Refika Sultan DOĞAN

Ph.D. Thesis

AGU 2023

# ARTIFICIAL INTELLIGENCE ASSISTED PROGNOSTIC MARKER DETERMINATION FROM COLONOSCOPY AND HISTOPATHOLOGY IMAGES FOR COLON POLYPS

A THESIS
SUBMITTED TO THE DEPARTMENT OF BIOENGINEERING
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
Refika Sultan DOĞAN
December 2023

# ARTIFICIAL INTELLIGENCE ASSISTED PROGNOSTIC MARKER DETERMINATION FROM COLONOSCOPY AND HISTOPATHOLOGY IMAGES FOR COLON POLYPS

A THESIS
SUBMITTED TO THE DEPARTMENT OF BIOENGINEERING
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
Refika Sultan DOĞAN
December 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: Refika Sultan DOĞAN
Signature :

## REGULATORY COMPLIANCE

Ph.D. thesis titled “Artificial Intelligence Assisted Prognostic Marker Determination from Colonoscopy and Histopathology Images for Colon Polyps” has been prepared in accordance with the Thesis Writing Guidelines of the Abdullah Gül University, Graduate School of Engineering & Science.

| Prepared By | Advisor |
| --- | --- |
| Refika Sultan DOĞAN | Prof. Bülent YILMAZ |
| Signature | Signature |

Head of the Bioengineering Program

Asst. Prof. Altan ERCAN

Signature

## ACCEPTANCE AND APPROVAL

Ph.D. thesis titled “Artificial Intelligence Assisted Prognostic Marker Determination from Colonoscopy and Histopathology Images for Colon Polyps” and prepared by Refika Sultan DOĞAN has been accepted by the jury in the Bioengineering Graduate Program at Abdullah Gül University, Graduate School of Engineering & Science.

19/12/2023

(Thesis Defense Exam Date)

**JURY:**

Advisor : Prof. Bülent YILMAZ

Member : Prof. Murat GÖK

Member : Assoc. Prof. Gökmen ZARARSIZ

Member : Assoc. Prof. Zafer AYDIN

Member : Asst. Prof. Bekir Hakan AKSEBZECİ

**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. İrfan ALAN

# ABSTRACT

## ARTIFICIAL INTELLIGENCE ASSISTED PROGNOSTIC MARKER DETERMINATION FROM COLONOSCOPY AND HISTOPATHOLOGY IMAGES FOR COLON POLYPS

Refika Sultan DOĞAN
Ph.D. in Bioengineering
Advisor: Prof. Bülent YILMAZ
December 2023

According to World Health Organization statistics for 2023, colorectal cancer is the third most common type of cancer worldwide, accounting for approximately 10% of all cancer cases. Most colon cancer begins with polyps that form as a result of abnormal cell proliferation in the colon mucosa. There are two types of colon polyps: neoplastic and nonneoplastic. Neoplastic polyps have malignant potential. Colonoscopy is the most common method for detecting polyps. It is possible to detect and remove polyps (polypectomy) with the tool at the end of the colonoscopy. Pathologists prepare and examine almost all removed polyps under a microscope using hematoxylin-eosin (H&E) stained tissue slides. When uncertainty arises, pathologists perform immunohistochemical (IHC) analyses to demonstrate significant expressions of cancer-specific antigens (proteins). This thesis achieved four main outcomes: First, the automatic determination of polyp type or subtype, stage, and malignant potential using colonoscopy videos, images, and frames and using pathology reports and IHC analysis results as labels were investigated. Secondly, the automatic characterization of colon polyps from histopathology images using features obtained from colonoscopy images, pathology reports, and IHC analysis results were examined. Statistical approaches were used to analyze polyp type or subtype, stage, and possible prognostic features (biomarkers) that may indicate malignant potential. Finally, we created a comprehensive database and shared it with the scientific community as an open-source repository. The database contains colonoscopy and histopathology images of more than 400 polyps, along with information on polyp type, location, stage, and IHC analysis results of Ki-67 (clone 30-9), CD34 (clone QBend/10), p53 (clone bp53-11), BRAF (clone V600E), VEGF (clone SP125), and PD-L1 (clone SP142) markers.

*Keywords: Colon Polyps, Computer Aided Diagnosis, Artificial Intelligence, Deep Learning, Machine Learning, Image Processing, Data Mining*

# ÖZET
# KOLON POLİPLERİ İÇİN KOLONOSKOPİ VE HİSTOPATOLOJİ GÖRÜNTÜLERİNDEN YAPAY ZEKÂ DESTEKLİ PROGNOSTİK BELİRTEÇ TESPİTİ

Refika Sultan DOĞAN
Biyomühendislik Anabilim Dalı Doktora
Tez Yöneticisi: Prof. Bülent YILMAZ
Aralık 2023

Dünya Sağlık Örgütü'nün 2023 yılı istatistiklerine göre kolorektal kanser dünya çapında en sık görülen üçüncü kanser türüdür ve tüm kanser vakalarının yaklaşık %10'unu oluşturmaktadır. Çoğu kolon kanseri, kolon mukozasında anormal hücre çoğalması sonucu oluşan poliplerle başlar. Kolon polipleri neoplastik ve neoplastik olmayan olarak iki türe ayrılır ve neoplastik polipler kanser potansiyele sahiptir. Kolonoskopi poliplerin tespitinde en yaygın kullanılan yöntemdir. Kolonoskopun ucundaki aletle poliplerin tespit edilip çıkarılması (polipektomi) mümkündür. Çıkarılan poliplerin neredeyse tamamının Hematoksilen ve Eozin (H&E) boyalı doku slaytları hazırlanıp patologlar tarafından mikroskop altında incelenir. Belirsizlik durumunda, kansere özgü önemli antijen (protein) ekspresyonlarını göstermek için immünohistokimyasal (İHK) analizler yapılır. Bu tezde dört ana çıktı elde edildi: İlk olarak, kolonoskopi videoları ve görüntüleri/kareleri kullanılarak polip tipi/alt tipi, evresi ve malignite potansiyelinin otomatik olarak belirlenmesi ve patoloji raporları ile İHK analiz sonuçlarının etiket olarak kullanılması araştırıldı. İkinci olarak kolonoskopi görüntülerinden, patoloji raporundan ve İHK analiz sonuçlarından elde edilen özellikler kullanılarak histopatoloji görüntülerinden kolon poliplerinin otomatik karakterizasyonu incelendi. Üçüncüsü, kanser potansiyeli gösterebilecek polip tipi/alt tipi, evresi ve olası prognostik özellikler (biyobelirteçler) istatistiksel yaklaşımlar kullanılarak analiz edildi. Son olarak Ki-67 (klon 30-9), CD34 (klon QBend/10), p53 (klon bp53-11), BRAF (klon V600E) , VEGF (klon SP125) ve PD-L1 (klon SP142) belirteçlerine ait 400'den fazla polipin kolonoskopi ve histopatoloji görüntülerini, polip tipini, lokasyonunu, evresini ve IHC analiz sonuçlarını içeren kapsamlı bir veri tabanı oluşturuldu ve bu veri tabanı oluşturuldu. açık kaynak kodlu bir depo olarak bilim camiasıyla paylaşılmaktadır.

*Anahtar Kelimeler: Kolon Polipleri, Bilgisayar Destekli Tanı, Yapay Zeka, Derin Öğrenme, Makine Öğrenmesi, Veri Madenciliği*

# Acknowledgments

Firstly, I would like to thank my research advisor, Prof. Bülent YILMAZ. His office was always open whenever I had a problem or needed advice about my research journey. I am especially grateful for his endless encouragement, which has been fundamental to my academic growth.

With the same level of gratitude, I would also like to thank Assoc. Prof. Gökmen ZARARSIZ and his family. As a member of my thesis committee and a supporter since my undergraduate years, his insightful feedback and understanding have been of immense value.

My gratitude extends to the Scientific and Technological Research Council of Turkey (TÜBİTAK) for their essential backing of this research under Grant 120E204, TÜBİTAK 1001. Beyond the financial assistance, their support has been a profound source of encouragement in my academic pursuits.

I am deeply thankful to Dr. Ebru AKAY and Dr. Serkan DOĞAN, alongside the entire team at the gastroenterology and pathology departments of Kayseri City Hospital. Their support in my data collection and curation process, even during the challenging times of the pandemic, played a pivotal role in the successful completion of this research.

I extend my gratitude to my dear friends Rukiye Nur KAÇMAZ, Kübra Nur ŞAHİN, and Gizem ARGIN for their support and the wonderful times we have shared over the years. Their fellowship has been a source of joy and strength, lifting me during moments of doubt and despair.

Finally, I would like to express my deep gratitude to my family, who raised me with love, provided me with unwavering support, and continuously encouraged me. Without them, nothing would have been possible. Special thanks to my beloved lovely mother and father, Sevgi and İsmail DOĞAN. I also want to express my gratitude to my dear sister, Emine DOĞAN ÇULHA, who has been my guiding light throughout my academic career and life, as well as to my dear siblings and nieces. Their love and support mean everything to me.

# TABLE OF CONTENTS


1. **INTRODUCTION ..... 2**

1.1 General Aim ..... 2
1.2 Motivations and Goals ..... 2
1.3 Thesis Contributions ..... 4

2. **BACKGROUND ..... 5**

2.1 Colon Anatomy ..... 5
2.2 Colon Polyps Types ..... 6
2.3 Colon Polyps Screening and Surveillance ..... 7
2.4 Colonoscopy Images Studies ..... 8
2.4 Histopathology Images Studies ..... 10
2.5 Immunohistochemistry Studies ..... 12

3. **MATERIALS ..... 14**

3.1 Patients ..... 14
3.2 Data Collection and Database Formation ..... 19
*3.2.1 Obtaining Colonoscopy Videos ..... 20*
*3.2.2 Preprocessing of Colonoscopy Images ..... 27*
3.2.2.1 CLAHE ..... 27
3.2.2.2 Dehazing ..... 29
*3.2.3 Obtaining Histopathology Images ..... 30*
3.2.3.1 Preprocessing of Histopathology Images, Color Normalization ..... 34
*3.2.4 Immunohistochemical Analysis with Prognostic Markers ..... 35*
3.3 Data Records ..... 41
*3.3.1 Colonoscopy Video Dataset ..... 41*
*3.3.2 Histopathological Image Dataset ..... 42*
*3.3.3 Immunohistochemistry (IHC) Results Dataset ..... 43*
3.4 Technical Validation ..... 43
3.5 Code and Dataset Availability ..... 45

4. **METHODS ..... 46**

4.1 Feature Extraction Methods ..... 46
*4.1.1 Border/Interior Pixel Classification ..... 46*
*4.1.2 Global Color Histogram ..... 46*
*4.1.3 First Order Measures ..... 47*
*4.1.4 Local Binary Pattern ..... 47*
*4.1.5 Haralick ..... 48*
*4.1.6 Threshold Adjacency Statistics ..... 49*
*4.1.7 Zernike Moments ..... 50*
4.2 Feature selection ..... 50
4.3 Classification Methods ..... 51
*4.3.1 Logistic Regression (LR) ..... 51*
*4.3.2 Random Forest (RF) ..... 52*
*4.3.3 Decision Trees (DT) ..... 52*
*4.3.4 K-Nearest Neighbors (KNN) ..... 52*
*4.3.5 Support Vector Machines (SVM) ..... 52*
*4.3.6 eXtreme Gradient Boosting (XGBoost) ..... 52*
*4.3.7 Artificial Neural Network (ANN) ..... 53*
*4.3.8 Gradient Boosting (GB) ..... 53*

*4.3.9 Light Gradient Boosting Machine (LGBM)* ..... *53*
*4.3.10 Naive Bayes (NB)* ..... *53*
4.4 Cross-validation techniques ..... 53
4.5 Performance evaluation metrics ..... 53
4.6 Optimizing Hyperparameters for Enhanced Model Performance ..... 58
4.7 Statistical Analysis and Evaluation ..... 59
**5. STUDIES AND ANALYSIS WITH RESULTS ..... 62**
5.1 Colonoscopy Based Studies ..... 60
5.2 Histopathology Based Studies ..... 76
5.3 IHC Based Studies ..... 80
**6. CONCLUSIONS ..... 89**
6.1 Colonoscopy Based Studies ..... 89
6.2 Histopathology Based Studies ..... 94
6.3 IHC Based Studies ..... 94
6.4 Societal Impact and Contribution to Global Sustainability ..... 95
**7. FUTURE PROSPECTS ..... 97**
7.1 Colonoscopy Based Studies ..... 101
7.2 Histopathology Based Studies ..... 102
7.3 IHC Based Studies ..... 103
**8. APPENDIX ..... 114**
Colonoscopy Examination Result Form ..... 114
Pathology Laboratory Examination Form ..... 111
Informed Consent Form from Volunteer Participants ..... 113

# LIST OF FIGURES

Figure 1.1 The comprehensive dataset to be obtained during this thesis work ..... 3
Figure 1.2 Estimating micro from macro..... 3
Figure 2.1 Anatomical parts of the colon (source: https://teachmeanatomy.info/abdomen/gi-tract/colon/)..... 5
Figure 2.2 Colon cancer progression on colon mucosa [15] ..... 6
Figure 2.3 Colonoscopy procedure..... 7
Figure 3.1 The flowchart of this study..... 15
Figure 3.2 Age distribution of patients ..... 15
Figure 3.3 Gender distribution of patients..... 16
Figure 3.4 Age distribution by gender..... 17
Figure 3.5 Number of polyp types..... 17
Figure 3.6 Number of polyp subtypes ..... 18
Figure 3.7 The comprehensive database created from KCH..... 20
Figure 3.8 Colonoscopy environment from KCH ..... 20
Figure 3.9 Total video durations for each patient's videos ..... 21
Figure 3.10 Total video durations for selected patient videos..... 22
Figure 3.11 Extracted frames from the selected videos..... 23
Figure 3.12 Examples from different polyp types ..... 23
Figure 3.13 Age distribution of the patients whose videos were selected..... 24
Figure 3.14 Age distribution by gender colon video input dataset..... 24
Figure 3.15 The distribution of polyp types and subtypes in colon video dataset..... 24
Figure 3.16 Examples of frames containing polyps from the colonoscopy dataset ..... 25
Figure 3.17 Colonoscopy images of the same polyp of a patient using different imaging modes, from left to right white light, FICE 0, and FICE 1..... 26
Figure 3.18 Labels used in the classification of polyps..... 26
Figure 3.19 Names of colonoscopy frames ..... 27
Figure 3.20 Sample frames from one polyp with different focusing levels ..... 27
Figure 3.21 Colonoscopy frames preprocessing..... 28
Figure 3.22 Same frames with different focusing levels, after preprocessing..... 29
Figure 3.23 Before and after preprocessing..... 29
Figure 3.24 The distribution of polyp types and subtypes of histopathology images input dataset ..... 31
Figure 3.25 Nonneoplastic, neoplastic, normal, and carcinoma examples from the input image set ..... 31
Figure 3.26 Different polyp types from the same patient with all magnification levels(2.5X, 5X, 10X, 20X from left to right)..... 32
Figure 3.27 Number of histopathology images with different magnification levels..... 33
Figure 3.28 Sample histopathology images from different polyp types/subtypes..... 33
Figure 3.29 Sample images of tubular adenoma polyp and normal tissue samples taken from the side of the same polyp, respectively ..... 33
Figure 3.30 Filename format for the histopathology slides/images..... 34
Figure 3.31 Application results of the Reinhard normalization technique on our images ..... 35
Figure 3.32 Comparison of biomarker expressions..... 36
Figure 3.33 Mean values of numerical variables..... 37
Figure 3.34 BRAF (clone V600E) distribution ..... 37

Figure 3.35 Distribution of PD-L1 epithelium ..... 37
Figure 3.36 Distribution of PD-L1 lymphocyte ..... 38
Figure 3.37 Distribution of p53 (clone bp53-11)..... 38
Figure 3.38 Distribution of CD34 (clone QBend/10)..... 38
Figure 3.39 Distribution of VEGF (clone SP125)..... 39
Figure 3.40 Distribution of Ki-67 (clone 30-9) ..... 39
Figure 3.41 Slides of BRAF (clone V600E) , Ki-67 (clone 30-9) and PD-L1 (clone SP142) antibodies, respectively..... 40
Figure 3.42 Individual representation of some of the BRAF (clone V600E) antibody slides ..... 40
Figure 3.43 CD34 (clone QBend/10), Ki-67 (clone 30-9), VEGF (clone SP125) and p53 (clone bp53-11) staining score results that came from 201 patients..... 41
Figure 4.1 The CV strategies in terms of forming training and test data [113]..... 54
Figure 4.2 Confusion matrix..... 56
Figure 4.3 Receiver operating characteristic curve ..... 57
Figure 5.1 Sample frames from one polyp with different focusing levels ..... 60
Figure 5.2 Same frames with different focusing levels, but preprocessing applied..... 61
Figure 5.3 Extraction of LBP features..... 61
Figure 5.4 Confusion matrix, precision, and recall results for each class using ResNet-18 ..... 63
Figure 5.5 Confusion matrix, precision, and recall results for each class using ResNet-50 ..... 63
Figure 5.6 Confusion matrix, precision, and recall results for each class using AlexNet ..... 64
Figure 5.7 Confusion matrix, precision, and recall results for each class using GoogleNet..... 64
Figure 5.8 Examples of two clinically important subtypes to distinguish from each other are a) hyperplastic polyps, and b) tubular adenomas..... 65
Figure 5.9 Sample frames from the datasets with different focusing levels..... 66
Figure 5.10 General workflow of this study..... 66
Figure 5.11 The same keyframe with different focusing levels before (first column) and after (second column) preprocessing was applied..... 67
Figure 5.12 AUC values and classifiers with optimum parameters and without feature selection method ..... 71
Figure 5.13 AUC values and classifiers with optimum parameters and with RFE feature selection method ..... 73
Figure 5.14 AUC values and classifiers with optimum parameters and with SelectFromModel feature selection method ..... 73
Figure 5.15 AUC values and classifiers with optimum parameters and with SelectKBest feature selection method..... 73
Figure 5.16 AUC values and classifiers with optimum parameters and with Boruta feature selection method..... 73
Figure 5.17 Distributions of biomarkers and age variable..... 78
Figure 5.18 Correlation matrix of various biomarkers ..... 79
Figure 5.19 VEGF (clone SP125) values ..... 81
Figure 5.20 Ki-67 (clone 30-9) values..... 81
Figure 5.21 p53 (clone bp53-11) values ..... 82
Figure 5.22 CD34 (clone QBend/10) values ..... 82
Figure 5.23 Test Train ratio 20%-80%, 5-fold CV, with all markers confusion matrix and ROC curve results from Linear SVM ..... 84

Figure 5.24 Test Train ratio 20%-80%, 5-fold CV, with selected markers confusion matrix and ROC curve results from Cubic SVM ..... 85
Figure 5.25 Test Train ratio 10%-90%, 5-fold CV, with all markers confusion matrix and ROC curve results from Efficient Linear SVM ..... 85
Figure 5.26 Test Train ratio 10%-90%, 5-fold CV, with selected markers confusion matrix and ROC curve results from the fine tree ..... 85
Figure 5.27 Test Train ratio 10%-90%, 10-fold CV, with all markers confusion matrix and ROC curve results from kernel naïve bayes ..... 86
Figure 5.28 Test Train ratio 10%-90%, 10-fold CV, with selected markers confusion matrix and ROC curve results from the fine tree ..... 86
Figure 5.29 Test Train ratio 20%-80%, 10-fold CV, with all markers confusion matrix and ROC curve results from the fine tree ..... 86
Figure 5.30 Test Train ratio 20%-80%, 10-fold CV, with selected markers confusion matrix and ROC curve results from decision tree ..... 87

# LIST OF TABLES

Table 4.1 The grid-search space used for tuning the hyperparameters of the machine-learning classifiers 58
Table 5.1 Using LBP features to classify neoplastic and nonneoplastic polyps 62
Table 5.2 Performance comparison of deep learning models in polyp classification 64
Table 5.3 Method, category, and the number of features extracted 68
Table 5.4 The best parameters after hyperparameters optimization 70
Table 5.5 Maximum AUC values and corresponding classifiers obtained by different feature selection methods on the dataset 74
Table 5.6 Method, category, and the number of features extracted 75
Table 5.7 Binary (neoplastic-nonneoplastic) and multiple (Tubular, Villous, Tubulovillous - Hyperplastic, inflammatory, hamartomatous) classification results of histopathology images obtained using all magnification ratios and different performance metrics 77
Table 5.8 Comparison detailed evaluation of the selection of markers and the performance of classification algorithms 84
Table 5.9 Mean ± standard deviation values markers by types/subtypes 87
Table 6.1 Polyp characterization studies using colonoscopy images/videos 90
Table 6.2 Method, dataset, and performance metric results 92

# LIST OF ABBREVIATIONS

| | |
|---|---|
| ANOVA | Analysis of Variance |
| AUC | Area Under Curve |
| ANN | Artificial Neural Network |
| BIC | Border/Interior Pixel Classification |
| CC | Colon Cancer |
| CLAHE | Contrast Limited Adaptive Histogram Equalization |
| CNN | Convolutional Neural Networks |
| CV | Cross Validation |
| DT | Decision Trees |
| XGBoosting | eXtreme Gradient Boosting |
| FOM | First Order Measures |
| FICE | Fuji Intelligent Chromo Endoscopy |
| GCH | Global Color Histogram |
| GBM | Gradient Boosting Machine |
| GLCM | Gray Level Co-occurrence matrix |
| H&E | Hematoxylin-Eosin |
| HD | High Definition |
| HSV | Hue Saturation Value |
| IHC | Immunohistochemistry |
| KNN | K-Nearest Neighbors |
| KCH | Kayseri City Hospital |
| LGBM | Light Gradient Boosting Machine |
| LBP | Local Binary Pattern |
| LR | Logistic Regression |
| NB | Naive Bayes |
| NBI | Narrow Band Imaging |
| PD-L1 | Programmed Death-Ligand 1 |
| RF | Random Forest |
| RFE | Recursive Feature Elimination |
| RGB | Red Green Blue |
| SFM | Select From Model |
| SVM | Support Vector Machines |
| TAS | Threshold Adjacency Statistics |
| UICC | Union Internationale Contre le Cancer |
| VEGF | Vascular Endothelial Growth Factor |

*To my lovely family*

# Chapter 1

# Introduction

## 1.1 General Aim

In this thesis study, we used macroscopic (obtained by colonoscopy) and microscopic (received pathological analysis) images to automatically differentiate which stage the polyp represents in colon cancer (CC) development. In colonoscopy images/videos obtained from patients with polyps, the morphological, shape, color, and tissue properties of the polyp were extracted with image/video processing approaches and analyzed together with the pathological evaluation results obtained from the biopsy samples.

Parameters and image features, which are explained in detail in the method section, show the presence of malignancy-risk, interpreted. At this stage, both statistical comparisons and automatic classification approaches are used. Automatic classification is made using the determined parameters and features, and in the first step, we develop a system that will help the physician's visual analysis. These steps assisted in identifying features with relatively high discriminating power and successful classification approaches. In this way, we develop a visualization system that can guide gastroenterology specialists while performing biopsy procedures during the routine colonoscopy examination. At the same time, this thesis will be a pioneering study to evaluate histopathological images. On the other hand, thanks to this study, immunohistochemical (IHC) analyses investigated the molecular level differences between the polyp and polyp side normal tissue samples taken from the patients. The IHC analysis explores a specific antigen (protein) or cell in a cell or tissue.

## 1.2 Motivations and Goals

The thesis aims to develop a comprehensive artificial intelligence-aided system to predict the stage of colon polyps known to be precursors of colon cancer. We expect this system to give endoscopists with and without advanced expertise the chance to better detect the stage of polyps on the screen during the colonoscopy examination aided to pathology. The most remarkable feature of this system is that it enables optical/virtual biopsy of colon polyps during colonoscopy, regardless of the endoscopist's skills. It ensures that the adenomatous (neoplastic)

polyps are entirely removed from the intestinal tissue and prevents unnecessary polypectomy. Manual analysis of histopathological images of these lesions taken during colonoscopy is still a time-consuming task for pathologists.

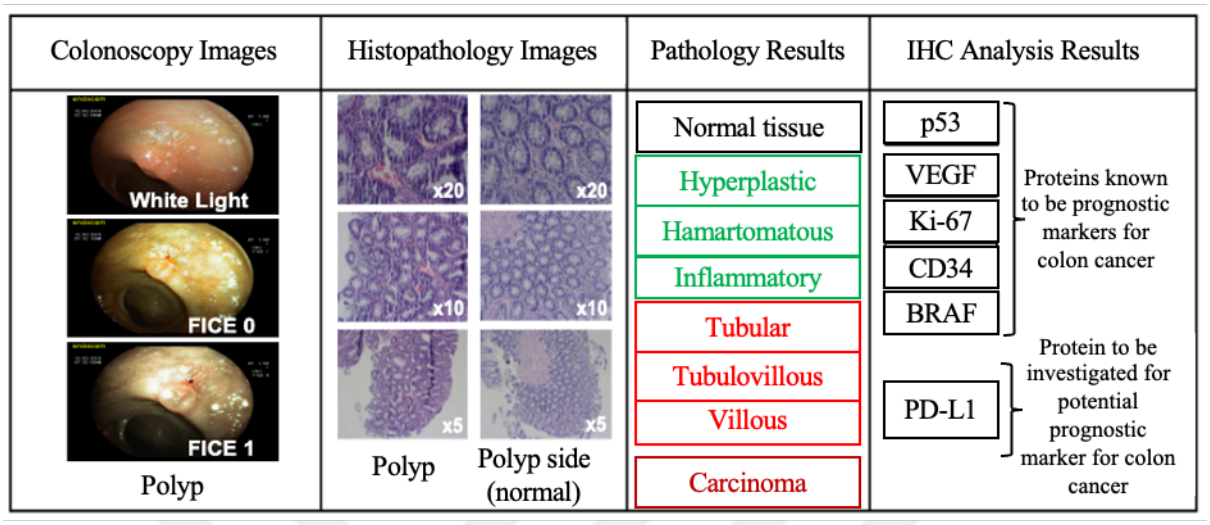



**Figure 1.1 The comprehensive dataset to be obtained during this thesis work**

Another aim of this thesis is to develop solutions to confirm optical/virtual biopsy during colonoscopy by developing medical models to classify tissues. Increasing the cost-efficiency of colonoscopy with the visual diagnosis of colon polyps, reducing the cost of pathology examinations, and increasing the efficiency of patient follow-up are among the essential outputs of the thesis. Using real-time visual AI to highlight cancerous lesions can ensure that an entirely concentrated second observer is always active. Studies have shown that having a second observer can increase polyp detection rates; a 1% increase in adenoma detection rate reduces the risk of colon cancer by 3 percent [1].

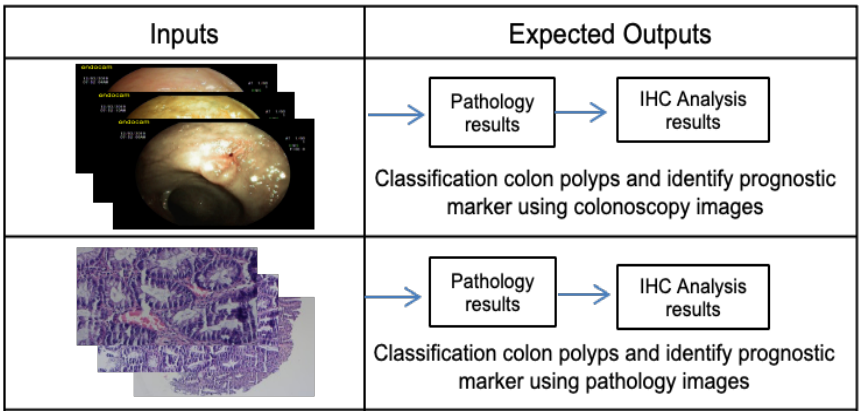



**Figure 1.2 Estimating micro from macro**

The final goal of our thesis is to determine a prognostic marker that gives information about the clinical course of the polyp from the colonoscopy images during the examination and to

contribute to the literature by analyzing routinely used histopathological images. The clinical meaning of the prognostic marker determination, which gives information about the clinical course of the polyp, is to detect the polyps that risk turning into cancer during the examination and preventing the risk of CC. On the other hand, we contributed to this field by presenting the lack of a comprehensive dataset, which is one of the essential needs in the literature and the study results as open source. This thesis aims to solve a significant problem: estimating macro and micro dimensions, and even proteins on the integrated dataset prepared (Figure 1.1 and Figure 1.2).

## 1.3 Thesis Contributions

This thesis provides significant advances in the field of colorectal cancer detection and characterization. Its main contributions are as follows:

• Automated Diagnosis and Characterization: The study investigates methods to automatically identify and characterize colon polyps using colonoscopy videos and images. This can reduce human errors in colorectal cancer detection and speed up the diagnostic process.

• Immunohistochemical (IHC) Analysis Integration: The study develops a method to determine the malignant potential of polyp types using valuable data such as pathology reports and IHC analysis results. This may help better evaluate the cancer risk of polyps.

• Examination of Biomarkers: In the study, biomarkers that may indicate cancer potential are analyzed statistically according to polyp type, subtype, and stage. This can provide a more accurate diagnosis and better patient management.

• Creating a Comprehensive Database: The study creates a comprehensive database containing more than 400 polyp samples. This database is shared with the scientific community so other researchers can conduct similar studies or evaluate the developed methods.

• Technology and Health Applications: This study highlights the potential of integrating technologies such as artificial intelligence, deep learning, and computer-aided diagnosis into medical applications. It contributes to the development of more effective and accurate methods in the field of colorectal cancer screening and diagnosis.

# Chapter 2

# Background

## 2.1 Colon Anatomy

The colon is an important part of the human digestive system and performs various physiological functions. The anatomy of the colon can be affected by factors such as gender, age, and body mass index [2]. Detailed anatomical and histological features of the colon and rectum provide a better understanding of the physiological and pathological processes of these organs. The colon, which is an important part of the digestive system, is generally divided into five main sections (Figure 2.1).

The cecum, which is the beginning part of the colon that merges with the ileum, the last part of the small intestine, is the starting point of the appendix, also known as appendicitis. The part that starts from the cecum and extends upwards in the right abdominal region and is located under the liver is the ascending colon.

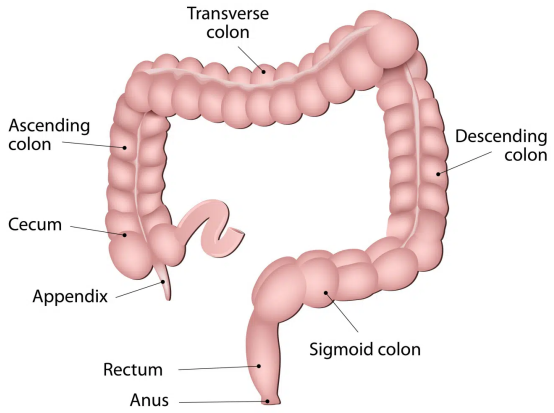



**Figure 2.1 Anatomical parts of the colon (source: https://teachmeanatomy.info/abdomen/gi-tract/colon/)**

The transverse colon starts from the right colon flexure and extends towards the left side in the upper part of the abdominal cavity and forms the splenic flexure. The descending part of the colon is called the descending colon and starts from the left colon flexure and extends downwards in the left abdominal region. The part that starts from the descending colon and extends in an S shape towards the rectum is the sigmoid colon. These parts change according

to the functional and anatomical characteristics of the colon [3]. The colon plays an important role in removing waste products from the body and absorbing water and electrolytes.

## 2.2 Colon Polyp Types

Colon cancer (CC) occurs in the colon, the last part of the digestive system, and is the second leading cause of cancer death worldwide for men and women [4], [5]. However, colon cancer can often be cured when detected early. In addition, detecting changes in the colon before turning into cancer reduces mortality rates, thanks to the improvements in treatment and the increase in screening technologies [6], [7]. Most CC cases begin with a small group of benign cells called adenomatous **polyps**, caused by abnormal cell proliferation in the colonic mucosa [8], [9]. According to histological and morphological features, colon polyps are divided into two main types: **neoplastic** and **nonneoplastic** [10], [11]. Neoplastic polyps, also known as adenomatous polyps, and it is crucial to differentiate the dysplasia level. According to their level of dysplasia, they are classified as tubular, villous, and tubulovillous adenomas [12]. Colon polyps can also be classified as sessile (flat, originating directly from the mucosal layer) or pedunculated (extending from the mucosa to a fibrovascular stem) by their macroscopic appearance [13]. Neoplastic polyps are of primary importance due to their potential for malignancy, representing a critical stage in CC development (Figure 2.2). For these reasons, by identifying polyps at an early stage, a simple outpatient procedure can stop cancer development and prevent disease and death [14].

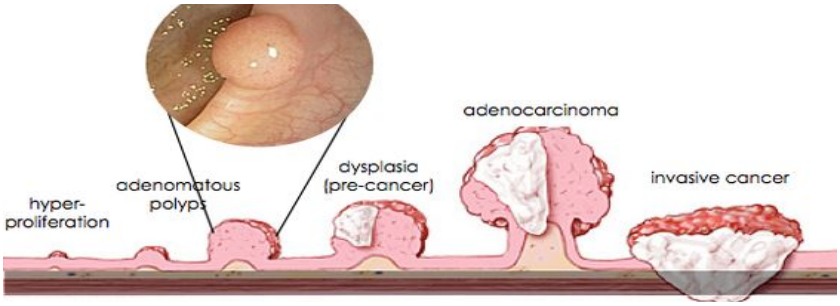



**Figure 2.2 Colon cancer progression on colon mucosa** [15]

After the polypectomy procedure, these biopsy samples are sent to the pathology department to analyze the tissue under the microscope. In the output of this analysis, the pathologist obtained hematoxylin-eosin (H&E) stained tissue slides. These slides of almost every extracted polyp tissue are routinely prepared and examined. If there is any doubt, tissues are analyzed by immunohistochemical (IHC) methods. However, it has been proven that about 50% of small polyps do not cause cancer [16]. Removal of polyps that do not turn into cancer can harm the

patient instead of benefit. Gastroenterologists should prefer not to touch polyps that are not malignant and do not have the risk of turning into cancer over time. Because, while removing such polyps increases medical costs, it may also cause complications such as perforation and bleeding. For this reason, the detection of polyp types and characterization of polyps during colonoscopy has a clinical meaning and requirement. Predicting the type of polyp from colonoscopy images is called optical/virtual diagnosis. It is thought that this approach has serious potential in the early detection of cancer and in helping patients recover their health as soon as possible. In addition, this method can significantly reduce the workload of pathologists. Studies in this area attract increasing attention because it is necessary to understand and quantify the differences in colonoscopy and histopathology images taken from patients with colon polyps.

## 2.3 Colon Polyps Screening and Surveillance

### 2.3.1 Colonoscopy

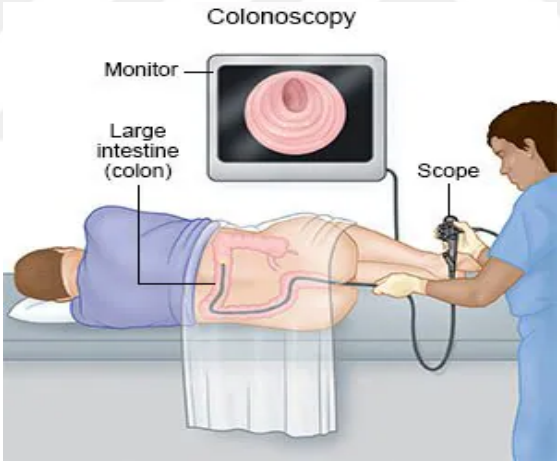



**Figure 2.3 Colonoscopy procedure**

The colonoscopy procedure is mainly used for the diagnosis and treatment of colon cancer. A colonoscopy is the most common procedure for detecting colon polyps. In this procedure (Figure 2.3), gastroenterologists visually examine the colon wall using a flexible probe with a camera and light source at the tip. It is possible to record videos and photographs via this system. In addition to determining the location of the polyps during colonoscopy, it is also possible to remove them with an apparatus at the tip of the probe, which is called polypectomy or polyp resection. The accurate detection of the polyp and the classification of its type depends on the experience of the gastroenterologist. Looking at the monitor for long hours may cause

mental or physical exertion, and in turn, misdiagnosis or missed polyps [17]. Currently, all polyps are resected during colonoscopy and identified whether they are neoplastic or nonneoplastic using histopathology techniques to prevent cancer. The resection procedure applies to all polyps, and this causes a significantly increased workload in the pathology department of hospitals. Therefore, discriminating polyp types in real-time during the colonoscopy procedure is critical to determine which polyp needs to be resected. In addition, this is a challenging issue for physicians because there are uninformative frames, varying illumination conditions of endoscopy, variant texture, and specular reflection due to the light source at the probe. Therefore, an accurate and effective computer-aided diagnosis system is necessary to help identify and classify polyp types during colonoscopy in real time.

## 2.4 Colonoscopy Images Studies

Gastroenterology is an area that requires physicians to have a wide range of clinical skills, from the precise operation of endoscopic devices to visual identification and classification of the disease as well as data-driven decision-making. In recent years, artificial intelligence tools have started to be designed to help doctors perform these tasks. Various research groups in different countries have shown how machine learning and deep learning can aid multiple skills, from detecting colon polyps to analyzing wireless capsule endoscopy images [18], [19]. As the number of applications of artificial intelligence in gastroenterology increases, the success and challenges of the method are better understood. Deep learning and artificial intelligence applications are precious in clinical medicine because medical data often consists of unstructured text, images, and videos that cannot be quickly processed [20]. Early studies were based on extracting or combining features such as texture and color from endoscopic images using pattern classification and analysis methods. One of these analysis methods was the region-growing method. Krishnan et al. used this method to detect abnormalities from the endoscopic images [21] and extract the colon lumen [22]. Iakovidis et al. employed this approach to detect the adenoma polyps [23]. Subsequent studies focused on the development of neural networks. Wang et al. reported using local binary pattern (LBP) texture features combined with the neural network to classify the characteristics of polyps using endoscopic images [24]. Another group developed an automatic polyp detection and classification system with 0.88 sensitivity using a hybrid context-shape approach; the structure of the lesion and shape features were related to the polyp localization [25]. These approaches depend on the conventional machine learning methods using manual feature extraction and classification. In

recent years, several research groups have been trying to develop optical biopsy systems that would enable early detection of cancer and polyps before they become a serious risk using new approaches, such as artificial intelligence. Deep learning and convolutional neural networks (CNN) have been used to detect and classify colon polyps. Urban et al. [26] designed a CNN system to detect and classify colonic polyps using 8641 manually labeled images and colonoscopy videos. Another study aimed to discriminate hyperplastic and adenomatous polyps by training neural networks to enhance the diagnosis using images obtained from 159 patients [27]. Byrne et al. achieved high accuracy only using narrow-band imaging (NBI) video frames while using the CNN model to classify adenomatous or hyperplastic polyps [28]. Different imaging modalities are used to increase the adenomatous polyp detection rate and aid the determination of polyp type. NBI is one of these modalities used for this purpose [29], [30]. Also, the magnification option is available in some colonoscopy systems [31]. Some research groups conducted polyp classification using the NBI and magnification options to avoid unnecessary tissue biopsy [32]–[34]. NBI also allows the visualization of mucosa and vascular structures on the surface of the polyp, and with these features, it has been used in classifying polyp histology [35]. However, endoscopists who use this system are not preferred due to its negative features, such as spending time to learn how to use it and costing more than routinely used devices. Therefore, academic studies with NBI are limited in their potential to turn into products in the commercial field [35]–[40]. In addition, some academic studies have shown that there is no difference in adenoma detection rates between the use of the NBI technique and white light colonoscopy [37]. Another imaging system is the chromoendoscopy, and the colon mucosa is colored with a special dye. Instead of chromoendoscopy and NBI systems, HD colonoscopy systems offer the opportunity to obtain images with higher quality and detail than standard images at reasonable costs. HD colonoscopy systems have image processing modes available. For example, an i-scan (Pentax) enhanced endoscopy image provides a real-time virtual image for a detailed view of mucosal and vascular structures with three different options highlighting specific anatomical features [41]. Another system that we used in this thesis: is Fujinon intelligent virtual chromoendoscopy (Fuji Intelligent Chromo Endoscopy, FICE) mode and high magnification. FICE digitally limits the wavelength range of light and offers images with brilliant color and light quality. A study by Sripatoorat and Angchuan [42] compared the pit pattern classification of 358 polyps with pathology outcomes using the high magnification FICE mode to predict the polyp's histology. High magnification FICE showed 82.1% accuracy in differentiating hyperplastic and adenomatous polyps. With this result, it has been shown that the FICE mode and high magnification make it easier to detect and classify colon polyps during

examinations. According to histological diagnosis, Wimmer et al. [43] used to classify colonic lesions as neoplastic and nonneoplastic. There is no comprehensive study comparing the FICE mode with other endoscopic imaging modalities on the automatic classification of colon polyps in the literature. We created the dataset for the literature for the first time by making comparative analyses between FICE mode and other imaging methods. Another novel aspect of this study is two different automatic classification problems with the features obtained from colonoscopy images and the use of the images themselves directly (as explained below in the deep learning approaches): 1) Polyps are not divided into two or three main types as in other studies (polyps or not, benign vs. malignant, etc.) it also be possible to divide them into subtypes. 2) As a unique study for colon cancer, classification be made by labels (expressed as a percentage of protein expression levels) obtained from IHC analyses on polyp tissues. It should also be noted that the aim of the thesis is not only classification, but also statistical comparison/association of various features obtained from the image with colonoscopy, histopathology and IHC analysis results. Thus, it is possible to detect biomarkers based on colonoscopy images, there are no examples in the literature. Optical biopsy is a cost-effective method because it is expected to save 33 million USD a year in health resources in the United States alone [44]. To develop these systems, it is critical to understand and quantify the differences in colonoscopy images.

