UiT The Arctic University of Norway

Faculty of Science and Technology
Department of Computer Science

Deep Learning in Precancerous Lesions Detection during Surveillance Colonoscopies of IBD Patients

Mayank Roy

INF-3990 Master's thesis in Computer Science, January 2024

Abstract

Deep Learning (DL) models have developed tremendously over the last couple of decades in their ability to train across large datasets and give fast and accurate results across a varied number of tasks like image classification and segmentation. This is the reason why DL models are being increasingly adopted for aiding medical professionals in the diagnosis and detection of various medical conditions like colorectal cancer (CRC). There happens to be a specific group of patients who suffer from Inflammatory Bowel Disease (IBD) who are at a significantly higher risk of developing CRC, which is why they undergo surveillance colonoscopies. However, precancerous lesions that have the potential of developing into cancer can sometimes be difficult to spot, identify and observe changes in, during colonoscopies. IBD patients can have internal scarring of tissue, which makes it even more difficult to detect precancerous lesions and spot the changes in them during surveillance colonoscopies. Modern DL models can be useful in aiding the detection and identification of these precancerous lesions, which is why in this thesis, various DL-based approaches for the detection and clustering of precancerous lesions during surveillance colonoscopies of IBD patients were tested. Both a supervised object-detection-based approach on a labelled dataset, and an unsupervised image-clustering-based approach were tried out using pre-designed DL models. Furthermore, it was investigated whether the colour channel separation and possible recombination of certain colour channels of the images in the dataset could help improve the detection of precancerous lesions, and make the object detection model more accurate.

Some results of the unsupervised image-clustering-based approach looked promising, but it was unable to segregate each type of potential precancerous findings into separate clusters. The supervised learning-based approach that did object detection worked very well with the labelled dataset used in this project. The colour channel separation and recombination of images in the dataset gave a significant improvement to the performance of the object detection model, particularly when the images in the dataset consisted of only the blue channel of the original RGB images.

Acknowledgements

I would like to express my deep gratitude to my supervisors Anne Håkansson and Peter Thelin Schmidt for their patient guidance, encouragement and constructive critique throughout the course of this project, despite the international borders and issues with Zoom that occasionally separated us.

I would also like to express my sincere thanks to the HMT Group and all its members, whose medical expertise and input were instrumental in guiding the work done in this project. Special thanks go to Thomas de Lange, whose experience in this field of research helped me immensely, and Camilla Wijkström, whose contribution towards providing me with the new set of colonoscopy videos to be used in this project was of great help.

I would also like to thank my friends and colleagues who have had to (sometimes begrudgingly) be my intellectual sparring partners over the last year. I acknowledge now that colonoscopies were perhaps not the best topic for dinner conversations.

Finally, this would not have been possible without the support of my family. Especially my Dad - who has always done a fantastic job of pretending to understand everything I tell him about my work with machine learning and colonoscopies; and my Mom - whose wish for me to become a Doctor was never fulfilled by me because of my squeamishness when it comes to invasive medical procedures (only for life to come a full circle over the last year when I had to work with colonoscopy videos on an almost daily basis for this Master thesis project).

Contents

List of Figures

List of Tables

1

Introduction

Deep Learning (DL) is a powerful subset of Machine Learning (ML) that leverages large datasets and sophisticated neural network architectures in order to train models that are capable of performing a large number of complex tasks. In this thesis, the use of DL models in the field of Colorectal cancer (CRC), for the task of detecting and identifying precancerous lesions in chronic IBD patients has been explored.

1.1 Background

Over the course of the last few years, the Deep Learning computing paradigm has steadily developed to become the Gold Standard when it comes to the Machine Learning (ML) community [7]. It has gradually developed into the most widely used computational approach in ML, delivering excellent results across a range of complex tasks, matching or even sometimes beating the performance of humans in some of these tasks [7]. DL models have outperformed many well-known traditional ML techniques and given excellent performances in domains ranging from cybersecurity [50] to natural language processing [35, 164], bioinformatics [71], speech processing [154, 4], computer vision [155, 102] and medical informational processing [89], amongst many others [7, 140, 74]. Deep Learning models are even used for specific applications like estimating the damage caused by natural disasters [142], the discovery of new drugs [25] or the diagnosis of various forms of cancer [124, 39, 31, 9][7], which has the potential for significantly improving human lives. The potential ability of DL models to aid medical professionals with the detection and diagnosis of precancerous lesions is what forms the focus of this thesis.

When it comes to the detection of cancer or precancerous lesions in videos and images, DL models like YOLO [114] that do real-time object detection have enormous potential to be developed and used [112, 94] for a full range of computer vision tasks including object detection, segmentation, pose estimation, tracking and classification which allows users and researchers to leverage the model's capabilities across diverse applications and domains [58]. One such area of interest in exploring YOLO and other similar object-detection models' use is for the detection and identification of precancerous lesions associated with CRC.

These object detection models also come with the requirement of needing large amounts of labelled datasets for the purpose of training these models, typically in the form of bounding box annotations [113]. So when there is a large dataset consisting of unlabelled images with no bounding box annotations, there is a need to make use of DL techniques that do not need labelled datasets, like deep clustering models [22, 163]. The ability of these kinds of DL-based clustering models to cluster frames of colonoscopy videos having the same or similar precancerous lesions into groups has also been explored in this thesis.

CRC is a disease where the cells in the colon or rectum grow out of control. Sometimes, abnormal growths called *polyps* or *lesions* form in the colon or rectum, which can occasionally turn cancerous [158]. Screening tests can help find these lesions so that they can be removed before turning into cancer, and can also help find CRC at an early stage when treatment works best [158].

Colorectal Cancer (CRC) is the third most common cancer worldwide, with over 1.9 million new cases reported in the year 2020 [27]. It also happens to be the second most common cause of cancer mortality in the world, with over 900,000 cases of CRC mortality in the year 2020 [27][26]. It has been estimated by the International Agency for Research on Cancer (IARC) that there will be a 56% increase in the global burden of CRC between the years 2020 and 2040, to over 3 million new cases worldwide per year, which will also be coupled with an estimated increase of 69% in the global CRC mortality per year to approximately 1.6 million deaths in 2040 [26].

About 90% of the people whose CRC is detected before it spreads to nearby lymph nodes or organs survive longer than 5 years after the diagnosis [77]. However, only 10% of those whose CRC has spread to distant parts of their body survive this same period of 5 years [77].

Given the significant numbers associated with CRC both in terms of the number of new cases as well as the number of deaths attributed to it every year, there is a heightened need to improve the current diagnostic and screening tools associated with precancerous lesions and CRC detection. This need gets amplified when it is observed just how effective early detection is when it comes to significantly boosting the 5-year survival rate of CRC patients, from 10% to almost 90%.

There are a few risk factors associated with CRC like a personal history of CRC or adenomatous polyps, a strong family history of CRC or polyps, a personal history of inflammatory bowel disease (IBD) and a family history of any hereditary CRC syndrome such as familial adenomatous polyposis (FAP) and Lynch syndrome [157]. It is recommended for people with an average risk to begin screening for CRC when they are over the age of 50 years [77]. One of the most commonly used tests for CRC screening is a colonoscopy, which is recommended to be done at regular intervals depending on the risk factors a patient has of developing CRC. For instance, experts recommend patients who have a history of CRC in multiple first-degree relatives (FDRs) to undergo colonoscopy (rather than any other screening methods for CRC) at an interval of every 5 years, starting at the age of 40 or 10 years before the age of the relative's diagnosis, whichever comes first [17]. Similarly, it is recommended for IBD patients to undergo surveillance colonoscopies beginning 8 years after diagnosis [17].

1.2 Problem

Since there is a human element involved in the process of colonoscopy, there is a need to automate and support real-time surveillance for patients at a higher risk of developing precancerous lesions, especially chronic IBD patients. Also, since colonoscopies are an operator-dependent procedure, human factors like fatigue, lack of sensitivity to visual characteristics associated with polyps and insufficient attentiveness during the colon examination can lead to potential mis-detection of polyps [133]. A retrospective observational study of patients who underwent a second colonoscopy within 6 months of the first evaluated the miss rate of colorectal polyps to be 17.24%, with 38.69% of the patients having at least 1 missed polyp [82].

Chronic IBD patients can have inflammation or scarring of tissue, which can make it even more difficult to detect precancerous lesions. Given the fact that early detection is of extreme importance in boosting the 5-year survival rate of CRC patients and that surveillance colonoscopy as a procedure is recommended to be done on IBD patients starting 8-10 years after initial diagnosis [17, 156] at an interval of 1-5 years depending on the risk factors [156], there is a significant need in the field of medicine to minimise the human error factor associated with colorectal screening of IBD patients during surveillance colonoscopies. DL-based tools might be able to help to reduce this human error factor, as well as support the surveillance colonoscopies of chronic IBD patients.

Problem to be investigated - How can DL-based systems support the detection and identification of precancerous lesions during Inflammatory Bowel Disease (IBD) surveillance colonoscopies?

1.3 Purpose

This thesis presents two primary tasks that were investigated over the course of the project in order to explore the problem of precancerous lesion detection and identification in IBD patients undergoing surveillance colonoscopies, using DL-based models. These two tasks are :

- **Object detection -** The training and testing of a pre-designed DL model (YOLOv8[69]/ YOLOv5[44]) for precancerous lesion detection in chronic IBD patients. This also involves testing out different pre-processing steps to improve the performance of the DL model, with a particular emphasis on the colour-channel separation and recombination in the images.

- **DL based image clustering -** The testing of a DL based clustering model on a dataset of images without bounding-box annotations, in order to see how well this model can identify and detect precancerous lesions and group them together into clusters, in surveillance colonoscopy patients.

1.4 Goal(s)

The goal of this thesis is a DL-based system that is best suited for the purpose of detection and identification of precancerous lesions in IBD surveillance colonoscopies.

DL models that can cluster these precancerous lesions together based on their visual characteristics as recorded in the colonoscopy video feed will be probed. Since the manual process of annotating colonoscopy datasets is a long, tedious and resource-intensive process, the clustering algorithm would also benefit medical professionals tasked with the labelling of these datasets by clustering together images and video-frames that have similar forms of visual anomalies to help distinguish them from other clusters.

The DL model for object detection of precancerous lesions would ideally be of a high enough accuracy as well as a decent enough frame rate to help aid medical professionals engaged in conducting colonoscopies. Processes like colour-channel separation and recombination of the split colour channels of the images will also be probed in order to test if they give any boost to the accuracy of the baseline DL model.

Medical health professionals, especially the ones associated with colonoscopies related to colorectal screening of IBD patients would be the primary beneficiaries of this work. If the results are promising, the DL-based systems developed in this project could be integrated into modern endoscopy systems used by these medical professionals while performing colonoscopies in order to aid them either during or after they conduct the procedure on patients, to help detect or identify precancerous lesions.

The work done in this thesis would also serve as an initial exploration and testing of DL models and techniques on the novel dataset of surveillance colonoscopy videos of chronic IBD patients that was used for the first time for this thesis.

1.5 Research method

On a fundamental level, there are two broad categories of basic research methods - *Quantitative research method* and *Qualitative research method*. The *Quantitative research method* supports experiments & testing by measuring certain variables in order to verify or falsify theories, hypotheses, computer systems' functionalities & interfaces [49]. Meanwhile, *Qualitative research method* concerns understanding meanings, opinions and behaviours in order to reach tentative hypotheses & theories or to develop computer systems and inventions [49]. There is also a third type of basic research method called *Triangulation* where both the previous research methods are used as complements in order to get a complete view of the research area [49].

Since this thesis investigates the performance of the DL-based system based on the input and insight of medical professionals in the field of CRC and IBD, the **Qualitative research method** [51] is used here. These insights from the group of medical professionals associated with this project were essential towards the further development of the DL-based system, especially the image clustering system, which is why the *Qualitative research method* was most appropriate for this project.

1.6 Research approach

Research approaches, with the most famous being *inductive* and *deductive*, are used for drawing conclusions and establishing what is true or false [49]. The *inductive approach* involves formulating theories and propositions with alternative explanations from observations and patterns; with the data collected & analysed in order to gain an understanding and establish different views on the phenomenon [49]. Meanwhile, the *deductive approach* involves testing theories to verify or falsify hypotheses by testing them using quantitative methods on large datasets [49]. The ***inductive approach*** was used in this thesis as the research approach for this thesis.

1.7 Benefits, Sustainability and Ethics

The **benefits** associated with the work done in this project are manifolds. It will benefit the medical professionals conducting colonoscopies by aiding the process of precancerous lesion detection, which would have otherwise been a completely manual process which relied solely on the medical professionals. The research work conducted in this project will also benefit society in general and chronic IBD patients in particular by improving the process of precancerous lesion detection during surveillance colonoscopies, hence helping in the early diagnosis and treatment of CRC, especially in chronic IBD patients who have a higher risk of developing CRC.

When it comes to the **sustainability** aspect of this research project, the model that is designed and tested, assuming that it performs satisfactorily in the evaluations, will be able to be used continuously to aid the detection of precancerous lesions for patients, including newer patients that do not feature in the testing dataset.

Since this research project involves working with medical data, there are strict **ethical** codes in place to ensure that patient data remains private and anonymous. There is a clear demarcation between the project members who will be collecting the patient data, and the project members who will be developing and testing the DL-based system for detecting and clustering the precancerous lesions. The doctors, nurses and other medical professionals collecting the colonoscopy videos make sure to anonymize each of these videos so that the patient cannot be identified on the basis of the colonoscopy videos collected.

1.8 Delimitations

This project was done with a few delimiting factors in mind, to help narrow down the focus of the study. These are :

1. Testing population - The models that were developed were tested on a limited dataset of colonoscopy videos obtained from a total of 32 patients, with 10 patients contributing to an older dataset from 2021, and 22 patients contributing to a newer, novel dataset used and explored in this project. This new patient population consisted only of chronic IBD patients undergoing surveillance colonoscopies. This was the target population to optimise the AI/DL models for.

2. Timeframe - The older dataset of 10 patients was collected over a time period of 16 months between October 2019 and January 2021, while the newer dataset of 22 patients was collected over the course of 4 months, between July 2023 and October 2023.

3. Dataset quality - The older dataset that was from 2021 had pictures of a lower resolution than the resolution of pictures/videos that modern endoscopy equipment produces. Moreover, a major delimitation associated with this thesis was that the newer dataset from 2023 of colonoscopy videos was not annotated with bounding box labels in time to be used for the purpose of this project's *object detection* task.

4. Colonoscopy Equipment - The testing dataset is collected using the Olympus Endoscopy System. So this project did not aim to test out how the DL systems that were developed here could aid colonoscopies performed with other endoscopy systems like the Fuji Endoscopy System or the PENTAX Endoscopy System.

5. DL techniques used - For the *Object Detection task*, just the YOLOv5 and YOLOV8 models were explored here. No other DL models were used for this task. Similarly, for the *Image Clustering task*, only the VGG-16 and VGG-19 models were used. Moreover, DL models for tasks like image segmentation could not be used in this project due to the lack of proper segmentation-masks-based annotations necessary for training these models, in the datasets used here.

1.9 Stakeholders

This project is the result of the collaboration and contribution of numerous institutes and individuals across Sweden and Norway. Doctors, nurses, PhD students, research fellows, professors and IBD patients from Ersta Hospital (Sweden), Sahlgrenska University Hospital (Sweden), Karolinska Institutet (Sweden), Uppsala University (Sweden), KTH Royal Institute of Technology (Sweden), University of Göthenburg (Sweden) and UiT The Arctic University of Norway (Norway) are the primary stakeholders of this multidisciplinary project between individuals involved in the field of computer science and medicine. In the long term, endoscopy device companies are also stakeholders in this project.

This thesis forms a part of the project "RS2021-0316, Ansökningsnummer FoUI-966813, *Artificiell intelligens for tidig upptäckt av premaligna och maligna förändringar i tjocktarmen.*", and is funded by the HMT funds for 2022 and 2023.

1.10 Outline

This thesis comprises of several chapters that explore and demonstrate various facets of this research topic. In Chapter 2, there is a focus on Deep Learning and related work, to gain a basic understanding of Deep Learning as well as to briefly explore the existing literature and research that has been done using DL in this field. Chapter 3 explores this project from a medical standpoint, while trying to introduce and explain terms and conditions associated with IBD, polyps and colonoscopies. In chapter 4, there is a focus on the methods and methodologies employed over the course of this thesis, while justifying the reasoning behind picking them. Chapter 5 focuses on the requirements and design of the system, exploring the DL-model training and pre-processing pipeline. Chapter 6 lays emphasis on the implementation of the system, highlighting how the datasets were processed and how the experiments were performed over the course of this thesis. Chapter 7 presents the results of the thesis, showcasing the performance of the DL-based system and also discussing these results. Chapter 8 is about the future works associated with the work done in the thesis, exploring avenues for further research and refinement. Finally, chapter 9 encapsulates this thesis with a conclusion, summarising the key findings and insights associated with the work done over the course of this project.

2

Deep Learning and Related Work

The term *Deep Learning* is believed to have been proposed by Rina Dechter Dechter in 1986 [42], while the first general, working algorithm for a deep, multilayered perception network was published in 1967 by Alexey Ivakhnenko and Lapa [63, 161]. However, in the 21st century, Hinton et al. (2006) [56], have been pioneers in the field of Deep Learning research, along with other scientists and researchers like Yann LeCun and Yoshua Bengio [161]. Geoffrey Hinton along with Yann LeCun and Yoshua Bengio were awarded the ACM A.M. Turing Award in 2018 "for conceptual and engineering breakthroughs that have made deep neural networks a critical component of computing" [1].

Deep Learning (DL) is based on the concept of Artificial Neural Networks (ANNs) [64] - networks that are created by simulating a network of model neurons in a computer [75]. DL is considered a subset of Machine Learning (ML) and Artificial Intelligence (AI) in the working domain, and can be seen as an AI function that mimics the human brain's processing of data [120]. DL is derived from conventional Neural Networks (NNs)[13] but has a tendency to outperform its predecessors, owing to its ability to learn from massive amounts of data [7, 120]. DL models are usually designed by stacking multiple layers of NNs consisting of artificial neurons, which are fundamentally nodes through which data and computations flow[149]. This stacking of layers of NNs forms a *deep* architecture, which enables the DL model to extract features associated with the training data it is fed, hence enabling it to discover complex patterns and relationships in data. This is why DL models tend to outperform traditional ML models in many tasks, as discussed earlier in Chapter 1. However, the deeper nature of these models leads to an increase in the number of training parameters associated with the models, which is why training DL models usually is a much longer and computationally more expensive process[162, 120].

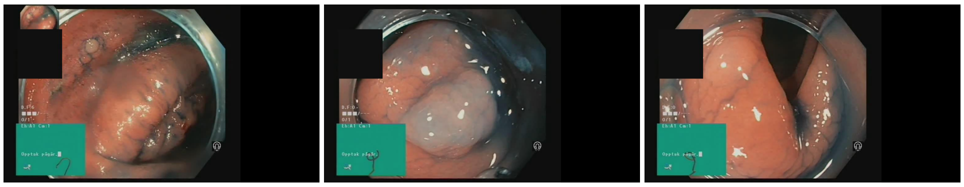
2.1 Classification of DL approaches

DL techniques are classified into three major categories - supervised, unsupervised and semi-supervised, although there is also a fourth category called Reinforcement Learning (RL) which is a form of semi-supervised (and sometimes unsupervised) learning technique [7].

2.1.1 Deep Supervised Learning

This DL technique deals with datasets that contain properly annotated training data [33]. Each training example has a *label* associated with it, which the DL model trains itself to predict. Supervised Learning techniques induce models from the training data, and these models can later be used to classify or predict the labels of unlabelled data after training [33]. Tasks like image classification (where the training data has a proper label associated with each image) or image segmentation (where the training data has a proper label associated with each pixel of the image) are suited for supervised learning techniques to be used in. This thesis employs pre-designed supervised-learning based DL models like YOLO [14, 58, 69] and VGG[127].

An example of a small sample of a labelled dataset can be seen in figure 2.1. Each image has an associated label, with a total of 3 different labels available. So a DL model that is trained to classify an image into one of these 3 labels can be considered to be a type of deep supervised learning model.



(a) Dysplasia **(b)** Hyperplasia **(c)** Inflammation

Figure 2.1: Sample of a dataset having images with their respective labels

2.1.2 Deep Unsupervised Learning

This technique lets the learning process be implemented even if the training dataset does not have any labelled data [7]. In this, the DL model tries to infer some underlying structure or patterns from the training data by trying to learn the significant features or interior representation required to discover the unidentified structure or relationships between the input data [7, 150]. Techniques of generative networks, dimensionality reduction and clustering are often considered to be within the category of unsupervised learning [7]. One of the most popular approaches of unsupervised learning is *clustering*, where the goal tends to be to categorize similar data into one cluster based on certain measures (like Euclidean distance) [7, 96].

Using figure 2.1 above as an example, if the dataset had all the images but without any of their associated labels, unsupervised learning techniques would be employed to try and cluster the images into separate clusters on the basis of similar visual characteristics. Aljalbout et al. (2018) [6] and Min et al. (2018) [96] have described techniques to create such a DL model that is used for unsupervised learning.

2.1.3 Deep Semi-Supervised Learning

The Deep Semi-Supervised Learning technique involves using labelled as well as unlabelled data to perform certain learning tasks [150]. Semi-supervised learning algorithms usually attempt to improve the performance of either a supervised learning task or an unsupervised learning task, by utilising information generally associated with the other [150].

2.2 Important terms and techniques

This section will briefly encapsulate a few terms associated with Deep Learning as well as a few additional techniques that are made use of over the course of this project, either in association with DL models, or independent of them.

2.2.1 Bounding Box

A bounding box is a rectangular box that is drawn around an object that needs to be identified/detected by a DL-based object detection model.

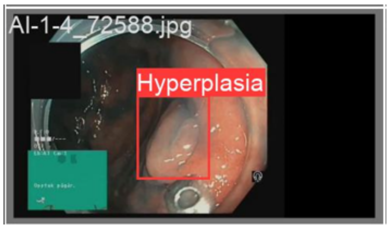


Figure 2.2: Image of an object (hyperplasia) labelled by a bounding box

In figure 2.2, there is a bounding box of the colour red around an object that needs to be detected, which in this case is hyperplasia. It helps localise the object of interest.

2.2.2 Object Detection

Object detection is a type of supervised learning task that aims to locate and classify the objects existing in any one image and label these objects with rectangular bounding boxes, to show the confidence of their existence [168].

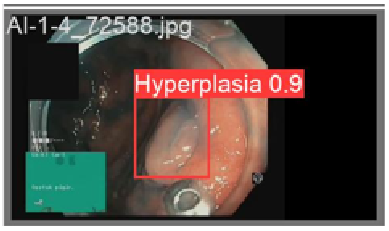


Figure 2.3: Output produced by an object detection model

Figure 2.3 shows the visual representation of the output an object detection model is supposed to produce. There is a red rectangular bounding box (generated by the model in the form of coordinates and dimensions of the bounding box), a class label which is *Hyperplasia* in this case, and a confidence score of 0.9 for the existence of that.

2.2.3 Activation functions

Activation functions are used in artificial Neural Networks (NNs) in order to transform an input signal to artificial neurons into an output signal [123], by doing some sort of a transformation on the input signal(s). The activation functions used by the DL models that are utilised in this project are :

1. **Sigmoid function** - It is one of the most widely used activation functions due to its non-linear nature [123]. It transforms the input signal/value to a value in the range of 0 to 1, and is calculated using the formula $\sigma(x) = \frac{1}{1+e^{-x}}$.

The graphical plot made by this activation function is seen in figure 2.4.

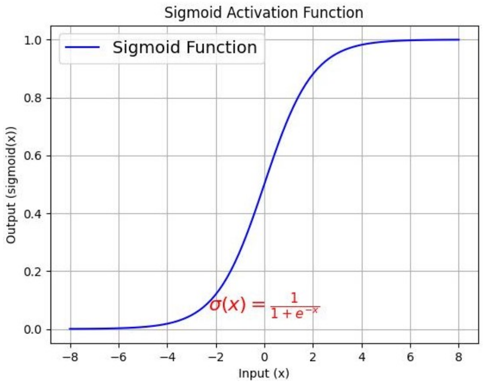


Figure 2.4: Plot of a Sigmoid activation function

2. **SiLU function** - SiLU stands for *Sigmoid-weighted Linear Unit* and is calculated using the formula $\text{SiLU}(x) = x \cdot \sigma(x)$.

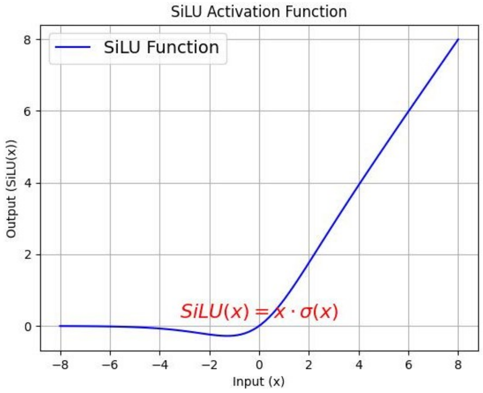


Figure 2.5: Plot of a SiLU activation function

Figure 2.5 shows the graphical plot made by this activation function for various values of input.

