data pool
$\mathcal{R}_N$: Subset of real images

$\mathcal{L} = \mathcal{R}_N$
for $k \leftarrow 0$ **to** $N \times C \times 4$ **by** $2 \times N$ **do**
  Set the learning rate lr to initial value
  if *from scratch* **then**
    Initialize classifier f_k
  else *fine tune*
    Initialize classifier f to the best $f_{k-N\times 2}$
    Set the learning rate to $lr = lr/10$
  end if
  Train a classifier f_k from $\mathcal{L}$
  Select the most informative $N/2$ instances per class using f_k, corresponding to a total of $2 \times N$ instances $\{x^*\} \in \mathcal{S}^*$
  $\mathcal{L} \leftarrow \mathcal{L} \cup \{x^*\}$
  $\mathcal{S}^* \leftarrow \mathcal{S}^* \setminus \{x^*\}$
end for

T-distributed Stochastic Neighbor Embedding (T-SNE) [79] plots have been created for visual analysis of samples after each training session. T-SNE maps image features to 2D space by minimizing Kullback-Leiber (KL) divergence.

Visualizations of T-SNE are obtained according to the following procedure.

- The Last layer of ResNet18 model is removed to obtain features.
- A model, which is trained using all of the real images in the training set, is used for feature creation.
- Using the features both from synthetic and real samples, 2D coordinates of features is computed with the help of T-SNE algorithm.
- 2D coordinates are saved and used as a look-up table for later experiments.
- After each run, the coordinates belonging to the images used in that run are shown in a figure.

Since T-SNE minimizes the cost function according to pairwise similarities, the look-up table of 2D coordinates is created to obtain a consistent plot.

3.3 Classification Metrics

Each experiment has been run 10 times and results have been obtained by taking the average results of 10 runs. Training samples must be fixed at each run since generative models were trained on those samples. Therefore, classification results are obtained by taking the average of 10 runs instead of cross-validation.

There are several classification metrics in the literature explained below. Experimental results have been calculated in terms of several metrics as described below, where TP, TN, FP and FN are True Positive, True Negative, False Positive and False Negative, respectively.

Algorithm 2 Quadratic Weighted Kappa

Definitions :
P : Predictions
G : Ground Truth
N : Number of Samples

$O = ConfusionMatrix(P, G)$
$H_P = Histogram(P)$
$H_G = Histogram(G)$
$E = (OuterProduct(H_P, H_G))/N$
for $i \leftarrow 0$ to C **do**
  for $j \leftarrow 0$ to C **do**

$$W_{i,j} = \frac{(i-j)^2}{(C-1)^2}$$

  end for
end for

$$QWK = 1 - \frac{\sum_{i,j} W_{i,j} O_{i,j}}{\sum_{i,j} W_{i,j} E_{i,j}}$$

1. Quadratic Weighted Kappa (QWK) [80], QWK is explained in Algorithm 2, measures the agreement between predictions. The term quadratic comes from the power term in the weight matrix. QWK firstly calculates the confusion matrix. Then:
   (a) The weight matrix W is used to penalize more as the difference between prediction and ground truth increases, which makes the score appropriate for an ordinal classification task.
   (b) Expected rating E is computed. Since the number of occurrence predictions and ground truths are taken into account, the metric is suitable for unbalanced data.
2. Accuracy is computed by dividing the true predictions by all predictions as in the Equation 3.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

3. Recall is computed by dividing the true positives by all of the positive samples as in Equation 4. Recall shows the ratio of positive class predictions to all positive samples. It is preferred to be used when the focus is on false negatives.

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

4. Precision is the ratio of true positive predictions to all predictions as in Equation 5. It is preferred to be used when the focus is on false positives.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

5. Ideally model is expected to predict all positive cases for better recall and only positive cases. However, in the real world, there is a trade-off between the two metrics. While recall misleads if all samples are predicted as positive, precision misleads if the model only predicts the samples very likely to be positive. Both metrics are not suitable for imbalanced data. F1 is a balanced metric that also works well in imbalanced data by combining precision and recall.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{6}$$

6. Mean Absolute Error is an evaluation metric that computes the average difference between the prediction and the ground truth as defined in the Equation 7, where y_i and x_i are the i^{th} samples of ground truth and prediction respectively, N is the total number of samples. Since the error increases as the distance increase between classes, this metric is suitable for ordinal data.

$$MAE = \sum_{i=1}^{N} |y_i - x_i| \tag{7}$$

Due to the fact that UC Mayo score that shows the severity of the disease is ordinal, Quadratic Weighted Kappa score is the most suitable metric for our case since it makes an ordinal assessment. Therefore, the main metric, over which the results have been evaluated, was selected to be Quadratic Weighted Kappa score. Use of this metric is also suggested in the LIMUC dataset [5].

CHAPTER 4

METHODOLOGY

4.1 DATASET

4.1.1 Labeled Images for Ulcerative Colitis (LIMUC) Dataset

The original dataset for the detection of Ulcerative Colitis, the LIMUC dataset [5], consists of 11276 images with the size of 352×288 from 564 patients, which were obtained by labeling the colonoscopy records throughout 8 years of studies in Marmara University Faculty of Medicine, Department of Gastroenterology. Example images from the dataset are shown in Figure 2.

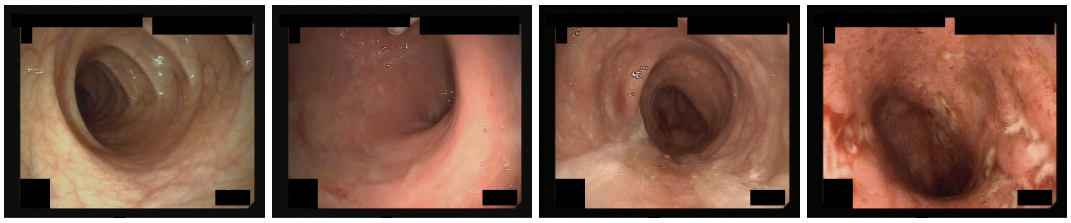


Figure 2: Example images from LIMUC dataset with Endoscopic Mayo Scores of 0,1,2 and 3 (from left to right).

There were 3 gastroenterologists involved in labelling. Two of which labeled each image according to the Endoscopic Mayo Score (EMS), containing 4 ordinal classes. When there was a discrepancy among the scores, the third, a more experienced gastroenterologist reviewed the samples without seeing previous annotations and the final decision was made by majority voting.

Class distributions of the original dataset given in Table 1 show the imbalanced nature of the dataset. The number of samples in a class decreases in relation to the severity of the disease with Mayo 0 having the most samples and Mayo 3 having the least, corresponding to 54.14% and 7.67% of all samples respectively.

Table 1: Class distributions in LIMUC dataset.

EMS Classes	Samples	Ratio
Mayo 0	6105	54.14%
Mayo 1	3052	27.07%
Mayo 2	1254	11.12%
Mayo 3	865	7.67%
Total	11276	100%

4.1.2 Pre-processing of the LIMUC Dataset

Visible medical instruments and markings overlayed in images may cause some undesired bias in image classification [7], especially considering these artefacts may potentially be unevenly distributed across the classes. In addition, when the images having visible instruments in them are used in training, it may cause image generation to underperform. Example images with instruments and markings are shown in Figure 3

Polat et al. [73] state that artifacts such as saturated or bright parts on images may cause the models to focus on undesired features, which may have negative effects in generalization and real-life performance. In the LIMUC dataset, the metadata were removed by masking those regions, as seen in Figure 3.

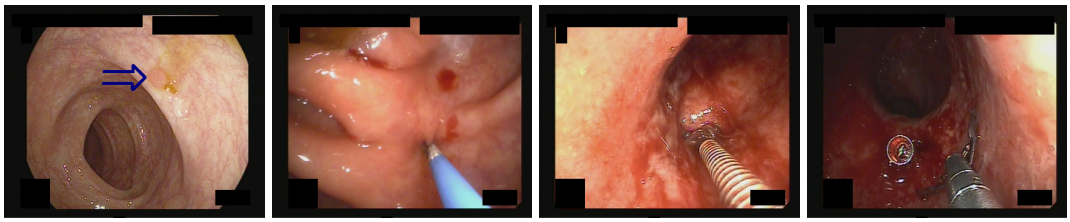


Figure 3: Example images with instruments and overlayed markings from LIMUC dataset.

On the other hand, Çağlar [7] got rid of those regions with metadata by cropping the images to the size of 256×256. In addition to this, images with signs are removed from the dataset. Since the GAN models trained in [7] to create a synthetic dataset were adapted in this thesis, the same training, validation and test subsets have been used. Example images after cropping are shown in Figure 4.

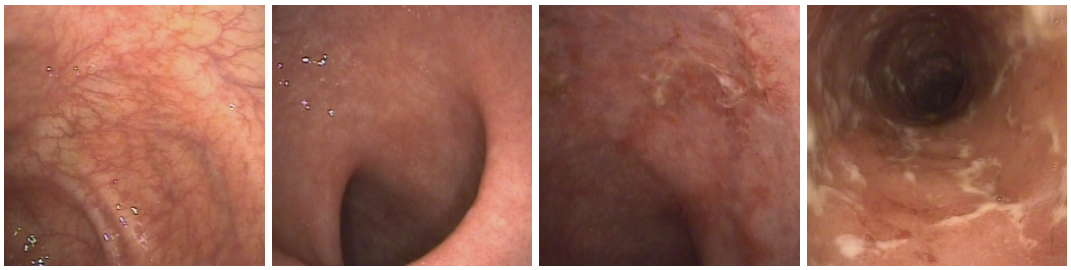


Figure 4: Example images from LIMUC dataset after pre-processing.

Çağlar applied 3 strategies for pre-processing on the LIMUC dataset.

- Removing the images in which medical instruments are visible or there are signs overlayed on the image that indicate anomaly regions.
- Cropping images to a size of 256×256 to eliminate regions with metadata and black areas.
- Creating subsets to simulate datasets having different number of images to use in the experiments.

In total, 992 images were removed from the dataset. Final dataset class distributions are given in Table 2.

Table 2: Class distributions of final dataset (images with instruments eliminated).

EMS Classes	Train	Validation	Test	Total
Mayo 0	3882	941	772	5595
Mayo 1	1924	448	406	2778
Mayo 2	832	151	146	1129
Mayo 3	542	121	119	782
Total	7180	1661	1443	10284

4.1.3 Generation of Subsets of LIMUC Dataset

Çağlar [7] created subsets from the dataset with equal number of samples in each class with the following purposes:

- Elimination of the class imbalance.
- Evaluation of the performance of GANs on different size datasets.
- Evaluation of image classification performance of the synthetic samples on different size datasets.

The class imbalance was eliminated by creating subsets from the reduced training set in Table 2. Each subset includes $\mathbf{50 \times n}$ samples where $\mathbf{n = 1..10}$. $\mathbf{n = 1}$ means that each class (i.e., $Mayo\ 0$, $Mayo\ 1$, $Mayo\ 2$ and $Mayo\ 3$) has 50 samples from the reduced training set. After the subsets were created, a GAN model was trained on each

subset. During training, weights of the GAN model with best FID score were saved. Each subset corresponds to a simulated dataset having different number of samples, as number of samples has a direct affect on the quality of synthetic samples generated by GAN. Different truncation parameters result in generation of samples with different characteristics. Therefore, the performance of the GANs with different truncation parameters and their effect on the image classification performance in the subsets are reported with an aim to investigate the importance of the truncation parameter and the success of the generated synthetic samples on the image classification performance.

The truncation parameter determines the variance of the synthetic samples. While 0 means that all of the generated outputs will be the same, 1 corresponds to the same diversity with the training set used in GAN training. When the truncation parameter is in the range of 0 and 1, generated samples come from the training dataset distribution. Values larger than 1 increase the diversity, on the other hand, they may result in generation of unrealistic samples. While different truncation values result in images having different properties, there is no best option for truncation value to obtain the best classification score for endoscopic image classification [7] and the best F1 scores are obtained with a different truncation value for each subset. For example, while a truncation value of 1.5 results in the best F1 score for the subset with 50 images per class, it generally fails to exceed the baseline performance on the subset with 200 images per class. Contrary to this, while truncation 0.5 results in the best F1 score for the subset with 200 images per class, it mostly fails to exceed the baseline performance on the subset with 50 images per class.

Preparing a labeled dataset is a time-consuming and challenging process, as it requires medical specialists and their unanimous decision, a sufficient number of patients permitting use of their data. The LIMUC dataset is a result of colonoscopy recordings within a 8-year time-span with significant labelling efforts of medical experts. Therefore, the main motivation of this thesis is improving the image classification performance using limited data by augmenting the dataset with synthetic samples. Working towards this aim, generation and informed selection of samples positively contributing to classifier performance is a fundamental aspect. For the evaluation on small-size datasets, GAN models trained on the subsets with 50 and 100 images per class have been used in this thesis to generate synthetic samples.