## 2.4 Histopathology Images Studies

Examination of histopathological images is the gold standard for diagnosing and screening for cancers in the digestive tract. Digital pathology has become increasingly popular in recent years and allows the examination of high-resolution histopathological images. However, manual analysis of histopathological images is a serious, time-consuming task for pathologists because images can be up to 100,000x100,000 pixels. Developing medical models to classify tissues in different pathological images automatically is envisioned as a promising solution. Previous studies include predicting whether lesions in the colon are cancerous or normal by extracting features from microscopic images. Esgiar et al. [45] studied features based on a co-occurrence matrix (texture analysis) on microscope images (44 normal and 58 cancerous cancers) and achieved a success of 90%. Histopathological images have become a routine diagnostic method with rapid image processing techniques and visualization software. For each patient, histopathology images contain quantitative information but are not routinely used to extract prognostic markers. With these potential findings, Kather et al. [46] investigated that the

predictions of the CNN approach about the course of the disease can be extracted from the available histopathology images. This retrospective study demonstrated that a CNN could assess the human tumor microenvironment and predict prognosis directly from histopathological images. The dataset used in this study includes open-source histopathology images. Based on the output neuron activations in the CNN, they calculated the "deep stroma (the connective tissue skeleton of the organ) score", a multivariate parameter, and an independent prognostic factor for overall survival. This score has been shown to predict better survival than the Union Internationale Contre le Cancer (UICC) staging system. As a result of their research, manual measurement of stromal areas in the same group and gene expression of cancer-related fibroblasts are prognostic factors, specifically in tumor stages. Vascularization was chosen as another prognostic factor to be used in the thesis. Pathological angiogenesis has been identified as a potential prognostic factor and therapeutic target in malignancy. In most cases, angiogenic analysis is based on measuring microvessel density detected by immunostaining of CD31 or CD34 (clone QBend/10) [47]. The role of microvessels in histopathology images has not been extensively studied due to their complexity and heterogeneity. A comprehensive analysis of the polyp, which also has a colonoscopy and histopathology images, is made with this thesis. Furthermore, the manual identification of microvessels is a time-consuming and challenging task for pathologists due to interobserver variability. Therefore, it is necessary to develop automated microvessel detection algorithms in histopathology images for clinical analysis. The first study to develop an algorithm for this purpose was Yi et al. [47], using CNN as a microvessel prediction method. Experimental results have demonstrated the feasibility of the proposed algorithm on histopathology images. In addition, the microvessel features identified were significantly associated with patients' clinical outcomes. In this study, they used an open-source dataset mainly consisting of lung cancer. The detection and analysis of vascularity from histopathology images have been studied with lung cancer but have not been reviewed on colon cancer. Vascular analysis of the same polyp from colonoscopy, histopathology, and IHC images were created for the first time with this thesis.

Recently, deep learning algorithms have shown great promise in pathology image analysis applications such as tumor site identification, metastasis detection, and patient prognosis [48], [49]. It is important to study that many machine learning algorithms automatically segment pathology images, make predictions from the image, and extract predictions [50], [51], [52]. Among these algorithms, segmentation-oriented deep learning algorithms such as CNN stand

out with accuracy, computational efficiency, and generalizability [53]. Thus, deep learning-based segmentation has become an essential tool in the analysis of histopathological images.

## 2.5 Immunohistochemistry Studies

Histopathology imaging and gene/protein expressions are two fundamental research techniques that allow the analysis of biological samples from different perspectives. Budinska et al. [54] studied how gene expression can be used directly for the feature extraction from the image. This study was carried out on the breast cancer dataset. It shows that molecular data (expression of genes or proteins) can be combined with pathological information represented by a categorical variable (e.g., tumor grade, etc.). In this study [54], raw pathology data can be transformed by the expert into a qualitative variable (e.g., tumor grade: "well-differentiated", "moderately differentiated" and "undifferentiated"). With such an approach, tumor grade is combined with a genomic proliferation score to determine the prognostic factor in breast cancer. As a result of this approach, it is considered that new prognostic or predictive features can be identified if pathology images and molecular data are analyzed together.

Various image features and multimodal (imaging and genomic) approaches aim to search for prognostic biomarkers in breast cancer. Combining histopathology images with the biomarker detection workflow could potentially outcome in the identification of new image-based markers and the development or even replacement of existing genomic and proteomic markers. Additionally, extracting meaningful and distinctive image features with genomic (and clinical, etc.) data is difficult due to the complexity of the images [55]. In another study by Popovici et al. [56], the joint analysis of histopathology image features and gene expression in breast cancer was studied.

This thesis extracts features from images, associating and detecting colon polyp-specific proteins with expression analysis results. Proteins known to play a role in cancer genetics are available in the literature. For this thesis, some proteins to be investigated in IHC analyses were determined as p53 (clone bp53-11), Ki-67 (clone 30-9), CD34 (clone QBend/10), VEGF (clone SP125), BRAF (clone V600E), PD-L1 (clone SP142)[57], [58], [59], [60]. However, studies on image detection are not available in the literature. Here, the uniqueness of this thesis is the determination of protein reactions in polyps and, therefore, malignancy potential from colonoscopy and histopathology images. However, previous studies did not perform independent prognostic marker analysis specific to polyp stages and focused on colon cancer.

The role and importance of mutations of p53 (clone bp53-11) genes in cancer have been demonstrated in many studies. The p53 (clone bp53-11) protein encoded by the tumor suppressor gene TP53, which is mutated in 50% of cancer cases, is one of the most studied proteins [61]. The TP53 gene contains the genetic code for making a protein called the tumor protein p53 (clone bp53-11). This protein functions as a tumor suppressor; it keeps the growth and proliferation of cells under control [62]. Another protein, Ki-67 (clone 30-9) is an antibody that can show proliferating cells within the tumor. It is used as a proliferation marker in colorectal and bladder cancer [58]. Vascular Endothelial Growth Factor (VEGF (clone SP125)) is a signaling protein produced by cells that stimulate vasculogenesis and angiogenesis. It has been shown that this protein has a key function in tumor development and progression [63], [64]. In addition, the relationship of these proteins, which are known to play a role in cancer genetics, with breast and lung cancer has been studied [65], [66]. A study on the investigation of its effect on colon polyps [67], Programmed Death-Ligand 1 (PD-L1), lacks the literature. In this study, the expression of PD-L1 increases as the degree of dysplasia increases. In the dataset, we created, unlike this study, since there are polyps with different degrees of dysplasia, more comprehensive and scientifically valuable, and original studies can be put forward to make it possible to compare the healthy tissues taken from the polyp side. This part of the thesis made the study significantly different from the studies in the literature. Without such an extensive dataset enriched with the results of expression analysis of colon polyp-specific proteins, understanding this cancer and investigating the applicability of automatic classification approaches are always incomplete.

# Chapter 3

# Materials

## 3.1 Patients

201 patients, 119 male and 82 female, obtained from Kayseri City Hospital (KCH) Gastroenterology Clinic were included in the study. Cohort studies are concerned with monitoring and evaluating the results of a group of people who share a particular trait or within a specified period. These types of studies are used in research such as understanding cause-effect relationships, assessing disease risk, or evaluating the effect of a particular treatment or intervention. In this study, a group of people with colon polyps were observed between 2019 and 2021 and their results were followed. The inclusion criteria were as follows: Patients aged 19 to 89 years were included as eligible participants in the study. Attention was paid to the fact that these patients were clinically without any significant symptoms or irregularities other than the presence of colon polyps. Individuals diagnosed with colon polyps by the routinely used and gold standard colonoscopy procedure were included in the study. The identification of colon polyps was based on a visual assessment by the gastroenterologist during the colonoscopy procedure. The study obtained voluntary informed consent forms from all participants stating that they were ready to participate in the study. The research conducted in this study complies with the ethical standards and regulations that govern research involving human subjects. The study received ethical approval from both the KCH Medical Specialization Education Board and the Erciyes University Clinical Research Ethics Committee, under the number with 96681246 at 9.10.2019.

Inclusion criteria were meticulously selected to ensure a study population that was both diverse and homogeneous in terms of study, age, gender, and clinical presentation. The study aimed to obtain reliable data for analysis and interpretation by including patients with confirmed colon polyps while excluding individuals with a genetic family history or contradictory factors. The design of the inclusion criteria aimed to prioritize patient safety and ethical considerations while ensuring the validity and generalizability of the study findings.

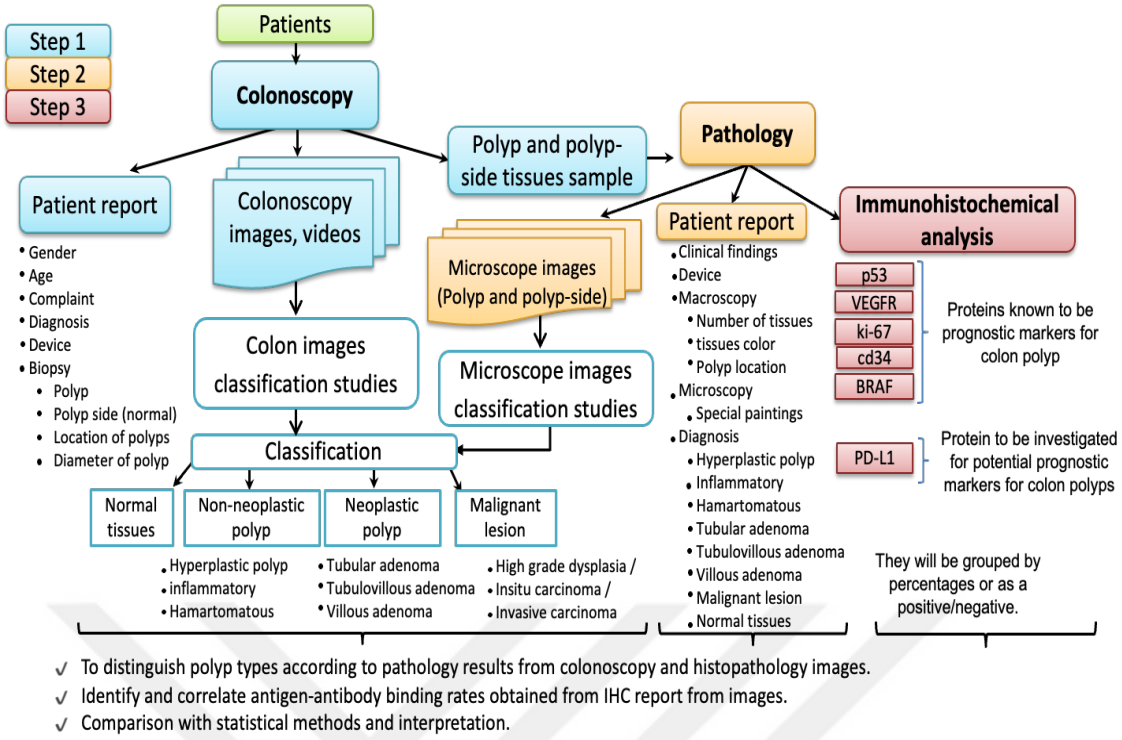



**Figure 3.1 The flowchart of this study**

According to the study plan presented in Figure 3.1, collecting data and creating a database was completed using patient reports, images, and videos obtained during the colonoscopy procedure, and patient reports and microscopy images acquired after the pathological examination. Additionally, IHC staining, and analysis were performed by the pathologist in KCH. The data collected are stored in an excel-based database.

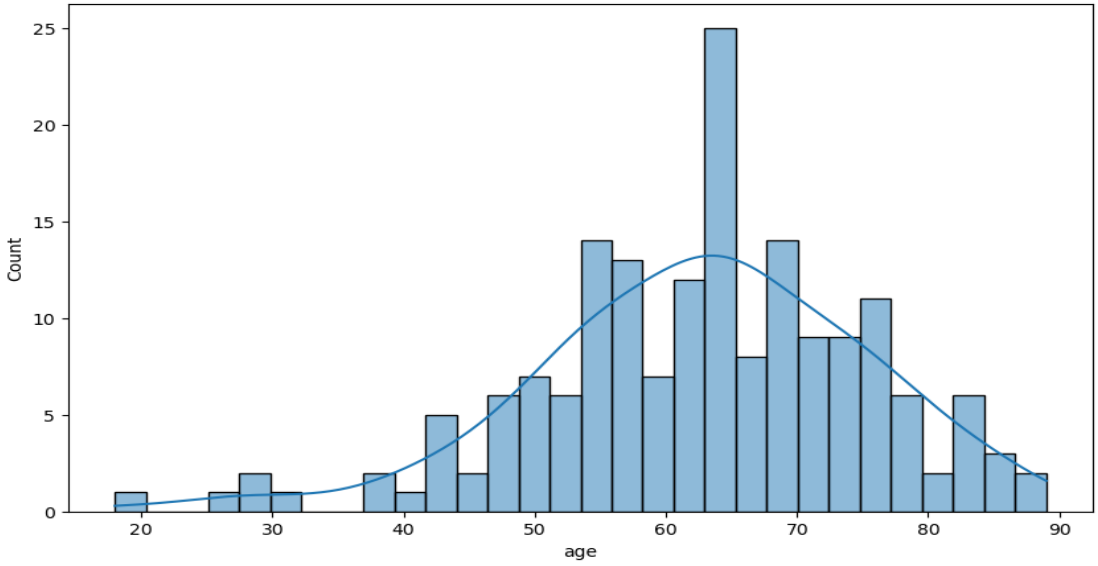



**Figure 3.2 Age distribution of patients**

The study included a total of 201 patients, with an average age of 62.74 (±12.39). There is a wide range of ages among the patients who were examined. Patients range in age from 18 to 91, with 18 being the youngest age group (Figure 3.2). This demonstrates the variability of the demographic structure under investigation. When we look at the age distribution of the patient population in more detail, the first quartile of age data is below the age of 55, while the median value is 63. This means that half of the patient population is 63 years old or younger. The third quartile value was determined at 71 years, indicating that 75% of patients were under 71 years of age. In light of these findings, we can observe that the patient population generally consists of older age groups. This demographic information obtained plays an important role in understanding the patient profile and conducting subsequent analyses correctly.

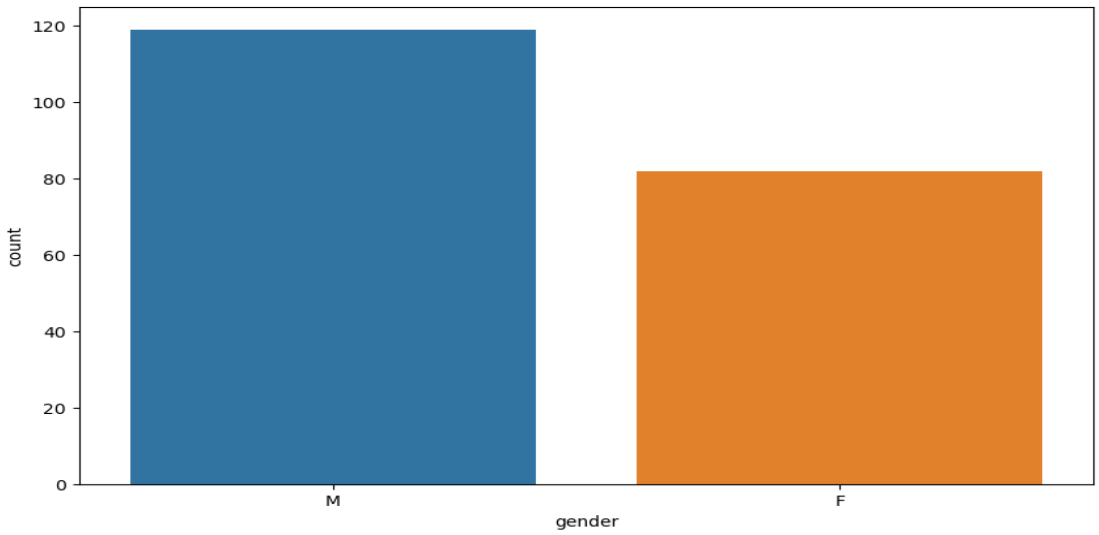



**Figure 3.3 Gender distribution of patients**

The Shapiro-Wilk test applied to age data is used to evaluate whether the data distribution is normal. The test result shows that the age data does not follow a normal distribution, since the p-value is less than 0.05. This means that the p-value is calculated as 0.0137, which is smaller than the generally accepted significance threshold of 0.05. This indicates that the age distribution of colon cancer patients deviates from the idealized normal distribution model used to understand clinical and demographic characteristics. These findings emphasize that the age distribution of patients is diverse, and this diversity should be taken into account to better understand the demographic profile of patients and the effects of the disease. In conclusion, analyzing the age distribution of the patient population and using this information to determine the course of the disease and treatment options is an important research and clinical practice step.

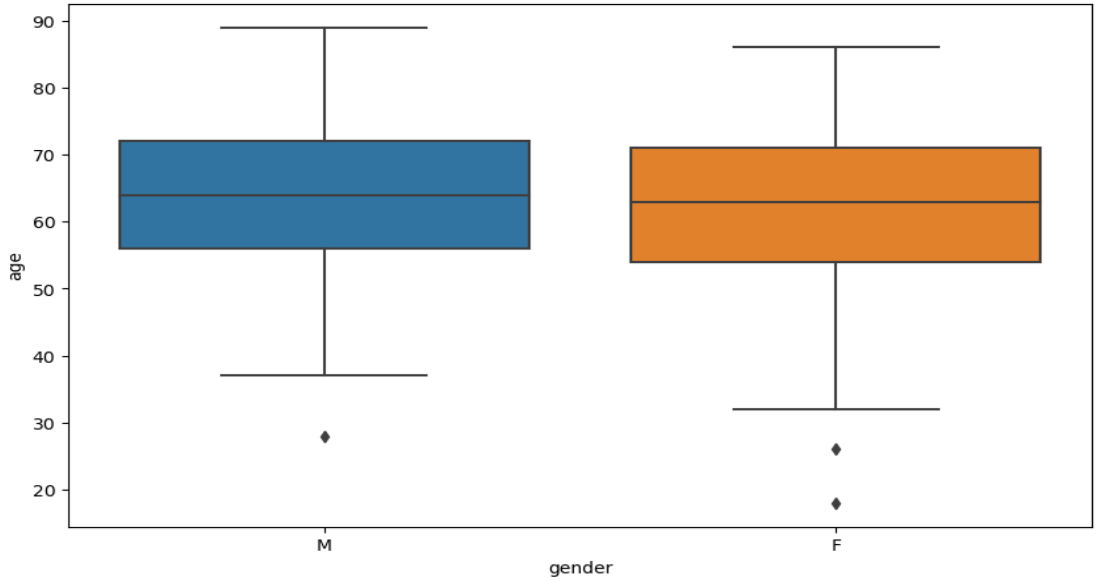



**Figure 3.4 Age distribution by gender**

The age distribution by gender is depicted in this graph. The discrepancies in age distribution seen between female (female) and male (male) groups are readily noticeable (Figure 3.4). The median age value is represented by the horizontal line in the center of the box graph. Women have a somewhat lower median age than men. The box graph's upper and lower boundaries depict the distribution of age values in the dataset. There is no significant variation in age values between men and women. The average age of women is, nevertheless, slightly lower than that of men.

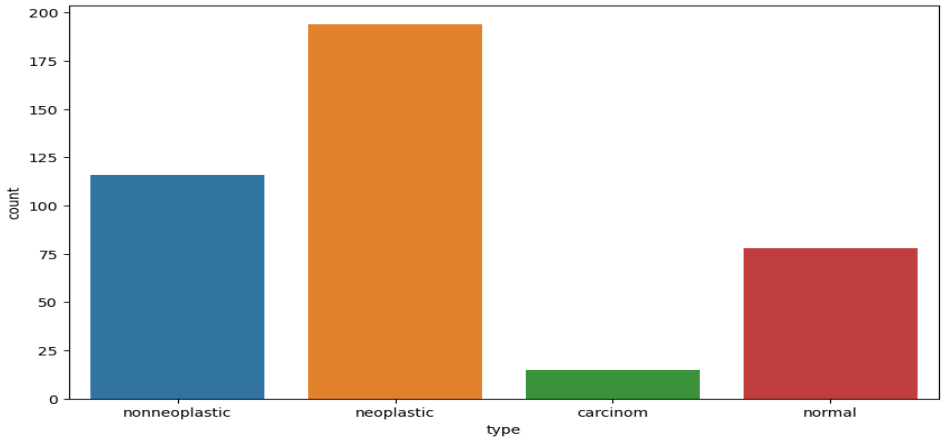



**Figure 3.5 Number of polyp types**

Analysis of tissue types showed that most of the polyps belonged to the neoplastic category. A total of 194 neoplastic polyps were detected. Nonneoplastic polyps were observed in 116 patients, while normal polyps were detected in 78 patients. In addition, 15 carcinomas were

found in a smaller group of patients (Figure 3.5). These findings highlight the diversity of polyp types in the study population and highlight the importance of evaluating and discussing each category separately in further analysis.

In this study, there is no class defined as normal in studies using colonoscopy videos because tissue samples labeled as normal from colonoscopy videos could not be detected accurately. Only microscope images and immunohistochemistry (IHC) results were used to determine the presence of normal tissues in the analyses. If it is desired to analyze normal tissues in the study, datasets consisting of microscope images and IHC results should be preferred. Because colonoscopy videos do not have labels for normal tissues, it is necessary to consult other datasets to analyze normal tissues. Thanks to this approach, it is possible to obtain more accurate and reliable results in an analysis of normal tissues.

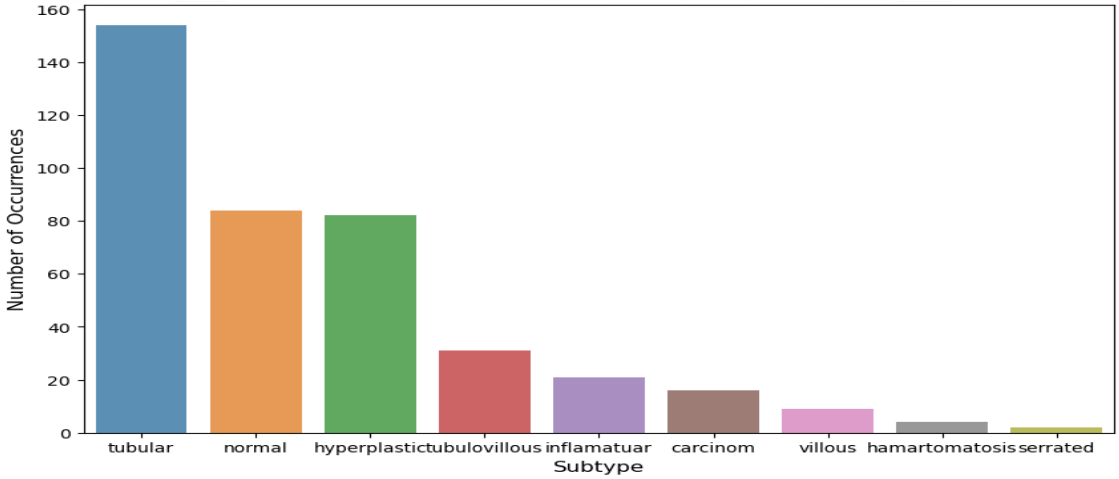



**Figure 3.6 Number of polyp subtypes**

The distribution of polyp subtypes in the study population is as follows: 16 cases of carcinoma, 4 cases of hamartomatous, 82 cases of hyperplastic polyps, 21 cases of inflammatory polyps, 84 cases of normal polyps, 2 cases of serrated polyps, 154 cases of tubular polyps, 31 cases of tubulovillous polyps, and 9 cases of villous polyps (Figure 3.6).

When the subtypes graph is examined, we can summarize the data. The most observed type of polyp was determined as tubular and a total of 154 of these types were found. The second most common type is the control group defined as normal and there are 84 samples in this category. Hyperplastic type polyps also take third place with 82. Among the less frequently encountered types of polyps, carcinoma is represented with 16 units, inflammatory with 21 units, tubulovillous with 31 units, and villous with 9 units. Hamartomatous and serrated types are rarely seen, 4 and 2 of these two species were found, respectively. According to the distribution of the data, we can say that there is significant variation in the frequency of polyp subtypes.

The most common tubular species has been observed more than any other species. These findings may guide further studies to examine and further understand polyp subtypes.

Using the same ID number, data were combined during data analysis to conduct a sustainable study on patients with comprehensive information. This method allowed us to determine that multiple data types, including videos, histopathology images, and IHC results, belonged to the same patient. Thus, polyp information was collected and analyzed from a unique source. This aggregation improved the consistency of the data and allowed for more reliable analyses. Using the same ID number to combine the data ensured data integrity and enabled a comprehensive review of each patient's data. This technique is essential to the study's validity and reliability. A more precise analysis of polyp characteristics and a more solid basis for conclusions were made possible by combining data from the same patient. It also reduced the problem of missing data, resulting in more consistent analysis results. This data aggregation procedure clarifies the study's methodology and makes the analysis procedure more transparent. It emphasizes the validity of the study by providing more information about how the data were combined and the dependability of the analyses.

## 3.2 Data Collection and Database Formation

The first step is to create a database using videos and reports from the patients who had a colonoscopy procedure in KCH. Producing pathological analysis reports generally takes a month after obtaining the tissue biopsy sample. According to the pathological report, each polyp is labeled with its diagnosis result. Firstly, colonoscopy video frames were extracted, and then we applied semi-automatic key-frame selection to use the informative frames from each video with a polyp. The frames that have no polyps were excluded, just the frames with polyp/s were used in the classification phase of the study. In the second step, the pathology reports and histopathology images (with different magnification levels) were obtained that belonged to the same polyp. Also, this information was appended to the database with the same polyp ID. After colonoscopy key-frame selection, each frame was manually labeled according to the pathology reports and applied same labeling procedure for histopathology images. The third step is the immunohistochemistry (IHC) studies, which were conducted in KCH. The IHC analysis performed by the pathologist is also completed this term. Protein staining results were collected as percentages and positive/negative labels for each polyp. This part also serves as the label in the further steps of this research to determine the prognostic marker.

Figure 3.7 shows a view of the first part of our database with colonoscopy, pathology, and IHC information combined for each polyp.

| | COLONOSCOPY | | | | | PATHOLOGY | | IHC | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | patient | polyp | gender | age | location | type | subtype | Ki67 | BRAF | PDL epitel | PDL lenfos | VEGF | cd34 | P53 |
| 1 | patient1 | polyp1 | M | 53 | descending | nonneoplastic | hyperplastic | 50 | negatif | negatif | negatif | 20 | 47 | 15 |
| | patient1 | polyp2 | | | descending | nonneoplastic | hyperplastic | 40 | negatif | negatif | negatif | 30 | 39 | 10 |
| | patient1 | polyp3 | | | descending | nonneoplastic | hyperplastic | 50 | negatif | negatif | negatif | 50 | 50 | 20 |
| 2 | patient2 | polyp1 | M | 49 | ascending | nonneoplastic | hyperplastic | 50 | negatif | negatif | zayıf | 40 | 63 | 5 |
| | patient2 | polyp2 | | | descending | neoplastic | villous | 60 | negatif | negatif | zayıf | 70 | 57 | 25 |
| 3 | patient3 | polyp1 | M | 70 | ascending | nonneoplastic | hyperplastic | 10 | negatif | negatif | negatif | 40 | 70 | 2 |
| 4 | patient4 | polyp1 | F | 57 | sigmoid | nonneoplastic | hyperplastic | 30 | negatif | negatif | zayıf | 75 | 50 | 25 |

**Figure 3.7 The comprehensive database created from KCH**

### 3.2.1 Obtaining Colonoscopy Videos



**Figure 3.8 Colonoscopy environment from KCH**

Colonoscopy videos and images were shot using high-definition (HD) colonoscopy technology at the KCH Gastroenterology Clinic. The Fujinon VP-3500HD processor and XL-4450 light source with a 300 W Xenon lamp were preferred in the system (Figure 3.8). HD colonoscopy devices are an essential feature in supporting studies in the literature by providing detailed vision and high image quality. Videos were recorded at a resolution of 1920x1080 pixels, providing high-resolution images for further research and analysis. An experienced gastroenterologist performed the colonoscopy procedures. During the procedure, a flexible colonoscope is inserted through the rectum, allowing a comprehensive examination of the inner lining of the colon. HD endoscopic device was used to record the entire colonoscopy procedure. Examination by the gastroenterologist has been the primary option for polyp discovery and characterization. Zoomed-in images were taken to precisely document the morphology and

characteristics of the polyps, and detailed polyp photographs were provided. Biopsy samples were taken from polyp areas using special equipment, such as forceps or snares, for histopathological examination.

In addition to obtaining images of the polyps, samples of normal colonic mucosa near the polyp were taken for comparison. These samples were taken from the locations around the polyps during the colonoscopy operation. The collection of polyp and normal tissue samples allows for a comprehensive examination and comparison of the two tissue types. Although these normal tissues are not available for every polyp, more detailed information is shared in the Histopathological image dataset section and indicated on the labels of the images. All colonoscopy video recordings and photographs are digitally securely preserved for further study and documentation. Appropriate measures have been taken to protect the confidentiality and privacy of patient data following applicable ethical and statutory laws.

Colonoscopy videos and images provided vital data for the characterization and examination of colon polyps using advanced HD technology and rigorous visualization techniques. High-definition videos and images made it possible to delve deeper into the size, shape, surface properties, and other important aspects of polyps. This comprehensive dataset serves as a foundation for additional analysis and research, providing a comprehensive understanding of colon polyps and their pathological consequences.

When a total of 202 colonoscopy videos were reviewed in our study, a total of 422 minutes and the average duration of each patient's video was determined as 2.14 minutes. However, the duration of each patient's video varied, and the standard deviation of these times was calculated as ±4.35 minutes. These data graphically indicate that the duration of each patient's colonoscopy video can be 2.14±4.35 minutes.

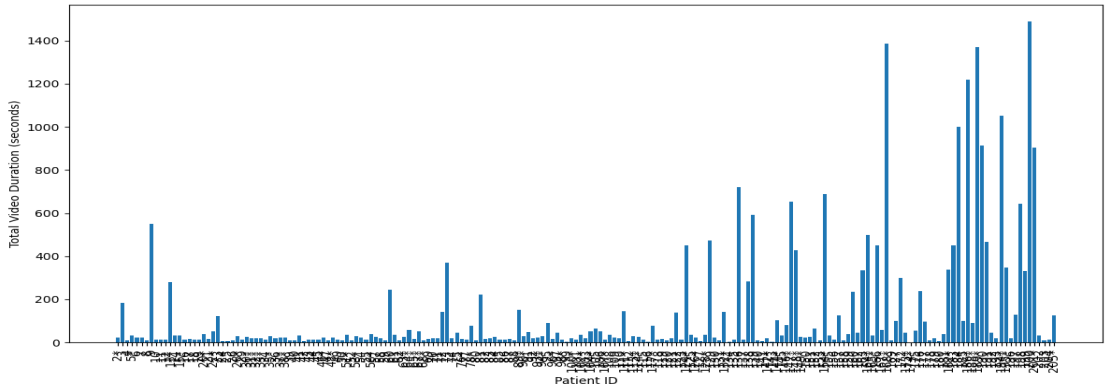



**Figure 3.9 Total video durations for each patient's videos**

Despite these data, such videos were excluded from our analysis due to the presence of more than one polyp in some videos and the inability to determine which polyp has which pathological result from the video recording. In addition, some videos were excluded from our analysis process for various reasons, such as videos that did not meet our quality standards, and videos that did not contain polyps. As a result, a total of 133 videos were selected and colon polyps in these videos were labeled. Labels have been added to video titles. When we reviewed the 133 selected colonoscopy videos, a total of 168 minutes, and the average duration of each patient's video was 1.29 minutes (Figure 3.10). However, the duration of each patient's video varied, and the standard deviation of these times was calculated as ±2.35 minutes. These data graphically indicate that the duration of each patient's colonoscopy video can be 1.29±2.35 minutes.

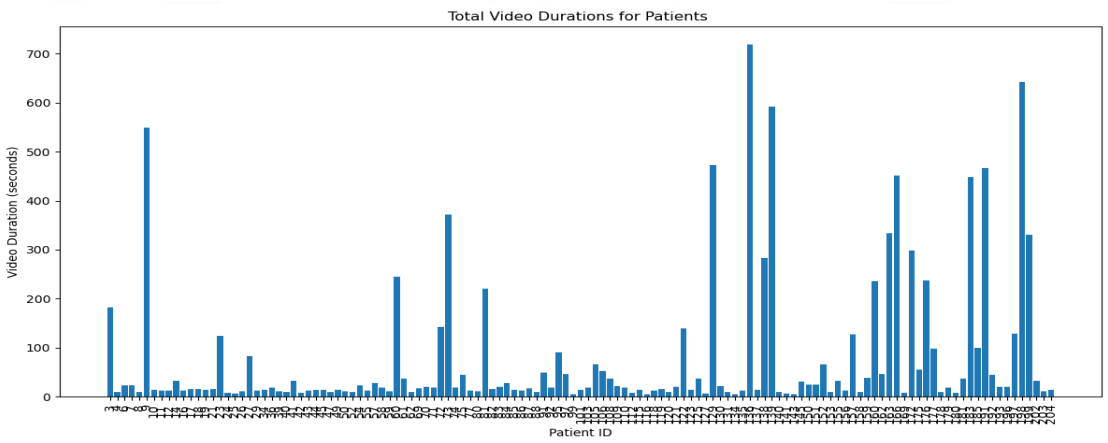



**Figure 3.10 Total video durations for selected patient videos**

However, information about the patient and the video was also provided in the Excel file. This information includes patients' demographic information and pathological results, as well as colonoscopy videos. This comprehensive set of information facilitates the use and analysis of colonoscopy videos while protecting patients' privacy. The specific study and data collection process may set a new standard for the analysis of colonoscopy videos and constitute a valuable resource for future research. Moreover, this data set's accuracy and reliability significantly contribute to the labeling and identification of colon polyps. This can improve the quality and accuracy of colonoscopy practices, thereby improving patient health outcomes.

Figure 3.11 shows sample frames from colonoscopy videos of different patients in the database. These frames contain various examples that provide essential information about the multiple characteristics of colon polyps. Each video offers a variety of images taken during the colonoscopy, including images of various colon polyps.

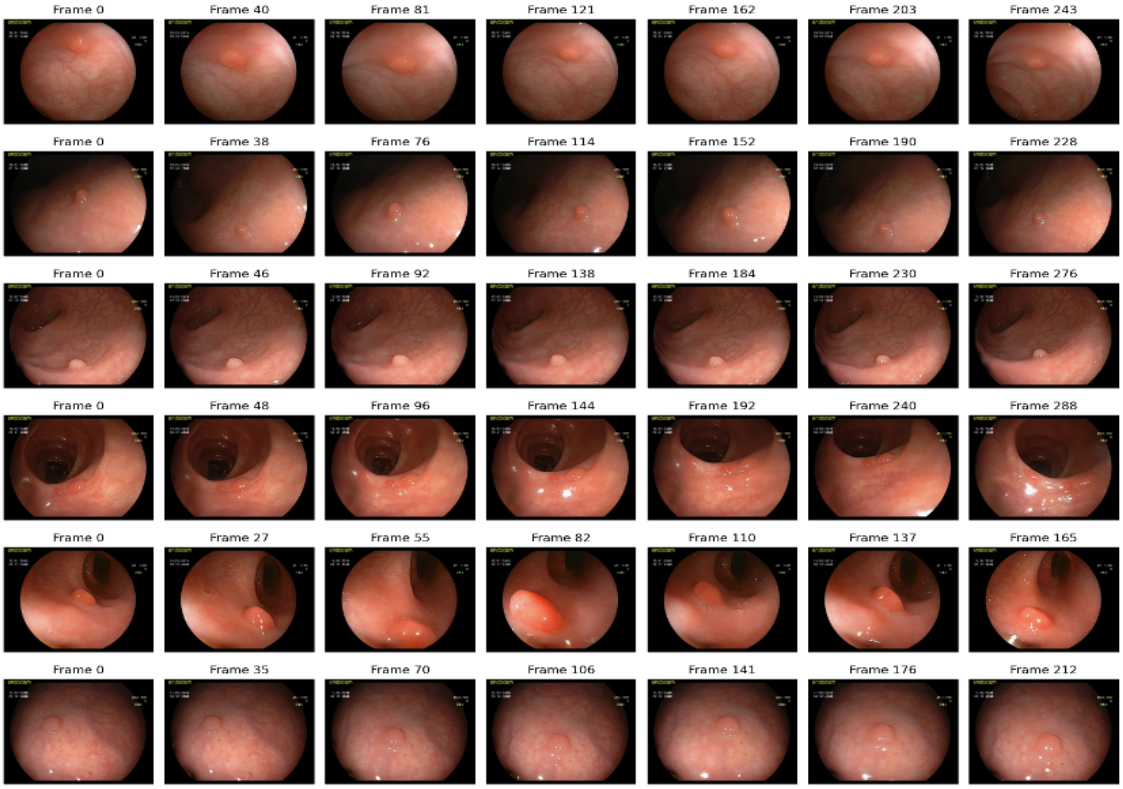



**Figure 3.11 Extracted frames from the selected videos**

These visual examples comprehensively cover colon polyp images, demonstrating that the database constitutes a large and diverse collection (Figure 3.12).

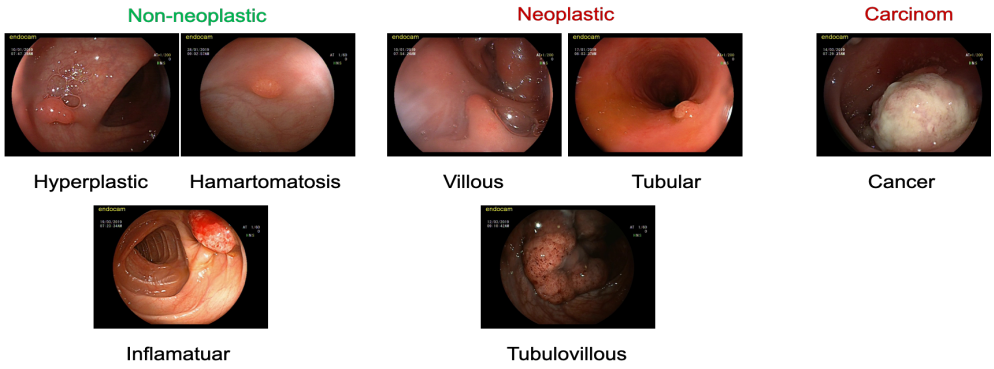



**Figure 3.12 Examples from different polyp types**

This visualization also provides information about the overall quality and resolution of our videos. High-quality and resolution images play an important role in increasing the accuracy and reliability of automated systems used for polyp detection and classification. These images provide researchers and clinics with an opportunity to better understand the appearance of colon polyps and the nature of the images to be used for training and evaluating automated polyp detection and classification systems.

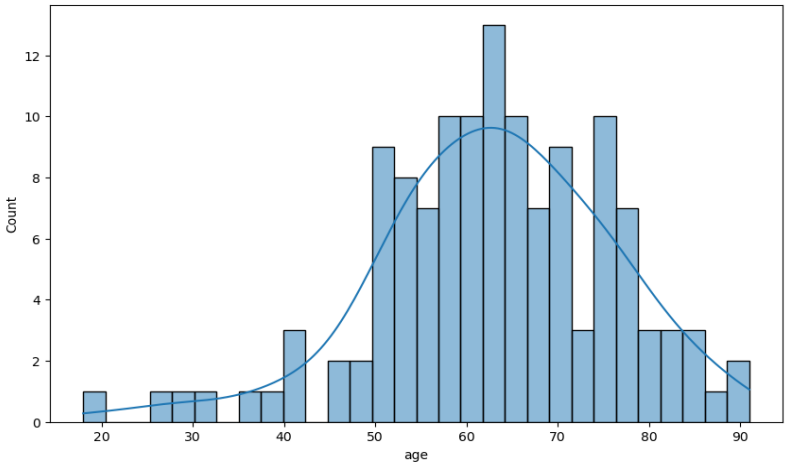



**Figure 3.13 Age distribution of the patients whose videos were selected**

In the task related to the colonoscopy videos, we have 133 polyp videos (72 male and 61 female) from 132 unique patients. These patients' age distribution is explained in Figure 3.13 and the mean age is 62.94±12.96 years.

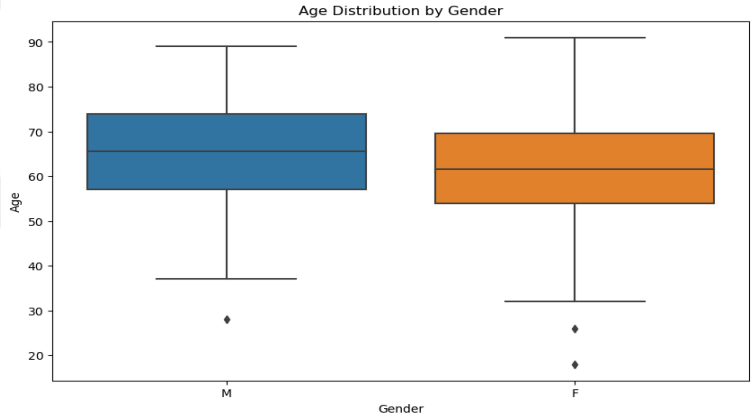



**Figure 3.14 Age distribution by gender colon video input dataset**

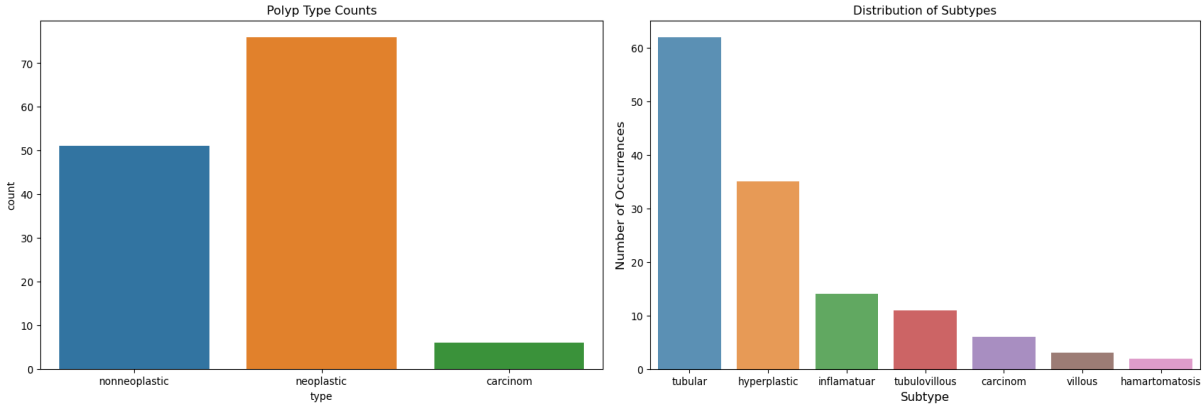



**Figure 3.15 The distribution of polyp types and subtypes in colon video dataset**

Figure 3.14 shows the age distributions of male and female genders with a box plot. Figure 3.15 indicates the distribution of polyps in the dataset according to their types and subtypes in the classification/characterization studies to be done using column videos. There are 133 colon polyps in total. Of these polyps, 76 were identified as neoplastic, 51 as nonneoplastic, and the remaining 6 as carcinomas. This video set represents the diversity of colonic polyp types, with tubular polyps being the most represented and hamartomatous and villous polyps being the least represented. This distribution provides important information about how the dataset can be used to train classification models. Class imbalance in the dataset can affect the model's ability to predict different classes and should be taken into account. As a result, this database of colonoscopy videos provides a valuable and comprehensive resource in colon polyp classification research.

