Sena B. YENGEÇ TAŞDEMİR

COMPUTER AIDED DETECTION OF CANCER USING HISTOPATHOLOGY IMAGES

AGU 2023

COMPUTER AIDED DETECTION OF CANCER USING HISTOPATHOLOGY IMAGES

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE OF ABDULLAH GUL UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

By

Sena Büşra YENGEÇ TAŞDEMİR

April 2023

COMPUTER AIDED DETECTION OF CANCER USING HISTOPATHOLOGY IMAGES

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE OF ABDULLAH GUL UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

By

Sena Büşra YENGEÇ TAŞDEMİR

April 2023

SCIENTIFIC ETHICS COMPLIANCE

I hereby declare that all information in this document has been obtained 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 materials and results that are not original to this work.

Name-Surname: Sena Büşra YENGEÇ TAŞDEMİR

Signature :

REGULATORY COMPLIANCE

Ph.D. thesis titled Computer Aided Detection of Cancer using Histopathology Images has been prepared in accordance with the Thesis Writing Guidelines of the Abdullah Gül University, Graduate School of Engineering & Science.

Prepared By	Advisor
Sena Büşra YENGEÇ TAŞDEMİR	Prof. Dr. Bülent YILMAZ
Signature	Signature

Head of the Electrical and Computer Engineering Program

Assoc. Prof. Zafer AYDIN

Signature

ACCEPTANCE AND APPROVAL

Ph.D. thesis titled Computer Aided Detection of Cancer using Histopathology Images and prepared by Sena Büşra YENGEÇ TAŞDEMİR has been accepted by the jury in the Electrical and Computer Engineering Graduate Program at Abdullah Gül University, Graduate School of Engineering & Science.

10 /April / 2023

JURY:

Advisor : Prof. Dr. Bülent YILMAZ

Co-Advisor: Assoc. Prof. Zafer AYDIN

Member : Prof. Dr. İsmail AVCIBAŞ

Member : Assoc. Prof. Kutay İÇÖZ

Member : Assist. Prof. Bekir Hakan AKSEBZECİ

Member : Assist. Prof. Özkan Ufuk NALBANTOĞLU

APPROVAL:

The acceptance of this Ph.D. thesis has been approved by the decision of the Abdullah Gül University, Graduate School of Engineering & Science, Executive Board dated ….. /….. / ……….. and numbered …………..……. .

……….. /……….. / ………..

(Date)

Graduate School Dean

Prof. Dr. İrfan ALAN

ABSTRACT

Computer Aided Detection of Cancer using Histopathology Images

Sena Büşra YENGEÇ TAŞDEMİR

Ph.D. in Electric and Computer Engineering Department

Advisor: Prof. Dr. Bülent YILMAZ

April 2023

Detecting colon adenomatous polyps early is crucial for reducing colon cancer risk. This thesis investigated various deep learning approaches for computer-aided diagnosis of colon polyps on histopathology images using deep learning. The thesis addressed key challenges in polyp classification, including differentiating adenomatous polyps from non-adenomatous tissues and multi-class classification of polyp types. Initially, a histopathology image dataset is collected and refined from Kayseri City Hospital. The first study used stain normalization algorithms and an ensemble framework for binary classification, achieving 95% accuracy on the custom dataset and 91.1% and 90% on UnitoPatho and EBHI datasets, respectively. The second study implemented a tailored version of the supervised contrastive learning model for multi-class classification, outperforming state-of-the-art deep learning models with accuracies of 87.1% on custom dataset and 70.3% on UnitoPatho dataset. The third study proposed a self-supervised contrastive learning approach for utilizing all data in cases of limited labeled images. This approach achieved better performance than transfer learning with ImageNet pre-trained models. In conclusion, this PhD thesis investigated deep learning approaches for computer-aided diagnosis of colon polyps on histopathology images, demonstrating high accuracy in binary and multi-class classification, outperforming state-of-the-art models. These findings contribute to improving colon polyp classification accuracy and efficiency, ultimately facilitating the early detection and prevention of colon cancer.

Keywords: Polyp Classification, Histopathology, Transfer Learning, Ensemble Learning, ConvNeXt, Supervised Contrastive Learning

ÖZET

Histopatoloji Görüntülerinden Bilgisayar Destekli Kanser Tespiti

Sena Büşra YENGEÇ TAŞDEMİR

Elektrik ve Bilgisayar Mühendisliği Anabilim Dalı Doktora

Tez Yöneticisi: Prof. Dr. Bülent YILMAZ

Nisan 2023

Kolon adenomatoz poliplerinin erken tespiti kolon kanseri riskini azaltmak için önem arz etmektedir. Bu tez, histopatoloji görüntüleri üzerinde kolon poliplerinin bilgisayar destekli teşhisi için çeşitli derin öğrenme yaklaşımlarını araştırmıştır. Tez çalışmaları sırasında, polip sınıflandırmasındaki ana zorluklar ele alınarak, adenomatoz polipleri adenomatoz olmayan dokulardan ayrılması ve polip tiplerinin sınıflandırması gibi konulara odaklanılmıştır. Tezin ilk kısmında, Kayseri Şehir Hastanesi'nden histopatoloji görüntü veri seti toplanıp iyileştirilmiştir. İkinci kısım, birinci çalışma sırasında, boya normalizasyon algoritmaları ve topluluk çerçevesi kullanılarak ikili sınıflandırma görevi için toplanan veri setinde %95, UnitoPatho ve EBHI veri setlerinde sırasıyla %91.1 ve %90 doğruluk elde edilmiştir. İkinci çalışmada, çoklu sınıflandırma için özelleştirilmiş bir denetimli kontrastif öğrenme modeli uygulanmış, ve geliştirilen modelin performansı önceden eğitilmiş derin öğrenme modellerini geçmiştir. Toplanan veri setinde %87,1, UnitoPatho veri setinde %70,3 doğruluk elde edilmiştir. Üçüncü çalışma, sınırlı sayıda etiketli görüntü olduğu durumda tüm verilerin kullanılması için bir kendinden denetimli kontrastif öğrenme yaklaşımı önermiştir. Bu yaklaşım ImageNet ile önceden eğitilmiş modellerle karşılaştırıldığında daha iyi performans göstermiştir. Sonuç olarak, bu doktora tezi, histopatoloji görüntüleri üzerinde kolon poliplerinin bilgisayar destekli teşhisi için derin öğrenme yaklaşımlarını araştırmıştır ve ikili ve çoklu sınıflandırmada yüksek doğruluk göstererek, mevcut modelleri geride bırakmıştır. Bu bulgular, kolon polip sınıflandırma doğruluğunun ve etkinliğinin geliştirilmesine katkıda bulunmaktadır ve sonuç olarak kolon kanserinin erken teşhisini ve önlenmesini kolaylaştırmaktadır.

Anahtar kelimeler: Polip Sınıflandırması, Histopatoloji, Transfer Öğrenme, Toplu Öğrenme, ConvNeXt, Denetimli Karşılaştırmalı Öğrenme

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Bülent YILMAZ, for his invaluable guidance, expertise, and unwavering support during my doctoral studies. His mentorship has been instrumental in shaping both my research and personal growth. Prof. YILMAZ's constructive criticism and guidance have been a constant source of inspiration and I feel privileged to have had him as my advisor.

I would like to express my appreciation to Dr. Zafer AYDIN for his valuable guidance and advice during both my doctoral and master's studies. His insights and expertise have been invaluable to me, and I am grateful for the support he has provided throughout my academic journey.

I extend my heartfelt gratitude to my dear spouse, Dr. Kasım TAŞDEMIR, for being a constant source of support in all aspects of my life. Throughout my doctoral studies, he has provided me with strength and encouragement, and I am immensely grateful for his presence in my life.

I am grateful for the unwavering presence of my feline friends, Yoğurt, Kekik and Puding, who brought joy and comfort to my life throughout my doctoral studies.

I would like to express my sincere gratitude to my dear friends Zeynep Şenel and Merve Balki Taş for their support and for being a source of joy and companionship throughout this journey.

Last but not least, I would like to extend my profound gratitude to my beloved father, mother, and sister for their unrelenting support and encouragement throughout my academic journey. Their love and guidance have been the driving force behind my achievements, and I am deeply appreciative of everything they have done for me. Without their constant support, none of this would have been possible.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
BiT	Big Transfer
CRC	Colorectal Cancer
CNN	Convolutional Neural Networks
CLR	Contrastive Learning
H&E Stain	Hematoxylin and Eosin Stain
HP	Hyperplastic Polyp
ML	Machine Learning
OD	Optical Density
ResNet	Residual Networks
ROC	Receiver Operating Characteristic
Sup-Con	Supervised Contrastive
SSL	Self-Supervised Contrastive Learning
SVD	Singular Value Decomposition
TBA	Tubular Adenoma
TVA	Tubulovillous Adenoma
WSI	Whole Slide Images

To my beloved family.

Chapter 1

Introduction

General

The large intestine includes cecum, colon, rectum, and anal canal. Most of the colorectal cancer cases begin as a polyp in the rectum or colon and removal of the polyps may prevent colorectal cancer [1]. In 2020, 1.93 million new colorectal cancer (CRC) cases were diagnosed worldwide, and 940,000 people died because of colorectal cancer [2]. In addition, it is estimated that during 2040, new CRC cases will reach 3.2 million. Early detection, and removal of cancerous tissue on the colon is critical to lower the mortality rates [3].

The gold standard for screening CRC cancer is colonoscopy followed by pathology examination since it allows the clinician to see the entire lining of the colon and the cellular structures in the tissue samples [4]. During a colonoscopy procedure, a specialist may examine the colon using a flexible tube with a camera and light source at the end. The found polyps are removed and tissue samples are sent for pathology examination for long-term follow-up and management of treatment. Then, an expert pathologist decides on the type of polyp after a microscopic examination [5].

Not all types of polyps may grow into colorectal cancer [4]. There are two main types of polyps, adenomatous polyps, and non-adenomatous polyps. Types of adenomatous polyps are villous, tubular, tubulovillous, and non-adenomatous polyps are hyperplastic and inflammatory [5]. Adenomatous polyps gradually show dysplastic changes, and high-grade dysplastic changes that occur over time become malignant. Therefore, early detection and removal of adenomatous polyps and planning of long-term treatment are important.

There is a growing demand for cancer screening programs, over the past decade pathological colon biopsy slide volumes are doubled [6]. Since the number of biopsies increases, the workload of the pathologist increases. Hence, detection of the early-stage disease becomes increasingly difficult. As a result, the computer-aided diagnosis systems can be used to ease this labor-intensive work and minimize the mistakes of the traditional approaches.

The main challenge in the clinical workflow of pathological polyp classification is differentiating adenomatous polyps from non-adenomatous tissues. In addition, a preliminary classification of the polyp types can facilitate the work of the pathologist. Therefore, this thesis explores various methodologies to distinguish adenomatous polyps from non-adenomatous tissues. Moreover, in this thesis, systems for polyp-type classification of tubular, tubulovillous, villous and hyperplastic polyps are proposed. The proposed computer-assisted diagnostic system is hoped to facilitate this labor-intensive decision-making process.

1.1 Objective and Scopes

Using histopathology images of a polyp tissue is the common way to distinguish adenomatous polyps from non-adenomatous tissues including hyperplastic polyps, inflammation, and normal tissue. However, the success of the histology examination largely depends on the experience of the expert pathologist. This makes CAD systems assistance helpful and necessary in this task.

In addition, the demand for cancer screening programs is considerably high and this causes an increase in the number of pathology images, which increases the workload of the pathologists. Therefore, it would be beneficial to use an automated system that distinguishes histopathology images to make it easier to make classification on pathology images.

In the literature, there are numerous studies on polyp classification using pathology images for computer-aided diagnosis systems. However, there is no complete solution yet. Indeed, most researchers propose methods that create tailored frameworks and demonstrate their performance in their datasets. In this thesis, various deep learning methods are explored, and frameworks are developed for the automatic classification of adenomatous polyps and the classification of polyp types and the performance is demonstrated for various datasets.

Two publicly available datasets were used to evaluate the generalization ability of the frameworks. More importantly, a new curated dataset is created.

This thesis covers four main parts. The first part of the thesis is composed of a collection of datasets and improvements on the datasets. The dataset is explored with baseline classifiers for different classification tasks of polyp classification on colon histopathology images. In the clinical workflow of polyp classification, a key diagnostic challenge is the differentiation of adenomatous polyps from non-adenomatous tissues. Therefore, in the second part of the thesis, we explored different approaches to solve the binary classification of adenomatous polyps and non-adenomatous tissues. Furthermore, adenomatous polyps have the potential to grow into cancer. In order to ease the classification of the adenomatous polyps and hyperplastic polyps, in the third study we proposed a method to make multi-class classification polyp types on histopathology images. Additionally, in supervised learning on medical image analysis, the need for a large number of medical images in a supervised training setting also brings the necessity of having labels. The problem here arises because medical image analysis and labelling are very time-consuming and laborious. Moreover, in most cases, the labelling should be done on-site due to the confidentiality of the patient information. In the fourth study, we tried to answer the research question of "How we can use all of the data when plenty of unlabeled images and a limited number of labelled images are obtained". To solve this problem, we used self-supervised contrastive learning for pre-training of the model on unlabeled data as a final work of the thesis.

1.2 Literature Overview

With the advancement of deep learning techniques, there has been a significant rise in research focused on classification in medical imaging. Specifically, for colon histopathology image classification, there exist several comprehensive studies addressing various problems such as polyp classification, gland classification, and colorectal cancer classification.

In their work Korbar et al. employed ResNet architecture variants to classify colon polyp types using histopathology images, and an overall accuracy of 93% was attained [7]. Moreover, in Korbar et al. visualized attention map on whole slide images (WSI) by highlighting areas on the images to provide more information about the classification.

Song et al. studied the importance of different patch sizes for polyp classification by comparing their framework results with those of pathologists [8]. Wei et al. achieved an AUC score of 88.2% by employing curriculum learning to discriminate sessile serrated polyps from hyperplastic polyps [9]. Building on their previous work, Wei et al. proposed Confidence-Aware Label smoothing to distinguish hyperplastic polyps from sessile serrated adenomas [10]. Nasir Moin et al. designed an AI augmented tool to assist pathologists by providing diagnostic information. This tool utilized ResNet-18 architecture to classify polyp types [11]. Zhou et al. designed a methodology to classify WSIs by localizing the region by employing the global labels of the WSIs [12]. Gupta et al. tailored Inception-ResNet-v2 model to classify and localize abnormal tissues in WSIs [13].

Even if the histopathology images are captured under the same conditions, variations in the images can occur due to the histological staining process of the tissue. To address this issue, researchers are exploring stain normalization and histogram equalization techniques. Perlo et al. proposed a framework for polyp classification and dysplasia grading, and compared its performance for Macenko Stain Normalized, RGB, and grayscale histopathology images [14]. Sarwinda et al. employed Contrast Limited Adaptive Histogram Equalization (CLAHE) to classify colorectal WSIs as malignant or benign using ResNet-50 architecture [15].

Recently, Byeon employed DenseNet-161 and EfficientNet-B7 to classify cancerous and non-specific tissues of adenomatous and hyperplastic polyps [16]. In their work, Bilal et al. classified WSIs as neoplastic or normal by using weakly supervised deep learning [6]. The model achieved an overall area under the receiver operating characteristic curve (AUROC) score of 97.46% is achieved on The Cancer Genome Atlas (TCGA) database. Ho et al. proposed a method to classify histopathology images as high risk and low risk and achieved an AUC of 91.7% [17]. In their work Yildirim et al. designed a CNN based framework to classify colon cancer on WSIs [18].

Chapter 2

2 Background

2.1 Anatomy of the Colon

The large intestine composed of cecum, colon, rectum, and anal canal. As it is illustrated in the Figure 2.1.1 colon can be anatomically divided into four sections: ascending, transverse, descending, and sigmoid [19]. The ascending colon is the beginning of the colon which extends from cecum. This is followed by transverse colon and descending colon. The last part of the large intestines is sigmoid colon which is followed by rectum and anus.

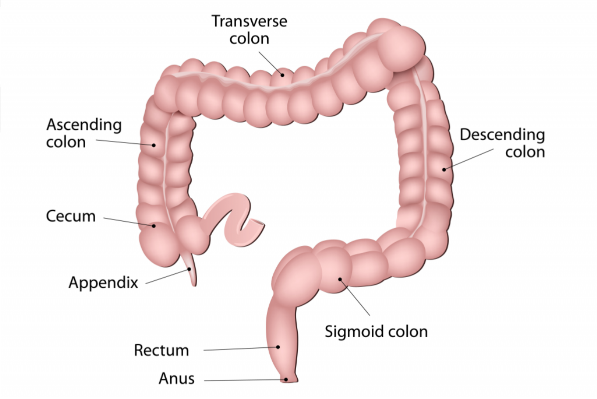


Figure 2.1.1 Anatomy of the Large Intestine [19]

2.2 Colon Polyp Types

A colon polyp is formed by the growth of tissue in the lining of the colon (Figure 2.2.1) Colon polyps can be classified into two main classes according to their behavior, which are benign or pre-malignant. Pre-malignant polyps are villous, tubulovillous, and tubular adenoma, while benign polyps are hyperplastic or inflammatory polyps. If left untreated, pre-malignant polyps have the potential to progress into colon adenocarcinoma [20].

Generally, most of the colon polyps do not have symptoms. Therefore, early diagnosis is challenging. The gold standard for the colon screening is colonoscopy procedure. When a polyp is found during the colonoscopy, it is removed. A specimen of the removed tissue is sent to the pathology department to diagnose the polyp type for long-term follow-up and management of treatment. Pathology examination is the only way to distinguish polyp type.

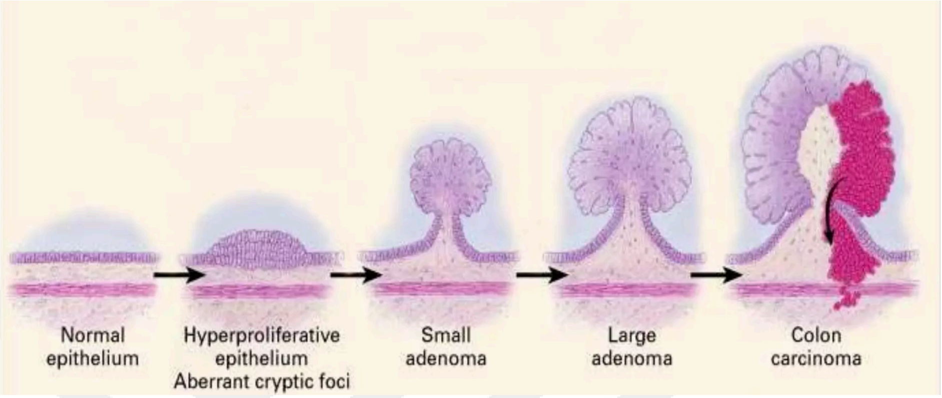
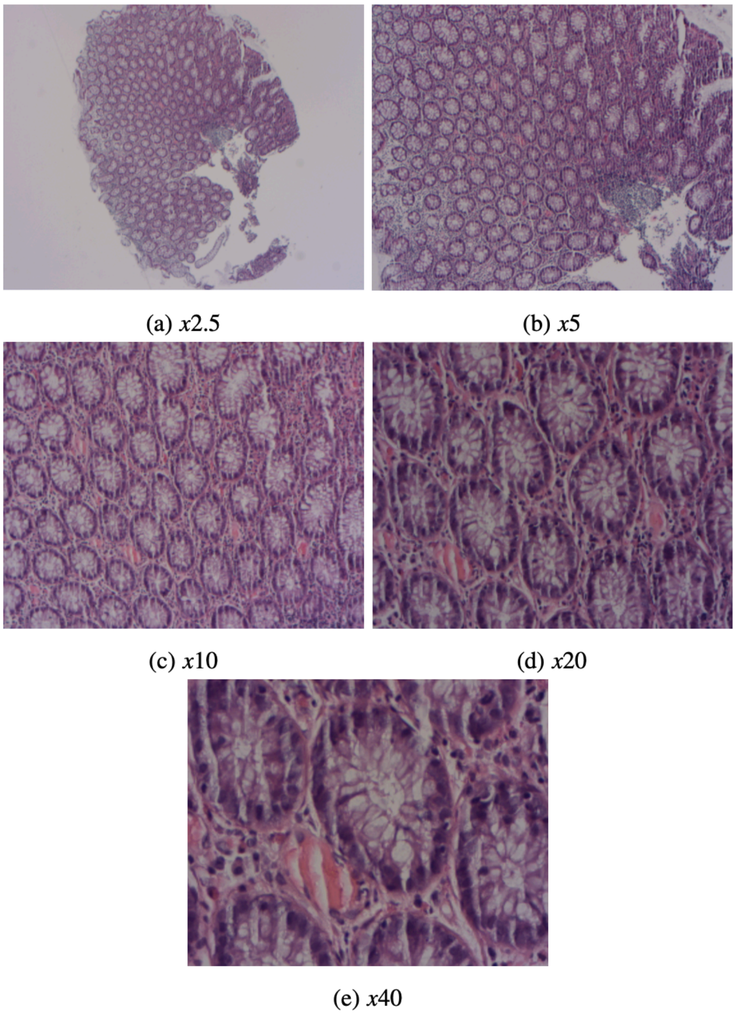


Figure 2.2.1 Colon polyp types and their life stages

2.3 Histopathology Images

The specimens are examined under a microscope to diagnose the polyp type by a pathologist. Usually, a pathologist examines a tissue under different magnification levels. These magnification levels might change from one device to another, but generally the used levels are x2.5, x5, x10, 20 and x40. As it is illustrated in Figure, 2.3.1, as the magnification level increases the microscope zooms into the sample.



(a) *x*2.5

(b) *x*5

(c) *x*10

(d) *x*20

(e) *x*40

Figure 2.3.1 Different magnification levels of an adenomatous histopathology image

2.3.1 H&E Staining

To prepare a sample for microscopic examination a pathologist uses tissue staining techniques. The staining process enables the tissue sample to show more fine details such as distribution of cells clearly and provides an overview of the structure. The gold standard for staining is hematoxylin and eosin (H&E) staining [21].

H&E staining is composed of hematoxylin and eosin stains. Eosin stain gives a pink color to the sample while hematoxylin stain gives a blue shade. The combination of hematoxylin and eosin gives different shades and hues to the overall cell structure (Figure 2.3.2). These general coloration patterns, which are formed by staining, highlight the general structure and arrangement of the cell, and provides fine information about the polyp type to the experts.

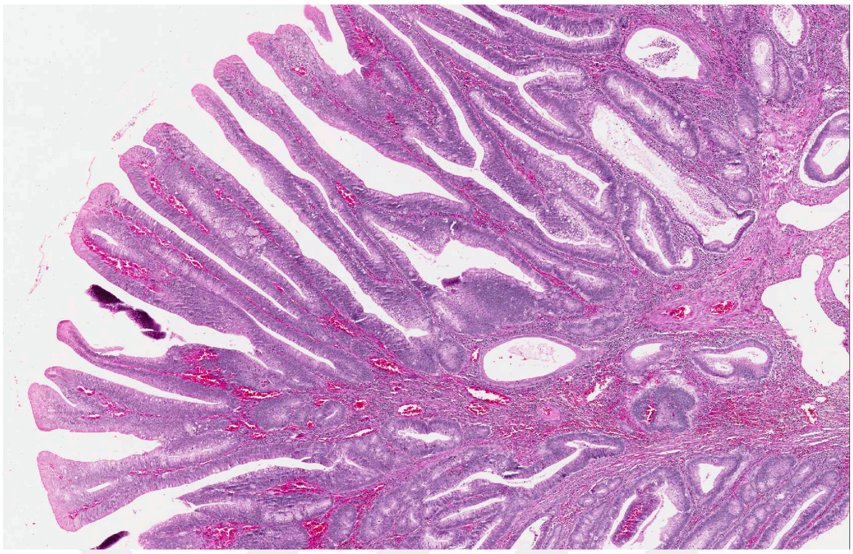


Figure 2.3.2 H&E stain instance of a villous adenoma [21]

2.3.2 Structure of Polyp Types for H&E Stained Images

The cellular architecture of tissue changes under the microscope depending on the type of polyp. A normal tissue sample of custom collected dataset shows an orderly and general structure (Figure 2.3.3), while a villous polyp exhibits a finger-like epithelial outgrowth structure (Figure 2.3.4) [22]. For tubular adenoma, cells display a more concentrated appearance (Figure 2.3.5). A tubulovillous polyp, has both the features of tubular and villous adenoma (Figure 2.3.6). If the villous appearance density of the cell structure is above 80%, it is classified as tubulovillous polyp [23]. Moreover, for hyperplastic polyp cells are mostly filled with more mucus (Figure 2.3.7). Considering the differentiation of cellular structure, it can be inferred that pattern differentiation of cellular structure can be used in computer-aided diagnosis of histopathology images.

