IMPERIAL

Enhancing Medical Diagnosis with Deep Learning-based Image Segmentation

Author
YANG, Xinchen
CID: 02431691

Supervised by
Dr Tania Stathaki

A Thesis submitted in fulfillment of requirements for the degree of
Master of Science in Communications and Signal Processing

Department of Electrical and Electronic Engineering
Imperial College London
2024

Abstract

The early detection of Colorectal Cancer (CRC) depends heavily on the accurate diagnosis of colorectal polyps. In recent years, Artificial Intelligence (AI) has been applied in Computer-Aided Diagnosis (CADx) to support colonoscopists during and after colonoscopy procedures. Previous studies have commonly combined segmentation and classification models for polyp margin delineation and the determination of polyp characteristics (e.g., hyperplastic or adenomatous). They mainly focused on White Light Endoscopy (WLE) images, while Narrow Band Imaging (NBI), an advanced imaging technique that is more effective in visualizing polyp surface features, is still underexplored in this context. Meanwhile, although deep learning-based classifiers usually offer high accuracy, their lack of explainability reduces their credibility among colonoscopists. Conversely, traditional machine learning classifiers offer more transparent feature design processes but lack accuracy. This master thesis proposes a cascaded model using Polyp-PVT for segmentation and a Support Vector Machine (SVM) for classifying polyps as adenomatous or hyperplastic. The model is specifically applied to NBI images. Based on the authoritative Japan NBI Expert Team (JNET) Classification standard, 26 features are designed to describe polyps' surface patterns, morphological features, and transition patterns to normal mucosa. These features are selected and ranked using Sequential Floating Backward Selection (SFBS) and subsequently used for classifier training. To implement the cascaded model, a polyp diagnosis platform is developed in Python. It integrates clear result visualization and multiple functions for colonoscopists to process medical images more efficiently. Experimental results achieved an overall segmentation Dice score of 0.8761 and a classification accuracy of 90.83% on a customized test dataset. The source code of this project is available at `https://github.com/Myosotis1111/Polyp-diagnosis-platform`

Declaration of Originality

I hereby declare that the work presented in this thesis is my own unless otherwise stated. To the best of my knowledge the work is original and ideas developed in collaboration with others have been appropriately referenced.

Copyright Declaration

Acknowledgments

The one-year Master's project has come to an end in the blink of an eye, marking the end of my student life. Over the past months, I've been fully devoted to both job hunting and completion of this thesis, which has left me exhausted, but still deeply fulfilled. This master thesis, representing everything I've learned over the years, serves as a conclusion to my academic journey and the start of a new chapter in my life.

I would like to express my sincere gratitude to Dr Tania Stathaki, my first instructor. Your invaluable insights and guidance have been useful in helping me complete this project. Your courses on digital image processing left a deep impression on me. While I often struggled with complex calculations and abstract concepts in other courses, your teaching style made everything clear and accessible. The concepts you taught are still vividly clear to me and have been directly applied in this thesis. In fact, much of the feature design in this work is based on the knowledge I gained from your classes. I would also like to extend my gratitude to Dr Athanasios Gkelias, for taking your valuable time to read and assess my thesis.

I would also like to thank Ms. Huang and Ms. Zhou, colonoscopists at Shanghai Fourth People's Hospital. As someone with no prior knowledge of colonoscopy, I greatly appreciate the clinical knowledge you provided for this thesis and your enthusiasm to answer all my questions with great detail. You also helped me with the collection and annotation of NBI images, which saved me a lot of time and ensured that the dataset was both valid and convincing for training and testing.

Finally, I want to express my deepest love and gratitude to my parents. Throughout my years as a student, your support and sacrifices allowed me to explore this meaningful chapter of my life on the other side of the continent. During countless moments of stress and sorrow, you offered me boundless patience and comfort, for which I am truly grateful. As I begin my career and start earning a living through my efforts, I hope to one day pass on the same love and support to my own family, just as you have done for me.

This master thesis is dedicated to all the extraordinary individuals I have encountered throughout my 24 years of life. May your future be filled with boundless success and joy.

Contents

List of Acronyms

AI Artificial Intelligence

CRC Colorectal Cancer

CADx Computer-Aided Diagnosis

ML Machine Learning

DL Deep Learning

NLP Natural Language Processing

ROI Region of Interest

NBI Narrow Band Imaging

JNET Japan NBI Expert Team

WLE White Light Endoscopy

IoU Intersection over Union

SVM Support Vector Machine

RF Random Forest

PCA Principal Component Analysis

SFS Sequential Feature Selection

SFFS Sequential Floating Forward Selection

SFBS Sequential Floating Backward Selection

GLCM Gray-Level Co-occurrence Matrix

CNN Convolutional Neural Network

ViT Vision Transformer

PVT Pyramid Vision Transformer

CCS Connected Component Size

HD Histogram Distance

TVS Transition Variability Score

VAR Variance

EVAR Energy Variance

BED Border-to-Ellipse Distance

GUI Graphical User Interface

MVC Model-View-Controller

YOLO You Only Look Once

List of Figures

List of Tables

1 Introduction

Contents

1.1 Motivation

According to the latest statistical data of the International Agency for Research on Cancer (IARC) [1], Colorectal Cancer (CRC) has become the third most common malignant tumor globally, and the second leading cause of cancer-related deaths. The development of CRC is strongly related to colorectal polyps, a common intestinal disorder. Colonoscopy, "an endoscopic procedure pioneered in Japan in the late 1950s that allows visualization of the entire mucosa of the large intestine and distal terminal ileum"[2], is regarded as the most effective and safest screening method for detecting colorectal polyps. However, colonoscopists' unbalanced skills and the reduction of concentration after consecutive inspections can lead to misdetection and misclassification, which may affect treatment decisions. Furthermore, images taken during the inspection need to be examined and categorized afterward for report generation and further reference. This also requires significant time and effort from the colonoscopists to do manually.

The rapid development of Artificial Intelligence has boosted CADe and CADx. Machine Learning (ML), a subset of artificial intelligence, has been widely applied in medical image processing for

cancer detection, segmentation, and classification since the mid-1960s [3]. To tackle the limitations of manual feature engineering in traditional machine learning approaches, Deep Learning (DL), a branch of ML, was proposed to enable the automatic extraction of complex features from medical images. DL-based segmentation methods have shown the ability to localize polyps and accurately provide reference information on polyp borders. Based on these segmentation results, classification algorithms are used then to analyze the extracted regions and predict whether a polyp is benign or malignant. Therefore, developing a robust algorithm represents a promising approach to increasing the accuracy and efficiency of the colonoscopy diagnosis workflow.

1.2 Goal

This research aims to achieve automatic colorectal polyp segmentation and classification. A DL-based segmentation model needs to be deployed to offer a clear separation between the background and the Region of Interest (ROI), ensuring that the masked images are accurate to be processed in subsequent classification. Based on this, a suitable classification method will be chosen to provide precise categorization for masked polyp images. The segmentation and classification modules will be integrated into a cascaded model to realize a complete automatic diagnosis workflow. The performance of the two modules will be evaluated both separately and in combination for further improvement.

1.3 Organization

This master thesis consists of six chapters: In **chapter 2**, a comprehensive review of relevant literature and potential methods are provided. In **chapter 3**, we first analyze task challenges, then we evaluate different methods and algorithms to determine the most effective approach. According to the proposed algorithm, new methods and metrics are defined. In **chapter 4**, we describe how the selected algorithm is implemented in a Python environment. In **chapter 5**, the algorithm is tested on a limited dataset, a detailed evaluation of the results is presented and the methodologies are revised. Finally, we provide a comprehensive summary of the research findings, draw meaningful conclusions, and outline recommendations for future research.

2

Review of Methodologies

Contents

This chapter aims to conduct a preliminary task analysis and paper review, and then a systematic introduction to possible methods for solving the given task is provided. A literature review on common methods will be presented to comprehensively understand the topic.

2.1 Narrow Band Imaging (NBI) in Colonoscopy

2

Although colonoscopy is considered the gold standard for early detection and removal of CRC, it is not perfect and can miss less significant lesions [2]. During the past decades, novel techniques and worldwide standards have been established to enhance the polyp diagnosis in colonoscopy. Among them, the NBI is one of the most commonly used techniques that is widely integrated into today's endoscopic system.

2.1.1 Narrow Band Imaging (NBI)

The National Cancer Center Hospital East originally developed the idea of NBI in 1999. Under the leadership of Dr. Shigeaki Yoshida, Sano and colleagues spearheaded this breakthrough. Monochrome NBI (red/green/blue) filter was a short-wavelength narrow-band filter that was used in a prototype setup in 2001 [4]. In 2003, color imaging of gastrointestinal microvascular architecture and tumor surfaces was achieved using 415nm and 540nm filters. After improved noise reduction, light, and color adjustment, Olympus launched the EVIS LUCERA SPECTRUM as the final mass-production model in 2006 [5].

NBI highlights the patterns of pits and tiny blood vessels using light with a shorter wavelength. This provides a clearer visualization of the surface pattern of the polyps, as shown in Figure 2.1. The border between polyps and colonic mucosa thus becoming easier to identify, and more surface features are captured. This gave rise to JNET, a new classification standard based on the surface pattern of the polyps, which will be introduced in the next subsection.

Compared to the conventional WLE, the use of NBI in the proximal colon for high-risk patients proves to improve the detection of adenoma, particularly for those with a flat morphology [6]. Unlike other advanced colonic imaging methods, such as chromoendoscopy and autofluorescence imaging, NBI is much simpler. Chromoendoscopy requires the application of dyes, while autofluorescence imaging demands special and expensive equipment. In contrast, NBI is a push-button technology that has been seamlessly integrated into most endoscopes. Therefore, NBI becomes a prevalent and useful supporting method for observing the endoscopic findings of early cancer today [7].

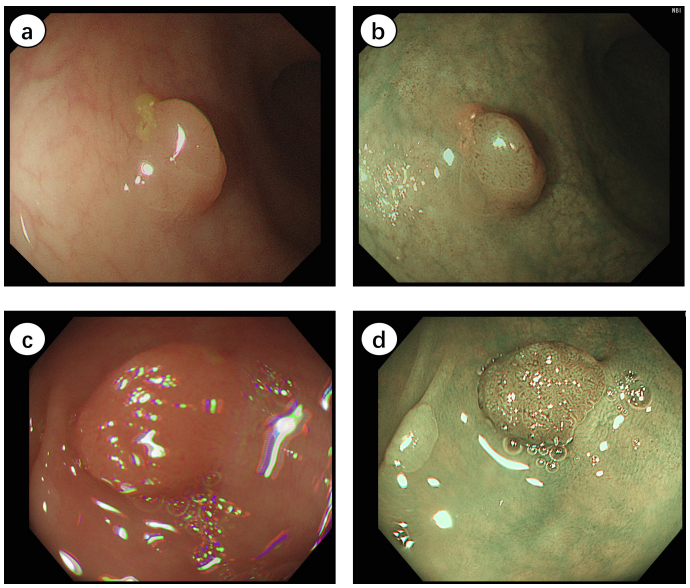
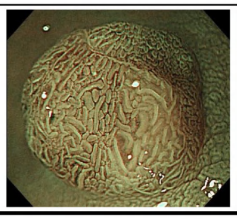
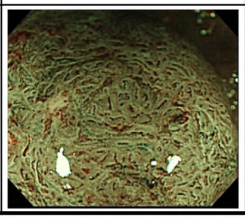
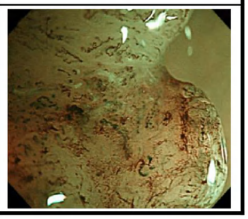


Figure 2.1: White Light Endoscopy (WLE) and NBI images of the colon. **a** , **c** | Example images of high-resolution WLE; **b** , **d** | Corresponding NBI images of **a** and **c**.

2.1.2 Japan NBI Expert Team (JNET) Classification

After the NBI technique was introduced, several issues arose. One issue was the need to use surface patterns as reference points for classification. Another was distinguishing between elevated and superficial lesions. To tackle these issues, in 2011, the JNET established a universal classification standard for colorectal tumors using NBI endoscopy, known as the JNET Classification [5].

Figure 2.2 depicts how tumors are classified according to JNET Classification. Based on the vessel pattern and surface pattern, four classes are established. According to the classification outcome, the treatment of the tumors is decided. For **Type 1** tumors, observation and conservative treatment are usually recommended; **Type 2A** tumors are generally removed by endoscopic resection using open forceps or endoscopic mucosal resection (EMR); **Type 2B** tumors require a further magnifying chromoendoscopy to determine the appropriate treatment method; **Type 3** tumors, however, are considered an early stage of cancer and cannot be directly removed. They require pathological analysis and subsequent surgery for removal [8].

	Type 1	Type 2A	Type 2B	Type 3
Vessel pattern	· Invisible ※1	・Regular caliber ・Regular distribution (meshed/spiral pattern) ※2	・Variable caliber ・Irregular distribution	・Loose vessel areas ・Interruption of thick vessels
Surface pattern	・Regular dark or white spots ・Similar to surrounding normal mucosa	・Regular (tubular/branched/papillary)	・Irregular or obscure	・Amorphous areas
Most likely histology	Hyperplastic polyp/ Sessile serrated polyp	Low grade intramucosal neoplasia	High grade intramucosal neoplasia/ Shallow submucosal invasive cancer ※3	Deep submucosal invasive cancer
Endoscopic image				

*1. If visible, the caliber in the lesion is similar to surrounding normal mucosa.
*2. Microvessels are often distributed in a punctate pattern and well-ordered reticular or spiral vessels may not be observed in depressed lesions.
*3. Deep submucosal invasive cancer may be included.

Figure 2.2: The Japan NBI Expert Team (JNET) classification [5].

2.2 AI-Enabled Colonoscopy: A Paper Review

2.2.1 Role of AI in endoscopy

As shown in Figure 2.3, Sung et al. stated we are currently in Level 1 of AI decision-making [9]. This means instead of providing automated operations on patients, AI is mainly used to provide information about lesions as a reference to aid endoscopists in decision-making. For example, during the endoscopy (with colonoscopy being a specific type), AI can highlight the area of potential lesions and categorize them with a certain confidence level. However, treatment methods (e.g., resection or conservative treatment) are still entirely decided by the endoscopist. Such a slow development of AI in healthcare is attributed to multiple factors. The most frequent question asked is where responsibility lies when negligence occurs. It is hard to hold AI developers legally accountable, thus hospitals and medical workers tend to doubt decisions made by algorithms until certain techniques are sufficiently mutual. Therefore, improving the accuracy and reliability of AI polyp diagnosis becomes a key focus area. Based on the research objectives, three directions are investigated here: **Polyp detection**, **Polyp segmentation**, and **Polyp classification**.

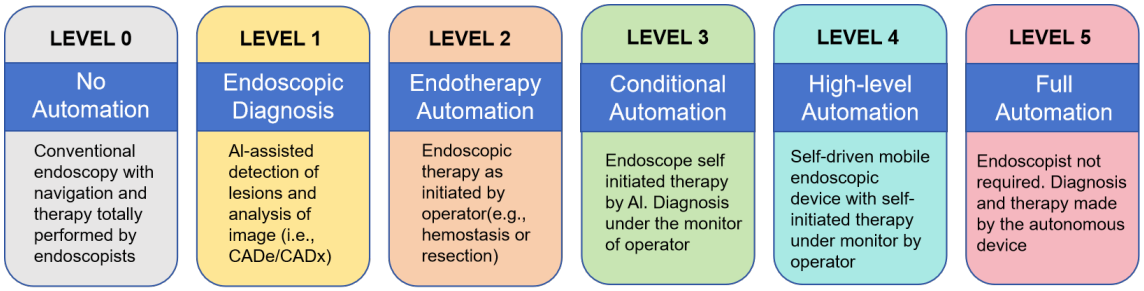


Figure 2.3: Level of AI decision-making in Endoscopy [9].

2.2.2 Polyp detection

When margin delineation is not considered critical, we prefer to choose object detection algorithms as they usually offer faster processing time and better detection ratios. This approach can help colonoscopists localize less significant tumors more efficiently compared to polyp segmentation.

Traditional machine learning methods proved to be less effective due to the variations in size, orientation, color, and texture of colorectal polyps[10]. However, the introduction of the Convolutional Neural Network (CNN), a deep learning network architecture, has revolutionized the field of medical image processing [11]. Polyp detection algorithms using CNNs have shown far better performance with higher sensitivity and precision, particularly for lesions with high confidence predictions [12].

Among all CNN-based object detection algorithms, the You Only Look Once (YOLO) algorithm [13] stands out for its capability to deliver real-time detection. This enables immediate feedback during colonoscopy procedures. In 2022, Nogueria et al. fine-tuned a YOLOv3-based model, achieving an F1 score of 0.88 and a polyp-based sensitivity of 89.91% with 54.97% specificity on colonoscopic video data with a detection rate exceeded 40 frames per second (fps) [14]. To further enhance YOLO's performance across versions, Karaman et al. introduced a potent hyperparameter tuning technique named "Artificial Bee Colony" (ABC) which can boost YOLO's efficiency across its iterations and is also adaptable to different YOLO variants[15]. YOLO algorithms have demonstrated quick development and good interoperability with supporting technologies, suggesting opportunities for further improvements in the identification of colorectal polyps.

2.2.3 Polyp segmentation

Polyp segmentation is a relatively difficult task compared to the detection in colonoscopy, as it not only requires localizing potential polyps but also needs to state the clear border between lesions

and the colonic mucosa. Object detection algorithms that use bounding boxes to indicate the ROI inevitably include additional information about the colonic mucosa. This extra information can make the subsequent classification of polyps more complicated. In such cases, pixel-level analysis enabled by segmentation algorithms becomes necessary to produce a masked image that contains only the information about the polyps.

In 2015, Ronneberger et al. proposed U-Net, a CNN-based network named after its U-shaped architecture [16], which has since become a milestone in medical image segmentation. Many variations developed based on U-Net were used for polyp segmentation, for example, U-Net++ [17], ResUNet [18] and ResUNet++ [19]. These models are tested on the Kvasir-SEG [20] dataset and achieve similar performance of mean Dice coefficients of around 0.8 [21]. In 2020, PraNet, proposed by Fan et al., used a two-stage approach: a parallel decoder to predict rough regions and an attention mechanism to enhance polyp edges and internal structure for detailed segmentation. Its proposal marked a significant step forward in the field with a mean Dice score of 0.898 on the Kvasir-SEG dataset [22]. Besides, the dataset used in their paper has been regarded as the benchmark for subsequent research on polyp segmentation as a consistent standard for evaluating different segmentation models.

The transformer architecture was initially designed for Natural Language Processing (NLP) tasks. It became significant in computer vision with the introduction of the Vision Transformer (ViT) by Dosovitskiy et al. in 2020 [23]. Since then, researchers have started to apply transformer-based networks to polyp segmentation and proved better performance compared to CNN-based networks. In 2021, Dong et al. proposed Polyp-PVT, employing Pyramid Vision Transformer (PVT), a variant of the ViT, as an encoder. On the benchmark dataset, it outperformed PraNet and showed greater resilience against noise and camouflage [24]. Subsequently, using the same test dataset, Tang et al.'s DuAT (Dual-Aggregation Transformer Network), another PVT-based network, performed even better than Polyp-PVT [25].

Notably, the development of large segmentation models has shown great promise in polyp segmentation. In 2023, the Segment Anything Model (SAM) was introduced as a foundational model designed to segment objects defined by users. It was trained on over 1 billion annotations, primarily focusing on natural images [26]. This training allowed for segmentation without the need for specific training and fine-tuning. In 2024, Li et al. fine-tuned SAM for polyp segmentation, resulting in Polyp-SAM. They compared it with state-of-the-art polyp segmentation models. Although Polyp-SAM did not surpass Polyp-PVT, it achieved high mean Dice Scores above 88% across all sub-datasets of the benchmark dataset [27]. Simultaneously, large medical datasets were utilized

to make SAM more specialized for medical segmentation tasks. MedSAM, introduced by Ma et al. in the same year, was developed using a medical image dataset comprising 1,570,263 image-mask pairs. This dataset included 10 imaging modalities and over 30 forms of cancer [28]. Furthermore, the combination of implicit segmentation methods and large segmentation models has enhanced segmentation performance. The newly proposed I-MedSAM [29] outperformed MedSAM on the test set. In summary, the advent of SAM has paved the way for a novel approach to polyp segmentation. As more medical images are added to the dataset and supporting methodologies are refined, SAM-based polyp segmentation models are expected to outperform traditional models in the near future.

2.2.4 Polyp classification

Polyp classification is crucial to help colonoscopists determine polyp characteristics and decide on treatment methods. It is usually a downstream task following polyp detection or polyp segmentation. Unlike detection and segmentation tasks where traditional machine learning methods perform poorly, polyp classification is based on clear standards and low-dimensional features. In such cases, manual feature engineering can sometimes be more effective and credible than deep learning methods.

In traditional machine learning, feature engineering significantly impacts classifier performance. In the past few years, researchers have explored various metrics to identify those that most accurately characterize the surface patterns of polyps. Pomeroy et al. [30] and Fu et al. [31] both employed the technique of GLCM (gray level co-occurrence matrix) for feature extraction in their papers. They proved GLCM's robust capability for pattern representation. In 2016, Hu et al. proposed an improved polyp classification method by combining the Haralick co-occurrence matrix and the Karhunen-Loeve (KL) transform to reduce feature dependency [32]. Besides features extracted from the texture domain, spatial and spectral domain features are also considered. Wimmer et al. utilized various wavelet-based techniques, including Discrete Wavelet Transform (DWT) and dual-tree Complex Wavelet Transform (DT-CWT) to extract spectral domain features and achieved satisfying results [33]. However, pattern domain features have proved dominant in polyp classification. In Pomeroy's study, two feature selection techniques are applied to discard less significant features. As a result, most of the spectral and spatial domain features are discarded [30]. In another comparative study by Engelhardt et al., GLCM and its variants are stated to perform well in polyp classification, while the discrete wavelet transform not have the expected

positive impact [34].