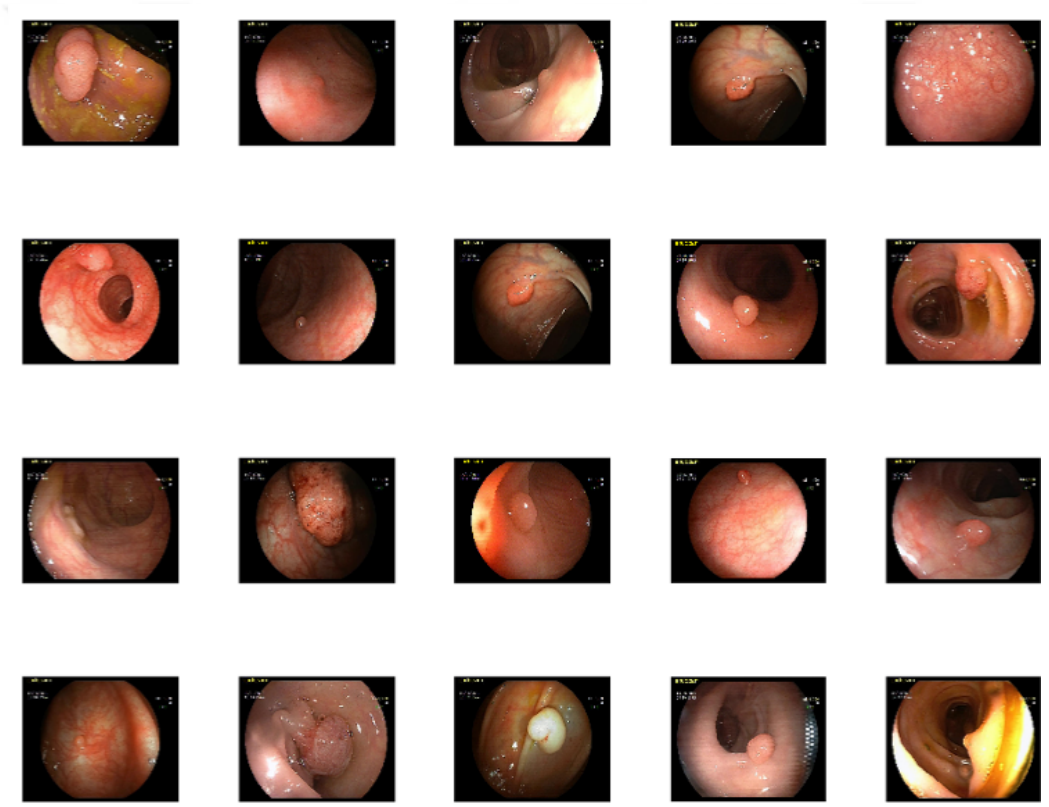


**Figure 3.16 Examples of frames containing polyps from the colonoscopy dataset**

Figure 3.16 shows several frames containing polyps from a colonoscopy dataset. In this way, the different sizes, shapes, and colors of polyps are showcased, giving a broad perspective of how they look visually. Each frame reveals in detail the location of the polyp on the colon wall and its relationship with its environment. Such visualizations may be important for training and validating algorithms used to detect and classify polyps. These frames also illustrate the diversity of polyps and the challenges that may be encountered during colonoscopy.

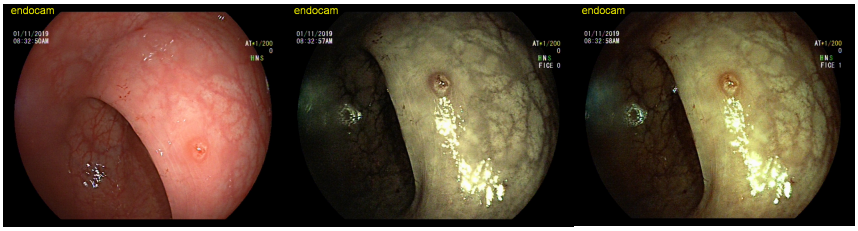



**Figure 3.17 Colonoscopy images of the same polyp of a patient using different imaging modes, from left to right white light, FICE 0, and FICE 1**

Figure 3.17 shows colonoscopy images of the same polyp of the same patient obtained using different imaging modes, with white light, FICE 0, and FICE 1 modes, respectively. In white light mode, the polyp is imaged using a standard light source, while FICE 0 and FICE 1 modes provide more detailed images of the polyp using light of different wavelengths. These modes help better distinguish textures and lesions. Each mode highlights different aspects of the polyp, providing additional information useful for diagnosis and treatment. This figure visually demonstrates the effectiveness of different imaging techniques in detecting polyps and how these techniques can produce different imaging results.

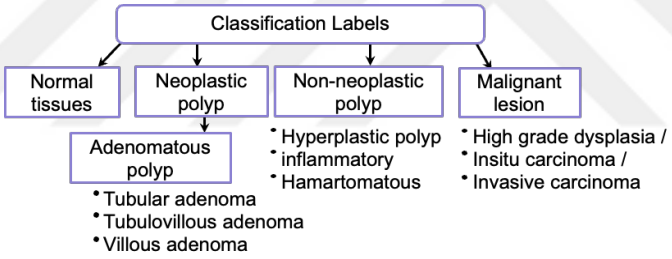



**Figure 3.18 Labels used in the classification of polyps**

According to the pathology result report, we labeled each polyp. We have two different classification tasks; one of them is the general classification like neoplastic or nonneoplastic; the other consists of polyp subtype classification. Figure 3.18 explains our labeling protocol for classification tasks. Normal tissues are coming from a location on the colon next to the polyp as a control tissue during the colonoscopy procedure; therefore, we do not have a normal tissue frame in the colonoscopy videos. However, we obtained these normal tissues in the histopathology images under the microscope. In our database, we have malignant lesions, but we did not use these lesions in the classification study. The malignancy of the polyp is obvious from the colonoscopy images and clinicians would immediately decide on the polyp type after a short visual examination on the screen, therefore, including the malignant lesions in the classification task does not have any clinical meaning. That is why we decided to focus on the discrimination of neoplastic and nonneoplastic polyps which has a great clinical significance.

After frame extraction from videos and labeling each informative frame, we created a file with a name that includes critical information about the polyp, such as patient number, polyp number, location of the polyp on the colon, type, and subtype information, and frame number. Figure 3.19 depicts the several image filenames that we used in our database.

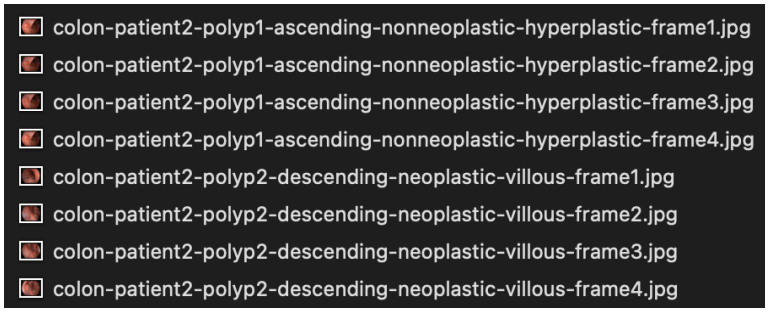



**Figure 3.19 Names of colonoscopy frames**

## 3.2.2 Preprocessing of Colonoscopy Images

The preprocessing part focuses on two different techniques to improve the quality of colonoscopy images.

### 3.2.2.1 CLAHE

We first investigated the effect of image content on the classification performance. For this purpose, we prepared four different colonoscopy image sets; original images, the same images without the black border, cropped polyp with 40% background, and cropped polyp with no background (Figure 3.20). Figure 3.21 shows the results of preprocessing applied on the same images using the approach called Contrast Limited Adaptive Histogram Equalization (CLAHE) [68]. This was utilized to minimize the contrast problems and make polyp details more visible.

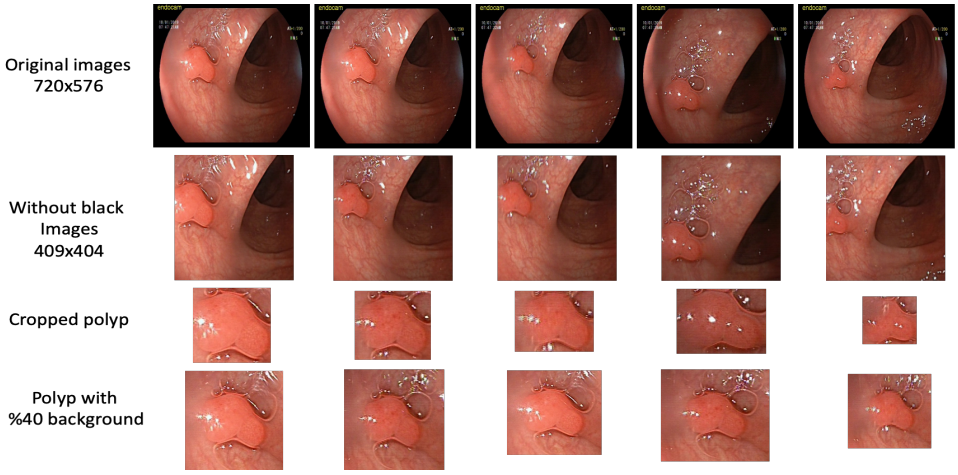



**Figure 3.20 Sample frames from one polyp with different focusing levels**

The first row shows six original frames/images from one polyp in the database, the second row depicts the same images with black border removal, the third row shows the image focused only on the polyp, and the last row depicts the images with 40% background and the polyp.

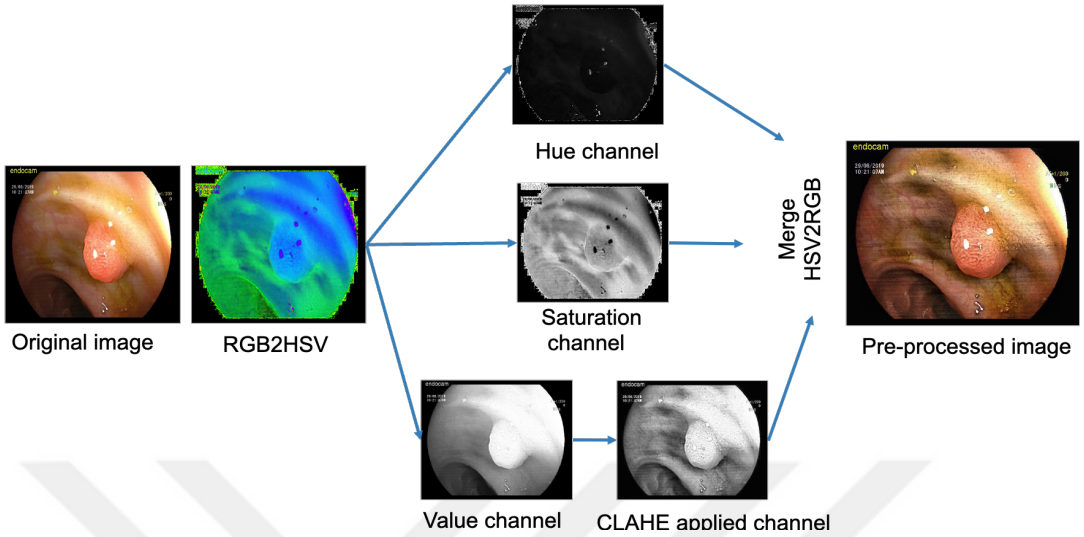



**Figure 3.21 Colonoscopy frames preprocessing**

During the preprocessing process of the colonoscopy frames, images in RGB (Red-Green-Blue) format were first converted to HSV (Hue-Saturation-Value) format. This transformation is a frequently used technique in image processing and is based on hue, saturation, and value (brightness) components. The HSV format is preferred in some image processing applications because it better reflects the human eye's color perception.

After conversion to HSV format, CLAHE was applied on the 'V' (Value) channel. CLAHE is a method that improves the contrast of the image and is especially effective in making details more visible in low-contrast images. This method is an advanced form of histogram equalization, limiting contrast to prevent excessive contrast enhancement in certain regions of the image [68]. In the final step, the processed 'V' channel was integrated back into the HSV format image and this updated HSV image was converted back to RGB format. This conversion provides an enriched image in the original RGB format with increased contrast and improved details. These preprocessing steps were performed to improve image quality for the detection and analysis of polyps in colonoscopy frames.

Figure 3.22 contains the same frames that have been pre-processed and have different focus levels. Each image in the figure shows different focus levels of the same frame taken during the colonoscopy. The preprocessing process can be applied to improve image quality or make certain features more prominent. During this process, images can be sharpened, contrast increased, or color balance adjusted.

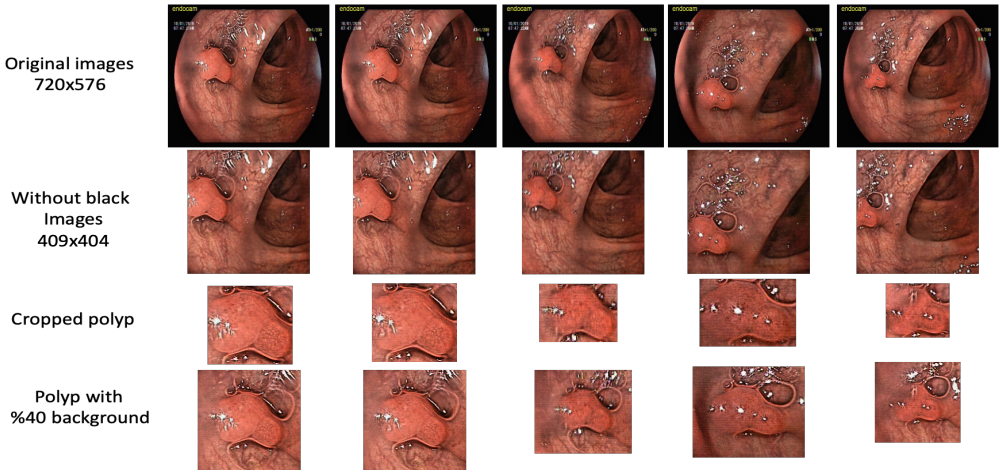



**Figure 3.22 Same frames with different focusing levels, after preprocessing**

Having different focus levels means that each image allows a specific area to be examined in more detail. This is especially important when examining polyps or other lesions. Different focus levels reveal depth information of the tissue, allowing for a more detailed analysis. Such images are especially important for specialists working in medical imaging and diagnostics because images at different focus levels more clearly show the structure of tissues and potential abnormalities. In this way, a more accurate and effective diagnosis can be made.

### 3.2.2.2 Dehazing

The dehazing [69], [70] algorithm was applied to the selected images for contrast enhancement as a preprocessing. Examples of the results of this process are in Figure 3.23. The previous method CLAHE has unwanted noisy parts that were also highlighted by the algorithm. On the other hand, dehazing just arranges the haze thickness, using this logic we can enhance the contrast of the dark parts like the lumen.

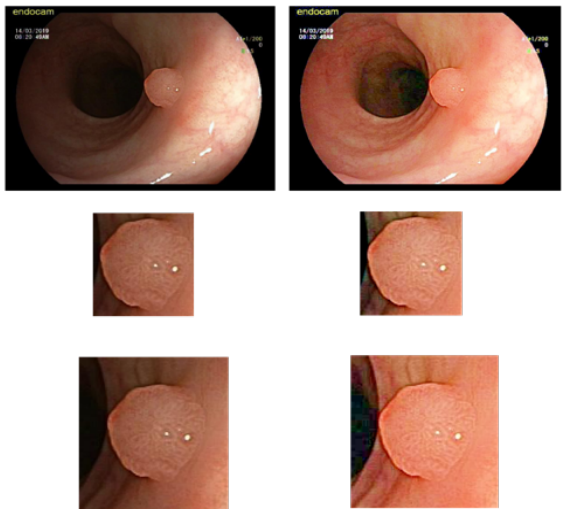


**Figure 3.23 Before and after preprocessing**

Figure 3.23 shows how images change before and after applying these preprocessing techniques. While images before preprocessing may be blurrier and less clear in detail, images after preprocessing will be clearer, have higher contrast and details will be more clearly visible. These changes help make polyps and other important tissue features easier to detect. In this way, the accuracy and efficiency of medical imaging and diagnostic processes are increased.

### 3.2.3 Obtaining Histopathology Images

In this study, clinical characteristics of polyp samples such as gender, location, type, subtype, BRAF (clone V600E), PD-L1 epithelial, and PD-L1 lymphoma were examined in detail. As a result of the analyses performed on 201 patients within the scope of the study, it was observed that the majority of the participants were male (59.2%). It has been determined that polyps are seen in different residential areas and are most frequently located in the descending area.

Type and subtype analysis of polyps makes an important contribution to a better understanding of their histopathological features. In the study, it was determined that the most common type of polyp was neoplastic and that these polyps had various subtypes. The tubular subtype was the most identified subtype in the study.

In order to evaluate the genetic and immunological characteristics of polyps, we focused on molecular markers such as BRAF (clone V600E), PD-L1epith, and PD-L1lymph. The results showed that the BRAF (clone V600E) mutation was generally negative, while PD-L1 epithelium and PD-L1 lymphocyte expression were generally negative or weak. These findings provide an important basis for better understanding the clinical and histopathological features of polyps and developing treatment strategies. The research results enable the analysis of a large dataset to understand the various features and molecular profiles of polyps.

A digital slide scanner was used to capture images at different magnifications such as 20X and 40X. Patient data includes demographic information such as age and gender, as well as pathology results while ensuring patient confidentiality is protected. Pathological diagnosis was made by two experienced pathologists. The image of each polyp is labeled with a diagnosis such as hyperplastic polyp, adenomatous polyp, or carcinoma.

Our dataset is divided into four main categories: neoplastic, nonneoplastic, normal, and carcinoma. These categories represent pathological conditions determined as a result of microscopic examination of patients' tissues. The largest group is neoplastic with 940 samples. The nonneoplastic category represented the second largest group with 519 samples, while the normal and carcinoma categories were less represented with 370 and 74 samples, respectively.

There are also more specific classifications under the neoplastic category. These subcategories are carcinoma, hamartomatous, hyperplastic, inflammatory, normal, tubular, tubulovillous, and villous. When the distribution of these subgroups is examined, it is seen that the largest group is tubular with 738 samples, while the smallest group is hamartomatous with 21 samples. These class distributions show the diversity of the dataset and the detailed classification options. This diversity enables the model to learn a wider range of pathology and thus be used in a wider range of applications. This is important for us to determine how effective our model is in classifying colon polyps.

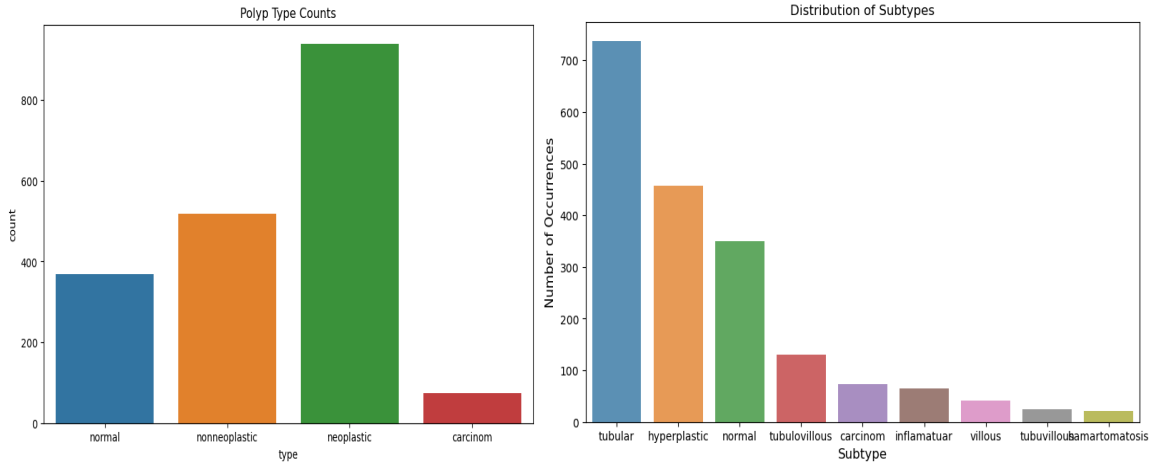



**Figure 3.24 The distribution of polyp types and subtypes of histopathology images input dataset**

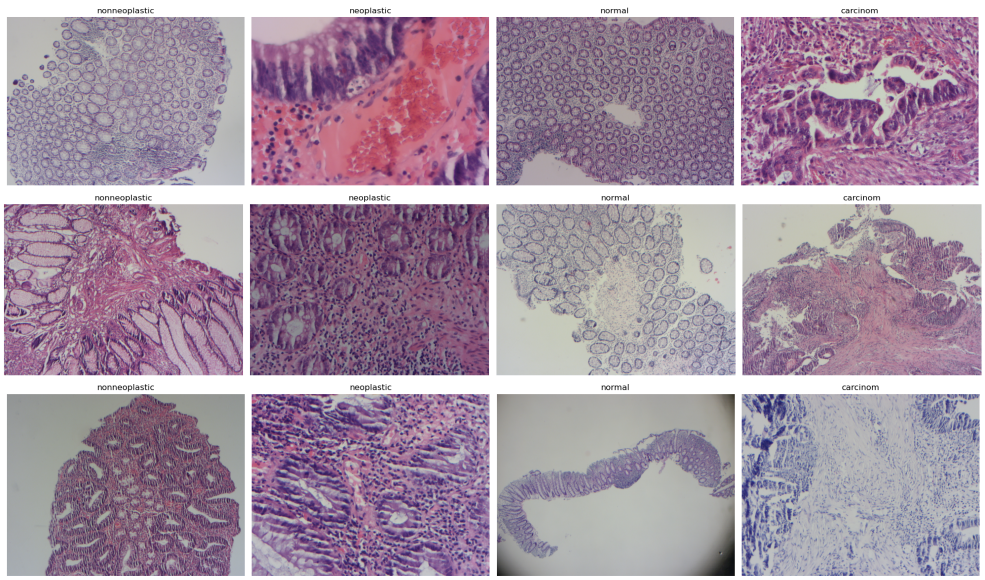



**Figure 3.25 Nonneoplastic, neoplastic, normal, and carcinoma examples from the input image set**

Figure 3.25 contains “nonneoplastic” (benign), “neoplastic” (malignant), “normal” and “carcinoma” (cancerous) samples taken from the input image set. Each category represents

different types of tissue that may be encountered during colonoscopy. Each category displays the typical appearance and characteristics of the relevant tissue or lesion type, providing critical information in diagnosis and treatment. This figure highlights the importance of distinguishing different tissue types and making a correct diagnosis.

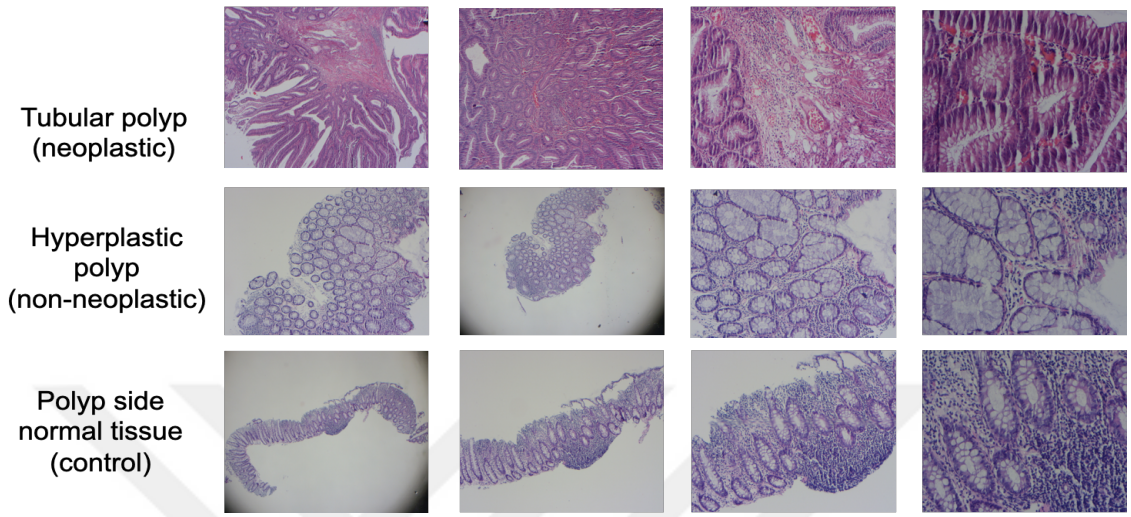



**Figure 3.26 Different polyp types from the same patient with all magnification levels(2.5X, 5X, 10X, 20X from left to right)**

Figure 3.26 shows various types of polyps from the same patient, images at different magnification levels. These images represent magnifications of 2.5X, 5X, 10X, and 20X, respectively, from left to right. Each level of magnification offers the opportunity to examine the different details and textures of the polyps more closely. 2.5X magnification level shows the general structure and the relationship of the polyp to the colon wall. Provides a wide view of the polyp. 5X magnification level, the more detailed structure of the polyp is revealed. Texture features and surface structure become more evident. 10X magnification level, the microscopic details of the polyp can be seen more clearly. Cellular structures and tissue patterns can be examined in more detail. 20X magnification makes it possible to examine the finest details of the polyp. At this level, even changes and abnormalities at the cellular level can be observed. Such detailed images are significant in identifying and classifying polyps. Different magnification levels reveal different aspects of the tissue, providing critical information for a more accurate diagnosis and treatment planning.

Within the scope of this study, microscopic images were obtained in a total of five different subfolders at different magnification levels. The folders are named 2.5X, 5X, 10X, 20X and 40X, respectively. The number of images in each subfolder is as follows: 2.5X folder contains 393 images, 5X folder contains 397 images, 10X folder contains 400 images, 20X folder

contains 399 images, and 40X folder contains 316 images. It aims to summarize the number of images obtained at different magnification levels by showing this information in Figure 3.27.

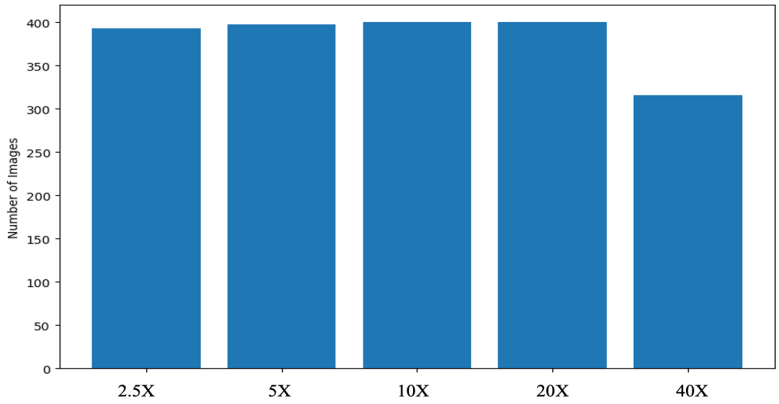



**Figure 3.27 Number of histopathology images with different magnification levels**

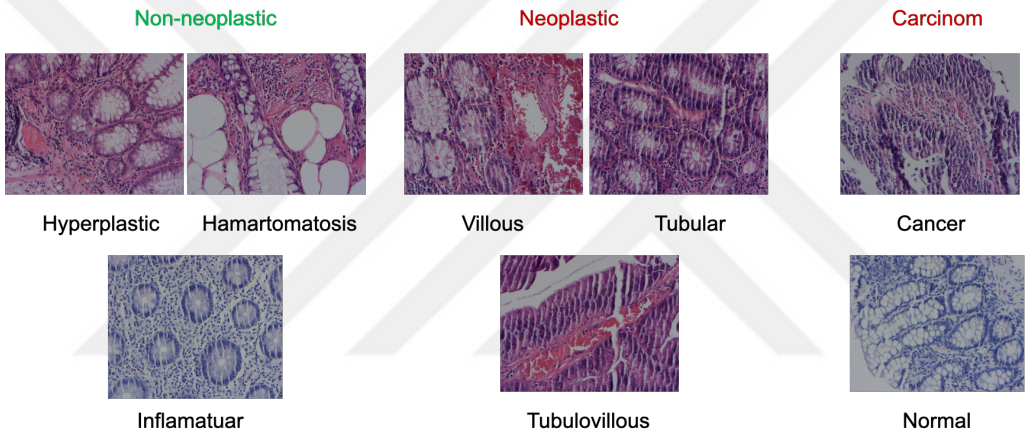



**Figure 3.28 Sample histopathology images from different polyp types/subtypes.**

The 2.5X magnification level provides the most general view of the tissue, revealing large-scale structures and the overall texture layout, while the 5X, 10X, 20X, and 40X magnification level images provide higher resolution, allowing cells and cellular allow for a more detailed examination of the structures (Figure 3.28).

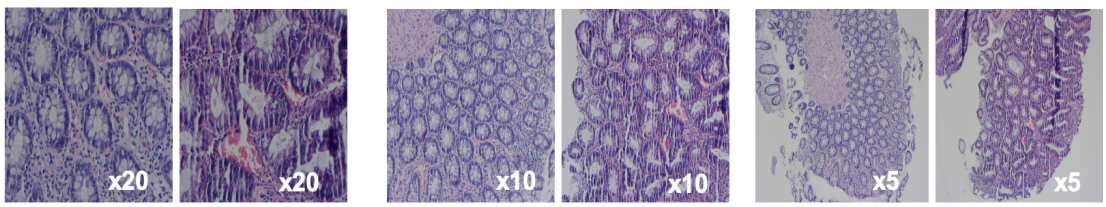



**Figure 3.29 Sample images of tubular adenoma polyp and normal tissue samples taken from the side of the same polyp, respectively**

Figure 3.29 shows the diversity of the various tissues in the dataset, which are examined under the microscope at different magnification levels, this diversity provides a wide field of view

for pathological analysis and diagnosis. The dataset includes both normal and pathological tissues, making it an ideal resource for training and testing machine learning models. In conclusion, the histopathological image acquisition procedure, which consisted of H&E staining and subsequent histopathology examination, enabled a thorough characterization of colon polyps. The high-resolution images acquired under the microscope were useful for describing the histological features identified in the polyp tissues.

The histopathology image dataset preparation process started with labeling. We labeled and arranged the histopathology slides/images in our database by assigning informative filenames. The names include similar information to the colonoscopy image filenames, but the magnification level is also made available for further analysis (Figure 3.30).

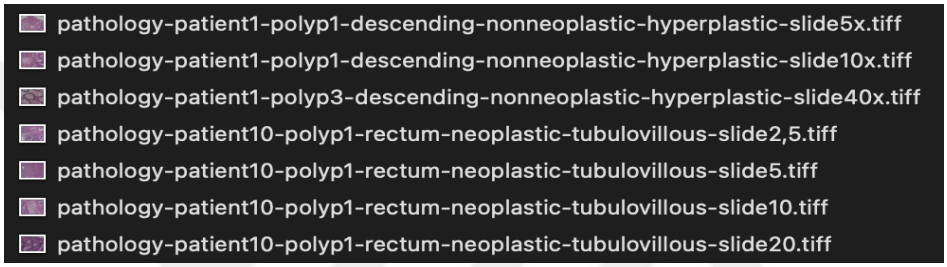



**Figure 3.30 Filename format for the histopathology slides/images**

### 3.2.3.1 Preprocessing of Histopathology Images, Color Normalization

The labeling and arranging of histopathology slides/images were finished in this database by assigning informative filenames. The filenames include similar information to the colonoscopy image filenames, but the magnification level is also available for further analysis. The color normalization was applied to all histopathology images; these images were also saved as files. Deep learning was applied to raw images without normalization. The result chapter shows classification results for these images

Before the feature extraction, image normalization is performed by taking into account histogram equalization, CLAHE image enhancement approach, and different luminance and staining factors. There are color normalization studies in the literature, especially histopathology images [71]–[73].

In Figure 3.31, a selected target image and three different source images using the Reinhard normalization technique [71] on both microscope and colonoscopy images. In addition, in deep learning studies on colonoscopy images, all different datasets are preprocessed. As a first step, the images are rescaled according to the input size of the desired model. Next, the pixel values in each image are normalized by subtracting the average pixel value from all pixels in the image and dividing the resulting values by the standard deviation measured in all pixels.

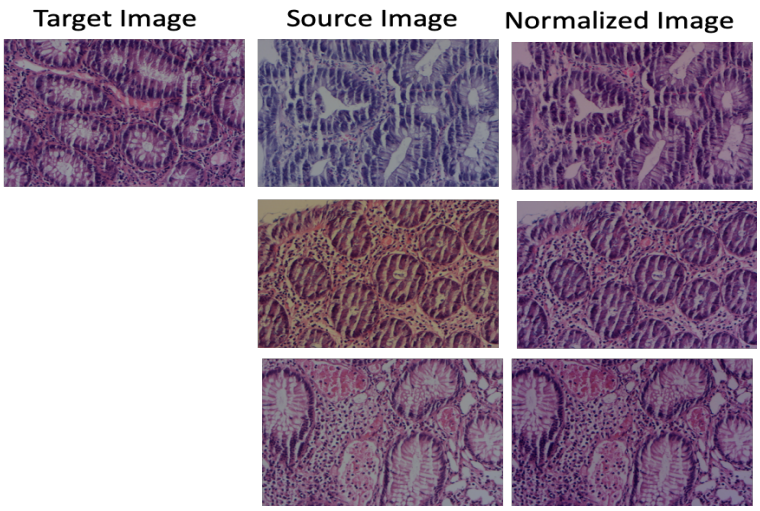



**Figure 3.31 Application results of the Reinhard normalization technique on our images**

This normalization process is called "z-score". This preprocessing and normalization approach allows us to apply the same neural network at different resolutions and data from various sources without further adjustments.

### 3.2.4 Immunohistochemical Analysis with Prognostic Markers

Immunohistochemical studies of specific proteins linked to colon polyps created a large and varied data set that included molecular and visual information. This particular data set was prepared using paraffin blocks and preparations from colon polyp cases. Sections cut from each paraffin block were stained immunohistochemically using antibodies specific to certain proteins. This multimodal dataset includes colonoscopy recordings, histopathology images, and immunohistochemistry staining results.

The methods used to ensure optimal antibody-antigen interaction followed typical cell conditioning procedures. Each primary antibody was carefully selected to ensure selective binding to target proteins in tissue samples. Antigen-antibody complexes were observed using secondary antibodies and a chromogen detection method. This special multimodal data set provided the opportunity to examine protein expression patterns and tissue structure simultaneously.

The resulting data set constitutes a valuable resource for computer vision and molecular-based researchers. This dataset combines the skills of computer scientists and molecular researchers, enabling interdisciplinary collaborations that will allow a comprehensive study of colon polyps.

Additionally, analysis of molecular markers provides essential information for understanding polyps' different characteristics and molecular characteristics. For example, the Ki-67 marker

emerged as an indicator reflecting the proliferation rate of polyps. VEGF and CD34 markers are associated with the vascular structures of polyps, and the levels of these markers provide important information about the vascular activity and status of polyps. CD34 score and p53 marker play an important role in the characterization of polyps because CD34 score expresses the degree of vascularization, while p53 is considered an indicator of genetic changes. This comprehensive data set offers researchers a broad evaluation opportunity to understand type of polyps, histological subtypes, and molecular profile. This dataset has the potential to make an impact in diverse fields such as cancer research, medical imaging, and computational pathology.

In this study, as a result of the analyses performed on 201 patients, it was determined that most participants were male (59.2%). It has been determined that polyps are seen in different residential areas and are most frequently located in the descending area. Types and subtype analysis is emerging as an essential component for understanding the histopathological features of polyps. The study observed that the most common type of polyp was neoplastic and that there were various subtypes of these polyps, and the most common subtype was defined as tubular. Molecular markers such as BRAF (clone V600E), PD-L1epith, and PD-L1lymph were examined to evaluate polyps' genetic and immunological characteristics. The findings showed that BRAF (clone V600E) mutation was frequently negative, while PD-L1epithelium and PD-L1lymph expression was generally negative or weak. These results provide an essential basis for better understanding polyps' clinical and histopathological features and developing treatment strategies. The study results enable the analysis of a large dataset to understand polyps' various features and molecular profiles.