Earlier experiments with each truncation value in different subsets demonstrated that there are no conclusive results [7]. Therefore, a synthetic dataset with 3 different truncation parameters, i.e., 0.5, 1.2 and 2.0, that control the variance of the samples for collective knowledge, is formed in this work. In each subset, 15.000 images per class are generated using these 3 different truncation values and a collective dataset with 180.000 synthetic images is obtained. Our idea is to create a more informative and representative dataset using different truncation parameters that control the variance of the generated samples. Then, the most informative ones are selected with the help of active learning approaches.

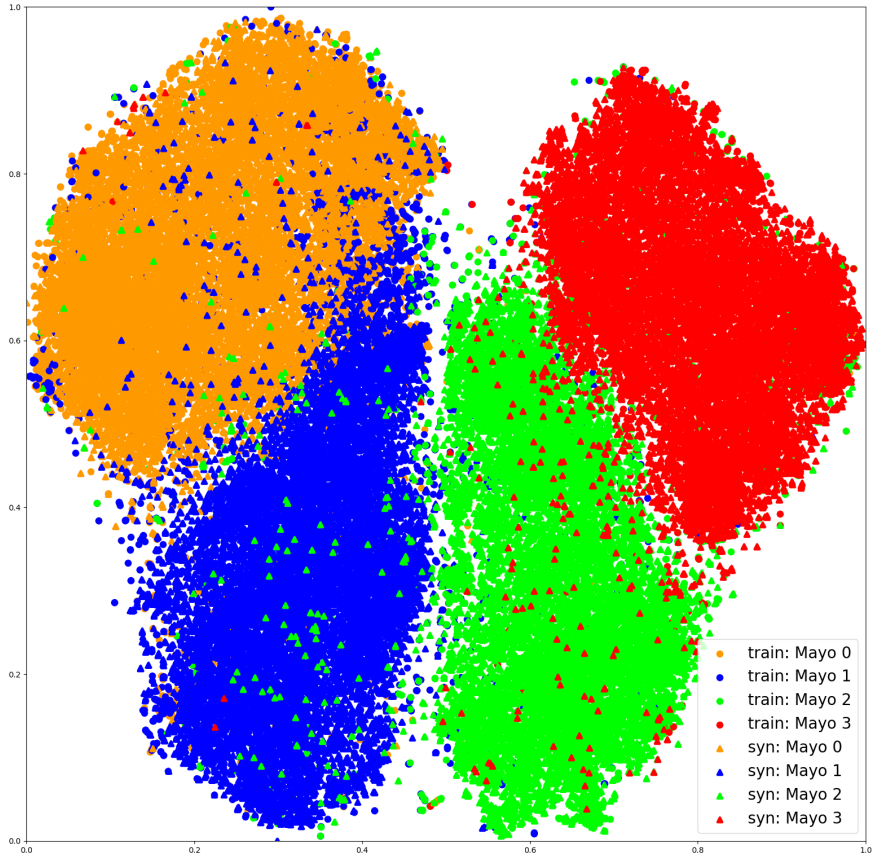
Active learning approaches are based on selecting the most informative and representative data to improve the model's performance. Uncertainty-based approaches select the informative samples using the output probabilities of a trained model. If

the dataset does not include sufficient number of uncertain samples, active learning approaches may not contribute to the model performance.

Samples generated with truncation 0.5 mostly come from the distribution center and classes are fairly well separated as seen in Figure 5a. These samples are still useful since they well represent the original data distribution, though this may not be sufficient.

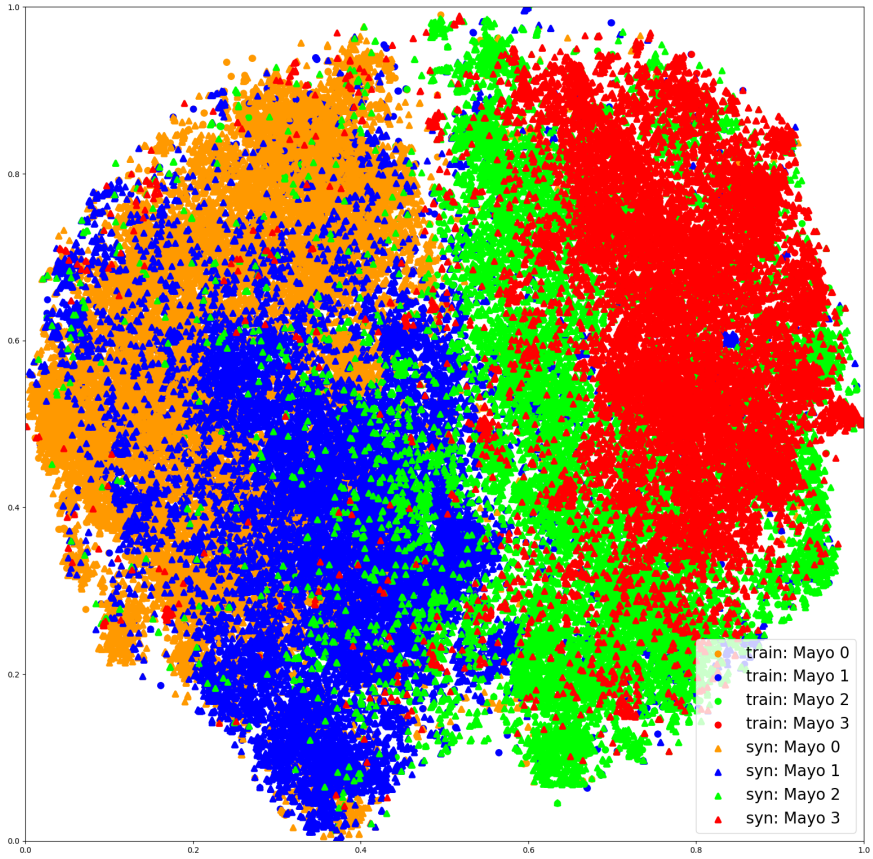
Truncation 1.2 provides a trade-off between 2.0 and 0.5. It is expected to generate a wider diversity of samples than 0.5 while it may produce out-of-distribution samples like 2.0. The T-SNE plot in Figure 5b indicates that the distributions overlap more. This overlap may be due to the unrealistic samples or informative realistic samples with high uncertainty scores.

Larger truncation means larger variance between the generated samples as shown in Figure 5c and they contain a greater variety of information that can contribute to the models' performance. However, a large number of produced samples may be from outside the distribution and these may not be realistic. Use of such unrealistic samples in training might adversely affect the model performance. This observation shows that it is necessary to sift some of the produced data with a data-centric methodology.



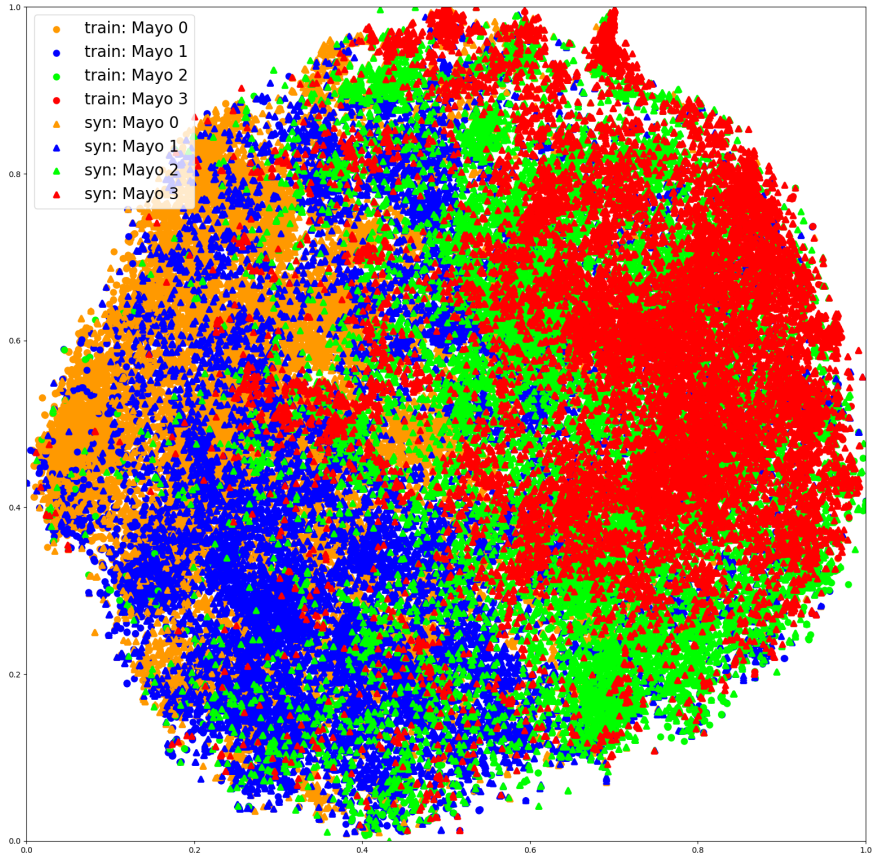
(a) Truncation = 0.5

Figure 5: Example T-SNE plots with different truncation parameters.



(b) Truncation = 1.2

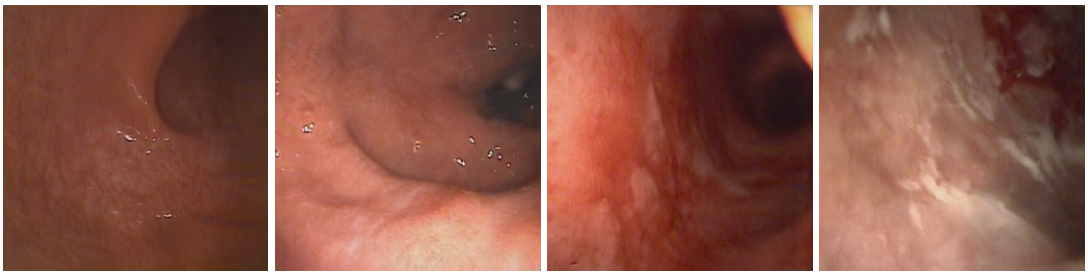
Figure 5: Example T-SNE plots with different truncation parameters.



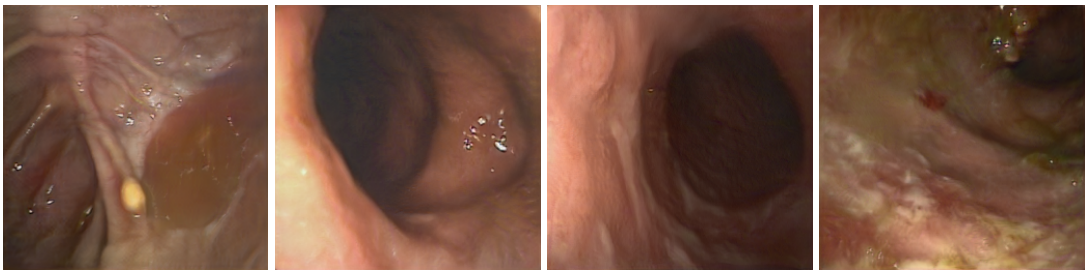
(c) Truncation = 2.0

Figure 5: Example T-SNE plots with different truncation parameters.

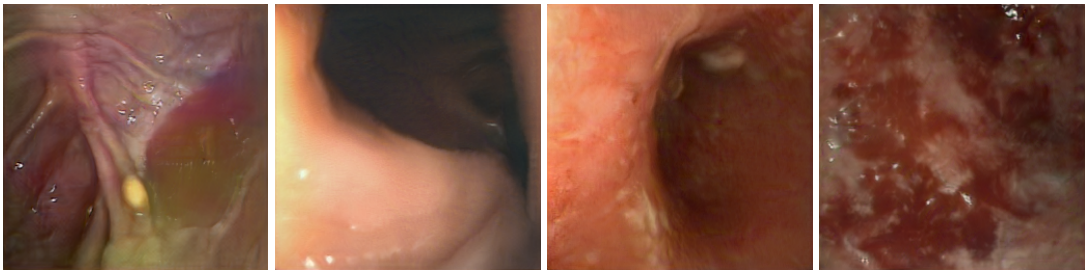
Some examples from the collective dataset are provided in Figure 6 belonging to the subset of 50 real images per class. Even though using a combination of different truncation values is expected to have a positive contribution, it may result in generation of some similar samples. Considering that the uncertainty-based algorithms are prone to selecting similar samples, use of diversity-based algorithms is necessary in conjunction with these algorithms. Therefore, we propose combining Coreset and Weighted Margin methods.