3. **ReLU function** - ReLU stands for *Rectified Linear Unit* and is a non-linear activation function which is represented by the formula $Relu(x) = max(0, x)$ [123].

Figure 2.6 shows the graphical plot made by this activation function.

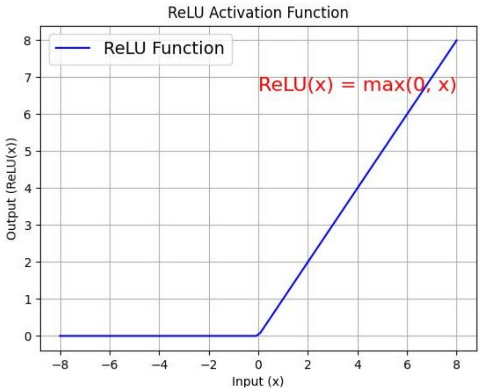


Figure 2.6: Plot of a ReLU activation function

4. **Softmax function** - It is a combination of multiple sigmoid functions, which are used to calculate the probabilities of different classes/categories in multiclass classification problems [123]. It can be represented with the following formula:

$$\sigma(x_i) = \left(\frac{e^{x_i}}{\sum\limits_j e^{x_j}} \right), j = 1, ..., n.$$

2.2.4 Loss functions

Loss functions are functions that are used to calculate the distance or difference between the present output of an algorithm or DL model, and the expected output for the same set of input values from the training dataset [107]. This calculated loss is then used to update the weights of the DL model in order to reduce the loss in the next evaluation [20]. There are several examples of loss functions that can be employed, like exponential loss, Mean Squared Error (MSE) etc. However, the important loss functions utilised by the DL models that were used in this project are briefly described below.

1. **Complete Intersection over Union (CIoU)** - This loss function pertains to bounding box regression, which is a crucial step in object detection [169]. For this loss function, there are three geometric factors associated with bounding box regression - overlap area, central point distance and aspect ratio [169] between the expected and the predicted bounding boxes that are taken into consideration and combined. This leads to faster convergence and better performance than other losses associated with bounding boxes like IoU loss or generalised IoU (GIoU) loss [169].

2. **Distribution Focal Loss (DFL)** [88] - It is a variant of focal loss that helps improve the model performance when the training data is imbalanced, which is why it is utilised in DL models like YOLOv8 to deal with class imbalance problems that might arise when training a model on a dataset that has some classes with very few training images [10, 45].

3. **Binary Cross Entropy (BCE)** - It is also known as the log loss, and it tracks the incorrect labelling of the data's class by a DL model [12]. The standard binary cross entropy function is given by the formula [57] -

$$J_{bce} = -\frac{1}{M}\sum_{m=1}^{M}\left[y_m \times \log\left(h_\theta\left(x_m\right)\right) + \left(1 - y_m\right) \times \log\left(1 - h_\theta\left(x_m\right)\right)\right]$$

where M is the number of training examples, y_m is the target label for the training example m, x_m is the input data for training example m, and h_θ is the model with neural network weights θ [57].

2.2.5 Batch Normalisation

Batch Normalisation refers to the process of normalising the inputs to a DL model's layers and making this a part of the model architecture [60]. It allows for the use of much higher learning rates during the training of a DL model and be less careful about the initialisation [60].

2.2.6 Underfitting

A DL model is termed to be *underfitting* when it is unable to learn the patterns in the data properly [98] and cannot create a mapping between the input and target variable [103].

2.2.7 Overfitting

It is a condition associated with Deep Learning where a DL model learns/gets trained to represent very well the training data and provides good results with it, but it does not perform well on new information like the test data [119].

2.2.8 Data Augmentation

Data Augmentation refers to a range of techniques that enhance the range and quality of training datasets that a DL model is trained on, so that better DL models can be built using these datasets [125]. There are various image augmentation algorithms that can be used to implement this, including (but not limited to) - geometric transformations, colour space augmentations, kernel filters, mixing images, noise injection, random cropping and random erasing [125]. This helps increase the size of the training dataset, while also providing some variance to the quality of the original dataset. Data Augmentation is an important process that is used to help DL models generalise well across different datasets, and to also help against overfitting.

2.2.9 Transfer Learning

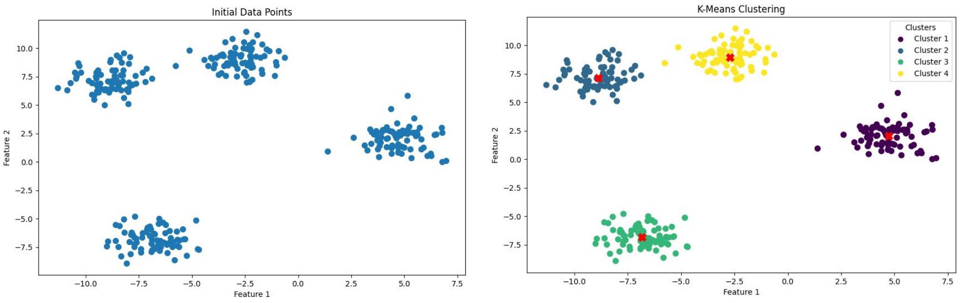
In transfer learning, the training data and testing data are not required to be *independent and identically distributed* [135], because of which there can be a transfer of knowledge from a source domain to a target domain [135].

From a more practical standpoint, transfer learning helps with the use of a pre-trained model that was trained for one task to be repurposed as the starting point for doing a new task [143]. For instance, a DL model that was pre-trained for the task of identifying the name of car brands based on an image given as input, can later be repurposed for the task of just identifying whether there was a car in an input image or not.

2.2.10 K-Means Clustering

It is an unsupervised algorithm whose aim is to separate out M data points having N dimensions (or features) into K clusters, such that the sum of squares is minimised within each cluster [52]. In the context of machine learning, it is the process of dividing a set of data points into a number of groups called *clusters* such that the data points in each group/cluster are more comparable to one another and different from the data points of other groups/ clusters [70].

Figure 2.7 shows how the algorithm (with the value of K set as 4) separates out the initial unlabeled data points having 2 features (as seen in figure 2.7a) into 4 separate clusters, with the centroid of each cluster also marked with a red cross, in figure 2.7b.



(a) Original data points

(b) After k-means clustering

Figure 2.7: Demonstration of K-Means Clustering algorithm's ability to separate the original data points in a dataset into different clusters

2.2.11 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a technique that analyses a data table that represents observations that are described by several dependent variables that tend to be correlated in nature, in general [2]. The technique's goal is to extract the important information from the aforementioned data table and to represent it as a set of new orthogonal variables known as *principal components* [2]. In the context of this project, it is used as a dimensionality reduction technique, in order to reduce the dimensionality of a dataset while preserving the most important patterns or relationships between the variables [110]. The *principal components* that are produced by this technique are linear combinations of the original variables in the dataset and are ordered in decreasing order of importance [110].

2.3 Deep Learning models used in this project

2.3.1 YOLOv8

YOLOv8 [69] is the latest version of YOLO (You Only Look Once) [114], a DL model for real-time object detection tasks that was originally developed in 2015 by Joseph Redmon et al [58, 114] that quickly gained popularity because of its balance between high speed and accuracy [139], which was obtained by leveraging a new approach to object detection where a single Neural Network (NN) was responsible for both predicting the bounding boxes for an object and its class probabilities directly from the full image in a single evaluation [114].

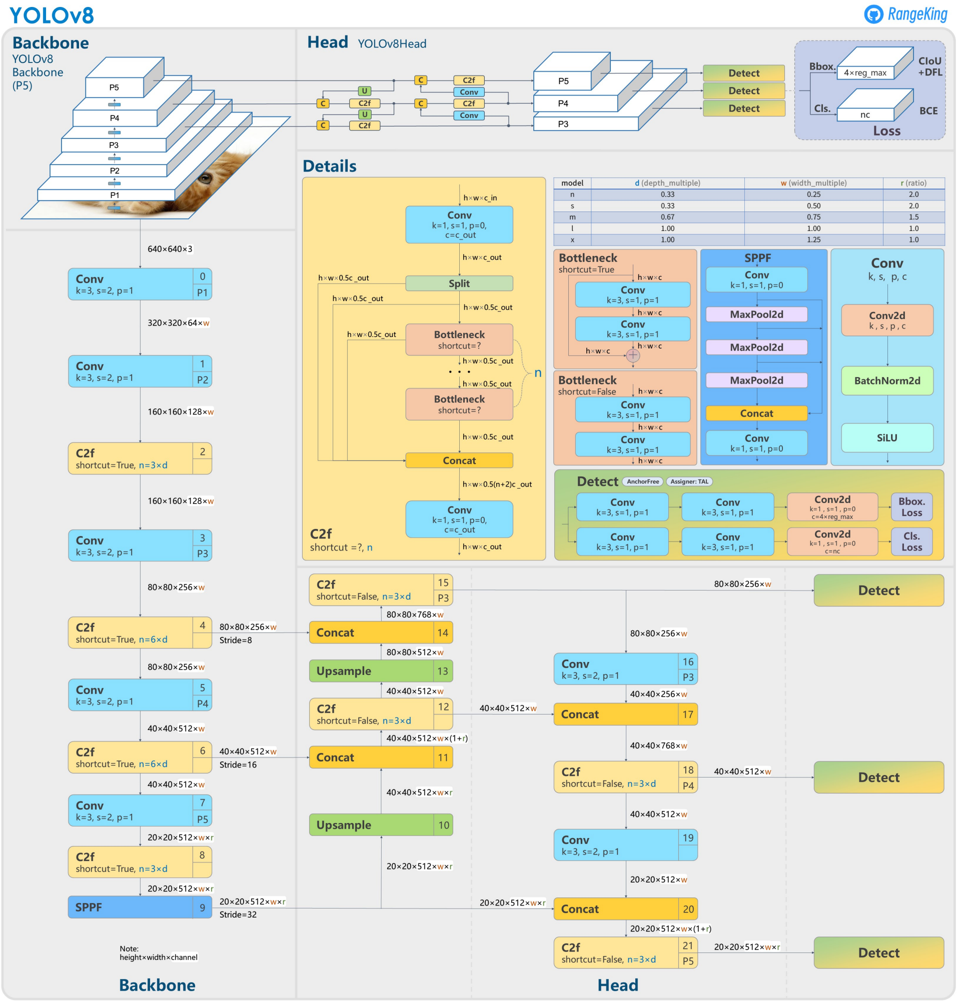
YOLO presented for the first time a real-time end-to-end approach for object detection, unlike other more time-consuming and computationally-heavy approaches for object detection like using sliding windows followed by a classifier [139]. This enabled it to be much faster than most other existing object detectors which however came with the disadvantage of a higher localisation error when compared to other state-of-the-art object detectors like Fast-RCNN [43] [139]. This, amongst other disadvantages of the initial YOLO model, was addressed over the course of the years that followed in the other versions of YOLO that were developed.

YOLOv8 builds on the limitations faced by the previous versions of YOLO in order to enhance its performance. It was developed by Ultralytics, who had also developed the YOLOv5 model, and supports a variety of tasks like object detection, image segmentation, pose estimating, object tracking and image classification [58, 139]. It also has 5 different-scale versions of the model to suit user needs - YOLOv8n (nano), YOLOv8s (small), YOLOv8m (medium), YOLOv8l (large) and YOLOv8x (extra large) [58].

Figure 2.8 visualizes the model architecture of a YOLOv8 model [111, 19], where the model can be seen as consisting of a *Backbone* and a *Head* that come together to form the whole model.

The *Backbone* consists of a series of convolutional layers that perform the task of extracting features from the input images. The C2f module that forms a part of the Backbone is the cross-stage partial bottleneck which combines high-level features with contextual information as can be seen in the figure, in order to improve detection accuracy [139, 19]. The Spatial Pyramid Pooling Fusion (SPPF) module is designed to speed up the network's computation speed by pooling different-scaled features into a fixed-size feature map [19] and consists of Convolutional and Max-Pooling layers, while the subsequent Upsample layers increase the resolution of the feature maps [19].

As for the model's *Head*, YOLOv8 uses an anchor-free model with a decoupled head [139], unlike the anchor-based models found in many other iterations of YOLO. The Detection module that forms a part of the *Head* uses a set of convolutional and linear layers in order to map the high-dimensional features from the C2f module to the output bounding boxes and object classes/labels [19]. The architecture has been designed in this manner in order to be fast and efficient while also achieving high detection accuracy [19]. The sigmoid function is used in the output layer as the activation function for the objectness score, which represents the probability that the bounding box contains an object [139]. Meanwhile, for calculating the class probabilities which represent the objects' probabilities for belonging to each possible class, the model uses a softmax function [139].



model	d (depth_multiple)	w (width_multiple)	r (ratio)
n	0.33	0.25	2.0
s	0.33	0.50	2.0
m	0.67	0.75	1.5
l	1.00	1.00	1.0
x	1.00	1.25	1.0

Figure 2.8: YOLOv8 model architecture [19, 69] (visualisation made by GitHub user RangeKing [111]). The rectangles with rounded corners represent the model's layers with their labels mentioning what kind of layers they are, along with associated parameters (like kernel size, number of channels, etc.) [19]. The arrows represent the data flow between the layers, with the direction of the arrow representing the flow of data [19].

The Complete Intersection over Union (CIoU) [169] and Distribution Focal Loss (DFL) [88] loss functions are used for bounding box prediction (by being utilised to reduce the discrepancy between the predicted and the ground-truth bounding boxes)[3], while binary cross-entropy is used for the classification loss calculation [139]. When training a YOLOv8 model, each of these losses is tracked for both the training and the validation datasets. The term *box_loss* that is generated by the model is calculated via this CIoU loss, while the *cls_loss* that is generated by the model is the classification loss [3].

2.3.2 YOLOv5

YOLOv5[44] is a predecessor of YOLOv8 [69] developed by Ultralytics (the company that also created YOLOv8). It was developed in PyTorch [139] and uses a modified CSPDarknet53 [146, 153] as the *Backbone*, which allows for more gradient flow through the network [32] and hence reduces the computation time [153].

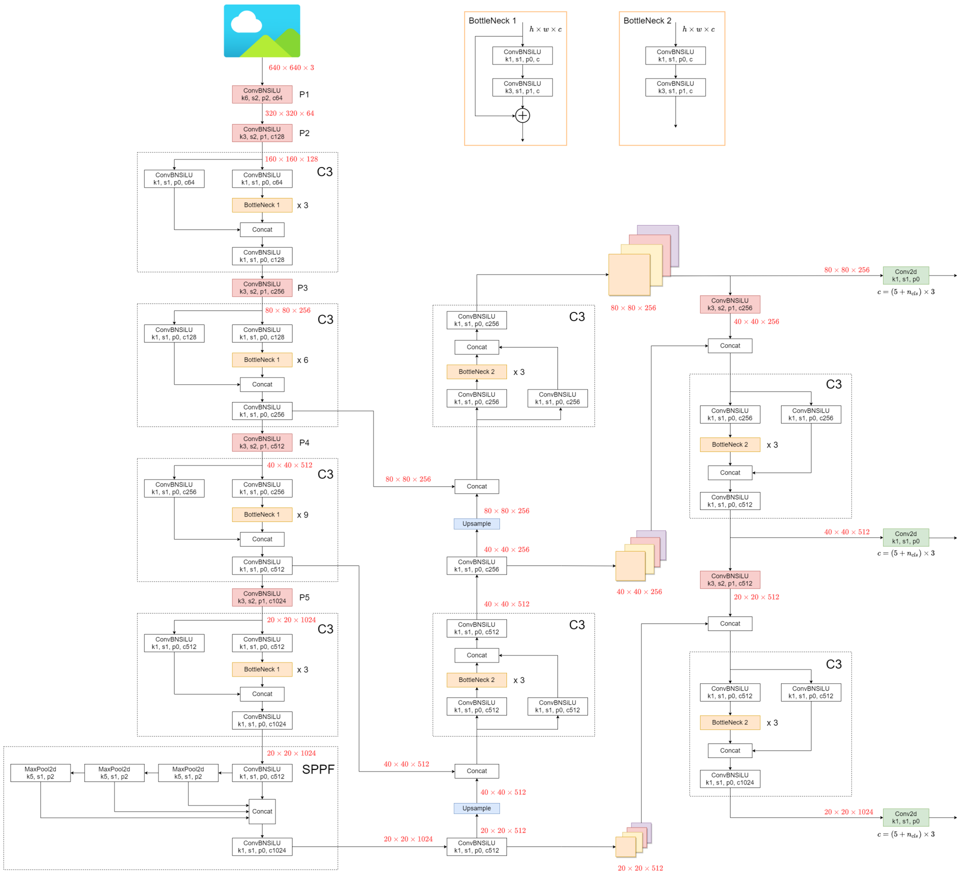

640 × 640 × 3
ConvBNSiLU k6, s2, p2, c64
P1
320 × 320 × 64
ConvBNSiLU k3, s2, p1, c128
P2
160 × 160 × 128
C3
ConvBNSiLU k1, s1, p0, c64
ConvBNSiLU k1, s1, p0, c64
BottleNeck 1
x 3
Concat
ConvBNSiLU k1, s1, p0, c128
ConvBNSiLU k3, s2, p1, c256
P3
80 × 80 × 256
C3
ConvBNSiLU k1, s1, p0, c128
ConvBNSiLU k1, s1, p0, c128
BottleNeck 1
x 6
Concat
ConvBNSiLU k1, s1, p0, c256
ConvBNSiLU k3, s2, p1, c512
P4
40 × 40 × 512
C3
ConvBNSiLU k1, s1, p0, c256
ConvBNSiLU k1, s1, p0, c256
BottleNeck 1
x 9
Concat
ConvBNSiLU k1, s1, p0, c512
ConvBNSiLU k3, s2, p1, c1024
P5
20 × 20 × 1024
C3
ConvBNSiLU k1, s1, p0, c512
ConvBNSiLU k1, s1, p0, c512
BottleNeck 1
x 3
Concat
ConvBNSiLU k1, s1, p0, c1024
20 × 20 × 1024
SPPF
MaxPool2d k5, s1, p2
MaxPool2d k5, s1, p2
MaxPool2d k5, s1, p2
ConvBNSiLU k1, s1, p0, c512
Concat
ConvBNSiLU k1, s1, p0, c1024
BottleNeck 1
$h \times w \times c$
ConvBNSiLU k1, s1, p0, c
ConvBNSiLU k3, s1, p1, c
BottleNeck 2
$h \times w \times c$
ConvBNSiLU k1, s1, p0, c
ConvBNSiLU k3, s1, p1, c
C3
ConvBNSiLU k1, s1, p0, c256
Concat
BottleNeck 2
x 3
ConvBNSiLU k1, s1, p0, c256
ConvBNSiLU k1, s1, p0, c256
80 × 80 × 256
Concat
80 × 80 × 256
Upsample
40 × 40 × 256
ConvBNSiLU k1, s1, p0, c256
40 × 40 × 512
C3
ConvBNSiLU k1, s1, p0, c512
Concat
BottleNeck 2
x 3
ConvBNSiLU k1, s1, p0, c512
ConvBNSiLU k1, s1, p0, c512
40 × 40 × 512
Concat
40 × 40 × 512
Upsample
20 × 20 × 512
20 × 20 × 1024
ConvBNSiLU k1, s1, p0, c512
80 × 80 × 256
40 × 40 × 256
20 × 20 × 512
80 × 80 × 256
Conv2d k1, s1, p0
$c = (5 + n_{cls}) \times 3$
ConvBNSiLU k3, s2, p1, c256
40 × 40 × 256
Concat
C3
ConvBNSiLU k1, s1, p0, c256
ConvBNSiLU k1, s1, p0, c256
BottleNeck 2
x 3
Concat
ConvBNSiLU k1, s1, p0, c512
40 × 40 × 512
Conv2d k1, s1, p0
$c = (5 + n_{cls}) \times 3$
ConvBNSiLU k3, s2, p1, c512
20 × 20 × 512
Concat
C3
ConvBNSiLU k1, s1, p0, c512
ConvBNSiLU k1, s1, p0, c512
BottleNeck 2
x 3
Concat
ConvBNSiLU k1, s1, p0, c1024
20 × 20 × 1024
Conv2d k1, s1, p0
$c = (5 + n_{cls}) \times 3$

Figure 2.9: YOLOv5 model architecture. Image sourced from Ultralytics Docs[146] (Authors - Glenn Jocher[45], Sergiu Waxmann [122])

The *Backbone* of the YOLOv5 model has convolutional layers that extract the relevant features from the input image [139]. Just like YOLOv8, the Spatial Pyramid Pooling Fusion (SPPF) module consisting of Convolutional and Max-pooling layers is used to speed up the network's computation speed by pooling different-scaled features into a fixed-size feature map [139], while the Upsample layers are used to increase the resolution of the feature maps [139]. Each convolution is followed by Batch Normalisation (BN) (to accelerate the training of the model [60]) and SiLU activation function[55] [139], and these together are represented as ‘ConvBNSiLU’ in figure 2.9. The *BottleNeck* modules that form a part of the architecture also consist of these ‘ConvBNSiLU’ layers, as seen in figure 2.9.

The *Neck* (which utilizes the *SPPF* and *New CSP-PAN* structures) [146] and *Head* (which resembles the head of YOLOv3 [139]) form the remainder of the model architecture. Unlike YOLOv8, it is not an anchor-free approach. Hence, the loss value calculation for this model is slightly different to YOLOv8.

The loss value in the YOLOv5 model is computed as a combination of 3 individual loss components [146]:

- Classes Loss - A Binary Cross Entropy (BCE) loss to measure the error in the classification task [146].

- Objectness Loss - A Binary Cross Entropy (BCE) loss to measure the error in detecting if an object was present in a particular grid cell or not [146].

- Location Loss - A Complete Intersection over Union (CIoU) loss to measure the error in localising the object within the grid cell [146].

Just like with YOLOv8, these losses are also kept track of while training a YOLOv5 model, for both the training and validation datasets.

There are a few terms defined for both YOLOv8 and YOLOv5 models that were needed to make predictions using these models -

- IoU: It is the ratio between the area of intersection to the area of union of the model’s predicted bounding box, and its ground truth bounding box. It has a value between 0 and 1.

- Box confidence: It is a measure of how certain the model is that a bounding box contains an object of interest, and is calculated by combining the objectness score (the model’s certainty that a predicted bounding box contains an object at all) with the IoU mentioned above [59].

- Class confidence: It is a measure of how certain the model is that the object detected by the model belongs to a certain class, and is calculated by taking the conditional probability of the class given that an object has been detected, and then multiplying it with the objectness score and the IoU [59].

- Confidence score: It is an output from the YOLOv8 model that is a combination of the Box confidence and Class confidence values, hence enabling it to balance between how certain it is that a box contains an object and how certain it is about which class that object belongs to [59]. This is one of the parameters this project will focus on, when evaluating the performance of the model.

In both YOLOv5 and YOLOv8 models, there is a certain *Confidence threshold* value and a *IoU threshold* value for a potentially-detected object's bounding box and class label to be over, in order for an object to be considered detected. These thresholds are used to determine the final predicted bounding box from multiple bounding boxes for a specific object [159].

2.3.3 VGG-16/ VGG-19

The Visual Geometry Group (VGG) at the University of Oxford developed the two models VGG-16 and VGG-19, first introduced in the paper "Very Deep Convolutional Networks for Large-Scale Image Recognition" [127] by K. Simonyan and A. Zisserman.

It is a deep CNN architecture with a large number of layers, with VGG-16 having 16 weight layers and VGG-19 having 19 weight layers, with each convolutional layer having small 3*3 filters which enables the model to have such an increased depth in the number of layers [127]. It was initially designed with a focus on the ILSVRC-2012 dataset for an image classification task with the dataset consisting of 1000 classes, which is why the final layer in the architecture is a softmax layer designed for a 1000-way classification task [127].