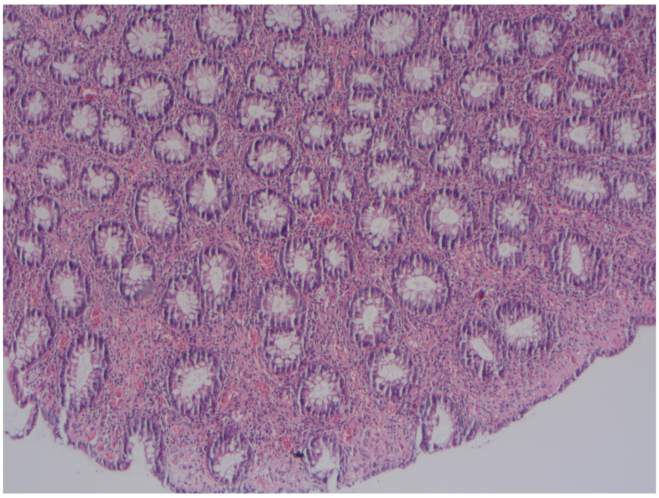


Figure 2.3.3 Normal Tissue Sample under microscopy

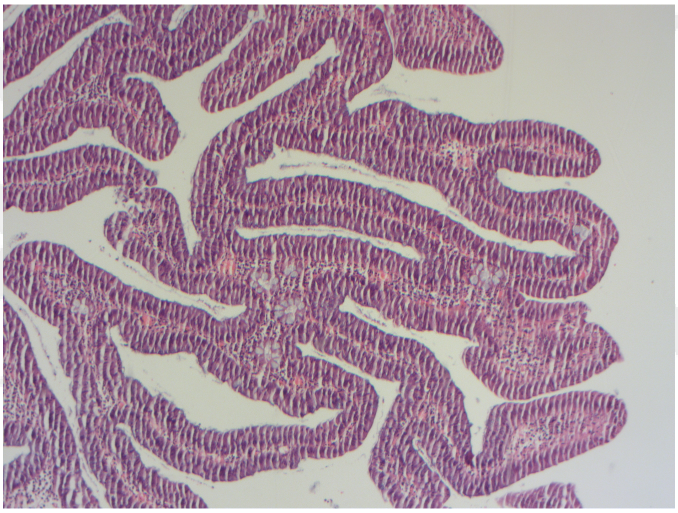


Figure 2.3.4 Villous Adenoma under microscopy

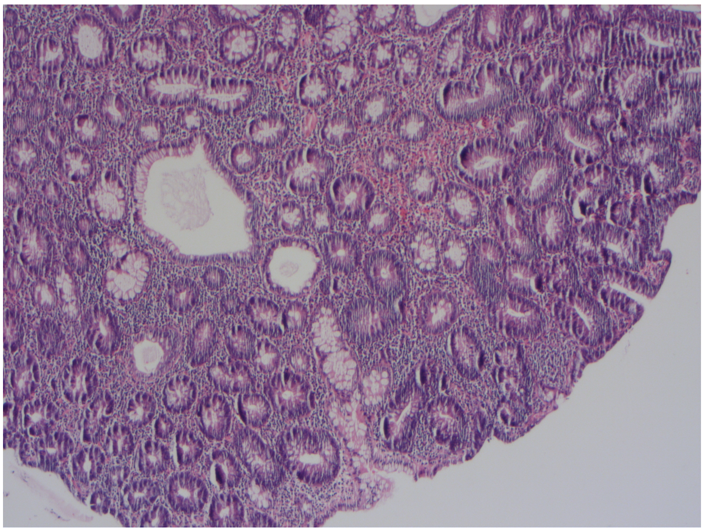


Figure 2.3.5 Tubular Adenoma under microscopy

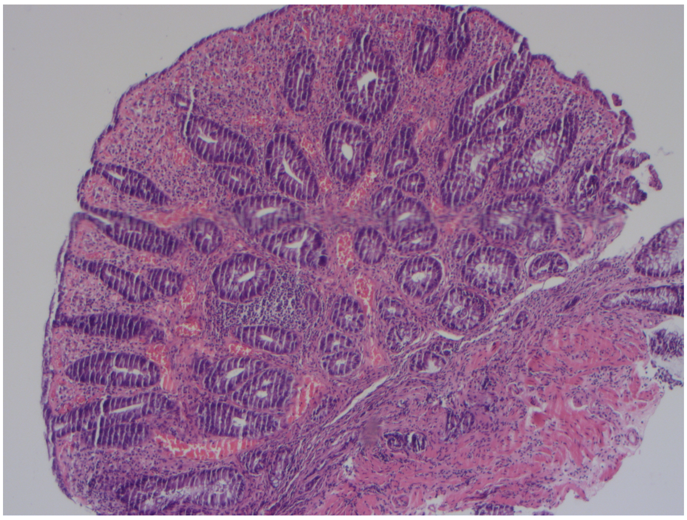


Figure 2.3.6 Tubulovillous Adenoma under microscopy

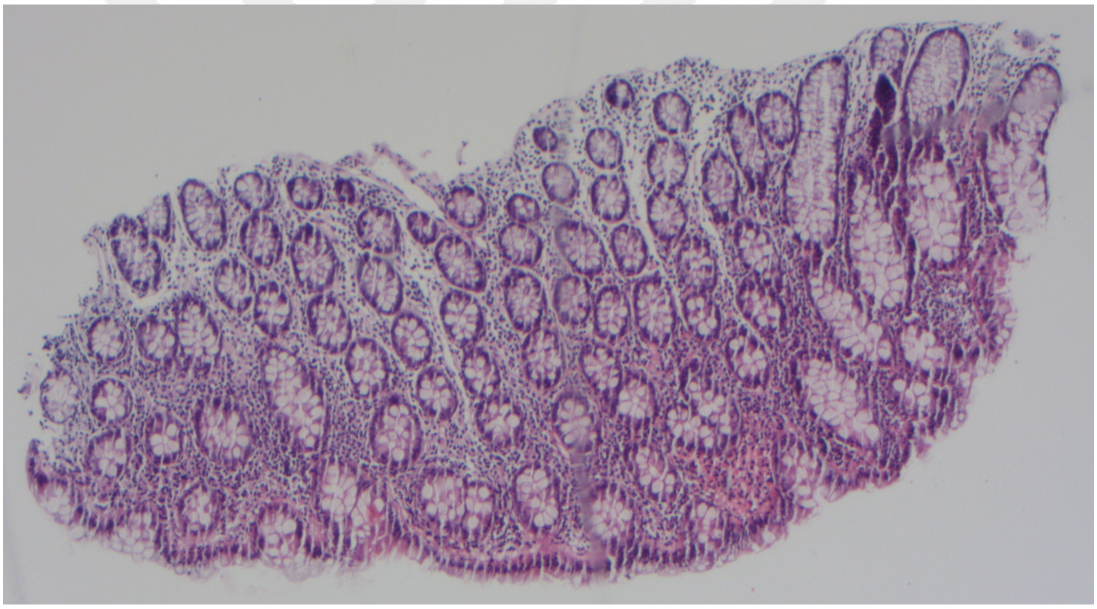


Figure 2.3.7 Hyperplastic Polyp under microscopy

2.4 Datasets

In this thesis to evaluate the generalization ability of the frameworks that are developed for the histopathology images provided by Kayseri City Hospital, two publicly available datasets are employed. The first dataset, UniToPatho contains 9536 H&E stained patches extracted [24]. The patches are from 292 whole-slide images which has a magnification level of 20. The patches are in size of 1812×1812 pixels. Moreover, the classes of the patches are normal tissue, hyperplastic polyp, tubular adenoma, and tubulovillous adenoma.

The second dataset is EBHI dataset which contains 5532 WSIs [25]. Furthermore, this WSIs belongs to following magnification levels: ×40, ×100, ×200 and ×400. The images on this dataset have a size of 2048×1536 pixels. This dataset contains two major categories, benign and malignant with the following classes: Normal, polyp, low grade adenoma, high grade adenoma, and adenocarcinoma.

2.5 Stain Normalization for WSIs

Deep Convolutional Neural Network algorithms have a great capacity to fit a dataset with high precision. However, this precision challenges the model to generalize for the unseen data. Moreover, if there is a domain shift in training and testing data, the model must be robust and reliable for real-world scenarios. In addition, domain shift is a common challenge faced by deep CNN structures, particularly in medical image analysis, where staining protocols and slide preparation can vary across different medical centers, leading to domain shifts [26]. Figure 2.5.1 illustrates a color intensity difference in histopathology images provided by Kayseri City Hospital. In order to deal with this issue, various stain normalization techniques are proposed by the researchers. The most common stain normalization algorithms in the literature are as follows: Reinhard, Macenko, Vahadane, Stain-GAN, and Stain-Net.

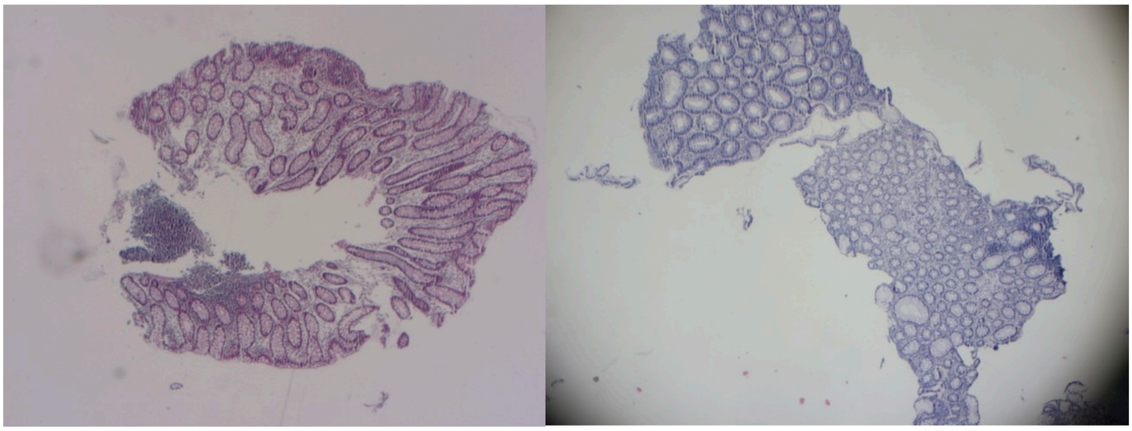


Figure 2.5.1 Color intensity difference in WSIs

2.5.1 Reinhard Color Normalization

Reinhard color normalization is a color matching method, which originally proposed for the real-world images [27]. However, it is widely used by stain normalization researchers. This methodology first changes color space of RGB to Ruderman perception-based color space lαβ. Transformation of a RBG to lαβ is composed of three stages. First, RGB image is transformed into LMS in two stages by using the Equations 1 and 2. Then LMS space is transformed to lαβ by employing Equations 3 and 4. After this color transition, mean and standard deviation of each channel are calculated to observe the distribution of data points. The mean value is subtracted from the data points of each channel as shown in the Equation 5 then each data point value is scaled with standard deviation Equation 6. Finally, lαβ image is transformed back to RGB image (Equation 7) [27].

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.5141 & 0.3239 & 0.1604 \\ 0.2651 & 0.6702 & 0.0641 \\ 0.0241 & 0.1228 & 0.8444 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{1}$$

$$\begin{bmatrix} L \\ M \\ S \end{bmatrix} = \begin{bmatrix} 0.3847 & 0.6890 & -0.0787 \\ -0.2298 & 1.1834 & 0.0464 \\ 0.0000 & 0.0000 & 1.0000 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{2}$$

$$\begin{gathered} \boldsymbol{L} = \log L \\ \boldsymbol{M} = \log M \\ \boldsymbol{S} = \log S \end{gathered} \tag{3}$$

$$\begin{bmatrix} l \\ \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{3}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{6}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -2 \\ 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{L} \\ \boldsymbol{M} \\ \boldsymbol{S} \end{bmatrix} \tag{4}$$

$$\begin{gathered} l^* = l - \langle l \rangle \\ \alpha^* = \alpha - \langle \alpha \rangle \\ \beta^* = \beta - \langle \beta \rangle \end{gathered} \tag{5}$$

$$l' = \frac{\sigma_t^l}{\sigma_s^l} l^*$$

$$\alpha' = \frac{\sigma_t^\alpha}{\sigma_s^\alpha} \alpha^* \qquad (6)$$

$$\beta' = \frac{\sigma_t^\beta}{\sigma_s^\beta} \beta^*$$

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 4.4679 & -3.5873 & 0.1193 \\ -1.2186 & 2.3809 & -0.1624 \\ 0.0497 & -0.2439 & 1.2045 \end{bmatrix} \begin{bmatrix} L \\ M \\ S \end{bmatrix} \quad (7)$$

2.5.2 Macenko Stain Normalization

By emphasizing that stain vectors of the pathology images are not linear in RGB channel, Macenko normalization first converts RGB channel into Optical Density (OD) [28]. The OD values of each pixel is calculated by using RGB color vectors with Equation 8 in which “I” represent image. Before processing further, low OD values are thresholded to ensure the stability of the algorithm. After thresholding, singular value decomposition (SVD) is calculated on OD vectors to have stain vectors (V) and saturation of each stain (S) by using the Equation 9. Two vectors that are corresponding to two largest singular values of SVD decomposition of previous step will be used to form a plane. Moreover, OD values of all the pixels is projected to this plane and values are normalized to unit length.

$$OD = -\log_{10} I \quad (8)$$

$$S = V^{-1} OD \qquad (9)$$

2.5.3 Vahadane Stain Normalization

Vahadane normalization is a structure-preserving color normalization technique which uses a source and a target image [29]. A source image is the input image which will be normalized, and a target image is the starting point which has the desired color intensity. In this framework color intensity of the source image is matched to the target image by following the steps: First the pixel values of an RGB image are converted into OD values by employing Equation 10 in which I represents image and I_0 represents illumination density. Thus, observation matrix of V is calculated. Then, stain density map matrix H and Stain color appearance matrix W is calculated with observation matrix and Equation 11 and 12. All matrices of V, S, and H is calculated for both source and target images. Moreover, by using Equation 13 and 14 normalized stain density map matrix (H_s^{norm}), normalized observation matrix of V_s^{norm} of the source image is calculated. Finally, normalized source image is calculated with Equation 15.

$$V = \log\frac{I_0}{I} \qquad (10)$$

$$I = I_0 \exp(-WH) \qquad (11)$$

$$V = WH \qquad (12)$$

$$H_s^{norm}(j,:) = \frac{H_s(j,:)}{H_s^{RM}(j,:)} H_s^{RM}(j,:), \quad j = 1, \dots r. \quad (13)$$

$$V_s^{norm} = W_t H_s^{norm} \qquad (14)$$

$$I_s^{norm} = I_0 \exp(-V_s^{norm}) \quad (15)$$

2.5.4 Stain-GAN Stain Normalization

The Stain-GAN stain normalization method normalizes an image without needing to have a target image [30]. This method learns whole distribution by employing a model which is based cycle-consistent generative adversarial networks. In order to learn the representation from the whole data distribution, this model employs two pairs of generators and discriminator for the source domain and target domain. Generator of source domain tries to match the images with target domain, while discriminator tries to distinguish whether the images are fake or not. Furthermore, generator and discriminator of the target domain repeat the same operation in their respective fields. The models are trained to match the objective function which sums the adversarial loss with cycle consistency loss as illustrated in Equation 16. The purpose of the adversarial loss is to match the distribution of the generated images with that of the target domain (Forward Cycle) and the distribution of the generated target domain matches back to the source domain (Backward Cycle). Cycle consistency loss controls that the generated images preserve their original structure.

$$\mathcal{L} = \mathcal{L}_{Adv} + \lambda\ \mathcal{L}_{Cycle} \qquad (16)$$

2.5.5 Stain-Net Stain Normalization

The Stain-Net employs a distillation learning scheme by using Stain-GAN [31]. For this distillation scheme, Stain-Gan is used as teacher network and Stain-Net is used as student network. This scheme is composed of three stages; the images that is normalized by the generator of the Stain-Gan is used as ground truth for the Stain-Net. The Stain-Net learns the mapping relationship of Stain-Gan by training on the normalized images with an optimizer of stochastic gradient descent and L1 loss. Then, it can transfer the source images to target domain. This method is 40 times faster than Stain-Gan methodology.

2.6 Machine Learning

Learning is a natural human behavior, while learning, people tend to learn through examples rather than formulating abstract rules or principles. Artificial Intelligence attempts to learn patterns and relationships by making observations and using examples to mimic the human learning path [32]. Moreover, a subset of the AI, Machine Learning tries to learn the patterns by using sophisticated algorithms with machine-accessible data [33]. In such a learning scheme, an algorithm is iteratively trained on a problem-specific data. This learning scheme allows machine to explore the hidden patterns of the data without needing to explicitly be programmed. Moreover, according to available data and problem there are four types of learning: Supervised Learning, Semi-Supervised Learning, Unsupervised Learning and Reinforcement Learning. During this thesis work, we will explore supervised learning and contrastive learning in unsupervised setting.

2.6.1 Classification Performance Assessment

The performance of the machine learning algorithms is evaluated by the classification evaluation metrics which are computed from the confusion matrix, which is illustrated in Figures 2.6.1 and 2.6.2. The rows of the confusion matrix show the number of samples for each class, while columns show the prediction of the model for each class. In this thesis following classification metrics are used: Overall Accuracy, Precision, Recall, F1-Measure. For binary classification task, precision and recall is calculated from the average of each class. For multi-class classification task, the precision, recall and consequently F1-measure is calculated with a weighted average of the classes.

The diagonal elements of the matrices in the Figure 2.6.1 and Figure 2.6.2 represents correct predictions, while other elements show the incorrect guesses. The false positive and false negative values for the binary classification is calculated with the number of samples that are classified as class A while the actual class is B or vice-verse, respectively. For the multi-class classification incorrect guesses are represented as Error BA, Error CA, Error AC and so on. Error BA represents the samples that belongs to class A, but incorrectly classified as class B [34].

	Actual Class A	Actual Class B
Predicted Class A	True Positive	False Positive
Predicted Class B	False Negative	True Negative

Figure 2.6.1 Confusion Matrix for Binary Classification

	Actual Class A	Actual Class B	Actual Class C
Predicted Class A	True Positive A	Error BA	Error CA
Predicted Class B	Error AB	True Positive B	Error CB
Predicted Class C	Error AC	Error BC	True Positive C

Figure 2.6.2 Confusion Matrix for Multi-class Classification

2.6.1.1 Overall Accuracy

The accuracy is one of the most widely used performance metric. It is calculated by using the Equation 17 for binary classification problem, Equation 18 for multi-class classification problem.

$$Binary\ Accuracy = \frac{TP+TN}{All\ Samples} \qquad (17)$$

$$Multi\ Class\ Accuracy = \frac{True\ Positive\ A+True\ Positive\ B+True\ Positive\ A}{All\ Samples} \qquad (18)$$

2.6.1.2 Precision

Precision value shows how well the classifier predicted samples of the class A with the Equation 19 for binary classification and Equation 20 for multiclass classification. Additionally, since the precision value is present for each class in the multiclass classification, weighted average of precision is calculated to have an overall value for the classifier by employing the equation 21, here Precision A is the precision value of class A and C_A is the number of samples in class A, and so on.

$$Precision = \frac{TP}{TP+FP} \qquad (19)$$

$$Precision\ Class\ A = \frac{TP_A}{TP_A+Error\ BA+Error\ CA} \qquad (20)$$

$$Weighted\ Average\ Precision = \frac{Precision_A C_A + Precision_B C_B + Precision_C C_C}{C_A + C_B + C_C} \qquad (21)$$

2.6.1.3 Recall

Recall value calculates ratio of true positive cases to true positive and false negative cases. For binary classification the ratio is calculated by using the Equation 22. For multiclass problem Recall value for each class is calculate with the Equation 23. Additionally, since the recall value is present for each class in the multiclass classification, weighted average of precision is calculated to have an overall value for the classifier by employing the Equation 24.

$$Recall = \frac{TP}{TP+FN} \qquad (22)$$

$$Recall\ Class\ A = \frac{TP_A}{TP_A+Error\ AB+Error\ AC} \quad (23)$$

$$Weighted\ Average\ Recall = \frac{Recall_A C_A + Recall_B C_B + Recall_C C_C}{C_A + C_B + C_C} \quad (24)$$

2.6.1.4 F1-Score

F1-Score is the harmonic mean of the precision and recall. By employing the Equation 25 the F1 score for binary classification is calculated. For multiclass problem F1 score for each class is calculate with the Equation 26. Additionally, weighted average of F1 score is calculated to have an overall value for the classifier by employing the Equation 27.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (25)$$

$$F1\ Class\ A = 2 \times \frac{Precision_A \times Recall_A}{Precision_A + Recall_A} \quad (26)$$

$$Weighted\ Average\ F1 = \frac{F1_A C_A + F1_B C_B + F1_C C_C}{C_A + C_B + C_C} \quad (27)$$

2.6.2 Supervised Learning

Supervised Learning is a strategy in which the label of the whole dataset is present and used during the training. In this thesis, we will explore the Deep CNN algorithms in a supervised setting.

2.6.2.1 Deep Convolutional Neural Networks

With the developments in machine learning, firstly artificial neural networks (ANN) have been proposed, which was inspired by the human neural structure. Moreover, the ANNs were evolved into Deep Learning (DL) and improved the learning capabilities of the algorithms. Indeed, for some specific applications, DL surpasses human performance and shows a superhuman performance [35], [36].

An ANN framework is composed of connected neurons which is illustrated in the Figure 2.6.3. Each neuron is connected to other neurons in the consecutive layers with specific weights. The weights are updated during the learning process. The learning starts with input to the first layers' neurons, the neurons calculate their own output by using the input and a non-linear activation function, then transmits the outputs to the next layers neurons. Similarly, Deep Convolutional Neural Network (Deep CNN) algorithms works in the same fashion with different types and number of hidden layers, activation functions and backpropagation.

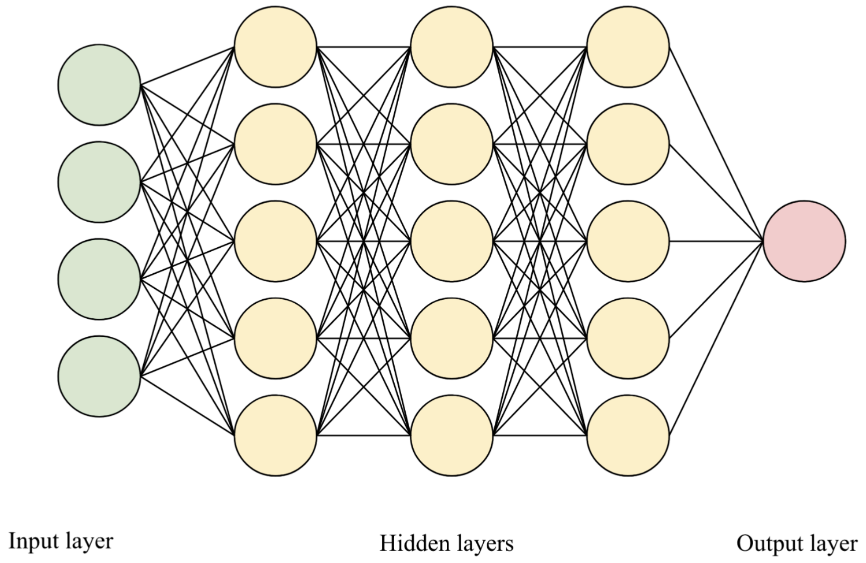


Figure 2.6.3 Artificial Neural Networks

Yann LeCun. et al. proposed Lenet-5, the CNN architecture which contains convolutional and pooling layers, during 2015 for a vision classification task. Following this work, novel CNN architectures inspired by Lenet-5 architecture have been presented in the literature, such as ResNet, AlexNet and so on. Moreover, different optimization algorithms are proposed to improve the Deep CNN's performance and speed of the algorithm. In this thesis, following optimization algorithms are explored:

- ADAM,
- SGD,
- RMSprop,
- AdaGrad.

2.6.2.2 Overfitting and Overconfidence

The Deep Learning algorithms encounter the major issue of Overfitting which later results in overconfidence [37]. Overfitting happens when a model starts to learn training set well by learning regularities in the data but does not perform on the test set. Moreover, as the model overfits to data, its confidence for its predictions increases. In order to tackle these issues, in this thesis dropout layers and weight regularization (L2 regularizer) are used for overfitting and label smoothing is used for overconfidence.

- Weight regulation (L2 regularizer): As the network is trained on training data, the weights specialize for the training data which results in bigger weights. These bigger weights have large variance and small bias, which is a sign of overfitting to the training data. To solve this problem, L2 regularizer is used which calculates the sum of the squared values of the weights [38].
- Dropout: The dropout method temporary removes a neuron from the network during the training which result in the network to learn sparse representations [39].
- Label smoothing regularizes the network from choosing a class with a high confidence [40].

2.6.3 Supervised Contrastive Learning

Supervised contrastive learning (Sup-Con) is a technique that fills the gap between fully supervised learning and self-supervised learning. This method allows contrastive learning to be applied in the supervised setting. In order to learn the representations, a model which uses contrastive learning typically tries to minimize the distance between the positive, while tries to maximize the distance between negative samples. In an unsupervised setting, the positive samples are produced from augmented images and negative samples are chosen from the other samples in the training mini batch. Since this method learns the common attributes of the data by comparing the samples, it mimics the way humans perceive. Furthermore, the loss function of this model boosts the learning of hard negatives and hard positives [41]. Hard positives are pairs of data points that belong to the same class but are difficult for the model to recognize as such. These data points might have subtle differences or variations that make them challenging for the model to identify as belonging together. In the context of image classification, hard positives could be images of the same object but with varying lighting, angles, or occlusions. Hard negatives, on the other hand, are pairs of data points that belong to different classes but are difficult for the model to recognize as different. These data points might have similar features or characteristics that make them appear similar to the model, even though they belong to distinct classes. In image classification, hard negatives could be images of different objects with similar shapes, textures, or colors. Incorporating hard negatives and hard positives into the loss function during training helps the model focus on these challenging pairs, ultimately improving its ability to distinguish between similar-looking data points from different classes and recognize subtle similarities within the same class. This approach can lead to better generalization and higher accuracy in classification tasks.

According to Khosla et al., suggest that the quality of representation may deteriorate as a result of random sampling, leading to false negatives. On the other hand, the Sup-Con uses labeled data, which simplifies the process of positive and negative sample selection. The Sup-Con framework is composed of an encoder network and projector network. Furthermore, the learning process have two stages (Figure 2.6.4). During the first stage, data augmentation is applied twice to a batch of inputs and the copies are passed to the encoder network. The encoder network produces embeddings. The embeddings are then forwarded through the projection network. By employing the normalized outputs of the projection network, the contrastive loss (Equation 28) is computed. In Equation 28, the query is compared with samples within the same class, and then the result is divided by the temperature parameter τ, which is also denoted in the equation. Additionally, at the second stage, a linear classifier is trained on the frozen representations of the first stage. The second stage is composed of a fully-connected layer on top of the encoder, and it is followed by a SoftMax layer with the target classes.