(a) truncation 0.5



(b) truncation 1.2



(c) truncation 2.0

Figure 6: Generated samples from Collective Dataset with different truncation parameters for 50 real images per class.

In this thesis, several active learning and dataset cleaning strategies are examined. After the evaluation of the baseline strategy, the naive method in [7] is applied to the collective dataset. Proposing a new method named Weighted Margin aims to increase classification performance considering the ordinal classification. Then, the dataset cleaning strategies are applied to remove the unrealistic examples from the dataset. Finally, combining these methods with Coreset, it is aimed to increase the diversity of the samples to solve the overlapping problem. Experiments can be itemized as follows:

- Baseline performance evaluation using subsets $50, 100$ of the dataset.
- Collective dataset creation using truncation values of $0.5, 1.2$ and 2.0
- Second baseline performance evaluation by adding synthetic samples randomly to the subsets.
- Adding synthetic samples with uncertainty-based active learning approaches, Entropy and Margin, to the subsets.
- Adding synthetic samples with a diversity-based active learning approach, Coreset, to the subsets.
- Adding synthetic samples with the proposed Weighted Margin method.
- Adding synthetic samples with the proposed Neighbor Margin method.
- Dataset cleaning strategies + Weighted Margin method.
- Dataset cleaning strategies + Weighted Margin score + Diversity Promoting

4.2 Training Procedure

For evaluation of image classification performance ResNet18 [77] model has been used throughout this thesis. All experiments have been repeated 10 times by using a random seed at each run.

The mean values of the experiments are reported in the figures. Two different baselines are used to evaluate our results. The first baseline is the image classification performance using the subsets of the original dataset in training. The second baseline is the naïve method in [7] which adds synthetic samples randomly from the generated cluster to the training set. The method applies the strategy to the subsets that are generated with different truncation values. We applied naive method to our collective dataset. Pseudo code of the training procedure is given in Algorithm 1 in Section 3.2. The following procedure is applied to add the samples.

- Samples are added until the total number of synthetic samples are 4 times the number of original samples.
- At each round, $N \times 2$ samples are added to the training dataset where N is the number of real samples in each class. For example, for the subset having a total of 200 original samples (50 original samples per class), 100 synthetic samples are added at each round until there are a total of 800 synthetic samples. Generally, models converge until the end of the training procedure. The purpose of the iterative addition of the synthetic samples and iterative training is to tune

the model in each round. In each round, the model is trained by selecting the new uncertain samples according to the model trained at the last round.

- For active learning approaches, the last trained model is used to determine the uncertainty scores and select diverse samples. Firstly, the baseline model is trained using the original samples. Then, the model is trained using both original and synthetic samples at the last round.
- Each of the 10 models obtained as a result of the baseline model training is trained once more with the newly added synthetic data and 10 different training results are obtained in the first round. After the first round, lastly trained models are used.
- For a better evaluation of the results 2 different training schemes are followed. From-scratch training is done in the case of model stacks in a local minimum and the first training results may highly affect the rest of the training. Fine-tuning is done to guide the model to boundary points according to lastly selected samples that the model is uncertain and to make sure of convergence. Also, the effect of randomness is decreased by giving the same initialization point to all of the methods.
  - The models are trained from scratch at each round.
  - The models are fine-tuned using the best checkpoint of the last trained model.

4.3 Dataset Cleaning

Generative models can be used to generate synthetic samples to train better and more robust models. However, the performance of traditional GAN models is highly dependent on the availability of high amounts of data. Recently proposed StyleGAN2-ADA shows impressive performance with relatively few data, i.e., thousands of images, and facilitates image generation when there are limited amounts of training data. On the other hand, synthetically generated samples and their labels may be unreliable and it is critical to eliminate unrealistic and undesired samples [7]. To alleviate this problem, Zhu et al. [38] adopted a scheme where the generated samples are labeled by annotators. However, in the medical domain, labeling needs to be done by domain experts, making it a time-consuming and costly operation.

The generated synthetic images may misguide the model training in 2 ways:

- Labels of the generated samples may be incorrect.
- Generated samples may be unrealistic and uninformative.

Here, 2 different strategies are followed to eliminate such samples:

1. Using a threshold according to the confidence and entropy scores.
2. Eliminating the samples which are not predicted correctly by 6 out of 10 models.

4.3.1 Confidence Threshold (CT)

For the first method, Confidence Threshold, our assumption is that unrealistic samples have high entropy and the model would fail to classify these. A threshold is selected according to the entropy vs. confidence graph shown in Figure 7 to eliminate these samples. The model fails to make reliable predictions with high entropy samples, i.e., it assigns a similar probability to all classes.

These samples are likely to be unrealistic as seen in Figure 8. The entropy score may be high, although a sample is predicted with high probability. For example, entropy score for a sample with class probabilities of $0.6, 0.15, 0.15, 0.1$ ($0.6log_26 + 0.15log_20.15 + 0.15log_20.15 + 0.1log_20.1$) is 1.6. Therefore, a threshold is determined according to the confidence score instead of the entropy score. The maximum entropy is obtained when all of the classes are predicted with a probability of 0.25. The threshold is selected by leaving a margin from the confidence value of 0.25. Margin-based uncertainty estimation methods explained in Section 4.4.2 compute the difference between the top two predicted classes. This value forces the margin strategy to select the samples that the model is confused between two classes. Lower values may cause the model to confuse between 3 or 4 classes which are more likely to be unrealistic. The baseline model predictions are used to remove samples from the synthetic dataset until the end of the training procedure.

Figure 8 shows eliminated examples by confidence threshold. These samples generally are not realistic and it is unclear which organ they belong to. Figure 9 shows the number of samples eliminated by confidence threshold. Totally 3023 samples eliminated for 50 images per class and 1208 samples eliminated for 100 images per class. As the number of real samples increases GAN is likely to generate better examples. Also, model converges better point. Thus, the number of eliminated samples decreases.

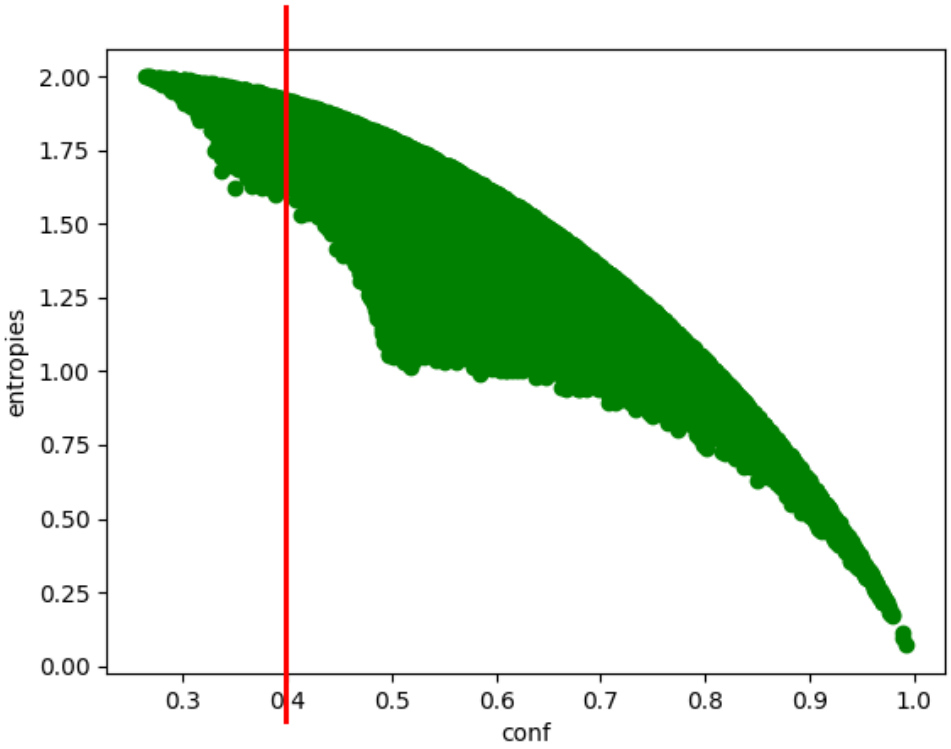


Figure 7: Entropy vs Confidence Score of Synthetic Samples.

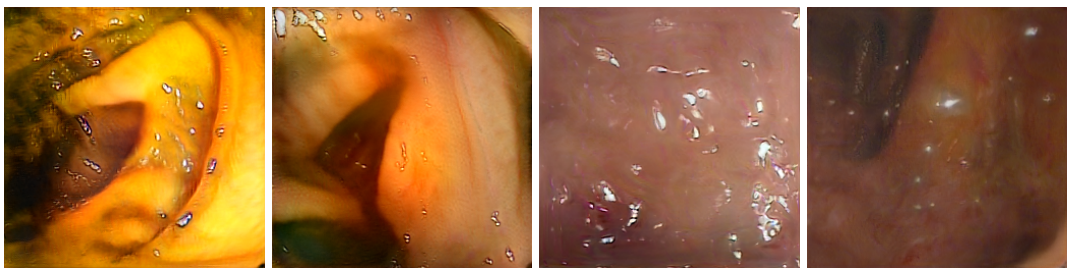
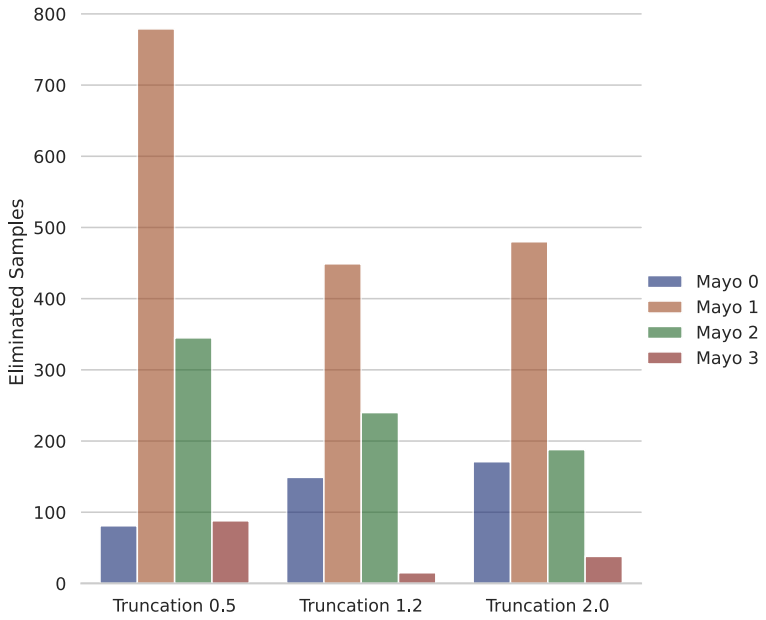
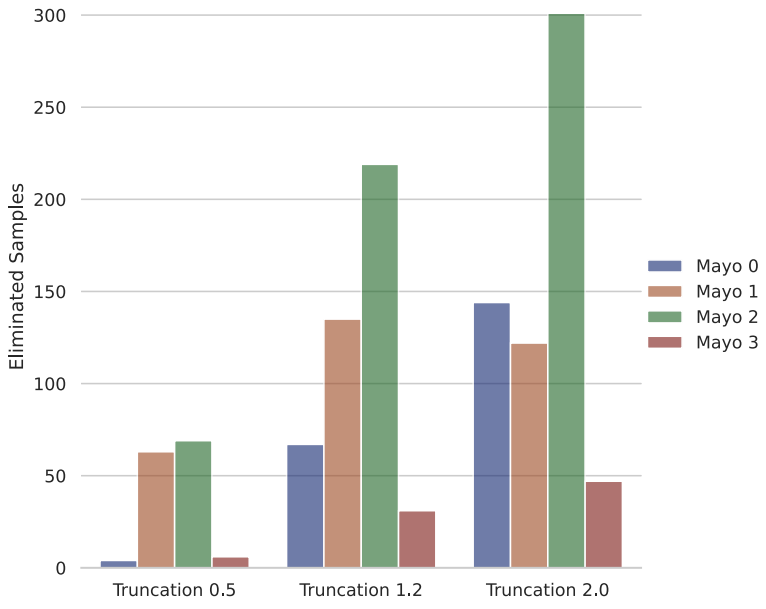


Figure 8: Examples that are eliminated using confidence threshold.