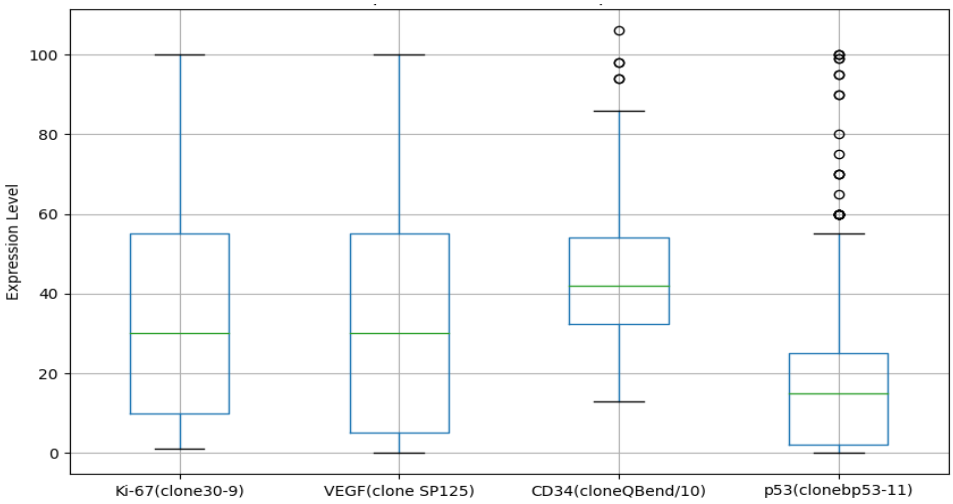



**Figure 3.32 Comparison of biomarker expressions**

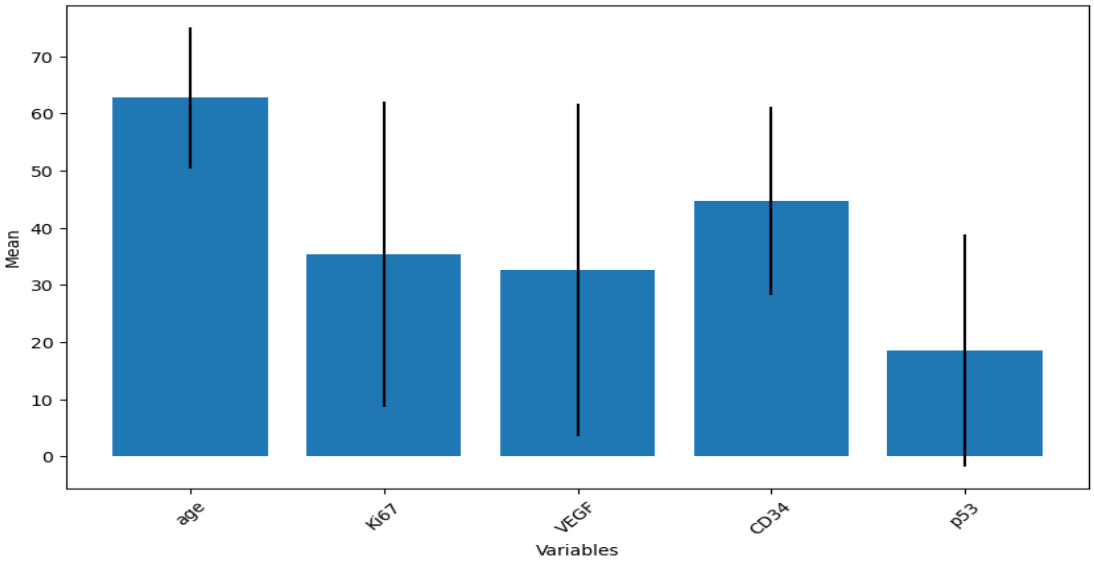



**Figure 3.33 Mean values of numerical variables**

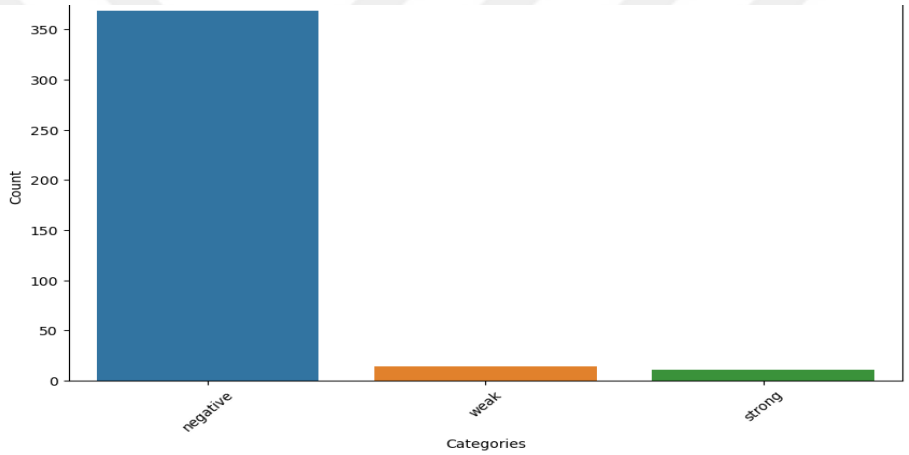



**Figure 3.34 BRAF (clone V600E) distribution**

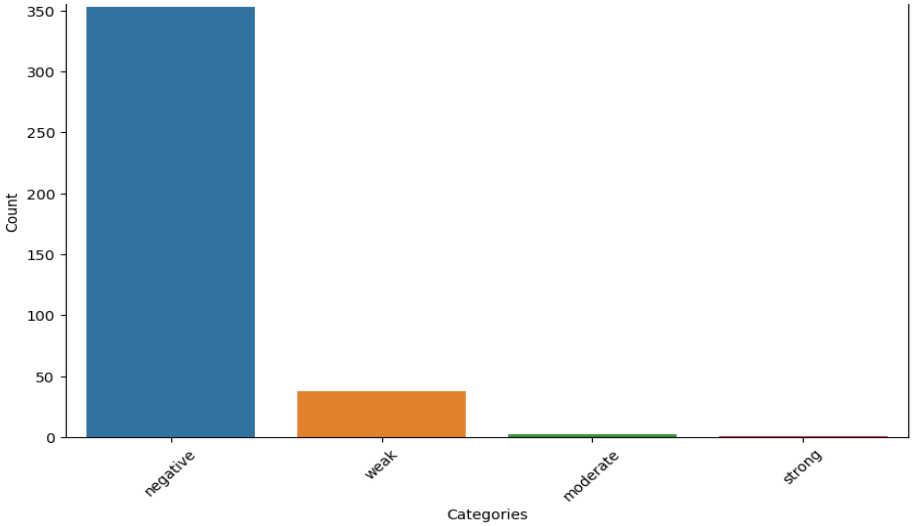



**Figure 3.35 Distribution of PD-L1 epithelium**

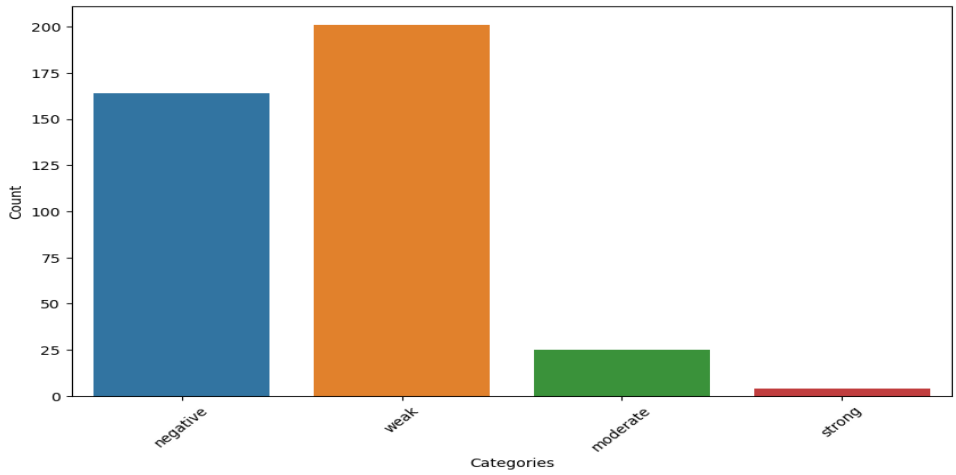



**Figure 3.36 Distribution of PD-L1 lymphocyte**

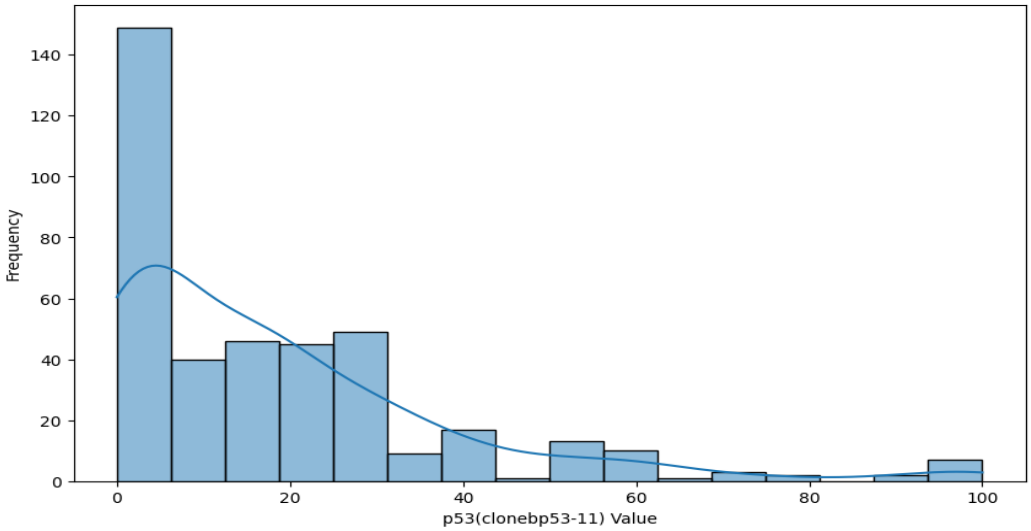



**Figure 3.37 Distribution of p53 (clone bp53-11)**

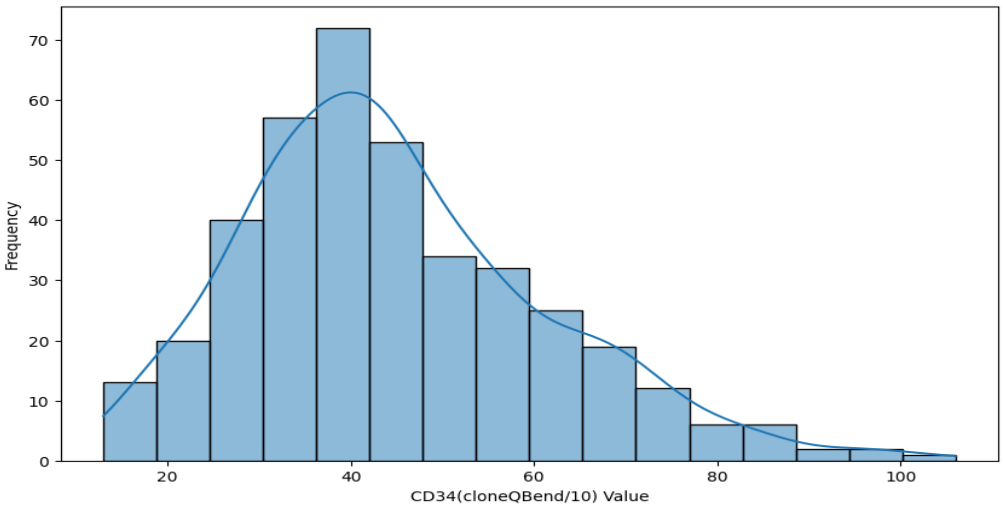



**Figure 3.38 Distribution of CD34 (clone QBend/10)**

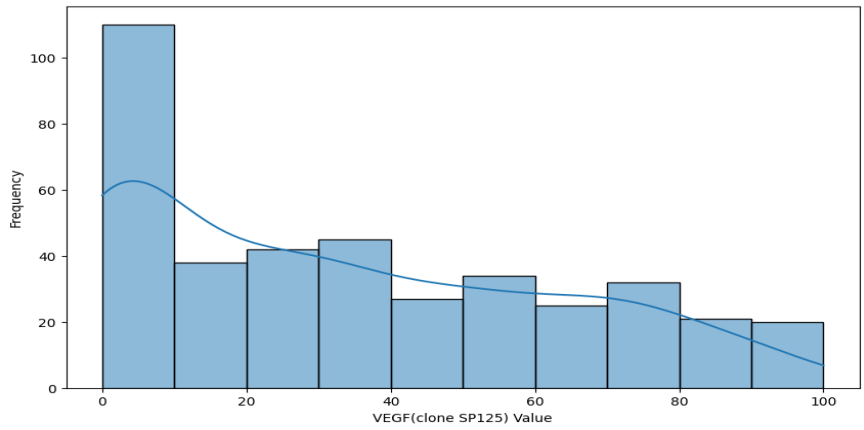



**Figure 3.39 Distribution of VEGF (clone SP125)**

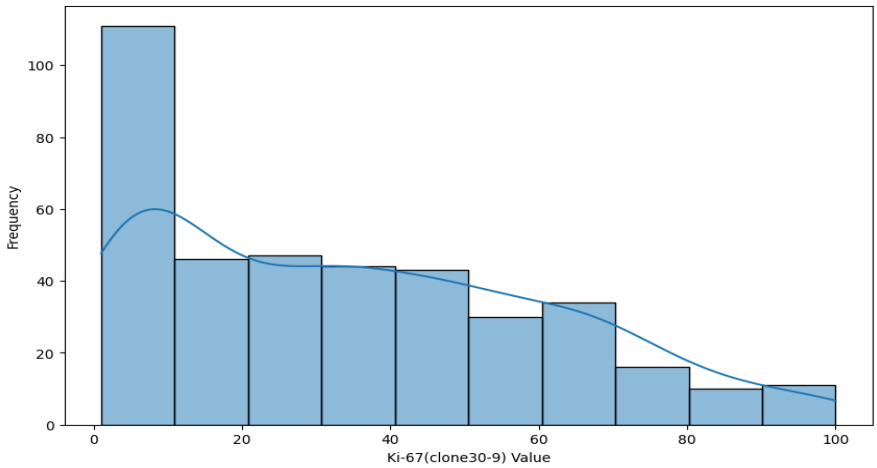



**Figure 3.40 Distribution of Ki-67 (clone 30-9)**

Shapiro-Wilk test results show that data for Ki-67 (clone 30-9), VEGF (clone SP125), CD34 (clone QBend/10), and p53 (clone bp53-11) variants do not follow a normal distribution. The test statistic obtained for Ki-67 (clone 30-9)is 0.9325, and the p-value is $2.53 \times 10^{-12}$, the test statistic for VEGF (clone SP125) is 0.9018 and the p-value is $2.94 \times 10^{-15}$, the test statistic for CD34 (clone QBend/10) is 0.9656 and the p-value is $5.30 \times 10^{-8}$, and the test statistic for p53 (clone bp53-11) is 0.79 and the p-value is $9.14 \times 10^{-22}$. Because these very low p-values are well below the 0.05 threshold, null hypotheses (the hypothesis that the data comes from a normal distribution) need to be rejected for all four variables. These results indicate that the datasets obtained for variants Ki-67 (clone 30-9), VEGF (clone SP125), CD34 (clone QBend/10), and p53 (clone bp53-11) do not follow a normal distribution. This requires the use of non-parametric statistical methods in the analysis of these variables. Furthermore, these findings require particular care in the analysis and interpretation of these variables. The non-normal distribution of these four biomarkers is typically encountered in clinical trials and should be considered in further analysis.

This research focuses on data obtained from identical polyps but collected by different methods. The data collection process was performed within a clearly defined experimental setting and covered a broad range of colon polyps and cancer biopsies. The study's design may include various factors such as patient information, lesion characteristics, and clinical parameters recorded during data collection.



**Figure 3.41 Slides of BRAF (clone V600E) , Ki-67 (clone 30-9) and PD-L1 (clone SP142) antibodies, respectively**

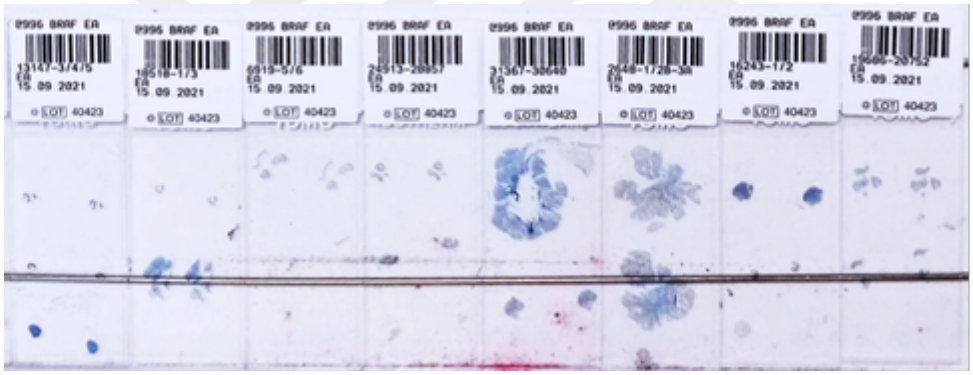


**Figure 3.42 Individual representation of some of the BRAF (clone V600E) antibody slides**

The method applied in the evaluation part of the stained slides is different for each antibody. While evaluating **BRAF (clone V600E)**, cytoplasmic staining was based on, nuclear positivity (i.e., staining) in the colon and could not be taken as a basis and evaluated. It is divided into negative and positive staining, and positive staining is divided into 3 classes strong and negative. In the microscope, the area was found with 5X magnification, and cytoplasmic staining was evaluated at 10X magnification.

Nuclear staining was taken as the basis for the evaluation of **Ki-67 (clone 30-9)** and **p53 (clone bp53-11)**. While evaluating the nuclear dyes, the area with the lesion was determined at a small magnification such as 0, all nuclei were counted at 40X magnification, and the percentage was given according to the stained ones.




**Figure 3.43 CD34 (clone QBend/10), Ki-67 (clone 30-9), VEGF (clone SP125) and p53 (clone bp53-11) staining score results that came from 201 patients**

While evaluating **PD-L1,** 2 different methods were applied. One of them is the staining percentage of the existing inflammatory cells in the lamina propria in the background, and the other is the cytoplasmic staining of the epithelium.

**VEGF (clone SP125)** was evaluated using 2 parameters vessel extent and density. In the diffuseness parameter, the stained cells are given as a percentage, and in the intensity parameter, the staining strength is indicated as negative, weak, and moderate.

Evaluation of **CD34 (clone QBend/10)**, the area with the lesion was determined with a small magnification like 5X, the section with the highest vein density in this area was selected, and the number of vessels was indicated with a magnification of 40X. The number of veins was divided into cut-off values according to previous similar studies.

## 3.3 Data Records

The data records used in this study include colonoscopy videos, histopathological images, and IHC results, providing a multidimensional and comprehensive representation of colonic polyps by combining several methods. Systematically combining these datasets creates a robust platform for in-depth research and study of colonic polyps.

### 3.3.1 Colonoscopy Video Dataset

The first dataset, called 'ColonVideosWithLabels', consists of 133 colonoscopy videos, each containing a personalized recording of a polyp. These videos are stored in AVI format and are presented following a systematic and descriptive file naming protocol. Accompanying the videos is a set of annotations that include information on detailed polyp characteristics as well as patient demographics.

These descriptions cover a variety of markers and include the patient's identity, gender, age, location of the polyp in the colon, type, and subtype, in addition to IHC staining results. An Excel file associated with the dataset archives this broad data summary in a structured way. In this way, researchers can easily access the desired polyp characteristics and patient demographics, analyze the data set, and use it in various studies.

Directory Architecture:

- colonVideosWithLabels

-colon-patient3-polyp1-M-70-ascending-nonneoplastic-hyperplastic-10-negative-negative-negative-40-3-70-2.avi

-colon-patient4-polyp1-F-57-sigmoid-nonneoplastic-hyperplastic-30-negative-negative-weak-75-2-50-25.avi

### 3.3.2 Histopathological Image Dataset

The second data set, 'pathoImagesWithLabels', consists of histopathological images of polyps. These high-resolution images were captured in TIFF format, including 2.5X, 5X, 10X, 20X, and 40X magnifications. Organized systematically according to the magnification level used during image acquisition, these images have filenames that follow the same convention as colonoscopy videos. These filenames cover a wide range of polyp information.

IHC staining results corresponding to these images are available in the Excel file 'pathoImagesWithLabels.xlsx'. This file includes IHC results associated with histopathological images, allowing for a more comprehensive analysis and evaluation.

Directory Architecture:

- pathoImagesWithLabels:
  - Magnification Level 1: 2.5X
    - patho-patient1-polyp2-descending-nonneoplastic-hyperplastic-slideX2,5.tiff
  - Magnification Level 2: 5X
    - patho-patient1-polyp2-descending-nonneoplastic-hyperplastic-slideX5.tiff
  - Magnification Level 3: 10X
    - patho-patient1-polyp2-descending-nonneoplastic-hyperplastic-slideX10.tiff

### 3.3.3 Immunohistochemistry Results Dataset

The third dataset, containing the file 'IHC_data.xlsx', has a unique feature because it contains IHC results and covers tissues without corresponding colon videos or histopathology images. This dataset can be used independently for research focusing specifically on IHC analysis. However, within the context of a larger study, unique patient identification numbers across the three datasets, which include colonoscopy videos, histopathological images, and IHC results, can be effectively used for data integration and cross-referencing purposes.

Each dataset provides a comprehensive view of polyp morphology, supporting the accuracy of polyp characterization. These carefully curated datasets lay the foundation for future research into the diagnosis and classification of polyps, charting a path for better patient care and clinical outcomes.

This comprehensive work aims to ensure long-term access and preservation by providing a persistent identifier (DOI) from the public repository where datasets are stored on Kaggle. You can access the DOI of this dataset via [https://www.kaggle.com/datasets/refikasultandogan/colonpolyp].

## 3.4 Technical Validation

In this section, we showcase the technical quality of our image set and the tests and analyses that support the pathological diagnosis of colon polyps. We have taken several steps to ensure our statistics are accurate and repeatable.

Initially, all images were captured using the same colonoscopy, microscope, and camera settings; they all had the same standards. Also, we used images of properly labeled healthy colon tissue as negative controls in our data collection process so that our classification system could express the difference between normal and abnormal tissue.

In this investigation, we applied a robust strategy for the classification and identification of polyp subtypes, integrating the strengths of two efficient staining techniques, H&E and IHC. This double staining approach was chosen considering its potential to improve the accuracy and reliability of our findings. To verify the consistency of the classification results produced by these two techniques, we used the well-known and widely adopted Cohen's Kappa statistical measure [74], [75] for its ability to precisely quantify the degree of agreement between two nominal-scale assessments. Our analysis revealed significant agreement between the two methods, with the Kappa coefficient being 0.9. This confirms the high consistency and reliability of our diagnostic approach. At the same time, we performed another statistical measure, called percent agreement calculation, to determine the degree of agreement between HE and IHC staining techniques [75], [76]. This measurement evaluates the agreement between two different measurement methods, and the result we obtained was approximately 90.41%. This high level of agreement indicates that similar results were obtained with both techniques in polyp classification. Therefore, it is possible to state that HE and IHC offer identical results regarding reliability and compatibility. In the context of our research, we concluded that the double-staining strategy not only produced consistent and reliable results but also strengthened the validity of the findings. The high level of agreement and similarity in results demonstrates an approach that supports internal validity. Therefore, we propose that the double staining strategy may be a valuable tool not only for current research but also for future histopathological studies and broad clinical applications.

In general procedures, we used blinding to avoid bias in the image classification. We also ensured accurate and reliable sample identification by tracking all samples from collection to analysis. The technical validation in this section supports the reliability of our data set and makes it amenable to further analysis and interpretation. This methodology maintains the validity of the findings by avoiding expected conclusions and allows for a more valid and reliable interpretation of the results.

Finally, we have followed standard procedures for reliable and unbiased data generation. All imaging and pathological diagnoses were performed without consideration of patient information, and all samples were tracked using a unique identifier system. Overall, our

technical validation experiments and procedures support our comprehensive image dataset's high quality and reliability.

## 3.5 Code and Dataset Availability

The custom code used to create and process the datasets mentioned in this research is available on Kaggle, a widely used machine learning and data science platform. The code repository has a clear and user-friendly organization equipped with explanations and documentation.

Interested researchers can obtain the code by visiting the link [https://www.kaggle.com/datasets/refikasultandogan/colonpolyp]. This code and data set are made available to the public free of charge and are available for download. The code, written using version 3.8 of the Python programming language specifically, includes machine learning tasks with the Scikit-learn library (version 0.24.2) and data visualization with the matplotlib library (version 3.4.3).

The code is designed to ensure reproducibility by using clear variable names, and the purpose and functionality in each section are explained with detailed comments. Additionally, the code contains information that specifies certain variables and parameters for creating, testing, and processing datasets.

# Chapter 4

# Methods

## 4.1 Feature extraction methods

### 4.1.1 Border/Interior Pixel Classification

The Border/Interior Pixel Classification (BIC) algorithm is generally preferred for these main reasons: an easy-to-follow and powerful image analysis algorithm to classify image pixels like border or interior; compare the histograms that came from the pixels using logarithmic distance; to extract visual features from the images to create compact representation. The logic of this algorithm; is RGB color-space homogeneously quantized in 4×4×4=64 colors so 64 bins. Following the quantization step, border or interior pixels are classified using image pixels, each BIC histogram contains 64 border and 64 interior pixels so a total of 128 bins. The dLog function is a mathematical function that simplifies the comparison of histograms by applying a logarithmic scale. Its definition is as follows:

$$\mathrm{dLog}(q,d) = \sum_{i=0}^{i<M} |f(q|i|) - f(d|i|)|, \quad f(x) = \begin{cases} 0, & \text{if } x = 0 \\ 1, & \text{if } 0 < x \leq 1 \\ \lceil \log_2 x \rceil + 1, & \text{otherwise} \end{cases} \tag{4.1}$$

In Equation 4.1, q and d represent two histograms, each consisting of M bins. The symbol q| i| denotes the $i^{th}$ bin of histogram q, while d| i| indicates the $i^{th}$ bin of histogram d. The histogram bins are standardized to a range of 0 to 255. The f(x) function result can be stored for comparing histograms using the dLog distance function [77].

### 4.1.2 Global Color Histogram

Global color distributions are efficiently used as descriptors for pattern recognition. Global Color Histogram (GCH) is also one of the global color distributions used for color feature extraction, to represent the images from other perspectives [78]. As explained in [79], the color histogram of an image is represented using vector H,

$$H = \{H(0), H(1), \ \dots \ , H(j), \ \dots \ , H(N)\} \tag{4.2}$$

where j denotes the color bin, N represents the number of bins, and H(j) denotes the total number of pixels at the $j^{th}$ color bin. During the feature extraction process 10 bins per channel.

Then the number of pixels for each color is counted and stored in the histogram's bins [80], [81].

### 4.1.3 First Order Measures

To extract the First Order Measures (FOM) statistics library used in Python. Totally 32 features were obtained from these measures, 8 of them belong to the gray level image statistics and the rest of the 24 features belong to each color channel: red, green, and blue uniformly. The first feature is average, to calculate the arithmetic mean of image statistics.mean() function used. statistics.mode() used for obtaining the most common value of an image. Then the spread of data using standard deviation and variance of population and sample using these functions; statistics.pvariance() used for population variance of data, pstdev() used for population standard deviation of data, statistics.variance() used for sample variance of data, statistics.stdev() used for sample standard deviation of data [82].

The first-order histogram P(I) is defined in Equation 4.3 [83]. I represent the gray levels of an image region. Entropy and energy of distribution are calculated using this histogram with given probability values (Equations 4.4 and 4.5).

$$\mathrm{P(I)} = \frac{\text{number of pixels with gray level (I)}}{\text{total number of pixels in the region}} \tag{4.3}$$

$$\text{Energy} = \sum_{\mathrm{I}} \mathrm{P(I)}^2 \tag{4.4}$$

$$\text{Entropy} = \sum_{\mathrm{I}} \mathrm{P(I)} \log \mathrm{P(I)} \tag{4.5}$$

### 4.1.4 Local Binary Pattern

We used the local binary pattern (LBP), which is a way to store information about the local image. LBP is one way to find out what kind of tissue something is, and it has many uses in picture classification and medical imaging [84], [85]. LBP parameters determine how the LBP method works. These factors are the radius, the number of neighbors, the features that don't change when rotated, and the interpolation. Histograms also have things like the size of the cells and the type of correction. The amount of neighbors parameter is set so that each pixel is arranged in a circle. By adding more pixels as friends, the algorithm can store more information about the area around the pixels. The radius is a circle that shows how each pixel's neighbors are chosen [86]. The logic of background LBP features with mathematical operation is shown in Equation 4.6.

$$\mathrm{LBP_{P,R}} = \sum_{\mathrm{p}=0}^{\mathrm{P}-1} \mathrm{s}(\mathrm{g_p} - \mathrm{g_c}) 2^{\mathrm{p}}, \ \mathrm{s(x)} = \begin{cases} 1, & \mathrm{x} \geq 0 \\ 0, & \mathrm{x} < 1 \end{cases} \tag{4.6}$$

In this equation $g_p$ is the gray-scale value of the neighboring pixels and $g_c$ is the gray-scale value of the center pixel. P represents the total number of involved neighbors, and R shows the radius of the neighborhood. The binomial parameter factor $2^p$, aims to convert to a unique $LBP_{P,R}$ number characterizing the spatial structure of the local image texture [85].

### 4.1.5 Haralick

Haralick texture features are similar to the LBP. Firstly gray-level co-occurrence matrix (GLCM) is computed. This computation obtains second-order statistical information on pixels from an image. Using these various statistical properties of the co-occurrence matrix Haralick features are computed. From GLCM, Haralick proposed 13 common statistical features, known as Haralick texture features [87]. These features are available in the mahotas.features.haralick module in Python [88], [89] and explained using the following equations. The first feature is the angular second-moment feature, asm or $f_1$, which is a measure of homogeneity of the image. This feature detects dominant tone transitions, if these transitions are very few this means this image is homogeneous. If pixels are very similar or the image has good homogeneity, then concluded that the asm feature is high [90].

$$f_1 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \left(\frac{P(i,j)}{R}\right)^2 = \sum_i \sum_j p(i,j)^2 \tag{4.7}$$

In Equation 4.7, Ng represents the number of distinct gray levels in the quantized image in which a square N-by-N image matrix. P(i,j) is a matrix of relative frequencies with which two neighboring cells, R represent normalizing constant and $p(i,j)$ or $P(i,j)/R$ represents (i,j)th entry in a normalized matrix. The contrast feature, $f_2$, is a difference moment of the P matrix and is the number of local variations existent in the image explained in Equation 4.8.

$$f_2 = \sum_{k=0}^{N_g-1} k^2 \left\{ \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \delta_{|i-j|,k} p(i,j) \right\} = \sum_{k=0}^{N_g-1} k^2 \, p_{x-y}(k), \qquad k = 0,1,\dots,N_g - 1$$

$$\delta_{i,j} = \begin{cases} 1, \text{when } i = j \\ 0, \text{when } i \neq j \end{cases} \tag{4.8}$$

The linear dependency of gray levels of neighboring pixels is measured using the correlation feature, explained in Equation 4.9 as a $f_3$.

$$f_3 = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (ij) p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y} \tag{4.9}$$

where $\mu_x, \mu_y, \sigma_x$ and $\sigma_y$ are the means and standard deviations of $p_x$ and $p_y$, respectively [90].

$$\text{Sum of squares variance} = f_4 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i-\mu)^2 p(i,j) \tag{4.10}$$

In Equation 4.11 is the Inverse Difference Moment (IDM) so it is the local homogeneity. This feature is high when the local gray level is uniform [90].

$$f_5 = \sum_{i=1}^{N_g}\sum_{i=1}^{N_g}\frac{1}{1+(i-j)^2}p(i,j) \tag{4.11}$$

$$\text{Sum average} = f_6 = \sum_{i=2}^{2Ng} i p_{x+y}(i) \tag{4.12}$$

$$\text{Sum variance} = f_7 = \sum_{i=2}^{2Ng}(i-f_8)^2 p_{x+y}(i) \tag{4.13}$$

$$\text{Sum entropy} = f_8 = -\sum_{i=2}^{2Ng} p_{x+y}(i)\log\left(p_{x+y}(i)\right) \tag{4.14}$$

$$\text{Entropy} = f_9 = -\sum_{i=1}^{Ng}\sum_{j=1}^{Ng} p(i,j)\log\left(p(i,j)\right) \tag{4.15}$$

$$\text{Difference variance} = f_{10} = \text{variance of } p_{x-y} \tag{4.16}$$

$$\text{Difference entropy} = f_{11} = -\sum_{i=0}^{Ng-1} p_{x+y}(i)\log\left(p_{x-y}(i)\right) \tag{4.17}$$

$$\text{Information measures of correlation 1} = f_{12} = \frac{f_9 - HXY1}{\max(HX, HY)} \tag{4.18}$$

where HX is the entropy of $p_x$, HY is the entropy of $p_y$ and HXY1 is $-\sum_i\sum_j p(i,j)\log\left(p_x(i)p_y(j)\right)$.

$$\text{Information measures of correlation 2} = f_{13} = \left[1-\exp\left(-2(HXY2-f_9)\right)\right]^{1/2} \tag{4.19}$$

where HXY2 is $-\sum_i\sum_j p_x(i)p_y(j)\log\left(p_x(i)p_y(j)\right)$.

### 4.1.6 Threshold Adjacency Statistics

For computational image analysis, the mahotas.features.tas module in Python [89] provides Threshold Adjacency Statistics (TAS) features that can be used to obtain image statistics. TAS was presented by Hamilton et al. in [91] to distinguish and classify medical images. This algorithm uses just three threshold values, these values can be symbolized by the letters E1,

E2, and E3, to generate three binary images. The basic idea behind TAS is to threshold the image and count the number of above-threshold pixels that are adjacent to a given number of above-threshold pixels [92].

### 4.1.7 Zernike Moments

Zernike moments are commonly described in the process of extracting and describing image shape features, in addition to being used in content-based image retrieval [93] and region-based matching [94]. The more specific cases include human face detection and retrieval [95], character and object matching and recognition [96], and object matching and recognition [97]. Two parameters are required to be supplied to compute Zernike moments: the radius of the region in polar coordinates and the degree of the polynomial; the default value for the degree parameter is 8. The dimensions of the feature vector that is returned are determined by the degree of the polynomial.

## 4.2 Feature selection

In this study four distinct feature selection methodologies this study: Recursive Feature Elimination (RFE), SelectKBest, SelectFromModel, and Boruta. The RFE technique was employed to identify the topmost significant features. It relied on the RF classifier prediction model and iteratively removed the least important features. RFE assessed the relative importance of features and systematically reduced those considered less relevant [98]. The SelectKBest approach utilizing statistical tests to compute feature scores, SelectKBest enabled the selection of the most promising 'k' features [99]. In our research, we compiled a list of the ten most prominent features using this method. The SelectFromModel approach allowed us to select features from a prediction model by considering feature severity levels and applying a specific threshold value [100]. In our study, we utilized the Random Forest (RF) classifier estimation model and set the threshold value to 'median' to identify impactful features. The Boruta built on the RF classifier prediction model, the Boruta technique was evaluated by comparing it to randomly selected iterations of all potential features [101]. By incorporating Boruta into our methodology, we automated the feature selection process, ultimately enhancing the overall efficiency of the model. Each of these feature selection techniques offered a unique approach to identifying relevant features. Engaging in a comparative analysis of these methods allowed us to determine the most suitable approach for a specific perspective. In conclusion,

our investigation into feature selection and model comparison underscores the critical role these processes play in the development of robust machine-learning models.

## 4.3 Classification Methods

In the context of machine learning, classification is the process of allocating observations to various groups based on the relationships that have been identified between input variables and results. The task consists of using knowledge extracted from the training dataset to predict categorical outcomes. Across a variety of fields, this discernment process makes a significant contribution to well-informed decision-making.

Classification algorithms are essential to machine learning because they make it possible to predict the class label that corresponds to a particular data point. This feature enables machines to recognize patterns and forecast outcomes for data, even in situations where class labels are not known. The application of classification in machine learning extends to both supervised and unsupervised tasks. Class labels are applied to training data points in supervised classification, where the algorithm is trained on labeled data. By preparing the algorithm to learn and generalize the connections between input features and the corresponding class labels, this training process makes it possible to make precise predictions.

On the other hand, unsupervised classification functions without a priori knowledge of the class labels. Rather than using predefined categories, it concentrates on grouping data points into clusters according to their similarities. This method of working without supervision is especially useful when the goal is to find inherent structures and patterns in the data without the assistance of labeled examples.

### 4.3.1 Logistic Regression (LR)

Logistic Regression acts as a classifier by estimating the likelihood that an instance belongs to a specific class. It accomplishes this by mapping the linear combination of input features into a probability score between 0 and 1 using a probabilistic framework and the logistic (sigmoid) function. This combination yields a linear decision boundary, which is a hyperplane that divides the feature space into various regions that correspond to different classes. To optimize the chance of observing the true class labels during training, Logistic Regression modifies its parameters by minimizing a cost function. For tasks involving binary classification, this method makes logistic regression a suitable tool.

### 4.3.2 Random Forest (RF)

The Random Forest (RF) classifier is an ensemble learning technique that generates many decision trees and determines the class mode based on the individual trees. The aforementioned methodology exhibits robustness against overfitting and demonstrates efficacy in effectively analyzing extensive datasets characterized by a high number of dimensions [102].

### 4.3.3 Decision Trees (DT)

The Decision Trees (DT) classifier is a very effective and straightforward supervised learning approach. The system can obtain and present associations among decision trees, characteristics, and the target class. The approach yields outcomes that are easily understandable by individuals [103].

### 4.3.4 K-Nearest Neighbors (KNN)

The K Nearest Neighbor (KNN) algorithm is a classification method that determines the class label of a sample based on the classes of its nearest neighbors within the dataset. The KNN approach is widely used in various applications and has shown to be effective, especially in scenarios involving datasets with a limited number of dimensions [104].

### 4.3.5 Support Vector Machines (SVM)

Support Vector Machines (SVMs) solve a constrained quadratic optimization problem to generate optimal possible boundaries between classes in a classification task. Determining a hyperplane that maximally divides the data points of various classes while preserving a margin of separation is the fundamental objective of a support vector machine (SVM). SVM seeks to maximize the margin, which is the distance between each class's closest data points and the hyperplane. The optimal hyperplane is the one that gives the maximum margin of separation and accurately classifies the data [105], [106]

### 4.3.6 eXtreme Gradient Boosting (XGBoost)

The eXtreme Gradient Boosting (XGBoost) classifier is a machine-learning technique that uses gradient-boosting decision trees to achieve speed and efficiency. A more effective model formulation was used to address the overfitting issue, which improved performance [107].

### 4.3.7 Artificial Neural Network (ANN)

Artificial Neural Networks are mathematical models that mimic the function of nerve cells (neurons) of the human brain, and these models are used in data processing and learning tasks. These networks consist of interconnected nodes (neurons), which process input data and produce output. Artificial Neural Networks are widely used in fields such as image and sound recognition and language processing, thanks to their ability to recognize, classify, and predict complex patterns and datasets [108], [109].

### 4.3.8 Gradient Boosting (GB)

The Gradient Boosting (GB) classifier is a machine learning technique that constructs a resilient predictive model by aggregating weak predictors, often in the shape of decision trees. The Gradient Boosting technique iteratively enhances the model by incorporating additional estimators that specifically address the flaws of the model. Consequently, the model can produce forecasts of higher accuracy [110].

### 4.3.9 Light Gradient Boosting Machine (LGBM)

The LGBM classifier is a gradient-boosting algorithm specifically developed to efficiently handle and achieve high performance on large-scale datasets. LightGBM (LGBM) utilizes a histogram-based methodology, which necessitates less memory use and results in expedited training and prediction durations [111].

### 4.3.10 Naive Bayes (NB)

Naive Bayes (NB) is a popular machine learning algorithm based on probability theory and used primarily in classification tasks. This algorithm is based on Bayes' Theorem and assumes that the effect of each feature on the class is independent of other features hence the name is "naive". NB is used effectively and efficiently in many fields such as text classification, spam filtering, sentiment analysis, and medical diagnosis [15].

## 4.4 Cross-validation techniques

Cross-validation is a method used to evaluate the generalization ability of a machine learning model.

The 5 k-fold cross-validation method divides the dataset into five equal (or approximately equal) parts. In each iteration, one fragment is used as the test set and the remaining four

fragments are used as the training set. This process is repeated five times so that each part is used as a test set exactly once. This method is used to evaluate the performance of the model on different subsets of data [112]. The 10 k-fold cross-validation method does the same process as the 5 k-fold but divides the dataset into ten parts. More folds generally provide a more stable estimate of the model's generalization performance because the model is tested on ten different combinations of training and test sets [113]. Stratified k-fold cross-validation is especially important for classification tasks and is a variation of the k-fold method. This method ensures that each layer is split to reflect the proportion of classes in the original dataset, preserving class proportions. This is important when the class distribution in the dataset is unbalanced and helps the model achieve more realistic performance measurements [114]. Repeated Stratified k-fold cross-validation refers to repeating stratified k-fold cross-validation many times. Each iteration involves a different random split. This method is used to obtain a more reliable estimate of the model's performance because the effect of randomness is reduced and the stability of the model on various subsets of data is evaluated [115].

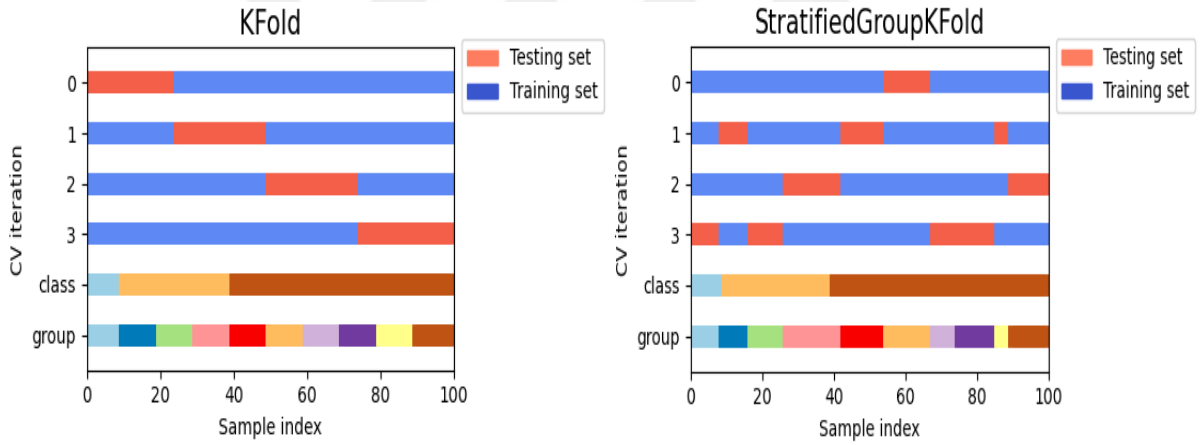



**Figure 4.1 The CV strategies in terms of forming training and test data [113]**

There are visual representations of k-fold and stratified group k-fold cross-validation methods shown in Figure 4.1. These methods are used in machine learning to evaluate the performance of a model on different subsets of data.

In the K-Fold CV method, the data set is randomly divided into 'k' equally sized subsets. In each iteration, a subset is reserved as the testing set, while the remaining 'k-1' subset is used as the training set. This process is repeated 'k' times such that each subset is used once as the test set. This method is effective for evaluating how the model can generalize to different parts of the data set because each data point is used for both training and testing. As you can see in the image, different colors can represent different classes or groups, and with each 'CV iteration' or cross-

validation iteration, there is a split between the test set shown in blue and the training set shown in red.

Stratified cross-validation is a variation of k-fold cross-validation and is used in classification problems. This method ensures that each subset is strategized to reflect the class distribution of the original data set. This ensures that the samples of each class have close representation rates in the original data set in each training and testing subset. Group K-Fold is used to ensure that samples (groups) that are related to each other in the data are included in the same subsets. For example, we may want different samples taken from the same patient to be included in the same training or test set so that the model does not learn these intra-group relationships and displays a more realistic generalization performance. 'StratifiedGroupKfold' in the image is a more complex cross-validation method that takes both stratification and grouping into account. In each iteration, the class distribution is preserved, and certain groups are reserved together, so we can more fairly evaluate the performance of the model on certain groups.

In both graphs, one can see how the test and training sets change and how the classes/groups are preserved in each cross-validation iteration. These methods are important for understanding how the model responds to different parts of the data and potentially detecting issues such as overfitting.

The **cross-validation** method with non-overlapping groups for each iteration. The folds were prepared by preserving the percentage of samples for each class [116]. According to Figure 4.1, in this thesis case, classes are polyp types, and groups are unique polyp IDs.

## 4.5 Performance evaluation metrics

Confusion Matrix is a table used for a model's perspective on seen problems. It consists of four basic components: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN).

Figure 4.2 shows a typical confusion matrix for a binary classification task. The confusion matrix is used to evaluate the performance of a classification model and visualizes the model's correct and incorrect predictions for each class. The term TP refers to the count of tubular polyps that are successfully predicted by the model, whereas FP denotes the count of polyps that are inaccurately identified as tubular. In a similar logic, the term TN refers to the count of accurately predicted hyperplastic polyps, while FN denotes the count of polyps that are incorrectly classified as hyperplastic.

| | | Predicted Class | | |
|---|---|---|---|---|
| | | **Positive** | **Negative** | |
| **Actual Class** | **Positive** | True Positive (TP) | False Negative (FN)<br>Type II Error | **Sensitivity**<br>$\frac{TP}{TP+FN}$ |
| | **Negative** | False Positive (FP)<br>Type I Error | True Negative (TN) | **Specificity**<br>$\frac{TN}{TN+FP}$ |
| | | **Precision**<br>$\frac{TP}{TP+FP}$ | **Negative Predictive Value**<br>$\frac{TN}{TN+FN}$ | **Accuracy**<br>$\frac{TN+TP}{TN+TP+FN+FP}$ |

**Figure 4.2 Confusion matrix**

The performance metrics, accuracy, precision, sensitivity (or recall), specificity, and F-score, are derived using the terms TP, FP, TN, and FN. These metrics are expressed as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{4.20}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{4.21}$$

$$\text{Recall (Sensitivity)} = \frac{TP}{TP+FN} \tag{4.22}$$

$$\text{F} - \text{score} = \frac{TP}{TP+(FP+FN)/2} \tag{4.23}$$

These metrics help you understand and improve different aspects of your model. For example, in a health screening, it may be important to reduce FN (missed cases of disease), while in a spam filter, it may be more important to reduce FP (important emails being mistakenly marked as spam).