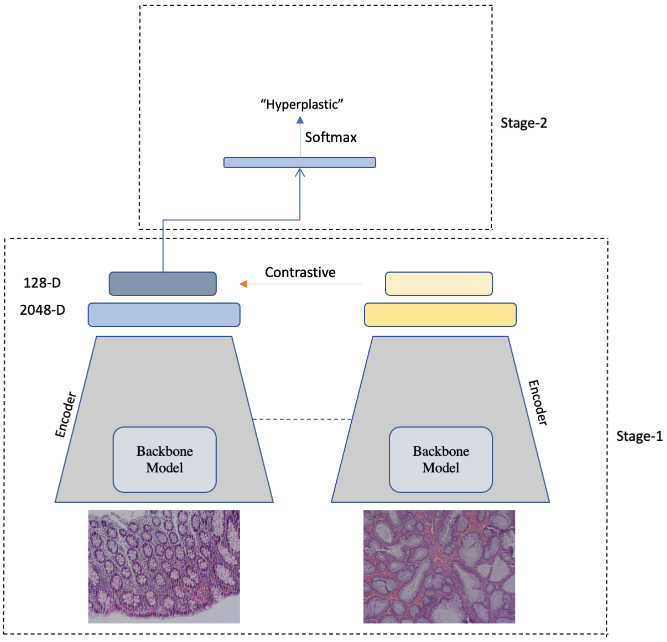


Figure 2.6.4 Supervised Contrastive Learning

Besides, contrastive learning approaches are widely used on histopathology images. There are extensive studies conducted with contrastive learning methods focusing on the histopathology image analysis, such as classification [26], [42]–[44], segmentation [45], [46] and stain normalization for the histopathology images [47].

$$\boldsymbol{\mathcal{L}(query)} = \sum_{\substack{1\le i\le M \\ y(query)=y(x_i)}} -\log\frac{\exp\,(query\cdot x_i/\tau)}{\sum_{j=1}^{M}\exp\,(query\cdot x_j/\tau)} \tag{28}$$

2.6.4 Self Supervised Contrastive Learning

Labeled data is essential for the success of supervised learning, especially for Deep CNN structures that require large amounts of data. However, in medical image analysis, obtaining a sufficient number of labeled medical images can be challenging and time-consuming, presenting a significant obstacle to achieving accurate performance. Moreover, for most of the cases, the labeling should be done on site due to confidentiality of the patient information. On the other hand, self-supervised learning methods use unlabeled data. In general, a self-supervised contrastive learning model uses augmented version of the same image to learn the latent features of the image. The reason why hidden features can be learned from self-supervised learning is because two different augmentations of the same image are expected to have similar representations.

In this thesis, self-supervised contrastive learning is used for pre-training of the Deep CNN models (Figure 2.6.5) After pre-training, the models are fine-tuned with the labeled data to explore the performance of self-supervised contrastive learning for colon histopathology images.

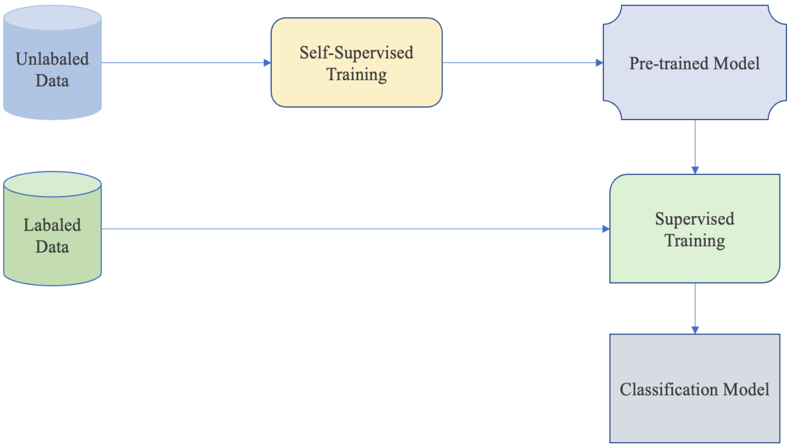


Figure 2.6.5 Self-supervised training

The self-supervised training framework is summarized in Figure 2.6.6. Two different augmentations of the same image are passed to a model. Then by employing a contrastive loss, a contrastive model tries to maximize the similarity of the representations for both images [48]. Moreover, the setup of the contrastive model varies for contrastive learning algorithms (i.e., SimCLR, SimSiam and Barlow Twins).

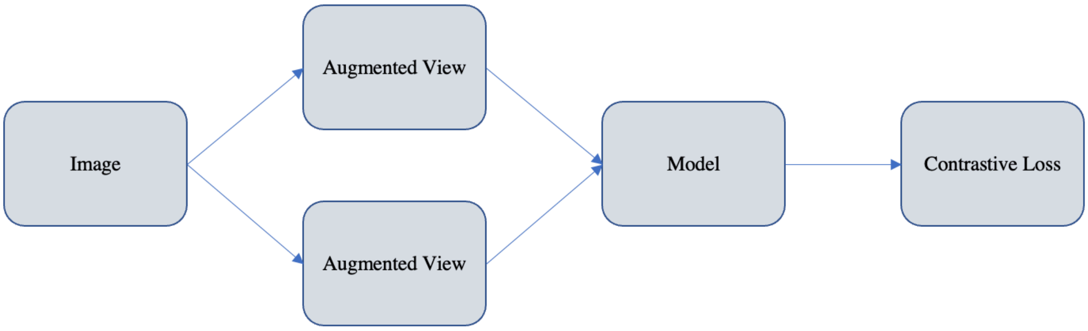


Figure 2.6.6 Self-supervised learning

A contrastive model is typically comprised of three components: Backbone, Projector and Predictor. The Figure 2.6.7 shows the contrastive model's architecture for SimCLR, SimSiam and Barlow Twins. The backbone is the base model which learns the representations. The projector project the representations as embeddings using a contrastive loss. The predictor, which is only used in SimSiam, increases the quality of the representations. In contrastive learning, different views of the same image are processed to the backbone model, the backbone model gives the representations to the projector, in which the similarity or dissimilarity of the representations are calculated with contrastive loss. Additionally, the contrastive loss differs for different contrastive learning algorithms. SimCLR, employs a contrastive cross entropy loss, SimSiam calculates the distance between views by using the cosine distance and Barlow Twins uses Barlow Twins Cross Correlation loss.

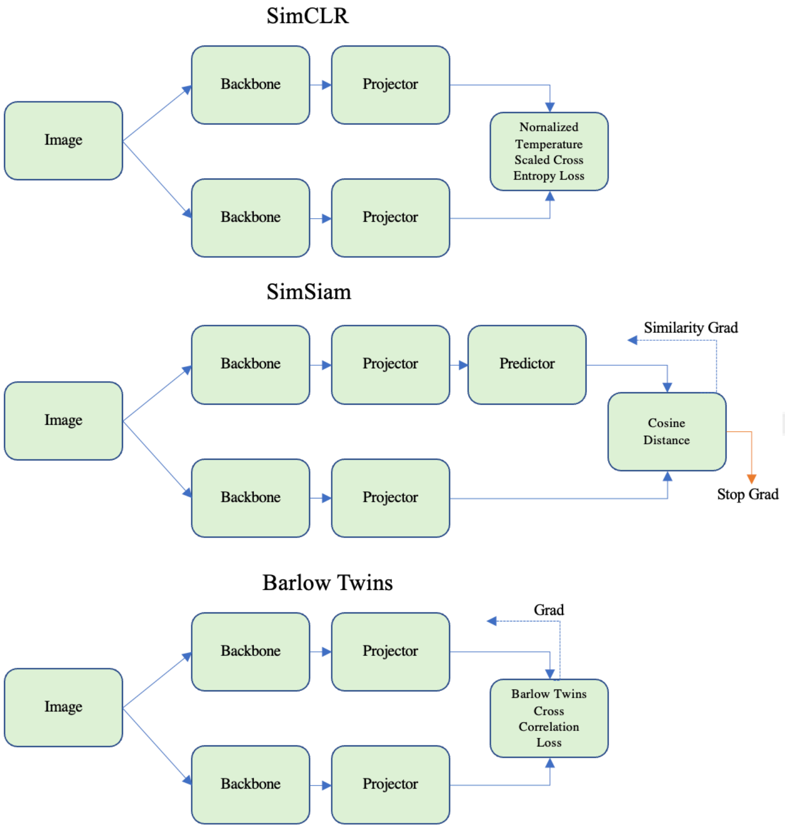


Figure 2.6.7 Self-supervised contrastive learning algorithms

2.6.5 Ensemble Learning

Ensemble learning is a technique which combines various classifiers to enhance classification performance. Different classifiers can capture different information and therefore, ensemble classifiers may result in better accuracy as compared to base learners. Furthermore, ensemble learning methods are widely used in different medical image classification tasks [49]. In [50], Kumar et al. suggested that different CNN classifiers can learn various levels of semantic image representation. In that work, AlexNet and LeNet architectures are fine-tuned on medical images. The proposed method achieved a greater accuracy than the AlexNet and LeNet architectures alone.

2.6.6 ConvNeXt

ConvNeXt architecture has recently been proposed by Liu et al. [51]. This architecture takes the advantages of both the attention-based classifiers and traditional ResNet architectures to compete with the performance of Vision Transformers (ViTs). ConvNeXt architecture is motivated to capture global dependencies by large receptive field and utilizes convolutions with large kernels as the main building block [52]. Moreover, ConvNeXt is a pure CNN architecture, that can outperform the Swin Transformer for ImageNet-1K classification. The architecture of this a ConvNeXt block is presented in Figure 2.6.8 and the ConvNeXt architecture is shown in Figure 2.6.9.

The following stages of the network are composed of ConvNeXt blocks. In each stage, the number of blocks has a ratio of 3:3:9:3. A ConvNeXt block contains a depth-wise convolution which is followed by 1×1 convolutions. The depth-wise convolution implements a special type of group-wise convolution by grouping the channels. The combination of depth-wise convolution and 1 × 1 convolutions performs a similar effect to a property that is shared between vision transformers. Additionally, ConvNeXt architecture implements a Gaussian Error Linear Unit (GELU) as an activation function between the two 1×1 convolution layers and uses layer normalization instead of batch normalization.

Furthermore, various ConvNeXt variants are suggested, namely, ConvNext-Tiny (T), -Small (S), -Base (B), -Large (L) and -X-Large (XL). The diversity of the variants differs as the number of channels and the number of blocks changes for each stage. Table 2.6.1 shows the different configurations of the variants.

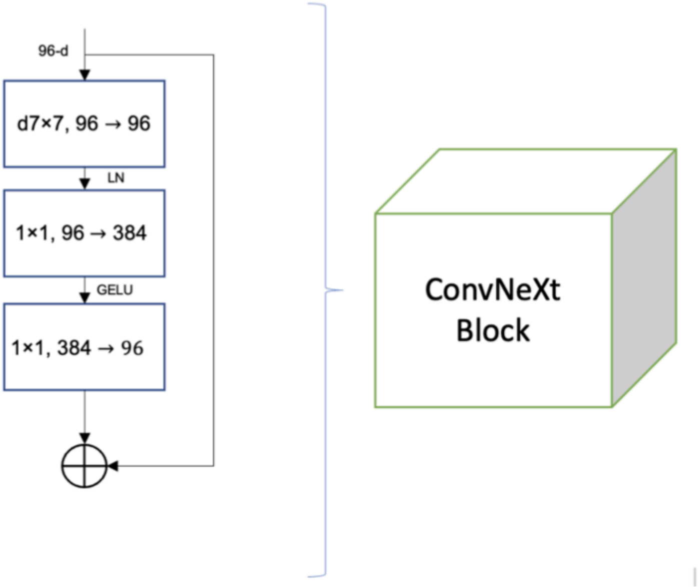


Figure 2.6.8 Structure of a ConvNext Block

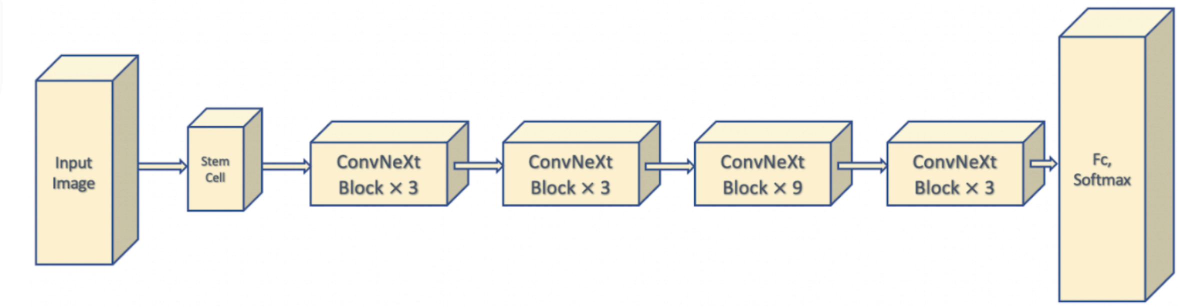


Figure 2.6.9 ConvNeXt Architecture

Table 2.6.1 Different configurations of the ConvNeXt variants

Model / Configurations	Number of Channels (C) of each stage	Number of Blocks (B) of each stage
ConvNeXt-T	(96, 192, 384, 768)	(3, 3, 9, 3)
ConvNeXt-S	(96, 192, 384, 768),	(3, 3, 27, 3)
ConvNeXt-B	(128, 256, 512, 1024)	(3, 3, 27, 3)
ConvNeXt-L	(192, 384, 768, 1536)	(3, 3, 27, 3)
ConvNeXt-XL	(256, 512, 1024, 2048)	(3, 3, 27, 3)

2.6.7 Big Transfer

Big Transfer (BiT) revisits the transfer learning paradigm by introducing some architectural modifications in which it reviews upstream and downstream components. Upstream components are used in pre-training, while downstream components are used during fine-tuning of a new task. The components of upstream tasks are scale, group normalization and weight standardization. Infrastructure of the BiT models are ResNet-v2 architectures of different sizes, pre-trained by supervised learning on natural datasets of different scales. The architectural basis is the same, except that Group Normalization is implemented instead of Batch Normalization and Weight Standardization is applied. Furthermore, BiT performs well on low data regime which contains limited number of samples per class. Additionally, BiT is extensively used in medical image classification tasks.

Chapter 3

3 Collection and Improvement of Dataset

3.1 Collection of Dataset

The histopathology image dataset used in this study was gathered and approved by both the Kayseri City Hospital Ethics Committee and the Erciyes University Clinical Research Ethics Committee, as well as being a part of the TUBITAK project. The images were obtained from 182 patients who underwent colorectal cancer screening at Kayseri City Hospital in Turkey starting from May 2018. Out of these patients, 80 were female and 102 were male, with an age range between 19 and 89 years, and an average age of 62 for both genders. During the colonoscopy procedure, a specimen of the polyp tissue and a neighboring normal tissue were extracted from most of the patients for long-term follow-up. Thus, the dataset contained samples from both adenomatous polyps and non-adenomatous tissues. The extracted tissues are examined with Hematoxylin and eosin staining method. The stained samples were evaluated with a light microscopy of Nikon Eclipse NI during the pathology examination, and images of each sample were taken for different magnification levels using Nikon DS-Fi2, including x2.5, x5, x10, x20, and x40. Each image has a size of 2560x1920. A higher magnification level corresponds to a more zoomed-in image, as shown in Figure 2.3.1. Additionally, some patients had more than one polyp, even with different types such as hyperplastic and tubulovillous. Each patient was assigned a unique patient number, and each sample from the same patient was named accordingly.

3.2 Initial Labeling of Dataset and Initial Experiments

In the initial phase of this thesis, patient reports were utilized to label each sample at different magnification levels. For example, if a patient report indicated that a particular specimen was from a hyperplastic polyp, all magnification levels of that specimen (x2.5, x5, x10, x20, and x40) were labeled as hyperplastic polyp. The initial focus was on binary classification of polyp tissue (including hyperplastic polyp, tubular polyp, tubulovillous, and villous polyp) and normal tissue. Therefore in order to build a baseline for this classification task, several models, such as ResNet18, ResNet50, SqueezeNet, AlexNet, VGG16, DenseNet-161, Vision Transformers with a backbone of Multi-Layer Perceptron, and EfficientNet, were employed in an initial experiment. The results are presented in Table 3.2.1.

Table 3.2.1 Performance for the classification of polyp and normal tissue

Classifiers	**Accuracy**
ResNet18	82%
ResNet50	86%
SqueezeNet	76%
Alexnet	73%
VGG16	70%
DenseNet-161	75%
Vision Transformers	70.67%
EfficinetNet	78%

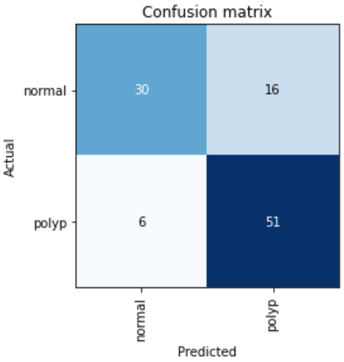


Figure 3.2.1 Sample Confusion Matrix of EfficientNet

During this thesis work, regular meetings were held with M.D. Ebru Akay, who is also a researcher in the TUBITAK project, to gain a better understanding of the pathology images. In our first meeting, the pathologist emphasized the importance of distinguishing hyperplastic polyps and normal tissues from adenomatous polyps, as well as the subtypes of adenomatous polyps (tubular, tubulovillous,villous) as a secondary task. In response to this, two distinct datasets using labels from the pathology reports were created.

In the first dataset, binary classes of adenomatous and non-adenomatous were labeled for each sample and magnification level using global labels extracted from the pathology report. The non-adenomatous class included normal, hyperplastic, and inflamed tissue samples, while the adenomatous class consisted of villous adenomas, tubular adenomas, and tubulovillous adenoma samples.

Similarly, the second dataset was composed of samples with global labels indicating hyperplastic, tubular, and tubulovillous/villous adenomas. Overall, these datasets were designed to address the pathologist's concerns and reflect the clinical importance of distinguishing between different types of polyps.

By using the first dataset, baseline models were initially explored for binary classification of the samples. The performance measures are presented in Table 3.2.2. Furthermore, the performance of the baseline models for different magnification levels was also explored using this dataset, and the results are presented in Table 3.2.1. The average accuracy achieved was between 65-70%. Additionally, the performance of the baseline models for different magnification levels was explored for binary classification. The results are presented in Table 3.2.3. The average accuracy for different magnification levels was found to be in the range of 65-70%. However, the performance of the models for different magnification levels separately was not very satisfactory. To better understand the dataset and customize the models accordingly, another meeting with the pathologist was held.

Table 3.2.2 Accuracy Results of Baseline Models for Binary Classification

Model	Accuracy
Inception Resnet	63.80%
Inceptionv3	54.60%
Efficientnet-b0	60.74%
EfficientNet-V2 Small	61.35%
ResNet50	61.35%
ResNet34	54.60%
SquuezeNet	65.03%

Vision Transformer	54.60%
CCT Transformer	65.54%
SWIN Transformer	57.06%

Table 3.2.3 Accuracy Results of Baseline Models on Different Magnification Levels

Model/Magnification	X2.5	X5	X10	X20	X40
Inception Resnet	62.50%	60.42%	68.75%	64.58%	68.75%
Inceptionv3	64.58%	67.27%	60.42%	67.35%	75.00%
Efficientnet-b0	64.15%	70.83%	72.92%	75.51%	60.75%
NasNet	66.04%	66.67%	63.64%	70.83%	65.62%
ResNet18	64.00%	65.45%	72.00%	71.00%	62.5%
SqueezeNet	71.00%	70.00%	52.00%	81.00%	53.12%
Vision Transformer	56.25%	58.18%	58.18%	60.42%	25.00%
CCT Transformer	54.26%	54.55%	63.24%	69.39%	74.29%
SWIN Transformer	39.62%	58.18%	56.36%	42.5%	59.38%

3.3 Detailed Labeling of the Dataset

During the second meeting with the pathologist, we realized that there was room for improvement in the dataset. This need arose from the use of global labels, which did not capture the fact that a histopathology image may contain samples from more than one class. For example, an adenomatous histopathology image may contain normal cell structures or even a different subtype of an adenomatous cell structure. Figure 3.1.1(a) illustrates a x5 magnification of an adenomatous sample, while Figure 3.1.1(b) shows a x10 magnification of the same image, which zooms in on the normal cell structure of the same sample. This realization led to the development of an improved dataset, where individual patches from different magnification levels were cropped and labeled separately, providing more detailed and accurate information about the samples.

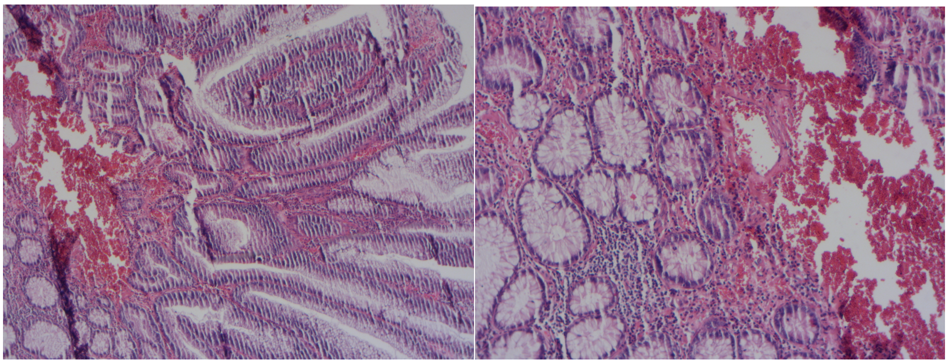


Figure 3.3.1 On the left, 5x magnification of an adenomatous sample, on the right 10x magnification of the same image is zoom into the normal cell structure of the same sample.

Therefore, to improve the dataset, an executable labeling program was developed for the expert pathologist to label each sample independently. Figure 3.3.2 shows a screenshot of the program.

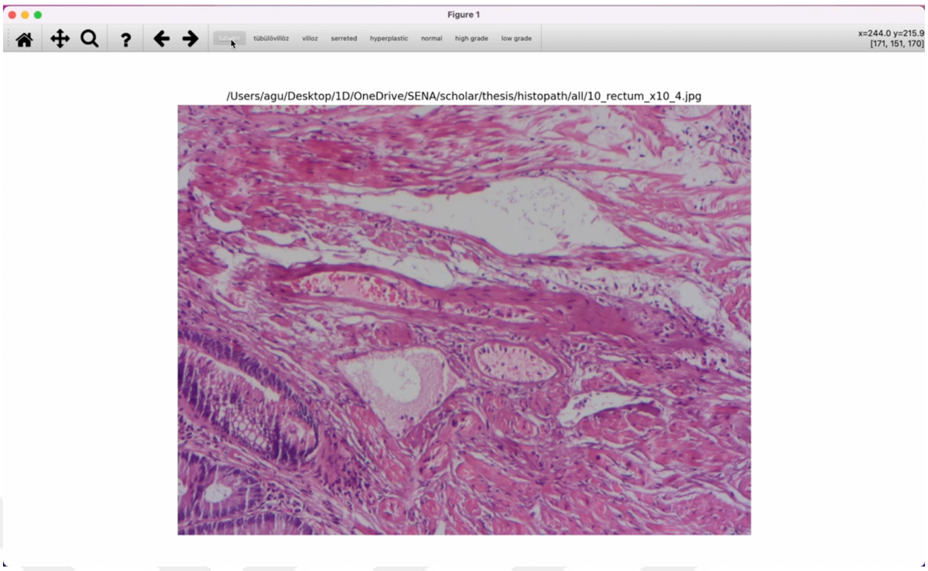


Figure 3.3.2. A screenshot of the executable labeling program

For every slide and magnification level, the samples in the dataset were meticulously labeled by two expert pathologists, namely Dr. Ebru Akay and Dr. Serdal Sadet Özcan. However, since the normal tissue samples only contain normal cell structures, detailed labeling was unnecessary for them. In total, 359 slides belonging to adenomatous polyps (i.e., tubular, tubulovillous/villous) and 181 slides belonging to hyperplastic polyps were used. Additionally, for larger magnifications (i.e., x2.5 and x5), the region of interest was manually annotated as rectangular bounding boxes around the polyps. The final distribution of the extracted samples is as follows: 346 samples belong to tubulovillous polyp type (TVA), 340 samples belong to tubular polyps (TBA), and 370 samples belong to hyperplastic polyps (HP), totaling 1056 samples. Table 3.3.1 provides the distribution of the train, test, and validation sets for each class.

Table 3.3.1 Number of Samples for Each of the Classes and Training/Validation/Test sets

	Hyperplastic	**Tubular**	**Tubulovillous**	**Total**
Train	254	222	242	718
Validation	56	51	50	157
Test	60	67	54	181

3.4 Experiments on Improved Dataset

After improving the dataset by detailed labeling, baseline models were utilized to evaluate the performance of the dataset for binary classification and multi-class classification tasks. The performance metrics for binary classification on the improved dataset are presented in Table 3.4.1. The table shows that Inception-v3 achieved an accuracy of 82.50%, compared to the same model's accuracy of 54.60% on the previous dataset. This indicates that the detailed labeling process has improved the overall accuracy by an average of 15%.

Table 3.4.1 Binary Classification Results for Baseline Models on Improved Dataset

Model	Acc (%)	F1 (%)
Inception-v3	82.50	68.60
ResNet-v2-50	76.25	60.92
ResNet-v2-101	79.81	65.52
InceptionResNet-v2	**86.25**	**76.19**
ViT	78.13	51.34
EfficientNet	86.25	85.00

Furthermore, the performance of the baseline models was also assessed for the multi-class classification task, specifically for distinguishing between hyperplastic polyps, tubular adenomas, and tubulovillous/villous adenomas. The corresponding results are presented in Table 3.4.2.

Table 3.4.2 Multi-class Classification Results for Baseline Models on Improved Dataset

MODEL	Overall Accuracy	Precision	Recall	F1
DenseNet-201	73.02%	72.98%	73.04%	73.00%
Inception-ResNet	67.01%	67.51%	67.79%	67.19%
Inception-v3	68.33%	68.36%	68.54%	68.28%
ResNet-50	70.24%	70.52%	70.41%	70.24%
Xception	73.45%	73.57%	73.43%	73.42%

Chapter 4

4 Study 1

4.1 Improved Classification of Colorectal Polyps on Histopathological Images with Ensemble Learning and Stain Normalization

In the clinical workflow of polyp classification, a key diagnostic challenge is the differentiation of adenomatous polyps from non-adenomatous tissues. The adenomatous polyp types are tubular, villous, and tubulovillous adenomas. Moreover, adenomatous polyps have the potential to develop into cancer, while hyperplastic (i.e. non-adenomatous or normal) polyps are usually not likely to show malignancy potential. Therefore, distinguishing adenomatous/neoplastic polyp tissue from non-adenomatous/hyperplastic/normal tissue is a significant step in cancer screening.