After the introduction of deep learning, CNN-based classification methods have been widely applied. In 2016, Ribeiro et al. provided the first review on applying CNN to polyp classification [35]. They compared early CNN networks such as AlexNet [36] on the i-Scan database, and stated that pre-trained CNN features can be effectively used for the automatic classification of colorectal polyps. In 2020, another comparative study investigated the performance of VGG [37], ResNet [38], SENet [39], DenseNet [40], and MnasNet [41] on a group of public colon datasets. They stated that VGG-19 outperforms other models in both test cases, regardless of whether the background of the polyps is included in the cropped polyp image or not [42]. Recently, as ViT became popular, ViT-based models also showed promising outcomes. In 2023, Hossain et al. introduced Deeppoly, a cascaded model consisting of DoubleU-Net for polyp segmentation and ViT for classification. Their approach, tested on a mixed dataset, achieved an exceptionally high classification accuracy of 99.6% [43].

Despite the impressive classification accuracy reported by the ViT-based model in Hossain et al. on their customized dataset [43], it has not yet been evaluated on a public and pre-annotated dataset. To claim that ViT provides the best performance based on this tends to be reckless. A thorough comparative analysis of these methodologies is therefore necessary.

2.3 Deep-learning Based Image Segmentation

Image segmentation is one of the most widely studied problems in computer vision [44]. Its main task is to find groups of pixels that "go together" and assign each pixel to a different category or object. This gives the so-called "semantic" understanding of the image. Early Image segmentation methods were based on digital image processing techniques such as thresholding, histogram-based bundling, and region-growing. After machine learning was applied to computer vision, we utilized techniques like K-means clustering, Conditional random fields (CRF), and Markov random fields (MRF). Over the past few years, deep learning segmentation algorithms have become the main trend with better generalization ability and higher accuracy. Notably, the introduction of U-Net [16] in 2015 became a milestone in the field of medical segmentation.

2.3.1 Categories of image segmentation

Based on different segmentation results, image segmentation can be categorized into three types: semantic, instance, and panoptic. An example is shown in Figure 2.4.

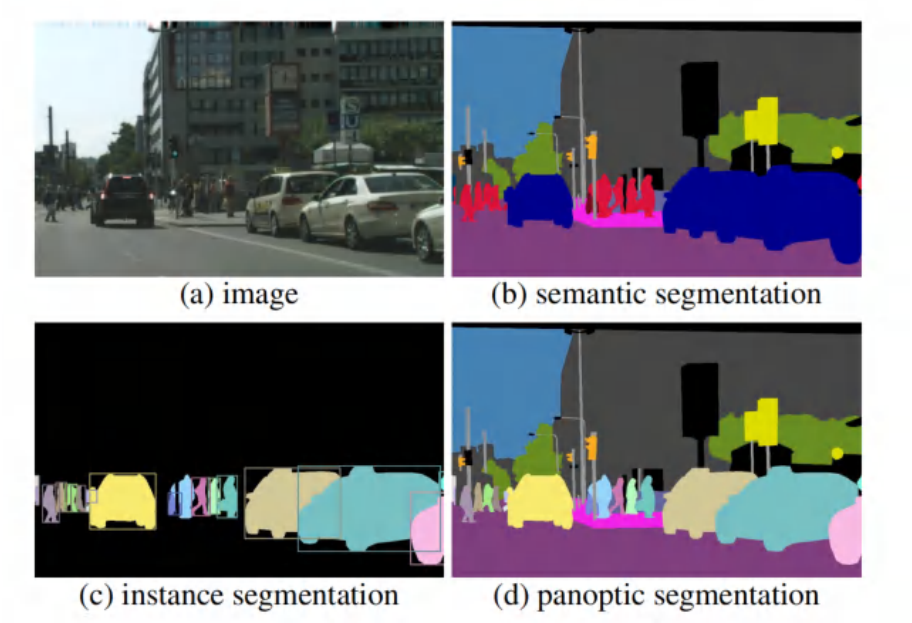


Figure 2.4: Examples of different segmentation tasks [45].

Semantic Segmentation is an approach "detecting, for every pixel, belonging class of the object" [46]. For example, in polyp segmentation, potential polyps are segmented as one object class, and the background is another object class. However, this approach does not provide information about which specific polyp each pixel belongs to. Therefore, semantic segmentation can also be considered a form of pixel-level classification. Currently, most polyp segmentation models are using semantic segmentation. This is mainly because it usually offers faster processing times, making it suitable for real-time applications in colonoscopy.

Instance Segmentation detects and segments each object instance. Using the example of polyp segmentation, instance segmentation further distinguishes different polyps, unlike semantic segmentation which treats all polyps as one class. The instance segmentation process can be seen as a combination of object detection and semantic segmentation. However, although instance segmentation provides more detailed information, it also requires more processing time. This trade-off is not always considered worthwhile in medical segmentation.

Panoptic Segmentation is proposed in 2019 [45]. It is a combination of semantic and instance segmentation which generate a comprehensive scene segmentation. It captures both class labels and

object instances. As a result, the computational complexity of panoptic segmentation is relatively higher compared to the other two segmentation methods, which makes it featured as a challenging track by COCO [47] and received high attention in community [45].

2.3.2 Evaluation metrics of semantic segmentation

We utilize various metrics to evaluate the performance of semantic segmentation algorithms. As semantic segmentation can be regarded as a pixel-level classification task, we need to first obtain the confusion matrices before further calculation.

Confusion matrices are commonly used to describe the outcome of each classification. Depending on the prediction and true label, the test result is classified into four categories: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Figure 2.5 shows a confusion matrix for a binary semantic segmentation job (e.g., allocating a pixel to the background or the ROI). We regard TPs and TNs to be correct forecasts, while FPs and FNs are incorrect. Using this, the pixel accuracy (PA) of each input image can be computed as follows:

$$\mathrm{PA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{2.1}$$

However, PA is not necessarily an appropriate statistic because it considers all pixel classifications equally. When background pixels account for a considerable proportion of the image, PA can be high, even if regions of interest are poorly predicted due to the high amount of TNs. Furthermore, PA may not accurately represent border detail prediction errors, which might be essential in real applications. To tackle this problem, we employ mean Intersection over Union (mIoU) and mean Dice coefficient (mDice) as evaluation metrics.

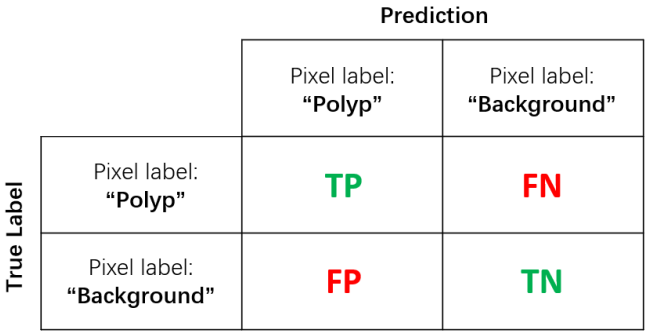


Figure 2.5: Confusion matrix for a binary semantic segmentation task: "Polyp" (positive) vs. "Background" (negative).

Intersection over Union (IoU) as its name indicates, is the ratio of the intersection to the union between the predicted results and the ground truth for a specific class. In other words, the intersection is the number of pixels in the overlapped area of the predicted mask and the ground truth (TP), while the union is the total number of distinct pixels in these two masks (TP+FP+FN). We suppose the prediction mask is X and the ground truth mask is Y, the IoU can be calculated as:

$$\text{IoU} = \frac{|X \cap Y|}{|X \cup Y|} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{2.2}$$

where $|\cdot|$ represents the number of pixels.

Dice coefficient focuses more on the overlapping area between the predicted and ground truth masks. It is calculated as:

$$\text{Dice} = \frac{2|X \cap Y|}{|X| + |Y|} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{2.3}$$

Compared to IoU, the Dice Coefficient doubles the intersection pixels, increasing the impact of these areas. This makes it better suited for smaller items and less frequent classes. To fully evaluate a segmentation model, we often employ both measures and average them across numerous classes. This provides us with mIoU and mDice.

2.4 Image Classification

Classification is the task in supervised learning that automatically assigns a discrete label to an unlabeled example [48], for instance, assigning a given colorectal polyp to the “hyperplastic” or “adenomatous” class.

2.4.1 Feature engineering

Image classification uses images as input and extracts features from these images. This process is automatically performed by deep learning models such as CNNs. In contrast, traditional machine learning algorithms require manual feature extraction. Feature engineering “transforms raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data” [49]. Although this manual process demands more prior knowledge in certain fields, it is still a common and effective method in medical image processing.

This will be introduced in Section 2.5. The choice of features simply sets the upper limit on the model's performance. However, such a limit is usually hard to achieve, and selecting the right classifier and optimizing feature selection is usually our main focus.

2.4.2 Traditional classifiers for supervised learning

A classifier assigns data to predefined categories or classes. In a classification task, the class of new instances is predicted based on patterns learned from labeled data. Support Vector Machine (SVM) and Random Forest (RF) are two commonly used classifiers in traditional machine learning.

Support Vector Machine (SVM) is a typical classifier to tackle two-group classification tasks. It conceptually maps input vectors non-linearly to a very high-dimension feature space [50]. The goal of SVM is to create a hyperplane during the training process which maximizes the space between itself and the nearest data points. This produces a decision boundary that separates the nearest samples from distinct classes by the greatest margin. Because of its strong generalization capabilities, SVM is a popular classifier for downstream classification in cascaded models.

Random Forests (RF) represents a collective of decision trees, where each tree is built using an independently sampled random vector. This vector maintains consistent distribution patterns across all trees within the forest structure [51]. A decision tree can be represented as a branching diagram resembling a flow chart. In this structure, each non-leaf node (internal node) represents an evaluation or a "test" of certain data characteristics. Random forest mixes many decision trees and employs a voting process to make decisions. This decreases the risk of overfitting and makes it appropriate for large-dimensional data. However, such a process is less interpretable.

2.4.3 Dimensional reduction and feature selection

Not all features contribute positively to classifier performance. Some features may offer little or even harmful contributions to distinguishing between classes while increasing the dimensionality of the feature space. This leads to the so-called "curse of dimensionality". Such features should be evaluated for integration or discarded to improve model performance.

Principal Component Analysis (PCA) is a dimensional reduction technique. It employs an orthogonal transformation to convert the original data into a new coordinate system with lower dimensions. Such transformation is designed to maximize the observed variance between data points in the new space. To perform PCA, the covariance matrix of the centered data is first

calculated. Then, Eigendecomposition is applied to this matrix to obtain eigenvalues and their corresponding eigenvectors. These eigenvectors are sorted based on their eigenvalues in descending order, becoming the principal components (PCs). Dimensionality reduction is achieved by retaining only the most significant PCs. At the same time, it preserves the majority of the data's variance. Instead of discarding original features like feature selectors which would introduced later, PCA integrates them into PCs. This effectively improves computational efficiency and may enhance classification performance in some situations. However, the PCs retained are less interpretable than the original features as they combine original features.

Sequential Feature Selection (SFS) is a wrapper-based feature selection approach for machine learning. It starts with an empty feature set and iteratively adds features to optimize model performance. In each stage, SFS examines all remaining features, choosing the one that improves performance the most, and adding it to the collection. The operation continues until it reaches a termination condition, such as achieving a predefined feature count or noticing minimal performance improvements. SFS captures feature interactions by evaluating the performance of feature subsets. It often outperforms filter techniques, which normally assess features separately. However, SFS increases computational cost owing to the frequent requirement for model retraining and can result in redundancy in the final subset because a feature cannot be removed once picked. To overcome these limitations, variants of SFS have been developed: Sequential Floating Forward Selection (SFFS), which allows for both feature addition and removal at each stage, and SFBS, which starts with the entire set of features and removes them instead of adding them from an empty set. These variants improve flexibility and efficiency in feature selection by allowing for dynamic modifications to the feature set.

2.4.4 Deep-learning based classification

The acquisition and selection of features are automatically performed by a deep learning network. A typical example is CNN, which involves three main types of layers: **Convolutional layers** use learnable kernels to extract various features from the input. **Pooling layers** reduce feature maps' spatial extent and lower computational demands. **Fully-connected layers** convert the 2D feature map into a 1D vector. The derived vectors can be fed into a certain number of classifications [52].

The training process consists of two main parts: In **forward propagation**, data is processed through convolutional layers, pooling layers, and other network components subsequently. It ends up with a prediction at the output layer. However, these predictions often deviate from the true

values. Loss functions are thereby used to quantify such deviations. **Back propagation** then uses the loss to update the weights and biases of each neuron to minimize the loss and improve the model's accuracy.

The CNN offers a robust solution for image classification tasks involving complex features. In recent years, the ViT has emerged as another popular choice for deep learning classification backbone. Its architecture will be discussed later.

2.4.5 Evaluation metrics of classification

The evaluation metrics for classification tasks are generally simpler compared to those used in semantic segmentation. Common metrics for classification are also derived from confusion matrices but focus on the accuracy of class labels rather than individual pixels. For each class, we look into three basic metrics: Accuracy, Precision, and Recall. Their calculation is shown below:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2.4}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.6}$$

It can be easily derived from the formula that Precision focuses on the accuracy of positive class predictions, while Recall on the ability to detect positive class instances. To consider both metrics to give a balanced evaluation of the model, the harmonic mean of precision and recall are calculated, as the F1 score:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.7}$$

The F1 score better reflects the model performance on the minority class, especially when given unbalanced datasets, as a result of balancing precision and recall. In contrast, accuracy is more direct and suitable for balanced datasets.

2.5 Radiomic Features of Medical Images

Radiomics is a novel technique that detects clinically relevant features from radiological imaging data, features that are often difficult for the human eye to perceive [53]. Although these features

are initially designed for radiological images, their feature classification can be applied to other fields, such as colorectal images. Radiomic features can be categorized into four parts, which are introduced below.

2.5.1 Statistical-based features

Statistical-based features are quantitative descriptors derived from the statistical properties of image data, like the intensity of each pixel. There are two main types of statistical-based features:

Histogram-based features: A histogram is a visual representation of the distribution of quantitative data [54]. Based on the global grayscale histogram, we can obtain the simplest feature descriptors, such as grayscale mean, variance, maximum, minimum, and percentiles. Since these features rely on single-pixel analysis, they are referred to as first-order features. Besides, more complex features like kurtosis and skewness can describe the shape of the intensity distribution. Other features include histogram entropy and energy, which provide insights into the texture and uniformity of the intensity distribution in the image.

Textural features: According to current research, textural features are among the most important imaging features for radiomics [53]. The texture of a lesion or tissue reveals details about its structure, composition, and arrangement. Haralick et al. proposed the **Gray Level Co-occurrence Matrix (GLCM)** technique for extracting representative textural features. in 1973 [55]. The principle of GLCM is simple: images' texture is caused by the repetition of gray-level patterns. This creates spatial associations between pixels at specific distances, representing gray-level correlations in the image. GLCM is a tool that investigates the relationship between pairs of pixels at a given distance. Haralick's paper proposes 14 statistical measures based on the GLCM: energy, entropy, contrast, homogeneity, correlation, variance, sum average, sum variance, sum entropy, difference variance, difference average, difference entropy, information measures of correlation, and maximal correlation coefficient [55]. This provides a diverse and thorough set of texture descriptors.

2.5.2 Model-based features

Model-based analysis focuses on analyzing spatial gray-level data to characterize objects or shapes. It involves creating parameterized models for texture generation, fitting them to the ROI, and utilizing the estimated parameters as radiomic features. However, because of their high computational

cost, complexity, and low interpretability, model-based features are less commonly used than other forms of radiomic characteristics. Autoregressive models are a common type of model-based technique.

2.5.3 Shape-based features

Shape-based features describe the geometry of the ROI directly, such as 2D and 3D diameters, axes lengths, and their ratios. In practice, these features are straightforward but require precise segmentation for accuracy. While shape-based features are valuable for characterizing morphology, they are often used with other radiomic features for comprehensive analysis.

2.5.4 Transform-based features

When the spatial and textural domain of an image cannot tell much information, transforming the image to a spectral domain would be a reasonable choice.

Wavelet Transform is a commonly used transform. The intensity and texture features of the original image are calculated through wavelet decomposition, which concentrates features within distinct frequency ranges inside the tumor volume. It gathers both frequency and spatial information concurrently, making it ideal for assessing textures and localized features in medical imaging.

Laplacian Transform is used to extract Laplacian of Gaussian (LoG) features. The LoG filter combines the Laplacian operator and a Gaussian filter to improve edges by emphasizing areas with substantial gray-level fluctuations. The filter's sigma parameter affects the level of texture detail that is highlighted: a lower sigma value emphasizes finer textures, while a larger sigma value emphasizes coarser textures. As a result, LoG features are recovered from areas with increasingly blurred texture patterns.

2.6 Transformer in Computer Vision

Transformer is a type of neural network that utilizes the self-attention mechanism [56] to extract intrinsic features and has shown significant potential in AI applications [57]. The Vision Transformer (ViT) is built on the Transformer architecture and is applied in the field of computer vision. The following paragraphs introduce the principles of ViT and its variation, Pyramid Vision Transformer (PVT) in different tasks.

2.6.1 Transformer architecture

The Transformer architecture employs an encoder-decoder structure instead of relying on recurrence or convolutions to generate output. Figure 2.6 illustrates the model structure, where the encoder and decoder serve as its core modules.

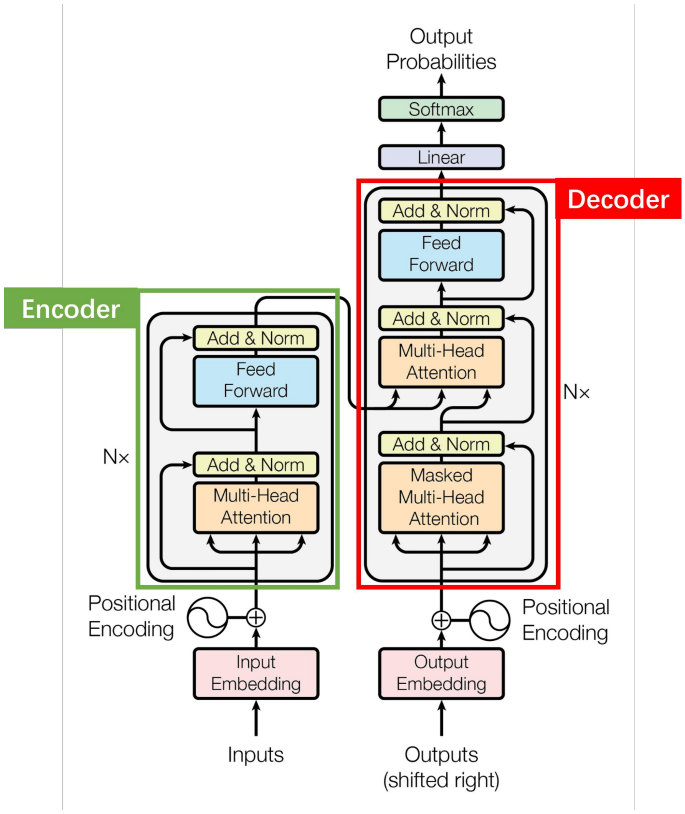


Figure 2.6: The encoder-decoder structure of Transformer [58].

Encoder: The encoder on the left side of the Transformer architecture converts the input sequence into a series of continuous representations. It has N layers, each with two sublayers: The **multi-head attention module**, based on the self-attention module [58], serves as the core of Transformer. This layer enables the model to focus on several parts of the input sequence at the same time by computing attention weights with multiple "heads". This allows the model to discover different types of relationships in the data. Another layer is a **fully connected feed-forward network**. It performs two linear transformations on the attention layer's output, with a ReLU activation in between. This brings nonlinearity into the model, thereby increasing its ability to learn complex functions. Continuous representations of the input sequence are then passed to the decoder.

Decoder: The decoder is located on the right side of the Transformer architecture. It generates

the output sequence using the encoder's representations and previously created outputs. It consists of N levels, each with three sublayers: The first sub-layer is a **masked multi-head self-attention mechanism**. It allows the model to focus on important portions of the previous output while avoiding future places. The second is a **multi-head attention layer** that makes the decoder capable of incorporating information from the input sequence. The third sublayer is a **feed-forward network**, just like the one used in the encoder. Normalization is applied to each sub-layer, including residual connections [59]. The output of the decoder is then subjected to a final linear transformation and softmax layer, which gives probabilities for the next token in the sequence.

The Transformer model starts with a sequence of tokens. It is initially transformed into embeddings and then coupled with positional encodings. The model then passes this information through the encoder and decoder. Finally, an output series of tokens is created, with each token generated based on the input and previously generated outputs.

Transformer architecture is frequently utilized in the field of Natural Language Processing. However, the proposal of ViT in 2020 gave rise to the use of transformers in computer vision.

2.6.2 Vision Transformer (ViT)

In a nutshell, ViT is a pure transformer architecture that processes images by converting them into vectors before feeding them into the transformer encoder. It retains most aspects of the original transformer architecture, treating input images similarly to sentences in NLP.

However, this approach introduces a challenge: converting 2D images into 1D vectors can result in extremely long vectors, especially for colored images with multiple channels. This significantly increases the model's complexity. To address this, ViT employs a key technique called patching. It splits the original images into smaller patches, treating each patch as a token instead of each pixel. The architecture of a ViT block is shown in Figure 2.7.

As seen in the figure, the task of a ViT block can separated into the following three tasks. We suppose the size of input images is $224 \times 224 \times 3$ and each patch has a fixed size of $16 \times 16 \times 3$.

Patch Embedding: Firstly, the input images are split into 196 patches, i.e. the length of the input sequence is 196. For each patch, the dimension is $16 \times 16 \times 3 = 768$ and is flattened to a 1D vector. This makes the final input to the linear projection layer 196×768. Until now, the problem has been transformed from a visual task into a sequence-to-sequence problem.