The input to the model is a fixed-size 224*224 RGB image. Figure 2.10 shows the architecture of a VGG-16 model, which consists of 13 convolutional layers and 3 fully-connected (FC) layers, giving a total of 16 weight layers. Similarly, the VGG-19 model has 16 convolutional layers and 3 fully-connected layers, giving it a total of 19 weight layers. Apart from the 3*3 convolution filters, there are also 1*1 convolution filters acting as linear transformation for the input [127, 15]. This is followed by the ReLU (Rectified Linear Unit) activation function [127, 15]. Convolution stride is set at 1 in order to preserve the spatial resolution, while the max-pooling is performed over a 2*2 pixel window with stride as 2 [127]. The number of channels for the convolutional layers starts at 64 and increases in multiples of 2 after every max-pooling layer till it reaches 512 [127], as seen in figure 2.10. The convolutional layers are followed by the fully-connected layers, with the first two FC layers having 4096 channels each, while the third FC layer has 1000 channels since it performs a 1000-way image classification, and so has 1 channel for each class of the dataset [127]. All hidden layers have the ReLU activation function [127]. The final layer is a softmax layer, which gives a value between 0 and 1 which corresponds to

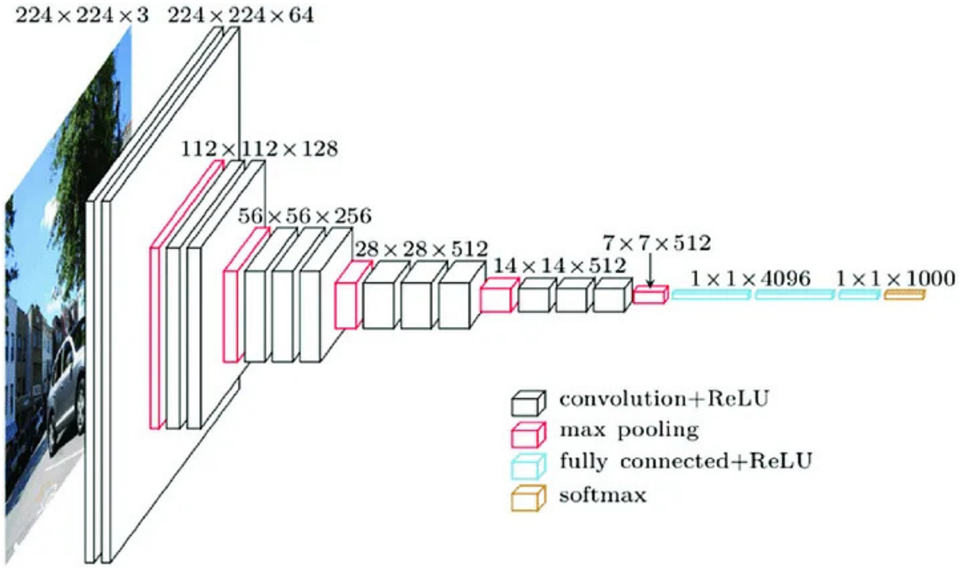


Figure 2.10: VGG-16 model [127] architecture (Reprinted from Computers in Industry, Volume 108, Solemane Coulibaly, Bernard Kamsu-Foguem, Dantouma Kamissoko, Daouda Traore, "Deep neural networks with transfer learning in millet crop images", Pages 115-120, (2019), with permission from Elsevier." [30])

the confidence of the model about the image belonging to each class of the dataset.

2.4 Evaluation Metrics

Evaluation metrics refer to a set of metrics and results that are used to assess and compare the results that are obtained. There were a few metrics associated with the *Object Detection* task that were used in the project to assess the quality of the YOLOv8 and YOLOv5 models being trained. These metrics apply to the bounding box and class label predictions generated by these models. These were -

- Precision (P): A measure of the accuracy of the detected objects, which indicates how many detections were correct [147]. It quantifies the proportion of true positive predictions amongst all the positive predictions [147].

- Recall (R): A measure of the ability of the model to identify all the instances of objects in the images [147]. It quantifies the proportion of true positives amongst all the actual positives [147].

- Precision-Recall Curve: It is a graphical representation of the trade-offs between the precision and recall at varied thresholds, in the form of the curve [147].

- Average Precision (AP): It computes the area under the precision-recall curve, hence providing a single value to indicate the model's precision and recall performance [147].

- Mean Average Precision (mAP): It extends the idea of AP by calculating the average AP values across multiple object classes in a dataset, and is especially useful in multi-class object-detection scenarios to gain an insight into the model's performance [147].

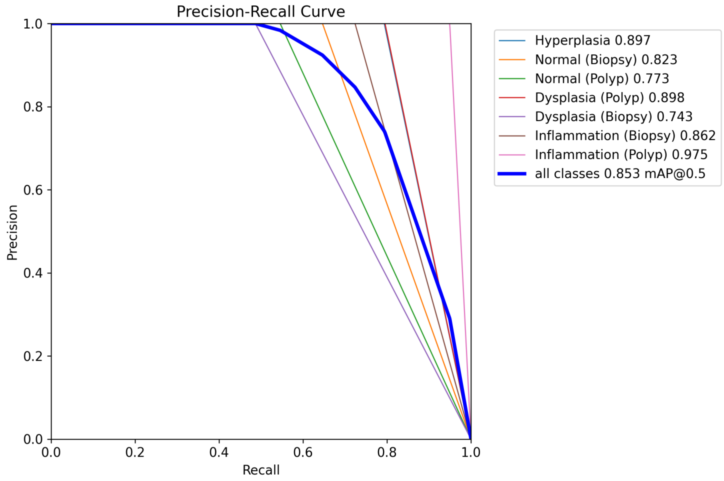


Figure 2.11: An example of a Precision-Recall curve. There are 7 different classes here represented with a uniquely coloured curve, along with a bold blue curve for all classes. The APs for the curve of each class, as well as the mAP associated with the curve for all classes, are mentioned in the legend.

- mAP50: It is the mAP calculated at an IoU threshold of 0.50 [147]. mAP50 gives more of an insight into the accuracy of the model when it comes to the easier-to-detect objects.

- mAP50-95: It is the average of the mAP values calculated at various IoU thresholds between 0.50 and 0.95, and gives a more comprehensive view of the model's performance [147].

- Confusion Matrix: It is a visual representation of a detailed view of the outcomes of the object-detection model, showcasing the number of True Positives, True Negatives, False Positives and False Negatives for each class [147]. With the YOLOv5 and YOLOv8 models, an additional class called ‘Background’ was added to the confusion matrix, to represent all the instances where no object was present.

- Normalised Confusion Matrix: It is just a normalised version of the confusion matrix, with the data in it present in proportions instead of raw counts [147]. It has been manipulated for the sake of this project to give the data in percentages instead of fractions.

2.5 Related Work

There has been some promising previous research work done in the field of both cancer detection in general, and CRC detection as well. These are discussed in detail in the following subsections.

2.5.1 Deep Learning in Cancer Detection

When it comes to aiding cancer diagnosis, DL models have shown some promising results over the years over multiple studies and surveys [91, 36, 39, 124, 11].

Mambou et al. (2018) [91] did a thorough comparative study of several Computer Vision (CV) and Deep Learning (DL) powered breast cancer detection models. Their work highlighted the importance of image processing which was not performed to adequately sufficient levels by Artificial Intelligence (AI) methods, relative to the human performance of the same. Hence, they ended up proposing a Computer Assist Device (CAD) that uses a new model based on a pre-trained Inception V3 [130] deep neural network with a Support Vector Machine (SVM) [100] classifier coupled to that. The CAD would take thermal images of breasts as input and give an output of whether the input was healthy (without cancer) or sick (with cancer), along with the associated probability of the same. Such a CAD system holds promise for other forms of cancer detection, like CRC.

Dildar et al. (2021) [36] also did a systematic review of various neural network techniques for skin cancer detection and classification. Various models based on Artificial Neural Networks (ANNs) [64], Convolutional Neural Networks (CNNs) [101], Kohonen Self-Organizing Neural Networks (KNNs) [148] and Generative Adversarial Networks (GANs) [46] were explored, along with their advantages and associated challenges. Their work noted how CNNs tended to perform better when it came to the task of image classification.

2.5.2 Deep Learning in polyp/precancerous-lesions detection

Even when it comes to CRC-related polyp detection, there has been a fair amount of research already done which show a few promising results [67, 104, 87, 90, 165, 83].

Li et al. (2021) [87] did a thorough evaluation and comparison of eight state-of-the-art object detectors based on DL models on a relatively large dataset for polyp detection that they had developed, where YOLOv4 [14], ATSS [166] and RefineDet [167] seemed to be the models that gave a well-balanced and consistent performance.

Liu et al. (2019) [90] meanwhile took a slightly different approach towards polyp detection in colonoscopy videos. They tried handling the challenge of limited training data for the training of DL models for polyp/CRC-detection by proposing a novel single-shot detector (SSD) framework-based method with 3 different feature extractors for polyp detection. Their experimentation on the ETIS-Larib [126] dataset showed a significantly high detection speed compared to the other methods that were tried, as well as a high detection rate for True Positives (TPs). Their work also showed the excellent performance of InceptionV3 [131] as a DL model for feature extraction in their SSD-based method.

Jha et al. (2019) [68] proposed a ResUNet++ based DL model for developing a fully automated model for pixel-wise polyp segmentation in colonoscopy videos and images. Comprehensive tests performed on different datasets using this architecture demonstrated how it outperformed State of the Art (SOTA) U-Net and ResUNet architectures when it came to producing semantically accurate predictions.

Jha et al. (2021) [65] later also built on this and demonstrated how Conditional Random Field (CRF) [5] and Test-Time Augmentation (TTA) can be used to further improve the performance of the previously discussed ResUNet++ architecture. This new framework showed an improved performance on various polyp segmentation datasets that it was tested on. Furthermore, this new framework overcame a major hurdle when it comes to CRC polyp detection, by showing good results for flat or sessile and smaller polyps. These are the polyps that are much harder to detect and contribute to a much higher polyp miss rate in colonoscopies.

Jha et al. (2021) [67] also used the Kvasir-SEG [66] dataset to benchmark the performance of various SOTA deep learning methods for polyp detection, localisation and segmentation tasks. They also proposed a DL model called

ColonSegNet for these tasks. It is an encoder-decoder model that uses residual blocks with squeeze and excitation network as the main component. This gives it the advantage of having very few trainable parameters compared to other SOTA models, hence resulting in a very light-weight network that can give real-time performance with a relatively high frame rate. For the Kvasir-SEG dataset on which the benchmark tests were done, the highest frame rate was obtained by the ColonSegNet model.

Pacal & Karaboga (2021) [104] proposed a YOLOv4[14]-based model for real-time automatic polyp detection that showed higher accuracy and performance compared to contemporary methods.

2.5.3 Deep Learning based image clustering

Caron et al. (2018) [23] presented a clustering method called DeepCluster which uses a standard clustering algorithm like k-means to cluster the features generated by a convolutional neural network on a dataset like ImageNet [34], followed by updating the weights of the neural network by using the cluster assignments produced earlier as pseudo-labels for prediction. This was a promising approach for domains where labels and annotations are scarce [23].

Yang et al. (2016) [163] proposed a framework for the joint unsupervised learning of deep representations as well as image clusters, with the image clustering done during the forward pass of the training process, while the deep representation learning was done during the backward pass of the training process. Their work showed that the model outperformed the state-of-the-art on image clustering across a large variety of datasets like MNIST [81] and COIL20[99].

2.5.4 Colour-channel separation of images

Apart from research work on just finding the most accurate or efficient DL models for CRC-related polyp or precancerous lesion detection, there has also been research conducted on the kind of pre-processing steps that colonoscopy videos can go through in order to improve the performance of the DL models in their tasks. Pre-processing of data (be it images or videos) has always been an important step most DL models need the data fed to them to go through, before the model can perform its task - be it classification, segmentation or anything else. Pre-processing in deep learning tasks related to images/video input usually involves steps like resizing the images to a certain fixed dimension, conversion to greyscale and normalisation of pixel values. However, there has been evidence to suggest that the separation of colour channels of 3-channel

RGB images and then combining a few channels together to form a new image to be fed into the DL network can improve the performance of the DL model [78, 73, 48, 62].

Lai et al. (2021) [78] did a study on this for endoscopy images of colon mucosa and polyps, for CRC detection. 3-channel RGB images that were acquired using Narrow-Band Imaging (NBI) [76] and White Light Endoscopy (WLE) were studied. From each image, 1-channel, 2-channel & full-colour versions of images were extracted separately, with the Deep Neural Network (DNN) being trained on each of these combinations separately to see which showed the best results. For WLE images, it was observed that the DNN performed much better using 2-channel Red+Green (R+G) channel images, when compared to the full-colour 3-channel RGB WLE images. Meanwhile, for NBI images, the performance of the R+G images was almost the same as that for the full-band images. The improvement in results for R+G WLE images suggests that colour-channel separation is a promising avenue to conduct further research on, to see if it can boost the performance of the DL model.

Kim et al. (2022) [73] also did a similar study, but this time for cervical cancer classification using a DL model. Here, apart from the original image, there were also Acetowhite Mask Images associated with each original image. The pre-processing model they proposed involved splitting both the original and the mask images into their constituent 3 channels each and then merging 2 channels from the original image and 1 channel from the mask image to form a superimposed image. This superimposed image would eventually get fed to a DL model for a classification task. Here the DL model used was ResNet [53]. Using this method helped increase the accuracy of the DL model from 72% using the original image to 81% using the superimposed image (R channel of acetowhite mask image + R&B channels of the original image, superimposed together). An approach similar to this could be tried in the case of colonoscopy images, to test out the potential of improving the performance of the DL models.

Similarly, Gupta & Manhas (2021) [48] did a similar study related to oral-cancer detection using DL. They proposed a DL framework that involved splitting the colour channels of the images and then extracting deep features from these individual channels rather than a single combined channel, using the Efficient Net B3 [136] DL model. The extracted features were later fused together using a fusion module layer in the DL model. Using this method helped give a much higher accuracy for oral cancer detection, compared to simply using the original 3-channel images.

All these studies conducted on the prospective benefits of colour-channel separation and the associated improvement in performance it provides suggest that trying different colour-channel combinations for the input to the DL model is as important a step as trying to find the best-performing DL model for that particular dataset.

3

IBD, Polyps and Colonoscopy

3.1 Inflammatory Bowel Disease (IBD)

Inflammatory Bowel Disease (IBD) is a term for two conditions - Crohn's Disease (CD) and Ulcerative Colitis (UC), that is characterised by chronic inflammation of the gastrointestinal (GI) tract [40]. IBD is associated with significant GI symptoms including diarrhoea, abdominal pain, bleeding, anaemia and weight loss [109]. IBD also happens to be associated with a spectrum of extraintestinal manifestations in the form of arthritis, ankylosing spondylitis, sclerosing cholangitis, uveitis, iritis, pyoderma gangrenosum, and erythema nodosum [109]. The risk of developing CRC for people with IBD increases by 0.5-1 % yearly, 8-10 years after diagnosis [97]. Approximately 5-10 % of IBD patients develop CRC after 20 years and 12-20 % after 30 years of the disease [97, 79].

IBD should not be confused with IBS (Irritable Bowel Syndrome), a condition that affects 10% of adults, with a female predominance [134]. It is a common, long-term condition of the digestive system and its symptoms include cramping, abdominal pain, bloating, gas, and diarrhoea or constipation, or sometimes both [61]. Although the condition can often be lifelong, the symptoms may change over time and it can be successfully managed with the right strategies [61]. Unlike IBD, IBS does not pose any serious threat to one's physical health and does not increase an individual's chances of developing CRC or other bowel-related conditions [61].

Endoscopy is considered to be the gold standard for the initial diagnosis of IBD [8]. Patients with long-standing UC and CD have an increased risk of developing colorectal cancer (CRC) [144]. This is why it is recommended for chronic IBD patients to undergo surveillance colonoscopies 8 years after diagnosis, except for the cases of ulcerative proctitis or proctosigmoiditis, which can be screened according to average-risk population guidelines [17]. Wehkamp et al (2016) [156] also mention how colonoscopy should be used to monitor for *dysplasia* (presence of abnormal cells within a tissue or organ) starting 8-10 years after the initial manifestation of either type of IBD. Chromoendoscopic surveillance with methylene blue dye [72] or indigo carmine dye [116] is better suited to this than white-light endoscopy [129]. However, chronic IBD patients tend to have scarring and inflammation of their colorectal tissue, which leads to something of a noisy background during colonoscopies, hence making it more challenging to do colorectal screening for precancerous lesions in IBD patients when compared to patients with a healthy colon. This can be further complicated by *intestinal fibrosis*, a condition where excessive scar tissue can accumulate on the intestinal walls [115].

3.2 Polyps

Polyps are a group of cells that abnormally grow on the inner surface of a colon, although they can occur anywhere in the gastrointestinal (GI) tract. They can grow to eventually lead to CRC [132].

Colorectal polyps are histologically classified into two major classes - *neoplastic* & *non-neoplastic*[28]. Non-neoplastic polyps are further subdivided into 4 other categories - hyperplastic polyps, hamartomas, lymphoid aggregates & inflammatory polyps, each of which are non-cancerous [28]. On the other hand, neoplastic polyps are similarly classified into tubular adenomas, tubulovillous adenomas & villous adenoma [28]. Each of these carries the potential to be cancerous [28].

3.2.1 Paris Classification of polyps

In 2002, an international consortium of endoscopists, surgeons, and pathologists gathered in Paris and developed the Paris classification of early and/or superficial tumours in the GI tract [80, 145, 106]. Figure 3.1 below shows how the Paris classification is used to segregate colorectal polyps into 6 different types [80, 106].

Out of these, types 0-Ip, 0-Is, 0-IIa & 0-IIb are less likely to be precancerous or cancer; type 0-IIc is likely to be precancerous/cancer (when associated with 0-Ip, 0-Is, 0-IIa or 0-IIb) while type 0-III is highly likely to be cancer [16].

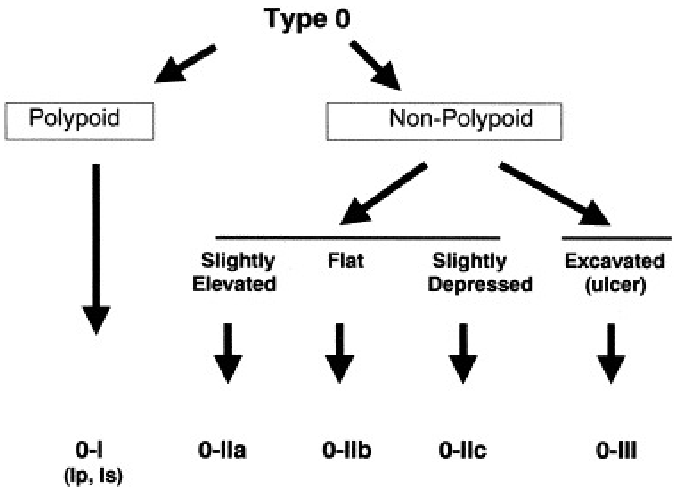


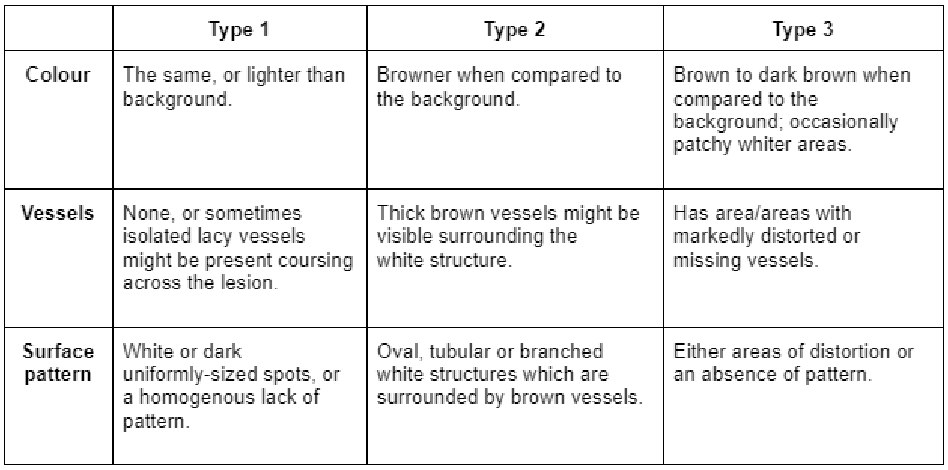
Figure 3.1: PARIS classification of polyps (Reprinted from Gastrointestinal Endoscopy, Volume 58, Issue 6, Supplement, Participants in the Paris Workshop, "The Paris endoscopic classification of superficial neoplastic lesions: esophagus, stomach, and colon: November 30 to December 1, 2002", Pages S3-S43, (2003), with permission from Elsevier)[106]

3.2.2 NICE Classification of polyps

Meanwhile, around 2010, an international consortium group consisting of members from Japan, the USA and Europe called the *Colon Tumor NBI Interest Group* developed a simple category classification for colorectal polyps [137]. This new classification was called the *NBI international colorectal endoscopic (NICE)* classification and it classified colorectal polyps into types 1-3 by close observation of colorectal tumours using a high-resolution videocolonoscope [137].

The table below (Table 3.1) shows how polyps are categorised differently according to the NICE classification, as well as the associated characteristics of each of these categories of polyps [137].

Using this classification, the most common pathology for each type of polyps is - Hyperplastic for NICE Type 1, Adenoma for NICE Type 2 & Deep submucosal invasive cancer for NICE Type 3 [38].



	Type 1	Type 2	Type 3
Colour	The same, or lighter than background.	Browner when compared to the background.	Brown to dark brown when compared to the background; occasionally patchy whiter areas.
Vessels	None, or sometimes isolated lacy vessels might be present coursing across the lesion.	Thick brown vessels might be visible surrounding the white structure.	Has area/areas with markedly distorted or missing vessels.
Surface pattern	White or dark uniformly-sized spots, or a homogenous lack of pattern.	Oval, tubular or branched white structures which are surrounded by brown vessels.	Either areas of distortion or an absence of pattern.

Table 3.1: NICE classification of polyps (Adapted with permission from Digestive endoscopy : official journal of the Japan Gastroenterological Endoscopy Society, 23 Suppl 1, 131–139. Tanaka, S., & Sano, Y. (2011), "Aim to unify the narrow band imaging (NBI) magnifying classification for colorectal tumors: current status in Japan from a summary of the consensus symposium in the 79th Annual Meeting of the Japan Gastroenterological Endoscopy Society", Copyright (2011)) [137]

3.3 Precancerous colorectal lesions

Premalignant or precancerous colorectal lesions are a collection of cells that have undergone changes that make them more likely to develop into cancer, although they are not yet cancer [84]. CRC is a tumour that develops from the progression of acquired or hereditary premalignant lesions [29]. About 70% of all CRCs are sporadic, perhaps attributed to unidentified genetic factors beyond the risk factors associated with CRC, and *colorectal adenomas* are a precursor of almost all sporadic CRCs [29].

Colorectal adenomas are typically asymptomatic lesions and are often found incidentally during colonoscopies [29]. Not all colonic polyps are adenomas and more than 90% of adenomas do not progress to cancer, yet it is concerning that colorectal adenomatous polyps develop in up to 40% of people over the age of 60 years [29, 86].

Advanced adenomas are usually considered to be the clinically relevant precursors of CRC and need to be removed once they are detected [18]. Advanced adenoma is considered to be present in a patient if they have at least one adenoma with at least one of the following features: size greater than 1 cm, tubulovillous or villous adenoma, high grade dysplasia or invasive cancer [18, 117]. The transformation rate of adenomatous polyps into *carcinoma* (cancer that forms in the epithelial tissue) is around 0.25% per year [29]. Moreover, the (projected) annual transition rate from advanced adenomas to CRC increases strongly with age (from 2.6% in the age group 55-59 years to 5.6% in the age group of over 80 years among women, and from 2.6% in the age group 55–59 years to 5.1% in the age group of over 80 years among men) [18].

There are a few pathological features such as size, architectural growth, type and dysplastic grade and organisation which are predictive of both the natural history of lesions and the time frame of their potential evolution from *adenoma* to *carcinoma* [29].

The size of the adenoma is a relevant determinant, since cancer develops in 1% of all adenomas < 1 cm, in 10% of adenomas > 1 cm & < 2 cm, and in 50% of adenomas > 2 cm in size [29].

The histological features determining the malignant potential of an adenoma are its growth pattern and the grade of dysplasia, with the risk of malignant transformation increasing to almost 50% in adenomas with a mainly villous architectural configuration [29].

Between 10-15% of sporadic CRCs are likely to have their origins in serrated polyps, which tend to have a significant malignant potential [29]. Serrated polyps include hyperplastic polyps (which form 80-90% of the cases) but also sessile serrated adenomas (which prevail in about 5-10% of the general population [95]), traditional serrated adenomas and mixed polyps displaying features of both [29].

There is a difference in the malignant potential of sessile serrated adenomas, traditional serrated adenomas, and conventional adenomas, reflecting the differences in their molecular pathways of carcinogenesis [29]. Histological assessment suggests that there is a significantly lower degree of high-grade dysplasia and carcinoma *in situ* for serrated adenomas in comparison to traditional adenomas, which is why serrated adenomas are much less likely to develop into CRC than traditional adenomas [85, 29].

The risk of cancer from colorectal adenoma is eliminated when it is completely removed, even if the discovery of the adenoma indicated the potential risk of *metachronous lesions* (cases in which the second primary cancer is diagnosed more than 6 months after the diagnosis of the first primary cancer) with variable potential for malignancy, depending on the endoscopic and histological features [29].