Since early diagnosis is vital, there is a growing demand for cancer screening programs. As the demand for screening increases, the workload of pathologists increases, and consequently it gets harder and harder to detect disease at an early stage. Indeed, over the past decade, according to [6], pathologic colon biopsy slide volumes are doubled. Furthermore, adenomatous lesions are distinguished from non-adenomatous lesions in 96% of cases by the experts, which shows that the problem is not completely solved yet [53]. In order to carry out this process faster and more accurately, Clinical Decision Support System (CDSS) can be employed, which can ease this labour-intensive work and minimize the mistakes of the traditional approaches. In this work, a CDSS is proposed to assist experts by providing the classes of each histopathology image and highlighting the most suspected areas with Grad-Cam method.

Ensemble learning methods are widely used in medical image classification tasks. In their work Kumar et al. proposed a model that ensembles AlexNet and LeNet architectures, which achieves a greater accuracy than the AlexNet and LeNet architectures alone [50]. Kallipolitis et al. ensemble EfficientNet variants to detect cancer from pathology images of breast and colon [54]. In another work, Kassani et al. employ a model which ensembled pre-trained VGG19, MobileNet, and DenseNet to detect cancerous regions from breast histology images [55]. Das et al. feed the wavelet transformed breast histology images to a model which ensembled three different CNN structures [49]. Kundu et al. design a network, which is composed of LeNet, ResNet-18 and DenseNet 121, to detect pneumonia from chest X-Ray images [56]. Nguyen et al. ensemble a model to detect polyps during the colonoscopy procedure. In their work, as input, they fed the model with endoscopy and pathology images [57].

Ensemble methods are widely employed in breast cancer classification tasks on histopathology images. They can make more robust decisions since they employ various classifiers as base learners, which can capture different levels of information contained in the latent features. However, they have not been used in colonic adenomatous polyp detection on histology images. To the best of our knowledge, this study is the first to use ensemble methods to classify colonic histological images as adenomatous and non-adenomatous. The main contributions of this study are as follows:

- In this study, we explore state-of-the-art pre-trained Deep CNN algorithms' performances on our custom dataset. To the best of our knowledge, this study is the first to comprehensively evaluate widely used stain normalization techniques namely, Stain-GAN, Stain-Net, Vahandane, Macenko and Reinhard by combining with state-of-the art Deep CNN models for classification of adenomatous and non-adenomatous colonic polyp tissues.
- This study is one of the first studies which employs ConvNeXt architecture on colon histopathology images for polyp classification task.
- We propose a model which ensembles the pre-trained ConvNeXt-Tiny and ConvNeXt-Base variants to classify adenomatous and non-adenomatous tissues on colonic histopathology images. Moreover, the variants are further tailored to the problem by network modifications at the image representation levels. In order to comprehensively evaluate and assess the generalizability of the proposed

model, we also employ publicly available UniToPatho and EBHI databases The proposed ensemble model achieves an accuracy of 95% on our custom dataset.

- Additionally, in order to ensure the explainability of the proposed model, the Grad-Cam method is used. The attention map of the model is explored for adenomatous and non-adenomatous images. We believe that these Grad-Cam visual outputs can help pathologists to see and judge the decision making process of the model.

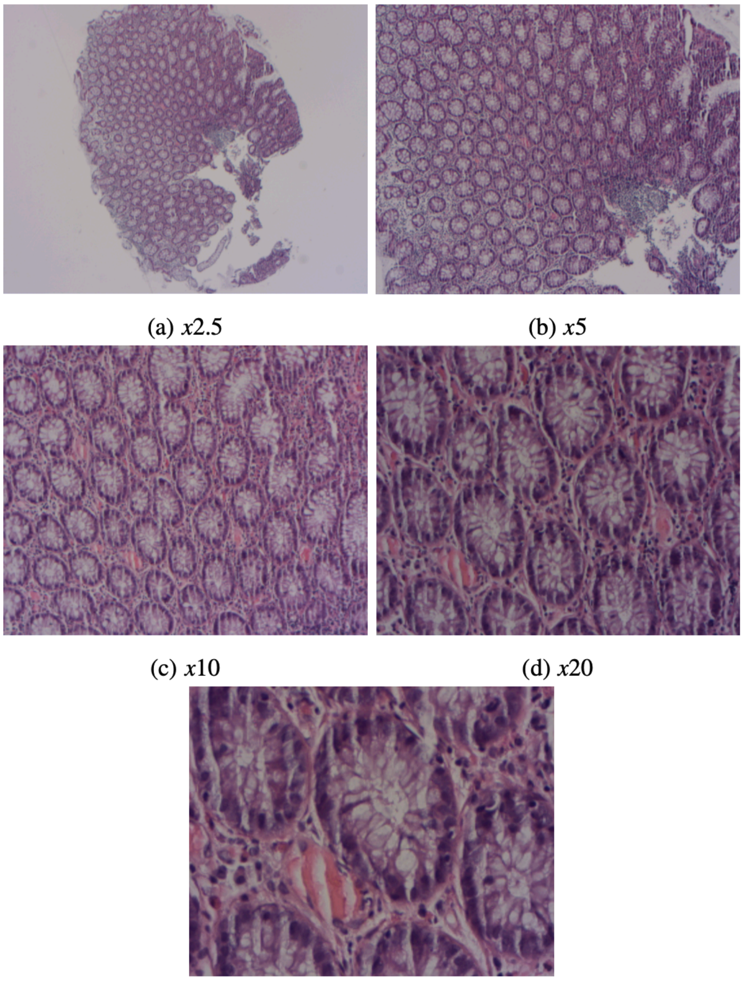
4.1.1 Material and Methods

4.1.1.1 Dataset

The histological slides used in this study were collected from 84 patients who underwent colorectal cancer screening since May 2018 at Kayseri City Hospital, Kayseri, Turkey. This study was approved by Kayseri City Hospital Ethics Committee and Erciyes University Clinical Research Ethics Committee. Forty-six of the 84 patients are male, while the other 38 are female.

Most of the patients have samples from both adenomatous polyp and non-adenomatous tissues. Each tissue sample was examined under the microscopy with 5 different magnification levels. Samples from different magnification settings can be seen in Figure 4.1.1. The magnification levels are: x2.5, x5, x10, x20 and x40.

A total of 671 slides were collected, with 359 classified as adenomatous polyps, and 312 classified as non-adenomatous tissues, including hyperplastic polyps, normal tissue, and chronic inflammation. The detailed labeling of the collected samples was done by two expert pathologists for each slide and magnification level.



(a) *x*2.5

(b) *x*5

(c) *x*10

(d) *x*20

(e) *x*40

Figure 4.1.1 Different magnification levels of a tissue

Four hundred seventy (470) slides were randomly selected for training set, 101 for validation set and 100 as test set. The train, test and validation sets were separated on a patient-based approach. That is, there were no whole-slide images that belong to the same patient in two different sets. A detailed description of the collected slides can be found in Table 4.1.1.

To evaluate the models' generalizability and stain normalization techniques, a total of 304 slides were randomly selected, with 152 slides acquired from each of the publicly available UniToPatho and EBHI databases. UniToPatho database contains 9536 hematoxylin and eosin stained patches extracted from 292 whole-slide images, where each of the slides have a magnification of ×20 [24]. The WSIs belong to the following classes: normal tissue, hyperplastic polyp, tubular adenoma and tubulo-villous adenoma. EBHI is composed of 5532 WSIs which has the categories of normal, low-grade and high-grade intra-epithelial neoplasm, and adenocarcinoma and divided into four magnifications of ×40, ×100, ×200 and ×400 [25].

Table 4.1.1 Number of Samples and Patients of Custom Collected Dataset

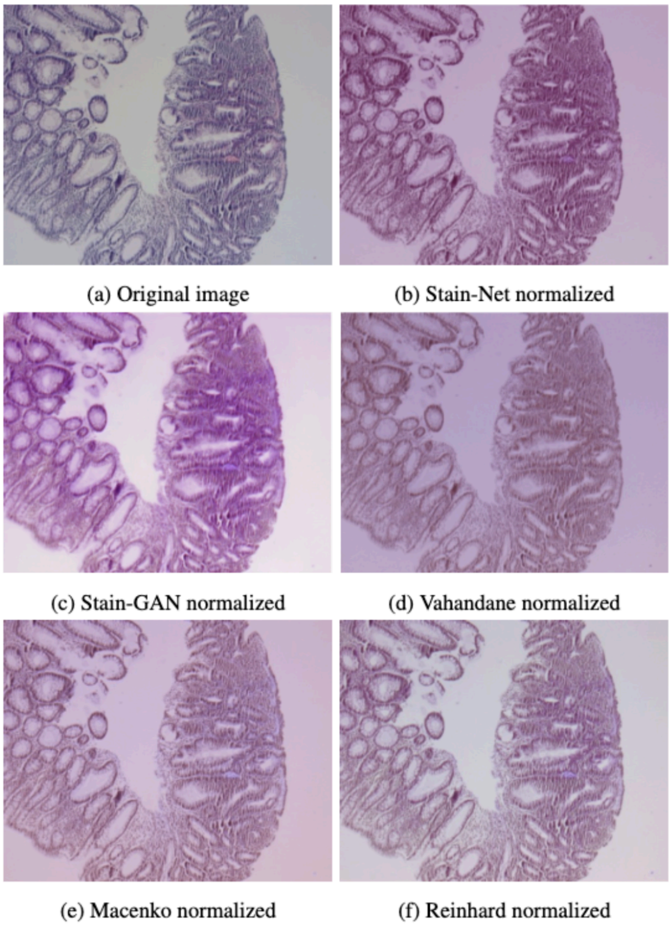
Class	Number of Samples	Number of Patients
Adenoma	359	52
Hyperplasia	181	29
Normal/Chronic Inflammation	130	30

In most cases, an adenomatous slide contains one or more different tissue structures, including both adenomatous and normal tissue. Additionally, expert pathologists typically determine whether a whole slide image (WSI) contains adenomatous tissue by examining the entire image. Thus, in this study, we classified each WSI individually rather than using manually cropped patches from the WSIs.

4.1.1.2 Stain Normalization

Deep CNN algorithms have a great capacity to fit a dataset with high precision. However, this precision challenges the model to generalize for the unseen data. Moreover, if there is a domain shift in training and testing data, the model must be robust and reliable for real-world scenarios. To address the issue of domain shift, we applied commonly used Stain Normalization techniques to our dataset in this study.

To ensure good generalization ability, a deep CNN algorithm must be resilient to domain shifts. Researchers have proposed various stain normalization techniques to address this issue. The literature commonly employs the following techniques: Vahandane, Macenko, Reinhard, Stain-GAN, and Stain-Net. Vahandane, Macenko, and Reinhard are more conventional methods, whereas Stain-GAN and Stain-Net utilize Generative Adversarial Network (GAN) architectures. Figure 4.1.2 is an adenomatous image from our dataset in which different stain normalization methods are applied.



(a) Original image

(b) Stain-Net normalized

(c) Stain-GAN normalized

(d) Vahandane normalized

(e) Macenko normalized

(f) Reinhard normalized

Figure 4.1.2 Results of the stain normalization techniques on a sample image on our dataset

4.1.2 Proposed Ensemble of ConvNeXt Framework

Figure 4.1.4 shows the proposed framework to classify colonic histological images as either adenomatous or non-adenomatous. Previous studies on colonic polyp classification problem predominantly employ a single deep CNN algorithm. According to previous studies, CNN architectures play an important role for the classifier performance, [58]–[61]. As stated in the [57], deep residual CNN architectures are used for more complex problems, while shallower CNNs are used for simple problems. Additionally, ensemble methods perform better than a single deep CNN algorithm, because its base classifiers can interpret various properties of an input image. Consequently, we designed the proposed framework by employing an ensemble of ConvNeXt variants.

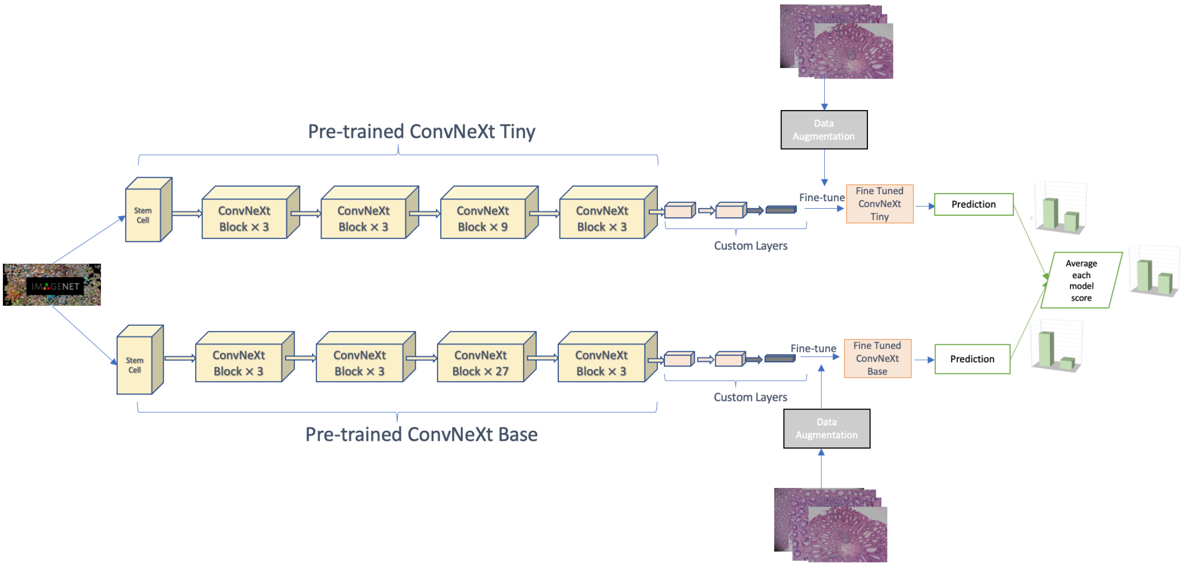


Figure 4.1.3 Proposed ensemble of ConvNeXts framework

We selected the ConvNeXt architecture since it is more suitable to classify adenomatous colonic WSIs than the attention-based networks or regular CNNs because of the following reasons: Unlike ViT [62] it does not require large amount of data in training. Since there are limited number of samples in our custom dataset, this makes ConvNeXt more convenient than data-hungry attention-based methods such as ViT. Moreover, in contrast to CNNs, it can capture longer dependencies because of its large receptive field. Similar to what experts do, it can recognize an adenomatous polyp structure in a histopathological image by visually inspecting spatially distant cell structures. This makes ConvNeXt more suitable for this task, while it is challenging for CNNs to capture those distant correlations.

The base classifiers of the ensemble model are ConvNeXt-Tiny and ConvNeXt-Base models, which are pre-trained on ImageNet-21k dataset. In order to make those networks more suitable to our task, we implemented a fine-tuning approach, which consists of unfreezing the entire model and re-training it on our data for each of the models separately. Subsequently, drop-out and a dense layer are added to the top layer. Adam optimizer with a learning rate of 0.001 is employed, and adaptive momentum optimization algorithm optimized the learning rate during the training. As the loss function, binary cross entropy is used with label smoothing with a smoothing coefficient having a value of 0.1. By this way, label smoothing regularizes the network to choose the class with high confidence. The training batch size is set to 64 and both of the networks are fine-tuned in 50 epochs individually. In the final decision step, the probabilistic outputs of each network for each of the classes are then averaged at the average layer to make a final decision.

Furthermore, in order to verify the proposed model's generalizability, the model is tested on three different sets of instances, which come from UnitoPatho, EBHI datasets and the testing set of the custom collected dataset.

Additionally, in order to show the explainability of the proposed model, the Grad-CAM method is used. The attention map of the model is explored for adenomatous and non-adenomatous images. Outputs of the Grad-CAM results of the proposed model can help the pathologist to see and judge the inner decision step of the model.

4.1.3 Experimental Setup

Figure 4.1.3 shows experimental setup of the proposed framework. As it is mentioned in the Section 4.1.2; each of the base classifiers of the ensemble model is fine-tuned on our dataset by adding a drop-out and a dense layer to the top layer.

The performance of the proposed ensemble method is compared against the extensively used pre-trained deep CNN methods and attention-based method. We further compared frequently used stain normalization techniques for each of the deep CNN-based, attention-based methods and the proposed ensemble method. For these ablation tests the following standard performance measures are used: F-1 score, accuracy, precision, and recall.

Initially, we first normalized the WSIs using the following stain normalization techniques separately and obtained different sets of normalized histological images: Stain-Net, Stain-GAN, Reinhard, Macenko and Vahadane. In order to apply Reinhard, Macenko and Vahadane techniques, we employed StainTools library [63]. The Stain-Net and Stain-GAN methods are implemented by using the source codes of [31], and [30].

The models that are used for the ablation study are employed from TensorFlow-Hub and the models are: Inception-V3 (InceptionV3 trained on ImageNet), ResNetV2-50 (ResNetV2-50 trained on ImageNet), ResNetV2-101, (ResNetV2-101 trained on ImageNet) , InceptionResNet-V2 (InceptionResNet-V2 trained on ImageNet) , ViT (fine-tuned on ImageNet 1k), EfficientNet-S (EfficientNet V2 pre-trained on ImageNet), EfficientNet-S-21k (EfficientNet V2 pretrained on ImageNet 21k), EfficientNet-S-21k-ft-1k (EfficientNet V2 pretrained on ImageNet-21k and fine-tuned on ImageNet-1k), ConvNeXt-Tiny (model pre-trained on the ImageNet-1k dataset), ConvNeXt-Small (model pre-trained on the ImageNet-1k dataset), ConvNeXt-Base-1k (model pre-trained on the ImageNet-1k dataset), ConvNeXt-Base-21k (model pre-trained on the ImageNet-21k dataset), ConvNeXt-Base 21k-ft-1k (model pre-trained on ImageNet-21k and finetuned on ImageNet 1k), ConvNeXt-Large (model pre-trained on the ImageNet-21k dataset).

We employed all the models as pre-trained; this is due to the fact that building those models from scratch needs huge amount of data. After adding custom layers at the end of the base models, we implemented a fine-tuning approach because the state-of-the-art CNN models are pre-trained on natural images, while our images belong to a different domain. Thus, in order to comprehensively compare the above-mentioned methods, we first fine-tuned them on our histological dataset. All the experiments were performed on Google Colab Platform with 52 GB of RAM and NVIDIA Tesla K80, NVIDIA Tesla T4 and NVIDIA Tesla P100 GPU accelerators. The application of the proposed experiments was implemented with Python v3.7.13 with the TensorFlow v2.8.0 framework. The details about the network parameters, optimizers, number of epochs, learning rates, and number of parameters are given Table 4.1.2 and Table 4.1.3.

Table 4.1.2 Network Parameters of The Proposed Method

Parameters	**Values**
Optimizer	ADAM
Learning Rate	0.001
Number of Epochs	50
Batch Size	64
Regularizer	L2 Norm

Table 4.1.3 Number of Parameters of The Models That are Used in This Work

Network	**Number of Parameters**
ConvNeXt-Large	229,843,637
ConvNeXt-Base	87,568,514
ConvNeXt-Small	49,456,226
ConvNeXt-Tiny	27,821,666
Inception-v3	21,806,882
ResNet-v2-50	23,568,898
ResNet-v2-101	42,630,658
InceptionResNet-v2	54,339,810
ViT	36,047,682
EfficientNet-v2-s	20,333,922
Proposed Method	115,390,180

4.1.4 Results and Discussions

For the curated sets of WSIs with different stain normalization techniques, we primarily experiment with the aforementioned baseline classifiers. The ResNet-50 is implemented as it is proposed by Korbar et al. [7]. oreover, other popular deep learning approaches are also implemented and the test performance of each model and stain normalization techniques are shown in Table 4.1.4 and Tables 4.1.5 to 4.1.10. The performance results of the top-performed model on the custom dataset are given in Table 4.1.5. This table shows the accuracy metrics of the proposed model on the custom-collected dataset with various normalization techniques.

Table 4.1.4 Accuracy results of the baseline models and the proposed model on the curated sets

Normalization	**Without Normalization**		**Stain-Net**		**Stain-GAN**		**Reinhard**		**Macenko**		**Vahande**	
Performance	**Acc (%)**	**F1 (%)**	**Acc(%)**	**F1 (%)**	**Acc(%)**	**F1 (%)**	**Acc (%)**	**F1 (%)**	**Acc (%)**	**F1 (%)**	**Acc (%)**	**F1 (%)**
ConvNeXt-Large	84.38	65.54	83.13	74.25	88.75	74.42	86.88	80.00	84.38	65.54	81.25	69.41
ConvNeXt-Base-21k	80.63	73.94	86.25	85.00	80.00	58.89	86.67	75.71	74.38	49.73	57.50	37.36
ConvNeXt-Base-21k-ft-1k	81.25	63.64	66.25	43.48	80.63	58.56	88.13	79.04	88.13	83.44	54.38	40.00
ConvNeXt-Small	78.13	63.58	81.25	63.64	83.75	65.91	86.88	77.84	83.75	67.82	71.88	67.59
ConvNeXt-Tiny	87.50	77.38	85.00	72.94	83.75	73.81	**91.36**	**89.02**	88.75	78.57	82.50	72.62
Inception-v3	82.50	68.60	80.00	68.24	85.53	69.77	86.25	74.12	82.50	66.67	75.74	63.44
ResNet-v2-50	76.25	60.92	78.13	63.58	82.50	64.77	82.50	64.77	75.63	54.14	77.50	62.07
ResNet-v2-101	79.81	65.52	77.36	56.98	83.13	72.19	80.00	58.89	74.21	53.93	71.25	49.45
InceptionResNet-v2	**86.25**	**76.19**	81.88	71.01	82.50	74.70	83.02	71.43	79.38	57.46	80.00	70.24
ViT	78.13	51.34	75.63	54.14	48.13	47.20	55.33	56.74	59.12	53.66	58.13	68.46
EfficientNet-v2-s	86.25	85.00	88.68	88.05	87.50	79.52	89.38	84.66	85.00	72.94	83.13	76.36
EfficientNet-v2-s-21k-ft-1k	85.00	83.75	86.88	82.21	**90.00**	**86.42**	89.87	82.93	**89.31**	**87.50**	83.75	81.08
EfficientNet-v2-s-21k	85.63	87.90	**89.38**	**78.11**	89.44	86.42	89.38	78.11	88.75	87.50	**84.13**	**71.62**
Proposed Method	**93.75**	**93.58**	**95.00**	**93.90**	**92.50**	**91.25**	**91.88**	**90.57**	**91.93**	**90.48**	**88.82**	**87.12**

Table 4.1.5 Performance results of the proposed model for different normalization methods

Normalization	Accuracy	Precision	Recall	F1
Original	**93.75%**	93.58%	93.58%	93.58%
Stain-Net	**95.00%**	92.77%	95.06%	93.90%
Stain-GAN	92.50%	92.41%	90.12%	91.25%
Reinhard	91.88%	92.31%	88.89%	90.57%
Vahandane	88.82%	87.65%	86.59%	87.12%
Macenko	91.93%	88.37%	92.68%	90.48%

From Table 4.1.4 and Figure 4.1.5 it can be seen that Stain-GAN and Reinhard normalization techniques perform better than the non-normalized dataset and other methods. Furthermore, accuracy and F1 scores of the Stain-GAN and Reinhard normalized datasets are approximately improved by 3-5% for the baseline models. The performance of the proposed model is evaluated for each stain normalization technique and is given in Tables 4.1.5 to 4.1.10.

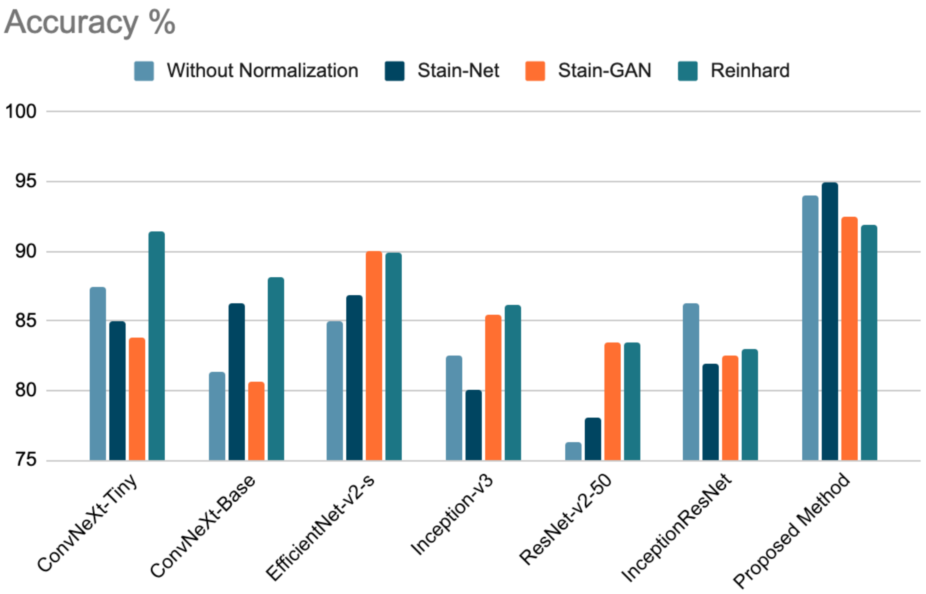


Figure 4.1.4 Comparison of different stain normalization techniques for the proposed method and baseline models

The proposed method on our custom dataset performs the best on Stain-Net normalized dataset and achieves the highest accuracy, precision, recall, and F-score with values of 95%, 92.8%, 95.1% and 93.9%, respectively. On the other hand, the performance of the ensemble model is relatively poor for the Vahandane normalized dataset with an accuracy, precision, recall, and F-score with values of 88.8%, 87.7%, 86.6% and 87.1%, respectively.

The performance of the proposed method and all the base-line classifiers are poor for the Vahandane normalized data. For the Vahandane normalized dataset, the same proportion of adenomatous and non-adenomatous images are confused by all the base classifiers and the proposed ensemble model. This may originate from the fact that the Vahandane normalized images have poorer contrast than the other normalized images. In order to address this problem, we implemented different data augmentation techniques that include random contrast, random brightness, and random hue, however, we observe that it is more suitable to employ image pre-processing techniques, such as adaptive histogram equalization, before the normalization of an image with Vahandane normalization algorithm.