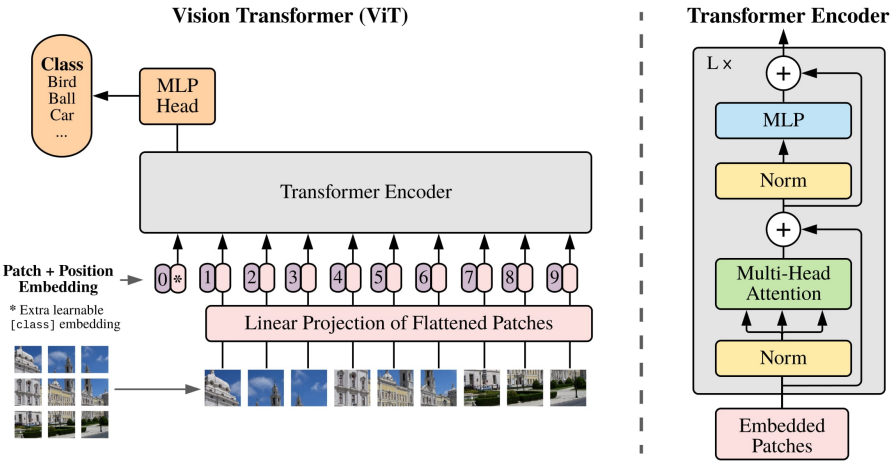


Figure 2.7: The architecture of Vision Transformer (ViT) [23].

Positional Encoding: In the original transformer model designed for NLP, positional embedding is necessary because tokens are input simultaneously into the encoder, lacking the inherent sequence order present in RNNs. Similarly, in ViT, just as each word in a sentence has a specific position, each flattened patch in an image must also be ordered to reflect its positional information within the original image. To achieve this, a positional embedding is generated for each patch and added to its corresponding patch embedding. Additionally, a special classification (cls) token is appended. This serves as a representation of the entire image, similar to the original BERT transformer. Since addition is used instead of concatenation, the input to the transformer encoder has dimensions of 197×768.

Transformer Encoder: The encoder in ViT has a few differences from the original version depicted in Figure 2.6. In ViT, normalization is applied after the multi-head self-attention mechanism and before the feed-forward network. This is a typical approach in Transformers, unlike in some NLP models where normalization might be applied differently. Additionally, padding is not required in ViT since each patch is of uniform size and the sequence length is fixed. The MLP (Multi-Layer Perceptron) serves as the feed-forward network. As the output dimension remains 197×768, multiple blocks can be stacked within the encoder. Finally, the output corresponding to the CLS token is used as the final output of the encoder, representing the overall image representation. An MLP head can then be attached for image classification tasks.

Dosovitskiy et al. stated in their paper that given a sufficiently large dataset for pre-training, the ViT is highly likely to outperform CNN. However, when the training dataset is not sufficiently large, ViT generally performs worse than similarly sized ResNets because transformers lack the inductive biases inherent to CNNs [23]. Therefore, in classification tasks, the choice between a

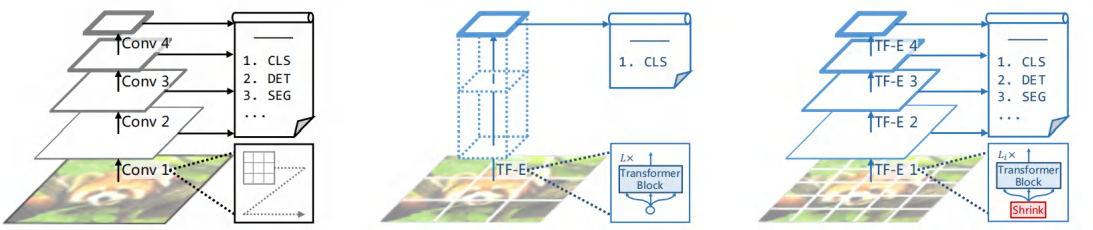
CNN-based model and a transformer-based model is generally influenced by the size of the training dataset.

2

2.6.3 Pyramid Vision Transformer (PVT)

The limitations of ViT are evident: Although it uses patch embedding to reduce the dimension of the input, the size of these patches remains unchanged as they are processed through multiple blocks in the encoder. This leads to high computational complexity. Additionally, the fixed patch size can hardly capture fine-grained spatial information. This makes ViT less suitable for pixel-level tasks such as detection and segmentation. To address these problems, Wang et al. drew inspiration from the feature pyramid (FPN) used in CNN-based networks and applied a pyramid architecture to the vision transformer [60].

The comparison of different architectures is shown in Figure 2.8. The pyramid structure in CNNs reduces the size of the feature maps through downsampling after each convolutional layer. This helps the model capture higher-level semantic information. Furthermore, feature maps from different stages can be fused through techniques such as concatenation or addition, which is known as feature fusion. PVT combines this concept with the patching approach used in ViT. As the network depth increases, PVT gradually reduces the sequence length in the Transformer, which significantly reduces computational complexity.



(a) CNNs: VGG, ResNet, *etc.* (b) Vision Transformer (c) Pyramid Vision Transformer

Figure 2.8: **Comparisons of different architectures**, where "Conv" and "TF-E" stand for "convolution" and "Transformer encoder", respectively. (a) The pyramid structure of many CNN-based models; (b) Vision transformer (ViT) without pyramid structure; (c) Pyramid vision transformer (PVT) [23].

To change the size of feature maps after each Transformer encoder layer, the PVT architecture is modified as shown in Figure 2.9. The PVT model is divided into four stages. Unlike ViT, which applies patch embedding and position embedding only once as a pre-processing step, PVT applies them at each stage, progressively splitting the input into smaller patches. The patch size is $4 \times 4 \times 3$ for stage 1 and $2 \times 2 \times 3$ for the subsequent stages.

We suppose the image input at stage 1 is $H \times W \times 3$, it is first split into $\frac{HW}{4^2}$ patches. These patches are then flattened and linearly projected to obtain embedded patches with a size of $\frac{HW}{4^2} \times C_1$. Here C_1 is determined by the size of patches, which equals $4^2 \times 3 = 48$. The embedded patches are then applied to positional encoding and fed into a transformer encoder. The output remains the same size as the input but is reshaped back to a $\frac{H}{4} \times \frac{W}{4} \times C_1$ 3D feature map. Similarly, the outputs of stages 2 to 4 have feature maps with strides of 8, 16, and 32, respectively. As the network deepens, the spatial resolution of the feature map is reduced while the number of channels increases. This provides richer high-level semantic information. These four feature maps form a feature pyramid that can be easily utilized for various downstream tasks, such as classification, detection, and segmentation.

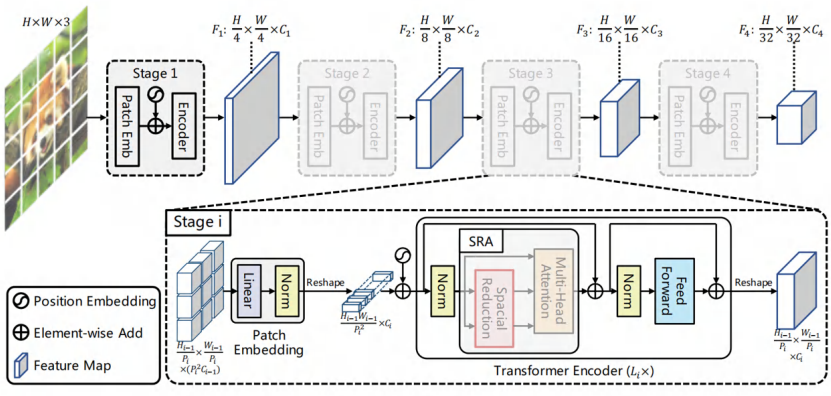


Figure 2.9: Overall architecture of Pyramid Vision Transformer (PVT) [23].

To process feature maps with high resolution, the PVT model develops a Spatial Reduction Attention (SRA) module to replace the multi-head attention module (MHA). The comparison between these two modules is shown in Figure 2.10. Similar to MHA, the proposed SRA also takes query Q, key K, and value V as inputs. The difference lies in that SRA reduces the height and width of K and V by a factor of R_i. In this way, the memory cost and computation cost in SRA reduces to $\frac{1}{R_i^2}$ of those in MHA, making it more efficient for processing larger feature maps with limited computational resources.

In general, the PVT model offers a more flexible and computationally efficient solution than the ViT model. It has demonstrated similar performance to ViT in classification tasks and can also be applied to detection and segmentation tasks. Moreover, numerous experiments have shown that PVT outperforms some well-designed CNN backbones. Although PVT cannot integrate certain modules specifically designed for CNNs, research on transformer-based models in computer vision has become the main trend, particularly in segmentation tasks.

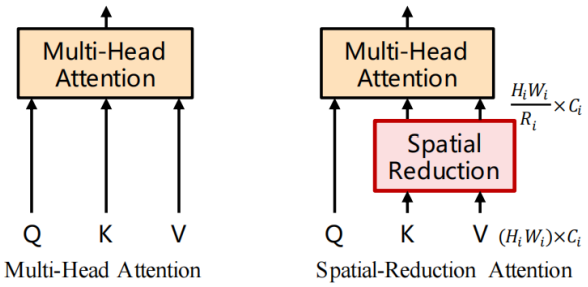


Figure 2.10: Multi-head attention (MHA) vs. spatial reduction attention (SRA). SRA introduces a spatial-reduction operation to reduce the computational and memory cost [23].

Polyp-PVT, proposed by Dong et al. [24] in 2021, is a PVT-based model designed for polyp segmentation. Its architecture is shown in Figure 2.11.

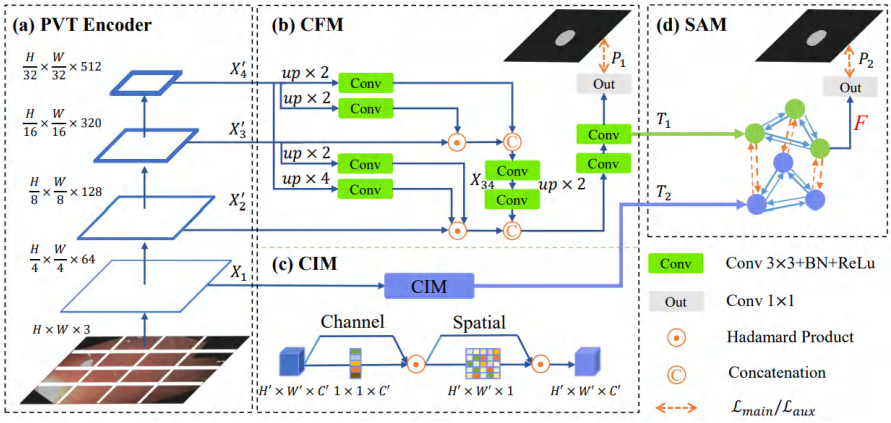


Figure 2.11: The architecture of Polyp-PVT [24].

As shown in the figure, the model integrates three components: the Cascaded Fusion Module (CFM), the Camouflage Identification Module (CIM), and the Similarity Aggregation Module (SAM). **CFM** combines high-level features through cascaded fusion, producing feature map T1. **CIM** extracts detailed polyp information from low-level features, such as texture and edges, resulting in feature map T2. **SAM** merges T1 and T2 using global self-attention to combine pixel-level and high-level semantic features of the polyp region. These modules capture polyp details from various dimensions, including texture, color, and edges, and use a global attention mechanism to integrate detailed appearance features with high-level semantic features. As a result, polyp-PVT is commonly used for polyp segmentation and similar lesions.

3 Concept

Contents

This chapter aims to identify promising methodologies by first conducting a thorough task analysis. This analysis provides a foundation for comparing and contrasting different methods and algorithms to determine the most effective approach. Based on the findings, new methodologies and metrics are designed if needed.

3.1 Task Analysis

This master thesis aims to develop a model to provide a suggestive diagnosis of colorectal polyps with two main goals. The first is to act as a supportive reference during colonoscopy inspection, providing prediction of the polyp border and category. The second is to automate the process of clinical image categorization after inspection. This gives three basic requirements when choosing the methodology:

Accuracy: The colonoscopist's role during inspection can be generally summarized as locating potential colorectal polyps and determining treatment methods based on their characteristics. Therefore, precise localization, margin delineation, and accurate classification are essential for a CADx model. In computer vision tasks, this can be interpreted differently depending on the type of model. For segmentation models, performance is often measured by a high Dice coefficient or IoU. For classification models, success is typically indicated by high accuracy or F1-score. These metrics need to be considered when choosing an algorithm. Additionally, advanced colonoscopy imaging techniques, such as NBI, can provide more detailed feature information about the lesions. This thereby potentially improves model performance.

Efficiency: Considering the time constraints during colonoscopy procedures and the need to reduce post-processing time on a large number of clinical images afterward, the model should operate with high computational speed. However, real-time diagnosis is not pursued in this master thesis due to hardware limitations. Also, the segmentation and classification performance is considered more crucial than processing speed. The trade-off between speed and accuracy is considered, with a focus on achieving high-quality results rather than real-time processing.

Interpretability: As introduced in subsection 2.2.1, AI-aided colonoscopy is not yet widely adopted in clinical practice. Its results typically serve as a supportive reference for colonoscopists' decisions. Therefore, enhancing the model's interpretability would significantly increase its reliability and credibility. This would also enable colonoscopists to double-check the results, thereby deciding to what extent they should trust the model.

Based on these requirements, we propose a cascaded model integrating two sequential tasks: polyp segmentation and classification. The subsequent sections will provide a comprehensive analysis of potential algorithms for both modules to determine the most suitable approach for each component of the proposed model.

3.2 Selection of Segmentation Algorithm

In subsection 2.2.3, different segmentation algorithms are introduced, which can be mainly categorized into CNN-based and Transformer-based models. To analyze their performance, these models are tested on four public datasets, which are Kvasir [20], ClinicDB [61], ColonDB [62], and ETIS. The comparison of these models is shown in Table 3.1.

It can be seen from the table that the transformer-based models Polyp-PVT and DUAT have

Table 3.1: Quantitative comparison of different methods on Kvasir, ClinicDB, ColonDB, and ETIS datasets. Polyp-PVT and DUAT are transformer-based models and achieve top-2 performance.

Algorithms	**Kvasir**		**ClinicDB**		**ColonDB**		**ETIS**	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
U-Net [16]	0.818	0.746	0.823	0.755	0.512	0.444	0.398	0.335
U-Net++ [17]	0.821	0.743	0.794	0.729	0.483	0.410	0.401	0.344
ResUNet++ [19]	0.813	0.793	0.796	0.796	N/A	N/A	N/A	N/A
PraNet [22]	0.898	0.840	0.899	0.849	0.712	0.640	0.628	0.567
MSNet [63]	0.907	0.862	0.921	0.879	0.755	0.678	0.719	0.664
SANet [21]	0.904	0.847	0.916	0.859	0.753	0.670	0.750	0.654
UACANet [64]	0.912	0.859	0.926	0.880	0.751	0.678	0.766	0.689
Polyp-PVT [24]	**0.917**	**0.864**	**0.937**	**0.889**	**0.808**	**0.727**	**0.787**	**0.706**
DUAT [25]	**0.924**	**0.876**	**0.948**	**0.906**	**0.819**	**0.737**	**0.822**	**0.746**

ranked 2nd and 1st respectively across all four datasets, outperforming the other CNN-based models. However, the code for DUAT is not open-source. Although Polyp-PVT is not the one that performs best in polyp segmentation, it remains one of the state-of-the-art algorithms. Therefore, this master thesis selects Polyp-PVT for addressing polyp segmentation.

3.3 Selection of Classification Algorithm

Unlike DL-based methods dominating today's segmentation tasks, traditional ML-based methods still have a place in polyp classification. In this section, both DL-based and ML-based algorithms are compared based on their accuracy. A detailed comparison is shown in Table 3.2.

It can be seen from the table that both ML-based and DL-based algorithms have reported high accuracy. However, unlike polyp segmentation, which has benchmark datasets, these proposed classification methods are tested on customized datasets with varying sources. Therefore, the reported classification accuracy serves only as a reference. It describes how well the models perform according to specific classification standards and imaging. Consequently, it does not provide clear evidence that one model outperforms others.

Among the DL algorithms, most researchers designed classification networks based on CNNs. However, The only study utilizing ViT, presented in Hossain's paper, achieved an accuracy of 99% on WLE images and outperformed the other models. For traditional ML algorithms, SVM classifiers are frequently used but often fail to achieve promising results. This is probably due to inadequate feature engineering done in their papers. In contrast, Fu's paper demonstrates that when over a hundred features are defined and an SFFS feature selector is applied, the SVM classifier achieves a considerable accuracy of 96% on WLE images.

Table 3.2: Comparison of classification algorithms: "Segmented Polyp" indicates whether images contain only the segmented polyp or include background. As all methods are tested on a combination of public and private datasets, the accuracy is just for reference.

Method	Algorithm	Data Imaging	Segmented Polyp (Yes/No)	Classes	Accuracy
Komeda et al. [65]	CNN	WLE & NBI	No	Adenomatous or Non-adenomatous	75.1%
Hsu et al. [66]	CNN	WLE & NBI	No	Neoplastic or Hyperplastic	82.8%
Ozawa et al. [67]	CNN	WLE & NBI	No	Hyperplastic or Adenomatous	83%
Tanwar et al. [68]	VGG-16	WLE & NBI	Yes	Benign, Non-malignant, or Malignant	84.1%
Hossain et al. [43]	ViT	NBI	No	Hyperplastic or Adenomatous	85.1%
Bour et al. [69]	ResNet50	WLE & NBI	Yes	Not Dangerous, Dangerous, or Cancer	87.1%
Krenzer et al. [70]	ResNet-18	WLE	No	Not Dangerous, Dangerous, or Cancer	89.35%
Ribeiro et al. [71]	CNN	i-Scan	No	Healthy or Abnormal	90.96%
Byrne et al. [72]	DCNN	NBI	No	Hyperplastic or Adenomatous	94%
Chung-Ming et al. [73]	AlexNet	WLE	No	Hyperplastic or Adenomatous	96.4%
Hossain et al. [43]	**ViT**	**WLE**	**No**	**Hyperplastic or Adenomatous**	**99%**
Chung-Ming et al. [73]	Ensemble Bagged Trees	WLE	No	Hyperplastic or Adenomatous	75.6%
Krenzer et al. [70]	SVM	NBI	No	Type 1 or Type 2 (NICE Classification)	81.34%
Zhang et al. [74]	SVM	WLE & NBI	No	Hyperplastic or Adenomatous	85.9%
Fu et al. [31]	**SFFS + SVM**	**WLE**	**Yes**	**Hyperplastic or Adenomatous**	**96%**

As can be seen from the "Segmented Polyp" column, most researchers chose not to exclude the background of the input polyp images. However, further investigation is needed to look into the impact of background exclusion. In DL methods, including background can provide more context but may introduce noise. On the other hand, excluding background focuses the model on the ROI and can potentially improve accuracy but may lose contextual information. In traditional ML methods, which rely on information such as textural features, excluding background may theoretically be a more advantageous approach. This could explain why Fu et al.'s model outperforms the other three ML models. Additionally, excluding background may require an extra step of segmentation; such computational cost should also be considered in practice.

Based on the previous literature review and information search, it is evident that both the traditional ML approach and DL approach have their respective strength and limitations. Table 3.3 provides an overview of how the two methods perform in fulfilling the task requirements and analyses are followed.

Table 3.3: Performance of DL and traditional ML approach in fulfilling task requirements.

Requirements	**Tradition ML Approach**	**DL Approach**
Accuracy	(−) Rely on feature engineering (−) Unsuitable for complex pattern (−) Weaker generalization ability (+) Less training data needed	(+) Automatic feature extraction (+) Suitable for complex pattern (+) Higher generalization ability (−) More training data needed
Efficiency	(+) Faster classification speed (−) Need time extracting feature (+) Less training time required	(−) Slower classification speed (+) Integrated Feature extraction (−) More training time required
Interpretability	(+) Highly interpretable features	(−) Less interpretable process

Note: (+) indicates the strength and (−) indicates the weakness. Evaluations are based simply on general rules and experience and the evaluation may not be reasonable in other scenarios.

The traditional ML method requires manual feature engineering. This demands specific domain knowledge and expertise. It is generally considered less capable of handling complex patterns and may have limited generalization to unseen data. Given the high accuracy requirement in clinical diagnosis, this method is not recommended unless the feature engineering is carefully designed. Despite these limitations, it offers faster processing speed and highly interpretable features, which are desirable aspects for the tasks in this master thesis.

The DL method enables automatic feature extraction. Its learning process is highly encapsulated, thus requiring a less in-depth understanding of polyp features and network architecture. Although it shows higher accuracy in polyp classification, the complexity of neural networks results in longer processing times for image classification. Additionally, the process is largely a "black box", from which limited information about the inferring process can be obtained.

Considering the analysis above, this master thesis employs traditional ML approaches in practice. This is because interpretability is considered more important and there is a lack of study and room for improvement in feature design for NBI images. However, the performance of DL and ML approaches will still be compared in this thesis based on the same dataset. For the DL algorithm, ViT is preferred, while for the traditional ML algorithm, we will evaluate various classifiers (such as SVM) with feature selectors (such as SFFS) to see if it improves model performance.

3.4 Feature Design for Traditional ML approach

To effectively utilize the traditional ML approach in classification, features that best differentiate between classes need to be carefully designed. The first step is to define classes based on an authentic polyp classification standard.

3.4.1 Class definition

In Section 2.1, NBI, a reliable colonoscopy imaging technique, is introduced. It offers a clearer visualization of the polyp's surface pattern, aiding colonoscopists in more accurate polyp classification. Correspondingly, the JNET classification standard has been established to complement this imaging technique (see Figure 2.2). It defines four types of polyps. In clinical practice, polyps of type 2B and type 3 are uncommon and can be easily identified from their appearance. Also, these types of polyps require pathology inspection for further surgery. Therefore, including these two classes in model training is not worthwhile. One reason is the limited number of NBI images available for these types, which would likely result in insufficient training in these classes. As a result, this could negatively impact the classification accuracy for the more common Type 1 and Type 2A polyps. Therefore, by just following recent research (see Table 3.2), two classes are defined: **Hyperplastic and Adenomatous**, corresponding to polyps of **Type 1 and Type 2A**, respectively, in the JNET classification.