3.4 Colonoscopy

One of the foremost ways to better prevent the occurrence of CRC is the early detection and removal of polyps [24]. The method most commonly used to do this is a ***colonoscopy***. *Colonoscopy* is a visual examination of the entire length of the colon and rectum with an endoscope that has a camera attached to its end which is connected to a monitor for the examiner to observe [77]. This test allows a direct mucosal inspection of the entire colon, and can be accompanied by the removal of polyps or a biopsy [77].

The current clinical practice for detecting colorectal polyps is by using conventional *White Light Endoscopy (WLE)*, which can be combined with dyes in order to enhance the visualisation of the tissues in the areas being inspected [108]. This process is called ***chromoendoscopy*** [108]. The removed polyps are then sent for a histopathological examination at a laboratory to determine if it is an adenoma (and so has a higher risk of cancer) or if it is hyperplastic (and so a lower risk of cancer) [108]. The stains used for dye-based chromoendoscopy are of 2 major categories - either absorptive stains (such as methylene blue [72]) or contrast stains (such as indigo carmine [116]) [21]. Methylene blue gets absorbed by the epithelial cells of the small or large intestine which stain blue, as opposed to dysplastic and cancerous lesions which tend to remain unstained [21]. Indigo carmine is a dark blue stain that tends to highlight mucosal topography by coating mucosal structures, pits, erosions and depressions [21]. The application of these agents appears to enhance the detection and discrimination of lesions by better defining the mucosal surface and light-absorptive patterns [21].

Virtual chromoendoscopy (VCE) is another technique which has been developed to provide an enhanced visualisation of tissues without needing any dyes, unlike conventional chromoendoscopy. VCE involves electronic endoscopic imaging techniques which help provide a much more detailed contrast enhancement of the mucosal surface and blood vessels of the colorectal region [108]. It usually makes use of an endoscope, a light source, a video processor and a visual display monitor [151].

VCE technologies can either be optical or digital in nature. *Optical VCE* consists of optical lenses being integrated into the endoscope's light source that helps selectively filter white light and results in narrow-band light [108, 92]. Meanwhile, *Digital VCE* involves using digital post-processing by the video processor in order to enhance the real-time image being displayed [92].

3.4.1 Commercial Endoscopy Systems

There are 3 major endoscopy systems used commercially, for the purpose of colonoscopies -

1. **The Olympus Endoscopy System** - Developed by Olympus Medical Systems Corp. Tokyo, Japan. Narrow Band Imaging (NBI), a type of optical chromoendoscopy technology is used in a few of the Olympus endoscopic video imaging systems like EVIS LUCERA ELITE, EVIS EXERA III & EVIS LUCERA SPECTRUM [108]. NBI helps enhance the contrast between the blood vessels & the surrounding mucosa, when compared to standard WLE. The system allows the endoscopist to switch between standard white light to NBI and vice versa at any time [108].

2. **Fuji Endoscopy System** - Developed by HC21 Aquilant Endoscopy & FujiFilm (Europe) GmbH, Willich, Germany. Flexible Spectral Imaging Colour Enhancement (FICE), a type of digital chromoendoscopy, is used in a few of these systems like EPX-4450HD, EPX-3500HD and EPX-4400 [108]. Here, white light is used to illuminate an area of interest before the images captured from the reflected light are processed by the system's software into spectral images [108].

3. **PENTAX Endoscopy System** - Developed by PENTAX Europe GmbH, Hamburg, Germany. A digital chromoendoscopy technique called *i-scan* is used in them, where white light illuminates an area of interest and then 3 different algorithms for surface enhancement, contrast enhancement and tone enhancement are applied for real-time image processing [108].

3.4.2 Advanced colonoscopy techniques

Once carcinoma is formed, there are five stages for CRC, as shown in figure 3.2 [118]. Stage 0 is when the tumour is located in the mucosal layer of the colon, stage I is when it reaches the muscularis layer, stage II is when it just perforates the serosa, stage III is when the surrounding lymph nodes are also involved, and lastly, stage IV is with distal metastasis [118]. In CRC screening, ideally, the precancerous lesions (or also up to Stage I cancer) should be discovered so that they can be resected endoscopically, and for this to be possible, better technology to discover earlier lesions is required [118].

It is of particular difficulty to discover mucosal flat lesions, especially in the right colon, hence leading to some of them getting overlooked during colonoscopies [118]. So, there is a need for extra modalities like chromoendoscopy to appreciate the small, flat lesions [118]. With time, endoscopy systems have

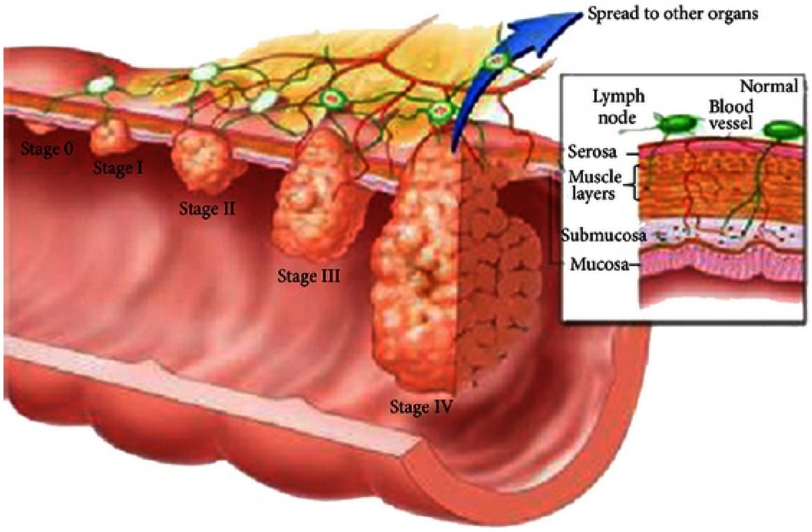


Figure 3.2: Stages of CRC (Reprinted from Sandouk, F., Al Jerf, F., & Al-Halabi, M. H. (2013). Precancerous lesions in colorectal cancer. Gastroenterology research and practice, 2013, 457901. https://doi.org/10.1155/2013/457901 [118])

developed in their functionalities by offering better resolution in their video feed and better zoom functionalities. For instance, high definition (HD) digital chromoendoscopy systems like Pentax i-Scan, Olympus NBI and Fujinon Fice can be used to help with detecting the small, flat lesions better [118]. This technology accentuates suspicious mucosal structures, hence providing us with better delineation of borders and better vasculature patterns by enhancing the minute mucosal and vessel structures [118]. Sometimes, with the superficial enhancement technology, it can give us ideas about the cytological pathology, especially when it is coupled with the advanced magnification/zooming abilities of these endoscopy systems [118]. All these advancements in modern chromoendoscopy systems better aid the detection of lesions.

4

Methods and Methodologies

4.1 Philosophical Paradigm

The **philosophical paradigm** that forms the basis of this project is ***critical realism***[37, 128]. *Critical realism* states that the evidence we observe can come close to reality but is always a fallible, social and subjective account of reality, and that human perspectives are always 'accounts of reality' [128]. This stands true in the case of this project too. With a focus on colonoscopies, the ability to observe a precancerous lesion in a colonoscopy video is subjective- influenced by the person performing the endoscopy, the tools that the person has available, the resolution of the video feed and other similar factors. Even with the DL System that is created and tested over the course of this project, the observed results from it will just be another 'account of reality' instead of being an objective representation of what is actually happening. This is why *critical realism* is the philosophical paradigm that this project builds from.

4.2 Research Methods

Research methods tend to provide procedures for accomplishing the research tasks associated with a project like this in order to support the process of conducting research [49].

A combination of ***applied research method*** and ***empirical research method*** will be used in this project. This project aims to solve a known & practical problem of precancerous lesion detection using DL in surveillance colonoscopies of IBD patients and builds on existing research, which is what applied research methods are suited for. However, the project also aims to test predictions by focusing on real people who will be contributing to the test dataset of colonoscopy videos. The observable evidence that is observed after the testing of the DL-based system on the image dataset, is later analysed in order to provide an explanation for it. Hence, the empirical research method [49] will also be involved in this project.

4.3 Research Methodology

The **research methodology** that this project uses is ***case study*** [54]. This project will be based on a mix of quantitative as well as qualitative evidence about the performance of the AI/DL models that are used. Moreover, it will be a study on a specific group of people (people suffering from chronic IBD) whose colonoscopy videos form a part of our testing dataset. The results observed on this limited number of people (and their associated dataset) should generalise over the rest of the chronic IBD patient population too. Hence, this project seems to be best suited to be a case study in terms of its research methodology [54, 49].

4.4 Data Collection Method

On a similar note, the **data collection method** used for this project is also a ***case study***, since the number of people whose colonoscopy videos form a part of our test dataset is fairly limited [49].

4.5 Data Analysis Method

Finally, the **data analysis method** used in this project was ***statistics*** since results were calculated over a sample of anonymous chronic IBD patients and later an evaluation was performed on the significance of these results[49].

5 Requirements and System Design

This project had a few requirements that needed to be followed throughout the course of the research.

- Dataset: the only datasets to be used for all training and testing purposes throughout this project were provided by the research group, and no external datasets were supposed to be used.

- DL Model: it was a requirement to use a YOLO-based DL model for the object detection task.

- Computer: the computer for all research purposes was a private one provided by the university (UiT) and no other computational resources could be used.

- GDPR [152]: due to the European privacy and data protection related regulations, only the colonoscopy videos consensually provided by patients (and later anonymized) could be used to generate the dataset and later be used in this project.

5.1 System Requirements

There were significant hardware, software and data needs associated with this project, that were necessary to be available in order to successfully undertake this project.

5.1.1 Hardware Requirements

Training and running Deep Learning models is a computational-resources-heavy process, especially when we take into consideration the large amount of data in the datasets that are being dealt with and processed. In order to make this process more efficient and less time-consuming, there is a prominent need to use a high-performance Graphics Processing Unit (GPU) along with an efficient computer. For the sake of this project, a computer with an 11th Gen Intel(R) Core(TM) i9 processor that comes with 64 GB of RAM and an NVIDIA GeForce RTX 3090 Graphics Card with 55.9 GB memory is being used.

5.1.2 Software Requirements

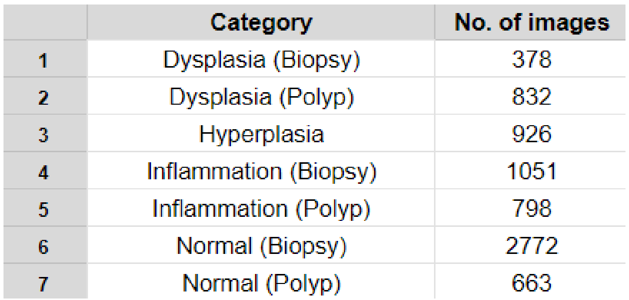
Python was used as the primary programming language because of its ease of use as well as how well-integrated it is with libraries that are used in Deep Learning like Tensorflow [105] and Keras [47]. For the object detection task, the pre-designed YOLOv8 model [69] and YOLOv5 model [44] were used, which come with an OSI-approved open-source license (AGPL-3.0 License) for students and enthusiasts [58, 44]. Meanwhile, for the image clustering task, the pre-designed and pre-trained VGG16 and VGG19 DL models were used, which were made available through *Keras Applications*[138].

5.1.3 Dataset

There are two primary datasets that are used for both training and testing purposes in this project. One of these is a labelled dataset with proper bounding-box annotations from 2021, while the other is a new, novel dataset without bounding-box annotations, that is being used for the first time in this project.

Dataset 1

Dataset 1 is a labelled, annotated dataset of 7420 images, each of resolution (1024 * 576) pixels. Table 5.1 shows the distribution of the images in the dataset across 7 categories that have been selected for the images in the dataset and their findings.



	Category	No. of images
1	Dysplasia (Biopsy)	378
2	Dysplasia (Polyp)	832
3	Hyperplasia	926
4	Inflammation (Biopsy)	1051
5	Inflammation (Polyp)	798
6	Normal (Biopsy)	2772
7	Normal (Polyp)	663

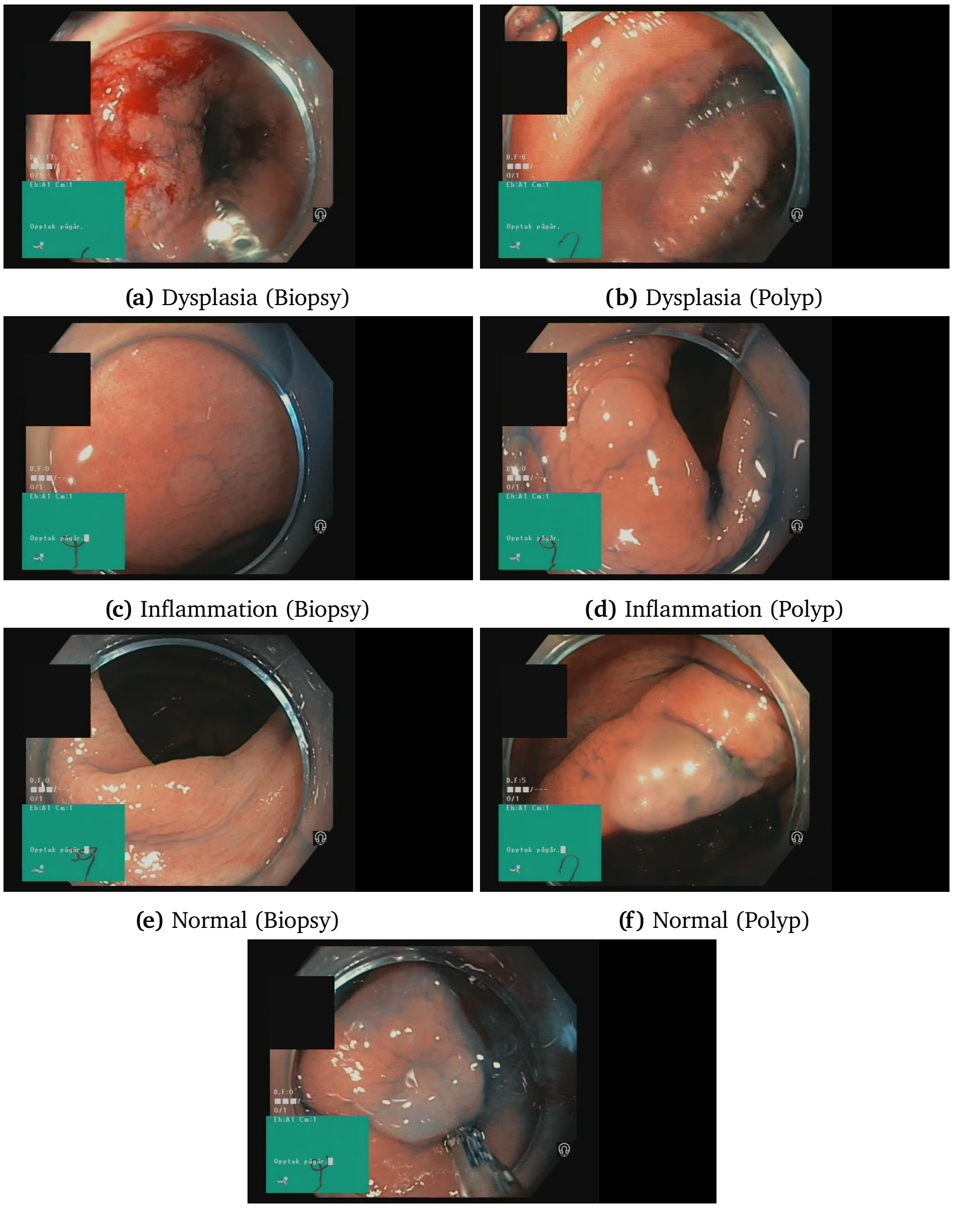
Table 5.1: *Dataset 1*

Olle Mannheimer (2021) in his degree project [93] mentions in further detail the aspect of creating this dataset, which he had the major responsibility for doing. The videos collected in order to create this dataset were collected between October 2019 and January 2021 at Akershus University Hospital, Oslo, Norway from 10 different patients diagnosed with Ulcerative colitis who each underwent colonoscopies as part of their regular surveillance program [93]. The colonoscopies were performed using the Olympus Endoscopy system and later annotated using a web-based video annotation interface provided by Augere Medical AS, Norway with bounding box annotations around each finding [93]. The bounding box annotations were labelled by masters students and then further validated by 2 senior experts in gastroenterology with more than 15 years of experience as specialists and more than 3000 colonoscopies performed [93]. The term *Biopsy* was used to label wherever a biopsy was taken with biopsy forceps, while the term *Polyp* was used for the areas being resected with a resection snare [93].

The dataset was exported to be used in this project in the YOLO Darknet TXT format [160], where each one of the 7420 images in this dataset had an associated text file which contains the bounding box annotations of the image along with a numeric representation of the label for that bounding box, along with a labelmap which mapped the numerical label ID to the string labels associated with them [160].

Figure 5.1 shows a sample image from each category of image in *Dataset 1.*

Dataset 1 was later used to generate 6 different datasets (*Dataset 1-R*, *Dataset 1-G*, *Dataset 1-B*, *Dataset 1-RG*, *Dataset 1-GB*, *Dataset 1-RB*) based on the colour channel separation and recombination of the RGB images in the dataset, the mechanism of which is described later in this chapter.



(a) Dysplasia (Biopsy)

(b) Dysplasia (Polyp)

(c) Inflammation (Biopsy)

(d) Inflammation (Polyp)

(e) Normal (Biopsy)

(f) Normal (Polyp)

(g) Hyperplasia

Figure 5.1: Images from *Dataset 1*

Dataset 2

Dataset 2 on the other hand is a dataset of 22 colonoscopy videos collected at a hospital in Stockholm that do not come with bounding-box annotations. The videos for this dataset were collected over the course of 2023 on IBD patients undergoing regular surveillance colonoscopies. The videos were anonymized before being shared, in accordance with the ethics associated to the collection and sharing of datasets in this project.

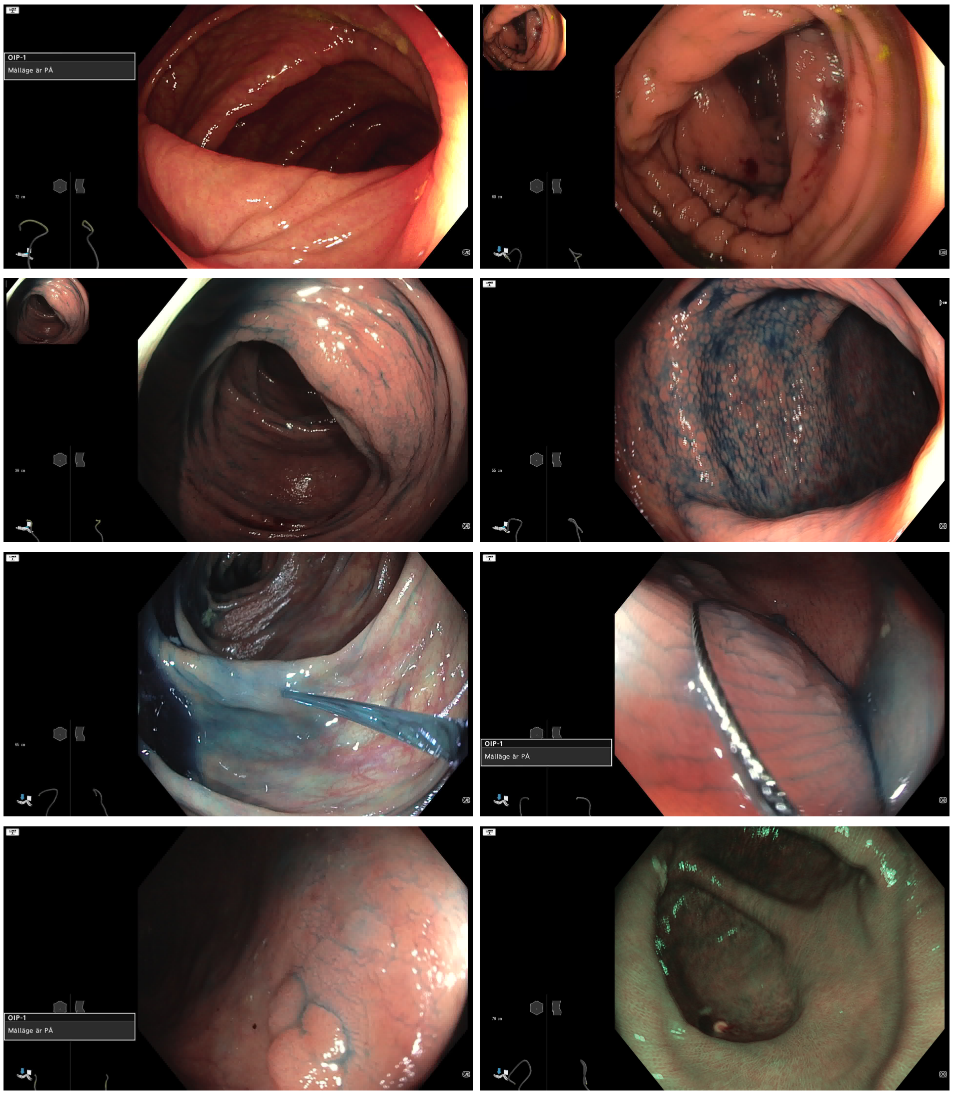


Figure 5.2: Sample frames of videos from *Dataset 2*

Each of these videos was collected using the Olympus Endoscopy System, with the videos being of resolution 1920*1080 pixels, or 3840*2160 pixels; and lasting in duration between 14 minutes to 64 minutes. The frame rate of these videos was 50 fps.

Some of the videos came without any form of labels at all except an anonymized unique video name, while some of the videos had time-stamp-based labels of the histopathological findings observed in the videos. This helped generate labels for those videos, although there were no bounding-box-based annotations, so this dataset could not be used for the sake of training or testing the DL model used for the *object detection* task.

The videos in this dataset capture frames from the colonoscopy feed consisting of things ranging from precancerous findings like dysplasia, hyperplastic mucosa to inflammatory activity, adenocarcinoma, stool, bubbles, and even instruments that are used to conduct biopsies during the colonoscopy. The pathological findings associated with certain videos in this dataset are presented in the next chapter.

Dataset 2 was later divided into 3 constituent datasets - *Dataset 2a*, *Dataset 2b* and *Dataset 2c* for further experiments and tests, for the *DL-based image clustering* task. This division of *Dataset 2* and further experiments with it are described in detail in the next chapter.

5.2 System Design & Architecture

The DL system designed as part of this project consists of primarily 3 subsystems within it, performing different tasks -

5.2.1 Supervised learning based object detector

This system solely used *Dataset 1* for its training step, and its purpose was to detect using bounding boxes and classify anomalies that are observed in a colonoscopy video or its associated image frames.

Dataset 1 was split up into 3 different subsets - 80% of images from each category formed a *training dataset*, while 10% of the images from each category formed a *validation dataset* and a *test dataset* respectively. Ideally, the test dataset would have come from a different source, but since the project had the constraint of just using *Dataset 1* for all the training and testing associated with the *Object Detection Task* and also because *Dataset 2* did not have bounding box annotations for the video frames in it, the *Test dataset* needed to be split from the same source as the *Training dataset* and the *Validation dataset*.

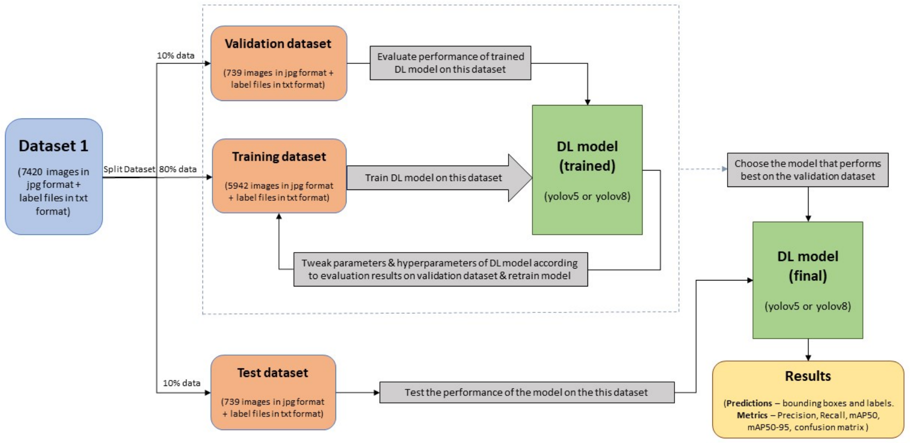

10% data
Validation dataset
(739 images in jpg format + label files in txt format)
Evaluate performance of trained DL model on this dataset
Dataset 1
(7420 images in jpg format + label files in txt format)
Split Dataset
80% data
Training dataset
(5942 images in jpg format + label files in txt format)
Train DL model on this dataset
DL model (trained)
(yolov5 or yolov8)
Choose the model that performs best on the validation dataset
Tweak parameters & hyperparameters of DL model according to evaluation results on validation dataset & retrain model
DL model (final)
(yolov5 or yolov8)
10% data
Test dataset
(739 images in jpg format + label files in txt format)
Test the performance of the model on the this dataset
Results
(Predictions – bounding boxes and labels. Metrics – Precision, Recall, mAP50, mAP50-95, confusion matrix)

Figure 5.3: Pipeline of splitting dataset & using it to train, validate and test the model

For the DL models to be used in this system - the focus was only on YOLOv8 [69], and one of its predecessors - YOLOv5 [44].