A comprehensive comparison of the top performed state-of-the-art Deep CNN classifiers' performance on the non-normalized, Stain-GAN normalized, Stain-Net normalized and Reinhard normalized data are presented in Figure 4.1.5. As it can be seen from the figure and the table, for the Reinhard normalized dataset, the maximum accuracies of 91.4%, 88.1%, 89.9%, 86.2%, 83.5% and 83.0% are produced by ConvNeXt-Tiny, ConvNeXt-Base, Efficient-Net-v2-S, Inception-v3, ResNet-v2-50 and InceptionResNet models, respectively. Moreover, we can see that the performance of the single deep CNN classifiers' accuracy results varies in terms of different normalization techniques. However, the variation of performance for the proposed method for different normalization techniques is relatively small. Thus, this minor variation shows that the proposed model is more robust to input variations in the given datasets and generalizes better than the single model classifiers.

As it can be seen in Tables Tables 4.1.5 to 4.1.11, the most significant performance gap for the proposed method and the base classifiers is obtained for the non-normalized and Stain-Net normalized data. The proposed method increased the overall accuracy for the non-normalized and Stain-Net normalized data by 6%, for Vahandane normalized data by 4%, for the Stain-GAN and Macenko normalized data by 2%. The best accuracy of the ensemble method is achieved for the Stain-Net normalized data with an accuracy of 95% which is followed by Efficient-Net-v2-S with 89%. The ROC curves of the proposed method for the non-normalized dataset and Stain-Net normalized dataset are presented on Figure 4.1.6 and Figure 4.1.7, respectively. The results of the experiments show that the performance of the proposed ensemble model is satisfactory on all the normalized datasets and non-normalized dataset. Especially, the results on the non-normalized dataset show that the proposed ensemble model has a sufficient generalization ability since the images on the dataset differ in terms of color intensity.

Table 4.1.6 Accuracy results of the baseline models and the proposed model on our dataset without normalization

Non-normalized Data	Accuracy	Precision	Recall	F-1 Score
ConvNeXt-Large	84.38%	73.42%	59.18%	65.54%
ConvNeXt-Base-21k	80.63%	77.22%	70.93%	73.94%
ConvNeXt-Base-21k-ft-1k	81.25%	70.89%	57.73%	63.64%
ConvNeXt-Small	78.13%	69.62%	58.51%	63.58%
ConvNeXt-Tiny	87.50%	82.28%	73.03%	77.38%
Inception-v3	82.50%	74.68%	63.44%	68.60%
ResNet-v2-50	76.25%	67.09%	55.79%	60.92%
ResNet-v2-101	79.81%	69.72%	61.79%	65.52%
InceptionResNet-v2	86.25%	81.01%	71.91%	76.19%
ViT	78.13%	60.76%	44.44%	51.34%
EfficientNet-v2-s	86.25%	86.08%	83.95%	85.00%
EfficientNet-v2-s-21k-ft-1k	85.00%	84.81%	82.72%	83.75%
EfficientNet-v2-s-21k	85.63%	87.34%	88.46%	87.90%
Proposed Method	93.75%	93.58%	93.58%	93.58%

Table 4.1.7 Accuracy results of the baseline models and the proposed model on our dataset with Stain-Net normalization

Stain-Net	Accuracy	Precision	Recall	F-1 Score
ConvNeXt-Large	83.13%	78.48%	70.45%	74.25%
ConvNeXt-Base-21k	86.25%	86.08%	83.95%	85.00%
ConvNeXt-Base-21k-ft-1k	66.25%	50.63%	38.10%	43.48%
ConvNeXt-Small	81.25%	70.89%	57.73%	63.64%
ConvNeXt-Tiny	85.00%	78.48%	68.13%	72.94%
Inception-v3	80.00%	73.42%	63.74%	68.24%
ResNet-v2-50	78.13%	69.62%	58.51%	63.58%
ResNet-v2-101	77.36%	64.56%	51.00%	56.98%
InceptionResNet-v2	81.88%	75.95%	66.67%	71.01%
ViT	75.63%	62.03%	48.04%	54.14%
EfficientNet-v2-s	88.68%	88.61%	87.50%	88.05%
EfficientNet-v2-s-21k-ft-1k	86.88%	84.81%	79.76%	82.21%
EfficientNet-v2-s-21k	89.38%	83.54%	73.33%	78.11%
Proposed Method	95.00%	92.77%	95.06%	93.90%

Table 4.1.8 Accuracy results of the baseline models and the proposed model on our dataset with Stain-GAN normalization

Stain-GAN	Accuracy	Precision	Recall	F-1 Score
ConvNeXt-Large	88.75%	81.01%	68.82%	74.42%
ConvNeXt-Base-21k	80.00%	67.09%	52.48%	58.89%
ConvNeXt-Base-21k-ft-1k	80.63%	67.09%	51.96%	58.56%
ConvNeXt-Small	83.75%	73.42%	59.79%	65.91%
ConvNeXt-Tiny	83.75%	78.48%	69.66%	73.81%
Inception-v3	85.53%	76.92%	63.83%	69.77%
ResNet-v2-50	82.50%	72.15%	58.76%	64.77%
ResNet-v2-101	83.13%	77.22%	67.78%	72.19%
InceptionResNet-v2	82.50%	78.48%	71.26%	74.70%
ViT	48.13%	47.50%	46.91%	47.20%
EfficientNet-v2-s	87.50%	83.54%	75.86%	79.52%
EfficientNet-v2-s-21k-ft-1k	90.00%	88.61%	84.34%	86.42%
EfficientNet-v2-s-21k	89.44%	88.61%	84.34%	86.42%
Proposed Method	92.50%	92.41%	90.12%	91.25%

Table 4.1.9 Accuracy results of the baseline models and the proposed model on our dataset with Reinhard normalization

Reinhard	Accuracy	Precision	Recall	F-1 Score
ConvNeXt-Large	86.88%	83.54%	76.74%	80.00%
ConvNeXt-Base-21k	86.67%	79.76%	72.04%	75.71%
ConvNeXt-Base-21k-ft-1k	88.13%	83.54%	75.00%	79.04%
ConvNeXt-Small	86.88%	82.28%	73.86%	77.84%
ConvNeXt-Tiny	91.36%	90.12%	87.95%	89.02%
Inception-v3	86.25%	79.75%	69.23%	74.12%
ResNet-v2-50	82.50%	72.15%	58.76%	64.77%
ResNet-v2-101	80.00%	67.09%	52.48%	58.89%
InceptionResNet-v2	83.02%	76.92%	66.67%	71.43%
ViT	55.33%	57.97%	55.56%	56.74%
EfficientNet-v2-s	89.38%	87.34%	82.14%	84.66%
EfficientNet-v2-s-21k-ft-1k	89.87%	86.08%	80.00%	82.93%
EfficientNet-v2-s-21k	89.38%	83.54%	73.33%	78.11%
Proposed Method	91.88%	92.31%	88.89%	90.57%

Table 4.1.10 Accuracy results of the baseline models and the proposed model on our dataset with Vahandane normalization

Vahandane	Accuracy	Precision	Recall	F-1 Score
ConvNeXt-Large	81.25%	74.68%	64.84%	69.41%
ConvNeXt-Base-21k	57.50%	43.04%	33.01%	37.36%
ConvNeXt-Base-21k-ft-1k	54.38%	44.30%	36.46%	40.00%
ConvNeXt-Small	71.88%	76.56%	60.49%	67.59%
ConvNeXt-Tiny	82.50%	77.22%	68.54%	72.62%
Inception-v3	75.74%	67.05%	60.20%	63.44%
ResNet-v2-50	77.50%	68.35%	56.84%	62.07%
ResNet-v2-101	71.25%	56.96%	43.69%	49.45%
InceptionResNet-v2	80.00%	74.68%	66.29%	70.24%
ViT	58.13%	64.56%	72.86%	68.46%
EfficientNet-v2-s	83.13%	79.75%	73.26%	76.36%
EfficientNet-v2-s-21k-ft-1k	83.75%	89.55%	74.07%	81.08%
EfficientNet-v2-s-21k	84.13%	75.23%	68.33%	71.62%
Proposed Method	88.82%	87.65%	86.59%	87.12%

Table 4.1.11 Accuracy results of the baseline models and the proposed model on our dataset with Macenko normalization

Macenko	Accuracy	Precision	Recall	F-1 Score
ConvNeXt-Large	84.38%	73.42%	59.18%	65.54%
ConvNeXt-Base-21k	74.38%	58.23%	43.40%	49.73%
ConvNeXt-Base-21k-ft-1k	88.13%	86.08%	80.95%	83.44%
ConvNeXt-Small	83.75%	74.68%	62.11%	67.82%
ConvNeXt-Tiny	88.75%	83.54%	74.16%	78.57%
Inception-v3	82.50%	73.42%	61.05%	66.67%
ResNet-v2-50	75.63%	62.03%	48.04%	54.14%
ResNet-v2-101	74.21%	61.54%	48.00%	53.93%
InceptionResNet-v2	79.38%	65.82%	50.98%	57.46%
ViT	59.12%	55.70%	51.76%	53.66%
EfficientNet-v2-s	85.00%	78.48%	68.13%	72.94%
EfficientNet-v2-s-21k-ft-1k	89.31%	88.61%	86.42%	87.50%
EfficientNet-v2-s-21k	88.75%	88.61%	86.42%	87.50%
Proposed Method	91.93%	88.37%	92.68%	90.48%

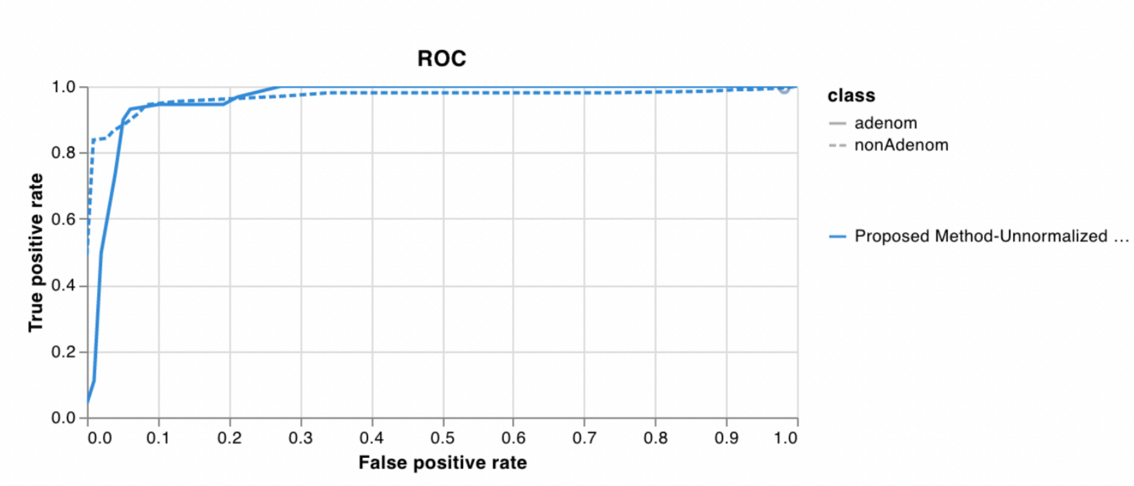


Figure 4.1.5 ROC of the proposed method on our dataset without normalization

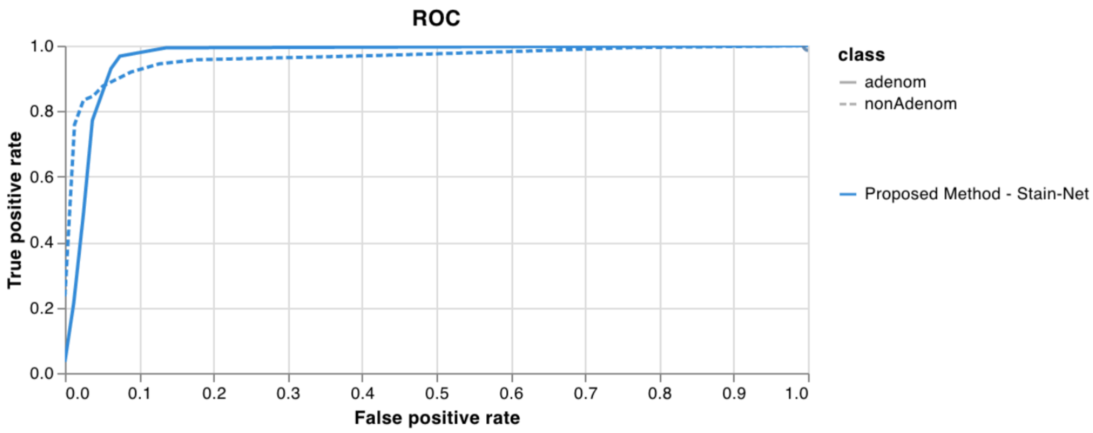


Figure 4.1.6 ROC of the proposed method on our dataset with Stain-Net normalization

4.1.5 Generalization Test

To evaluate the model's generalization ability, colonic adenomatous and non-adenomatous images from both the UniToPatho and EBHI datasets were used. The images obtained were fed into the model trained on our non-normalized dataset with fine-tuning. The achieved overall accuracies were 91.1% and 90% for the EBHI and UniToPatho datasets, respectively. The performance metrics of the proposed model on the UniToPatho and EBHI datasets can be found in the Table 4.1.12.

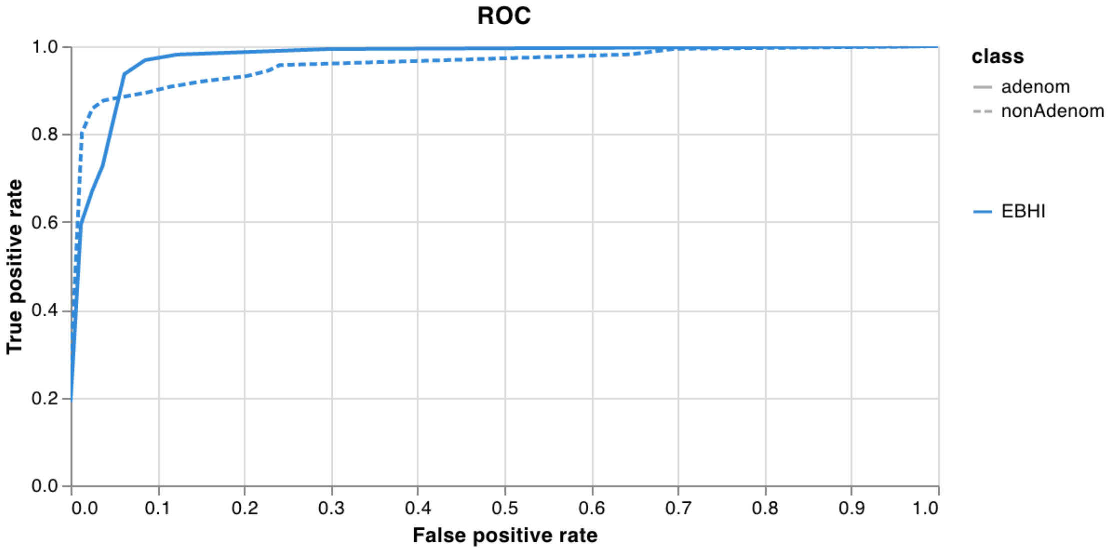


Figure 4.1.7 ROC of the proposed method on EBHI dataset

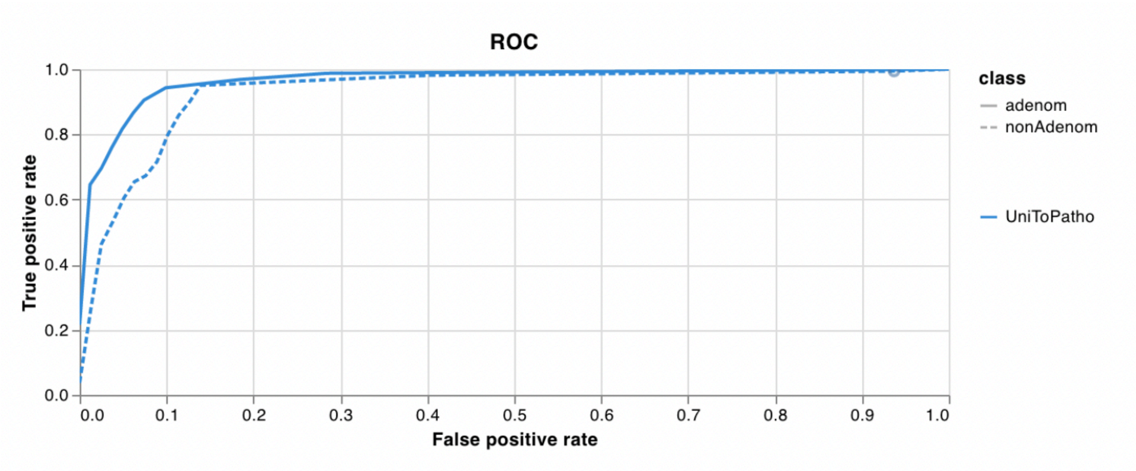


Figure 4.1.8 ROC of the proposed method on UniToPatho dataset

Table 4.1.12 Performance results of the proposed model on different datasets

	Proposed Method			
Dataset	**Accuracy**	**Precision**	**Recall**	**F1**
Custom	95.00%	92.77%	95.06%	93.90%
UnitoPatho	90.00%	91.83%	89.10%	90.45%
EBHI	91.1%	88.74%	94.36 %	91.46%

4.1.6 Grad-CAM Results of the Proposed Method

To see the proposed models' class activations maps, gradient-weighted class activation mapping (Grad-CAM) method is employed [64]. Grad-CAM method explains the operation of a deep model by using the activation maps of a model in which the more focused regions are highlighted with a red color while the less attention grasping regions highlighted with yellow to blue colors. The Figure 4.1.10 and Figure 4.1.11 shows the Grad-CAM results of the proposed models for adenomatous and non-adenomatous tissues, respectively. As it can be seen from the figures, the proposed model focuses on the spatially distant cell structures to classify an image. Furthermore, Grad-CAM outputs of the proposed model for the pathological image can provide insight to a pathologist by explaining why an image is classified as adenomatous or non-adenomatous by the model.

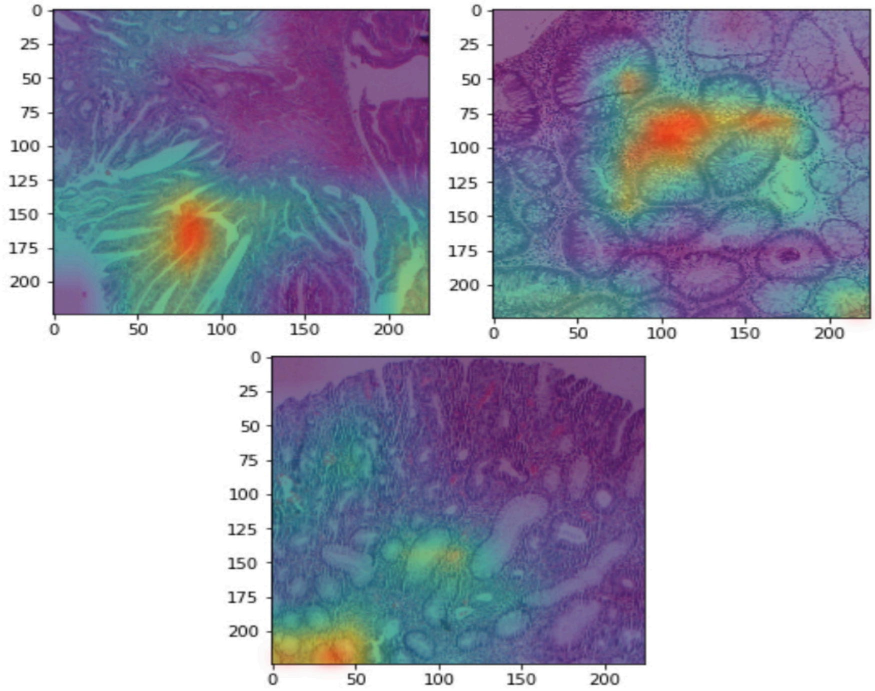


Figure 4.1.9 Grad-CAM results of adenomatous images

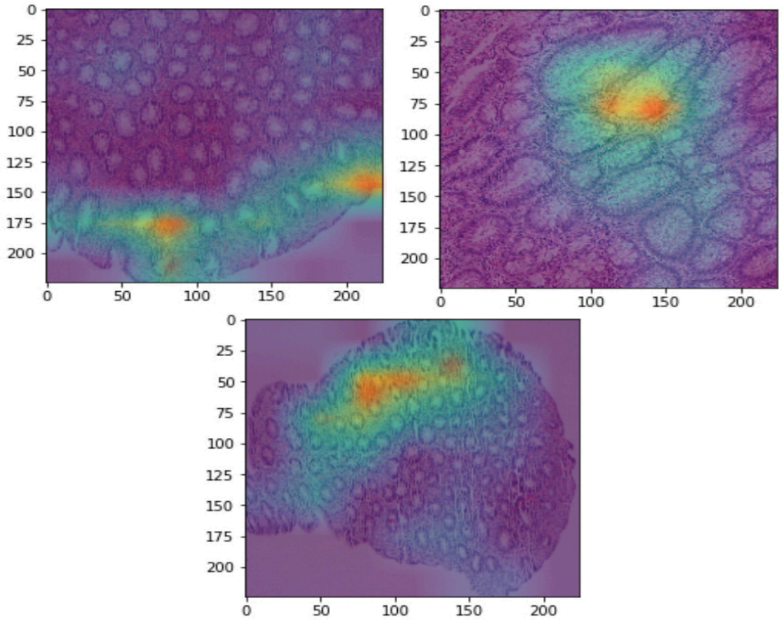


Figure 4.1.10 Grad-CAM results of non-adenomatous images

4.1.7 Conclusion of the Study

In this section, we propose an ensemble method which employs the recently proposed ConvNeXt variants for polyp classification on Stain-Net normalized histopathology images. The proposed method combines two separately fine-tuned ConvNeXt variants, namely ConvNeXt-Tiny and ConvNeXt-Base. The base models are tailored to the classification problem by network modifications at the image representation levels. The performance of the ensemble method is compared with the state-of-the-art deep CNN models and attention-based models on a custom colonic histological dataset. As a result, comprehensive experiments indicate that the ensemble of baseline models performs better than baseline models alone.

Ensemble methods are used in cancer classification tasks on breast, and colon histology images [50], [54], [65]–[68]. However, ensemble methods were not used in the colonic adenomatous polyp detection from the histology images. In the literature, researchers generally employ Deep CNN models alone. For example, in their work, Korbar et al. employed various ResNet-50 variants and selected the best-performed variant, which achieved 91.3% accuracy on their dataset [7]. Byeon et al. implemented EfficientNet for colon polyp subtype classification and achieved an overall F1 score of 98.8 on their dataset [16]. Iizuka et al. employed Inception-V3 to differentiate adenomatous, non-adenomatous and cancerous tissues on histopathology images and achieved an accuracy of 96% [69]. During the experimental setup, we implemented ResNet50, EfficientNet, Inception-v3 and compared the performance with our proposed method on the custom collected dataset. The proposed method achieves an accuracy of 93.75%, while ResNet50, EfficientNet and Inception-v3 achieve accuracies of 76.25%, 86.25% and 82.5% on the custom collected dataset, respectively. The gap in the model's performance on our custom-collected dataset and their dataset may be caused by the different domain distributions of the datasets.

In the literature, the models are generally tested against specific custom datasets. Since the models are developed for a specific dataset, they may not work well on the other datasets. Thus, this might be a drawback for real-world applications. To overcome this issue, researchers use publicly available datasets for benchmarking the models that are built for a custom dataset [6], [54], [55], [70], [71]. Therefore, in this work, additional experiments are conducted to explore the performance of the proposed model on two publicly available datasets, UniToPatho and EBHI. The proposed method outperforms the other methods by attaining 90% and 91.1% on UniToPatho and EBHI, while other methods in the literature achieve an accuracy of 64.29% and 66.55% on UniToPatho dataset [70], [71]. These accuracy results demonstrate that the proposed model has promising generalization ability for different datasets and has the potential to work in real-life scenarios.

To increase the models' generalization ability, stain normalization techniques are widely employed by researchers on HI. In contrast to the previous studies which make polyp classification on HI, in this study, stain normalization techniques are combined with an ensemble model. To the best of our knowledge, there is a limited number of studies which incorporates stain normalization methods for colon polyp classification on histopathology images [14]. In the literature, Perlo et al. use only Macenko normalization technique for polyp classification on histopathology images [14]. However, during the experiments, we observed that the performance of normalization techniques significantly differs for different classifiers. Therefore, combining various classifiers with different stain normalization techniques produced more beneficial outputs.

When it comes to medical image analysis, the black-box nature of the AI methods might restrict their usage in real applications. In recent years, this has sparked debates about the usage and necessity of explainability of opaque algorithms [72].There have been efforts to overcome this problem by introducing several tools for contemporary deep learning models [73]. The main approach to solve the problem is to provide the underlying reason for the decision as an auxiliary output to the clinician. This output could be either verbal or visual cues. This would be useful, especially when there is a mismatch between the clinician’s and the system’s decisions. In this case, the visual output might help to resolve the conflict. The clinician can check why the system diagnosed differently by evaluating the cues about the decision process of the system. As an interpretability method, the proposed system highlights the image regions which affected the decision most. As it was used in other medical image computing applications an attribution-based explainability method, Grad-CAM, is employed [74]. Various studies use Grad-CAM to assists during the decision-making process of pathologists [6]–[8], [14], [16], [69], [75], [76]. In their work, Wei et al. passed the Grad-CAM outputs to the expert pathologists to evaluate the models' performance to find a model that approximates most to the human interpreters [10]. Bilal et al. provided Grad-CAM outputs to experts for evaluation of their model [6]. Further, Perlo et al. used Grad-CAM method to provide the explainability of the model [14]. In their work, Iizuka et al. used Grad-CAM outputs to compare their models' performance with pathologists and medical school students [69]. Byeon et al. evaluated their model by using the Grad-CAM outputs for different polyp types [16]. Korbar et al.

annotated the region of interest using Grad-CAM [7]. Song et al. provided Grad-CAM results to pathologists for decision support [8]. In conclusion, while the black-box nature of AI methods in medical image analysis can limit their practical application, there have been various efforts to overcome this problem by introducing explainability tools such as Grad-CAM, which has been used in several studies to provide clinicians with visual and verbal cues about the decision process of the system, ultimately helping to resolve conflicts between the clinician's and the AI's diagnoses.