3.4.2 Determinants of polyp categories

Colonoscopists categorize polyps based on their visual characteristics. In this master thesis, we describe these features as either **evidence-based** or **experience-based**. Evidence-based features are those formally described in standardized classification systems such as Paris[75], Pit-

pattern, and JNET. These features have been extensively studied and validated in clinical research. Experience-based features, on the other hand, are derived from the practical observations and expertise of colonoscopists. While these features often follow general patterns, they may lack authoritative validation linking them to specific polyp categories. Table 3.4 lists common evidence-based and experienced-based features that are used to distinguish hyperplastic polyps and adenomatous polyps. As a reference, Figure 3.1 gives examples of these two types of polyps.

Table 3.4: Five evidence-based and three experience-based polyp features of hyperplastic and adenomatous polyps.

Polyp Features	**Polyp Categories**		**Basis**
	Hyperplastic	**Adenomatous**	
Size*	Generally smaller	Generally larger	Paris
Shape	Most Regular (ellipse-like)	Partly Irregular	Paris
Elevation	Slightly elevated, even flat	Significantly elevated	Paris
Vessel visibility	Mostly Invisible **	Mostly Visible	JNET
Surface pattern	Regular dark or white spots	Tubular/branched/papillary	JNET
Texture uniformity	Mostly uniform	Partly uneven	Experience
Colour richness	Single and low-contrast	Multiple and high-contrast	Experience
Transition to mucosa	Relatively smooth changes	Relatively radical changes	Experience

Note:
* Without real-world scaling, polyp size cannot be precisely measured via images. However, adenomatous polyps are empirically larger in colorectal images.
** If visible, the caliber in the lesion is similar to surrounding normal mucosa.

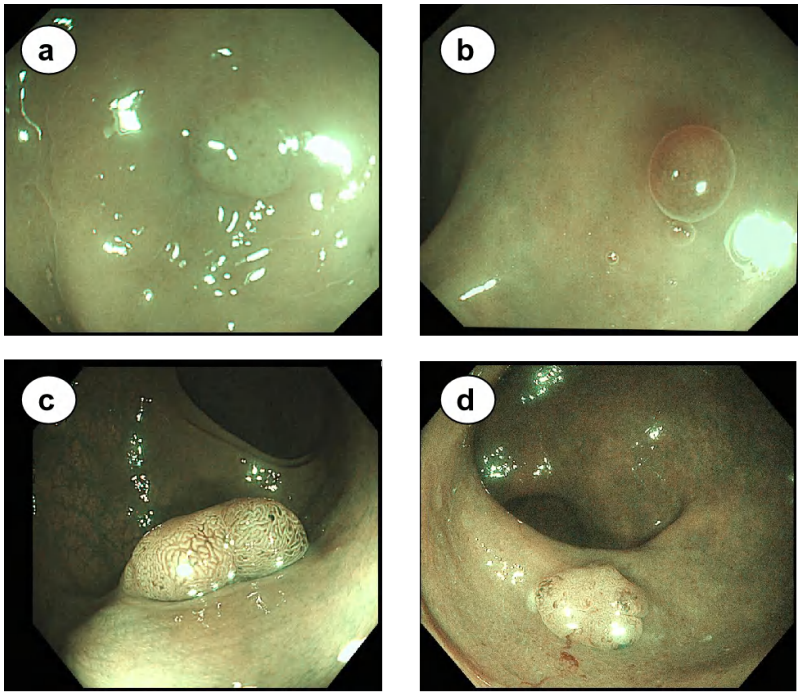


Figure 3.1: Example NBI images. **a** , **b** | Hyperplastic polyps; **c** , **d** | Adenomatous polyps.

Based on these features, colonoscopists can accurately classify polyps. However, translating these qualitative features into quantitative parameters that computers can process is a great chal-

lenge. Therefore, we employ advanced digital image processing techniques to extract and quantify these features from the original images. When necessary, novel metrics may need to be designed.

3.4.3 Edge-based feature design

Surface pattern is an important reference, especially in JNET classification as the surface is enhanced by NBI. To capture this feature, we employ edge detection techniques to highlight the surface features. The processing procedure is shown in Figure 3.2.

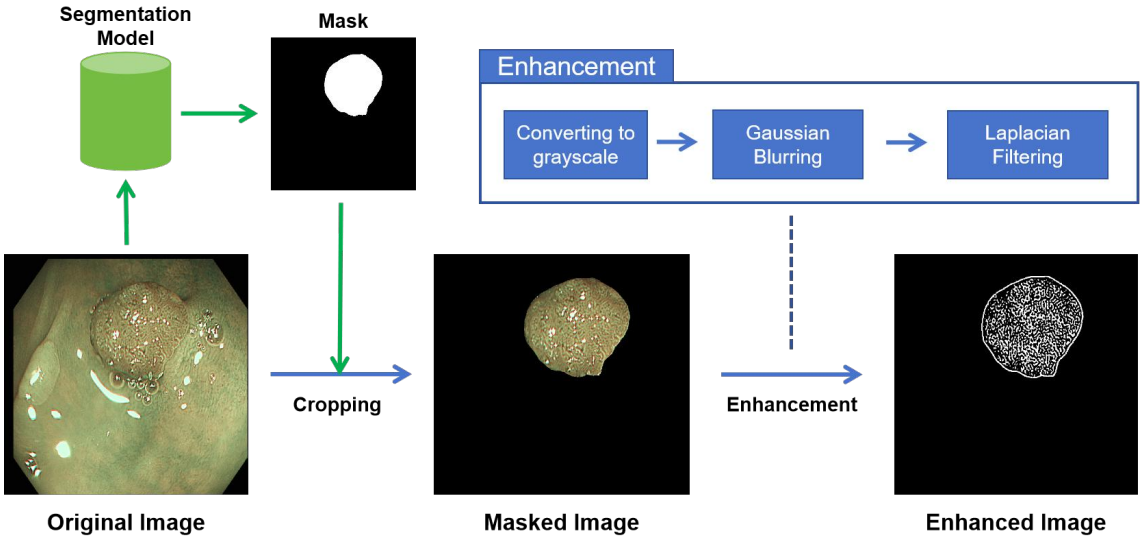


Figure 3.2: Image processing workflow to obtain an enhanced version that highlights vessel pattern.

The original image is first cropped according to the predicted mask to exclude background information. Image enhancement is then applied to improve edge detection, providing a clearer visualization of surface patterns. This enhanced version offers better differentiation between hyperplastic and adenomatous polyps.

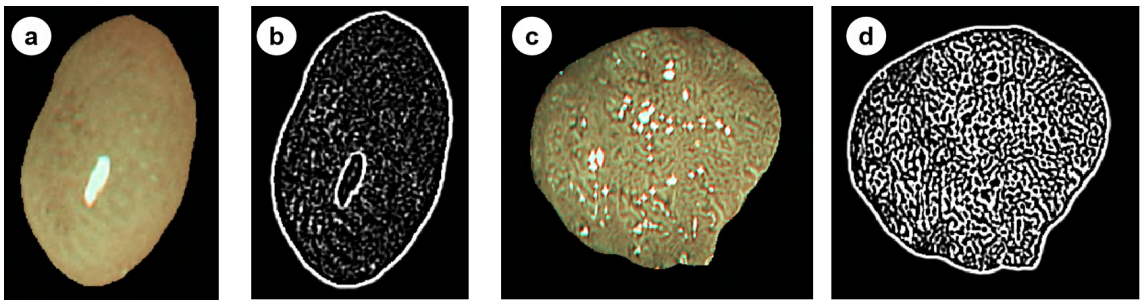


Figure 3.3: **a** , **b** | Example original image and its enhanced version of hyperplastic polyps; **c** , **d** | Example original image and its enhanced version of adenomatous polyps.

As demonstrated in the examples in Figure 3.3, the enhanced images of adenomatous polyps

generally show higher average pixel intensity and more segmented components than dot-like ones. To quantitate these characteristics, metrics are defined and subsequent steps are listed in Table 3.5 to obtain these metrics.

Table 3.5: **Above:** Steps following image enhancement to obtain edge-based metrics; **Below:** Definition of features and their meanings.

Step	Actions
1	Calculate **Edge Intensity** based on non-black pixels in the enhanced image.
2	Apply binary thresholding on the enhanced image.
3	To remove noise, apply median filtering.
4	Use `connectedComponentsWithStats()` in OpenCV to get all connected components.
5	Sort the connected components in descending order by their size (pixels).
6	Remove the largest 2% connected components (at least 3).
7	Remove the smallest half of connected components.
8	Calculate four statistic metrics: **CCS_mean**, **CCS_median**, **CCS_var**, **CCS_cv**
9	Calculate **Edge Density** based on the ratio of remaining connected components.

No.	Feature Name	Meaning
1	**Edge Intensity**	Average non-black pixel intensity, revealing surface vessel intensity
2	**CCS_mean**	Average of Connected Component Size (CCS)
3	**CCS_median**	Median of CCS
4	**CCS_var**	Variance of CCS
5	**CCS_cv**	Coefficient of variation (standard deviation over average) of CCS
6	**Edge Density**	Ratio of white pixels after binarization, revealing vessel density

Edge Intensity is calculated based on the average intensity of non-black pixels, which are edges that mainly represent vessels in the enhanced image. A higher intensity indicates more visible surface patterns, which are characteristic of adenomatous polyps. Let $I(x, y)$ represent the pixel intensity value at position (x,y) in the image, and S be the set of all non-zero pixels in the image. The equation can be expressed as:

$$\text{Edge Intensity} = \frac{1}{|S|} \sum_{(x,y)\in S} I(x, y) \tag{3.1}$$

Connected Component Size (CCS) depicts the size of each set of contiguous white pixels in an image. The enhanced images are first converted into a binary image and then median filtered to remove noise (Step 2&3). After Connected Components are obtained in Step 4 and are sorted based on size, the largest 2% are discarded to exclude the edges caused by cropping and specular highlights (see **b** in Figure 3.3). Besides, to better distinguish the two classes, the smallest half of connected components are also removed. The sizes of the remaining connected components are used to calculate four statistical parameters. These parameters are utilized to comprehensively evaluate the CCS feature, providing a multi-faceted assessment of the image texture and potential indicators of adenomatous polyps. Theoretically, adenomatous polyps tend to have larger average

and high variant CCS.

Edge Density is calculated based on the ROI of the binarized version of enhanced images. It simply represents the ratio of white pixels to the total number of pixels in the ROI. A higher edge density indicates richer content information of the polyp, which is more likely to be characteristic of adenomatous polyps. The equation can be expressed as:

$$\text{Edge Density} = \frac{\text{Number of white pixels in the ROI}}{\text{Total number of pixels in the ROI}} \tag{3.2}$$

3.4.4 Histogram-based feature design

In image processing, a histogram is a common technique for representing the distribution of pixel intensities within an image. For images with relatively uniform textures, the histogram is likely to show a narrow distribution, while those with rich or complex textures may have a wider distribution. Therefore, texture uniformity can be assessed by cropping a polyp image into several sub-images and calculating the differences between their histograms. Based on this concept, a feature extraction method has been designed, and its procedures are depicted in Figure 3.4 and Table 3.6.

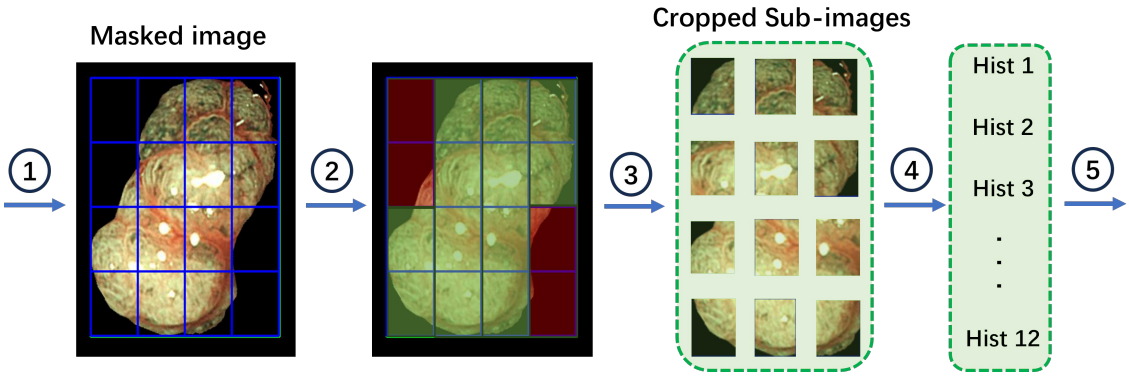


Figure 3.4: Procedures of extracting Histogram Distance (HD). Inputs are masked images and outputs are Histogram Distances between each pair of histograms. The numbers in circles correspond to steps in Table 3.6. Discarded sub-images are marked in red between Steps 2 and 3.

Table 3.6: Detailed steps to obtain the Histogram Distance (HD).

Step	Action
1	Input **the masked image** and calibrate it to ensure proper orientation.
2	Split the ROI into parts and discard those with more than 80% background pixels.
3	Crop the remaining parts and store them as sub-images.
4	Calculate the grayscale histogram of each sub-image.
5	Calculate the **Histogram Distances (HD)** between each histogram for further analysis.

The process begins with taking a masked image as input. The masked image is created using the original image and its corresponding binary mask. This masking technique sets the background pixel intensity to 0. It allows easy exclusion when computing histograms. The masked image is then divided into sub-images. However, some of these sub-images may contain insufficient information about the ROI. This could lead to a less representative histogram. They are therefore discarded, while the remaining sub-images are used to generate histograms. Finally, the **chi-square distance** is calculated between each pair of these histograms. The equation is described as:

$$D(H_i, H_j) = \sum_I \frac{(H_i(I) - H_j(I))^2}{H_i(I) + H_j(I)} \tag{3.3}$$

where H_i and H_j are two histograms compared and I represents bins of Histogram, normally ranging from 0 to 255. $H_i(I)$ represents the frequency of Ith bin in histogram H_i.

The larger the chi-square distance, the greater the difference between the two histograms. By calculating the distance across all histogram pairs, their statistical metrics can be obtained. Similar to CCS introduced in the previous subsection, four statistical metrics are considered for Histogram Distance, as described in Table 3.7.

Table 3.7: Definition of features based on Histogram Distance (HD).

No.	Feature Name	Meaning
7	**HD_mean**	Average of Histogram Distances (HD)
8	**HD_median**	Median of HD
9	**HD_var**	Variance of HD
10	**HD_cv**	Coefficient of variation of HD

Most hyperplastic and less severe adenomatous polyps tend to have a uniform surface pattern. This results in a small average and less variable HD. Although surface uniformity is not explicitly mentioned in the current classification standards, it remains a crucial empirical indicator. This feature may be especially useful in distinguishing polyps across a broad range of severities. However, since early cancer samples are not included in this master thesis, the performance of this feature may be limited.

3.4.5 Transition-based feature design

The transition between lesion and normal mucosa is another indicator. A more radical change is usually expected for adenomatous polyps. Additionally, this transition can somewhat reveal the elevation of the polyps, which is difficult to observe in a single image. In this master thesis, a specific score, namely the Transition Variability Score (TVS), is designed and calculated to

describe whether the transition is smooth or radical. To extract the TVS, the following procedure is performed, as illustrated in Figure 3.5 and detailed in Table 3.8.

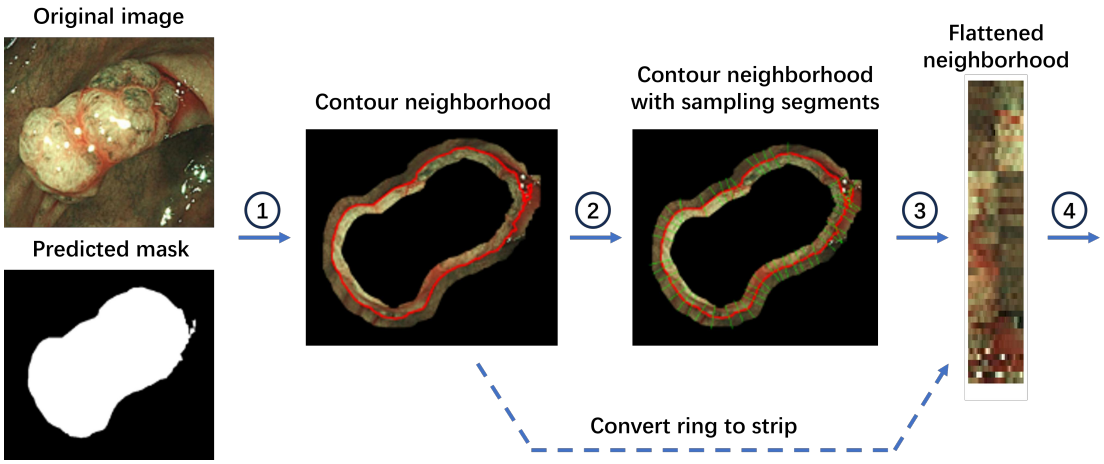


Figure 3.5: Procedures of extracting the Transition Variability Score (TVS). Sampling is conducted along the directions (green segments) perpendicular to the polyp border (red ring), and the TVS is calculated for each direction.

Table 3.8: Detailed steps to obtain the Transition Variability Score (TVS).

Step	Action
1	Find the polyp border and extend a neighborhood around it.
2	Select points on the contour and draw perpendicular segments.
3	Collect pixels along all segments and flatten to a strip image.
4	Calculate the Transition Variability Score (TVS) for each row in the strip image.
5	Discard the larger half of the scores and keep the others for further analysis.

The main idea of the workflow is to analyze the neighborhood region around the polyp border. This region roughly resembles a ring, where pixel intensity varies from the inside to the outside. To conduct a comprehensive analysis of these changes, we propose a method that involves first selecting random points along the contour and then sampling pixels along directions perpendicular to the contour at these points. The pixels collected from each perpendicular direction are used to form a row. By aggregating these rows from all directions, we create a strip image, as illustrated in Figure 3.5. For each row in the strip image, the TVS is calculated using the equation below:

$$\sigma_R = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(R_i - \mu_R)^2} \tag{3.4}$$

$$\sigma_G = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(G_i - \mu_G)^2} \tag{3.5}$$

$$\sigma_B = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(B_i - \mu_B)^2} \tag{3.6}$$

$$\text{TVS} = \sigma_R + \sigma_G + \sigma_B \tag{3.7}$$

where N denotes the number of pixels in each row, R_i, G_i, B_i represent pixel values of the red, green, and blue channels, respectively, for the ith pixel; μ_R, μ_G, μ_B and σ_R, σ_G, σ_B represent the mean and standard deviation of RGB pixel values, respectively.

A higher TVS indicates a more rapid change in one direction. To capture the overall characteristics in all directions, we calculate their metrics, which are shown in Table 3.9.

Table 3.9: definition of features based on Transition Variability Score (TVS).

No.	Feature Name	Meaning
11	**TVS_mean**	Average of Transition Variability Score (TVS)
12	**TVS_median**	Median of TVS
13	**TVS_var**	Variance of TVS

The coefficient of variation (CV) is not calculated for TVS. This is because in this case, the standard deviation of TVS is not directly related to the mean, making the corresponding CV redundant. Adenomatous polyps are usually elevated and have clear boundaries, thus expected to have larger average TVS. The variance of TVS, however, depends on the condition of the surrounding colon mucosa. When adenomatous polyps are so severe that they randomly invade the surrounding mucosa, a high variance in TVS is likely to occur.

3.4.6 GLCM-based textural features

The Gray-Level Co-occurrence Matrix (GLCM), introduced by Haralick et al. in 1973 [55], is a method for extracting textural features and is commonly used to analyze the texture of polyps. It reflects both the gray-level values and their spatial distribution in ROI. The GLCM is defined based on a specified direction and distance and is not generally used directly as a feature for texture classification due to its large size. Instead, various statistical measures are derived from the GLCM to serve as texture classification features. This master thesis selects five measures, which are **Contrast**, **Dissimilarity**, **Homogeneity**, **Energy**, and **Correlation**. We suppose L is the total number of gray levels in the image (or the size of each GLCM is $L \times L$); $P(i, j)$ is the joint probability of gray levels i and j in the GLCM. The equations of these measures are shown below:

$$\text{Contrast} = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} (i-j)^2 \cdot P(i,j) \tag{3.8}$$

$$\text{Dissimilarity} = \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} |i-j| \cdot P(i,j) \tag{3.9}$$

$$\text{Homogeneity} = \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} \frac{P(i,j)}{1+(i-j)^2} \tag{3.10}$$

$$\text{Energy} = \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} [P(i,j)]^2 \tag{3.11}$$

$$\text{Correlation} = \frac{\sum_{i=0}^{L-1}\sum_{j=0}^{L-1}(i-\mu_x)(j-\mu_y)\cdot P(i,j)}{\sigma_x \sigma_y} \tag{3.12}$$

where μ_x, μ_y and σ_x, σ_y represents the mean and standard deviations of $P(i,j)$ along x-axis and y-axis in GLCM, respectively.

Contrast and **Dissimilarity** both assess pixel value differences. Contrast focuses on local variations and Dissimilarity focuses on overall differences. These metrics typically increase with texture complexity. **Homogeneity**, conversely, reflects the uniformity of texture changes, often inversely related to contrast and dissimilarity. **Energy** indicates texture uniformity, with lower values suggesting more diverse textures. **Correlation** quantifies linear relationships between pixel pairs and a higher value indicates a finer and repetitive pattern.

GLCM assesses textures in various directions by setting angles to [0, 45, 90, 135] degrees, which correspond to horizontal, vertical, and diagonal orientations. As a result, it provides $5 \times 4 = 20$ features. This method is particularly effective for analyzing asymmetric textures. However, polyps generally have more symmetric textures, where the directional differences are less significant. To tackle this issue, averaging the GLCM features across these angles can be a reasonable approach to simplify the analysis. This reduces the number of features from 20 to 5, effectively lowering the dimensionality while still capturing the essential texture information. The five features are defined in Table 3.10.

Table 3.10: Definition of GLCM-based features.