The purpose of all these models is the same: to draw a rectangular bounding box around any object that they detect and also label the object in the bounding box based on the categories that the model has been trained on. In the case of this project, the categories are any of the 7 categories of images in *Dataset 1* - Dysplasia (Biopsy), Dysplasia (Polyp), Hyperplasia, Inflammation (Biopsy), Inflammation (Polyp), Normal (Biopsy) or Normal (Polyp).

While training these models, *online data augmentation* was used in order to increase the number of training image samples and to also provide a variance in the quality of pictures so that the model is able to generalise better to different datasets. *Online data augmentation* does not store the augmented images that are generated but instead makes the data augmentation a part of the pre-processing pipeline while training the model.

Figure 5.3 demonstrates the process of splitting the original *Dataset 1* into the *Training*, *Validation* and *Test* datasets, and how these datasets were used for training our models as well as tuning its parameters & hyperparameters according to the model's performance on the validation dataset, and finally for testing it on a separate test dataset.

5.2.2 Supervised learning based object detector with colour-channel separation and recombination

This system has fundamentally the same architecture and design as the last system, but instead of using *Dataset 1* like in the previous system, a variation of that dataset was used here.

Dataset 1 consists of 7420 images that are RGB in nature. So, instead of using all three channels of the images, the images were split up into their separate colour channels - R (Red), G (Green) and B (Blue). So, the R channel images form a separate dataset (*Dataset 1-R*), and similarly the G channel images form *Dataset 1-G* and the B channel images form *Dataset 1-B*.

Furthermore, as seen in subsection 2.5.4, the colour-channel recombination of images can also sometimes lead to improvements in the performance of DL models. So, each of these individual channels was also combined to form separate datasets. The R+G channel images form *Dataset 1-RG*, the G+B channel images form *Dataset 1-GB* while the R+B channel images combine to form *Dataset 1-RB*.

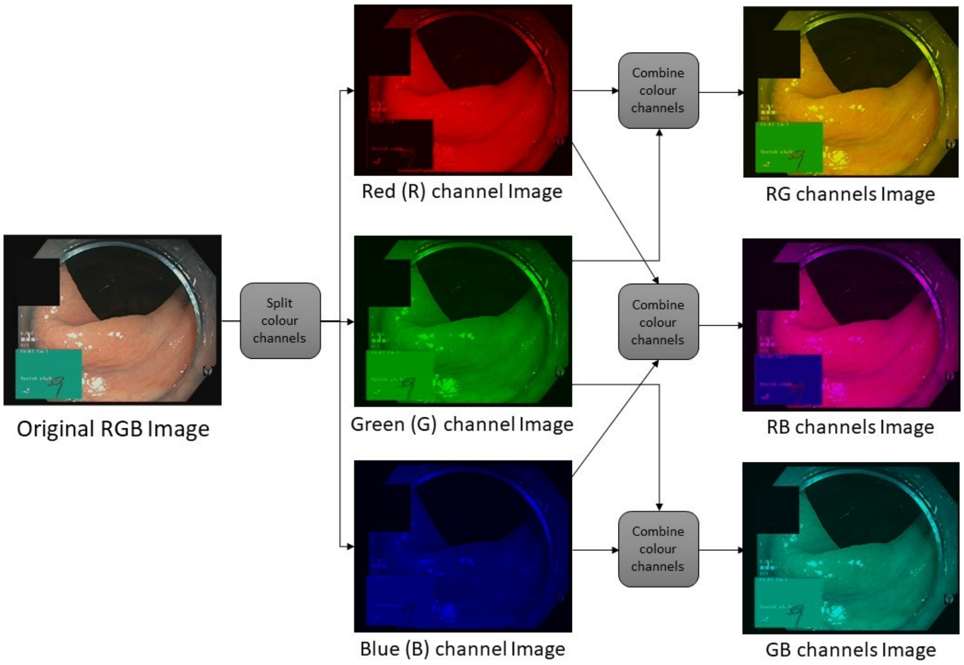


Figure 5.4: Pipeline of splitting RGB images to separate R, G, B channels & recombining it to form RG, RB and GB channel images

The entire pipeline for splitting the colour channels of the RGB images and then recombining them has been visualised in figure 5.4, with a sample image from *Dataset 1* being used to demonstrate the effects of colour channel splitting and recombination.

Each of these 6 new datasets is split up into 80% training dataset, 10% validation dataset and 10% test dataset like in the previous system. These datasets each had the same images in their *training*, *validation* and *test* datasets as *Dataset 1's training*, *validation* and *test* datasets. Just like the previous system, there is an online data-augmentation of the images involved in this system too.

Then, these new datasets are used for training and testing the object detection models mentioned in the previous system, to see if the colour-channel separation and/or recombination can give an improvement in results, compared to the standard RGB dataset.

5.2.3 Unsupervised learning based image-clustering

This system was fundamentally built to deal with the unlabelled data available in *Dataset 2*. Despite being unlabelled and hence not having any bounding-box annotations for the images in it, *Dataset 2* has images of a much higher resolution than *Dataset 1*, and also has a significantly larger number of images in it. Hence, an unsupervised learning based model was designed to cluster the images into different clusters.

For this, a pre-trained VGG-16[127] or VGG-19[127] model (pre-trained on the ImageNet [34] dataset) was used for feature extraction, by removing the final layer of the model in order to get a feature vector produced when the model runs on an image from the dataset on which clustering was being performed [41].

A variation of the clustering pipeline described by Gabe Flomo (2020) [41] was used in order to get a feature vector of length 4096 for each image that was passed through the model. The original image was resized to the dimensions 224*224 pixels before being processed through the model since the VGG-16 and VGG-19 models run with those dimension images as input. This was done for all images, so that a feature vector corresponding to every image was obtained.

However, since the number of components (or dimensions of each feature vector) was high which would make the process of clustering computationally very expensive, there was a need to reduce the number of components in the feature vector [41]. For this, the method of principal component analysis (PCA) [2] was harnessed in order to reduce the dimensions of the feature vector for all images from 4096 down to 1000.

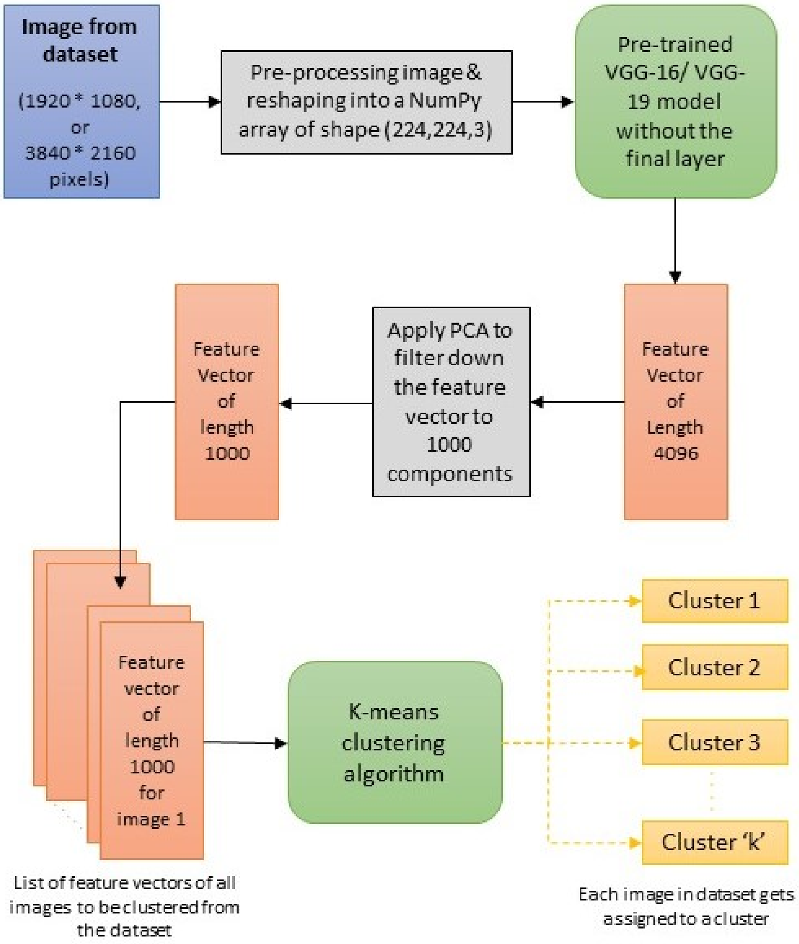


Image from dataset
(1920 * 1080, or 3840 * 2160 pixels)
Pre-processing image & reshaping into a NumPy array of shape (224,224,3)
Pre-trained VGG-16/ VGG-19 model without the final layer
Feature Vector of Length 4096
Apply PCA to filter down the feature vector to 1000 components
Feature Vector of length 1000
Feature vector of length 1000 for image 1
List of feature vectors of all images to be clustered from the dataset
K-means clustering algorithm
Cluster 1
Cluster 2
Cluster 3
Cluster 'k'
Each image in dataset gets assigned to a cluster

Figure 5.5: Pipeline of the unsupervised clustering algorithm based on VGG-16/ VGG-19

Finally, the K-Means clustering algorithm was used to group the feature vectors into 'k' numbers of groups called 'clusters'.

Figure 5.5 demonstrates the pipeline used by the unsupervised clustering algorithm in this project on the datasets that it was run on.

6

Implementation

In order to implement the system design discussed in Chapter 5, a few steps needed to be systematically followed, which are mentioned in the following sections.

6.1 Data preparation

For all the systems designed in Chapter 5, the datasets needed to be first prepared to work with that system.

6.1.1 For object detection task

Dataset 1 was going to be used for the training as well as evaluation of the object detector, and it was necessary for the dataset to be in the YOLO Darknet TXT format [160] for the YOLOv5 and YOLOv8 models to work with it. Since the dataset had already been provided in the correct format, there was only a need to verify whether each image had a proper label file associated with it or not.

Next, *Dataset 1* was split up into 3 different subsets as described in Figure 5.3, with 80% of images from each category formed a *training dataset*, while 10% of the images from each category formed a *validation dataset* and a *test dataset* respectively. It was again verified that each image had its associated text file with the proper annotations.

This was followed by implementing the colour-channel separation of the images in *Dataset 1*, and then the colour-channel recombination to form 6 new datasets - *Dataset 1-R*, *Dataset 1-G*, *Dataset 1-B*, *Dataset 1-RG*, *Dataset 1-GB*, *Dataset 1-RB*, as described in Subsection 5.2.2 and seen in Figure 5.4. These datasets each had the same images in their *training*, *validation* and *test* datasets as *Dataset 1's training*, *validation* and *test* datasets.

With this, the datasets for training and testing the *object detection models* were ready.

6.1.2 For DL based image clustering task

Dataset 2 needed a bit more pre-processing before it was ready for the experiments since originally it only consisted of videos that were shared in the MP4 video format, after being anonymized to remove any personal information of the patients.

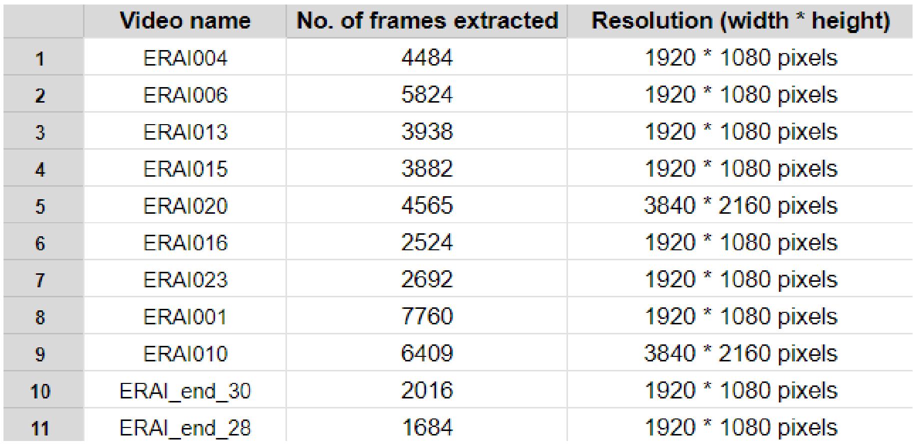
In order to make this dataset easier to work with and utilise for training and testing purposes later on over the course of this project, there was a need to convert the videos into an image dataset. This was done using the FFMPEG tool [141], where it was also possible to mention the frame rate to extract frames at, in case there was a need to not extract all the original frames of the 50 fps videos to reduce the number of similar-looking images in the dataset.

There were 3 different daughter-datasets that were created from this parent-dataset *Dataset 2*, based on the number of frames extracted and whether those videos had any time-stamp-based labels or not.

a) Dataset 2a -

This dataset was created towards the very beginning of the project, when the first colonoscopy videos were collected and shared by the medical professionals conducting the colonoscopies. At that point in time, none of the videos had any histopathological findings associated with them. So, this dataset was a completely unlabelled dataset.

The frames were extracted from the videos at the rate of 2 frames per second, which served the purpose of ensuring that the dataset size does not explode and become too big and memory-intensive to practically work with, and also that the dataset does not have too many similar-looking images. The images had the same resolution as the resolution of the videos they were extracted from - which was either 1920*1080 pixels, or 3840*2160 pixels. The extraction of frames was done for the entire duration of the original videos.



	Video name	No. of frames extracted	Resolution (width * height)
1	ERAI004	4484	1920 * 1080 pixels
2	ERAI006	5824	1920 * 1080 pixels
3	ERAI013	3938	1920 * 1080 pixels
4	ERAI015	3882	1920 * 1080 pixels
5	ERAI020	4565	3840 * 2160 pixels
6	ERAI016	2524	1920 * 1080 pixels
7	ERAI023	2692	1920 * 1080 pixels
8	ERAI001	7760	1920 * 1080 pixels
9	ERAI010	6409	3840 * 2160 pixels
10	ERAI_end_30	2016	1920 * 1080 pixels
11	ERAI_end_28	1684	1920 * 1080 pixels

Table 6.1: *Dataset 2a*

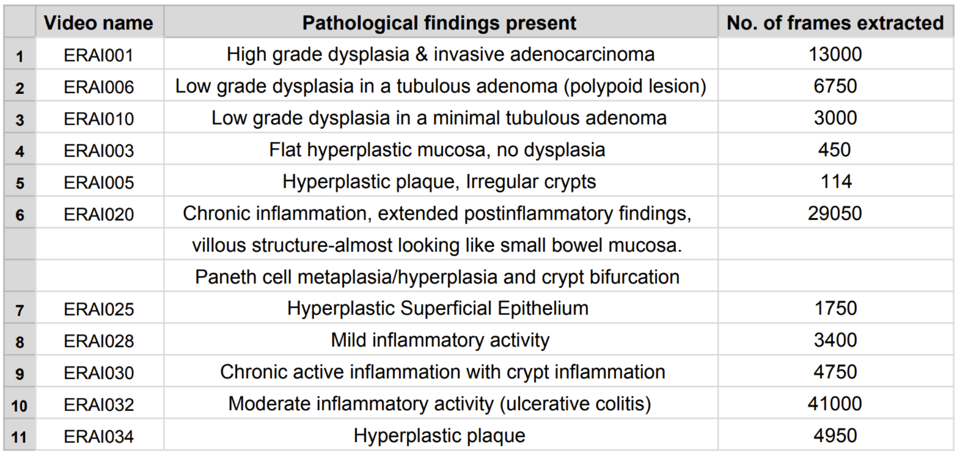
Table 6.1 gives a description of *Dataset 2a*, and the images in it. Each video's frames were saved in a separate folder, which had the same name as that of the video. In the end, the dataset consisted of 11 folders (each folder unique to 1 colonoscopy video) with the number of images in each folder varying.

b) Dataset 2b -

Later over the course of the project, some histopathological findings were obtained for the biopsies conducted for certain videos, based on which time-stamp-based labels were provided for the histopathological findings found in certain time-segments of the colonoscopy videos.

Based on these time stamps, the videos were first trimmed down to only those segments of the original videos where there were findings observed. Then, frames of videos were extracted from these video segments at the rate of 50 fps (which was the frame rate of the original video), hence ensuring that all the frames with findings were extracted. This formed *Dataset 2b*.

Table 6.2 gives a description of *Dataset 2b* and the findings found in it. Each video had its own folder, where the image frames containing its findings were extracted to. However, there was an extremely large variation between the number of frames extracted for each folder/video.



	Video name	Pathological findings present	No. of frames extracted
1	ERAI001	High grade dysplasia & invasive adenocarcinoma	13000
2	ERAI006	Low grade dysplasia in a tubulous adenoma (polypoid lesion)	6750
3	ERAI010	Low grade dysplasia in a minimal tubulous adenoma	3000
4	ERAI003	Flat hyperplastic mucosa, no dysplasia	450
5	ERAI005	Hyperplastic plaque, Irregular crypts	114
6	ERAI020	Chronic inflammation, extended postinflammatory findings,	29050
		villous structure-almost looking like small bowel mucosa.	
		Paneth cell metaplasia/hyperplasia and crypt bifurcation	
7	ERAI025	Hyperplastic Superficial Epithelium	1750
8	ERAI028	Mild inflammatory activity	3400
9	ERAI030	Chronic active inflammation with crypt inflammation	4750
10	ERAI032	Moderate inflammatory activity (ulcerative colitis)	41000
11	ERAI034	Hyperplastic plaque	4950

Table 6.2: *Dataset 2b*

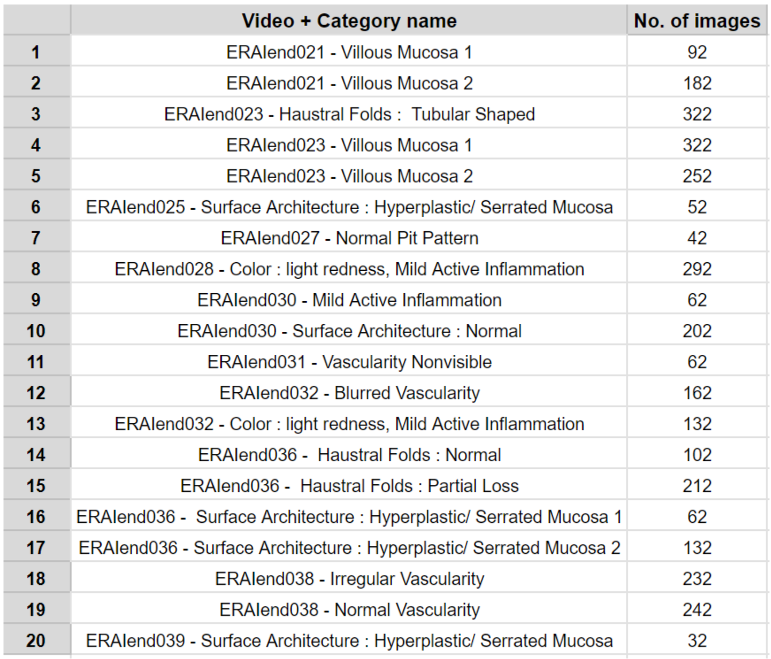
c) Dataset 2c -

Finally, towards the end of the project, the final list of time-stamp-based histopathological findings for some more videos was provided, based on which *Dataset 2c* was created.

Some of the videos from *Dataset 2a* and *Dataset 2b* were also used here, while there were some new videos. These videos were split into their individual frames at a rate of 10 frames per second, and then divided into different categories/folders mentioned in the table 6.3.

The *Video + Category name* in table 6.3 consists of 2 main components. The first part (which is the term starting with 'ERAIend') refers to the video name from which those frames were extracted. The remainder of the term, i.e. the *Category name*, is the actual histopathological finding mentioned by the medical professionals who provided the time-stamp-based labels. Each colonoscopy video can have different segments of the video where different findings are seen, which is why sometimes there are multiple categories for the same video. Also, the same finding can be found multiple times in the same colonoscopy video - like the 2 different time stamps where *Villous Mucosa* were observed in the video *ERAIend021*, which is why at least initially when extracting the video frames, they are each given a different *Video + Category name*.

However, these labels were only a textual description of the findings seen in different segments of the colonoscopy videos. These were not bounding-box-based labels. So, this dataset could not be used for the object-detection task or training. Instead, it was used to get an insight into the functioning of the image clustering model.



	Video + Category name	No. of images
1	ERAlend021 - Villous Mucosa 1	92
2	ERAlend021 - Villous Mucosa 2	182
3	ERAlend023 - Haustral Folds : Tubular Shaped	322
4	ERAlend023 - Villous Mucosa 1	322
5	ERAlend023 - Villous Mucosa 2	252
6	ERAlend025 - Surface Architecture : Hyperplastic/ Serrated Mucosa	52
7	ERAlend027 - Normal Pit Pattern	42
8	ERAlend028 - Color : light redness, Mild Active Inflammation	292
9	ERAlend030 - Mild Active Inflammation	62
10	ERAlend030 - Surface Architecture : Normal	202
11	ERAlend031 - Vascularity Nonvisible	62
12	ERAlend032 - Blurred Vascularity	162
13	ERAlend032 - Color : light redness, Mild Active Inflammation	132
14	ERAlend036 - Haustral Folds : Normal	102
15	ERAlend036 - Haustral Folds : Partial Loss	212
16	ERAlend036 - Surface Architecture : Hyperplastic/ Serrated Mucosa 1	62
17	ERAlend036 - Surface Architecture : Hyperplastic/ Serrated Mucosa 2	132
18	ERAlend038 - Irregular Vascularity	232
19	ERAlend038 - Normal Vascularity	242
20	ERAlend039 - Surface Architecture : Hyperplastic/ Serrated Mucosa	32

Table 6.3: *Dataset 2c*

6.2 Object detection model implementation

6.2.1 Training and evaluation on *Dataset 1*

After the dataset preparation was completed, the training and testing of the object detection models could take place, for which *Dataset 1* would be used first, in a manner similar to that described in Figure 5.3.

YOLOv5[44] and YOLOv8[58, 19] come in multiple model sizes - nano (YOLOv5n/ YOLOv8n), small (YOLOv5s/ YOLOv8s), medium (YOLOv5m/ YOLOv8m), large ((YOLOv5l/ YOLOv8l) and extra large (YOLOv5x/ YOLOv8x), with the model size and hence the number of parameters and training time increasing from nano to extra-large. Each of these models was trained with *Dataset 1's* training dataset, along with its validation dataset being used for validating the trained model after every epoch of training.

Since there were a large number of datasets to train and test the model on over the course of this project, and also because there were a high number of parameters and hyperparameters associated with the YOLOv8 and YOLOv5 models that could be altered and fine-tuned in order to improve the model's performance, it was necessary to set the same standard number of training epochs for each training experiment, which was 100 here. Another reason why the number of epochs for training was set as 100 was to minimise overfitting. One final reason for choosing this as the number of training epochs is the high training time associated with training the models - especially the models with a high number of parameters.

After training was complete, *Dataset 1's* test dataset was used to generate evaluation results in the form of a Confusion Matrix as well as Mean Average Precision (mAP) values.

6.2.2 Training and evaluation on colour-channel separated and recombined datasets

After *Dataset 1*, it was the turn of its daughter datasets *Dataset 1-R*, *Dataset 1-G*, *Dataset 1-B*, *Dataset 1-RG*, *Dataset 1-GB*, *Dataset 1-RB* that were obtained after the colour channel separation and recombination, to be used for training and evaluation of the object detection model.

This time, only the YOLOv8x model was used, with the training and evaluation taking place the same way as done on *Dataset 1*, as mentioned in section 6.2.1. The same metrics as that mentioned in section 6.2.1 were used for the evaluation of these models too. The number of epochs for training was also set as 100, for the same reasons as before.

6.3 DL-based image clustering model implementation

Finally, the DL-based image clustering model was implemented on *Dataset 2's* daughter datasets - *Dataset 2a*, *Dataset 2b* and *Dataset 2c*.