Chapter 5

5 Study 2

5.1 An Effective Colorectal Polyp Classification for Histopathological Images Based on Supervised Contrastive Learning

Clinical Decision Support System (CDSS) systems have been used to assist experts in decision-making to ease this labor-intensive work [77]. A benefit of this type of CDSS system is to help pathologists in these specific tasks. The number of histopathological analysis and cancer screening requests is rapidly increasing. According to Bilal et al. over the past decade, the number of pathological colon biopsy slide volumes has doubled [6]. With this rapidly expanding workload of experts, a computer-aided diagnosis system to automatize the differentiation of the polyp types is becoming the utmost important tool for the pathologist.

The development of the machine learning and deep learning algorithms brought a great interest in the research on the computer aided diagnosis systems for the medical image analysis. In addition, comprehensive studies such as colon adenocarcinoma classification [6], [12], [13], [15], [17], [18], [54], [69], [75], [78]–[80], colon polyp classification [7]–[11], [14], [76], [81], [82] and colon gland classification [78], [83], [84] are carried out on the individual diagnosis of colorectal cancer from histopathological images.

Contrastive learning (CLR) methods are commonly used in medical image classification tasks. For instance, Xu et al. used self-supervised contrastive pre-training to classify pleural effusion in chest X-ray images [85]. Similarly, Chen et al. employed an encoder trained with contrastive learning on public datasets to classify Covid-19 using chest X-rays [86]. Zhang et al. pre-trained medical image encoders with paired text data using contrastive loss[87]. In another work, Tian et al. proposed a Constrained Contrastive Distribution Learning approach to detect anomalies in medical images [88]. Azizi et al. used SimCLR with multiple instances of the same image for medical image classification [89] and compared it with baseline models such as Big Transfer (BiT) on various medical image datasets. Additionally, Stacke et al. evaluated the performance of contrastive learning methods on histology images, utilizing in-domain pre-training and ImageNet pre-training [26].

Azizi et al. utilized Big Transfer (BiT) as a baseline, along with traditional ResNet architectures, for medical image classification in mammography, chest X-rays, and dermatology images [90]. In another study, Galdran et al. proposed a methodology to enhance the classification performance in unbalanced medical image classification tasks by utilizing BiT [91]. Recently, Lu et al. aligned BiT with SimSiam and improved the classification performance in a skin cancer classification task [42]. Similarly, Shi et al. improved the classification performance in WSI classification of Eosinophilic esophagitis by using BiT [92]. Azizi et al. also presented a representation learning strategy for medical image classification, employing the weights of BiT as a backbone encoder [93].

Most of the colonic polyp classification methods employ a conventional supervised learning strategy. The downside of supervised learning is it requires an abundant amount of labelled data, which is exceptionally costly to obtain in the medical analysis field. The proposed method reduces the number of required labelled samples by applying Supervised Contrastive Learning (Sup-Con) approach on a different but same domain dataset prior to training the model for the downstream task. In this way, the representations are shifted into an in-domain space, which in turn gives quick learning and more accurate inference. To the best of the authors' knowledge, this study is the first to use Sup-Con Learning methodology with different state-of-the art CNN backbone architectures to classify colonic polyps on histopathology images The main contributions of this study can be summarized as follows:

- We develop an improved Sup-Con model and apply it for polyp classification on colon histopathology images for the first time in the literature. Unlike classical Sup-Con models, to increase the visual task adaptation, it uses a pre-trained BiT model as the encoder backbone rather than a conventional ResNet.

- To the best of our knowledge, this study is the first to comprehensively evaluate the performance of the Sup-Con method with various encoder structures for polyp classification problems on colon histopathology images.
- To support the experimental results empirically, a large custom data set is curated by experts and used in the experiments. Moreover, baseline tests on the data set using state-of-the-art pre-trained Deep CNN algorithms are provided for a fair comparison.
- Furthermore, this study is one of the first to investigate in-domain pre-training performance for the classification of colonic polyps during pre-training on a publicly available UnitoPatho database.
- In order to comprehensively evaluate and assess the generalizability of the proposed model, we also employ the publicly available UniToPatho database. The proposed Sup-Con model achieves an accuracy of 87% and 70.12% on our custom dataset and UnitoPatho, respectively. The accuracies on UnitoPatho are higher than other state-of-the-art methods.
- In the experiments, we also compare the performance of the proposed method and traditional Sup-Con on UnitoPatho and our custom data sets.

5.1.1 Material and Methods

5.1.1.1 Data Collection

Following the previous study, the number of patients in this study was increased, and the histological slides used were gathered from 184 patients who underwent colorectal cancer screening at Kayseri City Hospital in Turkey since May 2018.. For some cases, the patients had more than one polyp, even with different types such as hyperplastic and tubulovillous. Each tissue sample was examined under the microscopy with 5 different magnification levels. Samples of different magnification settings can be seen in Figure 5.1.1. The magnification levels are: x2.5, x5, x10, x20, and x40.

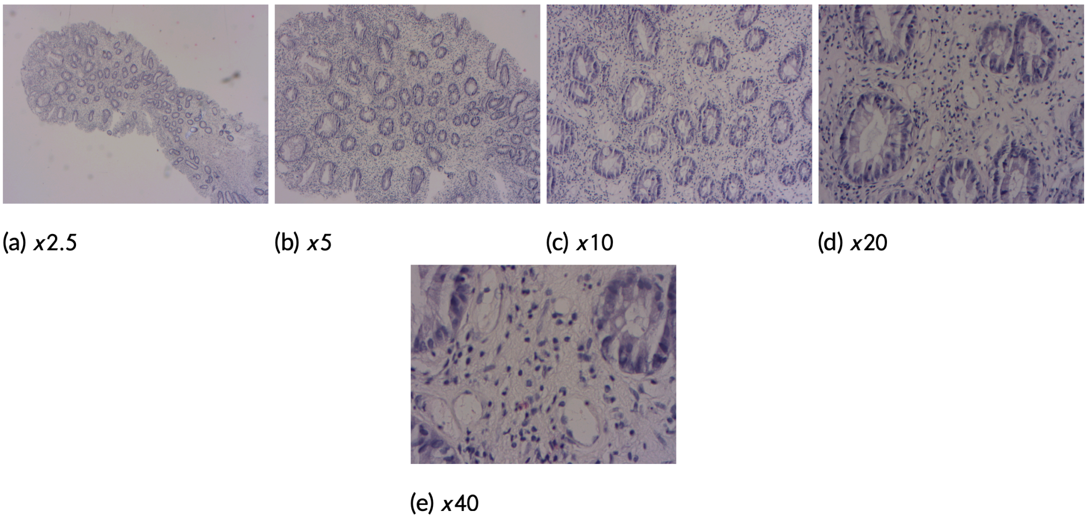

(a) *x*2.5 (b) *x*5 (c) *x*10 (d) *x*20 (e) *x*40

Figure 5.1.1 Different magnification levels of an adenomatous tissue

Out of the total 1056 samples used in this work, 359 slides belong to adenomatous polyps, including tubular, tubulovillous or villous types, while 181 slides belong to hyperplastic polyps. The detailed labeling of the samples was performed by two expert pathologists for each slide and magnification level. Additionally, at larger magnifications (i.e., x2.5 and x5), rectangular bounding boxes were manually annotated around the polyps as regions of interest. The final sample distribution is as follows: 346 samples belong to tubulovillous polyp type (TVA), 340 samples belong to tubular polyps (TBA), and 370 samples belong to hyperplastic polyps (HP).

Seven hundred and eighteen (718) samples were randomly selected for training, 157 for the validation set, and 181 as the test sample, detailed information is present in Table 5.1.1. Also, there is no overlap between samples from the same patient for training, validation, and test sets.

Table 5.1.1 Number of Samples for Each of the Classes and Training/Validation/Test sets used in this study

	Hyperplastic	**Tubular**	**Tubulovillous**	**Total**
Train	254	222	242	718
Validation	56	51	50	157
Test	60	67	54	181

5.1.2 Proposed Framework

Previous studies on polyp classification using histology images have predominantly utilized transfer learning methods. However, Raghu et al. conducted experiments demonstrating that the domain mismatch between medical and natural images can impede transfer learning [94]. In contrast, recent works have shown that the BiT method improves the performance of transfer learning approaches on medical image classification tasks due to its exceptional performance in domain adaptation. Additionally, the Sup-Con loss function accelerates the learning of hard negatives and hard positives, and the custom-collected histology database used in this study contains challenging samples.

The proposed framework was designed by combining the BiT and Supervised Contrastive (Sup-Con) Learning frameworks, as illustrated in Figure 5.1.2. Since our database contains a limited number of samples and BiT performs well on a low data regime, BiT was used as the encoder for Sup-Con. Additionally, the downstream components of BiT were designed to facilitate visual task adaptation. In the first stage, the encoder was trained in 100 epochs using the Sup-Con loss (Equation 28) with a batch size of 16. As demonstrated in Equation 28, the query is compared against the set of samples that are in the same class. After this training, the frozen representations were forwarded to the second stage. The second stage was designed as follows: the encoder was followed by a dropout layer, a fully connected layer using L2 kernel regularizers, another dropout layer, and a final Softmax layer with the target classes. During the second stage, stochastic gradient descent (SGD) optimizer with a learning rate of 0.001 and momentum of 0.9 was employed, and the adaptive momentum optimization algorithm optimized the learning rate during training. Categorical cross-entropy was used as the loss function with label smoothing having a value of 0.1. Label smoothing regularizes the network from choosing a class with high confidence.

In addition, we compared the proposed method against extensively used pre-trained Deep CNN methods. We employed these pre-trained methods as the encoder of the proposed Sup-Con framework and compared the performance of Supervised Learning and Supervised Contrastive Learning. We also compared the performance of the proposed method and traditional Sup-Con. Additionally, we conducted hyperparameter optimization to explore the model's performance for different settings of the hyperparameters.

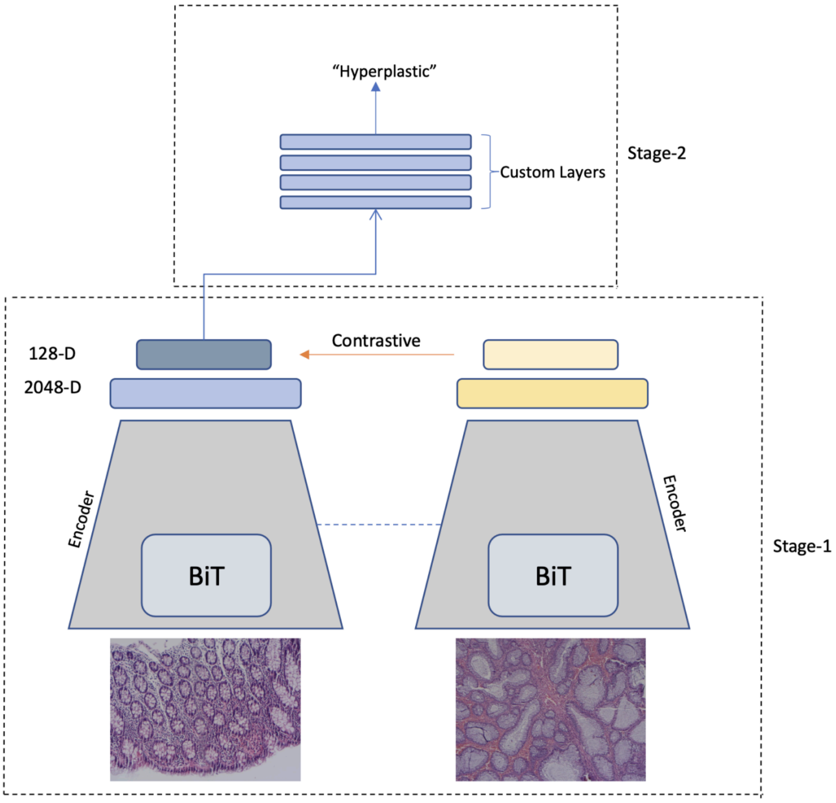


Figure 5.1.2 Proposed framework of this study

5.1.3 Experimental Setup

Figure 5.1.2 illustrates the setup of the proposed framework. The BiT-M model is used as the encoder of the Sup-Con method. In the first stage, we trained the encoder for 100 epochs using the Sup-Con loss with a batch size of 16. The output representations of the first stage were then forwarded to a fully connected layer that uses L2 kernel regularizer, and the classification was performed by a SoftMax layer with the target classes. In the second stage, stochastic gradient descent (SGD) optimizer with a learning rate of 0.001 and momentum of 0.9 was employed, and the adaptive momentum optimization algorithm optimized the learning rate during training. Categorical cross-entropy with label smoothing of 0.1 was used as the loss function. Additionally, we compared the proposed method against extensively used pre-trained Deep CNN methods. Furthermore, we employed these pre-trained methods as the encoder of the proposed Sup-Con framework and compared the performance of Supervised Learning and Supervised Contrastive Learning. As explained in section 5.1.3, the proposed Sup-Con framework is a modified version of the traditional Sup-Con. During the experiments, we also compared the performance of the proposed method, and the traditional Sup-Con. Hyperparameter optimization was implemented to explore the model's performance for different settings of the following hyperparameters:

- Learning rate: 0.0001, 0.0005, 0.001, 0.005, 0.01

- Optimizer: ADAM, SGD, RMSprop, AdaGrad

- Temperature Value: 0.03, 0.05, 0.08, 0.1

Additionally, to observe the effect of in-domain pre-training, we changed the pre-training set. For these ablation tests, we used standard performance measures such as weighted average and class-based F1-score, precision, recall, and overall accuracy. Finally, to verify the proposed model's generalizability, we tested it on a publicly available UnitoPatho database. The models used for the ablation study were employed from TensorFlow, including BiT-M (trained on ImageNet-21k), DenseNet-201 (trained on Imagenet), Inception-V3 (trained on ImageNet), ResNetV2-50 (trained on ImageNet), InceptionResNet-v2 (trained on ImageNet), and Xception (trained on ImageNet). All the employed models were pre-trained because building them from scratch requires a significant amount of data. Furthermore, we used a fine-tuning approach for the comparison of supervised learning because state-of-the-art CNN models are pre-trained on natural images, whereas our images belong to a different domain. Thus, to comprehensively compare the methods, we first fine-tuned them on our histological dataset. Additionally, all the above-mentioned pre-trained models were used as encoders during the first stage of the Sup-Con framework. To use them as encoders, the models were pre-trained on histology images using a Supervised Contrastive loss. All experiments were performed on the Google Colab platform with 52 GB of RAM and NVIDIA Tesla K80, NVIDIA Tesla T4, and NVIDIA Tesla P100 GPU accelerators. The experiments were implemented with Python v3.7.13 using the TensorFlow v2.8.0 framework.

5.1.4 Results and Discussions of the Chapter

The aim of this study was to train a customized BiT model using a modified version of the Sup-Con Learning framework to classify colorectal polyps using histopathological images. To achieve this, we first used the above-mentioned classifiers to build a baseline.

We implemented ResNet-50 as proposed by Korbar et al. [7], and other deep learning approaches were also utilized. The test performance of each model is presented in Table 5.1.2 and Figure 5.1.4.

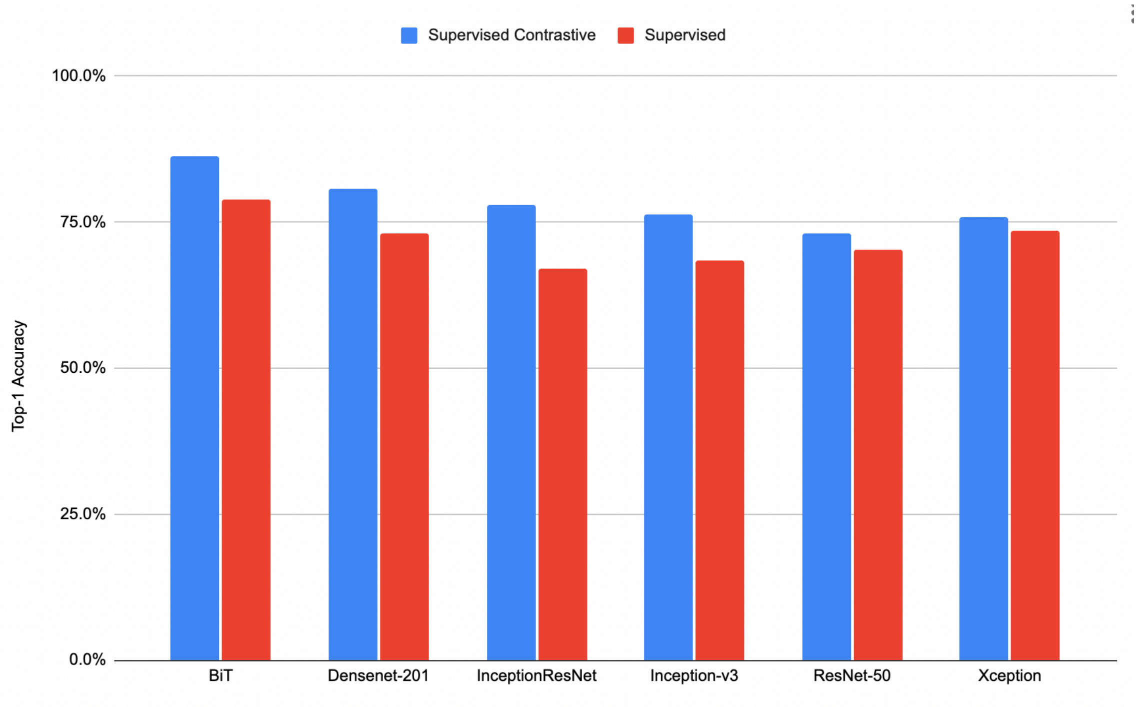


Figure 5.1.3 Accuracy Comparison of the Supervised and Supervised Contrastive Learning with Different Classifiers

In Table 5.1.2, Overall Accuracy is present for each model and precision, recall and F1 scores are given for each class. Additionally, a weighted average of the precision, recall and F1 scores are present for each model on the row that mentions the model’s name.

Table 5.1.2 Accuracy results for Supervised Learning and modified Supervised Contrastive Learning with various Deep CNN models

Supervised vs Supervised Contrastive Learning	**Supervised Contrastive**				**Supervised**			
MODEL	**Overall Accuracy**	**Precision**	**Recall**	**F-1**	**Overall Accuracy**	**Precision**	**Recall**	**F1**
BiT	**86.19%**	86.34%	86.19%	86.09%	78.91%	78.89%	79.16%	78.95%
HP		83.58%	93.33%	88.19%		78.18%	80.37%	79.26%
TBA		87.88%	86.57%	87.22%		81.03%	75.20%	78.01%
TVA		87.50%	77.78%	82.35%		77.01%	82.72%	79.76%
DenseNet-201	**80.69%**	80.17%	81.59%	80.44%	73.02%	72.98%	73.04%	73.00%
HP		87.85%	80.34%	83.93%		72.73%	74.07%	73.39%
TBA		80.77%	75.68%	78.14%		73.77%	72.00%	72.87%
TVA		70.89%	90.32%	79.43%		72.29%	73.17%	72.73%
Inception-ResNet	77.93%	78.10%	77.99%	77.94%	67.01%	67.51%	67.79%	67.19%
HP		74.77%	81.63%	78.05%		58.33%	70.79%	63.96%
TBA		80.77%	74.34%	77.42%		74.04%	60.63%	66.67%
TVA		78.48%	78.48%	78.48%		69.62%	73.33%	71.43%
Inception-v3	76.21%	76.32%	76.26%	76.20%	68.33%	68.36%	68.54%	68.28%
HP		73.83%	79.80%	76.70%		66.04%	70.71%	68.29%
TBA		78.85%	72.57%	75.58%		65.09%	69.70%	67.32%
TVA		75.95%	76.92%	76.43%		75.00%	64.71%	69.47%
ResNet-50	73.10%	73.26%	73.06%	73.07%	70.24%	70.52%	70.41%	70.24%
HP		70.09%	77.32%	73.53%		66.04%	75.27%	70.35%
TBA		75.96%	71.82%	73.83%		71.43%	71.43%	71.43%
TVA		73.42%	69.88%	71.60%		74.36%	63.74%	68.64%
Xception	75.86%	75.92%	75.82%	75.81%	73.45%	73.57%	73.43%	73.42%
HP		74.77%	80.00%	77.29%		71.70%	76.77%	74.15%
TBA		76.92%	72.73%	74.77%		73.33%	74.04%	73.68%
TVA		75.95%	75.00%	75.47%		75.95%	68.97%	72.29%

As it is illustrated in the Table 5.1.2, and Figure 5.1.4 the proposed method increases the overall accuracy for the BiT by 8%, for Inception-ResNet by 10%, for DenseNet201 by 5%, for Inception-v3 by 8%, for ResNet-50 by 3%, and for Xception by 2%. The proposed method on our custom dataset performs the best when BiT-M is used as the encoder and achieves the highest accuracy, weighted average precision, weighted average recall, and weighted average F1 score with values of 86.2%, 86.34%, 86.19% and 86.09%, respectively. On the other hand, performance of the model is relatively poor when ResNet-50 is used as the encoder with an accuracy, weighted average precision, weighted average recall, and weighted average F1 score with values of 73.10%, 73.26%, 73.06% and 73.07%, respectively. The structure of BiT-M is almost identical to ResNet-50, but the difference lies in the fact that BiT-M is an improved version of ResNet-50, specifically designed to facilitate domain adaptation.

Furthermore, to compare different pre-training sets performance, we employed three different settings. During the first experimental setting, we employed publicly available UniToPatho dataset as the pre-training dataset, for the second experimental setting we utilized ImageNet trained models and fine-tuned them on our custom dataset. Finally, we employed ImageNet pre-trained models to fine-tune the models with UniToPatho database. As classifiers, three top performed models of the Sup-Con methodology from the Table 5.1.2 are employed. The experimental result for this comparison can be seen in the Table 5.1.3 The results show that the accuracies of the ImageNet pre-trained DenseNet, Inception-ResNet-v2 is improved by 2% when ImageNet trained models are pre-trained with UniToPatho database. However, for the same experiment, performance of the BiT-M is decreased by 4%. The reason for this decrease might be the samples of UniToPatho are cropped from only one magnification of a WSI, while our custom dataset contains WSI of four different magnification levels. Moreover, the number of samples for tubulovillous polyp type, tubular polyps, and hyperplastic polyp types are unbalanced for the UniToPatho, this might severely affect the performance of the BiT. On the other hand, ImageNet pre-trained BiT-M achieves the best accuracy of 86.2%.

Table 5.1.3 Accuracy results for modified Supervised Contrastive Learning with various encoders that are pre-trained on different pre-train dataset

	UniToPatho				**ImageNet**				**UniToPatho + ImageNet**			
Model	**Ov. Acc. (%)**	**Pre. (%)**	**Rec (%)**	**F1**	**Ov. Acc. (%)**	**Pre. (%)**	**Rec (%)**	**F1**	**Ov. Acc. (%)**	**Pre. (%)**	**Rec (%)**	**F1**
DenseNet-201	63.10	62.45	64.44	62.81	80.69	80.17	81.59	80.44	**82.07**	82.06	82.71	82.31
HP		73.83	58.09	65.02		87.85	80.34	83.93		83.18	80.18	81.65
TBA		57.69	66.67	61.86		80.77	75.68	78.14		78.85	76.64	77.73
TVA		55.70	68.75	61.54		70.89	90.32	79.43		84.81	93.06	88.74
InceptionResNet-v2	72.41	71.97	72.97	72.03	77.63	77.76	78.12	77.78	80.69	80.30	81.11	80.47
HP		82.24	69.29	75.21		75.89	80.95	78.34		88.79	77.87	82.97
TBA		60.58	74.12	66.67		78.85	70.09	74.21		72.12	81.52	76.53
TVA		74.68	75.64	75.16		78.48	84.93	81.58		81.01	84.21	82.58
BiT-M	79.58	78.85	82.26	79.11	**86.19**	86.34	86.19	86.09	**82.41**	82.17	82.30	60.00
HP		88.68	74.02	80.69		83.58	93.33	88.19		85.98	85.98	93.12
TBA		84.62	78.57	81.48		87.88	86.57	87.22		81.73	78.70	94.60
TVA		60.76	96.00	74.42		87.50	77.78	82.35		78.48	82.67	99.13

Additionally, the hyperparameter optimization is implemented on the proposed method. Therefore, various learning rates, optimization methods and temperature values of the Sup-Con is implemented. In Figure 5.1.6, a standard box plot shows the Top-1 accuracy change for hyperparameters, learning rate, optimization method and temperature value. It can be observed that, the variance of the accuracy is low for the proposed method. Moreover, the best accuracy of 87.1% for the proposed method is achieved for the following parameters: Learning rate of 0.0005, Adam Optimizer, and temperature value of 0.05.

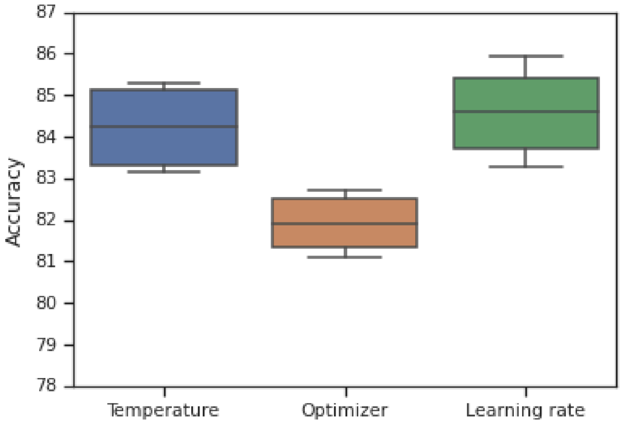


Figure 5.1.4 Hyperparamater stability analysis of the proposed method

5.1.5 Generalization Test

In order to assess the generalization ability of the proposed model, both the UniToPatho and custom collected datasets were employed. The performance of the proposed method was compared to that of the traditional Sup-Con method, as proposed by Khosla et al. [41], using both datasets. The models were fine-tuned separately on the two datasets. The results are presented in Table 5.1.4, and the ROC curves of the proposed method on the custom database and UniToPatho are shown in Figures 5.1.7 and 5.1.8, respectively. Additionally, the classification confusion matrices of the models for each dataset are displayed in Figure 5.1.9. The proposed method achieves an improvement of 6% and 9% in accuracy for the UniToPatho and custom databases, respectively, compared to the traditional Sup-Con method. Moreover, the proposed method outperforms other methods in the literature on the UniToPatho dataset, as those methods achieved accuracies of 64.29% and 66.55% [70], [71].