No.	Feature Name	Meaning
14	**GLCM_contrast**	Average contrast along four directions of GLCM
15	**GLCM_dissimilarity**	Average dissimilarity along four directions of GLCM
16	**GLCM_homogeneity**	Average homogeneity along four directions of GLCM
17	**GLCM_energy**	Average energy along four directions of GLCM
18	**GLCM_correlation**	Average correlation along four directions of GLCM

Adenomatous polyps usually have deeper and more complex surface texture. In that case, we expect higher GLCM contrast and dissimilarity, along with lower GLCM energy and correlation. Regarding homogeneity, hyperplastic polyps generally exhibit uniform textures with high

homogeneity, while adenomatous polyps are more likely to have less uniform surfaces with low homogeneity.

3.4.7 Morphological features

Morphological features play a crucial role in polyp assessment. In earlier years, the Paris classification standard mainly relied on these features to differentiate between types of polyps. However, with developments in imaging techniques and textural analysis, the Paris classification has become less commonly used. Despite this, morphological features remain important in the evaluation of polyps. In this master thesis, features are designed based on the **color**, **shape**, and **size** of polyps.

Color: To analyze the color of a polyp, various imaging techniques may represent different surface colors. As a result, average pixel intensity may not be an effective feature for classification. Instead, we can analyze color variance to detect color changes, as adenomatous polyps tend to exhibit more color variation. In this approach, we calculate the pixel intensity Variance (VAR) and Energy Variance (EVAR). Given a masked image with M non-black pixels in the ROI, the equations can be expressed as follows:

$$\text{VAR} = \frac{1}{M}\sum_{m=1}^{M}(I_m - \text{AVG})^2 \tag{3.13}$$

$$\text{EVAR} = \frac{1}{M}\sum_{m=1}^{M}(I_m^2 - \text{EAVG})^2 \tag{3.14}$$

where I_m represents the intensity of the mth pixel and

$$\text{AVG} = \frac{1}{M}\sum_{m=1}^{M} I_m \tag{3.15}$$

$$\text{EAVG} = \frac{1}{M}\sum_{m=1}^{M} I_m^2 \tag{3.16}$$

Shape: Regarding shape characteristics, hyperplastic polyps typically exhibit circular or elliptical forms. If we consider these shapes as regular, we can measure the irregularity of polyps by drawing a circumscribed ellipse and comparing the shape difference between the polyp border and the ellipse. In this master thesis, we define a metric called Border-to-Ellipse Distance (BED). Let $B = (x_i, y_i)|i = 1, ..., n$ and $E = (u_j, v_j)|j = 1, ..., m$ be the sets of point locations on polyp border

and ellipse, respectively. For each point (x_i, y_i) in set B, we calculate:

$$BED_i = \min_j \sqrt{(x_i - u_j)^2 + (y_i - v_j)^2} \tag{3.17}$$

From this, we can form a set of BED values: $BED = BED_i | i = 1, ..., n$. We can then analyze the statistical metrics based on its elements. Similar to CCS and HD, we chose mean, median, variance, and coefficient of variation as features.

Besides analyzing irregularity, circularity is useful to describe the shape of a polyp. As circularity approaches 1, the polyp is more likely to be circular. Circularity is calculated as:

$$\text{Circularity} = \frac{\text{Length of the major axis of the circumscribed ellipse}}{\text{Length of the minor axis of the circumscribed ellipse}} \tag{3.18}$$

Size: Precise measurement of polyp size from images is challenging without real-world scaling. However, in most cases, adenomatous polyps are visually larger in colorectal images. To approximate size, we can analyze the proportion of white pixels in the binary mask image. This is calculated as:

$$\text{ROI\_proportion} = \frac{\text{Number of white pixels in the binary mask image}}{\text{Total number of pixels in the binary mask image}} \tag{3.19}$$

Based on the analysis above, we define eight features in this master thesis, which are shown in Table 3.11. These features have their limitations and most are empirically based. Therefore, further analysis is needed to decide whether they are useful in classification.

Table 3.11: Definition of morphological features.

No.	Feature Name	Meaning
19	**ROI_var**	Variance of pixel intensity in ROI
20	**ROI_evar**	Variance of pixel energy in ROI
21	**BED_mean**	Average of Border-to-Ellipse Distance (BED)
22	**BED_median**	Median of BED
23	**BED_var**	Variance of BED
24	**BED_cv**	Coefficient of variance of BED
25	**Circularity**	Circularity of the circumscribed ellipse of polyp
26	**ROI_proportion**	Proportion of white pixels in the binary mask image

4 Implementation

Contents

This chapter explains how to implement the chosen algorithms and designed features to complete a computer-aided colonoscopy diagnosis workflow. The Python project and dataset have been uploaded on GitHub[1].

4.1 Workflow Overview

This master thesis aims to design a processing procedure that inputs NBI colonoscopy images and outputs predicted polyp mask and belonged category. This involves implementing segmentation and classification modules subsequently. The whole workflow is shown in Figure 4.1.

Colonoscopy images are first acquired from clinical inspections or medical records in the hospital database. These images are then processed through a segmentation model to produce predicted

[1] https://github.com/Myosotis1111/Polyp-diagnosis-platform

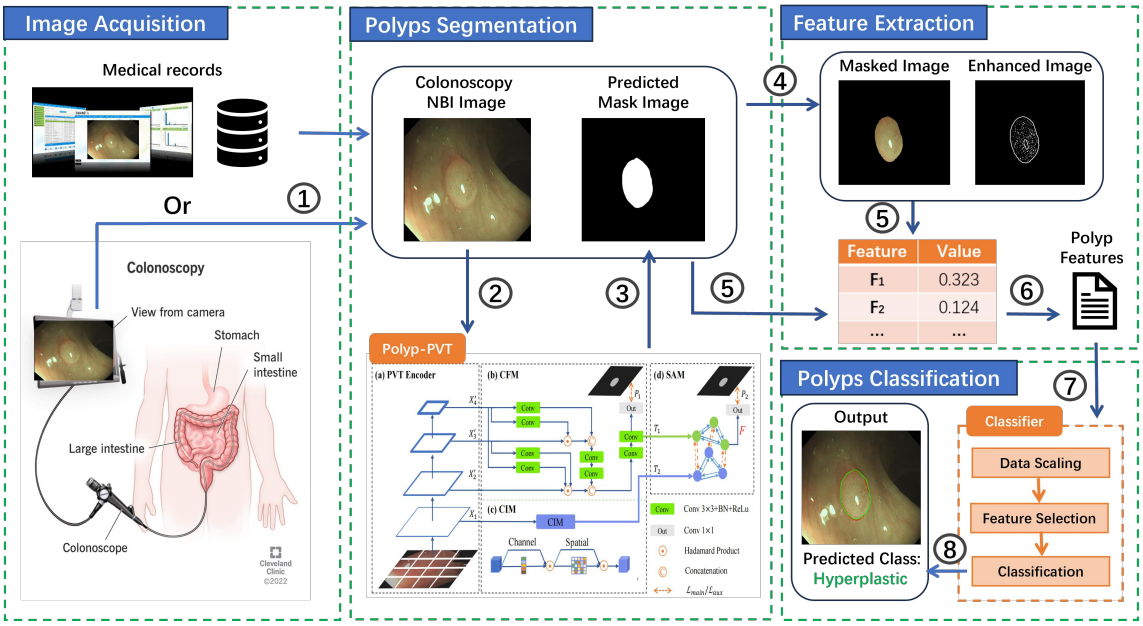


Figure 4.1: An overview of the workflow. Numbers in circles indicate step sequence. (Polyp-PVT architecture from [24]).

mask images that indicate their borders. Using digital image processing techniques like cropping and edge detection, we generate both a masked image and an enhanced version of the original image. Features are extracted from these four images and can be optionally stored in CSV or Excel files. Finally, these feature data are fed into a classifier to predict the final class. The output includes both the predicted mask and the category of the original image.

4.2 Data Management

4.2.1 Data collection

This master thesis collects images from the patient database of Shanghai Fourth People's Hospital Affiliated with Tongji University, China. All images have been de-identified and do not contain patient information or other sensitive content. The data collection flow is shown in Figure 4.2.

Unlike WLE images, which are commonly found in patient records regardless of whether polyps are detected, NBI images are only captured when at least one polyp is identified and require additional NBI techniques to characterize. Consequently, the number of available NBI images is limited, with only 525 images selected.

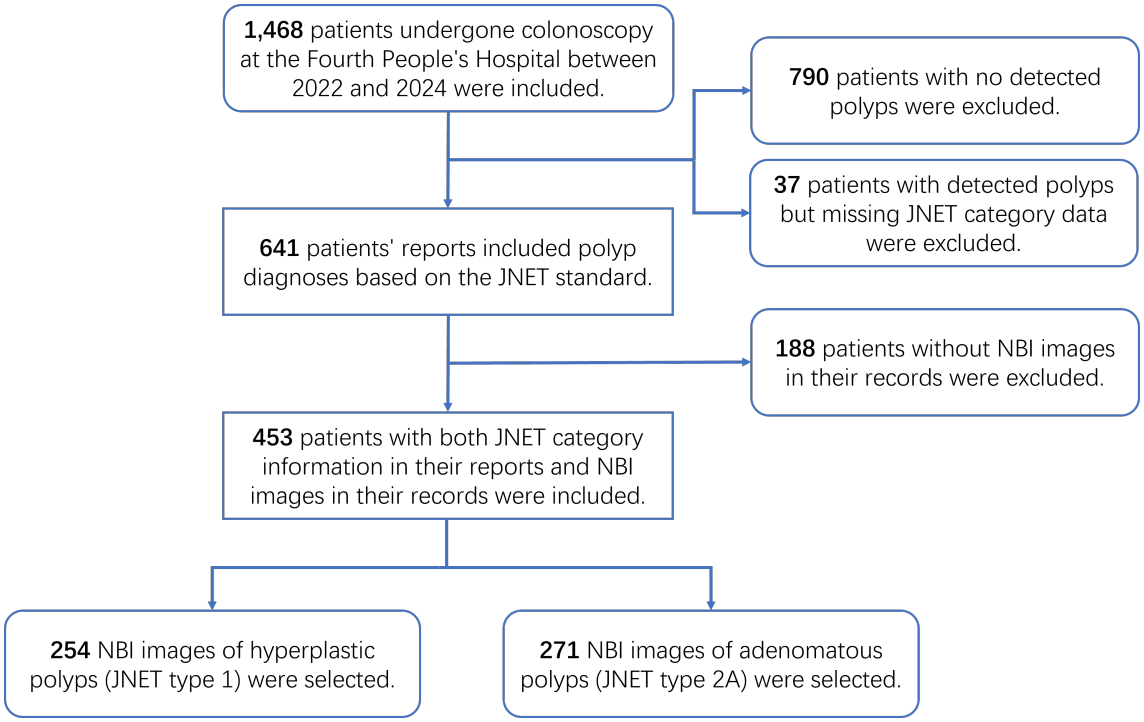


Figure 4.2: Data collection flow of NBI images with hyperplastic or adenomatous polyps.

4.2.2 Dataset setup

For model training and testing, 405 images were allocated for training and 120 images for testing (60 images per class). All the images are annotated under the guidance of a senior colonoscopist. To improve training quality, data augmentation techniques, including image flipping and rotation, were applied. Adjustments to saturation and exposure were not used. This is because they could potentially ruin the texture information of NBI images, which are carefully captured by colonoscopists. As a result, 806 images are generated, and a sum of 1211 NBI images are assigned for training.

Although public datasets with NBI images are unavailable due to issues with accessing such datasets, many datasets with WLE images can be accessed and involved in the segmentation model training. In this master thesis, the benchmark dataset defined by Fan et al.[22] was included to enhance model performance. The composition and split of the dataset used in this master thesis are summarized in Table 4.1.

Table 4.1: Composition of the training and test sets for the segmentation and classification model.

Source	Number of Images	Usage
Private Dataset	1211 (NBI)	Polyp-PVT and Classifier Training.
Kvasir-SEG[20]	1000	Polyp-PVT Training.
CVC-ClinicDB[61]	612	Polyp-PVT Training.
CVC-ColonDB	380	Polyp-PVT Training.
ETIS	196	Polyp-PVT Training.
CVC-300	60	Polyp-PVT Training.
Private Dataset	120 (NBI)	Polyp-PVT and Classifier Testing.
Training set Total		3459 (1211 for Classifier)
Test set Total		120

4.3 Model Training and Selection

In this master thesis, the segmentation model (Polyp-PVT) is trained on a CUDA device with an NVIDIA RTX 4090 GPU (24GB). The classification model is trained on an i5-12500H (2.50 GHz) CPU with a RAM of 16GB. Multiple versions of both the segmentation and classification models are obtained under different parameter settings. The following section provides an analysis of the performance of these models.

4.3.1 Segmentation models

The segmentation model is trained based on the pre-trained weight shared by Dong et al. in their project repository [24]. In this master thesis, four models are trained with different training settings shown in Table 4.2.

Table 4.2: Segmentation model with different training settings and Dice scores on the validation dataset. All models are trained for 100 epochs.

Model	Dataset	learning Rate	Decay Rate	Decay Epoch	Dice	Best Epoch
1	NBI only	1e-4	0.1	50	0.9035	89
2	ALL	1e-4	0.1	50	0.9127	82
3	ALL	1e-4	0.5	20	0.9053	91
4	ALL	2e-4	0.5	20	0.9116	89

As shown in the table, including public datasets improves the overall Dice score on the validation set compared to using only the private dataset, which consists of only NBI images. Adjusting the initial learning rate and its decay speed has little impact on model performance. Also, increasing the number of training epochs does not yield further improvements, as the optimal model weights are consistently achieved around 90 epochs. Beyond this point, there is little performance gain, with an increasing risk of overfitting instead. As shown in Figure 4.3, the validation loss starts to

see a slight change after 80 epochs, which indicates that 100 epochs is sufficient.

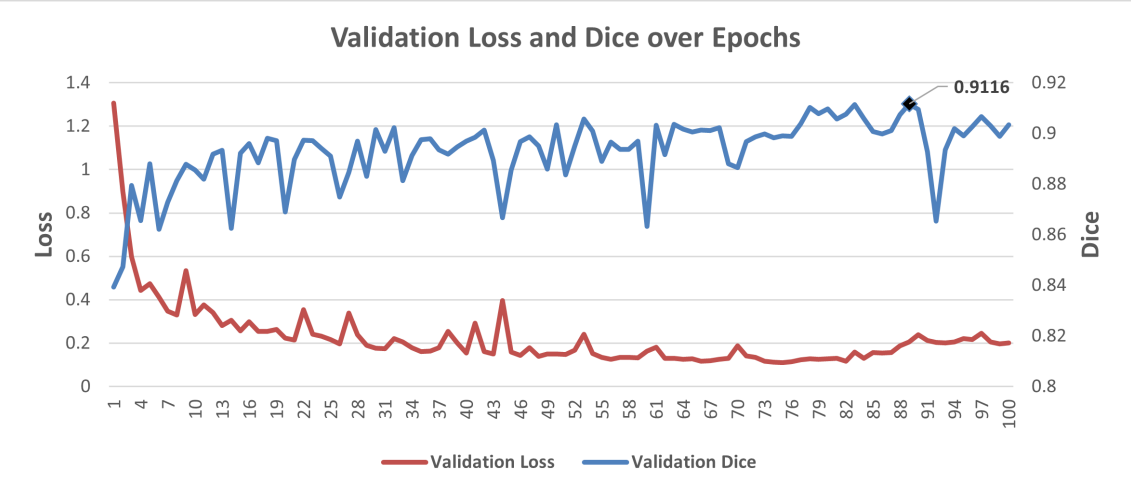


Figure 4.3: Validation Dice and loss of Model 4 over the 100 epochs.

All four models were evaluated on the test set. Their results are presented in Table 4.3. While Model 2 performs better in segmenting adenomatous polyps, Model 4 excels in segmenting hyperplastic polyps and has superior overall performance. These two models are further tested combining with classifiers in the subsequent subsections.

Table 4.3: Performance of segmentation models on the test dataset.

Metric(→) Model(↓)	Hyperplastic		Adenomatous		Average	
	IoU	Dice	IoU	Dice	IoU	Dice
1	0.7771	0.8425	0.8423	0.9004	0.8097	0.8714
2	0.7814	0.8502	**0.8449**	**0.9013**	0.8132	0.8758
3	0.7488	0.8248	0.8175	0.8800	0.7831	0.8524
4	**0.7938**	**0.8615**	0.8344	0.8907	**0.8141**	**0.8761**

4.3.2 Classification models

In this master thesis, several combinations of classifiers and feature selectors are tested to measure their usefulness in categorizing the dataset. We used features extracted from the training dataset to evaluate the model's classification accuracy on the test dataset. To assess the classifier's performance exclusively, the images in the test dataset are segmented using the ground truth mask. Figure 4.4 shows categorization accuracy for each model configuration.

As shown in Figure 4.4, selecting a SVM as the classifier yields better performance compared to a Random Forest (RF). Among the feature selectors, SFBS demonstrates the best performance with an overall accuracy of 93.3%. SFFS is slightly less effective but still shows improvement

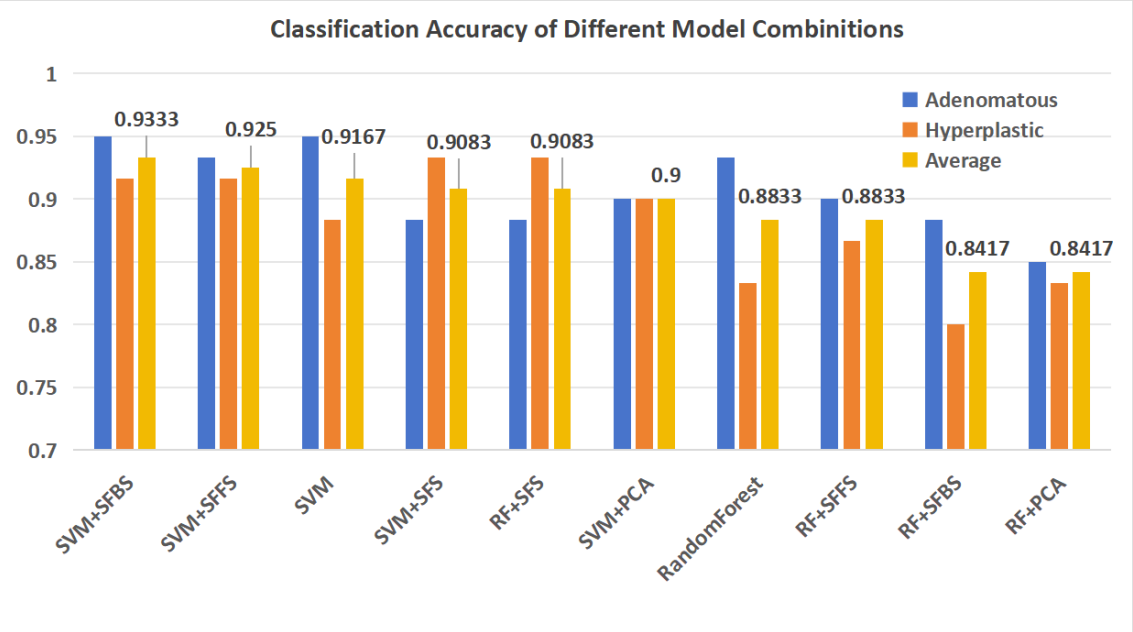


Figure 4.4: Classification accuracy of different model combinations (Classifiers combined with feature selector or PCA).

over using SVM without any feature selection. In contrast, SFS and PCA lead to a decrease in performance and are therefore not considered suitable for this task.

Notably, most models show higher prediction accuracy for adenomatous polyps compared to hyperplastic polyps. This indicates that the features associated with hyperplastic polyps are more challenging to capture. To tackle this problem, increasing the amount of training data for hyperplastic polyps would be beneficial. Also, it is important to focus on hyperplastic polyps that are visually more significant and could potentially be mistaken as adenomatous polyps.

Besides traditional ML methods, ViT models were also trained and their performance was evaluated. However, the results are not promising. As shown in Table 4.4 and Figure 4.5, the ViT model achieved an accuracy of 0.8906 by adjusting the learning rate and including background information from the poly images. However, the validation loss remained high and fluctuated throughout the training period. This indicates that the model converged early and failed to learn deeper patterns from the training set that could be generalized to unseen data.

To tackle this issue, increasing the number of training images could be a more effective solution for improving performance. However, given the limited size of the dataset in this study, the ViT model is not suitable as a classifier in this context. Therefore, we choose traditional ML-based classifiers for higher accuracy and interpretability.

Table 4.4: Training performance of ViT models with different settings. All models are trained for 200 epochs.

Model	Learning rate	Segmented Polyp (Yes/No)	Validation Accuracy	Best Epoch
1	3e-5	Yes	0.7997	112
2	3e-5	No	0.7997	114
3	1e-4	Yes	0.8447	184
4	1e-4	No	0.8598	166
5	3e-4	No	**0.8906**	146

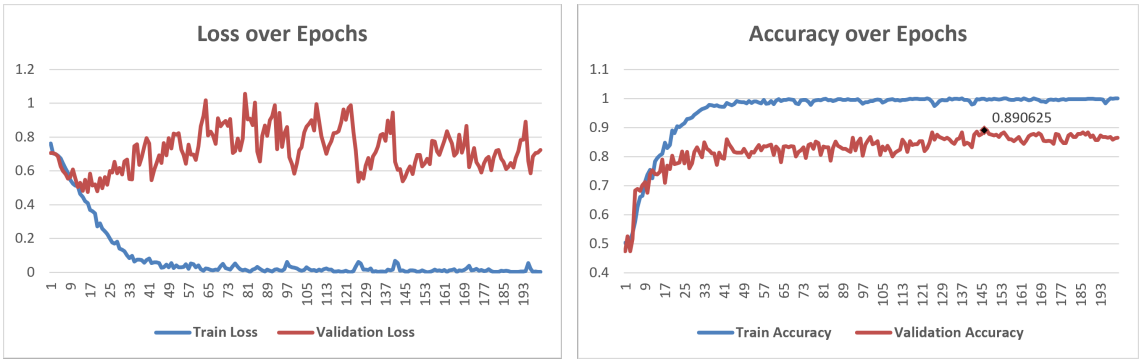


Figure 4.5: Training and validation performance of Model 5 over the 200 epochs.