6.3.1 Image clustering on *Dataset 2a*

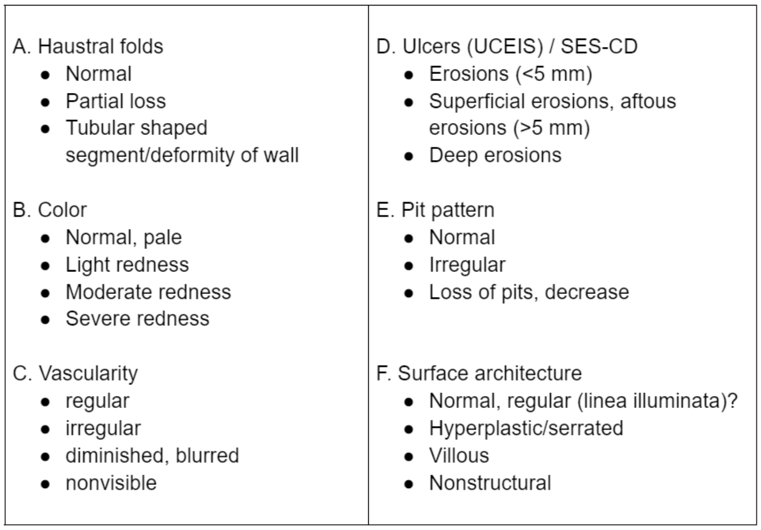
Dataset 2a came with no findings or labels associated with the videos in it, and no way of knowing how many clusters to segregate the images in the dataset into. After a bit of trial and error, it was decided to cluster the images in each folder of *Dataset 2a* into 48 different clusters each. Using a number of clusters that was too small had the potential to group together multiple different potential findings into the same clusters, while using a number of clusters that was too high had the potential of creating clusters that did not have similar images together.

The clustering pipeline seen in Figure 5.5 was implemented on each folder of *Dataset 2a* separately, with the VGG-16 model pre-trained on ImageNet [34] being used for the feature-extraction part of the pipeline to create the feature vector.

In order to assess the quality and performance of this clustering pipeline, the resulting clusters for each individual video were shared with the group of medical professionals in the field of colonoscopy and IBD in the *HMT Group* associated with the project, for their feedback and input, which was documented and will be discussed in the next chapter.

6.3.2 Image clustering on *Dataset 2b*

Since *Dataset 2b* came with time-stamp-based histopathological findings and the image frames were extracted only for the video segments where there were findings present, it helped ensure that this dataset had much fewer percentage of images with no finding present, compared to *Dataset 2a*.



A. Haustral folds
- Normal
- Partial loss
- Tubular shaped segment/deformity of wall

B. Color
- Normal, pale
- Light redness
- Moderate redness
- Severe redness

C. Vascularity
- regular
- irregular
- diminished, blurred
- nonvisible

D. Ulcers (UCEIS) / SES-CD
- Erosions (<5 mm)
- Superficial erosions, aftous erosions (>5 mm)
- Deep erosions

E. Pit pattern
- Normal
- Irregular
- Loss of pits, decrease

F. Surface architecture
- Normal, regular (linea illuminata)?
- Hyperplastic/serrated
- Villous
- Nonstructural

Table 6.4: Template to label the findings (Provided by Camilla Wijkström, Ersta Endoskopienhet, Sweden)

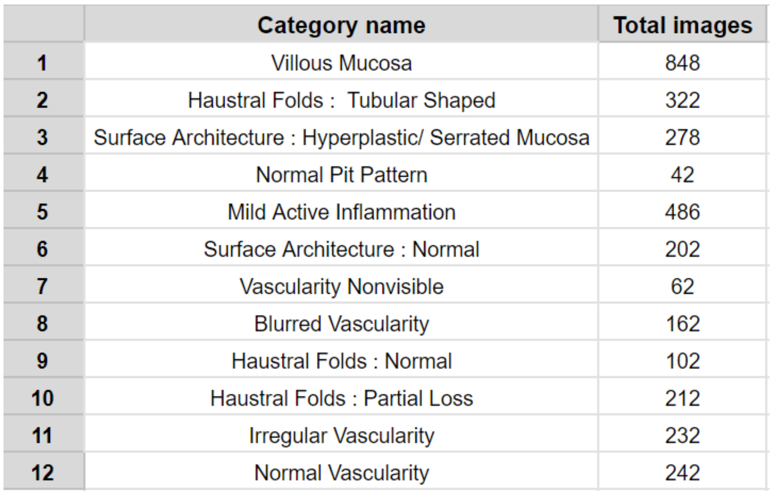
The template seen in Table 6.4 was used as a guiding tool to label the findings obtained in the videos, although some videos had findings or labels not seen in this table. However, most labels were based on this table. Assuming that 2 different types of findings did not appear on the video simultaneously, there could be 21 different categories that the potential findings could have. So, it was decided to cluster *Dataset 2b* into 23 different clusters (21 clusters for the 21 categories of potential findings, and 2 extra clusters - 1 for the frames where there is white light, and 1 for the frames where this is motion blur). The VGG-16 model pre-trained on ImageNet was again used as the feature extractor in order to generate the feature vector for the images in the dataset.

The clustering pipeline of Figure 5.5 was again run here for each individual folder in *Dataset 2b*. After the 23 clusters for each video's folder were generated, the resulting clusters were again shared with the team of medical professionals for their assessment and comments, which were documented and mentioned in the next chapter.

6.3.3 Image clustering on *Dataset 2c*

Dataset 2c is very similar to *Dataset 2b* in terms of the time-stamp-based histopathological findings based on which the dataset is generated. However, a different approach to the clustering and evaluation of results was done this time.

Dataset 2c was obtained from 11 individual videos, with 12 different categories of findings in it, as seen in Table 6.3 and Table 6.5.



	Category name	Total images
1	Villous Mucosa	848
2	Haustral Folds : Tubular Shaped	322
3	Surface Architecture : Hyperplastic/ Serrated Mucosa	278
4	Normal Pit Pattern	42
5	Mild Active Inflammation	486
6	Surface Architecture : Normal	202
7	Vascularity Nonvisible	62
8	Blurred Vascularity	162
9	Haustral Folds : Normal	102
10	Haustral Folds : Partial Loss	212
11	Irregular Vascularity	232
12	Normal Vascularity	242

Table 6.5: Unique categories in *Dataset 2c* used to set the number of clusters

In order to settle on these 12 different numbers of categories, the images with 'Villous Mucosa' present in videos 'ERAIend021' and 'ERAIend023' were considered to be the same category/potential cluster. Similarly, the images with 'Surface Architecture : Hyperplastic/ Serrated Mucosa' in videos 'ERAIend025', 'ERAIend036' and 'ERAIend039' were considered to be the same category/potential cluster. The same was done for the images with 'Mild Active Inflammation' present in videos 'ERAIend028', 'ERAIend030' and 'ERAIend32'.

For this implementation of the clustering pipeline of Figure 5.5, the VGG-19 model pre-trained on ImageNet was used. Moreover, all the images from each individual category of *Dataset 2c* were combined together into a common folder, and then the clustering pipeline was run on this folder of all images. The clustering pipeline was run to generate 12 clusters - corresponding to the 12 different categories of findings, as mentioned in table 6.5.

Moreover, since this time there were proper time-stamp-based labels for each frame in the dataset, there was no need to rely on the manual assessment and feedback done by the team of medical professionals, and instead, some numerical metrics were generated based on the clusters generated, which is discussed in the next chapter.

/7

Results and Discussion

7.1 Results

7.1.1 For Object Detection with *Dataset 1*

Dataset 1 was first used to train, validate and test various YOLOv5 and YOLOv8 models. Figure 7.1 shows a plot of the mAP50-95 values on the test dataset given by the various-sized YOLOv8 and YOLOv5 models, against the time taken to train these models.

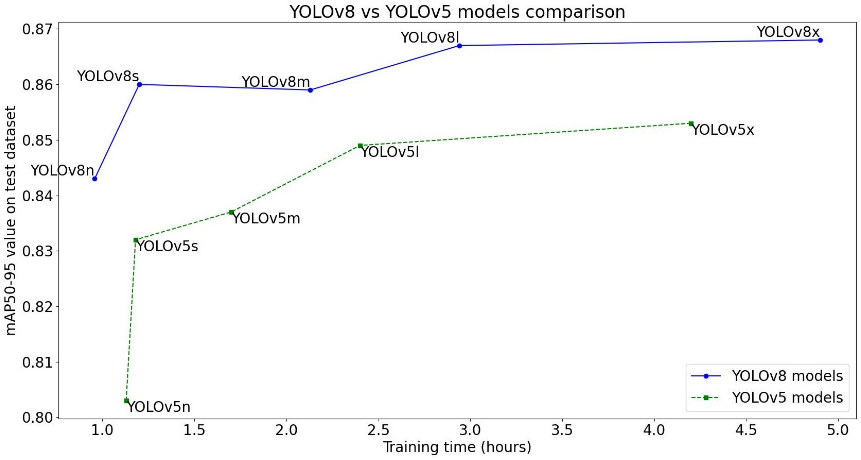
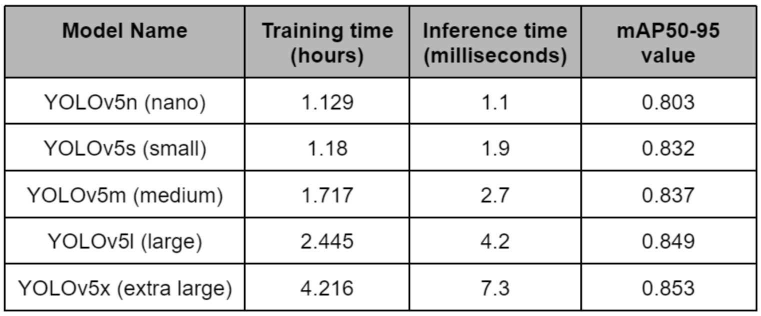


Figure 7.1: YOLOv8 vs YOLOv5 models' comparison on *Dataset 1*

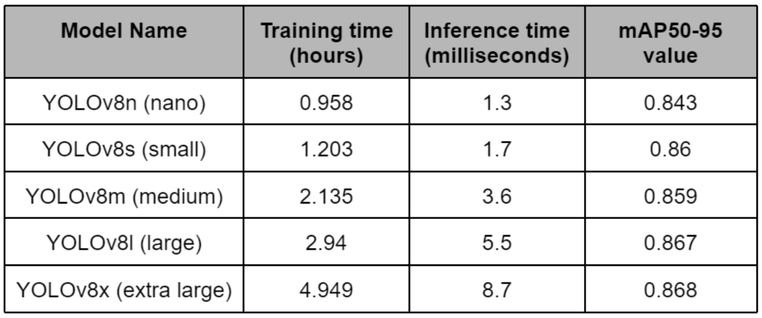
Tables 7.1 and 7.2 meanwhile also show the performance metrics of the YOLOv5 and YOLOv8 models respectively on the test dataset, while also mentioning their training time for **100 epochs**. The mAP50-95 value on these tables and plot are calculated for the Bounding Box that was predicted by the trained model, while the inference time is the time taken by the trained model to do inference for each image in the test dataset.

For testing the dataset and creating these metrics, the images in the test datasets were resized to 640*640 pixels, and the default parameters of the YOLOv8/YOLOv5 models were used. These default parameters included the parameter *object confidence threshold* for detection, which will henceforth be referred to as the *confidence threshold*. The confidence threshold's default value for this figure and these tables was 0.001.



Model Name	Training time (hours)	Inference time (milliseconds)	mAP50-95 value
YOLOv5n (nano)	1.129	1.1	0.803
YOLOv5s (small)	1.18	1.9	0.832
YOLOv5m (medium)	1.717	2.7	0.837
YOLOv5l (large)	2.445	4.2	0.849
YOLOv5x (extra large)	4.216	7.3	0.853

Table 7.1: YOLOv5 models' performance metrics with *Dataset 1*



Model Name	Training time (hours)	Inference time (milliseconds)	mAP50-95 value
YOLOv8n (nano)	0.958	1.3	0.843
YOLOv8s (small)	1.203	1.7	0.86
YOLOv8m (medium)	2.135	3.6	0.859
YOLOv8l (large)	2.94	5.5	0.867
YOLOv8x (extra large)	4.949	8.7	0.868

Table 7.2: YOLOv8 models' performance metrics with *Dataset 1*

Figure 7.2 shows the plot-based results of training and validating a YOLOv8x model with *Dataset 1*, more commonly referred to by the term *learning curves*. The training losses of the model after every epoch of training are represented by *train/box_loss*, *train/cls_loss* and *train/dfl_loss*; while the validation losses of the model on the validation split of *Dataset 1* are represented by *val/box_loss*, *val/cls_loss* and *val/dfl_loss*. The evaluation metrics of the model on the validation dataset after every epoch are also included in this figure.

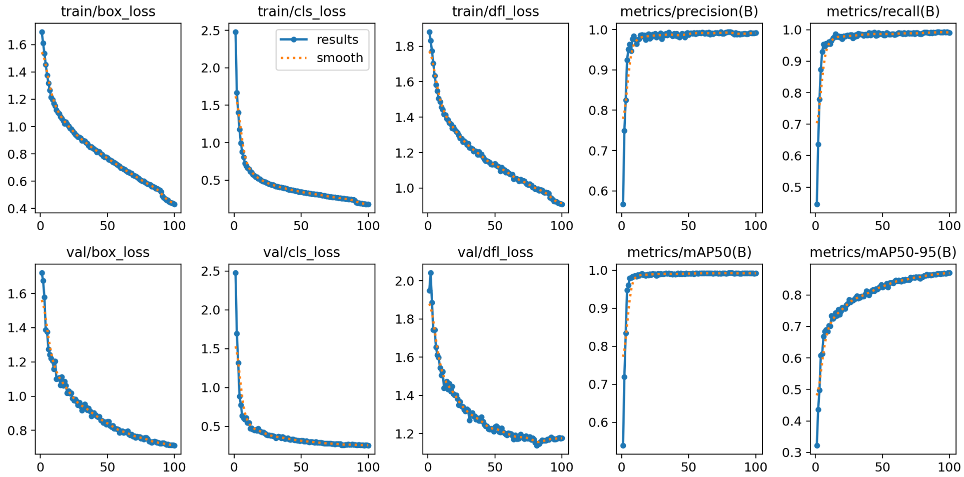


Figure 7.2: YOLOv8x model's learning curve, with *Dataset 1*

Figure 7.3 shows the mAP50-95 values for different-sized YOLOv8 models when trained with *Dataset 1*'s training split, and then tested on its testing split, while setting the confidence threshold for detections as 0.9. Setting this parameter so high means that the model only gives those objects as detections which have a confidence score of a value equal to or over 0.9, hence ensuring very accurate predictions for the object's class, and minimising mis-detection or wrong detections of objects. The x-axis marks the training time of these models for 100 epochs of training.

The corresponding performance metrics are mentioned in table 7.3. Figure 7.4 meanwhile shows the confusion matrices for the YOLOv8x and YOLOv8l models tested on the test split of *Dataset 1*, with the confidence threshold set as 0.9.

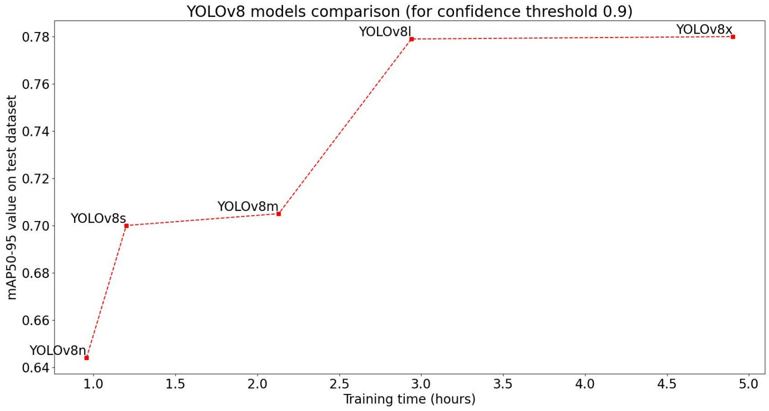
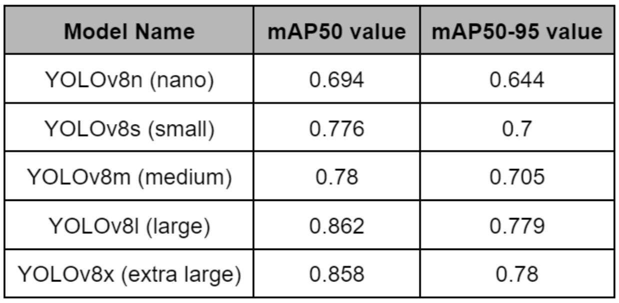
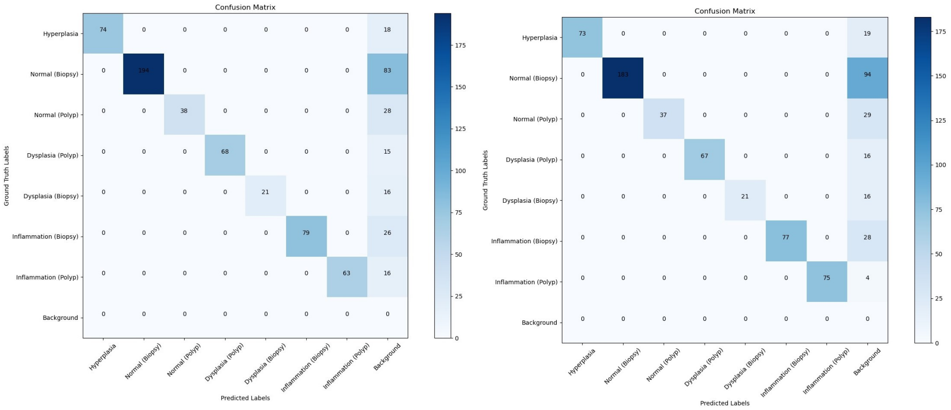


Figure 7.3: YOLOv8 models' comparison while training and testing on *Dataset 1*, with confidence threshold set as 0.9 for predictions



Model Name	mAP50 value	mAP50-95 value
YOLOv8n (nano)	0.694	0.644
YOLOv8s (small)	0.776	0.7
YOLOv8m (medium)	0.78	0.705
YOLOv8l (large)	0.862	0.779
YOLOv8x (extra large)	0.858	0.78

Table 7.3: YOLOv8 models' performance metrics on test split of *Dataset 1*, with confidence threshold set as 0.9



	Hyperplasia	Normal (Biopsy)	Normal (Polyp)	Dysplasia (Polyp)	Dysplasia (Biopsy)	Inflammation (Biopsy)	Inflammation (Polyp)	Background
Hyperplasia	74	0	0	0	0	0	0	18
Normal (Biopsy)	0	194	0	0	0	0	0	83
Normal (Polyp)	0	0	38	0	0	0	0	28
Dysplasia (Polyp)	0	0	0	68	0	0	0	15
Dysplasia (Biopsy)	0	0	0	0	21	0	0	16
Inflammation (Biopsy)	0	0	0	0	0	79	0	26
Inflammation (Polyp)	0	0	0	0	0	0	63	16
Background	0	0	0	0	0	0	0	0

(a) Confusion matrix for YOLOv8x

	Hyperplasia	Normal (Biopsy)	Normal (Polyp)	Dysplasia (Polyp)	Dysplasia (Biopsy)	Inflammation (Biopsy)	Inflammation (Polyp)	Background
Hyperplasia	73	0	0	0	0	0	0	19
Normal (Biopsy)	0	183	0	0	0	0	0	94
Normal (Polyp)	0	0	37	0	0	0	0	29
Dysplasia (Polyp)	0	0	0	67	0	0	0	16
Dysplasia (Biopsy)	0	0	0	0	21	0	0	16
Inflammation (Biopsy)	0	0	0	0	0	77	0	28
Inflammation (Polyp)	0	0	0	0	0	0	75	4
Background	0	0	0	0	0	0	0	0

(b) Confusion matrix for YOLOv8l

Figure 7.4: Comparison of YOLOv8x and YOLOv8l models' confusion matrices on test split of *Dataset 1*, with confidence threshold set as 0.9 for predictions

7.1.2 For Object Detection with colour-channel separation and recombination

Figure 7.5 and Table 7.4 show the performance of the YOLOv8x models that were trained, validated and tested on the different datasets created after colour channel separation and recombination - *Dataset 1-R*, *Dataset 1-G*, *Dataset 1-B*, *Dataset 1-RG*, *Dataset 1-GB*, *Dataset 1-RB*. Each of these datasets had the same sets of images in their training, validation and test splits, as stated earlier. Also, the confidence threshold for detection was set at 0.9 here for all these results.

The improvement in the object detection task by colour-channel separation and recombination was also easy to notice when observing the confusion matrices given by the YOLOv8x models trained and tested on these datasets, with confidence threshold set as 0.9 for detections. These confusion matrices are

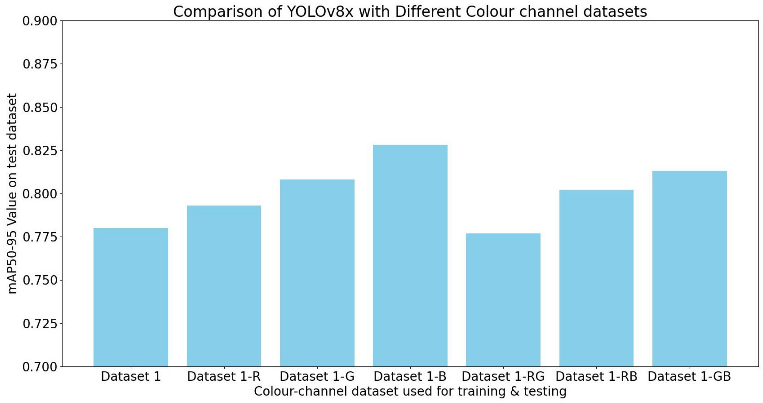
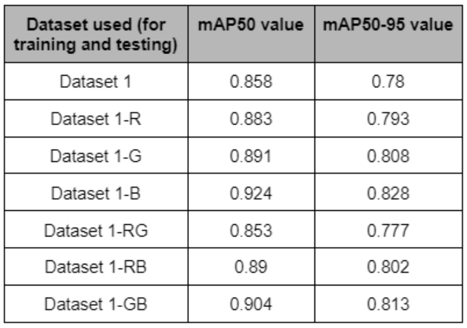


Figure 7.5: Comparison of YOLOv8x models trained and tested on different colour-channel datasets (with the confidence threshold for testing set at 0.9)

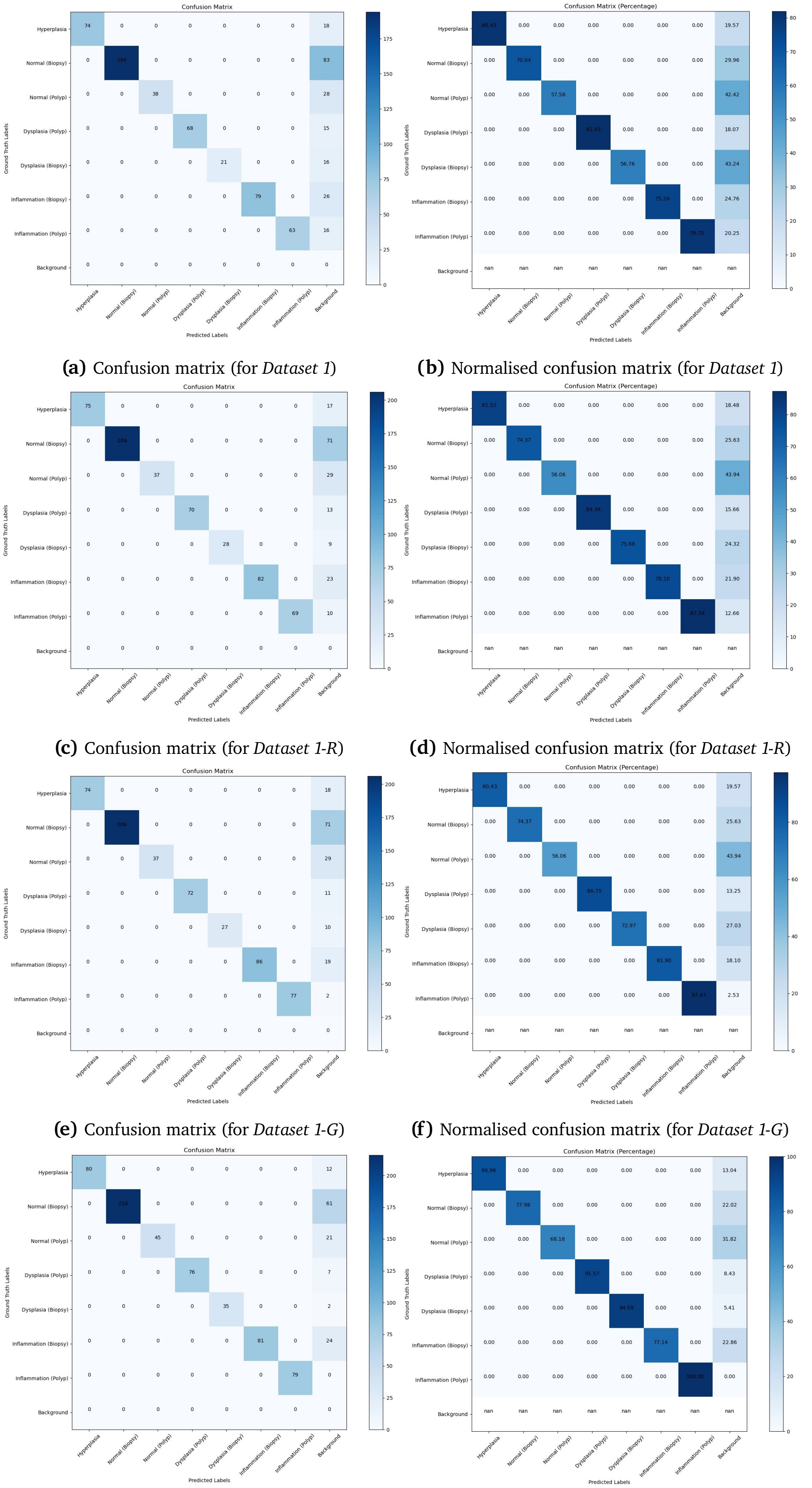


Dataset used (for training and testing)	mAP50 value	mAP50-95 value
Dataset 1	0.858	0.78
Dataset 1-R	0.883	0.793
Dataset 1-G	0.891	0.808
Dataset 1-B	0.924	0.828
Dataset 1-RG	0.853	0.777
Dataset 1-RB	0.89	0.802
Dataset 1-GB	0.904	0.813

Table 7.4: YOLOv8x models' performance metrics on different datasets (with confidence threshold set as 0.9 for testing)

included in figures 7.6 and 7.7. The confusion matrices and normalised confusion matrices show true positives (i.e. when the predicted label for an object and its original/ground truth label is the same) along their main diagonal. The confusion matrices also have each count numerically mentioned, as well as a representation in the form of a heat map.