Table 5.1.4 Accuracy results for proposed model and traditional Sup-Con on custom collected dataset and UnitoPatho for generalization test

	Collected Dataset				**UniToPatho**			
Model / Per Class	**Overall Accuracy**	**Precision**	**Recall**	**F1-Score**	**Overall Accuracy**	**Precision**	**Recall**	**F1-Score**
Traditional Supervised Contrastive Learning	75.17%	75.12%	76.04%	75.16%	63.95%	69.89%	63.98%	65.24%
HP		73.83%	77.45%	75.60%		88.24%	90.91%	89.55%
TBA		81.73%	69.11%	74.89%		74.04%	50.33%	59.92%
TVA		68.35%	83.08%	75.00%		32.74%	55.22%	41.11%
Proposed Method	**86.19%**	86.34%	86.19%	86.09%	70.13%	71.77%	70.13%	70.28%
HP		83.58%	93.33%	88.19%		70.59%	72.73%	71.64%
TBA		87.88%	86.57%	87.22%		78.57%	64.71%	70.97%
TVA		87.50%	77.78%	82.35%		57.78%	78.79%	66.67%

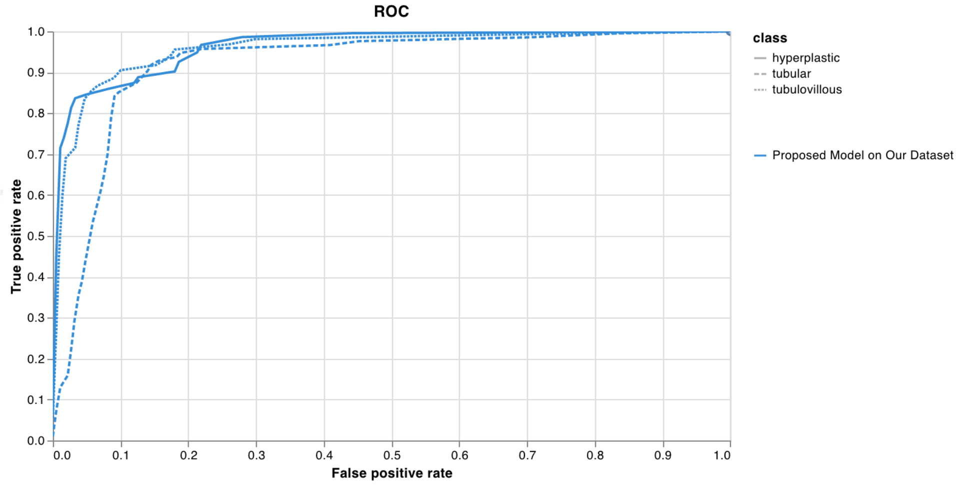


Figure 5.1.5 ROC curve of the proposed method on our dataset

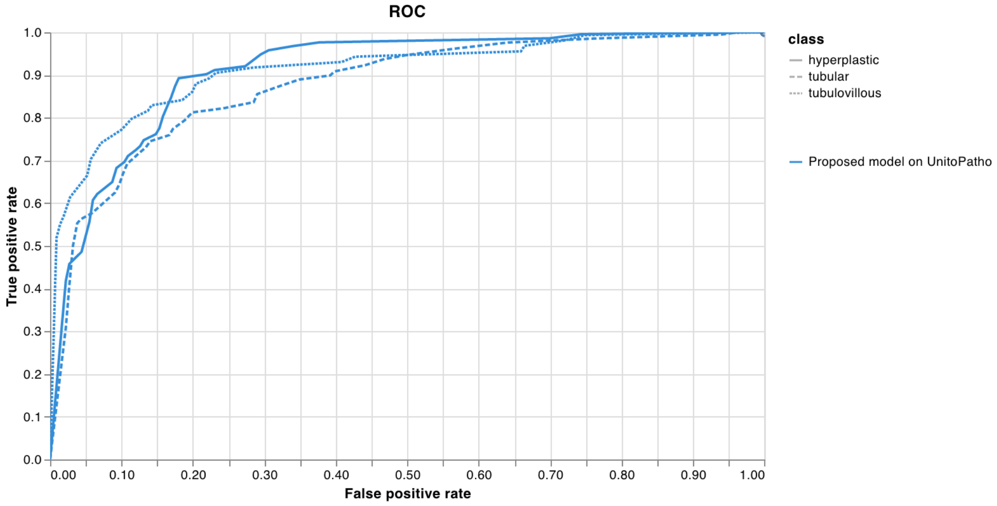


Figure 5.1.6 ROC curve of the proposed method on UnitoPatho

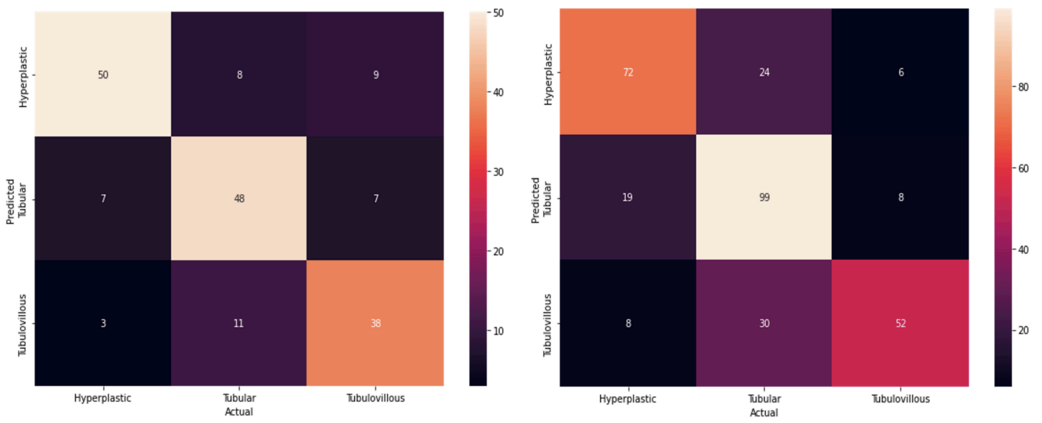


Figure 5.1.7 Confusion matrix of the proposed method on our custom collected database (left) and UnitoPatho (right)

5.1.6 Conclusion of the Study

In this study, we proposed a novel approach for polyp classification in histopathology images by combining the improved Supervised Contrastive (Sup-Con) Learning and Big Transfer (BiT) methodologies. Our tailored version of Sup-Con employed BiT-M architecture as an encoder and achieved superior performance in all metrics when compared to state-of-the-art Deep CNN models in a supervised setting. Furthermore, we compared the performance of Sup-Con and traditional supervised learning and found that Sup-Con improved the overall accuracy by 5%.

The generalization ability of the proposed method is tested on the publicly available UniToPatho database. The proposed model achieves the accuracies of 87.1% and 70.3% for the custom collected dataset and UniToPatho dataset, respectively, outperforming other methods in the literature on UniToPatho dataset.

Additionally, we explore the performance of various pre-training settings, which includes ImageNet pre-trained models as well as in-domain pre-trained models. To the best of our knowledge, this work is one of the first to explore in-domain pre-training with a publicly available dataset for polyp classification problem on histopathology images. The comprehensive experiments demonstrate that the performance of Deep CNN models increased when the ImageNet pre-trained models are fine-tuned on UniToPatho dataset. However, ImageNet pre-trained BiT model still achieves even higher accuracy. This shows the visual task adaptation performance of the BiT model on the proposed method.

The performance of the Deep CNN models decreased when they are pre-trained from scratch by only using UniToPatho dataset. This is due to the fact that samples of the UniToPatho database are extracted from a fixed magnification level, while our custom dataset contains samples of different magnification levels.

As a future work, other publicly available datasets which contains different magnification levels can be used as pre-training datasets. The grade of the polyps can be provided as diagnostic information for treatment in addition to the polyp classification.

Chapter 6

6 Study 3

6.1 Self-Supervised Contrastive Learning for Classification of Colon Polyps on Histopathology Images

The field of medical image analysis faces challenges in obtaining a large number of labeled medical images, as labeling is a time-consuming and laborious task. Additionally, on-site labeling is often necessary due to patient confidentiality concerns. Self-supervised learning methods offer a potential solution, as they can make use of unlabeled data. Previous research has demonstrated the effectiveness of self-supervised pre-training approaches for medical image classification [44]. Self-supervised contrastive learning is one such approach, where an augmented version of an image is used to learn the latent features of the image. The model is trained so that two different augmentations of the same image should have similar representations.

In previous studies of this thesis, supervised learning was performed on the labeled datasets. However, in this study, we aimed to address a significant challenge in the field of medical image analysis - namely, how to effectively use unlabeled data when only a limited number of labeled images are available for colon histopathology image classification. We believe that this is a crucial issue that needs to be addressed, and our study seeks to explore the use of self-supervised contrastive learning as a potential solution. By employing Task-specific Self-supervised Contrastive Learning (SSL) and pre-training Deep CNN models on the publicly available UniToPatho dataset without labels, we were able to fine-tune the models with our custom-collected data with labels and investigate the effectiveness of self-supervised contrastive learning for polyp classification. Furthermore, we varied the contrastive learning algorithms by employing different setups of the contrastive model, including SimCLR, SimSiam, and Barlow Twins, with multiple backbone models such as ResNet-18, ResNet-50, and EfficientNet, to gain insights into which method worked best for our application.

Self-supervised contrastive learning models have been widely used in medical image analysis tasks. For instance, Tellez et al. extracted patches from whole slide images and applied augmentations to learn the representations [95]. Azizi et al. used a SimCLR approach with multiple instances of the same image and compared it to baseline models such as BiT on various medical image datasets [89]. Stacke et al. examined the performance of contrastive learning approaches on histology images using in-domain pre-training and ImageNet pre-training [26]. Ciga et al. used histopathology images from different organs for self-supervised contrastive learning [44].

6.2 Methodology

For the methodology applied in this study, it is worth highlighting the use of three distinct algorithms, namely SimSiam, SimCLR, and Barlow Twins, in creating the contrastive model, as well as the adoption of multiple backbone models such as ResNet-18, ResNet-50, and EfficientNet. Additionally, to compare the efficacy of self-supervised learning versus supervised learning, the backbone models were pre-trained on the publicly available UniToPatho dataset in an unsupervised setting and subsequently fine-tuned on our dataset. An overview of the methodology is shown in Figure 6.2.1.

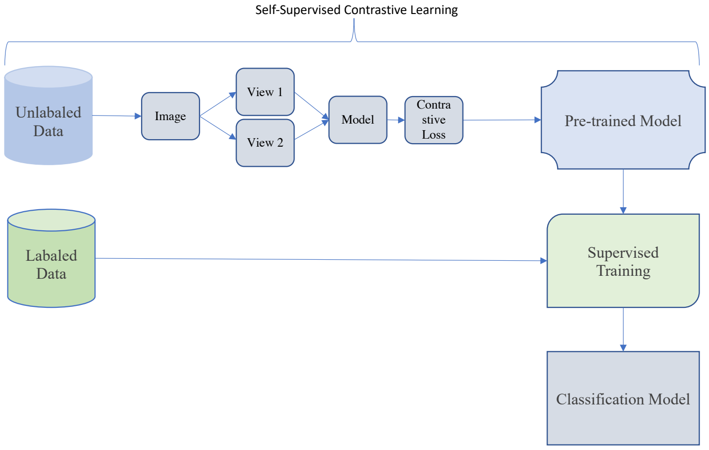


Figure 6.2.1 Self-supervised learning framework

6.3 Results and Discussions

The rationale behind this study is investigation of how to effectively utilize unlabeled data when there are limited labeled images available, as the labeling process is time-consuming and labor-intensive. Moreover, due to the confidentiality of patient data, labeling is often done on-site. Therefore, this study aims to explore the performance of SLL algorithms, which leverage both labeled and unlabeled data. The performance results of this study are presented in Tables 6.3.1, 6.3.2, and 6.3.3. Table 6.3.1 shows the results of various contrastive learning algorithms for different backbone classifiers, namely ResNet-18, ResNet-50, and EfficientNet. The overall accuracy, Precision, Recall, and F-1 scores for each class are presented for each model, and a weighted average is given for each model's performance. The results show that ResNet-18 performs the best with an overall accuracy of 66.21% for the SimCLR algorithm. ResNet-50 achieves an overall accuracy of 56.21% for the SimSiam and Barlow-Twins algorithms. However, the performance of EfficientNet is below average, which might be due to the unbalanced nature of the UniToPatho dataset. EfficientNet has a Precision, Recall, and F-1 score of 0% for the hyperplastic class, which is the minority class. This suggests that the classifier may be adversely affected if the unlabeled data used for self-supervised pretraining is unbalanced. Therefore, it may be necessary to use simpler backbone models when employing self-supervised learning, as the nature of the unlabeled data is often unknown.

Table 6.3.1 Accuracy results of the backbone models for different contrastive learning algorithms

	SimCLR				**SimSiam**				**Barlow Twins**			
Model	**Ov. Acc.**	**Pre.**	**Rec.**	**F1**	**Ov. Acc.**	**Pre.**	**Rec.**	**F1**	**Ov. Acc.**	**Pre.**	**Rec.**	**F1**
ResNet-18	**66.21%**	**67.30%**	**66.25%**	**65.89%**	51.03%	54.13%	52.63%	47.20%	54.14%	56.02%	54.50%	52.87%
HP		73.33%	61.68%	67.01%		63.33%	17.76%	27.74%		52.38%	41.12%	46.07%
TBA		59.85%	78.85%	68.05%		49.73%	87.50%	63.41%		51.88%	79.81%	62.88%
TVA		69.84%	55.70%	61.97%		49.35%	48.10%	48.72%		65.22%	37.97%	48.00%
ResNet-50	55.56%	57.44%	55.78%	54.88%	**56.21%**	**58.89%**	**56.27%**	**55.53%**	**56.21%**	**58.24%**	**56.40%**	**55.63%**
HP		54.74%	48.60%	51.49%		55.21%	49.53%	52.22%		56.04%	47.66%	51.52%
TBA		52.41%	74.51%	61.54%		52.35%	75.00%	61.66%		52.35%	75.00%	61.66%
TVA		66.67%	40.51%	50.39%		71.11%	40.51%	51.61%		68.00%	43.04%	52.71%
Efficient Net	39.20%	27.25%	39.43%	32.19%	41.38%	29.06%	43.73%	34.60%	41.38%	29.35%	43.78%	34.70%
HP		0.00%	0.00%	0.00%		0.00%	0.00%	0.00%		0.00%	0.00%	0.00%
TBA		49.65%	68.27%	57.49%		54.07%	70.19%	61.09%		54.62%	68.27%	60.68%
TVA		29.75%	47.47%	36.58%		30.32%	59.49%	40.17%		30.63%	62.03%	41.00%

In addition, we conducted experiments to compare the performance of self-supervised pre-training with training the backbone models from scratch on our custom collected data. The results are presented in Table 6.3.2. It is observed that self-supervised pre-training improved the accuracy of ResNet-18 by 16.9%, while the performance of ResNet-50 and EfficientNet are comparable with supervised learning. These results suggest that utilizing the unlabeled data with self-supervised pre-training can improve the performance, especially for ResNet-18, compared to training the models from scratch on a limited number of samples.

Furthermore, the performance of self-supervised learning was compared with supervised learning using transfer learning and fine-tuning on the ImageNet pre-trained classifiers. The results are presented in Table 6.3.3. Fine-tuning the ImageNet pre-trained models produced promising results, with the best accuracy of 75.17% achieved when EfficientNet was fine-tuned on the custom collected dataset. Transfer learning, on the other hand, did not perform well due to domain mismatch. However, fine-tuning handled domain adaptation well. It can be concluded that SSL outperforms transfer learning since there is no domain mismatch between the pre-training dataset and the downstream task of interest, as opposed to transfer learning where there may be domain mismatch.

Table 6.3.2 Accuracy results of the backbone models that are trained from scratch on labeled data

Model	Overall Accuracy	Precision	Recall	F1
ResNet-18	49.31%	53.48%	50.09%	48.85%
HP		69.39%	31.78%	43.59%
TBA		56.92%	71.15%	63.25%
TVA		31.53%	44.30%	36.84%
ResNet-50	**56.55%**	**58.26%**	**56.69%**	**55.79%**
HP		57.14%	48.60%	52.53%
TBA		53.02%	75.96%	62.45%
TVA		66.00%	41.77%	51.16%
EfficientNet	42.07%	29.92%	44.53%	35.28%
HP		0.00%	0.00%	0.00%
TBA		55.47%	68.27%	61.21%
TVA		31.48%	64.56%	42.32%

Table 6.3.3 Accuracy results of transfer learning and fine tuning for the pre-trained backbone models

	Transfer Learning				Fine Tuning			
Model	Overall Accuracy	Precision	Recall	F1	Overall Accuracy	Precision	Recall	F1
ResNet-18	50.69%	59.49%	48.87%	47.30%	71.38%	72.13%	71.44%	70.97%
HP		45.88%	83.18%	59.14%		70.80%	74.77%	72.73%
TBA		80.00%	26.92%	40.29%		76.92%	57.69%	65.93%
TVA		49.18%	37.97%	42.86%		67.68%	84.81%	75.28%
ResNet-50	27.15%	8.10%	29.83%	12.74%	74.48%	74.99%	74.30%	74.01%
HP		0.00%	0.00%	0.00%		73.73%	81.31%	77.33%
TBA		0.00%	0.00%	0.00%		78.75%	60.58%	68.48%
TVA		27.15%	100.00%	42.70%		71.74%	83.54%	77.19%
EfficientNet	45.17%	48.90%	45.21%	45.48%	**75.17%**	**75.60%**	**75.50%**	**75.24%**
HP		48.54%	46.73%	47.62%		77.55%	71.03%	74.15%
TBA		61.54%	38.46%	47.34%		80.21%	74.04%	77.00%
TVA		33.61%	51.90%	40.80%		67.71%	82.28%	74.29%

6.4 Conclusions of the Study

In this study, we investigated the effectiveness of task-specific Self-Supervised Learning (SSL) in leveraging unlabeled data. To comprehensively evaluate the performance of contrastive learning algorithms, experiments were conducted on different backbone classifiers, including ResNet-18, ResNet-50, and EfficientNet. The results showed that ResNet-18 achieved the best overall accuracy of 66.21% using the SimCLR algorithm, while ResNet-50 achieved an overall accuracy of 56.21% with the SimSiam and Barlow-Twins algorithms. However, Efficient Net's performance was below average, possibly due to the imbalanced nature of the UniToPatho dataset. Therefore, simpler backbone models may be necessary when using self-supervised learning since the unlabeled data's nature is usually unknown.

To compare self-supervised pre-training with training backbone models from scratch on custom collected data, experiments were conducted. The results showed that self-supervised pre-training improved ResNet-18's accuracy by 16.9%, while ResNet-50 and EfficientNet's performance were comparable with supervised learning. Hence, utilizing unlabeled data with self-supervised pre-training can improve performance, particularly for ResNet-18.

Moreover, the performance of self-supervised learning was compared with supervised learning using transfer learning and fine-tuning on ImageNet pre-trained classifiers. Fine-tuning the ImageNet pre-trained models produced promising results, with EfficientNet achieving the best accuracy of 75.17% on the custom collected dataset. On the other hand, transfer learning did not perform well due to domain mismatch. It can be concluded that SSL outperforms transfer learning since there is no domain mismatch between the pre-training dataset and the downstream task of interest. However, the SSL's performance was challenged when compared to fine-tuning with ImageNet pre-trained models. As mentioned in the literature, the limited number of labels available for generating accurate pseudo-labels reduces the effectiveness of SSL and can lead to misleading training of the backbone network [96], [97]

Chapter 7

7 Conclusions and Future Prospects

7.1 Conclusions

Histopathology images are crucial for distinguishing adenomatous polyps from non-adenomatous tissues, such as hyperplastic polyps, inflammation, and normal tissue. However, this classification heavily relies on the expertise level of pathologists and is a time-consuming process. An automated system that can accurately distinguish between histopathology images would therefore be advantageous. Although there have been numerous studies on polyp classification using pathology images in computer-aided diagnosis systems, there is no complete solution available yet. Most researchers propose methods that are specific to their particular datasets. In this thesis, different deep learning methods and frameworks are developed for the automatic classification of adenomatous polyps and polyp types. The histological slides used in this study were collected from Kayseri City Hospital, Kayseri, Turkey, and included samples from 182 patients who underwent colorectal cancer screening since May 2018, resulting in a total of 359 slides belonging to adenomatous polyps (tubular, tubulovillous/villous) and 181 slides belonging to hyperplastic polyps. It is important to note that new data was collected during this thesis work, with the first study involving 82 patients and the last two studies involving 182 patients. To ensure accurate labeling, two expert pathologists labeled each slide at different magnification levels. Additionally, rectangular bounding boxes were manually annotated around the polyps at larger magnifications (i.e., x2.5 and x5) to indicate the region of interest. The dataset is composed of 346 samples belonging to tubulovillous polyp type (TVA), 340 samples belonging to tubular polyps (TBA), and 370 samples belonging to hyperplastic polyps (HP), resulting in a total of 1056 samples used in this study. Additionally, in order to evaluate the generalization ability of the developed frameworks, two publicly available datasets were utilized. These datasets were used to test the performance of the models on histopathology images beyond the custom-collected dataset from Kayseri City Hospital.

The differentiation of adenomatous polyps from non-adenomatous tissues on histopathology images is a key diagnostic challenge in the clinical workflow of polyp classification. To address this, firstly, we developed a computer-aided diagnosis system for automatic detection of adenomatous polyps on colon histopathology images. Our proposed method integrated stain normalization techniques with ensemble variants of ConvNeXt, a recent and prominent convolutional deep learning architecture, and achieved improved generalization by using various stain normalization techniques. The proposed method also included network modifications at the image representation levels to tailor it to the problem. We evaluated the classification performance of the proposed method on three datasets and found that it outperformed state-of-the-art deep convolutional neural network models on our dataset. The model achieved an accuracy of 95% on our dataset and 91.1% and 90% on EBHI and UniToPatho datasets, respectively. Additionally, we investigated the Grad-Cam results of the proposed model, which revealed regions where cancer indicators potentially reside, demonstrating the model's high generalization ability across different datasets.

After the work on automatic detection of adenomatous polyps, a computer-aided diagnosis system was developed for multi-class classification of hyperplastic, tubular, and tubulovillous/villous polyps using colon histopathology images. The proposed framework combined Supervised Contrastive Learning with Big Transfer to improve learning of hard positives and negatives by leveraging the visual task adaptation of Big Transfer. The model outperformed state-of-the-art Deep Convolutional Neural Network models, achieving accuracies of 87.1% and 70.3% on our dataset and UniToPatho dataset, respectively, demonstrating its generalization ability. A comparison of the proposed method with traditional supervised contrastive learning algorithm on our dataset and UniToPatho dataset showed that the proposed model performed better by utilizing visual domain adaptation.

In the final study, the aim was to address the challenge of utilizing unlabeled images along with limited labeled data. To overcome this issue, a self-supervised contrastive learning approach was employed for pre-training of the model on the unlabeled data. Furthermore, the performance of the backbone models of SSL was compared with fine-tuning/transfer learning of the ImageNet pre-trained backbone models. The results showed that SSL outperformed transfer learning with ImageNet pre-trained models due to the absence of domain mismatch, and the downstream task of interest being the same for SSL.

Despite the promising results achieved in this thesis, there are still several challenges in the area of automatic classification of histopathology images of colorectal polyps. One major challenge is the lack of standardized datasets and benchmarks for evaluating different methods, which can hinder the comparison and reproducibility of results across studies. Another challenge is the interpretability of deep learning models, which can be critical in medical applications where decision-making must be transparent and explainable. Additionally, the high variability in histopathology images due to factors such as staining techniques and tissue preparation can pose a challenge for developing robust and generalizable models. Addressing these challenges will require collaboration between the medical and computer science communities, as well as the development of standardized datasets and tools for model interpretation and evaluation.

In conclusion, this study presents a deep learning-based approach for the automated classification of colorectal polyps using histopathology images. The results demonstrate the potential of the proposed method in accurately differentiating between adenomatous and non- adenomatous polyps and multi-class classification of polyp types. The use of transfer learning with pre-trained models and data augmentation techniques played a crucial role in achieving these results. Despite the promising results, the study has some limitations, such as the small dataset size and the absence of grading information for adenomatous polyps. It is important to note that while the proposed methods have been validated on the UnitoPatho dataset, further validation on larger and more diverse datasets could potentially improve the generalizability and robustness of the models. Nonetheless, this study serves as a foundation for future research towards developing an automated tool that can aid in the early detection and diagnosis of colon polyps, ultimately leading to improved patient outcomes.

7.2 Societal Impact and Contribution to Global Sustainability

In 2020, 1.93 million new cases of colorectal cancer (CRC) were diagnosed worldwide, resulting in 940,000 deaths. According to global cancer statistics published in 2021, CRC is the one of the most common cause of cancer death, and it is projected that by 2040, new cases will increase to 3.2 million. The increase in the number of pathological colon biopsy slides over the past decade has led to a growing demand for cancer screening programs. Early detection and removal of cancerous tissue for colon cancer can lower the mortality rate, but the increased workload of pathologists due to the rise in biopsy volumes makes it increasingly difficult to detect the disease at an early stage. Given the increasing workload of pathologists and the difficulty of detecting early-stage colon cancer, a system can be developed to alleviate the labor-intensive work and minimize the errors associated with traditional approaches. This study aimed to explore deep learning algorithms on histopathology images to assist pathologists in the decision-making process. Additionally, the use of the Grad-Cam method in this study not only improved the explainability of the classification model but also has the potential to guide pathologists during their diagnosis. The Grad-Cam method provides an attention map on histopathology images by coloring the important sections of the image. By highlighting the areas of interest in the histopathology images, the Grad-Cam printouts can provide additional diagnostic information and assist pathologists in making more informed decisions. As such, this study aligns with the third goal of the United Nations Sustainable Development Goals, which aims to promote good health and well-being.