4.3.3 Performance of combining segmentation and classification models

Based on the evaluation of the individual performance of the segmentation and classification models described above, this subsection explores the performance of combining these two modules. This means features for classification are extracted from masked images predicted by the segmentation models rather than the ground truth. We use Model 2 and Model 4 (see Table 4.2 and 4.3), which have higher Dice scores on the validation and test sets, respectively, as the segmentation models. For the classification model, we selected SVM+SFBS, SVM+SFFS, and SVM, which achieved the top three accuracies with the ground truth masked images. The performance of these model combinations is compared in Figure 4.6.

As observed from the comparison, after segmenting the images with either model, the classification accuracy generally decreases for SVM models with feature selectors (SFFS or SFBS). In contrast, for the model without feature selectors, the performance decreases only slightly. For Segmentation Model 4, the accuracy remains the same, achieving the highest accuracy of 0.9167. However, SVM+SFBS exhibits greater stability in classification accuracy across different classes. This suggests that the segmentation quality may have less impact on the predictions when using SVM with SFBS. Considering all these factors, SVM+SFBS is selected as the desired model, despite its slightly lower accuracy on the test dataset.

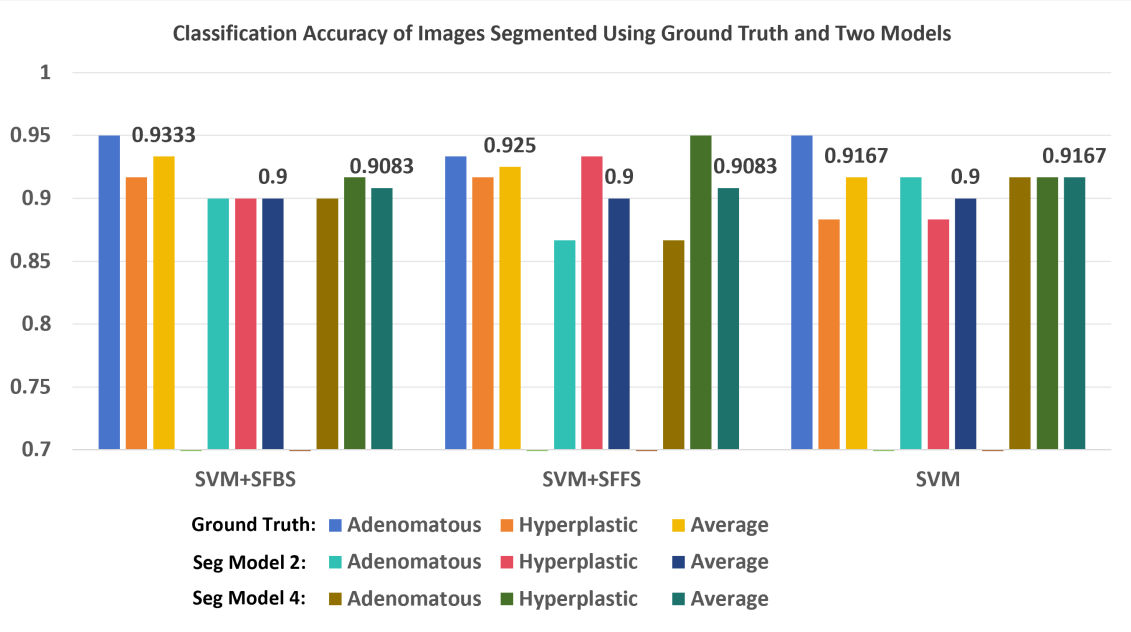


Figure 4.6: Comparison of classification accuracy on different combinations of selected segmentation and classification models.

Compared to Segmentation Model 2, Model 4 demonstrates higher accuracy on the test set across all three classifiers. Additionally, as Model 4 achieves a higher Dice score and IoU for segmenting hyperplastic polyps (see Table 4.3), it achieves improved accuracy for this class compared to Model 2. Despite a weaker performance on the other class, the accuracy for adenomatous polyps remains consistent with Model 2. Given its superior overall accuracy and the ability to better segment hyperplastic polyps, Model 4 is considered a better choice for the segmentation module.

Interestingly, after applying segmentation, the accuracy for the hyperplastic class is even higher than when using the ground truth masked images. This could be because the ROI identified by the segmentation model may sometimes include normal mucosa that has features similar to hyperplastic polyps. This makes the masked images look more likely to be hyperplastic, which inadvertently corrects those wrong predictions. To solve this problem, the boundary between hyperplastic and adenomatous polyps should be made clearer, and a better segmentation model is also desired.

In general, this master thesis selects SVM+SFBS as the preferred classification model, combined with an upstream Polyp-PVT model that has a Dice score of 0.9116 on the validation set. The cascaded model achieves a Dice score of 0.8761 and an accuracy of 90.83% on the test dataset.

4.4 Model Implementation

To enable an end-to-end implementation of the diagnosis process, a polyp diagnosis platform is designed based on Python Qt, and the cascaded model is applied to produce the border and category of the input polyp images. The Graphical User Interface (GUI) of the software is shown in Figure 4.7 and its main function is introduced in the following subsections.

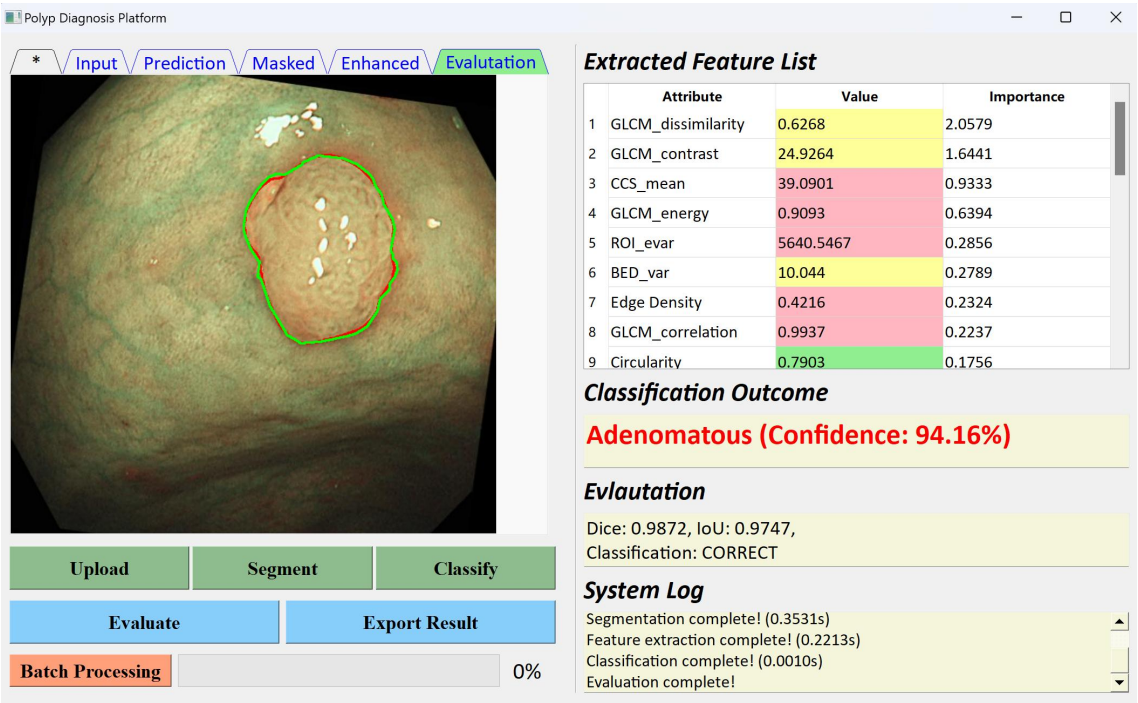


Figure 4.7: The GUI of the Polyp Diagnosis Platform.

4.4.1 Single image analysis

This GUI can provide detailed information for the image uploaded for analysis. The user's activity diagram and corresponding responses of the platform are shown in Figure 4.8.

Users first press the "Upload" button to upload medical images with various file types supported. This includes DICOM and BMP formats, which are commonly used in medical imaging. The platform will convert the image to PNG format and display it in the tab widget. Then, users can press the "Segment" button to perform image segmentation. After the segmentation is completed, the predicted mask and the processed masked image will be displayed. Finally, users can press the "Classify" button to extract and display the features in descending order of importance, as determined by the loaded classification model. Besides, the background color of each value cell

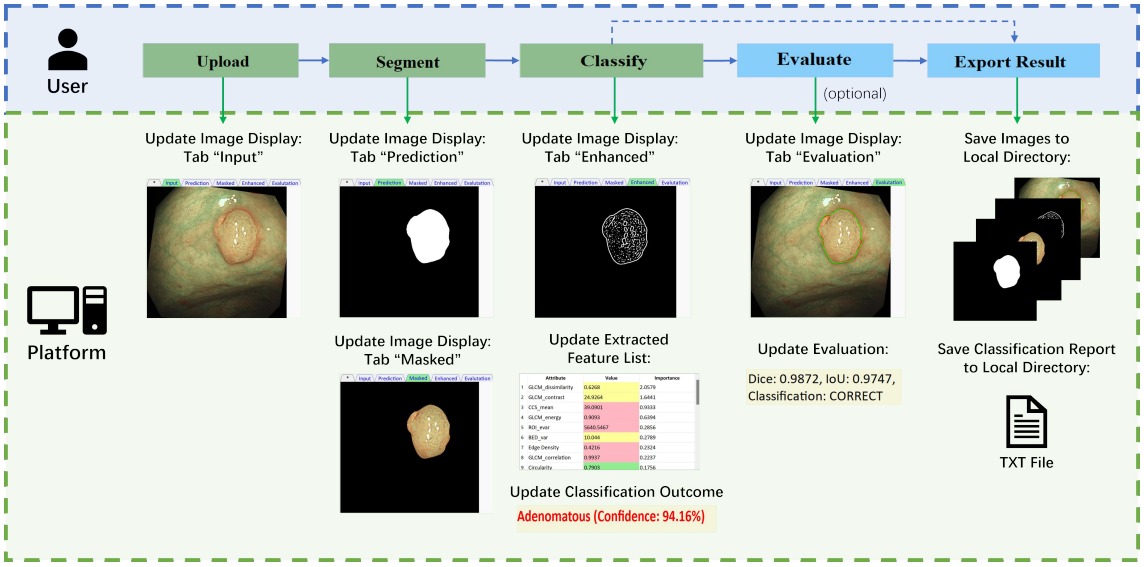


Figure 4.8: User actions and corresponding platform responses.

indicates the range in which certain feature value lies. Specifically, yellow indicates that the value is between the mean values of the hyperplastic and adenomatous classes; Red indicates the feature value is above the mean value of the adenomatous class; Green indicates it is below the mean value of the hyperplastic class. This serves as a helpful reference to determine whether a certain feature is more characteristic of an adenomatous polyp or a hyperplastic polyp.

After the diagnosis, users can choose to evaluate the outcome by pressing the "Evaluate" button if they have ground-truth data available. The platform will display a comparison by drawing the predicted border in red and the ground truth border in green. Additionally, the classification and segmentation metrics will be shown. These intermediate images and classification results can be exported to a local directory by pressing the "Export Result" button.

4.4.2 Batch processing

Users can also press the "Batch Processing" button to process all original polyp images within a specified directory. The platform will generate a predicted mask for each image and save them in a separate folder. Additionally, it will create an Excel report containing all the extracted feature values and classification outcomes for each image. In the current hardware environment, using only the CPU, the processing time is approximately half a second per image. This offers a more efficient solution for colonoscopists who need to process multiple images quickly.

5

Evaluation

Contents

This chapter evaluates the impact of feature engineering on the classification of polyps. The features designed for this master thesis will be analyzed, and suggestions for feature optimization will be made. Table 5.1 categorizes the 26 features into five groups for ease of understanding. Additionally, cases of incorrect predictions using the current model will be examined to assess segmentation and classification performance.

Table 5.1: Grouping of 26 features.

Group No.	Group Names	Number of Features	Detailed Information
1	Edge-based Features	6 (No.1 $\sim$ 6)	Table 3.5
2	Histogram-based Features	4 (No.7 $\sim$ 10)	Table 3.7
3	Transition-based Features	3 (No.11 $\sim$ 13)	Table 3.9
4	GLCM-based Features	5 (No.14 $\sim$ 18)	Table 3.10
5	Morphological features	8 (No.19 $\sim$ 26)	Table 3.11

5.1 Evaluation of Feature Design

To evaluate the quality of feature design for the task in this master thesis, we analyze the features in two ways. Firstly, we assess the discrimination ability of the features between the two target classes. Second, we evaluate the robustness or invariance of the features against changes in segmentation.

5.1.1 Distribution of feature values

The distribution of data plays an important role in splitting different classes, which shows the discriminability of a feature. A standard scaler is first applied to the raw data to perform normalization. The box plot of data used for training is shown in Figure 5.1.

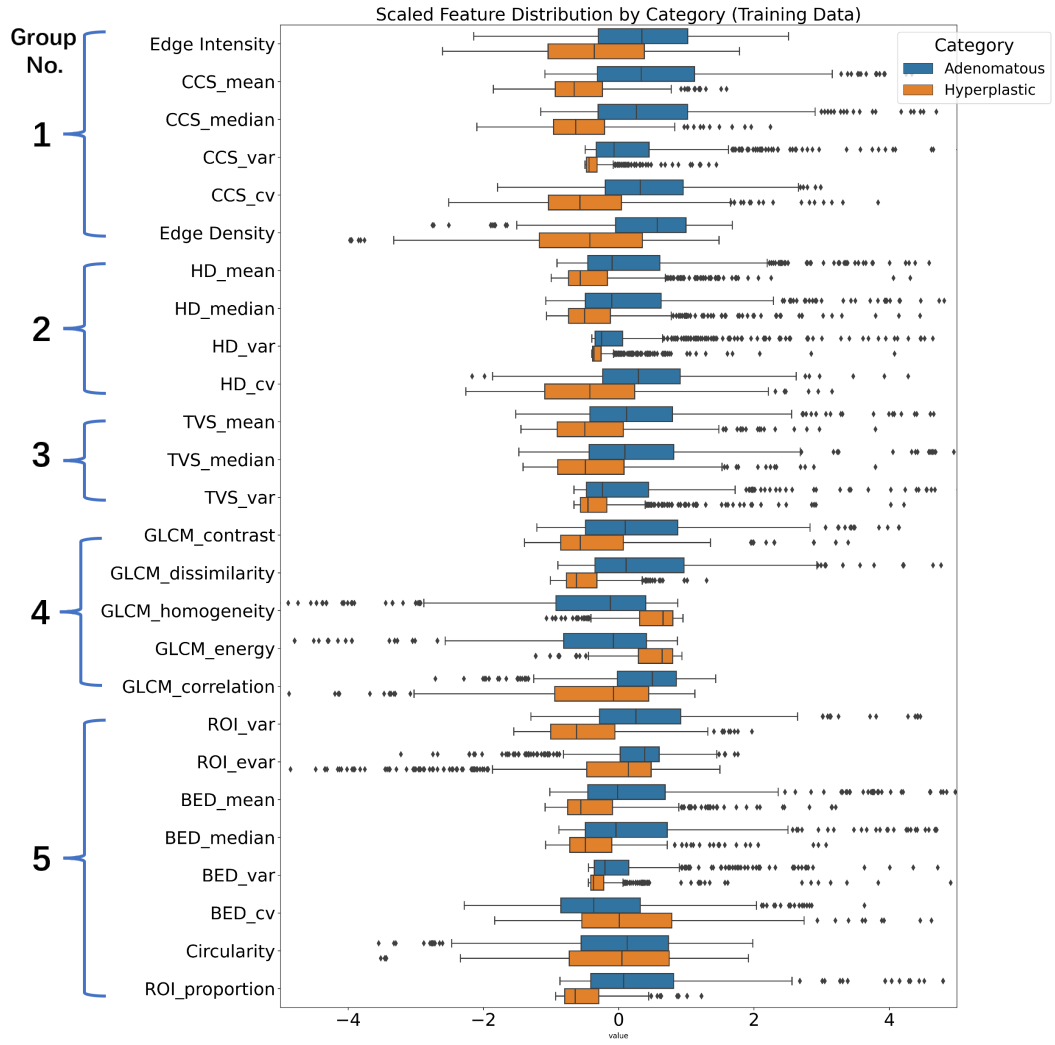


Figure 5.1: Comparative box plot analysis of feature value distributions across two classes.

The colored box pairs (blue for the “Adenomatous” class and orange for the “Hyperplastic”

class) in the box plot represent the 25th and 75th percentiles of the feature values. The degree of overlap between the boxes in each pair indicates the extent of value difference between the two classes. Smaller overlap suggests a better ability to separate these two classes.

Except for 4 features (GLCM_homogeneity, GLCM_energy, Circularity, and BED_cv), the other features generally had higher values for the Adenomatous class, as expected. For GLCM_homogeneity and GLCM_energy, these features should have higher values for the Hyperplastic class, which is also a reasonable observation. For the Circularity feature, the distribution is similar between the two classes. This indicates that it may not be as effective at separating the Adenomatous and Hyperplastic polyps. However, the BED_cv feature shows a different distribution compared to the other BED statistical metrics. This suggests a slightly more irregularity in the border of the hyperplastic polyps when polyp size is not considered, which is an unexpected result that goes against prior experience.

Notably, Groups 1, 2, and 4 show a generally better separation ability of the data. They are thus expected to have higher importance for the classification task. However, the other two groups have features that cannot split the data well, which requires further investigation.

5.1.2 Change of feature values due to segmentation

Classification is a downstream task in polyp diagnosis. The input image for feature extraction is first processed by the segmentation model. This makes the ROI vary to different extents compared to the ground truth. Such variation may affect the feature values and further affect the classification results. Therefore, it is crucial to design features that are robust and invariant to changes in the ROI to ensure stable classification performance across different segmentation outcomes.

To analyze how features change due to segmentation, we compared scaled feature values obtained from two different versions of input images: one cropped using the ground truth mask and the other segmented using the mask predicted by the Polyp-PVT model. For each sample, the feature change rate was calculated using the equation:

$$\textbf{Change Rate} = |\frac{v_2 - v_1}{v_1}| \tag{5.1}$$

where v_1 is the ground truth feature value and v_2 is the feature value obtained after segmenting by the model. We then computed the median and standard deviation of the change rates across all samples. The results are presented in Figure 5.2.

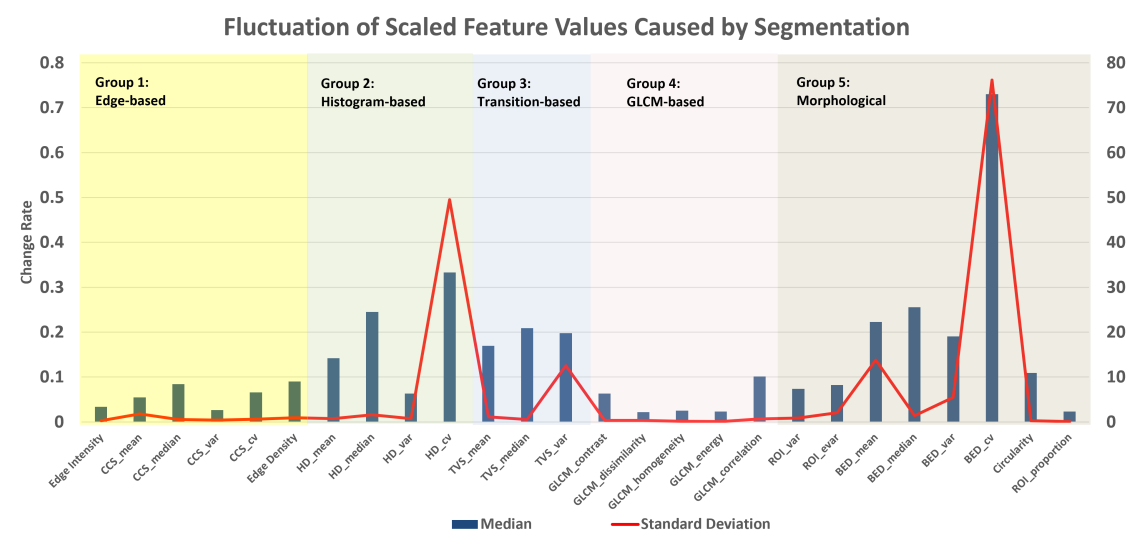


Figure 5.2: Change of scaled feature values obtained from masked images segmented using the proposed Polyp-PVT.

As seen from the figure, Groups 1 and 4 show a low median change rate of less than 0.1. Besides, their standard deviations are also low. This indicates little fluctuation and fewer outliers. It suggests that these two groups of features are robust against segmentation changes and are preferable for feature selection. Such a pattern is expected because these groups primarily involve textural features, which are less affected by changes in the ROI unless a significant number of background pixels are involved.

In contrast, Group 3 and some features in Group 5, such as BED, are highly sensitive to changes in the border of ROI. This results in higher change rates. The high change rate observed in Group 2 is also unexpected. This may be due to the changes in histograms obtained from sub-images near the polyp border. To address this issue, a stricter threshold could be applied to exclude sub-images with excessive background pixels, thereby reducing the impact of changes in the polyp border.

In conclusion, to ensure the stability of feature extraction, we should prioritize preserving Feature Groups 1 and 4. Future optimizations should focus on designing new features based on these two groups. Conversely, features with high change rates and standard deviations, such as HD_cv and BED_cv, should be excluded from the feature list due to their potential detrimental effects on the classification. Although the remaining features are still important, their calculation methods should be redesigned to enhance robustness against segmentation impacts.

5.2 Feature Importance

In addition to examining the pattern and robustness of the training data, feature importance is another clear and reliable indicator of measuring how much a feature contributes to the classification process. Features with high importance are generally more decisive in determining classification results. In contrast, features with low importance or those discarded by feature selectors from the feature list may be less significant or even detrimental to classification performance. In this master thesis, several combinations of classifiers and selectors are utilized to find the best classification models, their feature selection and feature importance rankings are shown in Table 5.2.