Figure 4.3 shows an ROC (Receiver Operating Characteristic) Curve. The ROC Curve is used to evaluate the performance of classification models and is often used in medical diagnostic testing, machine learning models, and similar fields. The ROC Curve has values of 1 - Specificity (FPR) on the x-axis and Sensitivity (TPR) on the y-axis. The width of the area under the curve (AUC) indicates how well the model performs compared to random guesses.

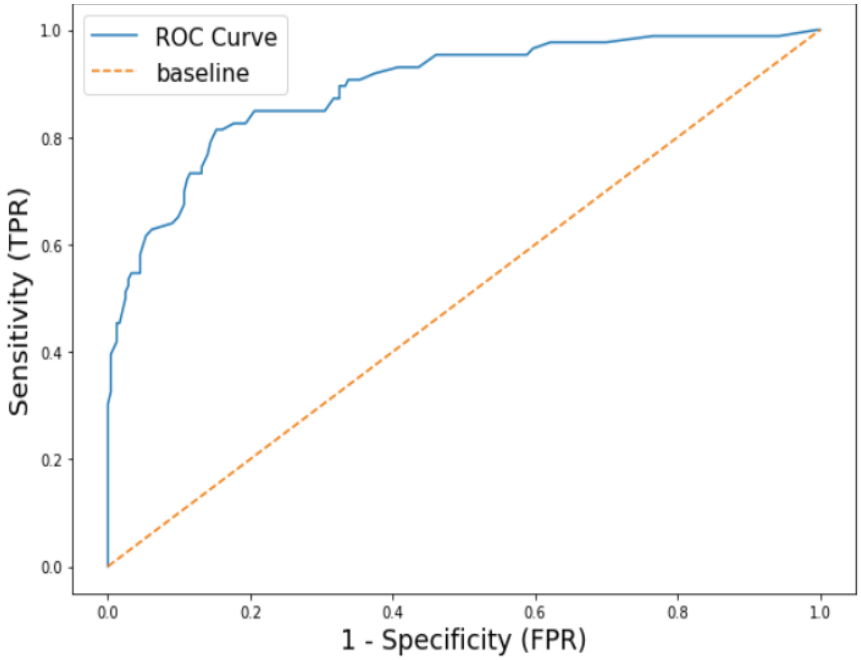



**Figure 4.3 Receiver operating characteristic curve**

The higher the AUC value, the better the performance of the model. The ROC curve of an ideal model would be near the upper left corner and the AUC value would approach 1. If a model's curve is above the 45-degree line, which is the dashed line, it indicates that the model is better than random guessing. This 45-degree line represents the 'baseline' where the model's predictions are as successful as a random guess.

The ROC Curve refers to:

- The further up and to the left the curve is, the better the model performs.

- The size of the area under the curve (AUC) is an indicator of the overall performance of the model. When the AUC value is close to 1, it means perfect classification, and when it is close to 0.5, it means a classification no different from a random prediction.

- The dashed orange line represents a representation of a random prediction, and an ideal ROC Curve is expected to lie above this line. The blue line in this graph shows the performance of the model at different threshold values. We can observe how the performance of the model changes by changing the sensitivity and 1-specificity of the model, that is, by using different cutoff points. When the blue line is well above the orange line on the chart, it indicates that the model is performing better than random guessing and is generally considered a good sign.

# 4.6 Optimizing Hyperparameters for Enhanced Model Performance

In this study, the hyperparameters of several machine learning models, such as Logistic Regression, Random Forest, Gaussian Naive Bayes, Decision Trees, K-Nearest Neighbors, Support Vector Machines, XGBoost, Neural Networks, Gradient Boosting, and LightGBM, were optimized. Multiple hyperparameter combinations were evaluated for each model, and the optimal values that yielded the most favorable outcomes were documented. The adjusted hyperparameters were subsequently used in certain classifier models to enhance their performance. The utilization of this method played a crucial role in producing optimal outcomes for every classifier model, hence improving the overall performance of the models and assuring their capacity to generalize.

**Table 4.1 The grid-search space used for tuning the hyperparameters of the machine-learning classifiers**

| Classifier | Parameters | Grid-search space |
|---|---|---|
| **LR** | C | [0.001, 0.01, 0.1, 1, 10, 100, 1000] |
| **RF** | n_estimators | [50, 100, 150, 200] |
| | max_depth | [None, 10, 20, 30] |
| | min_samples_split | [2, 5, 10] |
| **NB** | var_smoothing | [1e-9, 1e-8, 1e-7] |
| **DT** | max_depth | [None, 5, 10, 15] |
| | min_samples_split | [2, 5, 10] |
| | min_samples_leaf | [1, 2, 4] |
| **KNN** | n_neighbors | list(range(1, 30)) |
| | weights | ['uniform', 'distance'] |
| **SVM** | C | [0.1, 1, 10, 100] |
| | gamma | ['scale', 'auto', 0.1, 1, 10] |
| | kernel | Radial basis function, linear |
| **XGB** | n_estimators | [50, 100, 150, 200] |
| | max_depth | [3, 5, 7, 10] |
| | learning_rate | [0.01, 0.1, 0.2, 0.3] |
| **Neural Networks** | 'hidden_layer_sizes | [(50,50,50), (50,100,50), (100,)] |
| | activation | ['tanh', 'relu'] |
| | solver | ['sgd', 'adam'] |
| | alpha | [0.0001, 0.05] |
| **Gradient Boosting** | learning_rate | [0.01, 0.1, 0.5] |
| | n_estimators | [100, 200, 300] |
| | max_depth | [5, 10, 15] |
| | min_samples_split | [2, 5, 10] |
| | min_samples_leaf | [1, 3, 5] |
| **LGBM** | learning_rate | [0.01, 0.1, 0.5] |
| | n_estimators | [100, 200, 300] |
| | max_depth | [5, 10, 15] |
| | num_leaves | [20, 30, 40] |
| | min_child_samples | [10, 20, 30] |

## 4.7 Statistical Analysis and Evaluation

We used Turcosa software (Turcosa Analytics, Turkey) and Python programs for statistical analysis. Numerical variables are expressed as mean ± standard deviation. After the distribution tests are applied to all variables, the differences between the groups are evaluated by using the t-test for continuous variables with normal distribution and the Mann-Whitney U test for variables that do not. Correlation analysis is used to determine whether there is a relationship between the features extracted/obtained from colonoscopy and microscope images and the percentages of staining from the IHC study and, if there is, the direction of this relationship. The most used Pearson and Spearman correlations were used in the correlation analysis.

Analysis of Variance (ANOVA) is a method used to test whether the group means of data obtained from more than two independent groups are different. This analysis is tested using common variance whether more than two groups are drawn from normally distributed like-mean populations. For example, in a study where scores/features obtained from images in groups of different polyp types are desired to be compared, it would be appropriate to use a one-way analysis of variance. In the case of non-normal distribution, the Friedman test is used as an alternative to ANOVA.

The statistical significance level was determined as $p<0.05$. Here, we aim to determine the antibodies and antigen-binding scores/percentages/rates. For this, a correlation test was applied to correlate the features extracted from the images with the scoring of the IHC report. It investigated whether there is a significant difference between the features obtained from the images and the type of polyp (pathology result) and antigen-antibody binding score (IHC result). The powerful features considered prognostic markers for us to express the changes at the molecular level of the biological tissue from the macroscopic image and predict the clinical course of the polyp type. In this way, during the colonoscopy examination, the pathology result of the polyp and its relationship with specific proteins are analyzed in real-time and aid doctors in diagnose.

# Chapter 5

# Studies and Analysis with Results

## 5.1 Colonoscopy Based Studies

### 5.1.1 Study 1: Comparison Manual and Automatic Feature Extraction Based Colon Polyp Classification (Preliminary Study)

We first investigated the effect of image content on the classification performance (explained in Chapter 3.2.2.1). For this purpose, we prepared four different colonoscopy image sets; original images, the same images without the black border, cropped polyp with 40% background, and cropped polyp with no background (see Figure 5.1). Figure 5.2 shows the results of preprocessing applied to the same images using the approach called CLAHE [68]. This was utilized to minimize the contrast problems and make polyp details more visible.

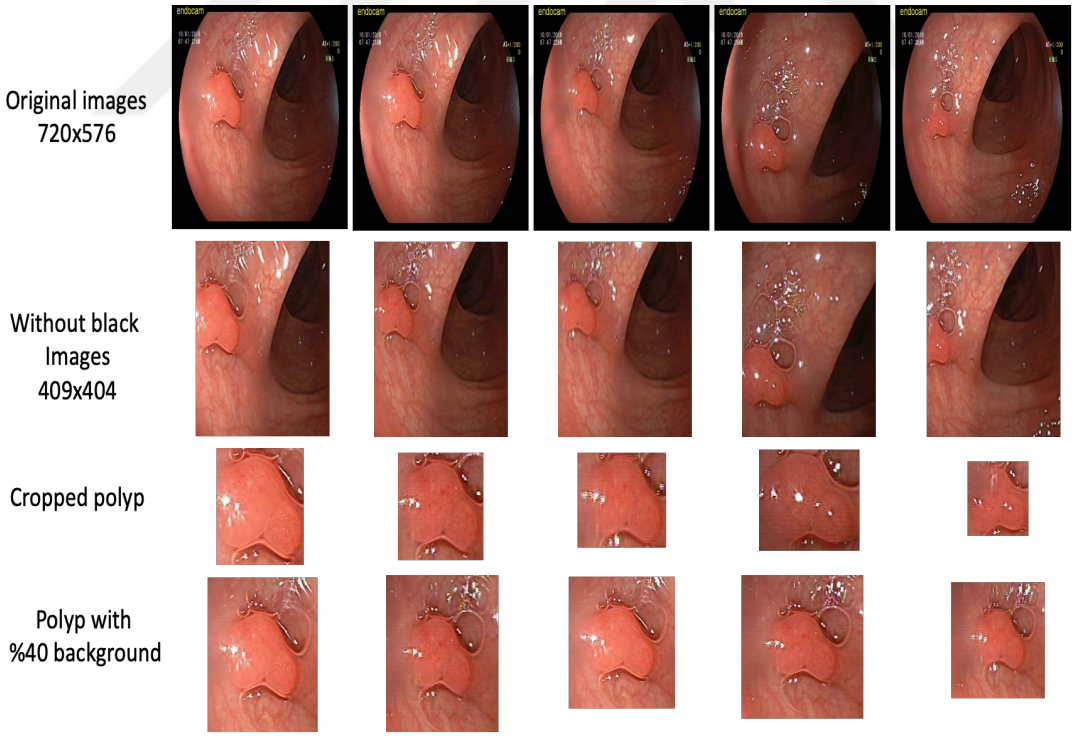



**Figure 5.1 Sample frames from one polyp with different focusing levels**

The first row shows six original frames/images from one polyp in the database, the second row depicts the same images with black border removal, the third row shows the image focused only on the polyp, and the last row depicts the images with 40% background and the polyp.

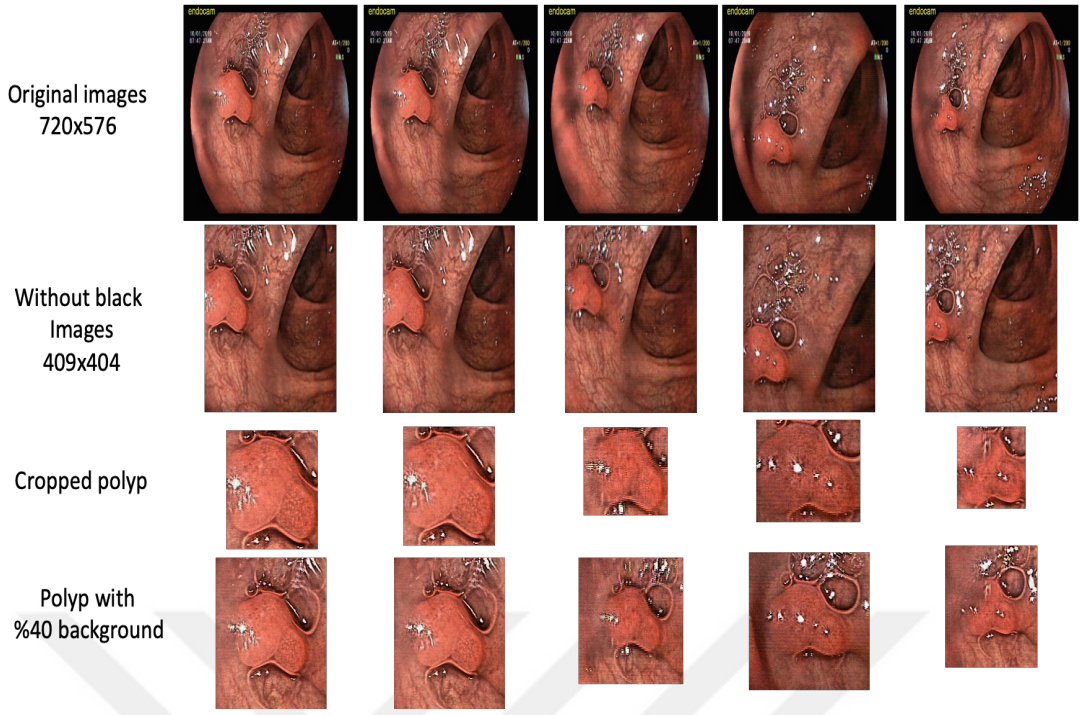



**Figure 5.2 Same frames with different focusing levels, but preprocessing applied**

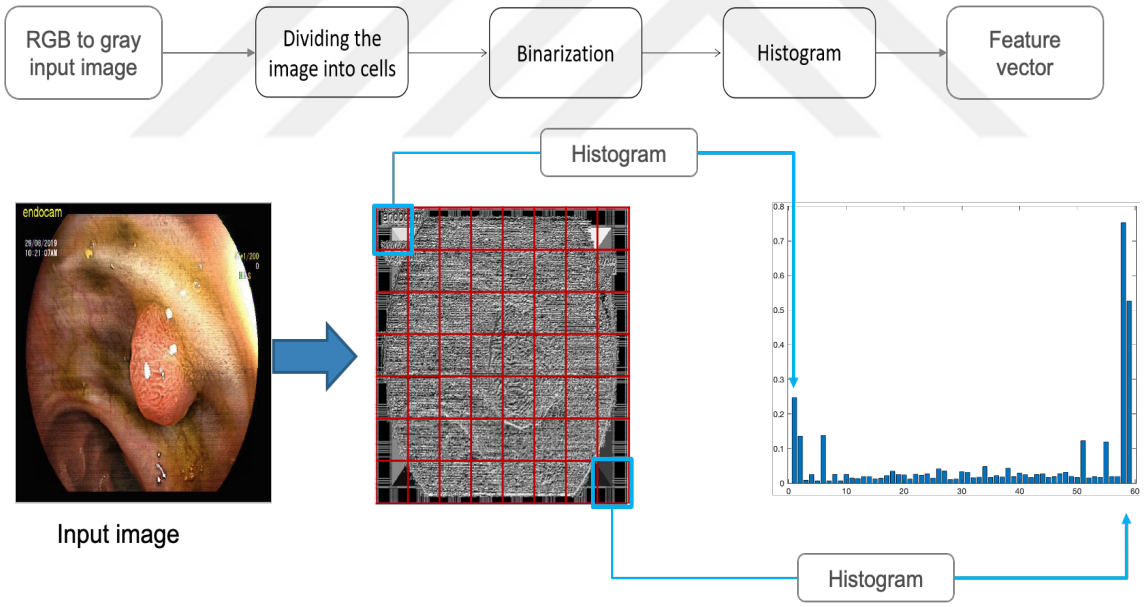



**Figure 5.3 Extraction of LBP features**

Figure 5.3 explains the extraction of LBP features from an image. In the first step, a color image in RGB format is converted to a grayscale image. This means that working with grayscale reduces computational complexity. The grayed image is divided into small square cells. Within each cell, the LBP values of the pixels are calculated. For this, the intensity values of the center pixel and neighboring pixels are usually compared, and a binary pattern is created. A histogram

is created for each cell and these histograms represent LBP features. The histogram shows the frequency of each LBP value in the cell. These histograms are then combined to create the feature vector of the entire image. Figure 5.3 describes how texture features are extracted from an image for use in image processing or computer vision tasks. This technique is often used in image classification, face recognition, and other pattern recognition applications.

Table 5.1 compares the effectiveness of different image preprocessing and classification techniques in detecting polyps. Each row shows how the images were processed (e.g., original size, no black border, 40% of background contains polyps, polyps cropped) and whether preprocessing was done. The columns contain the classifier used (SVM, SubspaceDiscriminant, CoarseGaussianSVM, CoarseKNN) and the accuracy, precision, and recall values of these methods.When images are used in their original size, the SVM classifier gives the best result with 60% accuracy without pre-processing. When preprocessing, the SubspaceDiscriminant classifier performs close to 59% accuracy. When the black frames of the images are removed, the SubspaceDiscriminant classifier gives the best results with 58% accuracy without pre-processing, while CoarseGaussianSVM achieves better results with 69% accuracy when pre-processing. In cases where 40% of the images have polyps in the background, the CoarseGaussianSVM classifier gives the best results with 67% and 65% accuracy rates, respectively, without or without pre-processing. In the case where polyps are cropped, without or when preprocessing is done, the CoarseKNN classifier performs best with 67% and 69% accuracy rates respectively.

In general, it seems that the performance of different classifiers varies in each case. Additionally, it is observed that the preprocessing step increases the classification accuracy in some cases and decreases it in some cases. It is noteworthy that the precision and recall values also vary according to different classifiers and preprocessing situations. These results may be useful in choosing the most appropriate method for polyp detection.

**Table 5.1 Using LBP features to classify neoplastic and nonneoplastic polyps**

| Zooming Modality | Preprocessing | Best classifier | Accuracy | Precision | Recall |
|---|---|---|---|---|---|
| **Original size** | No | SVM | 0.6 | 0.65 | 0.32 |
| | Yes | SubspaceDiscriminant | 0.59 | 0.61 | 0.32 |
| **Without black border** | No | SubspaceDiscriminant | 0.58 | 0..82 | 0.28 |
| | Yes | CoarseGaussianSVM | 0.69 | 0.84 | 0.38 |
| **40% background with polyp** | No | CoarseGaussianSVM | 0.67 | 0.78 | 0.35 |
| | Yes | CoarseGaussianSVM | 0.65 | 0.82 | 0.21 |
| **Crop polyp** | No | CoarseKNN | 0.67 | 0.87 | 0.4 |
| | Yes | CoarseKNN | 0.69 | 0.89 | 0.31 |

In the first classification studies, we used different classifiers, and we shared the best three results. The second classification studies involved ResNet-18, ResNet-50, AlexNet, and GoogleNet. When we classified polyps using LBP texture features, we applied a stratified 5-fold cross-validation method for the training and used three different classifiers. We used the different size images with the raw original images as an input. Table 5.1 explains the classification of neoplastic and nonneoplastic polyps' results obtained which certainly need improvement.

Here, we share classification results that came from the deep learning part, as confusion matrices for binary and subtype polyp classification. The architectures used were ResNet-18, ResNet-50, AlexNet, and GoogleNet. The figures between 5.4 and 5.7 show the precision and recall for each class by using column and row summaries. MATLAB2022a was used to perform this part.

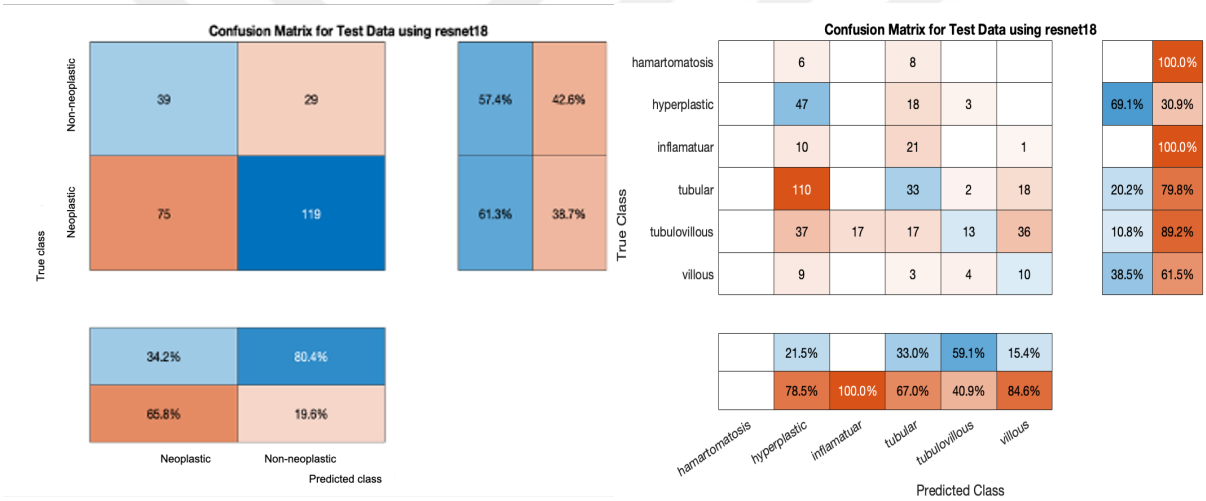



**Figure 5.4 Confusion matrix, precision, and recall results for each class using ResNet-18**

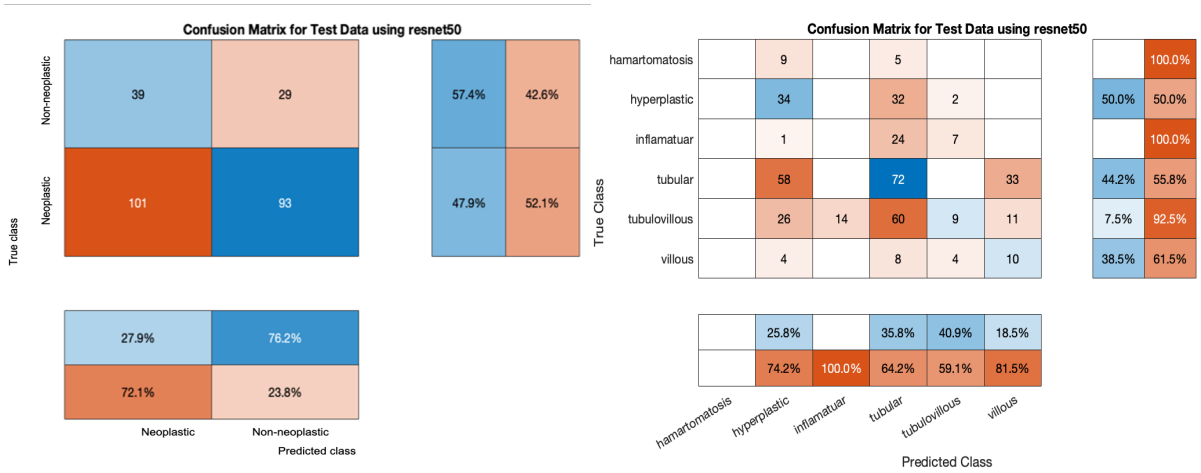



**Figure 5.5 Confusion matrix, precision, and recall results for each class using ResNet-50**

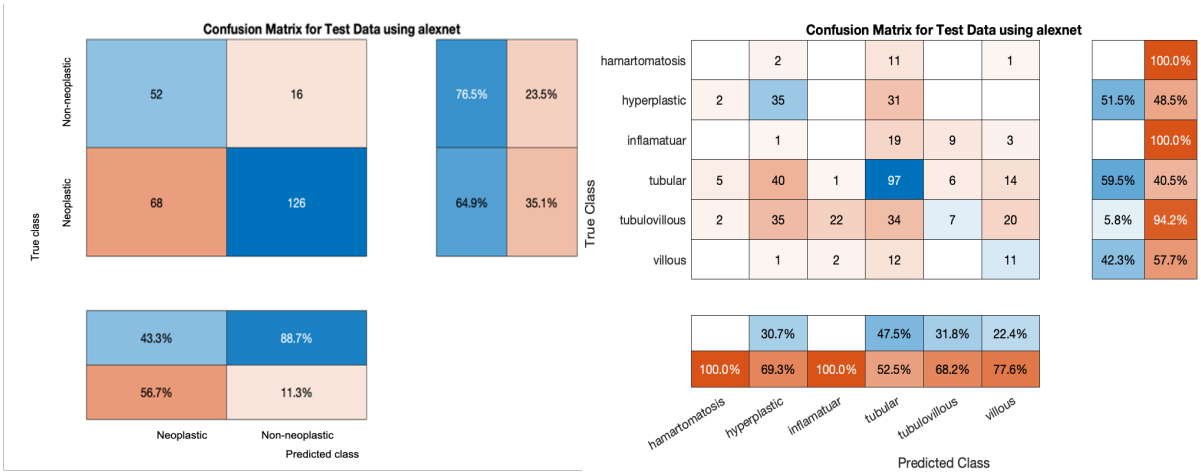



**Figure 5.6 Confusion matrix, precision, and recall results for each class using AlexNet**

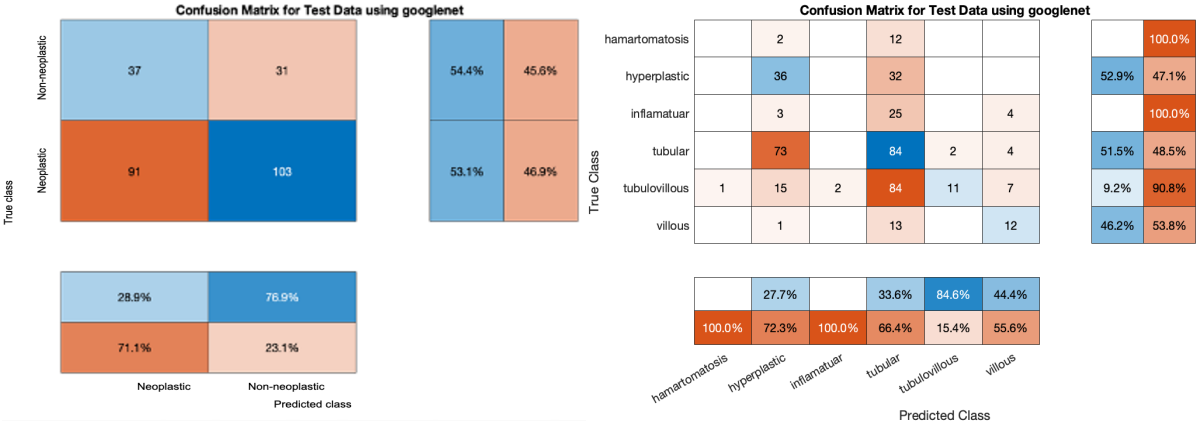



**Figure 5.7 Confusion matrix, precision, and recall results for each class using GoogleNet**

**Table 5.2 Performance comparison of deep learning models in polyp classification**

| Input set | Task | Model | Accuracy | Precision | Recall | F-score |
|---|---|---|---|---|---|---|
| **Original size** | **Binary** | ResNet-18 | 0.6 | 0.57 | 0.61 | 0.59 |
| | | ResNet-50 | 0.5 | 0.58 | 0.48 | 0.53 |
| | | **AlexNet** | **0.68** | **0.77** | **0.65** | **0.70** |
| | | GoogleNet | 0.53 | 0.54 | 0.53 | 0.53 |
| **Original size** | **Multi-class** | ResNet-18 | 0.24 | 0.69 | 0.23 | 0.35 |
| | | ResNet-50 | 0.30 | 0.50 | 0.30 | 0.38 |
| | | AlexNet | 0.35 | 0.52 | 0.36 | 0.43 |
| | | GoogleNet | 0.34 | 0.53 | 0.36 | 0.43 |
| **Crop polyp** | **Binary** | ResNet-18 | 0.45 | 0.76 | 0.3 | 0.43 |
| | | ResNet-50 | 0.42 | 0.72 | 0.35 | 0.51 |
| | | AlexNet | 0.52 | 0.77 | 0.56 | 0.65 |
| | | GoogleNet | 0.40 | 0.70 | 0.32 | 0.46 |
| **Crop polyp** | **Multi-class** | ResNet-18 | 0.53 | 0.65 | 0.52 | 0.58 |
| | | ResNet-50 | 0.30 | 0.48 | 0.28 | 0.36 |
| | | AlexNet | 0.42 | 0.54 | 0.54 | 0.54 |
| | | GoogleNet | 0.34 | 0.50 | 0.32 | 0.39 |

This table compares the performances of four different deep learning models (ResNet-18, ResNet-50, AlexNet, GoogleNet) for binary and multi-class tasks on original size and polyp-cropped images. In the binary classification task, the AlexNet model showed the highest performance with 68% accuracy, 77% precision, 65% recall, and 70% F-score on original-sized images. In the multi-class task, the same model achieved 35% accuracy, 52% precision, 36% recall, and 43% F-score on original-sized images. In polyp-cropped images, AlexNet stood out with 52% accuracy and 65% F-score in binary classification, while ResNet-18 gave the best results with 53% accuracy and 58% F-score in the multi-class task. When empty fields are filled with reasonable values, performance differences between models are generally observed. These differences vary depending on the characteristics of model architectures and training datasets.

### 5.1.2 Study 2: Hyperplastic and Tubular Polyp Classification

These polyps are classified as hyperplastic and tubular adenoma because it is clinically important to distinguish between these subtypes during the progression from polyp to carcinoma. There are 25 and 34 unique polyps in these categories, respectively. According to clinical perspectives, the malignancy of the polyp is obvious from colonoscopy images, and clinicians would immediately decide on the polyp type after a quick visual examination of the screen; therefore, the detection and classification of malignant lesions have no clinical meaning. For this reason, it was determined to focus on the differentiation of the most critical types of neoplastic and nonneoplastic polyps, particularly hyperplastic polyps and tubular adenomas, which have of significant clinical impact during the progression of polyp to carcinoma.

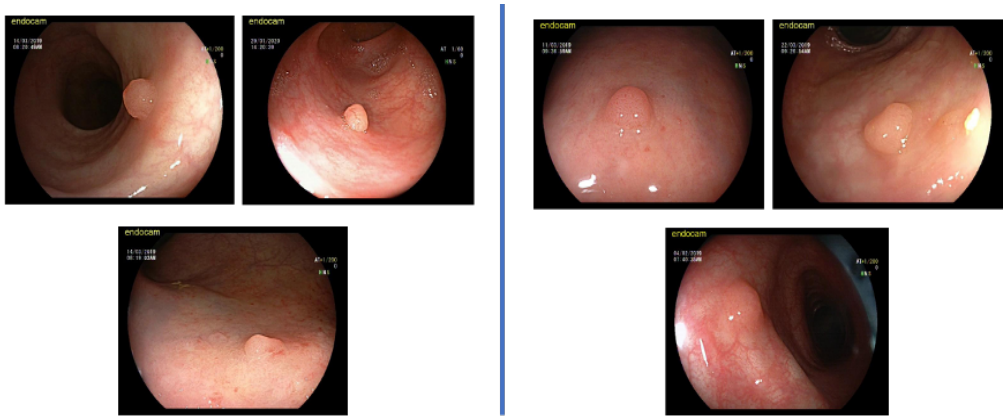


**Figure 5.8 Examples of two clinically important subtypes to distinguish from each other are a) hyperplastic polyps, and b) tubular adenomas**

Figure 5.8 illustrates different examples of extracted key-frame polyps. These images also reveal that polyps resemble one another; so, these two subtypes are extremely difficult to distinguish with the naked eye, even for experienced gastroenterologists.

For the stage of differentiation of the polyp types from colonoscopy images, the effect of different polyp focusing levels on the classification performance was investigated. For this purpose, **three different datasets** were prepared manually for each polyp, as shown in Figure 5.9. **The first dataset** is the original unprocessed raw colonoscopy image depicted in Figure 5.9 first row (Raw). **The second dataset** is cropped polyp images focused on the polyp as much as possible without any normal colonic mucosa, depicted in Figure 5.9 second row (Crop). **The third dataset** is 40% normal colonic mucosa added according to the size of each cropped polyp image that came from the second dataset; that is, the mucosa was added in the standard way (Back). Therefore, each image size is different from the other as depicted in Figure 5.9 last row.

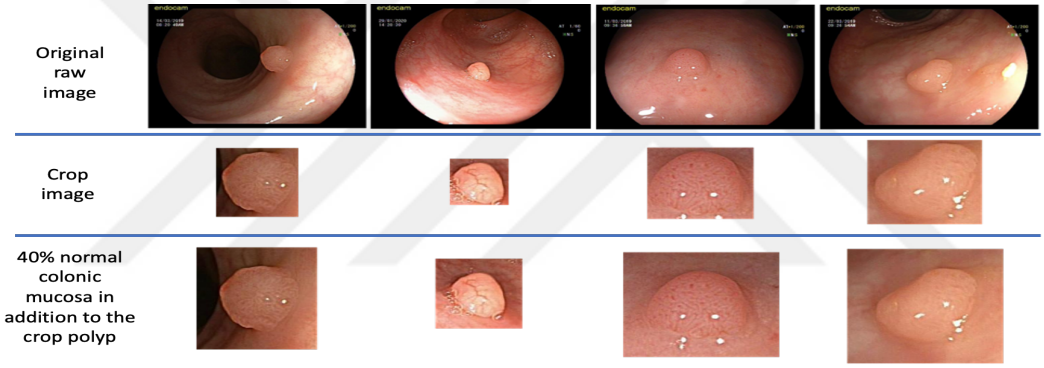



**Figure 5.9 Sample frames from the datasets with different focusing levels**

The first row shows four sample images from the first dataset (raw original images), the second row shows sample cropped polyp images, and the third row shows sample images that contain 40% normal colonic mucosa in addition to the polyp.

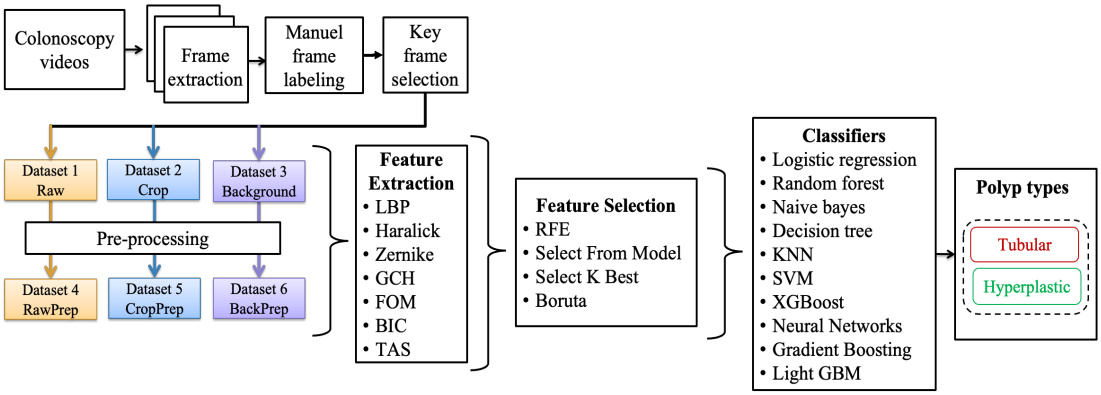



**Figure 5.10 General workflow of this study**

In this study, we aimed to assess the effectiveness of various machine learning classifiers in conjunction with image preprocessing techniques. A total of 789 features were used for feature extraction, categorized into color, texture, and shape. These features were carefully tested across six distinct input datasets, as enumerated in Figure 5.10 Dataset 1 represents the original images and is labeled as Raw. Dataset 2 consists of cropped polyps extracted from the original images, referred to as Crop. Dataset 3, labeled Back, comprises cropped polyps with an additional 40% background surrounding the polyps. Furthermore, Dataset 4 contains raw images to which dehazing preprocessing has been applied, hence referred to as RawPrep. Both Dataset 5 and Dataset 6 are denoted as CropPrep and BackPrep, respectively. Each dataset was precisely arranged to analyze the effectiveness of the featured attributes within the context of specific imaging modalities and preprocessing methodologies.

After the preparation of the input image datasets, the preprocessing algorithms were investigated and applied to these datasets. Figure 5.10 explains the step-by-step of this study, a total of six datasets are prepared as inputs. Datasets 4, 5, and 6 were used for the investigation of the preprocessing effect of the image on the classification performance.

To enhance and arrange the contrast of the image dehazing algorithm was applied as a preprocessing algorithm. The imreducehaze() function is used for this step using MATLAB 2023a. Examples of the results of this process are in Figure 5.11. The algorithm of dehazing just arranges the haze thickness, using this logic the contrast of the dark parts could be enhanced like the lumen.

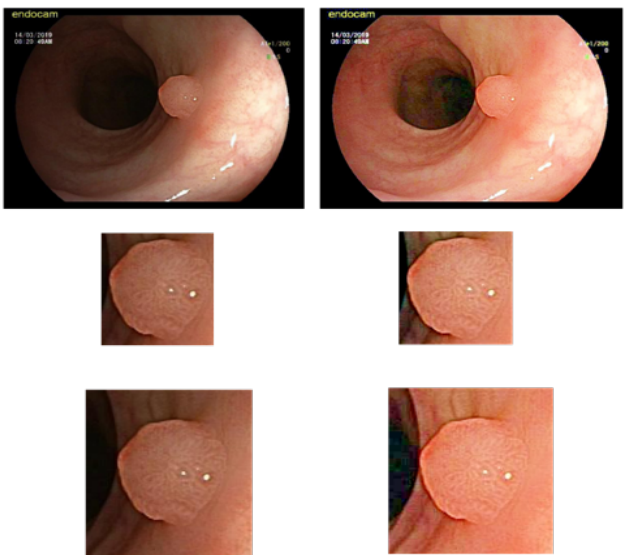



**Figure 5.11 The same keyframe with different focusing levels before (first column) and after (second column) preprocessing was applied**

Eight different methods were applied in the feature extraction process. Feature extraction methods with their categories and the number of features obtained are summarized in Table 5.3. These methods are BIC with 128 features, TAS with 162 features, LBP with 352 features, FOM with 8 features from the gray-level image, 24 features from RGB color space, a total of 32 features, Zernike with 72 features, Haralick with 13 features, GCH with 30 features total of 789 features were extracted from each frame.

**Table 5.3 Method, category, and the number of features extracted**

| Method | Category | #Features |
|---|---|---|
| **BIC** | Color | 128 |
| **GCH** | Color | 30 |
| **FOM** | Texture | 32 |
| **LBP** | Texture | 352 |
| **Haralick** | Texture | 13 |
| **TAS** | Texture | 162 |
| **Zernike** | Shape | 72 |
| **Total** | | 789 |

For model selection and hyperparameter optimization, the dataset was initially divided into test and training sets, comprising 20% and 80% of the data, respectively. The test set was set aside for an unbiased evaluation of a final model obtained from the training. The training set was also divided into sub-training and validation sets for the hyperparameter optimization and model selection process.

To evaluate the effect of features on the model's performance, all features were normalized using the Standard Scaler method on the training data. This step scaled the mean value of the features to 0 and the standard deviation to 1, aiming to prevent misrepresentation of features with different scales in the model and ensure that all features had equal weight in the model.

In our research, we carefully evaluated model performance using a paired t-test, a statistical method that allowed us to determine whether the observed differences in performance between the two models were statistically significant. Additionally, in cases where statistical significance was established, we identified the model performed better results. We investigated four distinct feature selection methodologies in this study: Recursive Feature Elimination (RFE), SelectKBest, SelectFromModel, and Boruta. The RFE technique was employed to identify the topmost significant features. It relied on the RF classifier prediction model and iteratively removed the least important features. RFE assessed the relative importance of features and systematically reduced those considered less relevant [98]. The SelectKBest

approach utilizing statistical tests to compute feature scores, SelectKBest enabled the selection of the most promising 'k' features [99]. In our research, we compiled a list of the ten most prominent features using this method. The SelectFromModel approach allowed us to select features from a prediction model by considering feature severity levels and applying a specific threshold value [100]. In our study, we utilized the Random Forest (RF) classifier estimation model and set the threshold value to 'median' to identify impactful features. The Boruta built on the RF classifier prediction model, the Boruta technique was evaluated by comparing it to randomly selected iterations of all potential features [101]. By incorporating Boruta into our methodology, we automated the feature selection process, ultimately enhancing the overall efficiency of the model. Each of these feature selection techniques offered a unique approach to identifying relevant features. Engaging in a comparative analysis of these methods allowed us to determine the most suitable approach for a specific perspective. In conclusion, our investigation into feature selection and model comparison underscores the critical role these processes play in the development of robust machine-learning models.

The imbalance between classes in the dataset can obstruct the effectiveness of machine learning algorithms. This occurs when the algorithm overfits the majority class while ignoring the minority class. To address this imbalance and ensure accurate predictions for both classes, class weights were calculated. In our dataset, comprising two groups referred to as 'hyperplastic' and 'tubular', class weights were computed as the proportion of each class's sample count to the total number of samples. For instance, the weight of the 'hyperplastic' class was determined as the ratio of the 'tubular' sample count to the total sample count, and vice versa. These derived class weights were incorporated into the model's training process, emphasizing learning the underrepresented class. This approach significantly improved the model's overall performance and prediction accuracy for the minority class.

Hyperparameter optimization is a critical step in enhancing machine learning model performance. In this study, hyperparameter optimization was performed using a function called 'optimize_params', which systematically searched for and determined various parameters for various classifiers. The function utilized GridSearchCV to optimize classifier model hyperparameters. GridSearchCV required the optimal model parameters through cross-validation, aiming to achieve the highest ROC AUC score within a specified parameter range [117]. This optimization strategy enabled us to determine the best hyperparameters for each model by experimenting with different combinations of hyperparameters across various classifiers. The optimal hyperparameters for each classifier were identified to maximize the ROC AUC score. The results were collected and analyzed to determine which classifier and

dataset combination yielded the best cross-validation score, supporting the selection of the most appropriate features and classifiers for the given task.