7.3 Future Prospects

A future research could involve exploring the grading of polyps by obtaining information on the grades of the adenomatous polyps. This additional diagnostic information could prove valuable for the classification of polyps. Furthermore, in this study, only the x2.5 and x5 magnifications of the histopathology images were used for region of interest cropping. To increase the dataset size and variety, manual cropping of patches can be performed for all magnification levels using overlapping windows and labeling of individual patches by an expert. This can potentially lead to the development of more robust and accurate classification models. To learn more fine details about the cancerous cells, various Deep CNN algorithms can be trained on these patches. As a future work, Self-Supervised Learning algorithms can be trained on the unlabeled data obtained from the same source as the labeled data (i.e., Kayseri City Hospital) to improve the performance of the model. Additionally, the fusion of colonoscopy images and histopathology images can be explored to provide more comprehensive diagnostic information. The network outputs of both modalities can be combined to make a more informed decision.

Another potential future research direction involves following up with patients over a longer period of time, such as one year, to obtain ground truth data on the true nature of the polyps. This would involve obtaining information on whether the polyps were benign or malignant and comparing it with the model's predicted classification. This would not only provide valuable information on the accuracy of the classification model but also help identify potential areas for improvement. Additionally, the long-term follow-up of patients would provide insight into the progression of colorectal cancer, leading to the development of more effective screening and diagnostic methods.

BIBLIOGRAPHY

[1] "Colorectal Cancer—Patient Version - NCI." https://www.cancer.gov/types/colorectal (accessed Jan. 12, 2023).

[2] Y. Xi and P. Xu, "Global colorectal cancer burden in 2020 and projections to 2040," *Transl Oncol*, vol. 14, no. 10, p. 101174, Oct. 2021, doi: 10.1016/J.TRANON.2021.101174.

[3] A. G. Zauber *et al.*, "Colonoscopic Polypectomy and Long-Term Prevention of Colorectal-Cancer Deaths," *New England Journal of Medicine*, vol. 366, no. 8, pp. 687–696, Feb. 2012, doi: 10.1056/NEJMOA1100370/SUPPL_FILE/NEJMOA1100370_DISCLOSURES.PDF.

[4] "Patient education: Colon polyps (Beyond the Basics) - UpToDate." https://www.uptodate.com/contents/colon-polyps-beyond-the-basics (accessed Jan. 12, 2023).

[5] M. Meseeha and M. Attia, "Colon Polyps," Aug. 2022, Accessed: Jan. 12, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK430761/

[6] M. Bilal *et al.*, "AI based pre-screening of large bowel cancer via weakly supervised learning of colorectal biopsy histology images," *medRxiv*, p. 2022.02.28.22271565, Feb. 2022, doi: 10.1101/2022.02.28.22271565.

[7] B. Korbar *et al.*, "Deep Learning for Classification of Colorectal Polyps on Whole-slide Images," *J Pathol Inform*, vol. 8, no. 1, 2017, doi: 10.4103/JPI.JPI_34_17.

[8] Z. Song *et al.*, "Automatic deep learning-based colorectal adenoma detection system and its similarities with pathologists," *BMJ Open*, vol. 10, no. 9, p. e036423, Sep. 2020, doi: 10.1136/BMJOPEN-2019-036423.

[9] J. W Wei *et al.*, "Deep neural networks for automated classification of colorectal polyps on histopathology slides: A multi-institutional evaluation."ArXiv, (2019). Accessed March 22, 2023. /abs/1909.12959.

[10] J. Wei *et al.*, “Learn like a pathologist: Curriculum learning by annotator agreement for histopathology image classification,” *Proceedings - 2021 IEEE Winter Conference on Applications of Computer Vision, WACV 2021*, pp. 2472–2482, 2021, doi: 10.1109/WACV48630.2021.00252.

[11] M. Nasir-Moin *et al.*, “Evaluation of an Artificial Intelligence-Augmented Digital System for Histologic Classification of Colorectal Polyps,” *JAMA Netw Open*, vol. 4, no. 11, pp. 1–12, 2021, doi: 10.1001/jamanetworkopen.2021.35271.

[12] C. Zhou *et al.*, “Histopathology classification and localization of colorectal cancer using global labels by weakly supervised deep learning,” *Computerized Medical Imaging and Graphics*, vol. 88, no. December 2020, p. 101861, 2021, doi: 10.1016/j.compmedimag.2021.101861.

[13] P. Gupta *et al.*, “Colon Tissues Classification and Localization in Whole Slide Images Using Deep Learning,” *Diagnostics (Basel)*, vol. 11, no. 8, Aug. 2021, doi: 10.3390/DIAGNOSTICS11081398.

[14] D. Perlo, E. Tartaglione, L. Bertero, P. Cassoni, and M. Grangetto, “Dysplasia Grading of Colorectal Polyps Through Convolutional Neural Network Analysis of Whole Slide Images,” *Lecture Notes in Electrical Engineering*, vol. 784 LNEE, pp. 325–334, 2022, doi: 10.1007/978-981-16-3880-0_34/COVER.

[15] D. Sarwinda, R. H. Paradisa, A. Bustamam, and P. Anggia, “Deep Learning in Image Classification using Residual Network (ResNet) Variants for Detection of Colorectal Cancer,” *Procedia Comput Sci*, vol. 179, pp. 423–431, Jan. 2021, doi: 10.1016/J.PROCS.2021.01.025.

[16] S. Byeon, J. Park, Y. A. Cho, and B.-J. Cho, “Automated histological classification for digital pathology images of colonoscopy specimen via deep learning,” *Scientific Reports 2022 12:1*, vol. 12, no. 1, pp. 1–8, Jul. 2022, doi: 10.1038/s41598-022-16885-x.

[17] C. Ho *et al.*, “A promising deep learning-assistive algorithm for histopathological screening of colorectal cancer,” *Scientific Reports 2022 12:1*, vol. 12, no. 1, pp. 1–9, Feb. 2022, doi: 10.1038/s41598-022-06264-x.

[18] M. Yildirim and A. Cinar, “Classification with respect to colon adenocarcinoma and colon benign tissue of colon histopathological images with a new CNN model: MA_ColonNET,” *Int J Imaging Syst Technol*, vol. 32, no. 1, pp. 155–162, 2022, doi: 10.1002/ima.22623.

[19] “The Colon - Ascending - Transverse - Descending - Sigmoid - TeachMeAnatomy.” https://teachmeanatomy.info/abdomen/gi-tract/colon/ (accessed Jan. 12, 2023).

[20] C. Walsh, “Colorectal polyps: cancer risk and classification,” *https://doi.org/10.12968/gasn.2017.15.5.26*, vol. 15, no. 5, pp. 26–32, Jun. 2017, doi: 10.12968/GASN.2017.15.5.26.

[21] “H&E stain - Wikipedia.” https://en.wikipedia.org/wiki/H%26E_stain (accessed Jan. 12, 2023).

[22] D. J. Myers and K. Arora, “Villous Adenoma,” Sep. 2022, Accessed: Jan. 12, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK470272/

[23] “Colonic Polyps,” *Imaging in Gastroenterology*, pp. 276–277, 2018, doi: 10.1016/B978-0-323-55408-4.50138-5.

[24] C. A. Barbano *et al.*, “UNITOPATHO, A LABELED HISTOPATHOLOGICAL DATASET FOR COLORECTAL POLYPS CLASSIFICATION AND ADENOMA DYSPLASIA GRADING,” *Proceedings - International Conference on Image Processing, ICIP*, vol. 2021-September, pp. 76–80, 2021, doi: 10.1109/ICIP42928.2021.9506198.

[25] W. Hu *et al.*, “EBHI:A New Enteroscope Biopsy Histopathological H&E Image Dataset for Image Classification Evaluation,” pp. 1–8, 2022, [Online]. Available: http://arxiv.org/abs/2202.08552

[26] K. Stacke, G. Eilertsen, J. Unger, and C. Lundstrom, “Measuring Domain Shift for Deep Learning in Histopathology,” *IEEE J Biomed Health Inform*, vol. 25, no. 2, pp. 325–336, 2021, doi: 10.1109/JBHI.2020.3032060.

[27] E. Reinhard, M. Ashikhmin, B. Gooch, and P. Shirley, “Color transfer between images,” *IEEE Comput Graph Appl*, vol. 21 (5), no. 5, pp. 34–41, Sep. 2001, doi: 10.1109/38.946629.

[28] M. Macenko *et al.*, “A METHOD FOR NORMALIZING HISTOLOGY SLIDES FOR QUANTITATIVE ANALYSIS”.

[29] A. Vahadane *et al.*, “Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images,” *IEEE Trans Med Imaging*, vol. 35, no. 8, pp. 1962–1971, Aug. 2016, doi: 10.1109/TMI.2016.2529665.

[30] M. Tarek Shaban, C. Baur, N. Navab, and S. Albarqouni, “StainGAN: Stain Style Transfer for Digital Histological Images.”

[31] H. Kang *et al.*, “StainNet: A Fast and Robust Stain Normalization Network,” *Front Med (Lausanne)*, vol. 8, p. 2002, Nov. 2021, doi: 10.3389/FMED.2021.746307/BIBTEX.

[32] “Pattern Recognition and Machine Learning,” *Pattern Recognition and Machine Learning*, Dec. 2006, doi: 10.1007/978-0-387-45528-0.

[33] C. Janiesch, P. Zschech, and K. Heinrich, “Machine learning and deep learning,” *Electronic Markets*, vol. 31, no. 3, pp. 685–695, Sep. 2021, doi: 10.1007/S12525-021-00475-2/TABLES/2.

[34] A. Tharwat, “Classification assessment methods,” *Applied Computing and Informatics*, vol. 17, no. 1, pp. 168–192, 2018, doi: 10.1016/J.ACI.2018.08.003.

[35] D. Silver *et al.*, “Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm,” 2017.

[36] A. Madani, R. Arnaout, M. Mofrad, and R. Arnaout, “Fast and accurate view classification of echocardiograms using deep learning,” *npj Digital Medicine 2018 1:1*, vol. 1, no. 1, pp. 1–8, Mar. 2018, doi: 10.1038/s41746-017-0013-1.

[37] “7.4 Overfitting‣ Chapter 7 Supervised Machine Learning ‣ Artificial Intelligence: Foundations of Computational Agents, 2nd Edition.” https://artint.info/2e/html/ArtInt2e.Ch7.S4.html (accessed Jan. 12, 2023).

[38] I. Goodfellow, Y. Bengio, and A. Courville, "[PDF] Deep Learning (Adaptive Computation And Machine Learning Series)".

[39] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014, Accessed: Jan. 12, 2023. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html

[40] R. Müller, S. Kornblith, G. H. Google, and B. Toronto, "When does label smoothing help?," *Adv Neural Inf Process Syst*, vol. 32, 2019.

[41] P. Khosla *et al.*, "Supervised contrastive learning," *Adv Neural Inf Process Syst*, vol. 2020-Decem, no. NeurIPS, pp. 1–13, 2020.

[42] Y. Lu, A. Jha, R. Deng, and Y. Huo, "Contrastive learning meets transfer learning: a case study in medical image analysis," *https://doi.org/10.1117/12.2610990*, vol. 12033, pp. 729–736, Apr. 2022, doi: 10.1117/12.2610990.

[43] P. Yang, Z. Hong, X. Yin, C. Zhu, and R. Jiang, "Self-supervised Visual Representation Learning for Histopathological Images," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12902 LNCS, pp. 47–57, 2021, doi: 10.1007/978-3-030-87196-3_5/COVER.

[44] O. Ciga, T. Xu, and A. L. Martel, "Self supervised contrastive learning for digital histopathology," *Machine Learning with Applications*, vol. 7, p. 100198, Mar. 2022, doi: 10.1016/j.mlwa.2021.100198.

[45] H. Wu, Z. Wang, Y. Song, L. Yang, and J. Qin, "Cross-patch Dense Contrastive Learning for Semi-supervised Segmentation of Cellular Nuclei in Histopathologic Images", Accessed: Jan. 12, 2023. [Online]. Available: https://github.com/zzw-szu/CDCL.

[46] N. Boserup and R. Selvan, "Efficient Self-Supervision using Patch-based Contrastive Learning for Histopathology Image Segmentation", doi: 10.7557/18.6798.

[47] J. Ke, Y. Shen, X. Liang, and D. Shen, “Contrastive Learning Based Stain Normalization Across Multiple Tumor in Histopathology,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12908 LNCS, pp. 571–580, 2021, doi: 10.1007/978-3-030-87237-3_55/TABLES/3.

[48] “Boost your model’s accuracy using self-supervised learning with TensorFlow Similarity — The TensorFlow Blog.” https://blog.tensorflow.org/2022/02/boost-your-models-accuracy.html (accessed Jan. 12, 2023).

[49] A. Das, M. N. Mohanty, P. K. Mallick, P. Tiwari, K. Muhammad, and H. Zhu, “Breast cancer detection using an ensemble deep learning method,” *Biomed Signal Process Control*, vol. 70, no. August, p. 103009, 2021, doi: 10.1016/j.bspc.2021.103009.

[50] A. Kumar, J. Kim, D. Lyndon, M. Fulham, and D. Feng, “An Ensemble of Fine-Tuned Convolutional Neural Networks for Medical Image Classification,” *IEEE J Biomed Health Inform*, vol. 21, no. 1, pp. 31–40, Jan. 2017, doi: 10.1109/JBHI.2016.2635663.

[51] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A ConvNet for the 2020s,” 2022, [Online]. Available: http://arxiv.org/abs/2201.03545

[52] J. Yang, C. Li, and J. Gao, “Focal Modulation Networks,” 2022, [Online]. Available: http://arxiv.org/abs/2203.11926

[53] J. K. Turner, G. T. Williams, M. Morgan, M. Wright, and S. Dolwani, “Interobserver agreement in the reporting of colorectal polyp pathology among bowel cancer screening pathologists in Wales,” *Histopathology*, vol. 62, no. 6, pp. 916–924, May 2013, doi: 10.1111/HIS.12110.

[54] A. Kallipolitis, K. Revelos, and I. Maglogiannis, “Ensembling efficientnets for the classification and interpretation of histopathology images,” *Algorithms*, vol. 14, no. 10, 2021, doi: 10.3390/a14100278.

[55] S. H. Kassani, P. H. Kassani, M. J. Wesolowski, K. A. Schneider, and R. Deters, “Classification of histopathological biopsy images using ensemble of deep learning

networks," in *CASCON 2019 Proceedings - Conference of the Centre for Advanced Studies on Collaborative Research - Proceedings of the 29th Annual International Conference on Computer Science and Software Engineering*, 2020, pp. 92–99.

[56] R. Kundu, R. Das, Z. W. Geem, G. T. Han, and R. Sarkar, "Pneumonia detection in chest X-ray images using an ensemble of deep learning models," *PLoS ONE*, vol. 16, no. 9 September. 2021. doi: 10.1371/journal.pone.0256630.

[57] D. T. Nguyen, M. B. Lee, T. D. Pham, G. Batchuluun, M. Arsalan, and K. R. Park, "Enhanced Image-Based Endoscopic Pathological Site Classification Using an Ensemble of Deep Learning Models," *Sensors (Basel)*, vol. 20, no. 21, pp. 1–24, Nov. 2020, doi: 10.3390/S20215982.

[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 770–778, Dec. 2015, doi: 10.48550/arxiv.1512.03385.

[59] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", Accessed: May 11, 2022. [Online]. Available: http://code.google.com/p/cuda-convnet/

[60] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, Sep. 2014, doi: 10.48550/arxiv.1409.1556.

[61] C. Szegedy *et al.*, "Going Deeper with Convolutions," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June, pp. 1–9, Sep. 2014, doi: 10.48550/arxiv.1409.4842.

[62] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Oct. 2020, doi: 10.48550/arxiv.2010.11929.

[63] "StainTools — StainTools documentation." https://staintools.readthedocs.io/en/latest/ (accessed May 11, 2022).

[64] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” *Int J Comput Vis*, vol. 128, no. 2, pp. 336–359, 2020, doi: 10.1007/s11263-019-01228-7.

[65] S. Ghosh, A. Bandyopadhyay, S. Sahay, R. Ghosh, I. Kundu, and K. C. Santosh, “Colorectal Histology Tumor Detection Using Ensemble Deep Neural Network,” *Eng Appl Artif Intell*, vol. 100, Apr. 2021, doi: 10.1016/J.ENGAPPAI.2021.104202.

[66] S. Mehmood *et al.*, “Malignancy Detection in Lung and Colon Histopathology Images Using Transfer Learning with Class Selective Image Processing,” *IEEE Access*, vol. 10, pp. 25657–25668, 2022, doi: 10.1109/ACCESS.2022.3150924.

[67] D. S. Luz, T. J. B. Lima, R. R. V. Silva, D. M. V. Magalhães, and F. H. D. Araujo, “Automatic detection metastasis in breast histopathological images based on ensemble learning and color adjustment,” *Biomed Signal Process Control*, vol. 75, May 2022, doi: 10.1016/J.BSPC.2022.103564.

[68] D. Albashish, “Ensemble of adapted convolutional neural networks (CNN) methods for classifying colon histopathological images”, doi: 10.7717/peerj-cs.1031.

[69] O. Iizuka, F. Kanavati, K. Kato, M. Rambeau, K. Arihiro, and M. Tsuneki, “Deep Learning Models for Histopathological Classification of Gastric and Colonic Epithelial Tumours,” *Scientific Reports*, vol. 10, no. 1. 2020. doi: 10.1038/s41598-020-58467-9.

[70] X. Wang *et al.*, “Transformer-based unsupervised contrastive learning for histopathological image classification,” *Med Image Anal*, vol. 81, p. 102559, Oct. 2022, doi: 10.1016/J.MEDIA.2022.102559.

[71] X. Wang *et al.*, “RetCCL: Clustering-guided contrastive learning for whole-slide image retrieval,” *Med Image Anal*, vol. 83, Jan. 2023, doi: 10.1016/J.MEDIA.2022.102645.

[72] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, “The role of text messaging intervention in Inner Mongolia among patients with type 2 diabetes mellitus: a randomized controlled trial,” 2020, doi: 10.1186/s12911-020-01332-6.

[73] A. Singh, S. Sengupta, and V. Lakshminarayanan, “Imaging Explainable Deep Learning Models in Medical Image Analysis”, doi: 10.3390/jimaging6060052.

[74] S. Pereira, R. Meier, V. Alves, M. Reyes, and C. A. Silva, “Automatic brain tumor grading from MRI data using convolutional neural networks and quality assessment,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11038 LNCS, pp. 106–114, 2018, doi: 10.1007/978-3-030-02628-8_12/COVER.

[75] M. Tsuneki and F. Kanavati, “Deep Learning Models for Poorly Differentiated Colorectal Adenocarcinoma Classification in Whole Slide Images Using Transfer Learning,” *Diagnostics 2021, Vol. 11, Page 2074*, vol. 11, no. 11, p. 2074, Nov. 2021, doi: 10.3390/DIAGNOSTICS11112074.

[76] J. Wei *et al.*, “Generative Image Translation for Data Augmentation in Colorectal Histopathology Images,” *Proc Mach Learn Res*, vol. 116, p. 10, Dec. 2019, Accessed: Aug. 03, 2022. [Online]. Available: /pmc/articles/PMC8076951/

[77] H.-G. Nguyen, A. Blank, H. E. Dawson, A. Lugli, and I. Zlobec, “Classification of colorectal tissue images from high throughput tissue microarrays by ensemble deep learning methods,” *Scientific Reports* |, vol. 11, p. 2371, 123AD, doi: 10.1038/s41598-021-81352-y.

[78] D. Bychkov *et al.*, “Deep learning based tissue analysis predicts outcome in colorectal cancer,” *Scientific Reports 2018 8:1*, vol. 8, no. 1, pp. 1–11, Feb. 2018, doi: 10.1038/s41598-018-21758-3.

[79] T. E. Tavolara, M. K. K. Niazi, V. Arole, W. Chen, W. Frankel, and M. N. Gurcan, “A modular cGAN classification framework: Application to colorectal tumor detection,” *Sci Rep*, vol. 9, no. 1, pp. 1–8, 2019, doi: 10.1038/s41598-019-55257-w.

[80] E. Terradillos *et al.*, “Analysis on the Characterization of Multiphoton Microscopy Images for Malignant Neoplastic Colon Lesion Detection under Deep Learning Methods,” *J Pathol Inform*, vol. 12, no. 1, p. 27, 2021, doi: 10.4103/jpi.jpi_113_20.

[81] B. Korbar *et al.*, “Looking under the Hood: Deep Neural Network Visualization to Interpret Whole-Slide Image Analysis Outcomes for Colorectal Polyps,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, vol. 2017-July, pp. 821–827, 2017, doi: 10.1109/CVPRW.2017.114.

[82] J. W. Wei *et al.*, “Evaluation of a Deep Neural Network for Automated Classification of Colorectal Polyps on Histopathologic Slides,” *JAMA Netw Open*, vol. 3, no. 4, pp. e203398–e203398, Apr. 2020, doi: 10.1001/JAMANETWORKOPEN.2020.3398.

[83] P. Kainz, M. Pfeiffer, and M. Urschler, “Segmentation and classification of colon glands with deep convolutional neural networks and total variation regularization,” *PeerJ*, vol. 5, no. 10, 2017, doi: 10.7717/PEERJ.3874.

[84] Y. Xu *et al.*, “Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features,” *BMC Bioinformatics*, vol. 18, no. 1, pp. 1–17, May 2017, doi: 10.1186/S12859-017-1685-X/FIGURES/8.

[85] Y. Nhi *et al.*, “MedAug: Contrastive learning leveraging patient metadata improves representations for chest X-ray interpretation,” *Proceedings of Machine Learning Research*, vol. 126. PMLR, pp. 755–769, Oct. 21, 2021. Accessed: Jan. 12, 2023. [Online]. Available: https://proceedings.mlr.press/v149/vu21a.html

[86] X. Chen, L. Yao, T. Zhou, J. Dong, and Y. Zhang, “Momentum Contrastive Learning for Few-Shot COVID-19 Diagnosis from Chest CT Images,” *Pattern Recognit*, vol. 113, Jun. 2020, doi: 10.1016/j.patcog.2021.107826.

[87] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, “Contrastive Learning of Medical Visual Representations from Paired Images and Text,” *Proc Mach Learn Res*, vol. 182, pp. 1–24, Oct. 2020, doi: 10.48550/arxiv.2010.00747.

[88] Y. Tian *et al.*, “Constrained Contrastive Distribution Learning for Unsupervised Anomaly Detection and Localisation in Medical Images,” *Lecture Notes in*

Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12905 LNCS, pp. 128–140, 2021, doi: 10.1007/978-3-030-87240-3_13/FIGURES/3.

[89] S. Azizi *et al.*, “Big Self-Supervised Models Advance Medical Image Classification,” *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3458–3468, 2021, doi: 10.1109/ICCV48922.2021.00346.

[90] A. Loh *et al.*, “Supervised Transfer Learning at Scale for Medical Imaging.” 2021. Accessed: Jan. 13, 2023. [Online]. Available: https://research.google/pubs/pub50267/

[91] A. Galdran, G. Carneiro, and M. A. González Ballester, “Balanced-MixUp for Highly Imbalanced Medical Image Classification,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12905 LNCS, pp. 323–333, 2021, doi: 10.1007/978-3-030-87240-3_31/TABLES/3.

[92] Y. Shi *et al.*, “Eosinophilic esophagitis multi-label feature recognition on whole slide imaging using transfer learning,” p. 40, Feb. 2022, doi: 10.1117/12.2611521.

[93] S. Azizi *et al.*, “Big Self-Supervised Models Advance Medical Image Classification”.

[94] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, “Transfusion: Understanding Transfer Learning for Medical Imaging,” *Adv Neural Inf Process Syst*, vol. 32, Feb. 2019, doi: 10.48550/arxiv.1902.07208.

[95] D. Tellez, G. Litjens, J. van der Laak, and F. Ciompi, “Neural Image Compression for Gigapixel Histopathology Image Analysis,” *IEEE Trans Pattern Anal Mach Intell*, vol. 43, no. 2, pp. 567–578, Feb. 2018, doi: 10.1109/TPAMI.2019.2936841.

[96] P. Cascante-Bonilla, F. Tan, Y. Qi, and V. Ordonez, “Curriculum Labeling: Revisiting Pseudo-Labeling for Semi-Supervised Learning,” *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, vol. 8A, pp. 6912–6920, 2021, doi: 10.1609/aaai.v35i8.16852.

[97] X. Xu *et al.*, “Revisiting Pretraining for Semi-Supervised Learning in the Low-Label Regime,” May 2022, doi: 10.48550/arxiv.2205.03001.

CURRICULUM VITAE

2012 – 2016	B.Sc., Electrical and Electronics Engineering, Turgut Özal University, Ankara, TURKEY
2017 – 2018	M.Sc., Electric and Computer Engineering, Abdullah Gül University, Kayseri, TURKEY
2019 – 2023	Ph.D., Electrical and Computer Engineering, Abdullah Gül University, Kayseri, TURKEY
2022 – Present	Research Assistant, Queen’s University Belfast, Belfast, UK

SELECTED PUBLICATIONS AND PRESENTATIONS

J1) S. Yengec-Tasdemir, K. Tasdemir, Z. Aydin. A review of mammographic region of interest classification published in Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery (Feb. 2020)

J2) S. Yengec-Tasdemir, Z. Aydin, E. Akay, S. Dogan, and B. Yilmaz. Improved Classification of Colorectal Polyps on Histopathological Images with Ensemble Learning and Stain Normalization published in Computer Methods and Programs in Biomedicine (Jan, 2023)

J3) S. Yengec-Tasdemir, Z. Aydin, E. Akay, S. Dogan, and B. Yilmaz An Effective Colorectal Polyp Classification for Histopathological Images Based on Supervised Contrastive Learning under Review in IEEE Access