Table 5.2: Feature importance rankings and selection frequencies (the number in each cell indicates the feature importance ranking using a certain model).

Models & Selectors(→)	SVM				RandomForest				Freq.
Features(↓)	/	SFS	SFFS	SFBS	/	SFS	SFFS	SFBS	
Edge Intensity	19	10	12		17				2
CCS_mean	8	3	3	3	2				3
CCS_median	7	9			5	1	1	2	4
CCS_var	21	12			1				1
CCS_cv	17	8		14	11	4	5	5	5
Edge Density	13	11	9	7	10				3
HD_mean	9		10	11	20	10	12	8	5
HD_median	24		13		24	14	14	12	3
HD_var	18				13	6	7		2
HD_cv	22				22	12			1
TVS_mean	15				12	5	10	6	3
TVS_median	16			13	15				1
TVS_var	26	2			23		13		2
GLCM_contrast	2	1	2	2	9	7	6		5
GLCM_dissimilarity	1		1	1	3		3	1	4
GLCM_homogeneity	20	4	6	10	7				3
GLCM_energy	4			4	4	2	2		3
GLCM_correlation	10			8	16	8	9	7	4
ROI_var	11	6	8		6	3	4	4	5
ROI_evar	12			5	21			11	2
BED_mean	5		7		18	9	8	9	4
BED_median	25	7		12	26				2
BED_var	3	13	4	6	19	11		10	5
BED_cv	23		14	15	25				2
Circularity	14	5	11	9	14	13	11		5
ROI_proportion	6		5		8			3	2
Num. of Features	26	13	14	15	26	14	14	12	

As shown in the table, GLCM-based features (Group 4) are selected by both SVM and Random Forest (RF) models and obtain high importance, particularly GLCM_contrast and GLCM_dissimilarity. Edge-based features are also commonly chosen, with CCS_mean being highly ranked in the SVM model and CCS_median preferred by the RF model. Morphological features (Group

5) are frequently selected as well, but their importance varies across selectors. BED_cv, which is selected less often and ranks low in importance, should be discarded.

Compared to the other groups, features in Groups 2 and 3 are less preferred. In Group 2, features are less frequently chosen by SVM models but are favored by RF models. HD_cv, selected only once with a rank of 12, should be removed from Group 2. Group 3 features are also less frequently selected, though TVS_mean is important in RF models. If SVM is selected for classier, to reduce the number of features, Groups 2 and 3 could be considered for removal. This could even potentially improve the classification performance.

To evaluate the effect of discarding single or multiple groups of features, SVM models are trained with different combinations of feature groups. As we have 5 feature groups, this gives 32 different combinations. The performance of models trained using these combinations is shown in Figure 5.3.

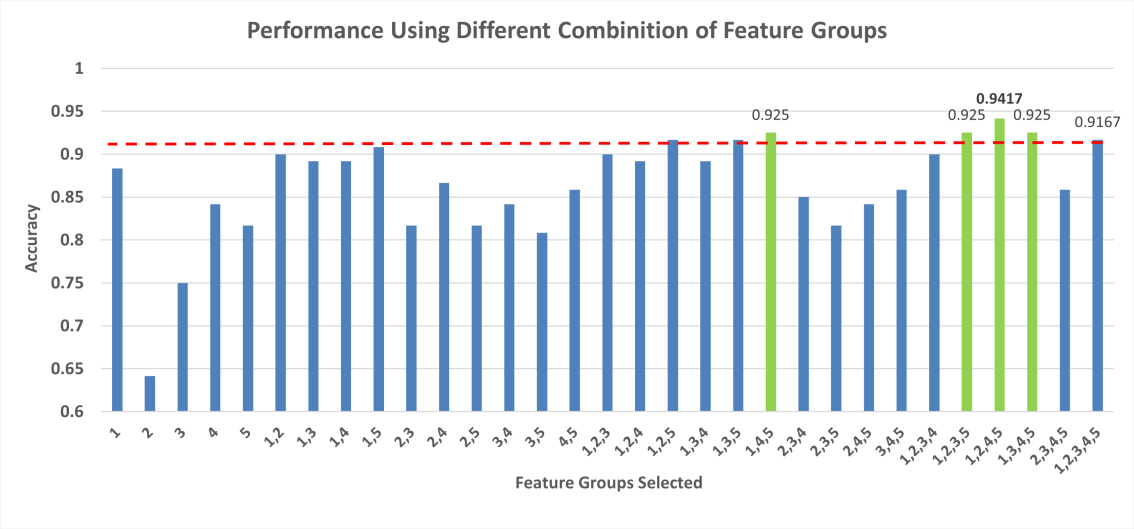


Figure 5.3: Comparison of SVM model performance with different feature group combinations, with the green bar indicating better performance on the test set compared to using all features.

The model trained using all features from the five groups serves as a standard. It achieves an accuracy of 91.67% on the test set. Four combinations outperform this baseline and are highlighted in green in the figure. Discarding Group 3, Group 4, or both results in higher accuracy. The highest accuracy of 94.17% is achieved when only Group 3 is excluded, as expected from the previous analysis. Interestingly, even when Group 4 is discarded, the model still performs better. This is likely because Group 1 alone is sufficient for analyzing textural features. This conclusion is supported by the "1, 5" case, where only two groups remain, yet still yield a high accuracy. However, when both Groups 1 and 4 are excluded, as in the "2, 3, 5" case, accuracy drops significantly to

81.17%. This is likely due to the absence of textural features. Another notable point is that retaining Group 4 while discarding Group 1 is not recommended. In all cases without Group 1, the best performance fails to exceed 86.67%. Therefore, although Group 4 features are theoretically the most significant, in practice, Group 1 is indispensable.

When only one feature group is used, the performance differences highlight each group's ability to separate the data. Groups 2 and 3 perform poorly, particularly Group 2. This indicates that the textural differences between hyperplastic and adenomatous polyps in the test set are not significant. It may be due to the absence of severe adenomatous polyps in the dataset, which have a relatively less uniform surface and more abrupt transitions to normal mucosa.

For Group 5, discarding it would decrease the accuracy slightly. It is more significant than Groups 2 and 3, but not as determinant as Groups 1 and 4. However, if only two groups can be chosen to train the model, selecting Groups 1 and 5 would be the best choice, as shown in case "1, 5".

Based on the analysis in this and the previous section, the importance of the five feature groups can be summarized as follows:

Groups 1 and 4 (Edge-based and GLCM-based features) are indispensable in feature extraction. Designing new features within these groups can likely improve model performance.

Group 5 (Morphological features) provides a considerable improvement to model performance. However, these features are highly sensitive to segmentation accuracy. To ensure their positive impact on classification, a more precise segmentation model is required, and the feature design should be made more robust. Nevertheless, BED_cv should be discarded from this group due to its undesirable pattern and high variability upon segmentation.

Groups 2 and 3 (Histogram-based and Transition-based features) are the least significant for classification. Discarding these two groups may not lead to a drop in performance and could potentially enhance it instead. However, certain features within these groups, such as HD_mean and TVS_median, could still be beneficial for classification and may be worth retaining.

5.3 Analysis of Misclassified Samples

In this master thesis, a platform implementing a cascaded model for polyp segmentation and classification is presented. The model achieves a Dice score of 0.8761 and an average classifica-

tion accuracy of 90.83%. The distribution of segmentation metrics and the confusion matrix are illustrated in Figure 5.4, providing detailed information about the model's performance.

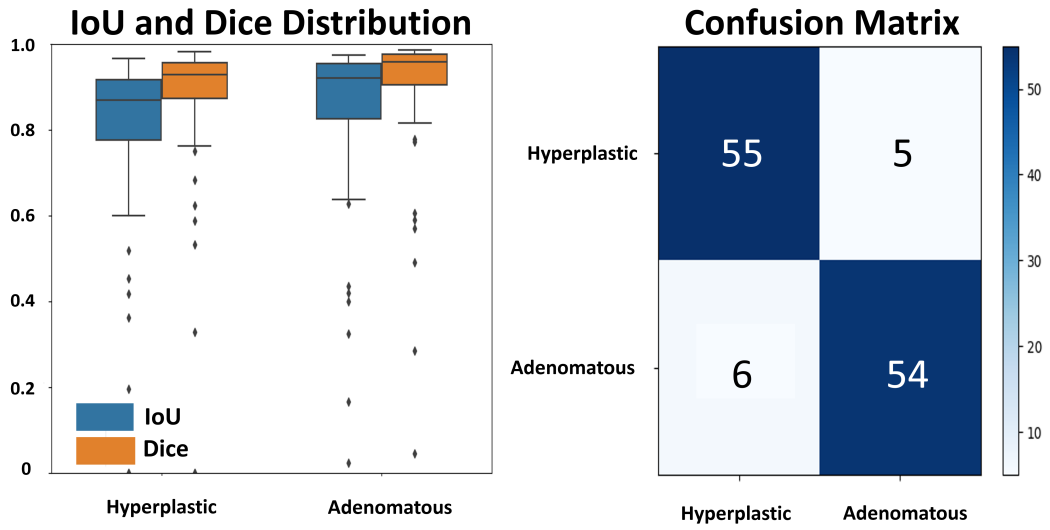


Figure 5.4: Distribution of IoU and Dice scores for segmentation (Left); Confusion matrix for classification (Right).

Generally speaking, the model performs relatively poorly in segmenting hyperplastic polyps. This is proved by lower Dice scores and IoU values. However, it shows slightly better performance in classifying hyperplastic polyps, with one fewer misclassified sample compared to the adenomatous class. This is partly because, even when hyperplastic polyps are missed and the mucosa is selected as the ROI, the classification can still be accurate due to the similarity of features between hyperplastic polyps and the mucosa. In contrast, adenomatous polyps are more sensitive to shifts in the ROI, which can affect classification accuracy.

Although several outliers are observed in the box plot, they do not necessarily contribute to misclassification. Similarly, good segmentation does not guarantee accurate classification. To further improve the model, analysis is conducted on five selected samples that were misclassified. The detailed information of the five cases are shown in Table 5.3 and Figure 5.5.

Firstly, we focus on the performance of the segmentation model. Cases 1 and 2 are well-segmented, with high Dice scores exceeding 0.96. There is little change when comparing the overlays. Misclassifications in these cases are mostly attributed to the classifier.

In Cases 3 and 4, where the metrics are lower, considerable differences exist between the predicted masks and the ground truth. These differences in the ROI lead to changes in polyp shape and borders, significantly affecting feature values and influencing classification. In these

cases, misclassification is likely due to both the segmentation model and the classifier.

For Case 5, the segmentation model only captures the specular highlight area while missing the entire polyp, which is a large adenomatous polyp. The misclassification in this case is entirely due to segmentation failure.

Table 5.3: Metrics of selected misclassified samples.

Case No.	Dice	IoU	True Label	Prediction	Confidence
1	0.9612	0.9253	Hyperplastic	Adenomatous	75.72%
2	0.9786	0.9581	Hyperplastic	Adenomatous	57.71%
3	0.5910	0.4194	Adenomatous	Hyperplastic	69.24%
4	0.8748	0.7775	Adenomatous	Hyperplastic	96.37%
5	0.0454	0.0232	Adenomatous	Hyperplastic	100.00%

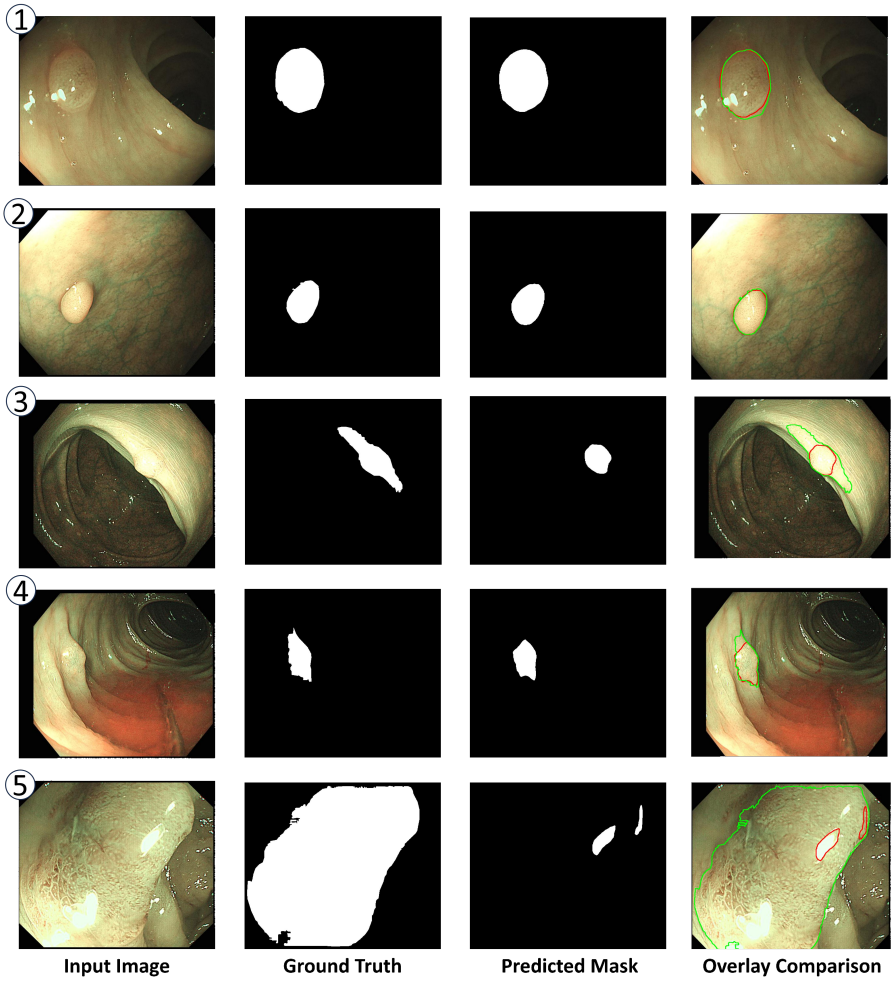


Figure 5.5: Segmentation outcomes of selected misclassified samples.

From these cases, we can conclude that the segmentation model struggles with sessile polyps, particularly those growing on colon folds, as seen in Cases 3 and 4. The model finds it challenging to accurately segment polyps as their borders are difficult to define. Also, large polyps with poorly

defined borders (Case 5) make it hard for the model to segment effectively. Besides the factor of the polyp itself, the presence of specular highlights significantly distracts the model, requiring special handling to improve segmentation accuracy.

Next, we analyze the classification results using SHAP values [76] to explain the contribution of each feature to the classification outcome. Firstly, the feature importances defined by the SVM+SFBS classifier are shown in Figure 5.6.

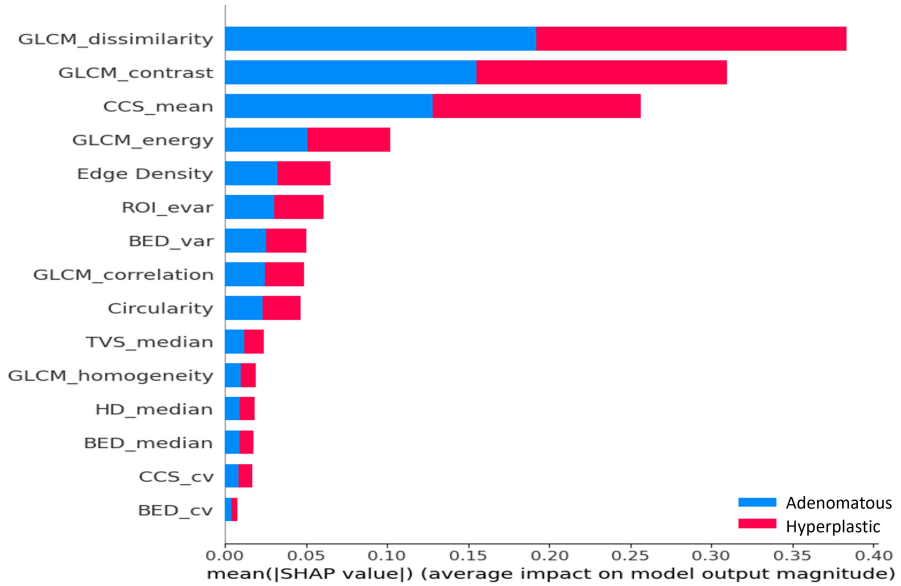


Figure 5.6: Feature importance of 15 selected features using SVM+SFBS.

The top three important features are GLCM_dissimilarity, GLCM_contrast, and CCS_mean. These features contribute most significantly to the overall feature importance. Changes in these features can substantially influence the classification results, while changes in rest feature values may have less effect. To analyze which features contribute to the misclassification, force plots are displayed for these five cases, as shown in Figure 5.7.

In the analysis of Cases 1 and 2, hyperplastic polyps are misclassified as adenomatous, which is mainly due to the contributions of GLCM-based features. These features highlight the richer texture content within the masked area. This makes the hyperplastic polyps appear texturally similar to adenomatous polyps. Such observation suggests that the texture patterns typically associated with adenomatous polyps are being picked up inappropriately in these hyperplastic samples. To tackle this issue, it is recommended to enlarge the training dataset with hyperplastic polyps with slightly richer texture features. This would help the model obtain a better ability to distinguish between the two classes when it comes to GLCM-based texture features. Meanwhile, the CCS_mean feature performs better in these cases, likely due to the presence of dot-like texture

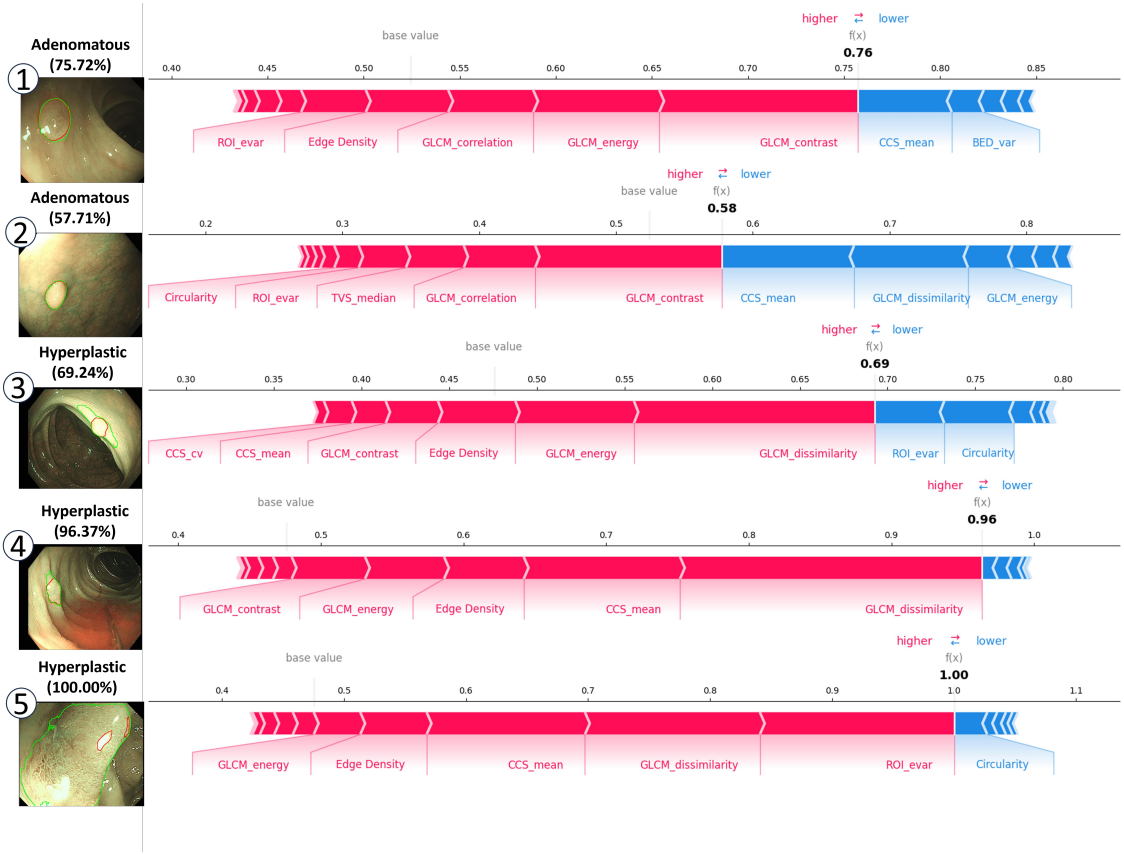


Figure 5.7: Force plots of misclassified samples: Features marked in red contribute to the misclassification, while those in blue aid correct classification.

within the polyps. This serves as an indicator for correctly classifying them as hyperplastic polyps, as hyperplastic polyps typically have fewer visible vessels compared to adenomatous polyps. Therefore, CCS_mean appears to be a more reliable feature for identifying hyperplastic polyps in these specific cases.

In Cases 3, 4, and 5, the misclassification occurs in the opposite direction where adenomatous polyps are incorrectly classified as hyperplastic. Here, both GLCM-based and Edge-based features contribute to the error. As can be seen from the images, All ROI has a uniform surface texture which is highly likely to be considered as hyperplastic polyps. This gives a low GLCM_dissimilarity and CCS value, which is not expected for adenomatous polyps. Furthermore, specular highlights obscure fine details of the texture by increasing local pixel intensity, making surface vessels less visible. In Case 5, the ROI is completely white. This leads to a 100% hyperplastic classification outcome as the CCS_mean, GLCM_dissimilarity, and ROI_var all reach their minimum values across all test samples.

Based on the case analysis above, three conclusions can be drawn to improve classification.

First, the boundary between hyperplastic and adenomatous polyps is not yet clearly defined. This leads to frequent misclassification of polyps near this boundary. Second, GLCM features are given too much importance, yet they are highly sensitive to changes, which significantly contributes to misclassification. Lastly, specular highlights greatly hinder both the classification and segmentation processes, so techniques should be developed and applied to address this issue.

5.4 Discussion

Based on the analysis in this chapter, several suggestions for feature design and model training are proposed. This section outlines these suggestions and discusses their feasibility and potential solutions. Also, the limitations of this research will be discussed.