(a) 50 images per class



(b) 100 images per class

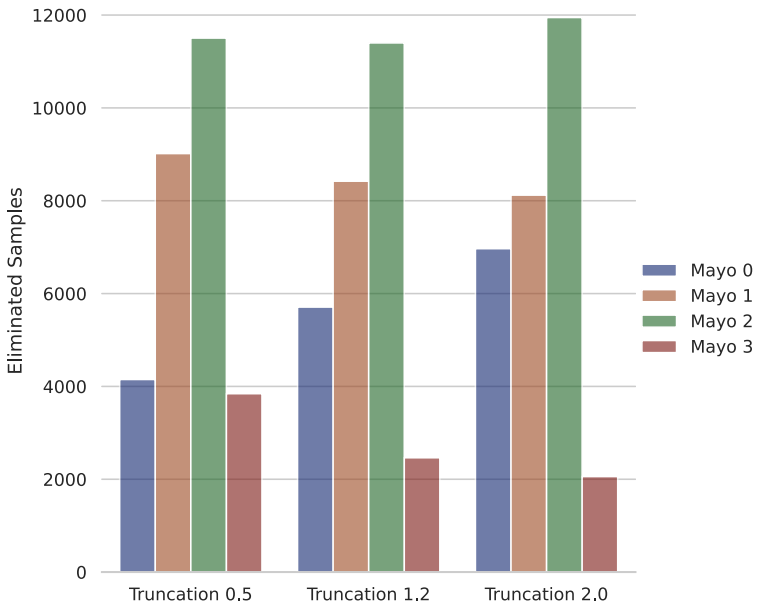
Figure 9: Distribution of eliminated samples with confidence threshold 0.4.

4.3.2 False Prediction Elimination (FPE)

The second method assumes that if 6 out of 10 models the model cannot classify a generated sample reliably, it is unrealistic.

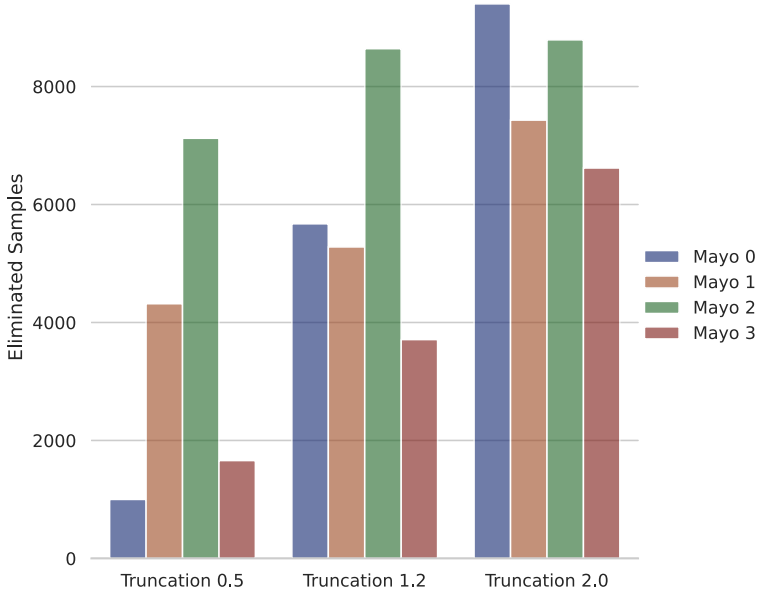
It has to be noted that either of these methods does not guarantee that the samples that do not contribute to the model performance are eliminated. Even though unrealistic samples are eliminated, highly informative uncertain samples may also be eliminated during this process.

The drawback of the method is that too many samples are eliminated after the procedure. Totally 85598 samples eliminated for 50 images per class and 69350 samples eliminated for 100 images per class. It is highly possible to eliminate highly informative samples. The Figure 10 shows the amount of eliminated samples. The same reason in confidence threshold is valid for here. As the number of real samples increases GAN is likely to generate better examples. Also, model converges better point. Thus, the number of eliminated samples decreases.



(a) 50 images per class

Figure 10: Distribution of eliminated samples with false prediction elimination.



(b) 100 images per class

Figure 10: Distribution of eliminated samples with false prediction elimination.

4.4 Active Learning Methods

Active learning methods can be grouped as diversity-based and uncertainty based. In this thesis, Coreset [12], which is a diversity-based approach, and Margin and Entropy, which are uncertainty-based approaches, are used. In addition, two new methods, Weighted Margin and Neighbor Margin are proposed. Since the Weighted Margin performs better than the Neighbor Margin, Weighted Margin is selected for further improvements. The Weighted Margin method is combined with dataset cleaning strategies and diversity-based methods.

Uncertainty-based sample selection steps are shown in Algorithm 3. Uncertainty score computation changes according to active learning strategies. While the highest entropy score corresponds to the highest uncertainty, the lowest margin score means the highest uncertainty score in margin-based methods.

Initially, the training set is a subset of real images. As described in the Algorithm 1 in Section 3.2, selected synthetic images are added to the training set after each training is completed.

Algorithm 3 Uncertainty based sample selection

Definitions :
N : Number of samples in a class
C : Number of classes
$\mathcal{L}$: Training set
$\mathcal{S}^*$: Synthetic data pool
$\mathcal{R}_N$: Subset of real images
$\mathcal{X}$: Selected samples

$\mathcal{L} = \mathcal{R}_N \cup \mathcal{X}$
for $c \leftarrow 0$ **to** C **by** 1 **do**
  Get samples $\mathcal{L}_c$ which belongs to class c from $\mathcal{L}$
  Get samples ${\mathcal{S}^*}_c$ which belongs to class c from $\mathcal{S}^*$
  Make inference on ${\mathcal{S}^*}_c$
  Compute uncertainty score
  Sort based on uncertainty score
  Select $2 \times N$ samples x_c^* with the highest uncertainty score
  $\mathcal{L}_c \leftarrow \mathcal{L}_c \cup \{x_c^*\}$
  $\mathcal{X} \leftarrow \mathcal{X} \cup \{x_c^*\}$
end for

4.4.1 Entropy

Uncertainty of a random variable can be computed by entropy. Entropy is computed by summing the product of the class probabilities and logarithms of the class probabilities as in Equation 8. Selection of samples are explained in Algorithm 3. Uncertainty score computation step uses Equation 8.

$$Entropy = \phi_{ENT}(x) = -\sum_{y} P_\theta(y \mid x) \log P_\theta(y \mid x) \tag{8}$$

$P_\theta(y \mid x)$ represents the probability of the model θ on the given sample x for the class y. ϕ_{ENT} represents the entropy based uncertainty score of the sample x.

The entropy score of a sample highly confused between two classes can be lower than a sample that is relatively more confident since the entropy score includes the probabilities of all classes. Therefore, it is not the best option for a classification problem. On the other hand, the highest entropy score is obtained when all class probabilities are equal, which means the model is not able to have a prediction on that sample. This may harm the model since those samples are synthetically generated and may be unrealistic.

4.4.2 Margin

Margin computes the difference between the top two class probabilities as in Equation 9. The method disregards the information from other classes since it only considers the probabilities of the top two classes.

$$\phi_M(x) = P_\theta\left(y_1^* \mid x\right) - P_\theta\left(y_2^* \mid x\right) \tag{9}$$

$P_\theta\left(y_1^* \mid x\right)$ represents the probability of the model θ on the given sample x for the class y_1^*. y_1^* and y_2^* represents the classes with the best and second best probabilities. ϕ_M represents the margin based uncertainty score of the sample x.

4.4.3 Weighted Margin

Weighted Margin computes the uncertainty score by promoting the samples with a class distance larger than 1 by taking the power of Margin score with class distance as in the Equation 10. Quadratic Weighted Kappa Score [80] penalizes more the wrong predictions with class distance larger than 1. The reason behind this is that class distance larger than 1 means a larger mistake because of the classes are ordinal values. Polat et al. [72] proposed a novel loss function considering the class-distance between prediction and ground-truth.

$$\phi_{WM}(x) = \left(P_\theta\left(y_1^* \mid x\right) - P_\theta\left(y_2^* \mid x\right)\right)^{|y_1^* - y_2^*|} \tag{10}$$

$P_\theta\left(y_1^* \mid x\right)$ represents the probability of the model θ on the given sample x for the class y_1^*. y_1^* and y_2^* represents the classes with the best and second best probabilities. ϕ_{WM} represents the Weighted Margin based uncertainty score of the sample x. $|y_1^* - y_2^*|$ is the power term, that prioritizes the samples with respect to the class distance.

4.4.4 Neighbor Margin

Neighbor Margin computes the Margin score on the samples for which the class distance is equal to 1 as in Equation 11.

$$\phi_{NM}(x) = \begin{cases} P_\theta\left(y_1^* \mid x\right) - P_\theta\left(y_2^* \mid x\right), & \text{if } |y_1^* - y_2^*| == 1 \\ \infty, & \text{otherwise} \end{cases} \tag{11}$$

$P_\theta\left(y_1^* \mid x\right)$ represents the probability of the model θ on the given sample x for the class y_1^*. y_1^* and y_2^* represents the classes with the best and second best probabili-

ties. ϕ_{NM} represents the Neighbor Margin based uncertainty score of the sample x. $|y_1^* - y_2^*|$ is a term that penalizes the samples with class distance larger than 1.

4.4.5 Coreset

Sener and Savarese [12] presented the Coreset approach that aims to find a diverse set of points with the maximum distance from others to represent the whole dataset. The approach computes the pairwise distances between the labeled set and the unlabeled set, which is the synthetic data pool in our case. The distance function uses image features that can be extracted with either PCA [14] or an image classification model. We extract task-aware features using our image classification model ResNet18 [77] trained on the training set $\mathcal{L}$. The labeled set covers both real examples and synthetic samples added to the training set $\mathcal{X}$. Then, the sample with the maximum distance is added to the labeled pool. This loop is iterated until the desired number of samples are obtained, which is $2 \times N$ in our case. The pseudo-code is given in Algorithm 4.

Algorithm 4 Coreset

Definitions :
N : Number of samples in a class
C : Number of classes
$\mathcal{L}$: Training set
$\mathcal{S}^*$: Synthetic data pool
$\mathcal{R}_N$: Subset of real images
$\mathcal{X}$: Selected samples
f : Classification model

$\mathcal{L} = \mathcal{R}_N \cup \mathcal{X}$
Create feature extractor f^* from f
for $c \leftarrow 0$ **to** C **by** 1 **do**
  for $k \leftarrow 0$ **to** $2 \times N \ / \ C$ **by** 1 **do**
    Get samples $\mathcal{L}_c$ belonging to class c from $\mathcal{L}$
    Get samples ${\mathcal{S}^*}_c$ belonging to class c from $\mathcal{S}^*$
    Compute $\max(pairwise\ distances(\mathcal{L}_c, {\mathcal{S}^*}_c))$
    Select sample x_c^* with maximum distance from ${\mathcal{S}^*}_c$
    $\mathcal{L}_c \leftarrow \mathcal{L}_c \cup \{x_c^*\}$
    ${\mathcal{S}^*}_c \leftarrow {\mathcal{S}^*}_c \setminus \{x_c^*\}$
    $\mathcal{X} \leftarrow \mathcal{X} \cup \{x_c^*\}$
  end for
end for

4.4.6 Weighted Margin + Diversity Promotion

Uncertainty-based active learning strategies frequently select similar samples since the trained model is likely to struggle to make decisions on almost identical samples. Therefore, uncertainty-based selection methods are prone to suffer from the overlapping problem. To alleviate this issue:

- Uncertainty scores are computed according to Weighted Margin.
- Uncertainty scores are sorted from highest to lowest.
- Normally $N \times 2$ are selected in each round. Here, $N \times 2 \times 2$ samples are selected according to uncertainty score.
- Then, $N \times 2$ samples are selected by promoting diversity with the help of Coreset algorithm.

CHAPTER 5

RESULTS AND DISCUSSION

In this thesis, Quadratic Weighted Kappa and F1 scores are used to evaluate the results and 2 different training procedures (i.e., training from scratch and fine-tuning), are used as explained in Section 3.2.

We compare the proposed methods against the naïve sample selection approach in [7] and active learning methods in the literature. We applied naïve sample selection to our collective dataset for a fair comparison and named it as Random in the following results.

The naming convention of the proposed algorithms, the details of which are explained in Chapter 4, and the ones in the literature are given in Table 3. These nomenclatures are used in the following charts and tables.

Table 3: Nomenclatures for following charts and tables.

Shortcut	Name
B	Baseline
E	Entropy
M	Margin
C	Coreset
R	Random
NM	Neighbor Margin
WM	Weighted Margin
FPE	False Prediction Elimination
CT	Confidence Threshold
DP	Diversity Promotion

Since the Quadratic Weighted Kappa penalizes more when class distances increase between the predicted class and ground truth class, QWK is selected as the main metric in this thesis.