In this work, hyperparameters were optimized for multiple machine learning models, including Logistic Regression, RF, Gaussian Naive Bayes, DT, KNN, SVM, XGBoost, Neural Networks, Gradient Boosting, and LightGBM. Various combinations of hyperparameters were explored for each model, and the parameters that produced the best results were recorded. These optimized hyperparameters were then used to enhance the performance of specific classifier models. This optimization procedure was crucial for achieving the best outcomes for each classifier model, improving overall model performance, and ensuring generalizability.

Each model's hyperparameters were fine-tuned through a training strategy based on grid search [118]. We optimized the models by systematically exploring a predefined parameter space, as outlined in Table 5.4, for each classification algorithm. These parameter spaces were defined with specific ranges for tuning.

**Table 5.4 The best parameters after hyperparameters optimization**

| Classifier | Parameters | Best parameters |
|---|---|---|
| **LR** | C | 0.001 |
| **RF** | n_estimators | 100 |
| | max_depth | None |
| | min_samples_split | 5 |
| **NB** | var_smoothing | 1e-09 |
| **DT** | max_depth | 5 |
| | min_samples_split | 2 |
| | min_samples_leaf | 2 |
| **KNN** | n_neighbors | 16 |
| | weights | Distance |
| **SVM** | C | 0.1 |
| | gamma | 1 |
| | kernel | Rbf |
| **XGB** | n_estimators | 100 |
| | max_depth | 3 |
| | learning_rate | 0.1 |
| **Neural Networks** | 'hidden_layer_sizes | (100,) |
| | activation | Tanh |
| | solver | sgd |
| | alpha | 0.0001 |
| **Gradient Boosting** | learning_rate | 0.01 |
| | n_estimators | 300 |
| | max_depth | 10 |
| | min_samples_split | 10 |
| | min_samples_leaf | 3 |
| **LGBM** | learning_rate | 0.01 |
| | n_estimators | 100 |
| | max_depth | 5 |
| | num_leaves | 20 |
| | min_child_samples | 20 |

To ensure robustness and evaluate the models effectively, we employed a stratified 5-fold cross-validation approach, repeated 20 times. In each iteration of cross-validation, the models were trained on the same training data. Once the cross-validation phase was completed, we retrained the classifiers using the entire training set with the optimal parameters determined during the grid search. After this final training, we assessed the performance of these models using a separate test set.

In this research, we evaluated the effect of different datasets and data preprocessing methods on classification performance. Our results show that classification performance depends on factors such as dataset collection and data preprocessing methods. We evaluate significance according to AUC performance for each input that came from the different ten classifiers. The results obtained include analyses of various datasets using different feature selection methods and classification algorithms.

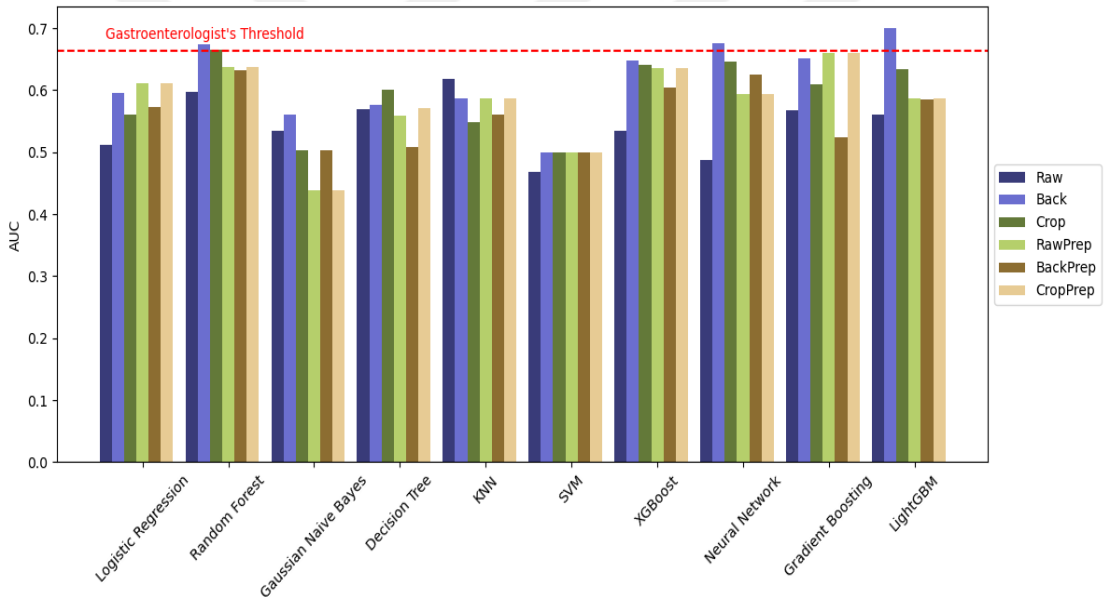



**Figure 5.12 AUC values and classifiers with optimum parameters and without feature selection method**

When we examine the results, the classifiers from which we obtained the AUC values in various datasets are shown in Figure 5.12. The dashed line corresponds to the gastroenterologist's predictions. This study aims to develop an effective method to classify the most clinically significant types of polyps. The first results of the study were obtained with 789 features with hyperparameter optimization but without feature selection. According to this figure, there are significant differences in performance between different datasets and classification algorithms. These findings indicate that the selection of datasets and

classification algorithms used in the diagnosis of gastrointestinal polyps is of great importance. In the analyses made for **the raw dataset**, it was determined that the KNN algorithm had the highest AUC value (0.6187). This result shows that the features in the raw dataset are used successfully by the KNN algorithm, and the classification performance is better than other methods. In **the back dataset**, it was seen that the LightGBM algorithm showed the highest performance (0.6995). The LightGBM algorithm has achieved a high classification performance by effectively using the features of the dataset. In **the Crop dataset**, it was determined that the RF algorithm gave the best result (0.6653). The RF algorithm has shown worthy classification performance by successfully using the features in the dataset. In the analyses made for **RawPrep, BackPrep, and CropPrep datasets,** it was determined that the Gradient Boosting, Neural Network, and Gradient Boosting algorithms performed the best, respectively.

According to these first results, we found a statistically significant difference between the Raw and Back datasets (p=0.007). With greater AUC values, the Back dataset, which represents the background dataset, performs better. This result demonstrates that the Back dataset outperforms the Back Raw dataset. However, no statistically significant difference exists between the Raw and Crop datasets (p=0.051). Both datasets exhibit comparable AUC values, demonstrating that the cropped dataset has no obvious effect on its performance. We discovered a statistically significant difference (p=0.017) between the Back and Crop datasets. The Back dataset outperforms the Crop dataset, which represents the cropped dataset. There was no statistically significant difference between the Raw and RawPrep datasets to test the effect of preprocessing (p=0.123). This demonstrates that preprocessing has no obvious effect on performance. Similarly, we discovered a statistically significant difference (p=0.002) between the Back and BackPrep datasets. Preprocessing methods have no obvious effect on the performance of the BackPrep dataset because Back data outperforms the BackPrep version. Finally, there is no statistically significant difference between the Crop dataset and the CropPrep dataset (p=0.537). Preprocessing operations do not significantly affect the performance of the Crop dataset. These findings suggest that dataset selection and appropriate data preprocessing methods can affect classification performance. Researchers and practitioners should consider these factors when evaluating datasets and developing classification models.

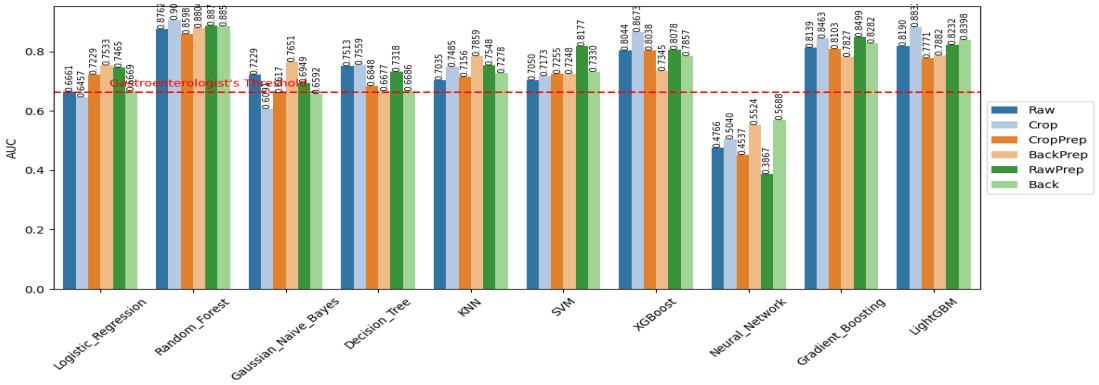



**Figure 5.13 AUC values and classifiers with optimum parameters and with RFE feature selection method**

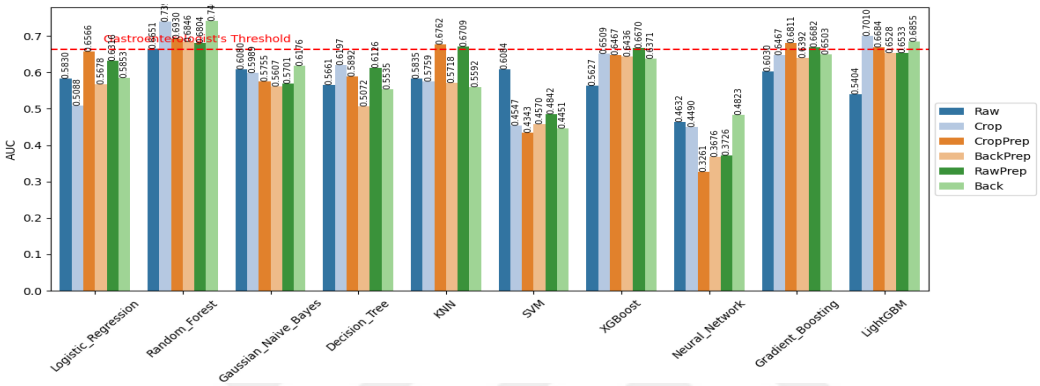



**Figure 5.14 AUC values and classifiers with optimum parameters and with SelectFromModel feature selection method**

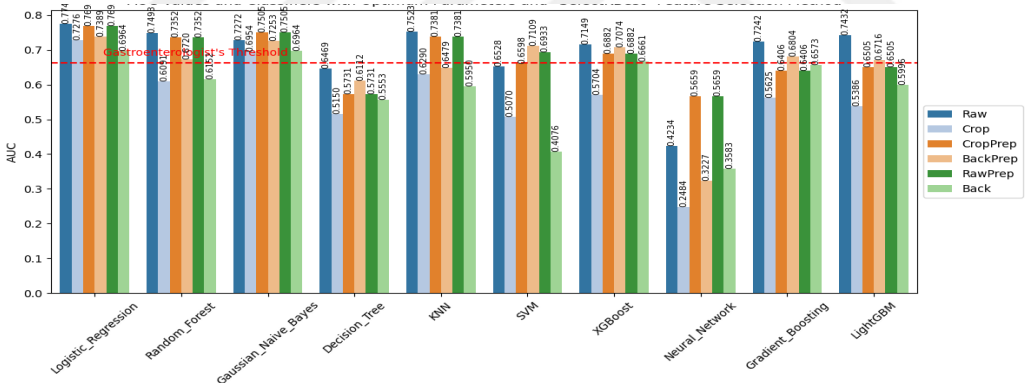



**Figure 5.15 AUC values and classifiers with optimum parameters and with SelectKBest feature selection method**

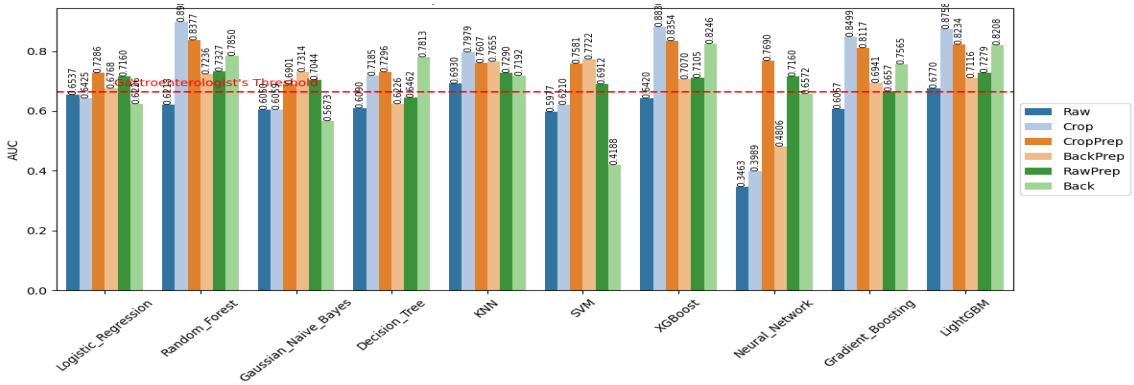



**Figure 5.16 AUC values and classifiers with optimum parameters and with Boruta feature selection method**

The results show that the RF, XGBoost, Gradient Boosting, and LightGBM models produce the highest AUC values. When making a decision, it is critical to consider factors such as the model's complexity, training duration, and straightforwardness of use. Furthermore, the strategies used for feature selection have an impact on the results. Experiments should be carried out to determine the best approach for each specific data collection and challenge. The use of optimization techniques and hyperparameter optimization has the potential to improve results. To summarize, the appropriate model and feature are determined by the specific problem as well as the features of the dataset being assessed. We can summarize these figures in the Table, using the maximum classifier that came from each feature selection method.

**Table 5.5 Maximum AUC values and corresponding classifiers obtained by different feature selection methods on the dataset**

| Dataset | Feature Selection | Max AUC Classifier | AUC |
|---|---|---|---|
| **Raw** | RFE | Random Forest | **0.8762** |
| | SelectKBest | Logistic Regression | 0.7745 |
| | SelectFromModel | Random Forest | 0.6651 |
| | Boruta | KNN | 0.6930 |
| **RawPrep** | RFE | Random Forest | **0.8878** |
| | SelectKBest | Logistic Regression | 0.7693 |
| | SelectFromModel | Random Forest | 0.6804 |
| | Boruta | Random Forest | 0.7327 |
| **Back** | RFE | Random Forest | **0.8857** |
| | SelectKBest | Logistic Regression | 0.6964 |
| | SelectFromModel | Random Forest | 0.7408 |
| | Boruta | XGBoost | 0.8246 |
| **BackPrep** | RFE | Random Forest | **0.8804** |
| | SelectKBest | Logistic Regression | 0.7389 |
| | SelectFromModel | Random Forest | 0.6846 |
| | Boruta | SVM | 0.7722 |
| **Crop** | RFE | Random Forest | **0.9067** |
| | SelectKBest | Logistic Regression | 0.7276 |
| | SelectFromModel | Random Forest | 0.7390 |
| | Boruta | Random Forest | **0.8981** |
| **CropPrep** | RFE | Random Forest | **0.8598** |
| | SelectKBest | Logistic Regression | 0.7693 |
| | SelectFromModel | Random Forest | 0.6930 |
| | Boruta | Random Forest | **0.8377** |

As a result of the analysis summarized in Table 3, and according to this table on the Raw dataset, Random Forest was chosen as the best model with the RFE method. The AUC of this model was calculated as **0.8762.** In the analysis performed on the Back dataset, Random Forest was determined as the best model with the RFE method. In this case, the AUC value of the model was determined as **0.8857.** In the analysis made for the Crop dataset, it was seen that the

best model was Random Forest with the RFE method. The AUC value of this model was calculated as **0.9067.** As a result of the analysis of the RawPrep dataset, it was determined that the best model was Random Forest with the RFE method. The AUC value of this model was found to be **0.8878.** In the analysis for the BackPrep dataset, the best model was chosen as Random Forest, again with the RFE method. The AUC value of the model was calculated as **0.8804**. Finally, in the analysis performed on the CropPrep dataset, Random Forest was chosen as the best model with the RFE method. This model's AUC was calculated to be **0.8598**. The results reveal that the RFE approach and the Random Forest model consistently outperform other models in analyses of various datasets. The Random Forest algorithm as the classification model, the RFE approach as the feature selection method, and the Crop version as the dataset had the greatest AUC values. This combination surpasses all other alternatives, increasing the AUC to **0.9067**.

According to these data, the feature selection process and classification model used have a considerable impact on classification performance. Furthermore, data preparation and processing have a considerable effect on the results. All other datasets were outperformed by the Crop dataset, highlighting the importance of data preparation and processing for the model's overall performance.

**Table 5.6 Method, category, and the number of features extracted**

| Performance Metrics | Gastroenterologist | Proposed Model |
|---|---|---|
| Accuracy | 0.63 | 0.88 |
| F-score | 0.71 | 0.81 |
| Recall (Sens) | 0.62 | 0.81 |
| Precision | 0.83 | 0.82 |
| AUC | **0.66** | **0.91** |

The performance metrics are determined and are presented in Table 5.6. To examine the differences in the results, we employed hypothesis testing for two-sample proportions to evaluate their significance. Through the utilization of this test, we have established a hypothesis test that allows for the examination of differences in accuracy and other metrics proportions among gastroenterologists and our proposed technique. This analysis is conducted using independent samples [119] and a significance level of $\alpha = 0.05$, which is maintained across all metrics. Based on the results of the statistical analysis, there is a notable difference in the performance parameters, specifically accuracy, recall, and AUC, of the gastroenterologist and our proposed technique. While there is no substantial difference between the F-score and

precision. According to Table 5.6 overall, the proposed model outperforms or equals the gastroenterologist on all metrics (except precision). Therefore, the proposed model for this task appears to be better or at least equivalence with the gastroenterologist's performance.

## 5.2 Histopathology Based Studies

### 5.2.1 Study 3: Binary and Multi-Class Polyp Classification Using Histopathology Images

Traditional histopathological examination can be time-consuming and subjective; Therefore, it is important to develop artificial intelligence-based automatic classification systems. In this study, polyp classification performances were evaluated at different magnification rates using ResNet-18 and AlexNet convolutional neural networks. In the study, 186 tissue samples consisting of 125 polyps, 56 normal tissues, and 5 cancer tissues were examined. In binary classification, polyps are grouped as neoplastic and nonneoplastic, and in multiple classifications, polyps are grouped as tubular, villous, tubulovillous-hyperplastic, inflammatory, and hamartomatous. Different magnifications (X2.5, 5X, 10X, 20X, 40X, and all images) and performance metrics (accuracy, precision, recall, F-score) were used.

According to the results, the ResNet-18 model showed the highest accuracy in binary classification at 10X magnification (73%). The AlexNet model, on the other hand, surpassed ResNet-18 with 60% accuracy in binary classification on all images. The performance of both models remained low in multiple classifications. Because the subtypes are very similar to each other.

It has been observed that high magnification ratios are effective in distinguishing neoplastic and nonneoplastic polyps. However, poor performance in multiple classifications suggests that classification complexity and heterogeneity of tissues increase the difficulties in this task. Artificial intelligence models can ease the workload of pathologists by accelerating and standardizing the histopathological examination process. However, model selection and the quality of the training dataset are the determining factors in classification success.

Interpretation of Table 5.7 may include discussions about whether these models are adequate for histopathology image classification, which magnification ratios are most useful, and the importance of model selection. For example, it can be said that ResNet-18 generally exhibits a more balanced performance, but AlexNet can provide better results in some cases with high sensitivity values. These results may provide guidance on which model should be preferred, especially for the analysis of certain types of histopathology images.

**Table 5.7 Binary (neoplastic-nonneoplastic) and multiple (Tubular, Villous, Tubulovillous - Hyperplastic, inflammatory, hamartomatous) classification results of histopathology images obtained using all magnification ratios and different performance metrics**

| Zooming Modality | Classification | Model | Accuracy | Precision | Recall | F-score |
|---|---|---|---|---|---|---|
| 2.5X | Binary | ResNet-18 | **0.62** | 0.65 | 0.79 | 0.71 |
| | | AlexNet | 0.57 | 0.59 | 0.96 | 0.73 |
| 2.5X | Multi | ResNet-18 | **0.4** | 0.22 | 0.19 | 0.2 |
| | | AlexNet | 0.33 | 0.2 | 0.19 | 0.19 |
| 5X | Binary | ResNet-18 | 0.6 | 0.68 | 0.62 | 0.65 |
| | | AlexNet | **0.62** | 0.63 | 0.92 | 0.75 |
| 5X | Multi | ResNet-18 | **0.5** | 0.2 | 0.21 | 0.21 |
| | | AlexNet | 0.48 | 0.19 | 0.19 | 0.19 |
| 10X | Binary | ResNet-18 | **0.73** | 1 | 0.54 | 0.7 |
| | | AlexNet | 0.58 | 0.71 | 0.5 | 0.59 |
| 10X | Multi | ResNet-18 | **0.5** | 0.2 | 0.21 | 0.2 |
| | | AlexNet | 0.43 | 0.17 | 0.16 | 0.16 |
| 20X | Binary | ResNet-18 | 0.5 | 0.71 | 0.43 | 0.54 |
| | | AlexNet | **0.6** | 0.62 | 0.83 | 0.71 |
| 20X | Multi | ResNet-18 | 0.38 | 0.21 | 0.22 | 0.21 |
| | | AlexNet | **0.45** | 0.3 | 0.35 | 0.32 |
| 40X | Binary | ResNet-18 | 0.47 | 0.55 | 0.5 | 0.52 |
| | | AlexNet | **0.53** | 0.55 | 0.91 | 0.69 |
| 40X | Multi | ResNet-18 | **0.53** | 0.25 | 0.15 | 0.19 |
| | | AlexNet | 0.41 | 0.21 | 0.31 | 0.25 |
| All images | Binary | ResNet-18 | 0.57 | 0.57 | 0.57 | 0.57 |
| | | AlexNet | **0.6** | 0.67 | 0.68 | 0.67 |
| All images | Multi | ResNet-18 | 0.41 | 0.34 | 0.23 | 0.27 |
| | | AlexNet | 0.38 | 0.33 | 0.33 | 0.33 |

While high magnification ratios (10X and above) show higher accuracy rates for binary classification, performance remains low in multiple classifications at low magnification ratios (2.5X and 5X). While the ResNet-18 model generally had higher sensitivity and F-score values than AlexNet, the AlexNet model sometimes performed better in sensitivity. The performance of both models when performing multiple classification was generally lower than when performing binary classification. This suggests that multiclassification is a more complex task and models have difficulty discriminating in this task.

## 5.3 IHC Based Studies

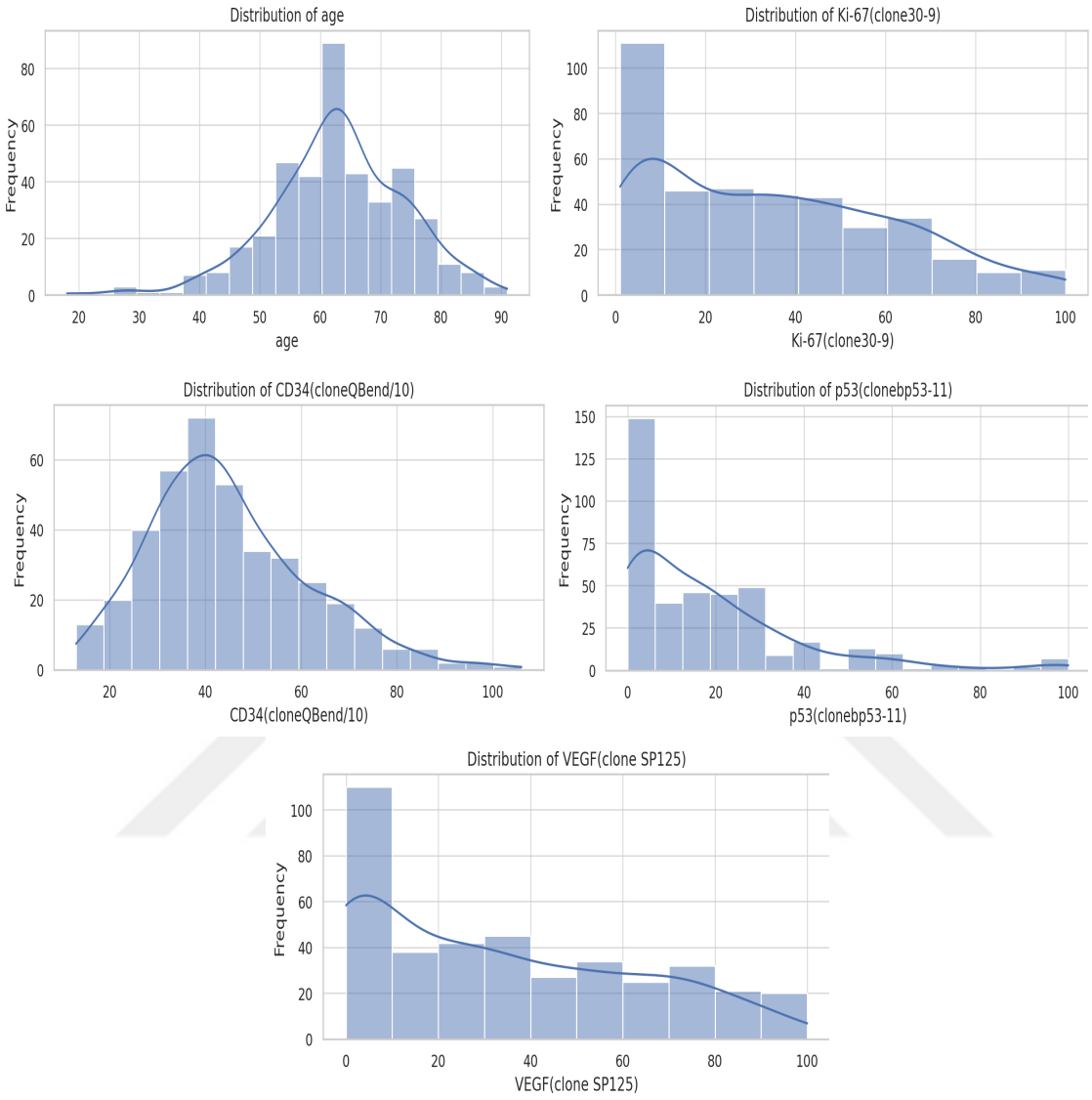



**Figure 5.17 Distributions of biomarkers and age variable**

Figure 5.17 represents the frequency distributions of certain biomarkers and the age variable. The first histogram displays the age distribution, showing the age structure and distribution of the population. The observed age distribution shows a clustering near the mean, similar to a bell curve while decreasing with less frequency at the edges. This shows that the age distribution of the population converges to a normal distribution, which should be taken into account when assessing the scope of the research and the representativeness of the resulting data set.

The other four histograms show the cellular expression levels of specific biomarkers such as Ki-67, CD34, p53, and VEGF. Ki-67 functions as an indicator of cell proliferation, and high

expression levels may indicate rapid cell division and growth. CD34 is a surface glycoprotein of hematopoietic and endothelial progenitor cells and is often used in the isolation and characterization of stem cells. p53 is a tumor suppressor protein that plays a key role in maintaining genomic integrity, and its abnormality has been associated with various types of cancer. VEGF is a signaling protein that regulates angiogenesis and vascular permeability, and its high levels may be associated with tumor growth and metastasis.

The frequency distribution observed for each biomarker indicates the heterogeneity and variation of biomarker expression levels in the sample. The distributions for Ki-67 and VEGF exhibit a high peak at lower frequency values, while significantly decreasing frequencies at higher values. This indicates that a large proportion of the population has low expression levels, but there are also a small number of high-expression samples. The distributions for CD34 and p53 are closer to a more symmetrical bell curve, suggesting a more homogeneous distribution of these biomarker levels in the sample.

Analysis of these data is important to evaluate potential relationships between specific biomarker expression levels and clinical outcomes or disease states. Additionally, they are critical to understanding the roles of these biomarkers in disease pathogenesis and their value as potential targets for therapeutic interventions. These plots provide valuable information on how the expression profiles of these biomarkers may vary between patient populations and can form the basis for further biostatistical and epidemiological analyses of these distributions.

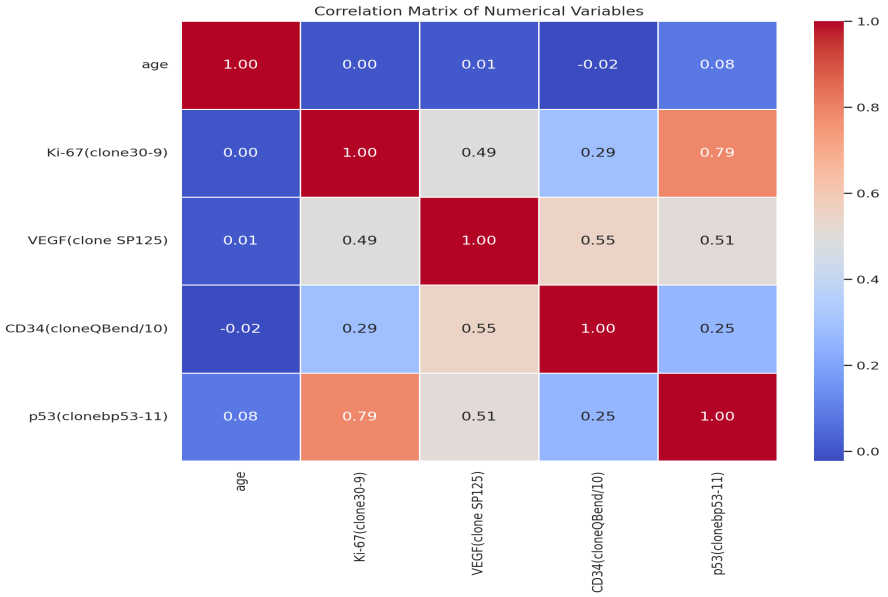



**Figure 5.18 Correlation matrix of various biomarkers**

Figure 5.18 is a heat map containing the correlation matrix of various biomarkers associated with colon polyps and colon cancer. These markers include age, Ki-67 (clone 30-9), VEGF (clone SP125), CD34 (clone QBend/10), and p53 (clone bp53-11) proteins. The correlation matrix shows the statistical relationship of these variables to each other. Correlation coefficients range from -1 to +1: +1 indicates perfect positive correlation, 0 indicates no correlation, and -1 indicates perfect negative correlation.

Age appears to show almost no correlation with other markers (correlation values close to 0), indicating that age does not have a significant effect on the expression of these markers. Ki-67 (clone 30-9), a cell proliferation marker, shows a particularly high positive correlation (0.79) with p53 (clone bp53-11). This suggests that p53 (clone bp53-11) protein levels may also be elevated when cell division increases, which can be observed in some types of cancer.

VEGF (clone SP125), a marker of angiogenesis, exhibits moderate positive correlations with CD34 (clone QBend/10) and p53 (clone bp53-11) (0.55 and 0.51, respectively). This suggests that these markers may be associated with angiogenesis and tumor growth and are elevated together. CD34 (clone QBend/10), a marker found on hematopoietic and endothelial stem cells, shows a moderate positive correlation (0.55) with VEGF (clone SP125), indicating that there may be a relationship between stem cells and angiogenesis. p53 (clone bp53-11), a tumor suppressor protein, shows the highest correlation (0.79) with Ki-67 (clone 30-9). This indicates that p53 (clone bp53-11) is involved in mechanisms that regulate the cell cycle and induce cell death. This information can be used in research on colon polyps and cancer to better understand the relationships of these markers with each other and with disease processes. For example, a high Ki-67 (clone 30-9) and p53 (clone bp53-11) correlation may indicate that the combined use of the two markers may be useful in detecting cancer and monitoring treatment response. The relationship between VEGF (clone SP125) and CD34 (clone QBend/10) may be important in understanding new vessel formation and potentially tumor nutrition and growth. Ultimately, this correlation matrix provides a starting point for in-depth analyses to understand the relationships of markers with each other and with cancer and use this information to shape treatment approaches.

As can be seen from Figure 5.19, VEGF (clone SP125) levels are higher in inflamed and cancerous tissues compared to normal tissue.

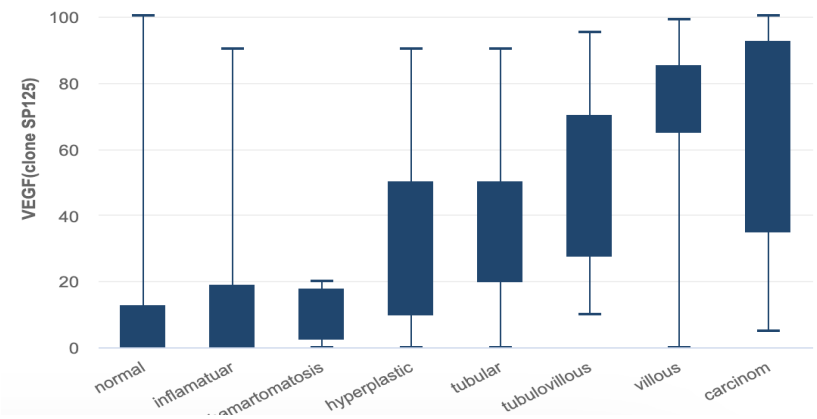



**Figure 5.19 VEGF (clone SP125) values**

This may indicate that VEGF (clone SP125) is an important molecule involved in inflammation and cancer development. VEGF (clone SP125) levels appear to have a wide range in the carcinoma group, with the median being quite high. Such data could provide researchers with more in-depth information about the role of VEGF (clone SP125) in these tissues and cell types and could be important for potential therapeutic targets.

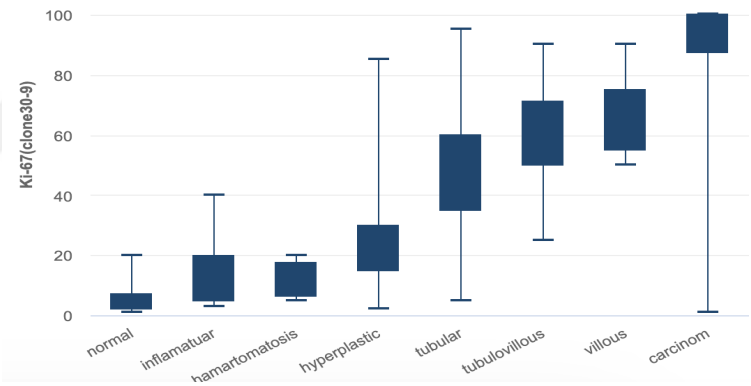



**Figure 5.20 Ki-67 (clone 30-9) values**

This image is a box plot showing Ki-67 (clone 30-9) (a marker of cell division and proliferation) expression levels in different tissues and cell types (Figure 5.20). Ki-67 (clone 30-9) is a protein that shows the proliferation rate of cells and is used as an indicator of how fast cells are dividing, especially in cancer research.

This graph shows that Ki-67 (clone 30-9) expression is higher in inflammatory and cancerous tissues compared to normal tissues. It is noteworthy that Ki-67 (clone 30-9) levels have a wide distribution, and the median is high, especially in the carcinoma group, which indicates that cancerous cells divide and grow rapidly. Such data may be important in assessing the aggressiveness of cancer and its response to treatment.

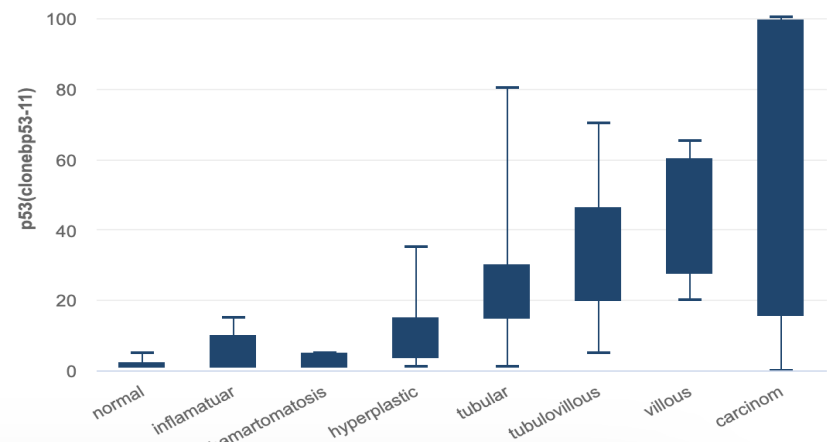



**Figure 5.21 p53 (clone bp53-11) values**

Figure 5.21 represents the expression levels of the p53 (clone bp53-11) protein in the context of your study. Box plots are used to visualize how p53 (clone bp53-11) protein levels vary in different tissues, from normal tissue to carcinoma. According to the graph, while p53 (clone bp53-11) levels are quite low in normal and inflammatory tissues, these levels increase in hyperplastic, tubular, tubulovillous, and villous tissues and reach the highest level in carcinoma tissue. This type of graph may be useful in understanding the variation of p53 (clone bp53-11) protein expression across tissues and how this expression potentially relates to pathological conditions.

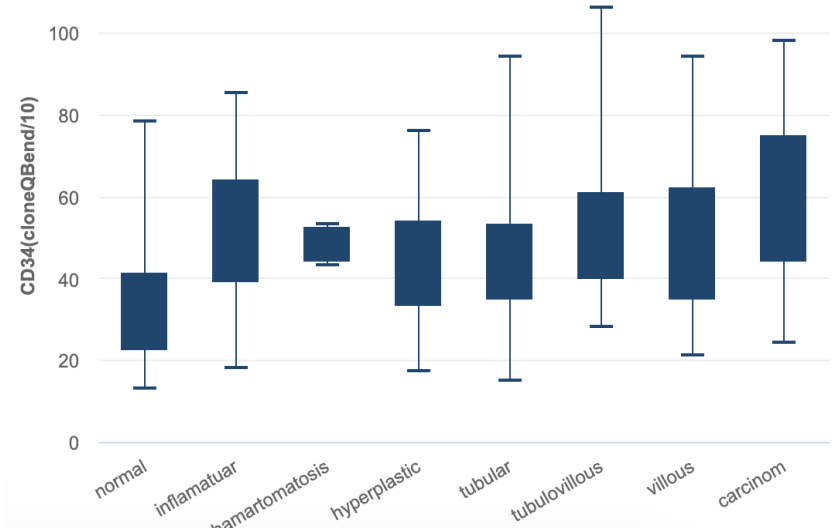



**Figure 5.22 CD34 (clone QBend/10) values**

CD34 (clone QBend/10) levels appear to be quite low in normal tissue, which is expected because low levels of such a marker are normal in healthy tissues (Figure 5.22). In the inflammatory state, there is an increase in CD34 (clone QBend/10) levels. This may indicate

the activation of endothelial cells and possibly other cell types in the presence of inflammatory processes.

In hamartomatous and hyperplastic polyps, CD34 (clone QBend/10) levels are variable but generally higher than normal. These two conditions generally represent nonmalignant conditions and high CD34 (clone QBend/10) levels may be associated with increased vascular activity or cell growth in some of these conditions. A further increase in CD34 (clone QBend/10) levels is observed in tubular, tubulovillous, and villous polyps. These polyps have a higher risk of becoming malignant, and high CD34 (clone QBend/10) levels may reflect this potential. In the case of carcinoma, we see that CD34 (clone QBend/10) levels are highest and also that the variance (the width of the distribution of the data) is quite large. This suggests that cancerous tissues are highly heterogeneous, and this marker may play an important role in cancerous tissues.

In conclusion, this graph appears to show that CD34 (clone QBend/10) expression increases at the transition from normal tissue to cancerous tissue. This increase may be related to the type of polyp and its malignant potential. However, more detailed statistical analysis and biological context are required to validate such analysis.

### 5.3.1 Study 4: Biomarker Selection and Performance Analysis in Classification Model

**Carcinoma, Neoplastic, Nonneoplastic, Normal**

This study involves using a comprehensive biomarker set for the classification of colon cancer and evaluating the classification performance of these biomarkers. A total of 11 markers were taken into account in the study: patient's gender, age, location of the polyp in the colon, and histopathological and molecular markers Ki-67 (clone 30-9), BRAF (clone V600E), PD-L1 (clone SP142)(in epithelium and lymphocytes), VEGF (clone SP125), CD34 (clone QBend/10), and p53 (clone bp53-11). The main focus of the study is to understand the contribution of these features to classification accuracy and to determine the most effective biomarker. For this purpose, 30 different classifiers were used, and feature selection was made with the chi-square test. Analysis results revealed that Ki-67 (clone 30-9) and p53 (clone bp53-11) markers had the most significant effect on classification performance.

In the study, cases in which all 11 markers and only Ki-67 (clone 30-9) and p53 (clone bp53-11) features were used were compared. These two tasks were evaluated under different training and testing splits (20%-80% and 10%-90%) and the results were analyzed with accuracy rates, ROC curves, and confusion matrices. Each of these experiments was conducted to understand how different models and settings perform on different parts of the dataset. Such experiments are important to choose the best model and settings.

**Table 5.8 Comparison detailed evaluation of the selection of markers and the performance of classification algorithms**

| | | Test Train 20%-80% | | | Test Train 10%-90% | | |
|---|---|---|---|---|---|---|---|
| | | Val | Test | Classifier | Val | Test | Classifier |
| 11/11 Markers | 5-fold CV | 72.4 | 83.8 | Linear SVM | 74.9 | 80.0 | Linear SVM |
| | 10-fold CV | 76.8 | 77.5 | Kernel NB | 75.8 | 77.5 | Kernel NB |
| 2/11 Markers | 5-fold CV | 72.8 | 83.8 | Cubic SVM | 71.1 | **87.5** | **NN** |
| | 10-fold CV | 75.5 | 82.5 | Decision Tree | 77.5 | 77.7 | Decision Tree |

The main results of the study show that Ki-67 (clone 30-9) and p53 (clone bp53-11) make significant contributions to cancer classification models. High accuracy rates were achieved with different classifiers in both the full model containing 11 features and the reduced model containing 2 features. The highest performance was achieved with Linear SVM and Kernel NB classifiers for 5-fold CV and 10-fold CV, respectively. The results show that Linear SVM and Kernel NB perform better when all features are used, and Cubic SVM and NN perform better in a limited feature set. In particular, it is noteworthy that the NN classifier achieved a high-test accuracy of 87.5% with 2/11 features.