5.4.1 Improvement in feature design

Four improvements could be made in feature design:

Discard less distinctive features. Based on the feature value distribution, features such as Circularity and BED_cv exhibit similar distributions for the two classes. This makes them less effective in distinguishing between them. Removing these features may not immediately enhance the model's performance on the current test set. However, with a larger set of features and a more extensive test set, the removal of these less distinctive features could improve the model's overall effectiveness.

Improve the robustness of features against segmentation. The features whose values fluctuate significantly before and after segmentation need to be redesigned. This is particularly important for morphological features, which play a crucial role in polyp classification. Future development efforts should focus on either enhancing the performance of the segmentation model to achieve higher Dice scores and IoU for both classes or modifying the feature calculation to make it more resilient to changes in border location, shape, and size. Notably, textural features (such as Edge-based and GLCM-based features) are inherently more robust to segmentation, so increasing the number of features in this category could be beneficial.

Redesign less significant feature groups. Analysis of different feature group combinations during training indicates that Histogram-based and Transition-based features are less significant or even detrimental to the current dataset. These features were originally designed for detecting

more severe polyps, such as early cancer, which are less common in NBI images. This is because they are significant enough under WLE and seldom require NBI to identify them. Given the small number of features in the current models, it is better to redesign these feature groups to capture subtle changes in textural uniformity and transition regions. As the dataset expands and includes an early cancer class (e.g., JNET Type 2B or C), the significance of these feature groups may become more apparent.

Increase the number of textural features. Textural features have shown significant value in polyp classification compared to other feature types. Currently, only five GLCM metrics are used, each with a single GLCM distance of 1. Future optimizations could involve selecting additional distances (e.g., 3 and 5) to capture more macro texture transformations. However, it is important to note that calculating GLCM with various distances and angles requires more computational resources and time. This requires us to find a balance between efficiency and the number of features. For Edge-based features, applying advanced image enhancement techniques, such as more sensitive edge detection algorithms, could improve the differentiation between hyperplastic polyps. This could also help avoid issues with outliers where few or even no edges are detected, making it difficult to calculate CCS.

5.4.2 Improvement in model training

Three improvements could be made in model training:

Incorporate striped and sessile Polyp Samples. The current segmentation model performs poorly with striped and sessile polyps, particularly adenomatous polyps located on the folds of the colon wall, which can be easily misclassified as hyperplastic polyps. To improve performance, it is essential to include samples of these polyp types in the dataset during further optimization.

Handle specular highlights. Specular highlights can be detected and processed using techniques such as inpainting. However, this method does not always ensure high-quality results. Often, the inpainted areas exhibit different texture features compared to other parts of the polyps, which can hinder the textural analysis in the classification process. Therefore, unless more advanced techniques for handling specular highlights become available, the most effective approach is to rely on colonoscopy hardware and imaging techniques designed to minimize specular highlights at the source.

Reconsider classification model selection. As can be seen from the feature importance

in Figure 5.6, the current SFBS feature selector process places the most importance on a small subset of features, suggesting that the results heavily depend on these values. As more training data becomes available and features are refined, the feature selector or even the classifiers may need to be reassessed and adjusted accordingly.

5.4.3 Limitations

Despite potential improvements to enhance model performance, several limitations need to be addressed in future work:

NBI-focused model. The cascaded model proposed in this thesis is specifically designed for NBI colonoscopy images. This limits its applicability to other imaging techniques, such as the widely used WLE. Although the model demonstrates high accuracy in polyp classification and is particularly beneficial for colonoscopists following the JNET standard, its specialized focus restricts its broader use. Moreover, NBI images are challenging to obtain, and there are fewer public datasets available, making training and further optimization more difficult compared to using WLE images.

Inadequate training data and model generalization. For the segmentation model, Polyp-PVT is currently trained with a combination of the public dataset and a private NBI dataset. While this approach has enhanced overall performance compared to using the NBI dataset alone, the specific impact of each sub-dataset remains unclear. A more detailed analysis is needed, and additional NBI images should be incorporated to improve performance further. Furthermore, the current input image size of 352 may lead to performance degradation when handling high-resolution medical images. For the classification model, the training and testing are based on a limited amount of data, which may not fully represent general patterns.

Feature interpretability and compliance with medical standards. I am an outsider with no direct involvement in colonoscopy and polyp classification. Many features have been designed based on my understanding of the JNET classification standard and insights from several experienced colonoscopists. These features may not fully align with established medical standards and may not provide strong evidence for polyp diagnosis. Besides, some features, such as GLCM features, are challenging to explain, which can make the results less interpretable for colonoscopists who lack specialized background knowledge.

Conclusions

In this master thesis, we enhance colorectal polyp diagnosis through a cascaded model. It consists of an upstream segmentation module and a downstream classification module. A polyp diagnosis platform is developed using Python and Qt Designer, integrating these modules to provide an efficient CADx system. The model classifies polyps into two categories: hyperplastic and adenomatous. On the test dataset, the model achieves an overall mDice score of 0.8761 and a classification accuracy of 90.83%.

The proposed model uses NBI images as input. These images are obtained during colonoscopy using NBI, a specialized technique that enhances the surface patterns of polyps through specific wavelengths. NBI images are commonly employed in clinical practice to classify polyps according to the JNET Classification standard. To translate clinical observations into features that can be extracted from medical images, new features are designed based on authoritative evidence and clinical experience.

This master thesis designs 26 features across 5 groups, focusing on enhanced textural edge patterns, histogram analysis, transition patterns, GLCM textural analysis, and morphology, respectively. Detailed analysis indicates that textural features are the most effective at distinguishing between hyperplastic and adenomatous polyps. They have also proved more robust to changes in masked images resulting from different segmentation processes. Morphological features, such as polyp size, shape, and color, are also crucial for the classification but are more dependent on the segmentation quality. However, histogram-based and transition-based features are found less significant for the current task.

The model training utilizes an annotated dataset of 1,211 NBI images for training and 120 for testing. To enhance the performance of the segmentation model, several public datasets are also involved. The Polyp-PVT and the traditional machine learning classifier SVM are employed for segmentation and classification, respectively. To optimize feature selection, SFBS is used as a feature selector. This combination achieves a classification accuracy of 93.33% on the ground truth feature values, which are unaffected by the segmentation model.

The polyp diagnosis platform represents the final contribution of this thesis. It offers an end-

to-end GUI with two main functions: visualization of the analysis for individual polyp images, and batch processing with report generation for multiple images. The processing speed in the current CPU environment (i5-12500H 2.50GHz) is approximately 0.5 seconds per image. The performance can be further improved with more advanced hardware. The platform's code structure is developed using the Model-View-Controller (MVC) architecture. This ensures a clear separation between business logic and the GUI. Besides, it enhances the maintenance and reusability of the methods.

However, several limitations have been identified during the evaluation. The diagnosis platform shows decreased performance when processing images with specular highlights; The distinction between hyperplastic and adenomatous polyps remains unclear due to the limited dataset and insufficient feature design; Additionally, the platform does not operate in real-time, which restricts its usability during colonoscopy procedures. To address these limitations, future work should focus on the following key areas for improvement:

Extend the dataset. The training dataset should be extended for both the segmentation and classification models to better capture the features of hyperplastic and adenomatous polyps under NBI imaging. Future studies should also consider incorporating WLE images and images from other imaging techniques. Additionally, including data from diverse hospital datasets and those more challenging to identify is essential to enhance the generalization ability of the models.

Optimize the feature design. The current study employs 26 features to characterize polyps, but some of these features have proven to be insignificant. To improve the model, additional features should be introduced. The calculation metrics for existing features should also be revised to enhance robustness against specular highlights and deviations caused by imprecise segmentation.

Extend functions of the diagnosis platform. The existing platform only supports batch processing of NBI images. However, hospital datasets typically include all images captured during a colonoscopy, necessitating a pre-selection process by colonoscopists. Furthermore, the platform should be enhanced to generate patient records automatically, thereby saving time for colonoscopists and improving workflow efficiency.

Generally speaking, to the best of my knowledge, this master thesis is the first to design a feature-based CADx methodology specifically for NBI colonoscopy images. It achieves commendably high classification accuracy and provides interpretable results that enable colonoscopists to understand and refine the analysis. Despite the model's limitations, the path for further optimization is clear, and the work presents a promising prospect for advancing CADx in polyp diagnosis through manual feature engineering and traditional machine learning methods.

A

List of Features

Table A.1: Complete feature list designed in this master thesis.

No.	Feature Name	Meaning
1	**Edge Intensity**	Average non-black pixel intensity, revealing surface vessel intensity
2	**CCS_mean**	Average of Connected Component Size (CCS)
3	**CCS_median**	Median of CCS
4	**CCS_var**	Variance of CCS
5	**CCS_cv**	Coefficient of variation (standard deviation over average) of CCS
6	**Edge Density**	Ratio of white pixels after binarization, revealing vessel density
7	**HD_mean**	Average of Histogram Distances (HD)
8	**HD_median**	Median of HD
9	**HD_var**	Variance of HD
10	**HD_cv**	Coefficient of variation of HD
11	**TVS_mean**	Average of Transition Variability Score (TVS)
12	**TVS_median**	Median of TVS
13	**TVS_var**	Variance of TVS
14	**GLCM_contrast**	Average contrast along four directions of GLCM
15	**GLCM_dissimilarity**	Average dissimilarity along four directions of GLCM
16	**GLCM_homogeneity**	Average homogeneity along four directions of GLCM
17	**GLCM_energy**	Average energy along four directions of GLCM
18	**GLCM_correlation**	Average correlation along four directions of GLCM
19	**ROI_var**	Variance of pixel intensity in ROI
20	**ROI_evar**	Variance of pixel energy in ROI
21	**BED_mean**	Average of Border-to-Ellipse Distance (BED)
22	**BED_median**	Median of BED
23	**BED_var**	Variance of BED
24	**BED_cv**	Coefficient of variance of BED
25	**Circularity**	Circularity of the circumscribed ellipse of polyp
26	**ROI_proportion**	Proportion of white pixels in the binary mask image

A

Bibliography

[1] The International Agency for Research on Cancer (IARC), *Global cancer observatory*, 2020. [Online]. Available: `https://gco.iarc.fr/en`.

[2] Y. Hazewinkel and E. Dekker, "Colonoscopy: Basic principles and novel techniques," *Nature reviews Gastroenterology & hepatology*, vol. 8, no. 10, pp. 554–564, 2011.

[3] G. Litjens, T. Kooi, B. E. Bejnordi, *et al.*, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[4] Y. Sano, M. Kobayashi, T. Kozu, *et al.*, "Development and clinical application of a narrow band imaging (nbi) system with builtin narrow-band rgb filters," *Stomach Intestine*, vol. 36, pp. 1283–1287, 2001.

[5] Y. Sano, S. Tanaka, S.-e. Kudo, *et al.*, "Narrow-band imaging (nbi) magnifying endoscopic classification of colorectal tumors proposed by the japan nbi expert team," *Digestive Endoscopy*, vol. 28, no. 5, pp. 526–533, 2016.

[6] J. E. East, N. Suzuki, M. Stavrinidis, T. Guenther, H. J. Thomas, and B. P. Saunders, "Narrow band imaging for colonoscopic surveillance in hereditary non-polyposis colorectal cancer," *Gut*, vol. 57, no. 1, pp. 65–70, 2008.

[7] K. Gono, T. Obi, M. Yamaguchi, *et al.*, "Appearance of enhanced tissue features in narrow-band endoscopic imaging," *Journal of biomedical optics*, vol. 9, no. 3, pp. 568–577, 2004.

[8] Y. Komeda, H. Kashida, T. Sakurai, *et al.*, "Magnifying narrow band imaging (nbi) for the diagnosis of localized colorectal lesions using the japan nbi expert team (jnet) classification," *Oncology*, vol. 93, no. Suppl. 1, pp. 49–54, 2017.

[9] J. J. Sung and N. C. Poon, "Artificial intelligence in gastroenterology: Where are we heading?" *Frontiers of medicine*, vol. 14, pp. 511–517, 2020.

[10] I. Pacal, D. Karaboga, A. Basturk, B. Akay, and U. Nalbantoglu, "A comprehensive review of deep learning in colon cancer," *Computers in Biology and Medicine*, vol. 126, p. 104 003, 2020.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[12] K. Waki, R. Ishihara, Y. Kato, *et al.*, "Usefulness of an artificial intelligence system for the detection of esophageal squamous cell carcinoma evaluated with videos simulating overlooking situation," *Digestive Endoscopy*, vol. 33, no. 7, pp. 1101–1109, 2021.

[13] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[14] A. Nogueira-Rodríguez, R. Domínguez-Carbajales, F. Campos-Tato, *et al.*, "Real-time polyp detection model using convolutional neural networks," *Neural Computing and Applications*, vol. 34, no. 13, pp. 10 375–10 396, 2022.

[15] A. Karaman, I. Pacal, A. Basturk, *et al.*, "Robust real-time polyp detection system design based on yolo algorithms by optimizing activation functions and hyper-parameters with artificial bee colony (abc)," *Expert systems with applications*, vol. 221, p. 119 741, 2023.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, Springer, 2015, pp. 234–241.

[17] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, Springer, 2018, pp. 3–11.

[18] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.

[19] D. Jha, P. H. Smedsrud, M. A. Riegler, *et al.*, "Resunet++: An advanced architecture for medical image segmentation," in *2019 IEEE international symposium on multimedia (ISM)*, IEEE, 2019, pp. 225–2255.

[20] D. Jha, P. H. Smedsrud, M. A. Riegler, *et al.*, "Kvasir-seg: A segmented polyp dataset," in *MultiMedia modeling: 26th international conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, proceedings, part II 26*, Springer, 2020, pp. 451–462.

[21] J. Wei, Y. Hu, R. Zhang, Z. Li, S. K. Zhou, and S. Cui, "Shallow attention network for polyp segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI*

2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24, Springer, 2021, pp. 699–708.

[22] D.-P. Fan, G.-P. Ji, T. Zhou, *et al.*, "Pranet: Parallel reverse attention network for polyp segmentation," in *International conference on medical image computing and computer-assisted intervention*, Springer, 2020, pp. 263–273.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[24] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, and L. Shao, "Polyp-pvt: Polyp segmentation with pyramid vision transformers," *arXiv preprint arXiv:2108.06932*, 2021.

[25] F. Tang, Z. Xu, Q. Huang, *et al.*, "Duat: Dual-aggregation transformer network for medical image segmentation," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, Springer, 2023, pp. 343–356.

[26] A. Kirillov, E. Mintun, N. Ravi, *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[27] Y. Li, M. Hu, and X. Yang, "Polyp-sam: Transfer sam for polyp segmentation," in *Medical Imaging 2024: Computer-Aided Diagnosis*, SPIE, vol. 12927, 2024, pp. 759–765.

[28] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024.

[29] X. Wei, J. Cao, Y. Jin, M. Lu, G. Wang, and S. Zhang, "I-medsam: Implicit medical image segmentation with segment anything," *arXiv preprint arXiv:2311.17081*, 2023.

[30] M. Pomeroy, H. Lu, P. J. Pickhardt, and Z. Liang, "Histogram-based adaptive gray level scaling for texture feature classification of colorectal polyps," in *Medical Imaging 2018: Computer-Aided Diagnosis*, SPIE, vol. 10575, 2018, pp. 507–513.

[31] J. J. Fu, Y.-W. Yu, H.-M. Lin, J.-W. Chai, and C. C.-C. Chen, "Feature extraction and pattern classification of colorectal polyps in colonoscopic imaging," *Computerized medical imaging and graphics*, vol. 38, no. 4, pp. 267–275, 2014.

[32] Y. Hu, Z. Liang, B. Song, *et al.*, "Texture feature extraction and analysis for polyp differentiation via computed tomography colonography," *IEEE transactions on medical imaging*, vol. 35, no. 6, pp. 1522–1531, 2016.

[33] G. Wimmer, T. Tamaki, J. J. Tischendorf, *et al.*, "Directional wavelet based features for colonic polyp classification," *Medical image analysis*, vol. 31, pp. 16–36, 2016.

[34] S. Engelhardt, S. Ameling, S. Wirth, and D. Paulus, "Features for classification of polyps in colonoscopy.," *Bildverarbeitung für die Medizin*, vol. 574, pp. 350–354, 2010.

[35] E. Ribeiro, A. Uhl, G. Wimmer, and M. Häfner, "Exploring deep learning and transfer learning for colonic polyp classification," *Computational and mathematical methods in medicine*, vol. 2016, no. 1, p. 6 584 725, 2016.

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

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

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

[39] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

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

[41] M. Tan, B. Chen, R. Pang, *et al.*, "Mnasnet: Platform-aware neural architecture search for mobile," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2820–2828.

[42] K. Patel, K. Li, K. Tao, *et al.*, "A comparative study on polyp classification using convolutional neural networks," *PloS one*, vol. 15, no. 7, e0236452, 2020.

[43] M. S. Hossain, M. M. Rahman, M. M. Syeed, *et al.*, "Deeppoly: Deep learning based polyps segmentation and classification for autonomous colonoscopy examination," *IEEE Access*, 2023.

[44] R. Szeliski, *Computer vision: algorithms and applications*. Springer Nature, 2022.

[45] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9404–9413.

[46] D. Guo, Y. Pei, K. Zheng, H. Yu, Y. Lu, and S. Wang, "Degraded image semantic segmentation with dense-gram networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 782–795, 2019.

[47] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, Springer, 2014, pp. 740–755.

[48] A. Burkov, *The hundred-page machine learning book*. Andriy Burkov Quebec City, QC, Canada, 2019, vol. 1.

[49] Jun. 2024. [Online]. Available: `https://en.wikipedia.org/wiki/Feature_engineering`.

[50] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, pp. 273–297, 1995.

[51] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[52] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Computational intelligence and neuroscience*, vol. 2018, no. 1, p. 7 068 349, 2018.

[53] A. Vial, D. Stirling, M. Field, *et al.*, "The role of deep learning and radiomic feature extraction in cancer-specific predictive modelling: A review," *Translational Cancer Research*, vol. 7, no. 3, 2018.

[54] K. Pearson, "Contributions to the mathematical theory of evolution," *Philosophical Transactions of the Royal Society of London. A*, vol. 185, pp. 71–110, 1894.

[55] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610–621, 1973.

[56] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[57] K. Han, Y. Wang, H. Chen, *et al.*, "A survey on vision transformer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 87–110, 2022.

[58] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[59] Z. Wu and Y.-G. Jiang, "Deep learning basics for video understanding," in *Deep Learning for Video Understanding*, Springer, 2024, pp. 7–20.

[60] W. Wang, E. Xie, X. Li, *et al.*, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 568–578.

[61] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015. DOI: 10.1016/j.compmedimag.2015.02.007.

[62] J. Bernal, J. Sánchez, and F. Vilarino, "Towards automatic polyp detection with a polyp appearance model," *Pattern Recognition*, vol. 45, no. 9, pp. 3166–3182, 2012. DOI: 10.1016/j.patcog.2012.03.002.

[63] X. Zhao, L. Zhang, and H. Lu, "Automatic polyp segmentation via multi-scale subtraction network," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, Springer, 2021, pp. 120–130.

[64] T. Kim, H. Lee, and D. Kim, "Uacanet: Uncertainty augmented context attention for polyp segmentation," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 2167–2175.

[65] Y. Komeda, H. Handa, T. Watanabe, *et al.*, "Computer-aided diagnosis based on convolutional neural network system for colorectal polyp classification: Preliminary experience," *Oncology*, vol. 93, no. Suppl. 1, pp. 30–34, 2017.

[66] C.-M. Hsu, C.-C. Hsu, Z.-M. Hsu, F.-Y. Shih, M.-L. Chang, and T.-H. Chen, "Colorectal polyp image detection and classification through grayscale images and deep learning," *Sensors*, vol. 21, no. 18, p. 5995, 2021.

[67] T. Ozawa, S. Ishihara, M. Fujishiro, Y. Kumagai, S. Shichijo, and T. Tada, "Automated endoscopic detection and classification of colorectal polyps using convolutional neural networks," *Therapeutic advances in gastroenterology*, vol. 13, p. 1 756 284 820 910 659, 2020.

[68] S. Tanwar, P. Goel, P. Johri, and M. Diván, *Classification of benign and malignant colorectal polyps using pit pattern classification. ssrn electron j. 2020.*

[69] A. Bour, C. Castillo-Olea, B. Garcia-Zapirain, and S. Zahia, "Automatic colon polyp classification using convolutional neural network: A case study at basque country," in *2019 IEEE international symposium on signal processing and information technology (ISSPIT)*, IEEE, 2019, pp. 1–5.

[70] A. Krenzer, S. Heil, D. Fitting, *et al.*, "Automated classification of polyps using deep learning architectures and few-shot learning," *BMC Medical Imaging*, vol. 23, no. 1, p. 59, 2023.

[71] E. Ribeiro, A. Uhl, and M. Häfner, "Colonic polyp classification with convolutional neural networks," in *2016 IEEE 29th international symposium on computer-based medical systems (CBMS)*, IEEE, 2016, pp. 253–258.

[72] M. F. Byrne, N. Chapados, F. Soudan, *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.

[73] C.-M. Lo, Y.-H. Yeh, J.-H. Tang, C.-C. Chang, and H.-J. Yeh, "Rapid polyp classification in colonoscopy using textural and convolutional features," in *Healthcare*, MDPI, vol. 10, 2022, p. 1494.

[74] R. Zhang, Y. Zheng, T. W. C. Mak, *et al.*, "Automatic detection and classification of colorectal polyps by transferring low-level cnn features from nonmedical domain," *IEEE journal of biomedical and health informatics*, vol. 21, no. 1, pp. 41–47, 2016.

[75] R. Lambert, "The paris endoscopic classification of superficial neoplastic lesions: Esophagus, stomach, and colon: November 30 to december 1, 2002," *Gastrointest Endosc*, vol. 58, S3–S43, 2003.

[76] S. Lundberg, "An introduction to explainable ai with shapley values," *Revision 45b85c18*, 2018.