Firstly, since the fine-tuning results are consistently better than from training from scratch, we compared the fine-tuning results of Neighbor Margin and Weighted Margin approaches. While the Weighted Margin (WM) outperforms the Neighbor Margin (NM) for 50 real images per class, both have similar results for 100 images per class

as seen in Tables 4 and 5. B_{50} and B_{100} shows the baseline performance 50 real images per class and 100 real images per class respectively. Since Weighted Margin has the best results overall, we have run the other experiments evaluating the dataset cleaning strategies and diversity promoting only on Weighted Margin method.

Previously, the amount of synthetic images to be added was determined to be 4 times the actual number of samples. But the experiments, which did not seem to have converged yet, were continued by adding synthetic images. Tables were cut at the previously determined location because the scores remained stable, the score started or continued to decrease after 4 times the synthetic images were added to the real samples. This approach applies to all shared experiments.

Table 4: Comparison of WM and NM in terms of Quadratic Weighted Kappa Scores in percentage, ($\mathbf{B_{50}} = 68.0$, $\mathbf{B_{100}} = 74.6$).

Approach	**Number of Synthetic Samples**							
	50 real samples per class							
	100	**200**	**300**	**400**	**500**	**600**	**700**	**800**
NM	69,8	71.4	71.8	72.8	73.2	73.5	**73.8**	73.7
WM	72.1	72.9	73.4	74.2	74.8	75.2	**75.7**	75.2
	100 real samples per class							
	200	**400**	**600**	**800**	**1000**	**1200**	**1400**	**1600**
NM	75.8	76.7	76.7	**77.2**	77.0	76.8	76.5	76.5
WM	75.9	76.4	**77.6**	77.0	76.8	76.8	76.7	76.3

Table 5: Comparison of WM and NM in terms of F1 Scores in percentage, ($\mathbf{B_{50}} = 54.3$, $\mathbf{B_{100}} = 59.5$).

Approach	**Number of Synthetic Samples**							
	50 real samples per class							
	100	**200**	**300**	**400**	**500**	**600**	**700**	**800**
NM	56.5	58.4	58.4	60.1	60.9	**61.3**	61.2	61.0
WM	59.4	60.6	60.9	61.5	61.7	61.8	**62.4**	61.9
	100 real samples per class							
	200	**400**	**600**	**800**	**1000**	**1200**	**1400**	**1600**
NM	61.3	62.1	62.9	63.3	**63.4**	62.9	63.0	62.6
WM	61.4	62.9	**63.4**	63.2	63.0	62.8	62.5	62.3

5.1 Fine-tuning Experiments

After the training of the baseline model, uncertain samples are selected with the help of the baseline model and active learning strategies. Firstly, the baseline model is fine-tuned with selected synthetic samples. Then, the last trained model, the ones trained with synthetic samples, are used to determine new samples to be selected.

The last trained model is fine-tuned with the new samples. This process goes in an iterative way.

5.1.1 Experiments on the Subset of 50 Real Images in Each Class

In this experiment, the baseline model has been trained using a total of 200 real images (50 images per class). 100 synthetic samples are added to the training set at each round until the training set contains 800 synthetic images in total.

Firstly, Weighted Margin and other improvement strategies are analyzed with an aim to select the best strategy to be used in the remainder of the experiments.

Results in terms of Quadratic Weighted Kappa Score and F1 Score are shown in Table 6. It shows that using confidence threshold and diversity approach together contributes the performance. The results are compared using fine-tuning strategy since, it reaches better classification scores.

Table 6: QWK and F1 Scores for WM in percentage (50 real images per class) ($\mathbf{QWK} : \mathbf{B_{50}} = 68.0, \mathbf{F1} : \mathbf{B_{50}} = 54.3$).

Approach				**Scores**	
Base Approach	**CT**	**FPE**	**DP**	**QWK**	**F1**
Weighted Margin				75.7	62.4
	X			75.5	62.3
		X		75.1	62.5
			X	75.4	62.3
	X		**X**	**75.9**	**62.9**
		X	**X**	75.0	62.4

Naïve method [7] with best F1 score in each subset can be taken as another baseline. While the highest F1 score 57.48% for the subset 50 images per class in naïve method, the highest F1 score in our work is 62.85% which is 5.37 percentage points higher. According to our experiments the F1 score is improved by 8.58 percentage points compared to baseline and 1.3 percentage points compared to Random. Compared to other active learning approaches, our approach 0.6 percentage points higher than Margin which is closest to ours. Results are shown in Table 8 and in Figure 12.

In Quadratic Weighted Kappa Score, our baseline is 68.03%. The naïve strategy, Random addition, improves the baseline 5.3 percentage points. Our method is 7.89 percentage points higher than baseline, 2.59 percentage points higher than Random. Again, we outperform the Margin by 0.59 percentage points. Results are shown in Table 7 and in Figure 11.

The best score of 75.66% in terms of QWK is obtained with the Weighted Margin method, which is 7.63% higher than the baseline and 2.33% higher than the random addition as percentage points.

It is shown in Figures 11, 12 and corresponding Tables 7, 8 that Entropy and Coreset approaches have worse results than even random addition. It proves that all of the active learning strategies are not applicable for synthetic sample selection. Weighted Margin combined with Diversity Promoting and Confidence Threshold yields the best results in both scores F1 and QWK.

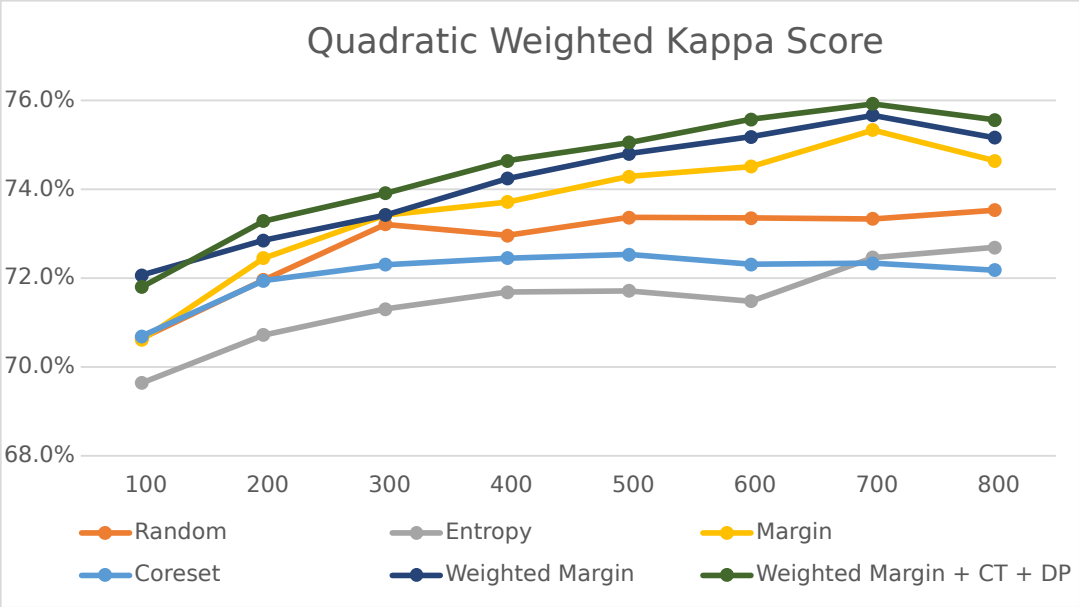


Figure 11: QWK Scores for fine-tuning (50 real images per class), $(\mathbf{QWK} : \mathbf{B_{50}} = 68.0)$.

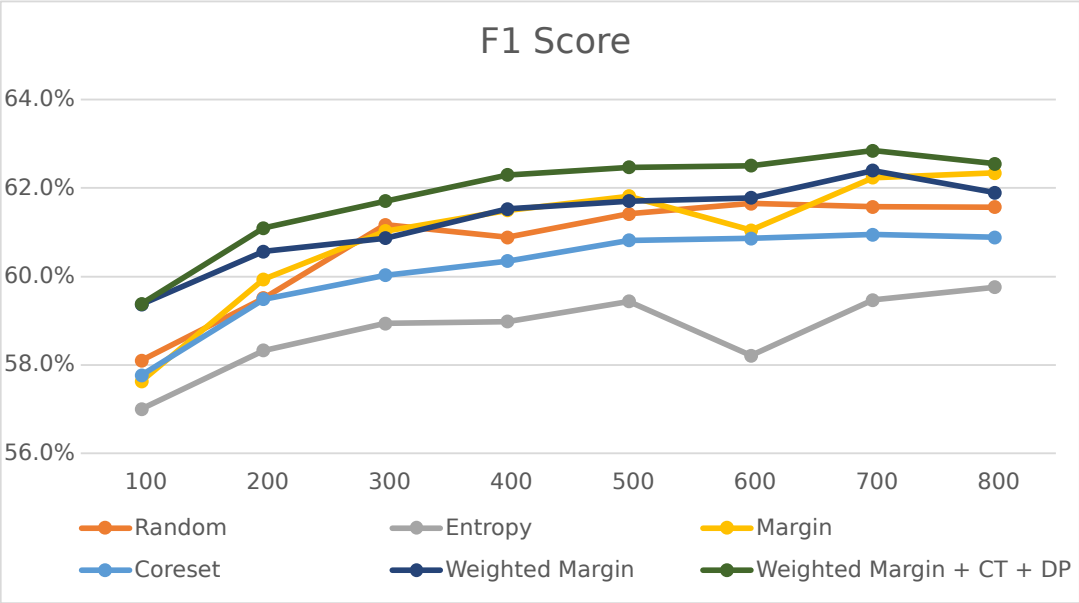


Figure 12: F1 scores for fine-tuning (50 Real Images Per Class), $(\mathbf{F1} : \mathbf{B_{50}} = 54.3)$.

Table 7: QWK Scores for fine-tuning in percentage (50 Real Images Per Class) ($\mathbf{B_{50}} = 68.0$).

Approach	Number of Synthetic Samples							
	100	**200**	**300**	**400**	**500**	**600**	**700**	**800**
Random	70.6	72.0	73.2	73.0	73.4	73.4	73.3	73.5
Entropy	69.6	70.7	71.3	71.7	71.7	71.5	72.5	72.7
Margin	70.6	72.5	73.4	73.7	74.3	74.5	75.3	74.6
Coreset	70.7	71.9	72.3	72.5	72.5	72.3	72.3	72.2
WM	72.1	72.9	73.4	74.2	74.8	75.2	**75.7**	75.2
WM+CT+DP	71.8	73.3	73.9	74.6	75.1	75.6	**75.9**	75.6

Table 8: F1 Score for fine-tuning in percentage (50 real images per class) ($\mathbf{B_{50}} = 54.3$).

Approach	Number of Synthetic Samples							
	100	**200**	**300**	**400**	**500**	**600**	**700**	**800**
Random	58.1	59.5	61.2	60.9	61.4	61.7	61.6	61.6
Entropy	57.0	58.3	58.9	59.0	59.4	58.2	59.5	59.8
Margin	57.6	59.9	61.0	61.5	61.8	61.1	62.2	62.3
Coreset	57.8	59.5	60.0	60.4	60.8	60.9	61.0	60.9
WM	59.4	60.6	60.9	61.5	61.7	61.8	**62.4**	61.9
WM+CT+DP	59.4	61.1	61.7	62.3	62.5	62.5	**62.9**	62.6

5.2 Experiments on the Subset of 100 Images in Each Class

In this experiment, the baseline model is trained using a total of 400 real images (100 images per class). 200 synthetic samples are added to the training set at each round until the training set contains 1600 images in total.

Firstly, Weighted Margin and other improvement strategies are analyzed. Best resulted strategy is selected for later tables and charts.

Results belongs to Quadratic Weighted Kappa Score and F1 Score are shown in the Table 9. The results are compared according to fine-tuning strategy since, it reaches better classification scores than from-scratch training. Weighted Margin and Confidence Threshold reaches the highest score in QWK which is 0.6 higher than Margin and Random, 0.4 higher than Coreset and 3.1 higher than baseline as percentage points. For consistency in the plots and charts, we shared the Weighted Margin + CT + DP in Tables 10 and 11, in Figures 13 and 14. The tables and Figures shows that our approaches outperform others. Entropy yields the worst results which means all of the active learning approaches are not suitable for this task.

Best F1 score obtained with Naïve method [7] is 61.49%. Random addition, corresponding to Naïve approach, reaches 63.13% using our collective synthetic dataset and fine-tuning training procedure. According to our experiments F1 score is improved by 4.29, 0.67, 1.56, 0.39, 0.19 percentage points compared to Baseline, Random, Entropy, Margin and Coreset respectively. Results are shown in Table 10 and in Figure 14.

Table 9: F1 and QWK Scores for WM in percentage (100 real images per class) ($\mathbf{QWK} : \mathbf{B_{100}} = 74.6$, $\mathbf{F1} : \mathbf{B_{100}} = 59.5$).