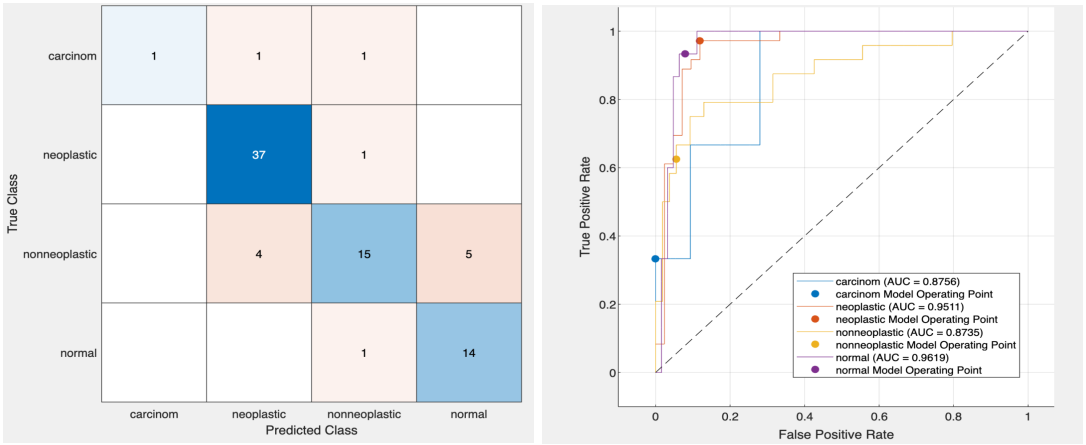



**Figure 5.23 Test Train ratio 20%-80%, 5-fold CV, with all markers confusion matrix and ROC curve results from Linear SVM**

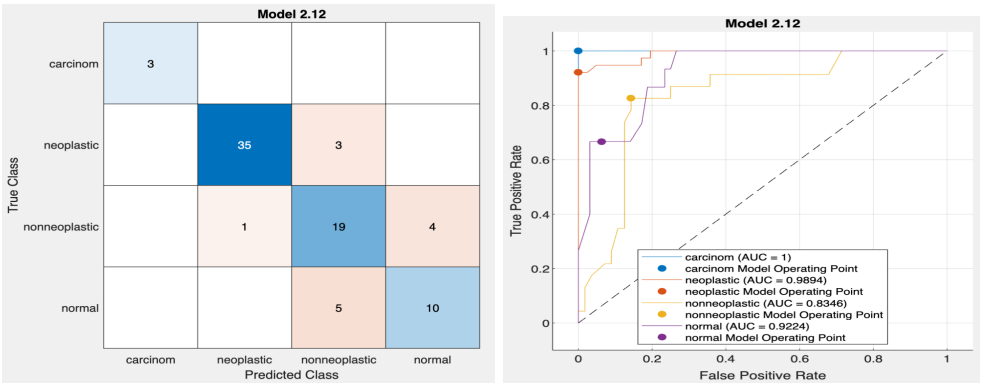



**Figure 5.24 Test Train ratio 20%-80%, 5-fold CV, with selected markers confusion matrix and ROC curve results from Cubic SVM**

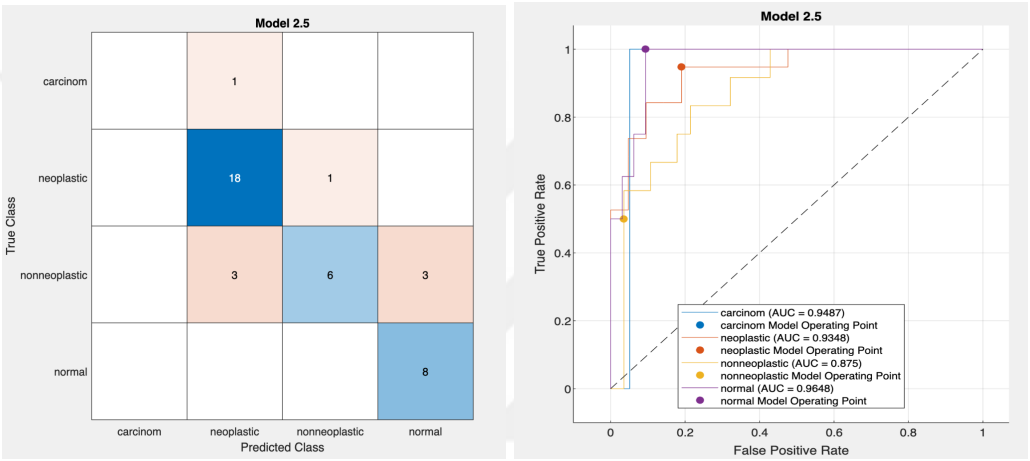



**Figure 5.25 Test Train ratio 10%-90%, 5-fold CV, with all markers confusion matrix and ROC curve results from Efficient Linear SVM**

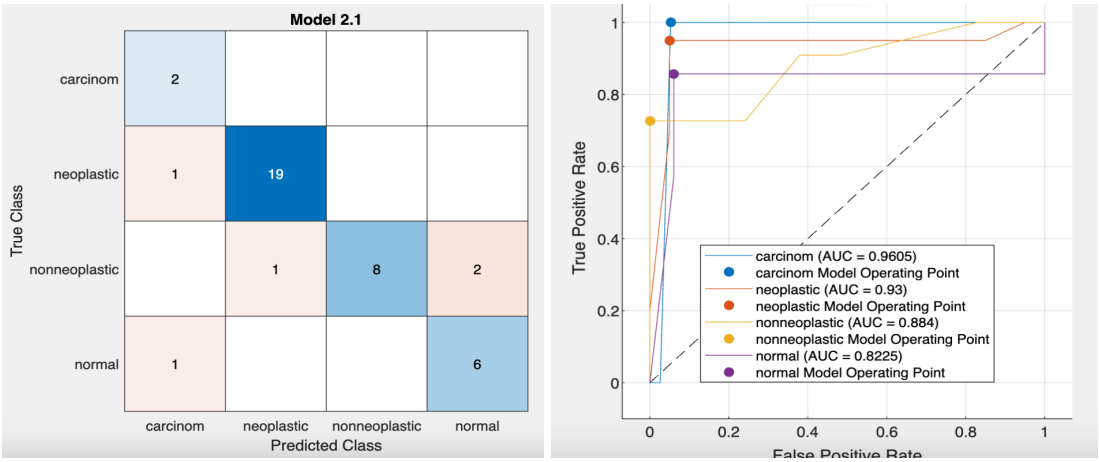



**Figure 5.26 Test Train ratio 10%-90%, 5-fold CV, with selected markers confusion matrix and ROC curve results from the fine tree**

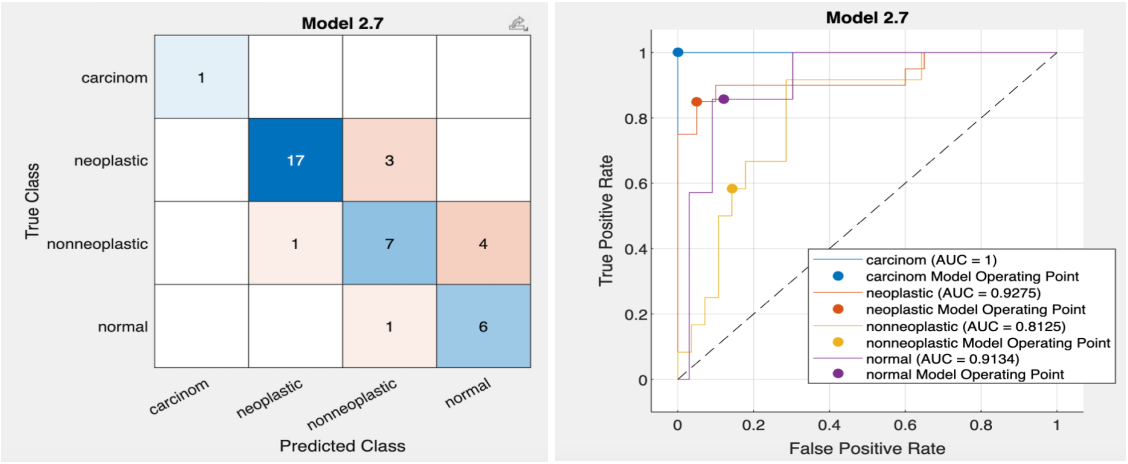



**Figure 5.27 Test Train ratio 10%-90%, 10-fold CV, with all markers confusion matrix and ROC curve results from kernel naïve bayes**

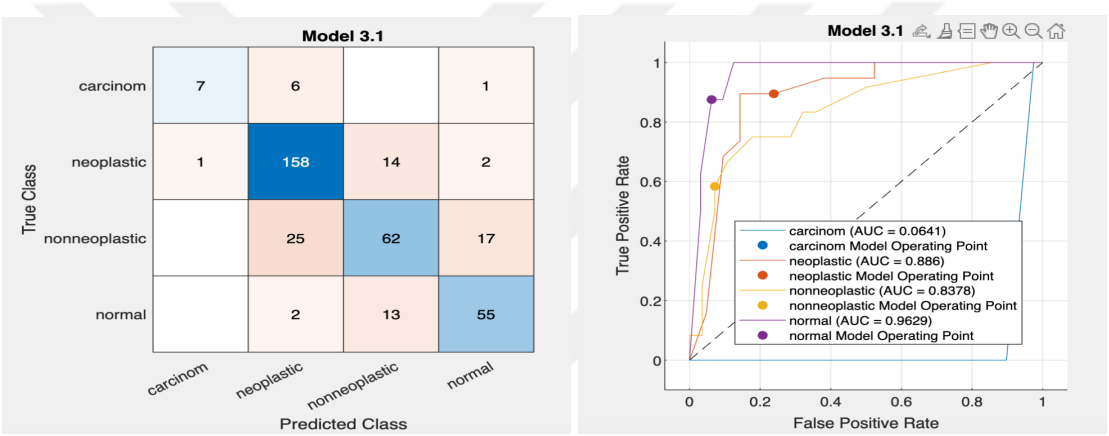



**Figure 5.28 Test Train ratio 10%-90%, 10-fold CV, with selected markers confusion matrix and ROC curve results from the fine tree**

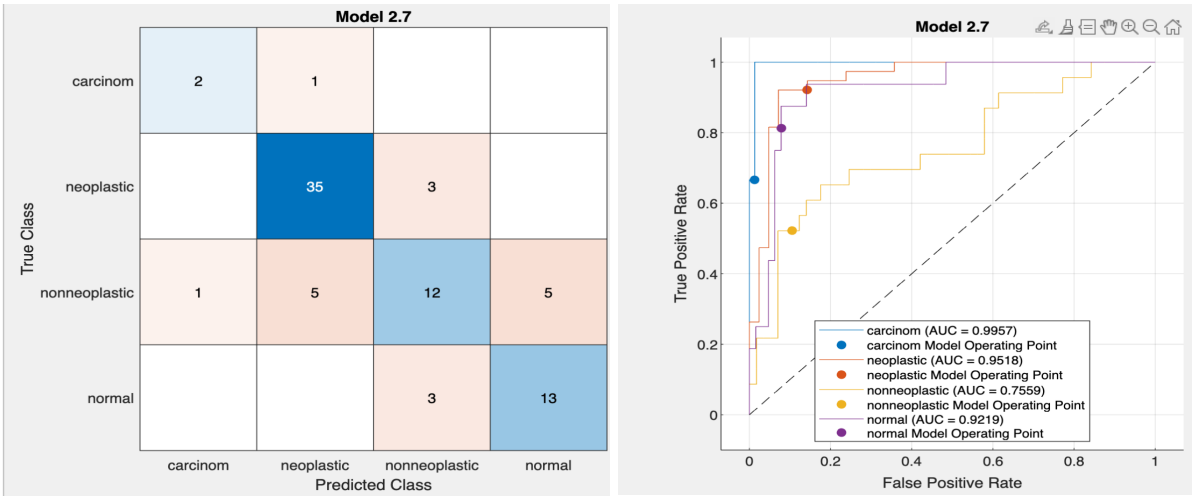



**Figure 5.29 Test Train ratio 20%-80%, 10-fold CV, with all markers confusion matrix and ROC curve results from the fine tree**

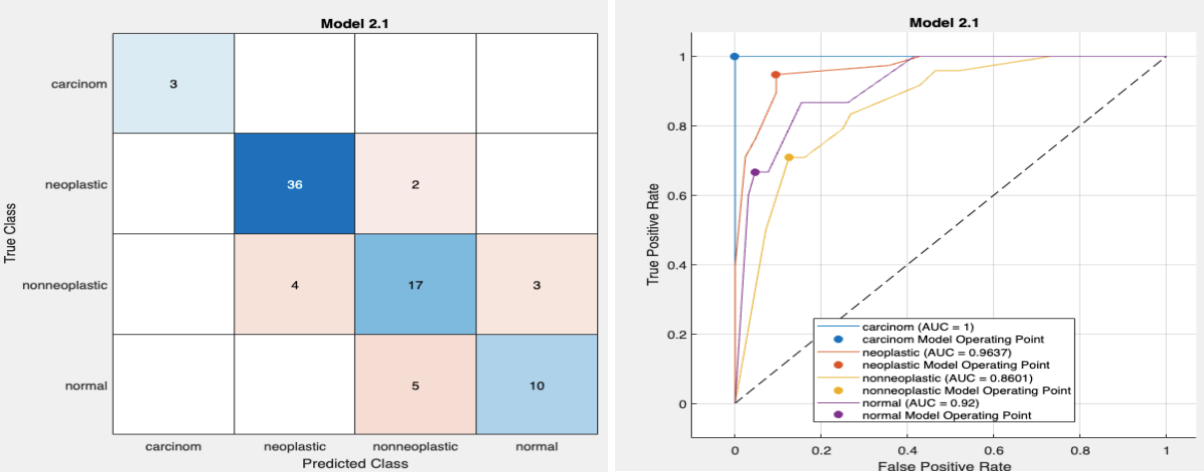



**Figure 5.30 Test Train ratio 20%-80%, 10-fold CV, with selected markers confusion matrix and ROC curve results from decision tree**

These findings suggest that markers, especially Ki-67 (clone 30-9) and p53 (clone bp53-11), play a critical role in colon cancer classification and that accurate selection of such markers can significantly improve classification performance. In addition, the importance of the contribution of demographic and anatomical information to the classification process is emphasized by this study. In conclusion, this study provides a detailed evaluation of the selection of markers and the performance of classification algorithms used in diagnosing colon cancer. This type of analysis can be considered an important step in the development of early diagnosis and treatment strategies.

**Table 5.9 Mean ± standard deviation values markers by types/subtypes**

| Molecular Markers | Ki-67 (clone 30-9) | VEGF (clone SP125) | CD34 (clone QBend/10) | p53 (clone bp53-11) |
|---|---|---|---|---|
| Main types | | | | |
| Carcinoma (#16) | 85.62±27.82 | 70.77±35.41 | 63.15±19.19 | 64.69±43.16 |
| Neoplastic (#195) | 51.95 ± 18.46 | 41.18±25.68 | 46.88±16.09 | 27.89±16.20 |
| Nonneoplastic(#116) | 22.17 ± 15.92 | 26.31±27.10 | 45.24±15.68 | 8.78 ± 8.08 |
| Normal (#78) | 5.01 ± 4.00 | 14.17±24.26 | 35.40±13.37 | 1.75±1.36 |
| Subtypes | | | | |
| Carcinoma (#16) | 85.62±27.82 | 70.77±35.41 | 63.15±19.19 | 64.69±43.16 |
| Villous (#9) | 65.56±13.33 | 67.67±28.04 | 51.33±21.66 | 47.78±17.70 |
| Tubulovillous (#31) | 61.00±18.31 | 52.50±24.87 | 54.77±19.87 | 35.00±17.37 |
| Tubular (#154) | 49.29±17.89 | 37.27±24.18 | 45.01±14.38 | 25.24±14.62 |
| Hyperplastic (#82) | 25.38±16.39 | 31.70±26.77 | 45.33±14.65 | 10.34±8.56 |
| Inflamatory (#21) | 16.00±11.64 | 15.11±26.66 | 50.63±18.11 | 5.53±4.98 |
| Hamartomatous (#4) | 11.25±6.29 | 10.00±8.16 | 48.25±4.27 | 3.00±2.31 |
| Normal (#78) | 5.01±4.00 | 14.17±24.26 | 35.40±13.37 | 1.75±1.36 |

Table 5.9 presents the mean ± standard deviation values of molecular markers in various species and subspecies. Markers include Ki-67 (clone 30-9), VEGF (clone SP125), CD34 (clone QBend/10), and p53 (clone bp53-11).

Main Types: In the carcinoma group, all markers have high values. Particularly high mean values (85.62 and 70.77, respectively) are observed in Ki-67 and VEGF. Neoplastic has intermediate mean values, the mean values of Ki-67 and VEGF are lower than in carcinoma (51.95 and 41.18, respectively). Nonneoplastic group, the mean values of the markers are lower, especially the p53 marker (8.78) is remarkably low. Normal has the lowest mean values for all markers, indicating that these markers are present at low levels in normal tissues.

Subtypes: Carcinoma, Villous, Tubulovillous, Tubular: In these subtypes, the values of the markers appear in decreasing order. There are highest values in carcinoma and lower values in tubules. Hyperplastic, Inflammatory, and Hamartomatous: In these three subtypes, the values of the markers are generally lower. Especially in the hamartomatous group, the lowest mean values for all markers stand out. Normal: Again, it has the lowest mean values for all markers, indicating low levels of these markers in healthy tissues.

Table 5.9 compares the average levels of molecular markers in tissue samples from different types and subtypes and reveals the differences in these markers between diseased and healthy tissues. These data can play an important role in medical diagnosis and research.

# Chapter 6

# Conclusions

## 6.1 Colonoscopy Based Studies

### 6.1.1 Study 1: Comparison Manual and Automatic Feature Extraction Based Colon Polyp Classification

In the binary classification task, the AlexNet model achieved the highest accuracy (68%), precision (77%), and F-score (70%) with original dimensional inputs. This shows that AlexNet is more successful than other models in detecting the presence of polyps. On the other hand, ResNet-50 showed the lowest performance in this task. In the multi-class classification task, the performance of all models decreased significantly. This shows that multiclass classification is a more difficult task than binary classification. AlexNet and GoogleNet showed similar performance on this task.

Polyp cropping generally reduced the performance of the models. This suggests that cropping may lead to the loss of some important information. However, the ResNet-18 model performed better in multi-class classification with trimmed inputs, indicating the sensitivity of the model to the input size. This study compares how different deep learning models perform in polyp classification. AlexNet gave better results in binary classification, and generally, AlexNet and GoogleNet gave better results in multi-class classification. It has been observed that input size and cropping have significant effects on model performance. These findings should be taken into account in model selection and data preparation in future studies. Determining the most appropriate model and data processing technique for polyp classification should continue to be the focus of research in this field.

According to Table 6.1, polyp characterization studies in the literature, Tischendorf and colleagues [120] achieved an accuracy rate of 85.3% on a dataset of 209 polyps using augmentation NBI. This study highlights the potential of narrowband imaging technology in differentiating neoplastic and nonneoplastic polyps. Gross and colleagues [121] achieved an impressive accuracy rate of 93.1% using a similar technique and a larger dataset (434 polyps).

**Table 6.1 Polyp characterization studies using colonoscopy images/videos**

| Classification | Input Type | Lesion Type | Accuracy | Sensitivity | Specificity | Study |
|---|---|---|---|---|---|---|
| Binary | 209 polyps/ 128 patients | 160 neoplastic 49 nonneoplastic | 85.3% | 90% | 70.2% | [120] |
| Binary | 434 polyps/ 214 patients | 258 neoplastic 176 nonneoplastic | 93.1% | 95.0% | 90.3% | [121] |
| Binary | 118 lesions | 73 neoplastic 45 nonneoplastic | 94.9% | 95.9% | 93.3% | [122] |
| Multi | 76 videos | 41 adenomas 21 hyperplastic 15 serrated | 82.5% | 72.7% | 85.9% | [123] |
| Binary | 100 images | 50 neoplastic 50 nonneoplastic | 90.0% | 84.5% | 97.6% | [124] |
| Binary | 205 polyps/ 123 patients | 147 neoplastic 58 nonneoplastic | 89% | 89% | 88% | [125] |

Takemura et al. [126] achieved a remarkable accuracy of 98.5% using chromo-endoscopy with magnification. In another study by Kominami et al [122], 94.9% accuracy was achieved on 118 colorectal lesions using magnification NBI technology. Mesejo et al. [123] studied 76 videos using white light and NBI and obtained an accuracy of 82.5%. Misawa et al. [124] and Mori et al. [125] achieved accuracy rates of 90.0% and 89.0% using NBI with endocytosis and endocytosis only techniques, respectively.

### 6.1.2 Study 2: Hyperplastic and Tubular Polyp Classification

In this study, we evaluated the performance of different data preparation methods and classification models for the classification of gastroenterology images. During data preparation, we processed our original images in different ways: Raw, Crop, Back, RawPrep, CropPrep and BackPrep. These processing methods include adding background to polyp images, cropping polyp, and preprocessing. We also implemented the feature selection steps. Four different feature selection methods were used: RFE, SelectKBest, SelectFromModel, and Boruta. These methods were used to identify the most important features in the dataset and improve model performance. For model training, 10 different classification models were used: LR, RF, Gaussian Naive Bayes, DT, KNN, SVM, XGBoost, Neural Network, Gradient Boosting, and LightGBM. Each model was trained separately for each dataset and feature selection method and the AUC value was used to measure performance. The 5-fold cross-validation procedure was used to improve the reliability of the results for each model and the

AUC value was calculated in each experiment, repeated 20 times. In this way, statistically significant results were obtained. Paired t-test was used to compare model performance. This analysis allowed us to evaluate the statistical difference between the performance of both models. If the difference was statistically significant, it was determined which model performed better.

This study uses various classification methods on colon polyp subtype classification data to distinguish tubular and hyperplastic polyps. However, there are aspects of the system that could be improved. By increasing the number of input images of the dataset and features, better performance can be achieved when testing classification methods. However, this study shows that even features obtained from this number of samples and features reached 90% AUC after feature selection and optimum parameters were selected. In the clinically the most challenging classification types are the hyperplastic and tubular adenomas. Considering the results, some general evaluations can be made. The Random Forest generally appears to have the highest AUC scores, indicating that this model performs best overall on the various feature selection methods used. However, the performance of the Random Forest model may vary depending on the feature selection method and dataset. For example, the Random Forest model generally performed better on the RawPrep, CropPrep, and Crop datasets and with the RFE and Boruta feature selection methods, while it had lower AUC scores with the SelectFromModel. The Neural Network model has the lowest overall AUC scores, indicating that it generally performs the worst over the various feature selection methods used. However, the low performance of this model could be attributed to inappropriately adjusted model parameters or the model's capacity to model more complicated feature dependencies being limited by the dataset and features employed. When it comes to feature selection approaches, RFE typically has the greatest AUC scores, while SelectFromModel typically has the lowest. However, while RFE normally outperforms, this can vary depending on the model and dataset. The performance of a model can vary greatly depending on the dataset employed. In the Crop dataset, for example, the Random Forest and LightGBM models have better AUC scores, whereas the SVM and Neural Network models have lower AUC values. Overall, Random Forest, XGBoost, Gradient Boosting, and LightGBM are among the best-performing models; the capacity of these models to describe complicated feature dependencies may explain why they frequently have better AUC scores than others. This study contributes to the comprehension of the significance of feature selection methods and classifiers in the context of polyp classification tasks. Additionally, it offers potential direction for future researchers and developers involved in clinical applications. The findings of our study underscore the practicality of the suggested

model as a valuable instrument for differentiating between various types of polyps, hence demonstrating its potential for future implementation in clinical settings. This study has the potential to facilitate future investigations in the field of polyp categorization and make valuable contributions toward enhancing medical diagnostic procedures in this domain. As a result of these analyses, various performance differences in data preparation procedures, feature selection approaches, and classification models were discovered. These results and comprehensive information could enlighten the literature and other research on this topic.

**Table 6.2 Method, dataset, and performance metric results**

| Methods | Dataset | Sensitivity | Specificity | Study |
|---|---|---|---|---|
| CNN | 11802 image patches<br>35 videos | 86% | 85% | [127] |
| Window Median Depth of Valleys Energy Maps | 24 videos / 31 polyps | 70.4% | 72.4% | [128] |
| Hybrid Context-Shape | Internal dataset: 10 videos<br>External dataset: 15 videos | 48% (internal)<br>88% (external) | - | [25] |

What these studies collectively show is how effective advanced imaging techniques and computer-aided diagnostic systems are in identifying and classifying colon polyps. Additionally, the potential of these technologies to improve discrimination between different types of polyps is considered a significant advance in colonoscopy practices. These studies collectively highlight the advances of learning-based algorithms in classifying colon polyps, the importance of dataset selection and preprocessing, and the development of models and comparative tests for better diagnostic accuracy in colonoscopy. Thus far, there have been very few learning studies focused on the most challenging characterization of hyperplastic and tubular adenoma specimens. Therefore, in this study, we focused on this task, which is the most difficult to distinguish clinically and is not focused on in the literature, by creating different input datasets for feature selection and machine learning. We created a holistic dataset by taking videos and detailed pathology reports of patients who underwent colonoscopy biopsy. We aimed to develop and validate the best-performing model by combining different feature extraction, feature selection, and machine learning methods.

This study aims to develop an effective approach to distinguish hyperplastic and tubular adenoma colon polyps, which is one of the most challenging tasks in colonoscopy procedures. In this context, it is important to discuss the existing literature and the innovations taken by this study. The study [72] evaluated the histological and endoscopic features of colorectal polyps

and showed that narrow-band imaging (NBI) technology is effective in real-time prediction of polyp histology. However, this study did not specifically focus on the classification of hyperplastic and tubular adenoma polyps. Another study [129] developed a new classification using blue light imaging (BLI) to distinguish polyps. This study identified several descriptors for polyp characterization and achieved high interobserver agreement but did not focus on identifying differences between hyperplastic and tubular adenoma polyps. Zhang et al. [130] aimed to automatically detect and classify hyperplastic and adenomatous polyps using deep learning and transfer learning methods. This study showed that the algorithms achieved higher accuracy, precision, and recall rates compared to endoscopist observations. However, this work did not focus on specialized datasets and feature selection methods. The study [131] developed an artificial intelligence model that can distinguish adenomatous and non-adenomatous polyps in real-time. This model achieved high accuracy rates by analyzing standard colonoscopy videos. This study was an important step towards the applicability of the “resect and discard' paradigm but did not examine in detail the impact of feature selection and data preprocessing methods on classification.

In this study, we highlight the importance of dataset selection, data preprocessing, and feature selection in classifying hyperplastic and tubular adenoma polyps. We state that the Random Forest model and the Recursive Feature Elimination (RFE) feature selection approach are superior to other models and reach the highest AUC value. This approach exceeded the performance of a gastroenterologist in terms of F-score, recall, and AUC as well as accuracy. This demonstrates the importance of dataset selection and feature selection in classifying hyperplastic and tubular adenoma polyps. This study combines different lines of research, integrating advanced data preprocessing and classification algorithms into the colonoscopy process.

In conclusion, the application of machine learning methods, especially the combination of feature selection with Random Forest models and Recursive Feature Elimination (RFE), has shown significant potential in precisely categorizing tubular adenoma and hyperplastic polyps during colonoscopy procedures. The performance of this approach has surpassed traditional methods with high accuracy rates and area under curve (AUC) values. These results highlight the importance of careful feature selection, data preprocessing, and dataset selection in improving the effectiveness of classification models. In addition to improving the existing knowledge base in the fields of medical imaging and computational pathology, this development is practically applicable and has the potential to improve diagnostic accuracy and patient outcomes in the field of gastroenterology. However, the successful outcome of this

approach does not diminish the importance of a gastroenterologist's expertise; On the contrary, it strengthens it by providing a valuable tool to assist in complex diagnostic decisions. In the field of gastroenterology, this line of inquiry could lead to more sophisticated diagnostic methods powered by artificial intelligence, thereby transforming patient care and early cancer detection.

## 6.2 Histopathology Based Studies

### 6.2.1 Study 3: Binary and Multi-Class Polyp Classification Using Histopathology Images

In this study, the performances of the ResNet-18 and AlexNet models for binary and multiple classification of histopathological images were evaluated at different magnification rates. While binary classification distinguishes neoplastic and nonneoplastic lesions, multiple classification aims to categorize more specific types of polyps (tubular, villous, tubulovillous, hyperplastic, inflammatory, and hamartomatous). The results showed that model selection and training had a significant impact on performance, especially at low magnifications and in binary classification scenarios. AlexNet demonstrated superior performance with a high recall rate in binary classification at 2,5X magnification. However, it was determined that this success was not observed in multiple classifications. Optimizing deep learning models used in histopathological image analysis for various magnification rates and classification challenges is crucial, as highlighted by this study. In addition, the advantages and limitations of each model should be examined in detail, and this information should be a guide for the advancement of artificial intelligence applications in the field of healthcare.

## 6.3 IHC Based Studies

### 6.3.1 Study 4: Biomarker Selection and Performance Analysis in Classification Model

In this study, the effectiveness of various features used in classifying colon cancer was evaluated using a wide range of classifiers. The results of the study revealed that certain classifiers performed better with certain feature sets. In particular, when using 11 features, Linear SVM and Kernel Naive Bayes classifiers achieved high accuracy rates with both 5-fold and 10-fold CV methods. On the other hand, in the case where only two key features (Ki-67

(clone 30-9) and p53 (clone bp53-11)) are used, Cubic SVM and NN, especially NN, stand out with a remarkable accuracy rate of 87.5%.

These results show that the correct selection of features used in cancer classification and the performance of classification algorithms are critical in the development of diagnostic and treatment strategies. The findings of this study suggest an important direction for future research. First, testing these features and classifiers on larger patient populations is important for the general validity of the results obtained. Secondly, the effects of markers such as Ki-67 (clone 30-9) and p53 (clone bp53-11) on cancer classification and prognosis can be examined in more detail. This is an important step in developing biomarker-based classification systems and supporting personalized treatment approaches.

Additionally, the development of artificial intelligence and machine learning algorithms can contribute to further improving cancer classification by allowing the analysis of more complex datasets and interactions. This may enable early diagnosis as well as a better understanding of the course of the disease and response to treatment. Finally, the methodology and results of this study can be applied to other types of cancer and contribute to knowledge in these areas. This comprehensive approach will contribute to the development of a holistic and integrated perspective in cancer research.

## 6.4 Societal Impact and Contribution to Global Sustainability

**SDG 3: "Good Health and Well-Being"**

The outcomes of this thesis are directly aligned with the SDG 3 goal, defined as "ensuring healthy living and well-being for all ages." Developing early diagnosis and treatment methods is aimed at reducing the risk factors of non-contagious diseases and improving the general health status. This study contributed to protecting patient health and improving quality of life, especially through the use of artificial intelligence and computer vision technologies in the early diagnosis of serious health problems such as colon cancer.

**SDG 9: "Industry, Innovation, and Infrastructure"**

The thesis also contributes to the SDG 9 goal by promoting innovation and the use of high technology in the field of health technologies. Developed artificial intelligence-based systems

increase the quality and accessibility of healthcare services by supporting digital transformation in medical diagnosis and treatment processes.

**SDG 17: "Achieving Goals Through Partnerships"**

This thesis demonstrates a model aligned with SDG 17 by promoting collaboration between academic institutions, healthcare, and research sectors. This collaboration between various disciplines and sectors plays an important role in achieving sustainable development goals.

These outcomes show that the thesis is more than just a scientific success; it is a study that makes concrete contributions to global health and sustainable development goals. The thesis has taken important steps towards the development of early diagnosis and treatment methods, increasing the quality of health services, and strengthening interdisciplinary collaborations.

# Chapter 7

# Future Prospects

## 7.1 Colonoscopy-Based Studies

Artificial intelligence and machine learning technologies will shape the future of colonoscopy-based studies. Advanced image processing algorithms and deep learning models can provide higher accuracy and precision in polyp detection and classification. Artificial intelligence-supported colonoscopy will have an important role, especially in the diagnosis of early-stage colon cancer. In addition, with real-time data analysis, it will be possible to instantly evaluate polyps and make treatment recommendations. The future perspective of this study focuses on the importance of the development of artificial intelligence-based colonoscopy image classification systems. The study evaluated the polyp classification performance of ResNet-18 and AlexNet convolutional neural networks at different magnifications. The results show that these models can offer high accuracy and precision in some cases. In particular, the ResNet-18 model generally exhibits a more balanced performance, while the AlexNet model stands out with high sensitivity values in some cases.

Going forward, the development and improvement of such models may play an important role in accelerating and standardizing histopathological examinations. Furthermore, it is crucial to emphasize that the selection of the model and the quality of the training dataset determine the success of classification. In this context, training models using more diverse and comprehensive datasets can improve classification performance.

The low multi-classification performance shows that this area is more complex, and the models have difficulty discriminating between these tasks. Future work may focus on improving model architectures and training techniques to improve multiclassification accuracy. In particular, since polyp subtypes are very similar to each other, it is important to develop algorithms that can extract more sensitive and detailed features in this field.

Finally, further examination may provide more precise guidance on which magnification ratio to prefer in different situations by assessing the impact of different magnification ratios on model performance. This can be especially critical when analyzing complex and heterogeneous

tissues. Such detailed analyses can make significant contributions to the future development of artificial intelligence-based histopathological image classification systems.

## 7.2 Histopathology-Based Studies

- **More Advanced Deep Learning Models:**

Analysis of the performance of existing models such as ResNet-18 and AlexNet demonstrates the potential benefits of more complex deep learning architectures and training strategies. For instance, researchers can create more complex network structures with GANs (Generative Adversarial Networks), capsule networks, or attention mechanisms to help classify polyps. These models can help researchers gain a deeper understanding of tissue heterogeneity and make more accurate distinctions between similar subtypes [132].

- **Multiple Display Modes and Fusion Techniques:**

Some classification tasks have shown higher magnifications to be more effective. This could lead to the development of multimodal image fusion techniques that combine information from images obtained at different magnifications. These techniques can improve classification performance by enabling the integration of important features at different resolutions [133].

- **Integration with IHC and Colonoscopy Data:**

Integrating IHC and colonoscopy findings with histopathological images may further improve polyp classification. For example, a unified model can utilize the synergistic information of biomarker profiles obtained from IHC and morphological features in colonoscopy images. This integration may provide a more comprehensive perspective from both diagnostic and prognostic perspectives [134].

- **Personalized Treatment Approaches:**

Improving existing models may contribute to personalized treatment planning. For example, deep learning models can analyze the treatment response profiles of different polyp subtypes. Deep learning models can utilize the patient's own histological and molecular data [135] to determine more targeted treatment options.

- **Real-Time Classification and Diagnosis:**

Integrating the developed models into real-time classification and diagnostic applications can support the decision-making processes of pathologists and clinicians. This can provide significant time and cost savings, especially in busy pathology laboratories [136].

- **Big Data and Transfer Learning:**

Applying information obtained from large histopathological data sets to new models through transfer learning can enable the training of robust models on larger and more diverse data sets. This may increase the general applicability of the model and its adaptation to different populations [137].

These future expectations suggest that researchers should constantly improve existing models and approaches and that the field of pathology will increasingly rely on artificial intelligence and machine learning.

## 7.3 IHC Based Studies

The future of IHC-based studies, particularly in cancer research, appears to hold great promise but also presents complex challenges. Scientists expect that the discovery of more accurate and sensitive biomarkers will substantially impact the course of these studies, leading to a more comprehensive and inclusive approach to cancer research. Scientists aim to develop personalized strategies for cancer diagnosis and treatment by leveraging various combinations of biomarkers to improve efficacy. Personalization is crucial, especially considering the diverse characteristics of cancerous growths and their different responses to treatment.

The future study in this sector will heavily rely on the enhanced utilization of IHC to accurately diagnose and develop treatment options in the early stages of diseases. Prompt identification is particularly vital in improving the survival rates and general welfare of patients with cancer. The timely incorporation of IHC into the diagnostic procedure might significantly impact the efficacy of treatment.

The inquiry into the classification of colon cancer using several biomarkers and classifiers is indicative of the ongoing advancements in the field. Some biomarkers, such as Ki-67 (clone 30-9) and p53 (clone bp53-11), can make diagnostics more accurate. Classification models with SVM, NB, and NN greatly enhance the accuracy of diagnostics when utilizing biomarkers such as Ki-67 (clone 30-9) and p53 (clone bp53-11).

Future investigations will focus on expanding these findings. Testing the features and classifiers on larger patient groups is crucial to evaluating the general applicability of the results. We should also carefully look at how important markers like Ki-67 and p53 affect the classification and prognosis of cancer. This will help in the creation of biomarker-based classification systems. These technologies possess the capacity to significantly improve

personalized therapy approaches, thereby enhancing treatment effectiveness and reducing the likelihood of adverse effects.

Another topic with potential for future research involves integrating AI and machine learning algorithms in the classification of cancer. These technologies can analyze datasets of higher complexity and understand the interactions within them. This has the potential to lead to early detection of illnesses, a more thorough understanding of disease progression, and improved effectiveness in treatment outcomes.

Applying the ongoing study's strategy and findings to various types of cancer broadens the scope of exploration and deepens our comprehensive understanding of cancer as a disease. An upcoming trend in cancer research is the development of a comprehensive and unified approach that takes into account various types of cancer and their unique characteristics.

The future of IHC-based studies in cancer research centers on advancing more precise, personalized, and streamlined methods for identifying and managing cancer. Integrating state-of-the-art technologies, performing more extensive investigations, and adopting a complete research methodology are expected to drive anticipated advancements in the discipline.

# BIBLIOGRAPHY


[1] D. A. Corley *et al.*, "Adenoma Detection Rate and Risk of Colorectal Cancer and Death," *http://dx.doi.org/10.1056/NEJMoa1309086*, vol. 370, no. 14, pp. 1298–1306, Apr. 2014, doi: 10.1056/NEJMOA1309086.

[2] J. Fernando *et al.*, "Influence of Sex, Age and Body Mass Index on Healthy Colon Anatomy," *Actual Med*, vol. 107, no. 815, pp. 18–26, 2022, doi: 10.15568/am.2022.815.or02.

[3] B. Shen, "Colon: Anatomy and Structural Anomalies," *Yamada's Textbook of Gastroenterology*, pp. 93–113, Apr. 2022, doi: 10.1002/9781119600206.CH6.

[4] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer Statistics, 2023," *CA Cancer J Clin*, vol. 73, no. 1, pp. 17–48, Jan. 2023, doi: 10.3322/CAAC.21763.

[5] "Colorectal cancer." Accessed: Nov. 28, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/colorectal-cancer

[6] "Advances in Colorectal Cancer Research - NCI." Accessed: Nov. 28, 2023. [Online]. Available: https://www.cancer.gov/types/colorectal/research

[7] K. D. Miller *et al.*, "Cancer Treatment and Survivorship Statistics, 2022," *CA Cancer J Clin*, vol. 72, no. 5, pp. 409–436, Sep. 2022, doi: 10.3322/CAAC.21731.

[8] R. Butrum, "American Institute for Cancer Research," *Neoplasia*, vol. 2, no. 6. pp. 569–570, 2000. doi: 10.1038/sj.neo.7900109.

[9] W. Chen, D. Jones, and W. L. Frankel, "Molecular Pathology of Colorectal Tumors," *Molecular Surgical Pathology*, pp. 133–154, 2023, doi: 10.1007/978-3-031-35118-1_5.

[10] A. Leslie, F. A. Carey, N. R. Pratt, and R. J. C. Steele, "The colorectal adenoma-carcinoma sequence," *British Journal of Surgery*, vol. 89, no. 7, pp. 845–860, 2002, doi: 10.1046/j.1365-2168.2002.02120.x.

[11] P. M. Colucci, S. H. Yale, and C. J. Rall, "Colorectal Polyps," *Clin Med Res*, vol. 1, no. 3, pp. 261–262, 2003, doi: 10.3121/cmr.1.3.261.

[12] N. L. Bertelson *et al.*, "Colectomy for Endoscopically Unresectable Polyps," *Dis Colon Rectum*, vol. 55, no. 11, pp. 1111–1116, 2012, doi: 10.1097/dcr.0b013e3182695115.

[13] N. Shussman and S. D. Wexner, "Colorectal Polyps and Polyposis Syndromes," *Gastroenterol Rep (Oxf)*, vol. 2, no. 1, p. 1, Feb. 2014, doi: 10.1093/GASTRO/GOT041.

[14] J. N. Kather *et al.*, "Predicting Survival from Colorectal Cancer Histology Slides using Deep Learning: A retrospective Multicenter Study," *PLoS Med*, vol. 16, no. 1, pp. 1–22, 2019, doi: 10.1371/journal.pmed.1002730.

[15] G. Shobha and S. Rangaswamy, "Machine Learning," *Handbook of Statistics*, vol. 38, pp. 197–228, Jan. 2018, doi: 10.1016/bs.host.2018.07.004.

[16] A. Picon *et al.*, "Novel Pixelwise Co-Registered Hematoxylin-Eosin and Multiphoton Microscopy Image Dataset for Human Colon Lesion Diagnosis," *J Pathol Inform*, vol. 13, p. 100012, 2022, doi: 10.1016/J.JPI.2022.100012.

[17] M. K. Gangwani, A. Aziz, D. S. Dahiya, M. Nawras, M. Aziz, and S. Inamdar, "History of Colonoscopy and Technological Advances: A Narrative Review," *Transl Gastroenterol Hepatol*, vol. 8, no. 0, p. 18, Apr. 2023, doi: 10.21037/TGH-23-4/COIF.