Finally, figure 7.8 shows the comparison of the predictions generated by the YOLOv8x model for different datasets. None of the images had the image name in the white text on the top-left corner, originally. Figure 7.8a shows the ground truth images with the respective target bounding box and target class labels. Figure 7.8b shows the bounding box predictions along with the class prediction for the object in the bounding box and the confidence score for it, generated by a YOLOv8x model that was trained on *Dataset 1* and then tested on images from *Dataset 1*, with the confidence threshold set as 0.9. Figure 7.8c meanwhile shows the bounding box predictions along with the class prediction and the confidence score for it, generated by a YOLOv8x model that was trained on



(a) Confusion matrix (for *Dataset 1*)

(b) Normalised confusion matrix (for *Dataset 1*)

(c) Confusion matrix (for *Dataset 1-R*)

(d) Normalised confusion matrix (for *Dataset 1-R*)

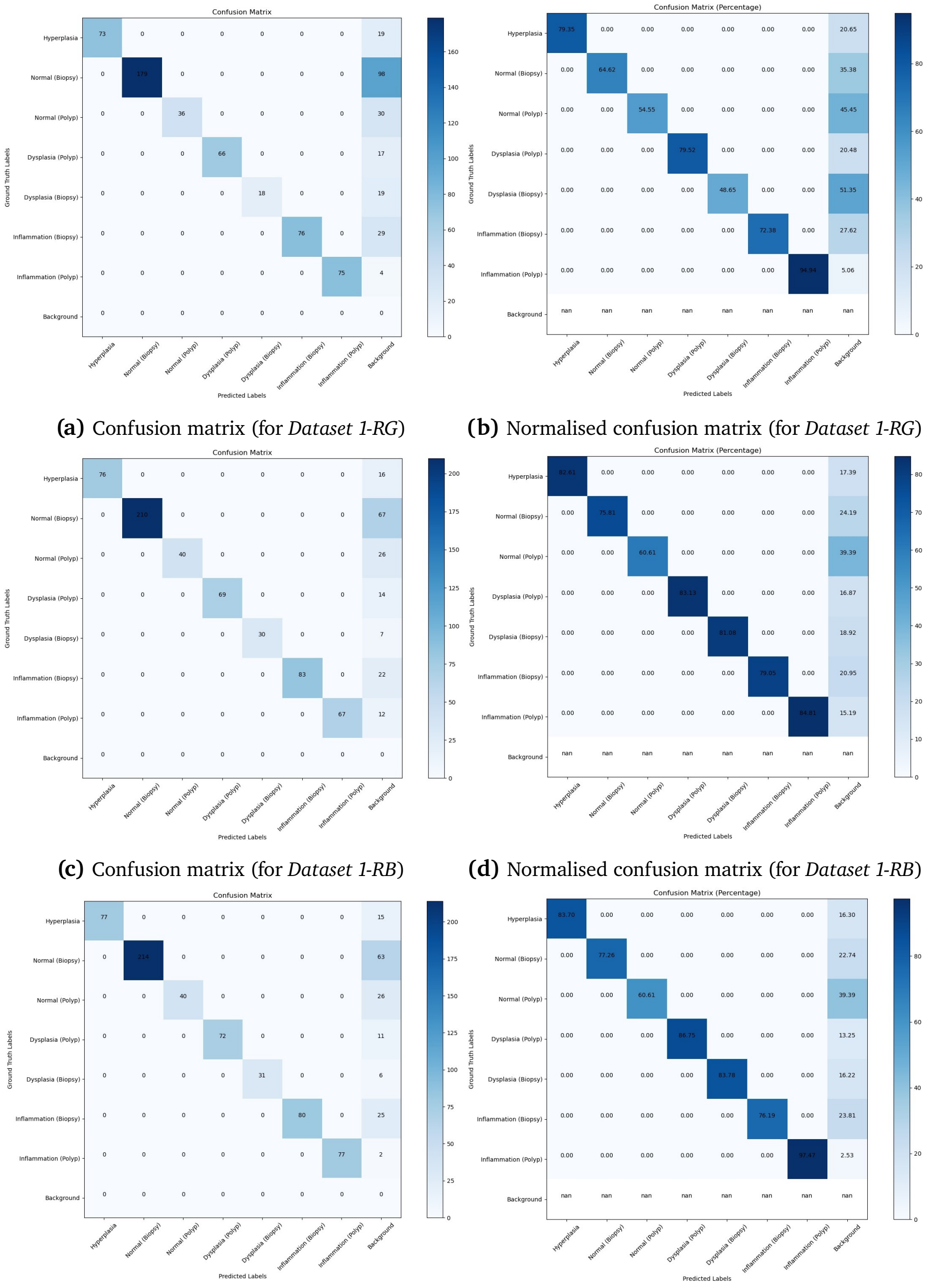
(e) Confusion matrix (for *Dataset 1-G*)

(f) Normalised confusion matrix (for *Dataset 1-G*)

(g) Confusion matrix (for *Dataset 1-B*)

(h) Normalised confusion matrix (for *Dataset 1-B*)

Figure 7.6: Confusion matrices for YOLOV8x model trained and tested on different colour channel datasets (with confidence threshold 0.9) - part I



(a) Confusion matrix (for *Dataset 1-RG*)

(b) Normalised confusion matrix (for *Dataset 1-RG*)

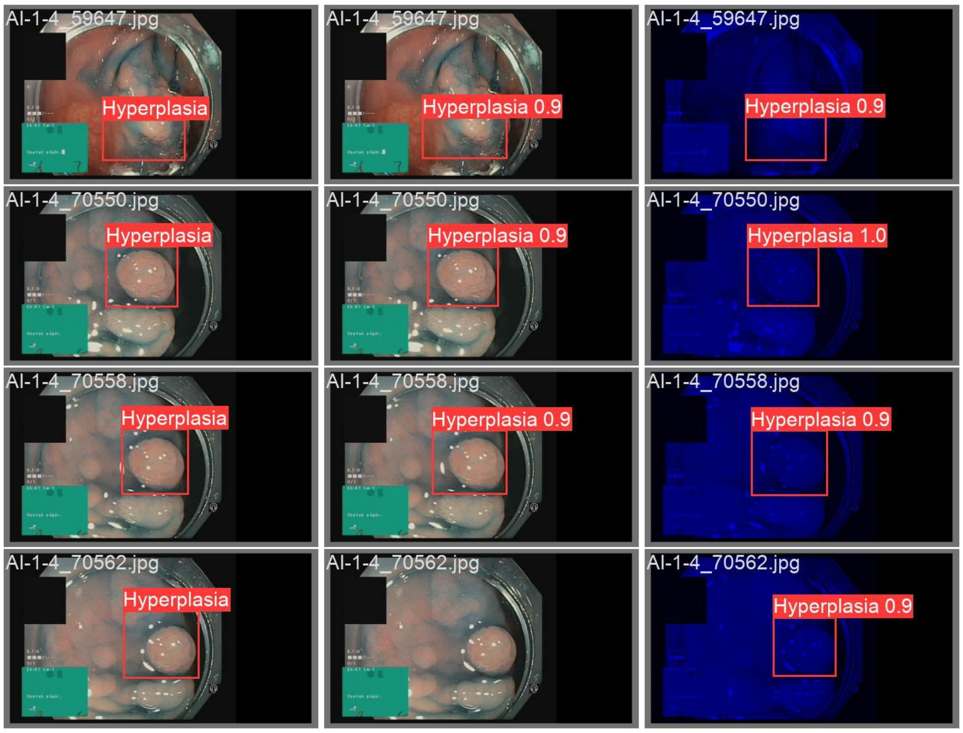
(c) Confusion matrix (for *Dataset 1-RB*)

(d) Normalised confusion matrix (for *Dataset 1-RB*)

(e) Confusion matrix (for *Dataset 1-GB*)

(f) Normalised confusion matrix (for *Dataset 1-GB*)

Figure 7.7: Confusion matrices for YOLOv8x model trained and tested on different colour channel datasets (with confidence threshold 0.9) - part II



(a) Ground Truth **(b)** *Dataset 1* prediction **(c)** *Dataset 1-B* prediction

Figure 7.8: Comparison of predictions of YOLOv8x model that was trained and tested on *Dataset 1* and *Dataset 1-B* for the same set of images, with confidence threshold 0.9

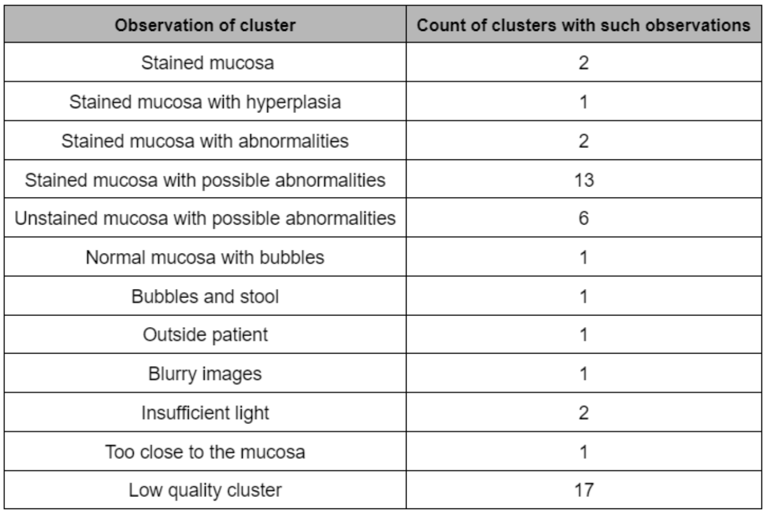
Dataset 1-B and then tested on the same set of test images, but from *Dataset 1-B*, with the confidence threshold again set as 0.9.

7.1.3 For DL-based Image Clustering Task

1. Image clustering on *Dataset 2a*

Given the high number of images in this dataset which totalled 45778 total images, it was impractical to assess the clustering results on the entire dataset in the limited time frame of this project, especially because there were **48 different clusters** that each video's frames were clustered into. However, some clusters were analysed to assess the performance of the clustering model, which are highlighted here-

- ERAI004 - Table 7.5 shows an assessment of the clusters observed for this video.

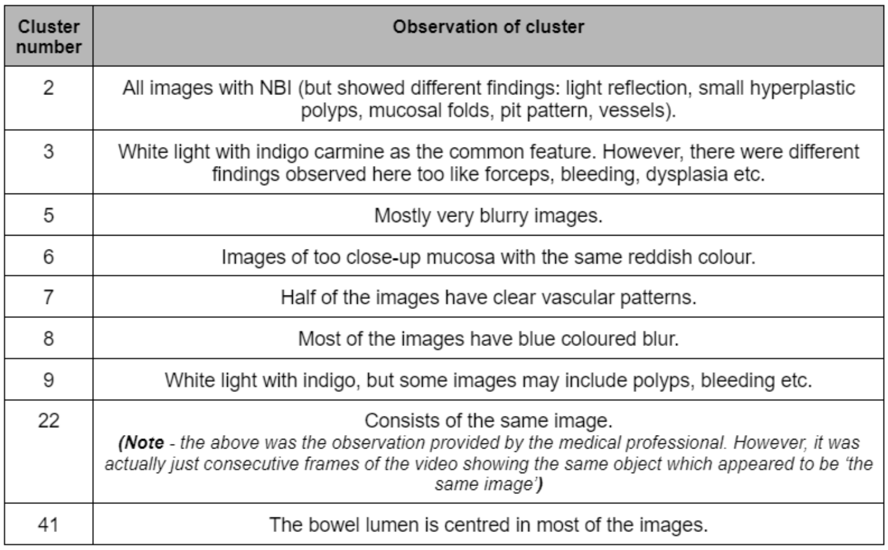


Observation of cluster	Count of clusters with such observations
Stained mucosa	2
Stained mucosa with hyperplasia	1
Stained mucosa with abnormalities	2
Stained mucosa with possible abnormalities	13
Unstained mucosa with possible abnormalities	6
Normal mucosa with bubbles	1
Bubbles and stool	1
Outside patient	1
Blurry images	1
Insufficient light	2
Too close to the mucosa	1
Low quality cluster	17

Table 7.5: *Dataset 2a* (Video ERAI004) clustering assessment

- ERAI006 - Not all clusters were thoroughly assessed for this video. However, there were some interesting findings observed in some clusters, as seen in Table 7.6.

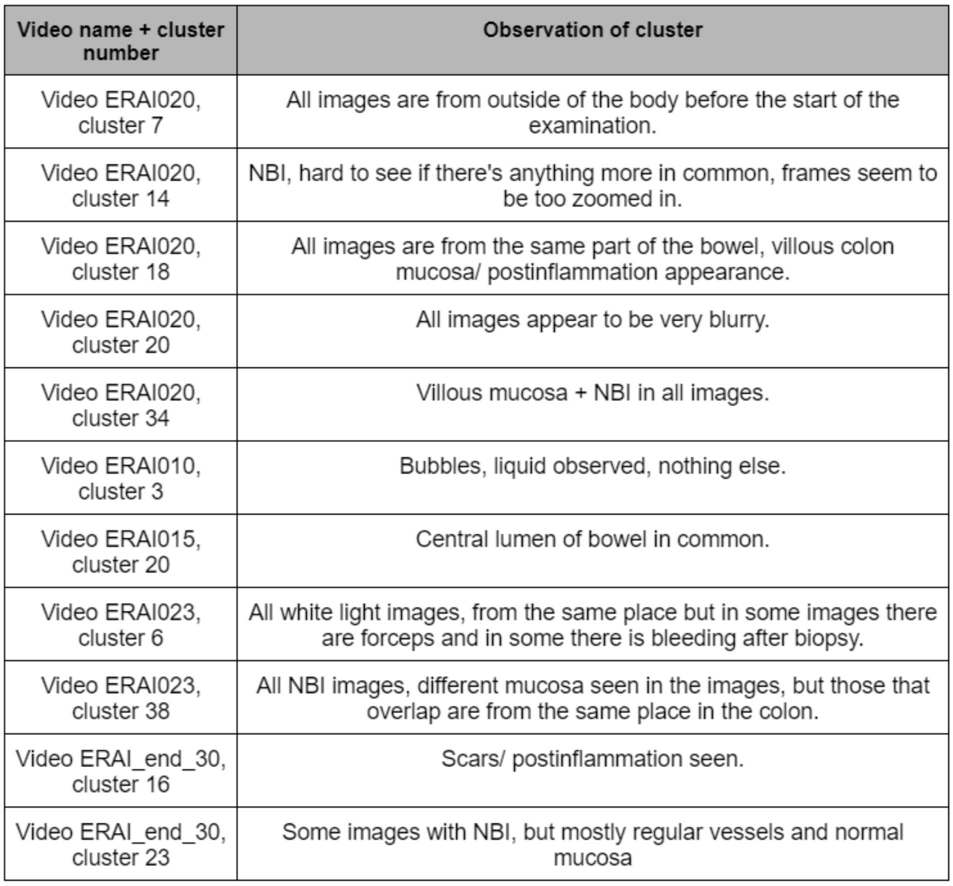
  The 48 clusters generated were numbered from 0-47, based on which the 'Cluster number' is mentioned in this table.



Cluster number	Observation of cluster
2	All images with NBI (but showed different findings: light reflection, small hyperplastic polyps, mucosal folds, pit pattern, vessels).
3	White light with indigo carmine as the common feature. However, there were different findings observed here too like forceps, bleeding, dysplasia etc.
5	Mostly very blurry images.
6	Images of too close-up mucosa with the same reddish colour.
7	Half of the images have clear vascular patterns.
8	Most of the images have blue coloured blur.
9	White light with indigo, but some images may include polyps, bleeding etc.
22	Consists of the same image. ***(Note** - the above was the observation provided by the medical professional. However, it was actually just consecutive frames of the video showing the same object which appeared to be 'the same image'**)***
41	The bowel lumen is centred in most of the images.

Table 7.6: *Dataset 2a* (Video ERAI006) clustering assessment

- Rest of the videos - The patterns and observations seen in some other clusters in the remainder of the videos are encapsulated in Table 7.7.



Video name + cluster number	Observation of cluster
Video ERAI020, cluster 7	All images are from outside of the body before the start of the examination.
Video ERAI020, cluster 14	NBI, hard to see if there's anything more in common, frames seem to be too zoomed in.
Video ERAI020, cluster 18	All images are from the same part of the bowel, villous colon mucosa/ postinflammation appearance.
Video ERAI020, cluster 20	All images appear to be very blurry.
Video ERAI020, cluster 34	Villous mucosa + NBI in all images.
Video ERAI010, cluster 3	Bubbles, liquid observed, nothing else.
Video ERAI015, cluster 20	Central lumen of bowel in common.
Video ERAI023, cluster 6	All white light images, from the same place but in some images there are forceps and in some there is bleeding after biopsy.
Video ERAI023, cluster 38	All NBI images, different mucosa seen in the images, but those that overlap are from the same place in the colon.
Video ERAI_end_30, cluster 16	Scars/ postinflammation seen.
Video ERAI_end_30, cluster 23	Some images with NBI, but mostly regular vessels and normal mucosa

Table 7.7: *Dataset 2a* (other videos) clustering assessment

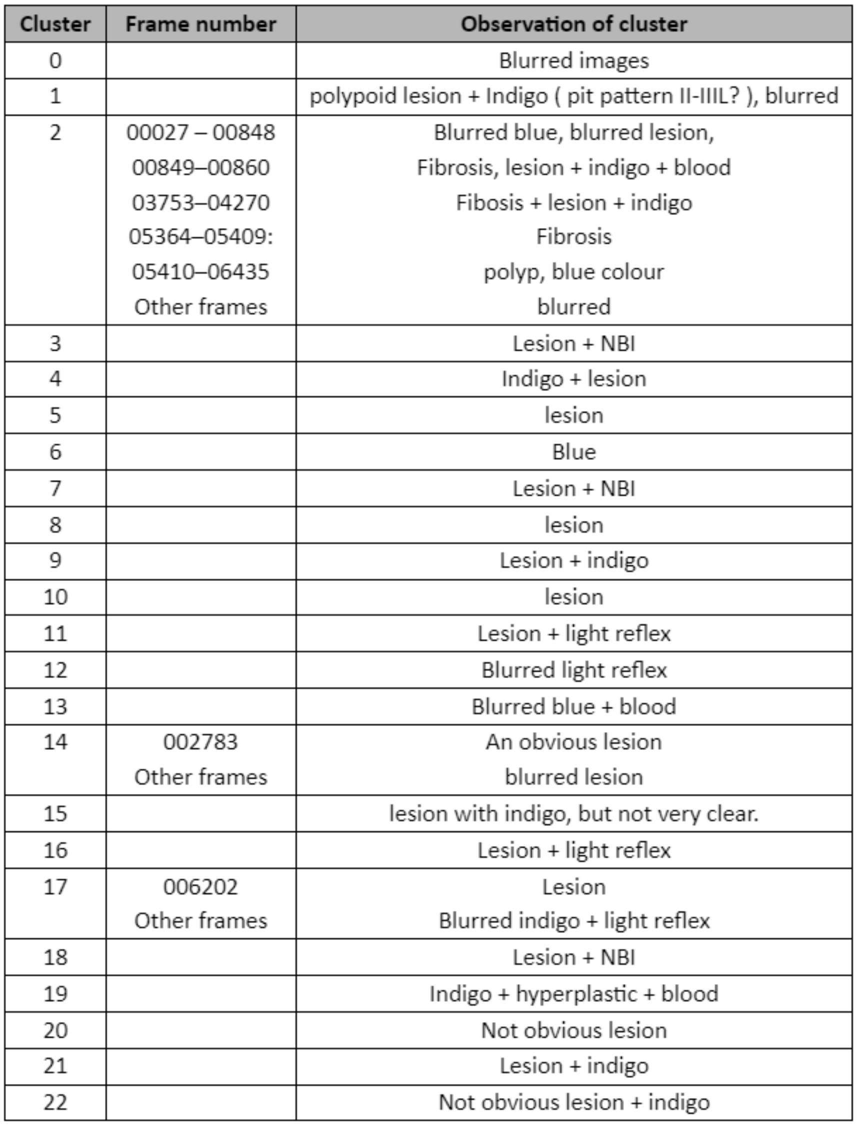
2. Image clustering on *Dataset 2b*

Just like *Dataset 2a*, with even *Dataset 2b* it was difficult to completely assess the clustering results because of the high number of video frames in it which totalled 108,214 total images.

Each video's frames were clustered into **23 separate clusters**, making the assessment of the quality of clustering even more difficult. However, it was possible to provide an assessment of some videos and clusters, which have been summed up in the tables below for different videos of the dataset.

The clusters here were indexed from 0-22. The column titled 'Frame number' refers to the particular video frame/frames to which the observation mentioned in that row of that table pertains to.

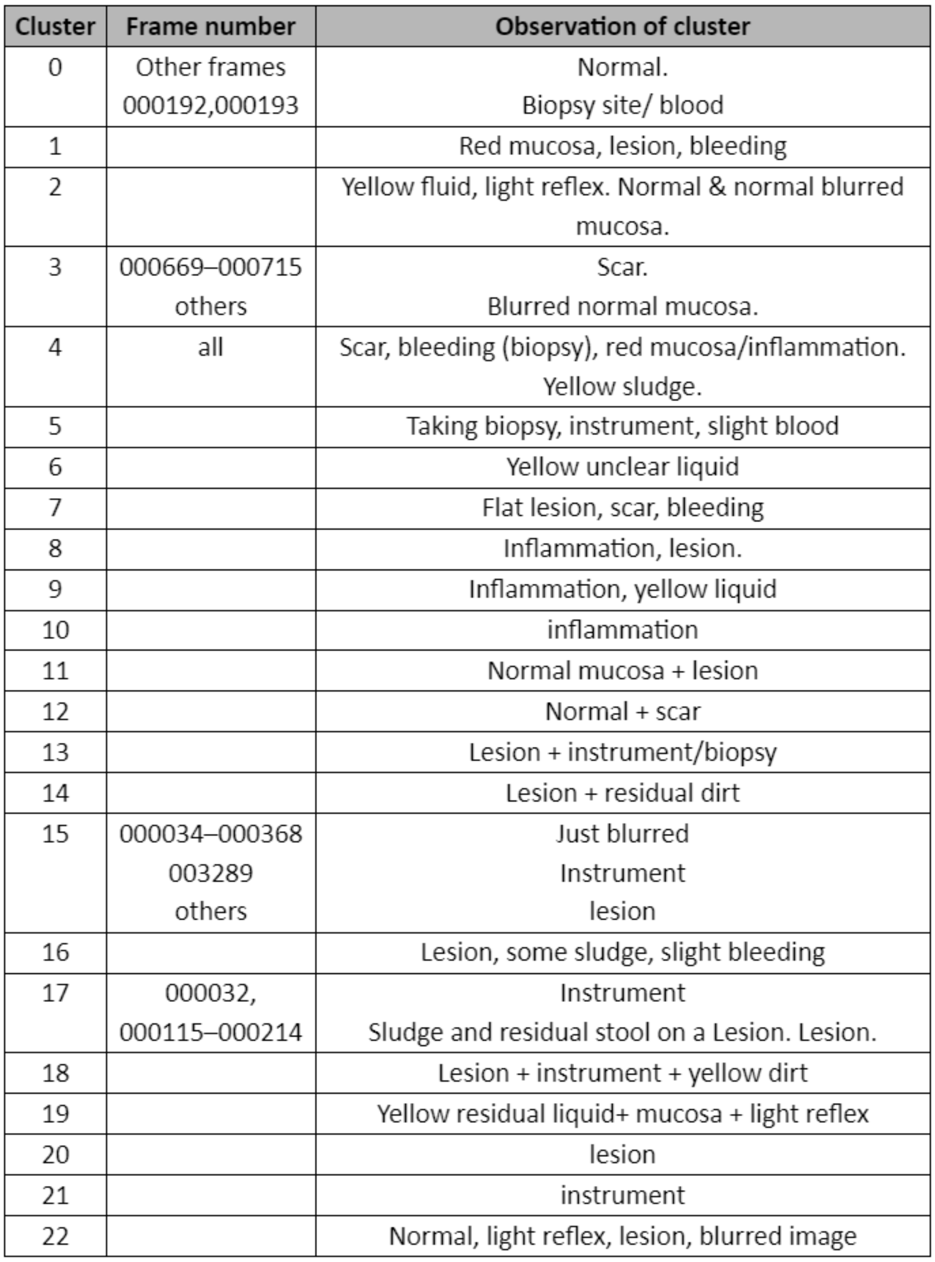
- ERAI006 - According to the histopathological findings associated with *Dataset 2b* seen in table 6.2, *Video ERAI006* had the finding 'Low grade dysplasia in a tubulous adenoma (polypoid lesion)'. The assessment for the clusters obtained from video ERAI006 is seen in Table 7.8.