Approach				**Scores**	
Base Approach	**CT**	**FPE**	**DP**	**QWK**	**F1**
Weighted Margin				77.6	63.4
	X			**77.7**	**63.8**
		X		77.5	63.6
			X	77.3	63.6
	X		**X**	77.5	**63.8**
		X	**X**	77.0	63.3

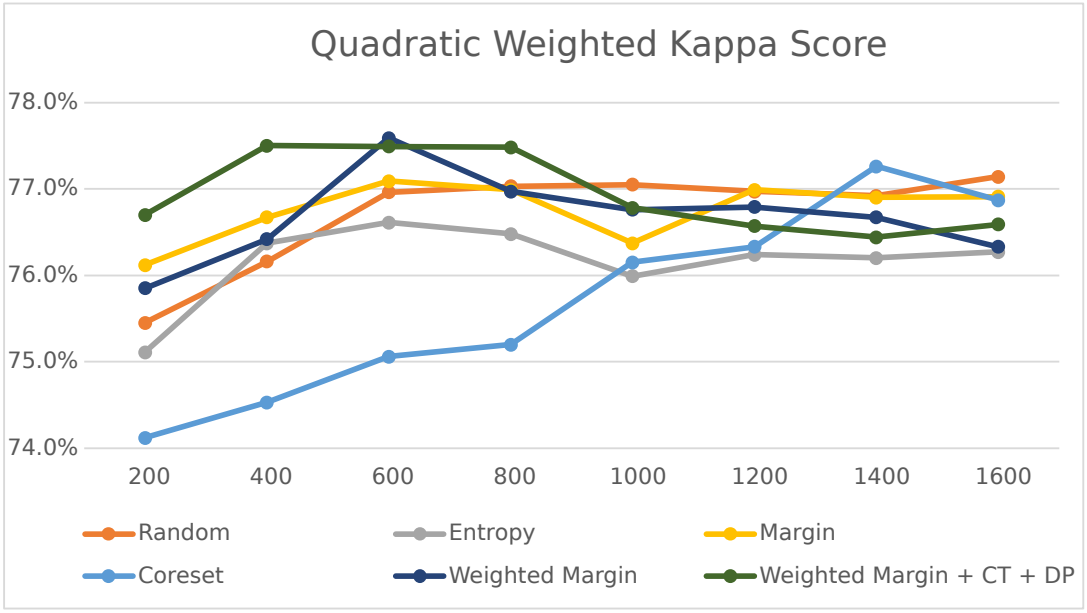


Figure 13: QWK scores for fine-tuning (100 real images per class), ($\mathbf{QWK} : \mathbf{B_{100}} = 74.6$).

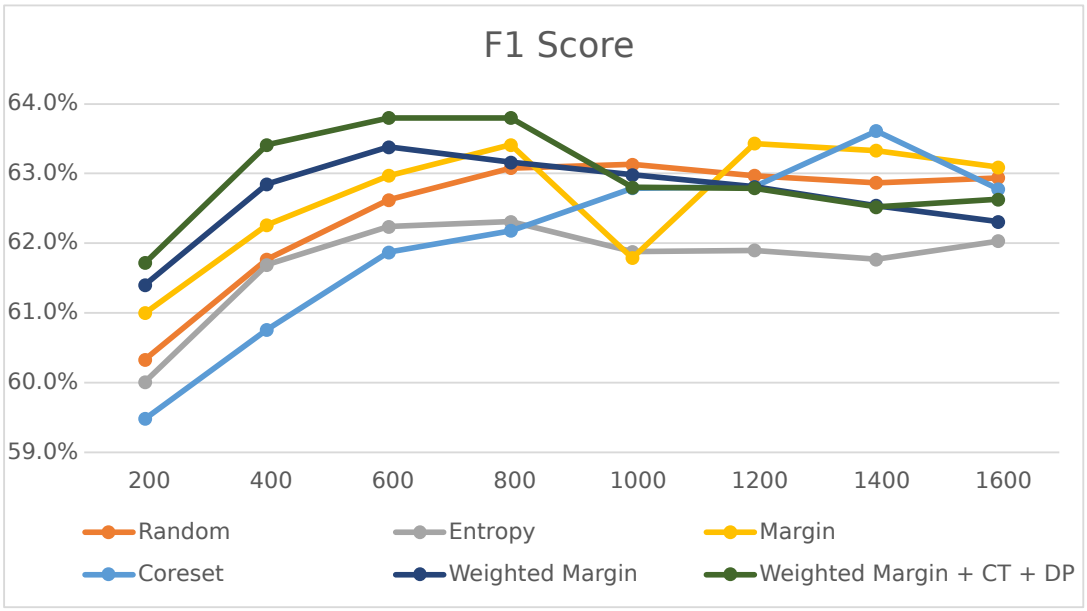


Figure 14: F1 cores for fine-tuning (100 real images per class), ($\mathbf{F1} : \mathbf{B_{100}} = 59.5$).

Table 10: QWK scores for fine-tuning in percentage (100 real images per class), $(\mathbf{QWK} : \mathbf{B_{100}} = 74.6)$.

Approach	Number of Synthetic Samples							
	200	**400**	**600**	**800**	**1000**	**1200**	**1400**	**1600**
Random	75.5	76.2	77.0	77.0	77.1	77.0	76.9	77.1
Entropy	75.1	76.4	76.6	76.5	76.0	76.2	76.2	76.3
Margin	76.1	76.7	77.1	77.0	76.4	77.0	76.9	76.9
Coreset	74.1	74.5	75.1	75.2	76.2	76.3	77.3	76.9
WM	75.9	76.4	**77.6**	77.0	76.8	76.8	76.7	76.3
WM+CT+DP	76.7	77.5	77.5	77.5	76.8	76.6	76.4	76.6

Table 11: F1 scores for fine-tuning in percentage (100 real images per class), $(\mathbf{F1} : \mathbf{B_{100}} = 59.5)$.

Approach	Number of Synthetic Samples							
	200	**400**	**600**	**800**	**1000**	**1200**	**1400**	**1600**
Random	60.3	61.8	62.6	63.1	63.1	63.0	62.9	62.9
Entropy	60.0	61.7	62.2	62.3	61.9	61.9	61.8	62.0
Margin	61.0	62.3	63.0	63.4	61.8	63.4	63.3	63.1
Coreset	59.5	60.8	61.9	62.2	62.8	62.8	63.6	62.8
WM	61.4	62.9	63.4	63.2	63.0	62.8	62.5	62.3
WM+CT+DP	61.7	63.4	**63.8**	**63.8**	62.8	62.8	62.5	62.6

5.3 From-scratch Experiments

After the training of the baseline model, uncertain samples are selected with the help of the baseline model and active learning strategies. Then, a ResNet18 model is initialized with the weights pre-trained on ImageNet to follow from-scratch training. Then, the model is trained. This process goes in an iterative way.

5.3.1 Experiments on the Subset of 50 real images per class

Entropy, an active learning approach, underperform Random in F1 score. However, in QWK, Random underperform Entropy which means that Random fails to distinguish distant classes compared to other approaches.

According to Table 12 and Figure 15 the proposed method Weighted Margin outperforms other approaches in QWK score. Confidence Threshold and Diversity Promoting takes the score a step further. Even though with CT + DP reaches the best scores for both F1 and QWK, margin is better than Weighted Margin in F1 score according to Table 13 and Figure 16. This means that WM helps models to distinguish distant

classes better as expected. However, it underperform Margin in neighbour classes since focuses more on samples where model predictions have difficulty distinguishing distant classes.

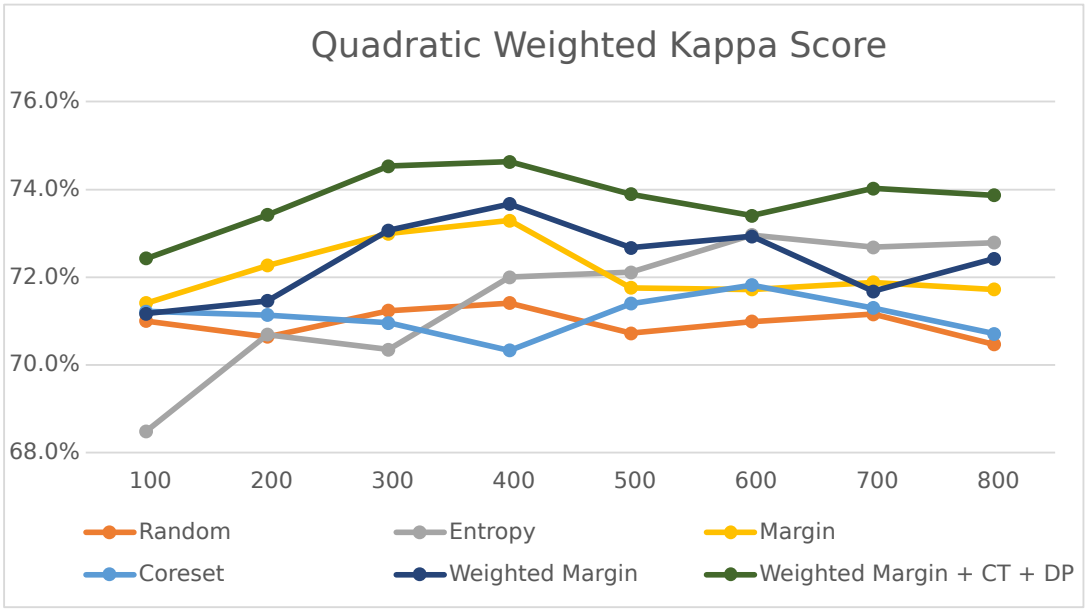


Figure 15: QWK scores for from-scratch (50 real images per class), ($\mathbf{QWK}$: $\mathbf{B_{50}} = 68.0$).

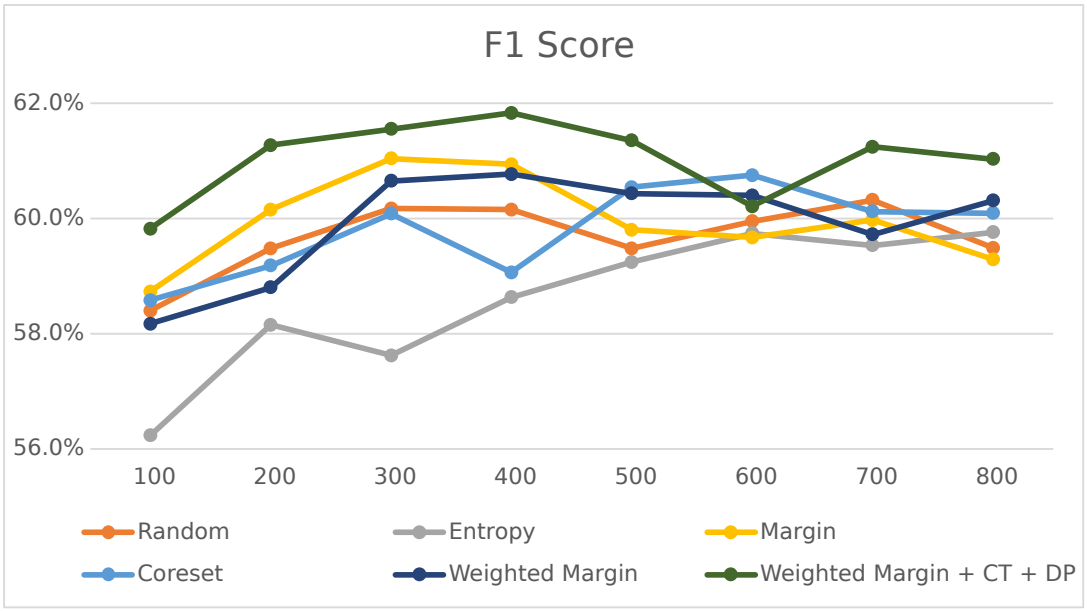


Figure 16: F1 scores for from-scratch (50 real images per class), ($\mathbf{F1}$: $\mathbf{B_{50}} = 54.3$).

Table 12: QWK scores for from-scratch in percentage (50 real images per class), $(\mathbf{QWK} : \mathbf{B_{50}} = 68.0)$.

Approach	Number of Synthetic Samples							
	100	**200**	**300**	**400**	**500**	**600**	**700**	**800**
Random	71.0	70.6	71.2	71.4	70.7	71.0	71.2	70.5
Entropy	68.5	70.7	70.4	72.0	72.1	73.0	72.7	72.8
Margin	71.4	72.3	73.0	73.3	71.8	71.7	71.9	71.7
Coreset	71.2	71.1	71.0	70.3	71.4	71.8	71.3	70.7
WM	71.2	71.5	73.1	73.7	72.7	72.9	71.7	72.4
WM+CT+DP	72.4	73.4	74.5	**74.6**	73.9	73.4	74.0	73.9

Table 13: F1 scores for from-scratch in percentage (50 real images per class), $(\mathbf{F1} : \mathbf{B_{50}} = 54.3)$.