[18] Y. Mori *et al.*, "Computer-Aided Diagnosis for Colonoscopy What Can We Expect of CAD in Clinical Automated Detection of Colorectal Polyps," *Endoscopy*, vol. 49, pp. 813–819, 2017.

[19] Y. Yuan and M. Q. H. Meng, "Deep Learning for Polyp Recognition in Wireless Capsule Endoscopy Images," *Med Phys*, vol. 44, no. 4, pp. 1379–1389, 2017, doi: 10.1002/mp.12147.

[20] A. Le, M. Salifu, and I. McFarlane, “Artificial Intelligence in Colorectal Polyp Detection and Characterization,” *Int J Clin Res Trials*, vol. 6, no. 1, Mar. 2021, doi: 10.15344/2456-8007/2021/157.

[21] S. M. Krishnan, X. Yang, K. L. Chan, S. Kumar, and P. M. Y. Goh, “Intestinal Abnormality Detection from Endoscopic Images,” vol. 20, no. 2, pp. 895–898, 1998, doi: 10.1109/iembs.1998.745583.

[22] S. M. Krishnan, C. S. Tan, and K. L. Chan, “Closed-Boundary Extraction of Large Intestinal Lumen,” *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*, vol. 16, no. pt 1, pp. 610–611, 1994, doi: 10.1109/iembs.1994.411878.

[23] D. K. Iakovidis, D. E. Maroulis, and S. A. Karkanis, “An Intelligent System for Automatic Detection of Gastrointestinal Adenomas in Video Endoscopy,” *Comput Biol Med*, vol. 36, no. 10, pp. 1084–1103, 2006, doi: 10.1016/j.compbiomed.2005.09.008.

[24] P. Wang, S. M. Krishnan, C. Kugean, and M. P. Tjoa, “Classification of endoscopic images based on texture and neural network,” *Annual Reports of the Research Reactor Institute, Kyoto University*, vol. 4, pp. 3691–3695, 2001, doi: 10.1109/iembs.2001.1019637.

[25] N. Tajbakhsh, S. R. Gurudu, and J. Liang, “Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information,” *IEEE Trans Med Imaging*, vol. 35, no. 2, pp. 630–644, Feb. 2016, doi: 10.1109/TMI.2015.2487997.

[26] G. Urban *et al.*, “Deep Learning Localizes and Identifies Polyps in Real Time With 96% Accuracy in Screening Colonoscopy,” *Gastroenterology*, vol. 155, no. 4, pp. 1069–1078, 2018, doi: 10.1053/j.gastro.2018.06.037.

[27] F. Pace *et al.*, “Artificial neural networks are able to recognize gastro-oesophageal reflux disease patients solely on the basis of clinical data,” *Eur J Gastroenterol Hepatol*, vol. 17, no. 6, pp. 605–610, 2005, doi: 10.1097/00042737-200506000-00003.

[28] M. F. Byrne *et al.*, “Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model,” *Gut*, vol. 68, no. 1, pp. 94–100, 2019, doi: 10.1136/gutjnl-2017-314547.

[29] M. Ganz, X. Yang, and G. Slabaugh, “Automatic segmentation of polyps in colonoscopic narrow-band imaging data,” *IEEE Trans Biomed Eng*, vol. 59, no. 8, pp. 2144–2151, 2012, doi: 10.1109/TBME.2012.2195314.

[30] T. Tamaki *et al.*, “Computer-aided colorectal tumor classification in NBI endoscopy: Using local features,” *Med Image Anal*, vol. 17, no. 1, pp. 78–100, 2013, doi: 10.1016/j.media.2012.08.003.

[31] M. Häfner *et al.*, “Computer-aided classification of zoom-endoscopical images using Fourier filters,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 4, pp. 958–970, 2010, doi: 10.1109/TITB.2010.2044184.

[32] S. Tetsuji Takayama, M. Katsuki, Yasuo Takahashi, K. Ohi, Shuichi Nojiri, Sumio Sakamaki, Junji Kato, and Y. N. Kogawa, Hirotsugu Miyake, “Aberrant Crypt Foci of The Colon as Precursors of Adenoma and Cancer,” *N Engl J Med*, vol. 339, no. 18, pp. 1277–1284, 1998.

[33] K. Gono *et al.*, “Appearance of enhanced tissue features in narrow-band endoscopic imaging,” *J Biomed Opt*, vol. 9, no. 3, p. 568, 2004, doi: 10.1117/1.1695563.

[34] H. Dg, T. Kaltenbach, and Y. Sano, “Validation of a simple classification system for endoscopic diagnosis of small colorectal polyps using narrow-band imaging,” *Gastroenterological Endoscopy*, vol. 54, no. 11, p. 3642, 2012, doi: 10.11280/gee.54.3642.

[35] T. Stehle *et al.*, “Classification of colon polyps in NBI endoscopy using vascularization features,” *Medical Imaging 2009: Computer-Aided Diagnosis*, vol. 7260, pp. 1–13, 2009, doi: 10.1117/12.808103.
[36] S. Gross *et al.*, “Polyp segmentation in NBI colonoscopy,” *Informatik aktuell*, pp. 252–256, 2009.
[37] D. K. Rex, “Narrow-Band Imaging Without Optical Magnification for Histologic Analysis of Colorectal Polyps,” *Gastroenterology*, vol. 136, no. 4, pp. 1174–1181, 2009, doi: 10.1053/j.gastro.2008.12.009.
[38] A. Rastogi *et al.*, “Recognition of surface mucosal and vascular patterns of colon polyps by using narrow-band imaging: interobserver and intraobserver agreement and prediction of polyp histology,” *Gastrointest Endosc*, vol. 69, pp. 716–722, 2009, doi: 10.1016/j.gie.2008.09.058.
[39] A. Rastogi *et al.*, “High accuracy of narrow band imaging without magnification for the real-time characterization of polyp histology and its comparison with high-definition white light colonoscopy: A prospective study,” *American Journal of Gastroenterology*, vol. 104, no. 10, pp. 2422–2430, 2009, doi: 10.1038/ajg.2009.403.
[40] S. K. McGill, E. Evangelou, J. P. A. Ioannidis, R. M. Soetikno, and T. Kaltenbach, “Narrow band imaging to differentiate neoplastic and nonneoplastic colorectal polyps in real time: A meta-analysis of diagnostic operating characteristics,” *Gut*, vol. 62, no. 12, pp. 1704–1713, 2013, doi: 10.1136/gutjnl-2012-303965.
[41] N. H *et al.*, “Learning curve of virtual chromoendoscopy for the prediction of hyperplastic and adenomatous colorectal lesions: a prospective 2-center study,” *Gastrointest Endosc*, vol. 78, no. 1, pp. 115–120, Jul. 2013, doi: 10.1016/J.GIE.2013.02.001.
[42] P. Sripatoorat and S. Angchuan, “Feasibility of Fujinon Intelligent Color Enhancement (FICE) with Magnification in Differentiating between Diminutive Adenomatous and Hyperplastic Polyps,” *Medicine*, 2014.
[43] G. Wimmer *et al.*, “Quest for the best endoscopic imaging modality for computer-assisted colonic polyp staging,” *World J Gastroenterol*, vol. 25, no. 10, pp. 1197–1209, 2019, doi: 10.3748/wjg.v25.i10.1197.
[44] C. Hassan, P. J. Pickhardt, and D. K. Rex, “A resect and discard strategy would improve cost-effectiveness of colorectal cancer screening,” *Clinical Gastroenterology and Hepatology*, vol. 8, no. 10, pp. 865–869, 2010, doi: 10.1016/j.cgh.2010.05.018.
[45] A. N. Esgiar, R. N. G. Naguib, B. S. Sharif, M. K. Bennett, and A. Murray, “Microscopic image analysis for quantitative measurement and feature identification of normal and cancerous colonie mucosa,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 2, no. 3, pp. 197–203, 1998.
[46] J. N. Kather *et al.*, “Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer,” *Nat Med*, 2019, doi: 10.1038/s41591-019-0462-y.
[47] F. Yi *et al.*, “Microvessel prediction in H&E Stained Pathology Images using fully convolutional neural networks,” *BMC Bioinformatics*, vol. 19, no. 1. 2018. doi: 10.1186/s12859-018-2055-z.
[48] M. Macenko *et al.*, “A method for normalizing histology slides for quantitative analysis,” *Proceedings - 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2009*, pp. 1107–1110, 2009, doi: 10.1109/ISBI.2009.5193250.
[49] P. Mobadersany *et al.*, “Predicting cancer outcomes from histology and genomics using convolutional networks,” *Proc Natl Acad Sci U S A*, vol. 115, no. 13, pp. 2970–2979, 2018, doi: 10.1073/pnas.1717139115.

[50] J. N. Kather *et al.*, "Multi-class texture analysis in colorectal cancer histology," *Sci Rep*, vol. 6, pp. 1–11, 2016, doi: 10.1038/srep27988.
[51] T. J. Alhindi, S. Kalra, K. H. Ng, A. Afrin, and H. R. Tizhoosh, "Comparing LBP, HOG and Deep Features for Classification of Histopathology Images," *Proceedings of the International Joint Conference on Neural Networks*, vol. 7, pp. 1–7, 2018, doi: 10.1109/IJCNN.2018.8489329.
[52] D. Komura and S. Ishikawa, "Machine Learning Methods for Histopathological Image Analysis," *Comput Struct Biotechnol J*, vol. 16, pp. 34–42, 2018, doi: 10.1016/j.csbj.2018.01.001.
[53] B. Korbar *et al.*, "Deep-Learning for Classification of Colorectal Polyps on Whole-Slide Images," *Med Image Anal*, pp. 1–24, 2017.
[54] E. Budinská *et al.*, "Gene expression-guided selection of histopathology image features," *2015 IEEE 15th International Conference on Bioinformatics and Bioengineering, BIBE 2015*, no. November, 2015, doi: 10.1109/BIBE.2015.7367653.
[55] R. Colen *et al.*, "NCI workshop report: Clinical and computational requirements for correlating imaging phenotypes with genomics signatures," *Transl Oncol*, vol. 7, no. 5, pp. 556–569, 2014, doi: 10.1016/j.tranon.2014.07.007.
[56] V. Popovici *et al.*, "Joint analysis of histopathology image features and gene expression in breast cancer," *BMC Bioinformatics*, vol. 17, no. 1, 2016, doi: 10.1186/s12859-016-1072-z.
[57] N. Eti and İ. Gürses, "Kolon Poliplerinde Adenom - Karsinom Sekansinda Sox- 2 Ve B - Katenin Ekspresyonun Yeri ve Tanıya Katkısının Araştırılması," Mersin Universitesi, 2016.
[58] A. Akyol, "Kolorektal Kanserdeki Biyobelirteçler," *Güncel Gastroenteroloji*, vol. 20, no. 2, pp. 119–129, 2016.
[59] V. Das, J. Kalita, and M. Pal, "ScienceDirect Predictive and prognostic biomarkers in colorectal cancer : A systematic review of recent advances and challenges," *Biomedicine et Pharmacotherapy*, vol. 87, pp. 8–19, 2017, doi: 10.1016/j.biopha.2016.12.064.
[60] T. K. Nikolouzakis *et al.*, "Improving diagnosis, prognosis and prediction by using biomarkers in CRC patients (Review)," *Oncol Rep*, vol. 39, no. 6, pp. 2455–2472, 2018, doi: 10.3892/or.2018.6330.
[61] M. Gezen and B. Erman, "DNA Hasar Yanıtlarında p53 Tümör Baskılayıcı Proteinin Görevi," *Turkiye Klinikleri Immunology Allergy - Special Topics*, vol. 11, no. 3, pp. 17–23, 2018.
[62] A. I. Robles and C. C. Harris, "Clinical outcomes and correlates of TP53 mutations and cancer.," *Cold Spring Harb Perspect Biol*, vol. 2, no. 3, pp. 1–15, 2010.
[63] H. Wang *et al.*, "STAT3-mediated upregulation of lncRNA HOXD-AS1 as a ceRNA facilitates liver cancer metastasis by regulating SOX4," *Mol Cancer*, vol. 16, no. 1, pp. 1–15, 2017, doi: 10.1186/s12943-017-0680-1.
[64] S. H. Aktaş and H. Akbulut, "Kolorektal kanserde anjiyogenez ve anti-anjiyogenik tedaviler," *Turk Onkoloji Dergisi*, vol. 29, no. 2, pp. 67–79, 2014, doi: 10.5505/tjoncol.2014.1066.
[65] H. K. Emine Bozkurtlar, "Akciğer Kanserinde Moleküler Patoloji," *Nükleer Tıp Semineri*, vol. 4, pp. 26–31, 2018, doi: 10.4274/nts.2018.004.
[66] F. C. Santini and M. D. Hellmann, "PD-1/PD-L1 axis in lung cancer," *Cancer J*, vol. 24, no. 1, p. 15, Jan. 2018, doi: 10.1097/PPO.0000000000000300.
[67] M. Mostafa, C. Hartley, A. Mackinnon, and C. Hagen, "Evaluation of PD-L1 and PD-1 in Colorectal Polyps," *Am J Clin Pathol*, vol. 146, Sep. 2016, doi: 10.1093/AJCP/AQW159.076.

[68] K. Zuiderveld, “Contrast Limited Adaptive Histogram Equalization,” in *Graphics Gems*, Academic Press, 1994, pp. 474–485. doi: 10.1016/b978-0-12-336156-1.50061-6.
[69] D. Park, H. Park, D. K. Han, and H. Ko, “Single image dehazing with image entropy and information fidelity,” *2014 IEEE International Conference on Image Processing, ICIP 2014*, pp. 4037–4041, Jan. 2014, doi: 10.1109/ICIP.2014.7025820.
[70] Kaiming He, Jian Sun, and Xiaoou Tang, “Single image haze removal using dark channel prior,” pp. 1956–1963, Mar. 2010, doi: 10.1109/CVPR.2009.5206515.
[71] E. Reinhard, M. Ashikhmin, B. Gooch, and P. Shirley, “Color transfer between images,” *IEEE Comput Graph Appl*, vol. 21, no. 5, pp. 34–41, Sep. 2001, doi: 10.1109/38.946629.
[72] V. A *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.
[73] S. Roy, A. kumar Jain, S. Lal, and J. Kini, “A study about color normalization methods for histopathology images,” *Micron*, vol. 114, pp. 42–61, Nov. 2018, doi: 10.1016/J.MICRON.2018.07.005.
[74] X. Zhao, G. C. Feng, S. H. Ao, and P. L. Liu, “Interrater reliability estimators tested against true interrater reliabilities,” *BMC Med Res Methodol*, vol. 22, no. 1, Dec. 2012, doi: 10.1186/s12874-022-01707-5.
[75] M. L. McHugh, “Interrater reliability: the kappa statistic,” *Biochem Med (Zagreb)*, vol. 22, no. 3, p. 276, 2012, doi: 10.11613/bm.2012.031.
[76] J. R. Landis and G. G. Koch, “The Measurement of Observer Agreement for Categorical Data,” *Biometrics*, vol. 33, no. 1, p. 159, Mar. 1977, doi: 10.2307/2529310.
[77] R. O. Stehling, M. A. Nascimento, A. X. Falcão, and F. Falcão, “A Compact and Efficient Image Retrieval Approach Based on Border/Interior Pixel Classification,” *International Conference on Information and Knowledge Management*, 2002.
[78] P. Stanchev, D. Green, B. Dimitrov, and P. L. Stanchev, “High level color similarity retrieval,” *International Journal Information Theories & Applications*, vol. 10, 2003.
[79] R. Mehta, N. Mishra, and S. Sharma, “Color Texture based Image Retrieval System,” *Int J Comput Appl*, vol. 24, no. 5, pp. 24–29, Jun. 2011, doi: 10.5120/2958-3910.
[80] J. Narwade and B. Kumar, “Local and global color histogram feature for color content-based image retrieval system,” in *Advances in Intelligent Systems and Computing*, Springer Verlag, 2016, pp. 293–300. doi: 10.1007/978-981-10-0767-5_32.
[81] B. Malinga, D. Raicu, and J. Furst, “Local vs. Global Histogram-Based Color Image Clustering,” *Computer Science*, 2006.
[82] K. Nyein and N. Hlaing, “First order statistics and glcm based feature extraction for recognition of myanmar paper currency,” in *Proceedings of 31st The IIER International Conference*, 2015.
[83] V. Kumar and P. Gupta, “Importance of Statistical Measures in Digital Image Processing,” *Computer Science*, 2012.
[84] N. Loris, L. Alessandra, and B. Sheryl, “Survey on LBP based texture descriptors for image classification,” *Expert Systems with Applications*, vol. 39, no. 3. pp. 3634–3641, 2012.
[85] T. Ojala, M. Pietikäinen, and T. Mäenpää, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Trans Pattern Anal Mach Intell*, vol. 24, no. 7, pp. 971–987, 2002, doi: 10.1109/TPAMI.2002.1017623.
[86] “Extract local binary pattern (LBP) features - MATLAB extractLBPFeatures.” Accessed: Nov. 28, 2023. [Online]. Available: https://www.mathworks.com/help/vision/ref/extractlbpfeatures.html
[87] T. Anh and P. Abstract, “Optimization of Texture Feature Extraction Algorithm,” *Computer Science*, 2010.

[88] R. M. Haralick, I. Dinstein, and K. Shanmugam, “Textural Features for Image Classification,” *IEEE Trans Syst Man Cybern*, vol. SMC-3, no. 6, pp. 610–621, 1973, doi: 10.1109/TSMC.1973.4309314.
[89] T. E. Oliphant, “Python for scientific computing,” *Comput Sci Eng*, vol. 9, no. 3, pp. 10–20, May 2007, doi: 10.1109/MCSE.2007.58.
[90] R. M. Kumar, “A Survey on Image Feature Descriptors,” *International Journal of Computer Science and Information Technologies*, vol. 5, no. 6, pp. 7668–7673, 2014.
[91] N. A. Hamilton, R. S. Pantelic, K. Hanson, and R. D. Teasdale, “Fast automated cell phenotype image classification,” 2007, doi: 10.1186/1471-2105-8-110.
[92] M. Tahir, B. Jan, M. Hayat, S. U. Shah, and M. Amin, “Efficient computational model for classification of protein localization images using Extended Threshold Adjacency Statistics and Support Vector Machines,” *Comput Methods Programs Biomed*, vol. 157, pp. 205–215, Apr. 2018, doi: 10.1016/j.cmpb.2018.01.021.
[93] M. Liu, Y. He, and B. Ye, “Image Zernike moments shape feature evaluation based on image reconstruction,” *Geo-spatial Information Science 2007 10:3*, vol. 10, no. 3, pp. 191–195, Sep. 2007, doi: 10.1007/S11806-007-0060-X.
[94] Z. Chen and S. K. Sun, “A zernike moment phase-based descriptor for local image representation and matching,” *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 205–219, Jan. 2010, doi: 10.1109/TIP.2009.2032890.
[95] E. Sariyanidi, V. Dağli, S. C. Tek, B. Tunç, and M. Gökmen, “Local Zernike Moments: A new representation for face recognition,” *Proceedings - International Conference on Image Processing, ICIP*, pp. 585–588, 2012, doi: 10.1109/ICIP.2012.6466927.
[96] A. Aggarwal and K. Singh, “Zernike moments-based retrieval of CT and MR images,” *12th IEEE International Conference Electronics, Energy, Environment, Communication, Computer, Control: (E3-C3), INDICON 2015*, Mar. 2016, doi: 10.1109/INDICON.2015.7443132.
[97] C. Singh and A. Aggarwal, “An Effective Approach for Noise Robust and Rotation Invariant Handwritten Character Recognition Using Zernike Moments Features and Optimal Similarity Measure,” *Applied Artificial Intelligence*, vol. 34, no. 13, pp. 1011–1037, Nov. 2020, doi: 10.1080/08839514.2020.1796370.
[98] B. F. Darst, K. C. Malecki, and C. D. Engelman, “Using recursive feature elimination in random forest to account for correlated variables in high dimensional data,” *BMC Genet*, vol. 19, no. 1, pp. 1–6, Sep. 2018, doi: 10.1186/S12863-018-0633-8/FIGURES/1.
[99] M. Ayyanar, S. Jeganathan, S. Parthasarathy, V. Jayaraman, and A. R. Lakshminarayanan, “Predicting the Cardiac Diseases using SelectKBest Method Equipped Light Gradient Boosting Machine,” *2022 6th International Conference on Trends in Electronics and Informatics, ICOEI 2022 - Proceedings*, pp. 117–122, 2022, doi: 10.1109/ICOEI53556.2022.9777224.
[100] P. Stefanidou-Voziki, D. Cardoner-Valbuena, R. Villafafila-Robles, and J. L. Dominguez-Garcia, “Feature Selection and Optimization of a ML Fault Location Algorithm for Low Voltage Grids,” *21st IEEE International Conference on Environment and Electrical Engineering*, 2021, doi: 10.1109/EEEIC/ICPSEUROPE51590.2021.9584731.
[101] M. B. Kursa and W. R. Rudnicki, “Feature Selection with the Boruta Package,” *J Stat Softw*, vol. 36, no. 11, pp. 1–13, Sep. 2010, doi: 10.18637/JSS.V036.I11.
[102] L. Breiman, “Random forests,” *Mach Learn*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324/METRICS.
[103] J. R. Quinlan, “Induction of decision trees,” *Machine Learning 1986 1:1*, vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: 10.1007/BF00116251.

[104] T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification," *IEEE Trans Inf Theory*, vol. 13, no. 1, pp. 21–27, 1967, doi: 10.1109/TIT.1967.1053964.
[105] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 2018. doi: 10.7551/MITPRESS/4175.001.0001.
[106] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000. doi: 10.1017/CBO9780511801389.
[107] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-August-2016, pp. 785–794, Aug. 2016, doi: 10.1145/2939672.2939785.
[108] P. Mars, J. R. Chen, and R. Nambiar, *Artificial Neural Networks (ANN)*. CRC Press, 2018. doi: 10.1201/9781351073974-3.
[109] A. F. Gad and F. E. Jarmouni, *Introduction to artificial neural networks (ANN)*. Academic Press, 2021. doi: 10.1016/B978-0-323-90933-4.00007-3.
[110] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," *https://doi.org/10.1214/aos/1013203451*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: 10.1214/AOS/1013203451.
[111] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *31st Conference on Neural Information Processing Systems*.
[112] N. A. Diamantidis, D. Karlis, and E. A. Giakoumakis, "Unsupervised stratification of cross-validation for accuracy estimation," *Artif Intell*, vol. 116, no. 1–2, pp. 1–16, 2000, doi: 10.1016/S0004-3702(99)00094-6.
[113] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Stat Surv*, vol. 4, pp. 40–79, 2010, doi: 10.1214/09-SS054.
[114] "sklearn.model_selection.StratifiedKFold — scikit-learn 1.3.2 documentation." Accessed: Dec. 01, 2023. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html
[115] "sklearn.model_selection.RepeatedStratifiedKFold — scikit-learn 1.3.2 documentation." Accessed: Dec. 01, 2023. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RepeatedStratifiedKFold.html
[116] L. Buitinck *et al.*, "API design for machine learning software: experiences from the scikit-learn project," Sep. 2013.
[117] J. Bergstra, J. B. Ca, and Y. B. Ca, "Random Search for Hyper-Parameter Optimization Yoshua Bengio," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.
[118] B. H. Shekar and G. Dagnew, "Grid search-based hyperparameter tuning and classification of microarray cancer data," *2019 2nd International Conference on Advanced Computational and Communication Paradigms, ICACCP 2019*, Feb. 2019, doi: 10.1109/ICACCP.2019.8882943.
[119] D. C. Montgomery, G. C. Runger, and J. Wiley, *Applied Statistics and Probability for Engineers*, 3rd Edition. Wiley, 2003.
[120] J. J. W. Tischendorf *et al.*, "Computer-aided classification of colorectal polyps based on vascular patterns: a pilot study," *Endoscopy*, vol. 42, no. 3, pp. 203–207, 2010, doi: 10.1055/S-0029-1243861.
[121] S. Gross *et al.*, "Computer-based classification of small colorectal polyps by using narrow-band imaging with optical magnification," *Gastrointest Endosc*, vol. 74, no. 6, pp. 1354–1359, Dec. 2011, doi: 10.1016/J.GIE.2011.08.001.

[122] Y. Kominami *et al.*, "Computer-aided diagnosis of colorectal polyp histology by using a real-time image recognition system and narrow-band imaging magnifying colonoscopy," *Gastrointest Endosc*, vol. 83, no. 3, pp. 643–649, Mar. 2016, doi: 10.1016/J.GIE.2015.08.004.
[123] P. Mesejo *et al.*, "Computer-Aided Classification of Gastrointestinal Lesions in Regular Colonoscopy," *IEEE Trans Med Imaging*, vol. 35, no. 9, pp. 2051–2063, Sep. 2016, doi: 10.1109/TMI.2016.2547947.
[124] M. Misawa *et al.*, "Characterization of Colorectal Lesions Using a Computer-Aided Diagnostic System for Narrow-Band Imaging Endocytoscopy," *Gastroenterology*, vol. 150, no. 7, pp. 1531-1532.e3, Jun. 2016, doi: 10.1053/j.gastro.2016.04.004.
[125] Y. Mori *et al.*, "Impact of an automated system for endocytoscopic diagnosis of small colorectal lesions: an international web-based study," *Endoscopy*, vol. 48, no. 12, pp. 1110–1118, Dec. 2016, doi: 10.1055/S-0042-113609.
[126] Y. Takemura *et al.*, "Quantitative analysis and development of a computer-aided system for identification of regular pit patterns of colorectal lesions," *Gastrointest Endosc*, vol. 72, no. 5, pp. 1047–1051, Nov. 2010, doi: 10.1016/J.GIE.2010.07.037.
[127] S. Y. Park and D. Sargent, "Colonoscopic polyp detection using convolutional neural networks," *Medical Imaging 2016: Computer-Aided Diagnosis*, vol. 9785, p. 978528, Mar. 2016, doi: 10.1117/12.2217148.
[128] G. Fernández-Esparrach *et al.*, "Exploring the clinical potential of an automatic colonic polyp detection method based on the creation of energy maps," *Endoscopy*, vol. 48, no. 9, pp. 837–842, Sep. 2016, doi: 10.1055/S-0042-108434.
[129] R. Bisschops *et al.*, "BASIC (BLI Adenoma Serrated International Classification) classification for colorectal polyp characterization with blue light imaging," *Endoscopy*, vol. 50, no. 3, pp. 211–220, Mar. 2018, doi: 10.1055/S-0043-121570.
[130] R. Zhang *et al.*, "Automatic Detection and Classification of Colorectal Polyps by Transferring Low-Level CNN Features From Nonmedical Domain," *IEEE J Biomed Health Inform*, vol. 21, no. 1, pp. 41–47, Jan. 2017, doi: 10.1109/JBHI.2016.2635662.
[131] M. F. Byrne *et al.*, "Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model," *Gut*, vol. 68, no. 1, pp. 94–100, 2019, doi: 10.1136/gutjnl-2017-314547.
[132] M. E. Paoletti, S. Moreno-Alvarez, and J. M. Haut, "Multiple Attention-Guided Capsule Networks for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, 2022, doi: 10.1109/TGRS.2021.3135506.
[133] Y. Dai, Y. Gao, F. Liu, Y. ; Dai, Y. ; Gao, and F. Liu, "TransMed: Transformers Advance Multi-Modal Medical Image Classification," *Diagnostics 2021, Vol. 11, Page 1384*, vol. 11, no. 8, p. 1384, Jul. 2021, doi: 10.3390/DIAGNOSTICS11081384.
[134] A. Ramon-Michel Schreuder *et al.*, "Algorithm combining virtual chromoendoscopy features for colorectal polyp classification," *Endosc Int Open*, vol. 09, no. 10, pp. E1497–E1503, Oct. 2021, doi: 10.1055/A-1512-5175.
[135] Z. Wen *et al.*, "Deep learning in digital pathology for personalized treatment plans of cancer patients," *Semin Diagn Pathol*, vol. 40, no. 2, pp. 109–119, Mar. 2023, doi: 10.1053/J.SEMDP.2023.02.003.
[136] A. J. Schaumberg *et al.*, "Interpretable multimodal deep learning for real-time pan-tissue pan-disease pathology search on social media," *Modern Pathology*, vol. 33, no. 11, pp. 2169–2185, Nov. 2020, doi: 10.1038/s41379-020-0540-1.
[137] Y. Zheng *et al.*, "Application of Transfer Learning and Ensemble Learning in Image-level Classification for Breast Histopathology," *Intelligent Medicine*, vol. 3, no. 2, pp. 115–128, Apr. 2022, doi: 10.1016/j.imed.2022.05.004.

# APPENDIX

# Colonoscopy Examination Result Form

A sample of the colonoscopy examination result reports made by Gastroenterologist Assoc. Dr. Serkan Doğan specific to each patient is attached. The reports include detailed findings observed during each patient's colonoscopy, diagnoses made, and recommended treatment approaches. The presentation of the reports in this section is intended to support the scope of the research and the applied methodologies with concrete examples. Each report is prepared with personal information suppressed to protect the patient's privacy. Detailed review of the reports is important in understanding the diversity of colonoscopy procedures and clinical situations encountered.

A typical colonoscopy report includes the following sections:

- Patient Information: Patient name, age, gender, and other personal information.
- Examination Date: The date the colonoscopy was performed.
- Reference Doctor: The doctor who orders or directs the colonoscopy for the patient.
- Procedure Information: Details of how the colonoscopy procedure is performed, equipment used, and anesthesia information.
- Findings: Abnormal findings were observed during colonoscopy (e.g., polyps, inflammation, ulcers, etc.).
- Diagnosis: Diagnosis or possible diagnoses made based on observations.
- Recommendations: Necessary treatment methods, medications, dietary recommendations, or follow-up examinations.
- Additional Notes: Any complications or additional information encountered during the procedure.
- Signature: Signature and date of the doctor performing the procedure.

T.C.
SAĞLIK BAKANLIĞI
Kayseri Şehir Hastanesi

| Dk.No: | | Yayin Tarihi: | | Revize No: | | Revize Tarihi: | | Sayfa No: | 1 |
|---|---|---|---|---|---|---|---|---|---|

**TETKIK SONUÇ RAPORU**

| | | | |
|---|---|---|---|
| **Adi Soyadi** | : | **Islem No** | : |
| **T.C. No** | : | **Istem Doktor** | : |
| **Cinsiyet** | : | **Onaylayan Doktor** | : |
| **Yaş** | : | **Istem Servis** | : |
| **Dogum Tarihi** | : | **Istem Tarihi** | : |
| **Birim** | : | **Onay Tarihi** | : |
| **Tetkik** | : | **Çekim Tarihi** | : |
| **Danışman Dr** | : | | |
| **Klinik Bilgi** | : ..... | | |

**SONUÇ AÇIKLAMA;**

Süre: 20 DK

**Şikayeti:**

**Premedikasyon:**

**Cihaz:**

**BULGULAR:**

**BİYOPSİ :+**

**TANI:**

# Pathology Laboratory Examination Form

A sample of the medical pathology laboratory examination result reports filled out by Specialist Pathologist Dr. Ebru Akay is included in this section. These reports contain the pathological analysis results of biopsy samples taken during colonoscopy examinations performed by Gastroenterologist Assoc. Dr. Serkan Doğan for each patient.

Each pathology report provides detailed histopathological examinations of biopsy samples taken during the patient's colonoscopy. These examinations include the microscopic structure of the tissues as well as immunohistochemical (IHC) test results. Reports detail the pathological changes observed, cell types identified, and any abnormal findings.

The presentation of pathology reports aims to reveal the depth of the research and the methodological approaches applied with concrete examples. These reports reveal the complexity of colonoscopy procedures and the different clinical situations encountered. Each report is presented with personal information suppressed to protect patient privacy.

A detailed review of the reports highlights the diversity of colonoscopy and subsequent pathological evaluations and the clinical importance of these procedures. This represents a particularly valuable resource for researchers in the fields of gastroenterology and pathology.

A typical pathology laboratory examination result report includes the following sections:

- Patient Information: Patient name and surname, age and gender, patient identification number or unique identifier, history
- Sample Information: Sample number, sample collection date and time, sample type (biopsy, surgical sample, etc.)
- Clinical Information and History: Clinical diagnosis, patient's medical history, and symptoms, requested tests or special requests.
- Macroscopic Findings: Physical appearance and measurements, details such as number of sections and color.
- Microscopic findings: Cellular structures, morphology of tissues, detailed description of pathological changes.
- Immunohistochemistry (IHC) Findings: Antibodies used and results, cellular markers detected by IHC.
- Diagnosis and Comment: Histopathological diagnosis based on microscopic and IHC findings, additional comments, and suggestions if necessary.
- Pathologist Signature and Information: Name and signature of the pathologist performing the examination, and history.

T.C.
SAĞLIK BAKANLIĞI
Kayseri Şehir Hastanesi

| Dk.No: | LAB.PL.FR.30 | Yayin Tarihi: | 5/15/18 12:00 | Revize No: | 0 | Revize Tarihi: | | Sayfa No: | 1 |
|---|---|---|---|---|---|---|---|---|---|

## TIBBİ PATOLOJİ LABORATUVARI TETKİK SONUÇ RAPORU

| | | | |
|---|---|---|---|
| İşlem No | : | Rapor No | : |
| Hasta Adı Soyadı | : | İsteyen Servis | : |
| TC Kimlik No | : | İsteyen Doktor | : |
| Cinsiyet / Yaş | : | İstem Tarih/Saat | : |
| Doğum Tarihi | : | Numune Alım Tarih/Saat | : |
| Kurumu | : SOSYAL GÜVENLİK KURUMU | Numune Kabul Tarih/Saat | : |
| Telefon | : | Onay Tarih/Saat | : |
| Doku Alınış Yeri | : Peritonun spesifik bölgeleri | Numune Kabul Eden | : |
| Numune Türü | : BIOPSI | | |
| ICD-0 | : 8000/3 - Malign neoplazi | | |
| Eski Biyopsi No | : | | |
| Ön Tanısı | | | |
| Frozen Tanı | | | |

**KLİNİK BULGU**

dispespi

**MAKROSKOPİ**

**MİKROSKOPİ**

**HİSTOKİMYA/İMMÜNOHİSTOKİMYA**

İMMÜNOHİSTOKİMYA:

**TANI**

**2- Asit sıvısı, yaymalar:**

**3- İnen kolon proksimali, endoskopik biyopsi:**

**4- İnen kolon distali, endoskopik biyopsi:**

# Informed Consent Form from Volunteer Participants

| BİLGİLENDİRİLMİŞ SAĞLIKLI GÖNÜLLÜ OLUR FORMU (BGOF) | **Tarih:** |
|---|---|
| | **Sayfa No:** 1/2 |

**Bilgilendirme Bölümü:**

Değerli Gönüllü,

İştirak etmenizi talep ettiğimiz "**Kolon Polipleri için Kolonoskopi ve Histopatoloji Görüntülerinden Yapay Zekâ Destekli Prognostik Belirteç Tespiti"** başlıklı bu çalışma, muayene esnasında hastadan elde edilen kolon poliplerinin görüntü ve video kayıtlarından polipin damarlanma ve diğer özellikleri çıkarılarak, bu polipin sağlıklı mı, eğer değilse kansere dönüşme riskinin bilgisayar ortamında analiz edilmesini ve gerçek zamanlı sınıflandırılmasını sağlayacaktır. Ayrıca yapılan bu analizler, moleküler testler ile de desteklenecektir.

Amacımız doktorun kolonoskopi muayenesi esnasında karşılaştığı polipleri analiz için biyopsi alıp, patoloji laboratuvarına göndermeye gerek kalmadan gerçek zamanlı olarak analiz edip, polipin bağırsak dokusundan alınıp alınmamasına karar vermesi için bilgisayar destekli bir yazılım geliştirmektir. Araştırmaya katılması için tahmini gönüllü sayısı 200 kişi ve çalışmanın öngörülen süresi 2 yıldır.

Çalışma, polipi olan gönüllü hastalardan, görüntü ve video kayıtları alınmasının yanı sıra patolojiye gönderilen şüpheli doku ve onun yanından kontrol için alınan normal doku ile gerçekleştirilecektir.

Bu projede Fujinon marka VP-3500HD video işlemcisi ve 300 W Xenon lambalı XL-4450 ışık kaynağı olan yüksek çözünürlüklü (HD) kolonoskopi sistemi cihazı kullanılarak rutin muayene işleminiz yapılacak ve ek olarak hiçbir şey uygulanmayacaktır.

Sizden beklenen, rutin muayene işleminiz esnasında kalın bağırsağınızda polip tespit edilmesi durumunda, bu poliplerin görüntü ve video kayıtlarının bizimle paylaşılmasına ve karşılaştırma yapabilmek için yanındaki normal dokudan da biyopsi için örnek alınmasına razı olmanızdır. Analizler Kayseri Şehir Hastanesi bünyesinde Gastroenteroloji Kliniği, Patoloji Laboratuvarı ile Abdullah Gül Üniversitesi bünyesinde Biyomedikal Enstrümantasyon ve Sinyal Analizi laboratuvarında yapılacaktır.

Araştırmaya dâhil olma durumunuzda, hastalığınız için aldığınız tedavi değişmeyecek veya etkilenmeyecek ve size, yeni bir tedavi verilmeyecek veya uygulama yapılmayacaktır. Ayrıca, araştırmaya katılımınız isteğe bağlı olup ve istediğiniz zaman, bir cezaya veya yaptırıma maruz kalmaksızın ve hiçbir hakkınızı kaybetmeksizin, araştırmaya katılmayı reddedebilir veya araştırmadan çekilebilirsiniz; kimliğinizi ortaya koyacak kayıtlar gizli tutulacak, kamuoyuna açıklanamayacak olup araştırma sonuçlarının yayımlanması halinde bile kimliğiniz gizli kalacaktır.

Araştırma süresince 24 saat ulaşılabileceğiniz kişilerin adı ve soyadı:

| BİLGİLENDİRİLMİŞ SAĞLIKLI GÖNÜLLÜ OLUR FORMU (BGOF) | **Tarih:** |
|---|---|
| | **Sayfa No:** 2/2 |

**Gönüllü Oluru Bölümü**:

*"Bilgilendirilmiş Gönüllü Olur Formundaki tüm açıklamaları okudum. Bana, konusu ve amacı belirtilen araştırma ile ilgili yazılı ve sözlü açıklama, aşağıda adı belirtilen hekim tarafından yapıldı. Araştırmaya gönüllü olarak katıldığımı, istediğim zaman gerekçeli veya gerekçesiz olarak araştırmadan ayrılabileceğimi ve kendi isteğime bakılmaksızın araştırmacı tarafından araştırma dışı bırakılabileceğimi biliyorum".*

*"Söz konusu araştırmaya, hiçbir baskı ve zorlama olmaksızın, özgür irademle (kendi rızamla) katılmayı kabul ediyorum. Çalışma kapsamında benden alınan kayıtların ileride yapılması planlanan tüm araştırmalarda kullanılmasına izin veriyorum".*

- Gönüllünün Adı / Soyadı / İmzası / Tarih

- Açıklamaları Yapan Hekimin Adı / Soyadı / İmzası / Tarih

- Gerekiyorsa Olur İşlemine Tanık Olan Kişinin Adı / Soyadı / İmzası / Tarih

- Gerekiyorsa Yasal Temsilcinin Adı / Soyadı / İmzası / Tarih

# CURRICULUM VITAE

| | |
|---|---|
| 2014 – 2017 | B.Sc., Biomedical Engineering, Erciyes University, Kayseri, TURKEY |
| 2018 – 2023 | Ph.D., Bioengineering Department, Abdullah Gül University, Kayseri, TURKEY |
| 2020 – Present | Research Assistant, Bioengineering Department, Abdullah Gül University, Kayseri, TURKEY |

**SELECTED PUBLICATIONS AND PRESENTATIONS**

**J1)** F. Uslu, K. Icoz, K. Tasdemir, **R. S. Doğan,** B. Yilmaz, Image-Analysis Based Readout Method for Biochip: Automated Quantification of Immunomagnetic Beads, Micropads, and Patient Leukemia Cell published in Micron (2020).

**J2) R. S. Doğan,** B. Yilmaz, Comparison of Deep Learning and Conventional Machine Learning Methods for Classification of Colon Polyp Types published in The EuroBiotech Journal (2021).

**J3) R. S. Doğan,** B. Yılmaz, Histopathology Image Classification: Highlighting the Gap Between Manual Analysis and AI Automation published in Frontiers in Oncology (2023).

**J4**) R.N. Kaçmaz, **R.S. Doğan,** B. Yılmaz, A Comprehensive Study on Automatic Non-Informative Frame Detection in Colonoscopy Videos published in International Journal of Imaging Systems and Technology (2023).

**J5) R. S. Doğan,** E. Akay, S. Doğan, B. Yılmaz, Hyperplastic and Tubular Polyp Classification Using Machine Learning and Feature Selection under review in Computerized Medical Imaging and Graphics.

**J6) R.S. Doğan,** E. Akay, S. Doğan, B. Yılmaz, A Benchmark Dataset for Colon Polyps: Colonoscopy Videos, Histopathology Images, and IHC Results under review in Nature Scientific Data.

**C1) R. S. Doğan,** B. Yilmaz, Polyp Localization in Colonoscopy Images Using Vessel Density in IEEE Medical Technologies National Congress (2018).

**C2) R. S. Doğan,** S. Saka, B. Güngör, Investigation of Hepatocellular Carcinoma Molecular Mechanisms via in Silico Analyses in IEEE 5th International Conference on Computer Science and Engineering (2020). **(Best paper award)**

**C3) R. S. Doğan,** H. İ. Baynal, B. Yilmaz, The Effect of Local Binary Pattern Parameters on the Classification Performance of Polyps using Colonoscopy Images in 8th International Medical and Health Sciences Research Congress (2021).

**C4) R. S. Doğan,** B. Yılmaz, Binary and Multiclass Classification of Histopathological Images Using Deep and Machine Learning-Based Techniques in 5th International Conference on Data Science and Applications ICONDATA'22 (2022).