Approach	Number of Synthetic Samples							
	100	**200**	**300**	**400**	**500**	**600**	**700**	**800**
Random	58.4	59.5	60.2	60.2	59.5	60.0	60.3	59.5
Entropy	56.2	58.2	57.6	58.6	59.2	59.7	59.5	59.8
Margin	58.7	60.2	61.0	60.9	59.8	59.7	60.0	59.3
Coreset	58.6	59.2	60.1	59.1	60.5	60.8	60.1	60.1
WM	58.2	58.8	60.7	60.8	60.4	60.4	59.7	60.3
WM+CT+DP	59.8	61.3	61.6	**61.8**	61.4	60.2	61.2	61.0

5.3.2 Experiments on the Subset of 100 Real Images Per Class

According to Table 14 and Figure 17 Random, Margin, and WM+CT+DP reaches the same highest score. However, proposed approach, WM+CT+DP reaches the highest score.

Table 15 and Figure 18 shows that WM+CT+DP reaches the best performance, while Margin follows it.

No method falls below baseline performance for any amount of synthetic images added.

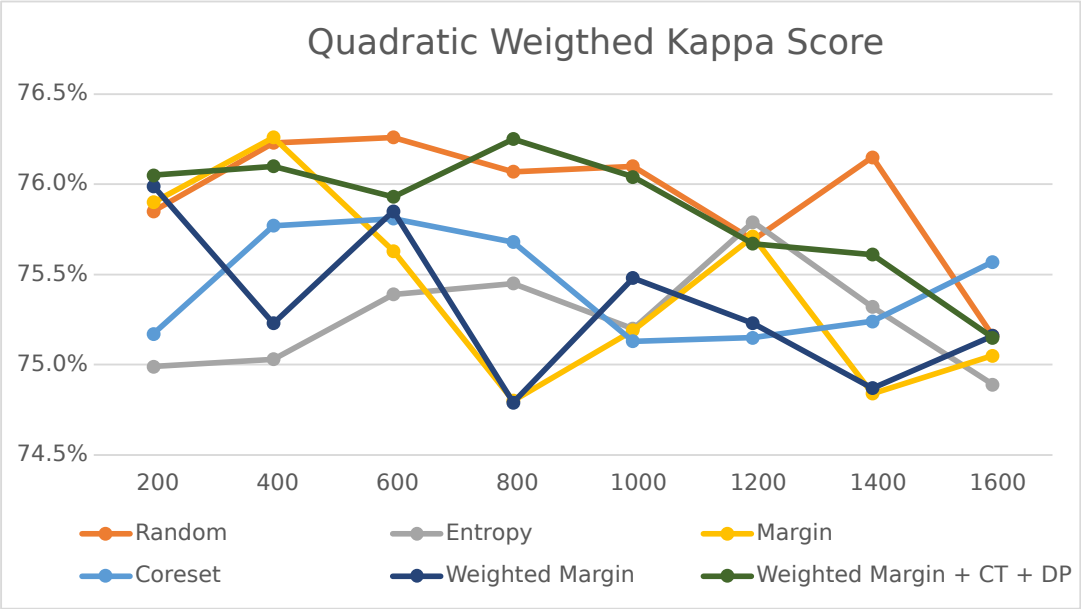


Figure 17: QWK scores for from-scratch (100 real images per class), ($\mathbf{QWK}$: $\mathbf{B_{100}} = 74.6$).

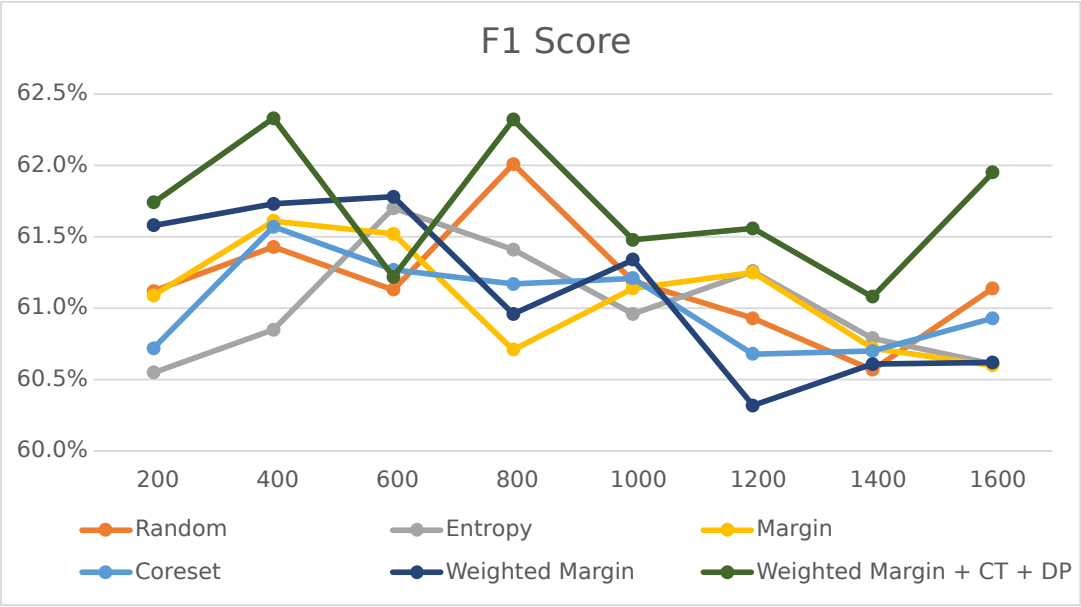


Figure 18: F1 scores for from-scratch (100 real images per class), ($\mathbf{F1}$: $\mathbf{B_{100}} = 59.5$).

Table 14: QWK scores for from-scratch in percentage (100 real images per class), $(\mathbf{QWK} : \mathbf{B_{100}} = 74.6)$.

Approach	Number of Synthetic Samples							
	200	**400**	**600**	**800**	**1000**	**1200**	**1400**	**1600**
Random	75.9	76.2	**76.3**	76.1	76.1	75.7	76.2	75.2
Entropy	75.0	75.0	75.4	75.5	75.2	75.8	75.3	74.9
Margin	75.9	**76.3**	75.6	74.8	75.2	75.7	74.8	75.1
Coreset	75.2	75.8	75.8	75.7	75.1	75.2	75.2	75.6
WM	76.0	75.2	75.9	74.8	75.5	75.2	74.9	75.2
WM+CT+DP	76.1	76.1	75.9	**76.3**	76.0	75.7	75.6	75.2

Table 15: F1 scores for from-scratch in percentage (100 real images per class), $(\mathbf{F1} : \mathbf{B_{100}} = 59.5)$.

Approach	Number of Synthetic Samples							
	200	**400**	**600**	**800**	**1000**	**1200**	**1400**	**1600**
Random	61.1	61.4	61.1	62.0	61.2	60.9	60.6	61.1
Entropy	60.6	60.9	61.7	61.4	61.0	61.3	60.8	60.6
Margin	61.1	61.6	61.5	60.7	61.1	61.3	60.7	60.6
Coreset	60.7	61.6	61.3	61.2	61.2	60.7	60.7	60.9
WM	61.6	61.7	61.8	61.0	61.3	60.3	60.6	60.6
WM+CT+DP	61.7	**62.3**	61.2	**62.3**	61.5	61.6	61.1	62.0

5.4 Discussion

Results show that wisely selection of generated samples improves the image classification performance. The proposed Weighted Margin approach outperforms the baseline and random sample selection methods as well as other active learning approaches in the literature. We speculate that this is mainly due to the it taking the ordinal structure of the classes into account, since the classes in question (i.e., EMS) are ordinal. Predicting Mayo 0 as Mayo 3 is a larger mistake than predicting Mayo 0 as Mayo 1. The proposed Weighted Margin approach tries to select samples by promoting the ones that model might make a larger mistake. On the other hand, the model used in this thesis uses Cross-entropy loss function which does not consider ordinal structure of the classes and the results may be further improved with an appropriate loss function considering ordinal structure of the classes, such as the one proposed in [72]. We observed that QWK and loss sometimes increase together during training, which is an indicator of the inconsistency of the QWK score and loss function. Using the Class Distance Weighted Cross-Entropy Loss in [72] will possibly give better results for Weighted Margin approach. It will make the results more stable.

Confidence Threshold value reduced performance when used alone for the subset with 50 real images per class. When the selected samples were examined, it was observed

that the number of similar samples increased after the threshold value was applied. The benefit was observed when the Confidence Threshold was used with Diversity Promoting, with even better results.

The experimental results suggest that training from scratch is not the best approach for this task and fine-tuning is more preferable. This is in-line with our expectations, since when training-from-scratch, there is no record of where the model converges at the last iteration. Active learning strategies aim to identify the samples for which the model fails in an iterative way. This causes a conflict with training-from-scratch and makes it less likely to get the best results out of active learning. In addition to this, it is fair to compare algorithms according to from scratch results, since algorithms can not reach their potential improvement in from scratch training.

The reason that decrease in the results after some point is faster in from-scratch than fine-tune is that fine-tune resumes from the best checkpoint of the model that is trained in the last iteration. Hence, after one iteration, if a drop in score is observed, it could be reverted back to the best checkpoint, preventing a reduction in scores and making the results more stable.

We also tested a strategy where Coreset and Weighted Margin approaches are used in association: 50% of the samples are selected with Coreset and 50% of the samples are selected with Weighted Margin to promote diversity as suggested in the literature [13]. However, since the classification scores got worse, we tried to increase diversity by selecting diverse samples from uncertainty pool as explained in Section 4.4.6.

Results of combination of the Confidence Threshold and False Prediction Elimination is not shared. The union of CT and FPE is almost same with FPE since too many samples already removed in FPE. Therefore, results are almost same. FPE did not contributed the performance. It eliminates almost half of the generated samples which means a loss of information.

We tried to select most similar samples with the real samples in the training set to make sure that selected samples comes from the original distribution. This approach got worse than other margin based approaches since it is contrary to uncertainty and diversity based active learning strategies.

CHAPTER 6

CONCLUSION AND FUTURE WORK

UC is a chronic bowel disease that physicians have difficulty in determining its severity. This shows that it would be beneficial to automatically identify the disease with the help of deep learning. Sufficient data is needed to train deep learning models. However, the difficulty of accessing labeled data in the medical field shows the necessity of methods for training robust models with small datasets. Although synthetic data generation is offered as a solution to this need, it is possible to further improve performance by selecting samples wisely. In this thesis, firstly, a synthetic pool with more information was created using different GAN parameters. Then, performance improvements were achieved by using active learning methods. Since disease severity is an ordinal data, a more appropriate selection method, Weighted Margin, have been proposed. It has been seen that synthetic data also produces unrealistic samples that will not contribute to the model performance, and several methods have been proposed to eliminate them. Finally, it has been seen that the selected samples can be similar. Therefore, it has been shown that it is beneficial to add diversity-based algorithms to add more diverse samples to the training set. Table 16 shows the performance improvements of our approach over other methods.

Table 16: Performance improvements of our approach over other methods

Methods	**Performance Increase as Percentage Points**			
	50 real images per class		**100 real images per class**	
	QWK Score	**F1 Score**	**QWK Score**	**F1 Score**
Baseline	7.9	8.6	3.1	4.3
Naive method [7]	N/A	7.1	N/A	2.3
Random [7]	2.4	1.2	0.6	0.7
Margin [10]	0.6	0.6	0.6	0.5
Entropy [11]	3.2	3.1	1.1	1.5
Coreset [12]	3.4	1.9	0.4	0.2

Even though our approach contributes the literature in several ways, there is a room for improvement. In our case, GAN is prone to generate similar samples. Feeding similar samples to dataset will result in overlapping problem. Therefore, very similar samples need to be eliminated either. Similar images can be eliminated using a similarity metric and image features. For the dataset cleaning procedure, another purpose

is to detect outliers in synthetic samples. Coreset approach select the samples with highest distance to the training set. Entropy score defines the how uncertain model is about the sample. The outliers are with high entropy score and high distance from the training distribution. Therefore, the combined distance score can be used to find and eliminate the outliers. The distance score can be thresholded in 2 ways, using the scores of real samples, or using the scores of synthetic samples. In both scores Inter Quartile Range (IQR) can be used to determine threshold. If the threshold is selected using real samples, the threshold value should be increased by a margin since we would like to increase the diversity and uncertainty of the dataset. If synthetic samples are used to determine threshold, the threshold can be defined by IQR with appropriate margin. The margin can be chosen with an easy test. It is expected that the model performance decreases if samples eliminated by threshold are fed into the network instead of eliminating. However, it is better to test these after similar samples are eliminated from the dataset. Because, they effect both IQR computation and model performance by creating bias on some samples.

Training, validation and test set may include mislabelled samples. These samples may be obtained according to model prediction results. The wrongly predicted samples with low entropy and high confidence score are likely to be mislabelled. These samples can be collected for re-annotation, can be removed from the dataset or their labels can be changed according to model prediction. Same approach can be applied on the synthetic dataset. However, uncertainty based approaches try to select most uncertain samples which do not have high confidence. Therefore, we did not evaluate this approach on the synthetic dataset throughout this thesis.

Active learning strategy can be applied to real examples as well. Reusing the informative sample in model training might improve the model performance.

Active learning and dataset cleaning strategies can be also used with training of Generative Adversarial Networks. For example, this can be done by guiding GAN training to generate informative and representative samples by adding a classification network and loss considering uncertainty score and similarity score.

REFERENCES

[1] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, B. B. Gupta, X. Chen, and X. Wang, "A survey of deep active learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 9, pp. 1–40, 2021.

[2] G. D'haens, W. J. Sandborn, B. G. Feagan, K. Geboes, S. B. Hanauer, E. J. Irvine, M. Lémann, P. Marteau, P. Rutgeerts, J. Schölmerich, *et al.*, "A review of activity indices and efficacy end points for clinical trials of medical therapy in adults with ulcerative colitis," *Gastroenterology*, vol. 132, no. 2, pp. 763–786, 2007.

[3] N. M. Vashist, M. Samaan, M. H. Mosli, C. E. Parker, J. K. MacDonald, S. A. Nelson, G. Zou, B. G. Feagan, R. Khanna, and V. Jairath, "Endoscopic scoring indices for evaluation of disease activity in ulcerative colitis," *Cochrane Database of Systematic Reviews*, no. 1, 2018.

[4] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," *arXiv*, 2020.

[5] G. Polat, H. T. Kani, I. Ergenc, Y. O. Alahdab, A. Temizel, and O. Atug, "Labeled Images for Ulcerative Colitis (LIMUC) Dataset," Mar. 2022.

[6] S. Karamcheti, R. Krishna, L. Fei-Fei, and C. D. Manning, "Mind your outliers! investigating the negative impact of outliers on active learning for visual question answering," in *ACL*, 2021.

[7] M. Çağlar, "Improving classification performance of endoscopic images with generative data augmentation," Master's thesis, Middle East Technical University, 2022.

[8] P. Liu, L. Wang, R. Ranjan, G. He, and L. Zhao, "A survey on active deep learning: from model-driven to data-driven," *ACM Computing Surveys (CSUR)*, 2021.

[9] X. Zhan, Q. Wang, K.-h. Huang, H. Xiong, D. Dou, and A. B. Chan, "A comparative survey of deep active learning," *arXiv preprint arXiv:2203.13450*, 2022.

[10] A. J. Joshi, F. Porikli, and N. Papanikolopoulos, "Multi-class active learning for image classification," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2372–2379, 2009.

[11] C. E. Shannon, "A mathematical theory of communication," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 5, p. 3–55, jan 2001.

[12] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *International Conference on Learning Representations*, 2018.

[13] S. K. Thapa, P. Poudel, B. Bhattarai, and D. Stoyanov, "Task-aware active learning for endoscopic image analysis," *arXiv preprint arXiv:2204.03440*, 2022.

[14] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1, pp. 37–52, 1987. Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists.

[15] M. Bernhardt, D. C. Castro, R. Tanno, A. Schwaighofer, K. C. Tezcan, M. Monteiro, S. Bannur, M. P. Lungren, A. Nori, B. Glocker, *et al.*, "Active label cleaning for improved dataset quality under resource constraints," *Nature communications*, vol. 13, no. 1, pp. 1–11, 2022.

[16] Y. Gal, R. Islam, and Z. Ghahramani, "Deep bayesian active learning with image data," in *International Conference on Machine Learning*, pp. 1183–1192, PMLR, 2017.

[17] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[18] W. H. Beluch, T. Genewein, A. Nürnberger, and J. M. Köhler, "The power of ensembles for active learning in image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9368–9377, 2018.

[19] J. Choi, I. Elezi, H.-J. Lee, C. Farabet, and J. M. Alvarez, "Active learning for deep object detection via probabilistic modeling," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10264–10273, 2021.

[20] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[22] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

[23] D. Yoo and I. S. Kweon, "Learning loss for active learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 93–102, 2019.

[24] S. Wang, Y. Li, K. Ma, R. Ma, H. Guan, and Y. Zheng, "Dual adversarial network for deep active learning," in *European Conference on Computer Vision*, pp. 680–696, Springer, 2020.

[25] S. Sinha, S. Ebrahimi, and T. Darrell, "Variational adversarial active learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5972–5981, 2019.

[26] D. Wang and Y. Shang, "A new active labeling method for deep learning," in *2014 International Joint Conference on Neural Networks (IJCNN)*, pp. 112–119, 2014.

[27] K. Kim, D. Park, K. I. Kim, and S. Y. Chun, "Task-aware variational adversarial active learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8166–8175, 2021.

[28] Y. Saquil, K. I. Kim, and P. Hall, "Ranking cgans: Subjective control over semantic image attributes," *arXiv preprint arXiv:1804.04082*, 2018.

[29] G. Citovsky, G. DeSalvo, C. Gentile, L. Karydas, A. Rajagopalan, A. Rostamizadeh, and S. Kumar, "Batch active learning at scale," *Advances in Neural Information Processing Systems*, vol. 34, pp. 11933–11944, 2021.

[30] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal, "Deep batch active learning by diverse, uncertain gradient lower bounds," *arXiv preprint arXiv:1906.03671*, 2019.

[31] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, (USA), p. 1027–1035, Society for Industrial and Applied Mathematics, 2007.

[32] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[34] J. Z. Bengar, J. van de Weijer, L. L. Fuentes, and B. Raducanu, "Class-balanced active learning for image classification," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1536–1545, 2022.

[35] E. Haussmann, M. Fenzi, K. Chitta, J. Ivanecky, H. Xu, D. Roy, A. Mittel, N. Koumchatzky, C. Farabet, and J. M. Alvarez, "Scalable active learning for object detection," in *2020 IEEE intelligent vehicles symposium (iv)*, pp. 1430–1435, IEEE, 2020.

[36] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018.

[37] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[38] J.-J. Zhu and J. Bento, "Generative adversarial active learning," *arXiv preprint arXiv:1702.07956*, 2017.

[39] C. Mayer and R. Timofte, "Adversarial sampling for active learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3071–3079, 2020.

[40] M. Huijser and J. C. van Gemert, "Active decision boundary annotation with deep generative models," in *Proceedings of the IEEE international conference on computer vision*, pp. 5286–5295, 2017.

[41] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[42] D. Mahapatra, B. Bozorgtabar, J.-P. Thiran, and M. Reyes, "Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 580–588, Springer, 2018.

[43] Q. Kong, B. Tong, M. Klinkigt, Y. Watanabe, N. Akira, and T. Murakami, "Active generative adversarial network for image classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 4090–4097, 2019.

[44] H. A. Duran, "Wasserstein generative adversarial active learning for anomaly detection with gradient penalty," Master's thesis, Middle East Technical University, 2021.

[45] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," *Advances in neural information processing systems*, vol. 30, 2017.

[46] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), vol. 70 of *Proceedings of Machine Learning Research*, pp. 214–223, PMLR, 06–11 Aug 2017.

[47] Y. Liu, Z. Li, C. Zhou, Y. Jiang, J. Sun, M. Wang, and X. He, "Generative adversarial active learning for unsupervised outlier detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 8, pp. 1517–1528, 2019.

[48] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[49] L. Nanni, M. Paci, S. Brahnam, and A. Lumini, "Comparison of different image data augmentation approaches," *Journal of Imaging*, vol. 7, no. 12, p. 254, 2021.

[50] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[51] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[52] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

[53] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, 2020.

[54] C. Bowles, L. Chen, R. Guerrero, P. Bentley, R. Gunn, A. Hammers, D. A. Dickie, M. V. Hernández, J. Wardlaw, and D. Rueckert, "Gan augmentation: Augmenting training data using generative adversarial networks," *arXiv preprint arXiv:1810.10863*, 2018.

[55] Z. Zhao, Z. Zhang, T. Chen, S. Singh, and H. Zhang, "Image augmentations for gan training," *arXiv preprint arXiv:2006.02595*, 2020.

[56] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[57] 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," in *Proceedings of the 9th international conference on information management and engineering*, pp. 139–144, 2017.

[58] 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.

[59] 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.

[60] G. E. Tontini, A. Rimondi, M. Vernero, H. Neumann, M. Vecchi, C. Bezzio, and F. Cavallaro, "Artificial intelligence in gastrointestinal endoscopy for inflammatory bowel disease: a systematic review and new horizons," *Therapeutic advances in gastroenterology*, vol. 14, p. 17562848211017730, 2021.

[61] 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, 2019.

[62] S. P. Travis, D. Schnell, P. Krzeski, M. T. Abreu, D. G. Altman, J.-F. Colombel, B. G. Feagan, S. B. Hanauer, M. Lémann, G. R. Lichtenstein, *et al.*, "Developing an instrument to assess the endoscopic severity of ulcerative colitis: the ulcerative colitis endoscopic index of severity (uceis)," *Gut*, vol. 61, no. 4, pp. 535–542, 2012.

[63] D. o. Rachmilewitz, "Coated mesalazine (5-aminosalicylic acid) versus sulphasalazine in the treatment of active ulcerative colitis: a randomised trial.," *British Medical Journal*, vol. 298, no. 6666, pp. 82–86, 1989.

[64] T. Lobatón, T. Bessissow, G. De Hertogh, B. Lemmens, C. Maedler, G. Van Assche, S. Vermeire, R. Bisschops, P. Rutgeerts, A. Bitton, *et al.*, "The modified mayo endoscopic score (mmes): a new index for the assessment of extension and severity of endoscopic activity in ulcerative colitis patients," *Journal of Crohn's and Colitis*, vol. 9, no. 10, pp. 846–852, 2015.

[65] M. Naganuma, H. Ichikawa, N. Inoue, T. Kobayashi, S. Okamoto, T. Hisamatsu, T. Kanai, H. Ogata, Y. Iwao, and T. Hibi, "Novel endoscopic activity index is useful for choosing treatment in severe active ulcerative colitis patients," *Journal of gastroenterology*, vol. 45, no. 9, pp. 936–943, 2010.

[66] 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.

[67] S. Mazzuoli, F. W. Guglielmi, E. Antonelli, M. Salemme, G. Bassotti, and V. Villanacci, "Definition and evaluation of mucosal healing in clinical practice," *Digestive and Liver Disease*, vol. 45, no. 12, pp. 969–977, 2013.

[68] H. T. Kani, I. Ergenc, G. Polat, Y. Ozen Alahdab, A. Temizel, and O. Atug, "P099 Evaluation of endoscopic mayo score with an artificial intelligence algorithm," *Journal of Crohn's and Colitis*, vol. 15, pp. S195–S196, 05 2021.

[69] G. Polat, E. Işık Polat, K. Kayabay, and A. Temizel, "Polyp detection in colonoscopy images using deep learning and bootstrap aggregation," IEEE International Symposium on Biomedical Imaging, Endoscopy Detection and Segmentation Workshop (EndoCV2021), 2021.

[70] 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.

[71] 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.

[72] G. Polat, Ergenç, H. T. Kani, Y. Özen Alahdab, Atuğ, and A. Temizel, "Class distance weighted cross-entropy loss for ulcerative colitis severity estimation," 26th UK Conference on Medical Image Understanding and Analysis, 2022.

[73] G. Polat, D. Sen, A. Inci, and A. Temizel, "Endoscopic artefact detection with ensemble of deep neural networks and false positive elimination.," in *EndoCV@ ISBI*, pp. 8–12, 2020.

[74] H. Nosato, H. Sakanashi, E. Takahashi, and M. Murakawa, "An objective evaluation method of ulcerative colitis with optical colonoscopy images based on higher order local auto-correlation features," in *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pp. 89–92, IEEE, 2014.

[75] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.

[76] 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* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), pp. 8024–8035, Curran Associates, Inc., 2019.

[77] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[78] L. Biewald, "Experiment tracking with weights and biases," 2020. Software available from wandb.com.

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

[80] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.