Cluster	Frame number	Observation of cluster
0		Blurred images
1		polypoid lesion + Indigo (pit pattern II-IIIL?), blurred
2	00027 – 00848 00849–00860 03753–04270 05364–05409: 05410–06435 Other frames	Blurred blue, blurred lesion, Fibrosis, lesion + indigo + blood Fibosis + lesion + indigo Fibrosis polyp, blue colour blurred
3		Lesion + NBI
4		Indigo + lesion
5		lesion
6		Blue
7		Lesion + NBI
8		lesion
9		Lesion + indigo
10		lesion
11		Lesion + light reflex
12		Blurred light reflex
13		Blurred blue + blood
14	002783 Other frames	An obvious lesion blurred lesion
15		lesion with indigo, but not very clear.
16		Lesion + light reflex
17	006202 Other frames	Lesion Blurred indigo + light reflex
18		Lesion + NBI
19		Indigo + hyperplastic + blood
20		Not obvious lesion
21		Lesion + indigo
22		Not obvious lesion + indigo

Table 7.8: *Dataset 2b* (Video ERAI006) clustering assessment

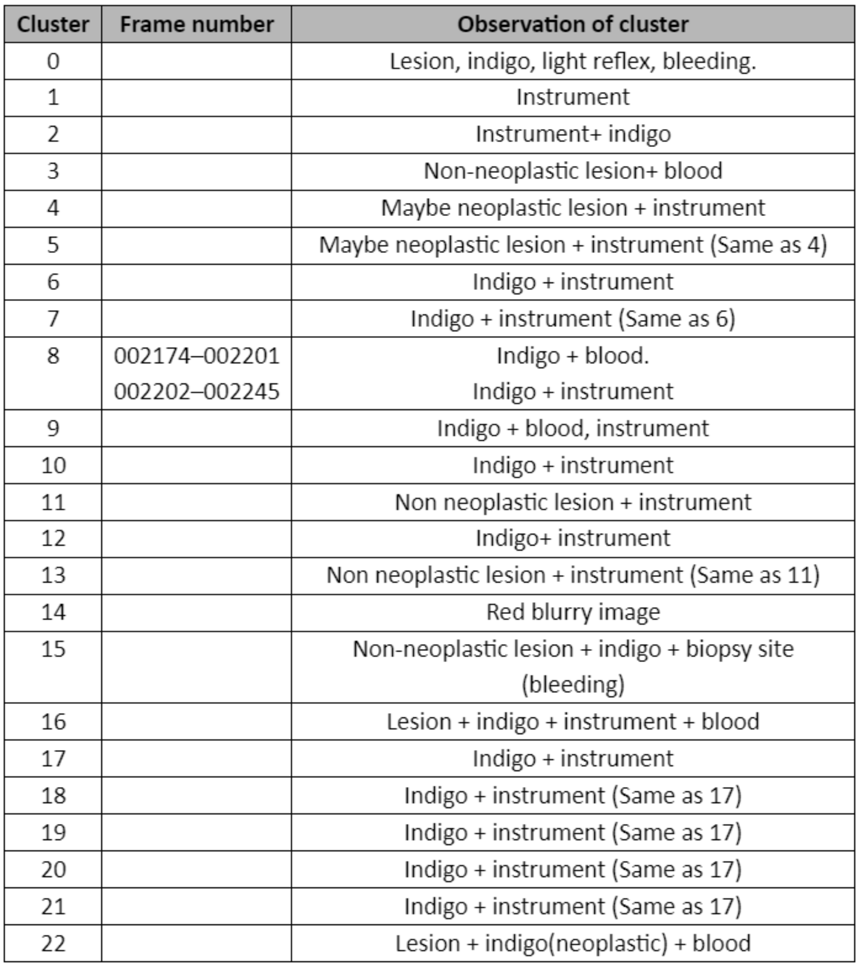
- ERAI028 - The histopathological findings associated with this video according to Table 6.2 is ‘Mild inflammatory activity’. The assessment of the clusters seen with this dataset is summed up in Table 7.9.



Cluster	Frame number	Observation of cluster
0	Other frames 000192,000193	Normal. Biopsy site/ blood
1		Red mucosa, lesion, bleeding
2		Yellow fluid, light reflex. Normal & normal blurred mucosa.
3	000669–000715 others	Scar. Blurred normal mucosa.
4	all	Scar, bleeding (biopsy), red mucosa/inflammation. Yellow sludge.
5		Taking biopsy, instrument, slight blood
6		Yellow unclear liquid
7		Flat lesion, scar, bleeding
8		Inflammation, lesion.
9		Inflammation, yellow liquid
10		inflammation
11		Normal mucosa + lesion
12		Normal + scar
13		Lesion + instrument/biopsy
14		Lesion + residual dirt
15	000034–000368 003289 others	Just blurred Instrument lesion
16		Lesion, some sludge, slight bleeding
17	000032, 000115–000214	Instrument Sludge and residual stool on a Lesion. Lesion.
18		Lesion + instrument + yellow dirt
19		Yellow residual liquid+ mucosa + light reflex
20		lesion
21		instrument
22		Normal, light reflex, lesion, blurred image

Table 7.9: *Dataset 2b* (Video ERAI028) clustering assessment

- ERAI010 - The histopathological findings associated with this video according to Table 6.2 is ‘Low grade dysplasia in a minimal tubulous adenoma’. The assessment of the clusters seen with this dataset is summed up in Table 7.10.



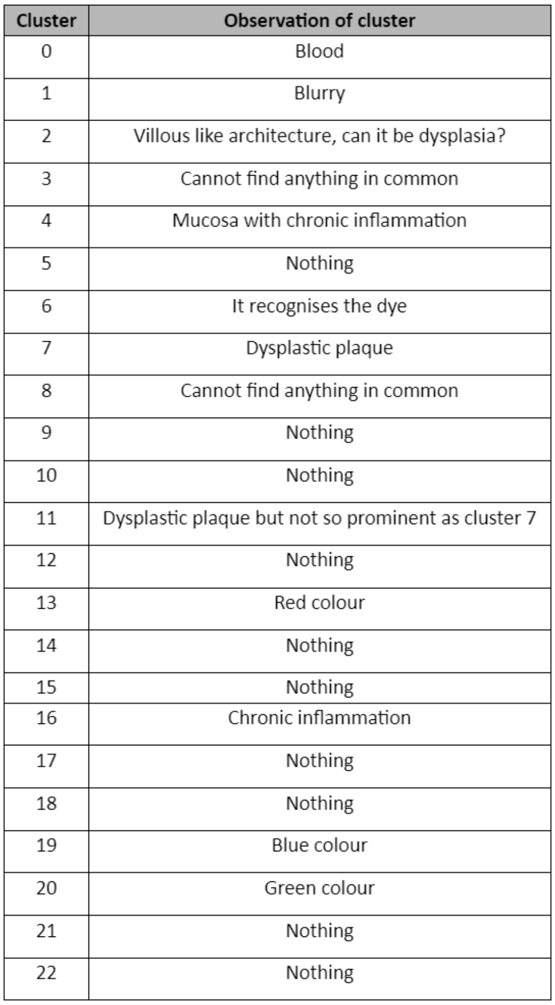
Cluster	Frame number	Observation of cluster
0		Lesion, indigo, light reflex, bleeding.
1		Instrument
2		Instrument+ indigo
3		Non-neoplastic lesion+ blood
4		Maybe neoplastic lesion + instrument
5		Maybe neoplastic lesion + instrument (Same as 4)
6		Indigo + instrument
7		Indigo + instrument (Same as 6)
8	002174–002201 002202–002245	Indigo + blood. Indigo + instrument
9		Indigo + blood, instrument
10		Indigo + instrument
11		Non neoplastic lesion + instrument
12		Indigo+ instrument
13		Non neoplastic lesion + instrument (Same as 11)
14		Red blurry image
15		Non-neoplastic lesion + indigo + biopsy site (bleeding)
16		Lesion + indigo + instrument + blood
17		Indigo + instrument
18		Indigo + instrument (Same as 17)
19		Indigo + instrument (Same as 17)
20		Indigo + instrument (Same as 17)
21		Indigo + instrument (Same as 17)
22		Lesion + indigo(neoplastic) + blood

Table 7.10: *Dataset 2b* (Video ERAI010) clustering assessment

- ERAI005 - This video had only 114 video frames extracted having the findings ‘Hyperplastic plaque, Irregular crypts’ according to Table 6.2.

  On clustering the frames of this video into 23 clusters, these were the observations:
  i) 13 clusters had the finding ‘Irregular crypts with chromo’,
  ii) 8 clusters had the finding ‘Granulate mucosa with chromo’,
  iii) 2 clusters had just a singular image each.

- ERAI020 - This video had 29050 video frames which was a very large number of frames to visually assess and note patterns for. The histopathological findings associated with this video according to table 6.2 is ‘Chronic inflammation, extended postinflammatory findings, villous structure-almost looking like small bowel mucosa. Paneth cell metaplasia/hyperplasia and crypt bifurcation.’ The assessment of the clusters seen with this dataset is summed up in Table 7.11.



Cluster	Observation of cluster
0	Blood
1	Blurry
2	Villous like architecture, can it be dysplasia?
3	Cannot find anything in common
4	Mucosa with chronic inflammation
5	Nothing
6	It recognises the dye
7	Dysplastic plaque
8	Cannot find anything in common
9	Nothing
10	Nothing
11	Dysplastic plaque but not so prominent as cluster 7
12	Nothing
13	Red colour
14	Nothing
15	Nothing
16	Chronic inflammation
17	Nothing
18	Nothing
19	Blue colour
20	Green colour
21	Nothing
22	Nothing

Table 7.11: *Dataset 2b* (Video ERAI020) clustering assessment

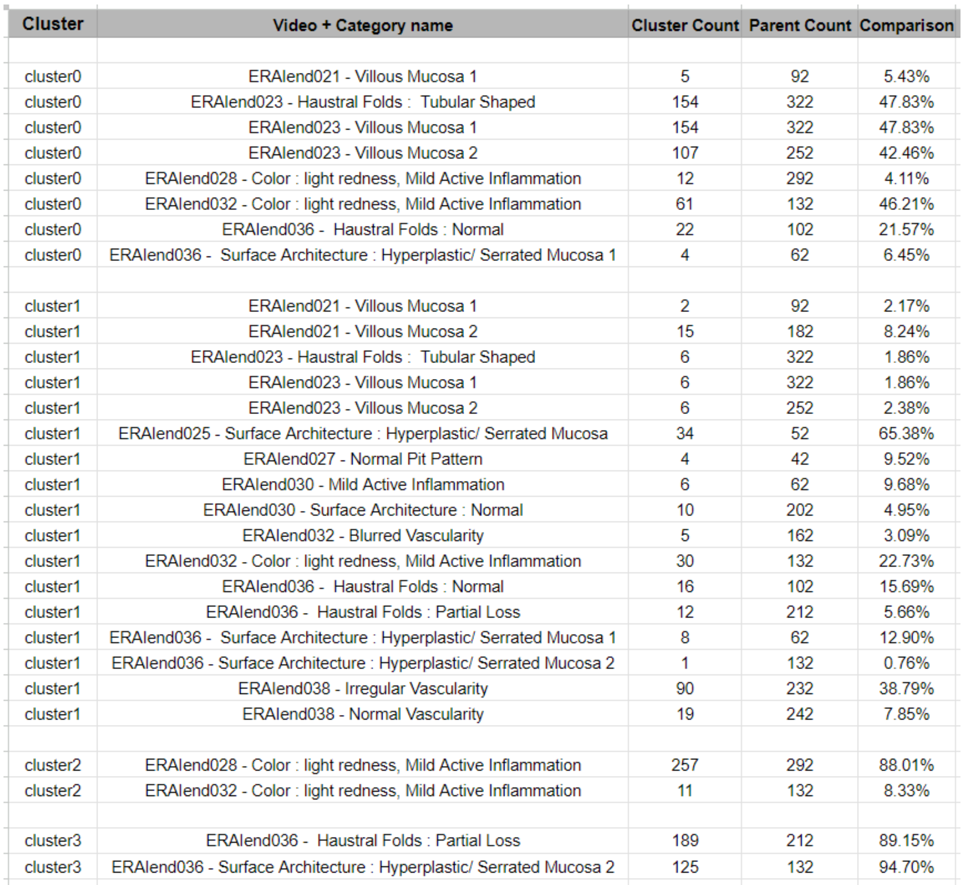
- Rest of the videos -

  1. ERAI003: This video had the histopathological findings ‘Flat hyperplastic mucosa, no dysplasia’ in 450 total video frames, according to table 6.2. In the resulting clusters, ‘Hyperplastic polyp and biopsy forceps’ was seen in clusters 2, 5, 6, 10, 11, 12, 13, 15, 16 and 18.

2. ERAI030: This video had ‘Chronic active inflammation with crypt inflammation’ as the histopathological findings, according to table 6.2. In the resultant clusters, Inflammation (in white light) was seen in clusters: 3, 7, 9, 12 and 16; while Inflammation (in NBI) was seen in clusters 5 and 8.

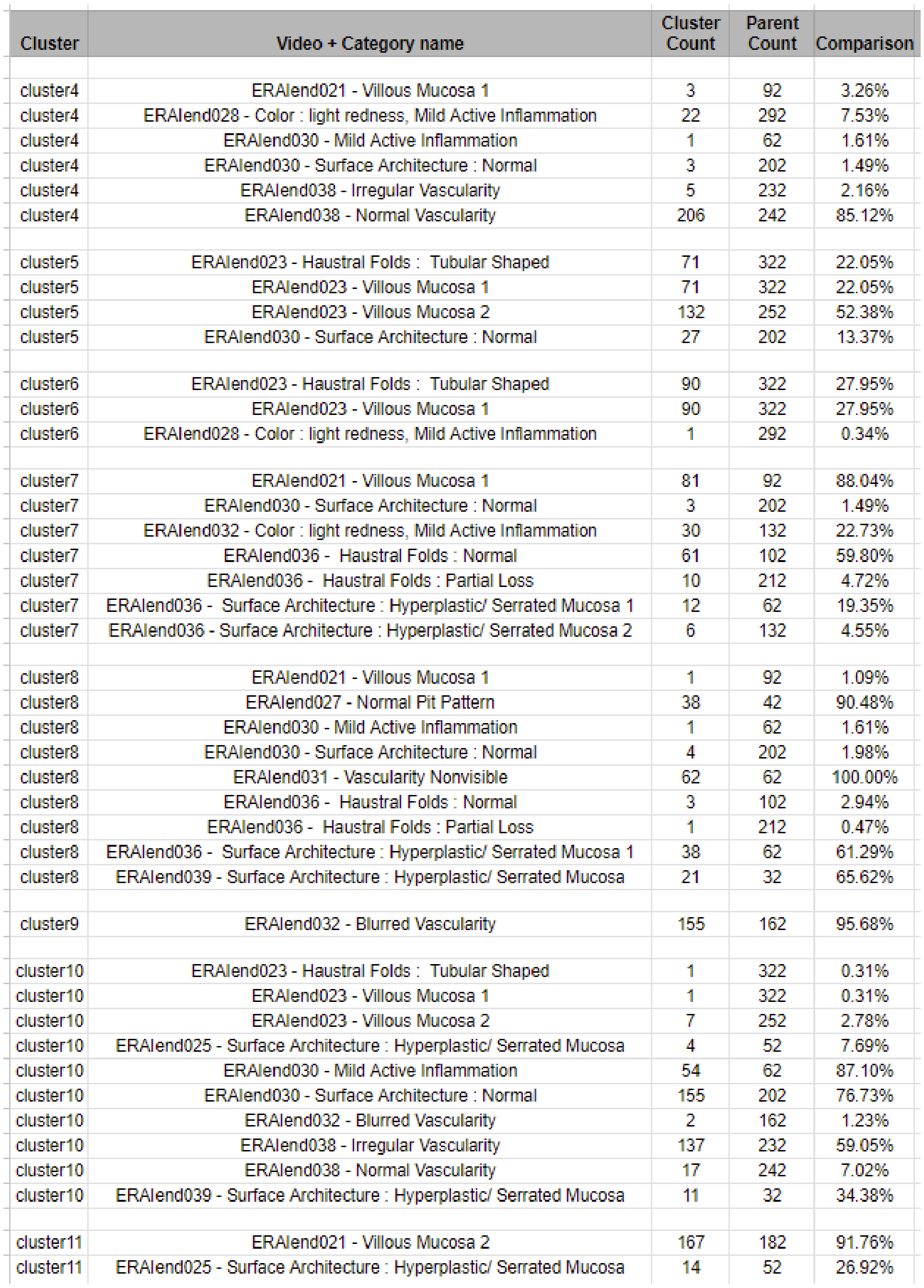
3. Image clustering on *Dataset 2c*

Dataset 2c came with proper time-stamp-based labels for the video frames, which is why it was easier to assess the performance of the clustering system on it, since there was no need to rely on the observations of a team of different medical professionals for their assessment of the clustering results. Tables 7.12 and 7.13 together provide an assessment of the results of the image clustering system when *Dataset 2c* is clustered into 12 different clusters.



Cluster	Video + Category name	Cluster Count	Parent Count	Comparison
cluster0	ERAIend021 - Villous Mucosa 1	5	92	5.43%
cluster0	ERAIend023 - Haustral Folds : Tubular Shaped	154	322	47.83%
cluster0	ERAIend023 - Villous Mucosa 1	154	322	47.83%
cluster0	ERAIend023 - Villous Mucosa 2	107	252	42.46%
cluster0	ERAIend028 - Color : light redness, Mild Active Inflammation	12	292	4.11%
cluster0	ERAIend032 - Color : light redness, Mild Active Inflammation	61	132	46.21%
cluster0	ERAIend036 - Haustral Folds : Normal	22	102	21.57%
cluster0	ERAIend036 - Surface Architecture : Hyperplastic/ Serrated Mucosa 1	4	62	6.45%
cluster1	ERAIend021 - Villous Mucosa 1	2	92	2.17%
cluster1	ERAIend021 - Villous Mucosa 2	15	182	8.24%
cluster1	ERAIend023 - Haustral Folds : Tubular Shaped	6	322	1.86%
cluster1	ERAIend023 - Villous Mucosa 1	6	322	1.86%
cluster1	ERAIend023 - Villous Mucosa 2	6	252	2.38%
cluster1	ERAIend025 - Surface Architecture : Hyperplastic/ Serrated Mucosa	34	52	65.38%
cluster1	ERAIend027 - Normal Pit Pattern	4	42	9.52%
cluster1	ERAIend030 - Mild Active Inflammation	6	62	9.68%
cluster1	ERAIend030 - Surface Architecture : Normal	10	202	4.95%
cluster1	ERAIend032 - Blurred Vascularity	5	162	3.09%
cluster1	ERAIend032 - Color : light redness, Mild Active Inflammation	30	132	22.73%
cluster1	ERAIend036 - Haustral Folds : Normal	16	102	15.69%
cluster1	ERAIend036 - Haustral Folds : Partial Loss	12	212	5.66%
cluster1	ERAIend036 - Surface Architecture : Hyperplastic/ Serrated Mucosa 1	8	62	12.90%
cluster1	ERAIend036 - Surface Architecture : Hyperplastic/ Serrated Mucosa 2	1	132	0.76%
cluster1	ERAIend038 - Irregular Vascularity	90	232	38.79%
cluster1	ERAIend038 - Normal Vascularity	19	242	7.85%
cluster2	ERAIend028 - Color : light redness, Mild Active Inflammation	257	292	88.01%
cluster2	ERAIend032 - Color : light redness, Mild Active Inflammation	11	132	8.33%
cluster3	ERAIend036 - Haustral Folds : Partial Loss	189	212	89.15%
cluster3	ERAIend036 - Surface Architecture : Hyperplastic/ Serrated Mucosa 2	125	132	94.70%

Table 7.12: Assessment of clustering *Dataset 2c* into 12 clusters - Part 1



Cluster	Video + Category name	Cluster Count	Parent Count	Comparison
cluster4	ERAlend021 - Villous Mucosa 1	3	92	3.26%
cluster4	ERAlend028 - Color : light redness, Mild Active Inflammation	22	292	7.53%
cluster4	ERAlend030 - Mild Active Inflammation	1	62	1.61%
cluster4	ERAlend030 - Surface Architecture : Normal	3	202	1.49%
cluster4	ERAlend038 - Irregular Vascularity	5	232	2.16%
cluster4	ERAlend038 - Normal Vascularity	206	242	85.12%
cluster5	ERAlend023 - Haustral Folds : Tubular Shaped	71	322	22.05%
cluster5	ERAlend023 - Villous Mucosa 1	71	322	22.05%
cluster5	ERAlend023 - Villous Mucosa 2	132	252	52.38%
cluster5	ERAlend030 - Surface Architecture : Normal	27	202	13.37%
cluster6	ERAlend023 - Haustral Folds : Tubular Shaped	90	322	27.95%
cluster6	ERAlend023 - Villous Mucosa 1	90	322	27.95%
cluster6	ERAlend028 - Color : light redness, Mild Active Inflammation	1	292	0.34%
cluster7	ERAlend021 - Villous Mucosa 1	81	92	88.04%
cluster7	ERAlend030 - Surface Architecture : Normal	3	202	1.49%
cluster7	ERAlend032 - Color : light redness, Mild Active Inflammation	30	132	22.73%
cluster7	ERAlend036 - Haustral Folds : Normal	61	102	59.80%
cluster7	ERAlend036 - Haustral Folds : Partial Loss	10	212	4.72%
cluster7	ERAlend036 - Surface Architecture : Hyperplastic/ Serrated Mucosa 1	12	62	19.35%
cluster7	ERAlend036 - Surface Architecture : Hyperplastic/ Serrated Mucosa 2	6	132	4.55%
cluster8	ERAlend021 - Villous Mucosa 1	1	92	1.09%
cluster8	ERAlend027 - Normal Pit Pattern	38	42	90.48%
cluster8	ERAlend030 - Mild Active Inflammation	1	62	1.61%
cluster8	ERAlend030 - Surface Architecture : Normal	4	202	1.98%
cluster8	ERAlend031 - Vascularity Nonvisible	62	62	100.00%
cluster8	ERAlend036 - Haustral Folds : Normal	3	102	2.94%
cluster8	ERAlend036 - Haustral Folds : Partial Loss	1	212	0.47%
cluster8	ERAlend036 - Surface Architecture : Hyperplastic/ Serrated Mucosa 1	38	62	61.29%
cluster8	ERAlend039 - Surface Architecture : Hyperplastic/ Serrated Mucosa	21	32	65.62%
cluster9	ERAlend032 - Blurred Vascularity	155	162	95.68%
cluster10	ERAlend023 - Haustral Folds : Tubular Shaped	1	322	0.31%
cluster10	ERAlend023 - Villous Mucosa 1	1	322	0.31%
cluster10	ERAlend023 - Villous Mucosa 2	7	252	2.78%
cluster10	ERAlend025 - Surface Architecture : Hyperplastic/ Serrated Mucosa	4	52	7.69%
cluster10	ERAlend030 - Mild Active Inflammation	54	62	87.10%
cluster10	ERAlend030 - Surface Architecture : Normal	155	202	76.73%
cluster10	ERAlend032 - Blurred Vascularity	2	162	1.23%
cluster10	ERAlend038 - Irregular Vascularity	137	232	59.05%
cluster10	ERAlend038 - Normal Vascularity	17	242	7.02%
cluster10	ERAlend039 - Surface Architecture : Hyperplastic/ Serrated Mucosa	11	32	34.38%
cluster11	ERAlend021 - Villous Mucosa 2	167	182	91.76%
cluster11	ERAlend025 - Surface Architecture : Hyperplastic/ Serrated Mucosa	14	52	26.92%

Table 7.13: Assessment of clustering *Dataset 2c* into 12 clusters - Part 2

The idea behind using 12 as the number of clusters was to try to cluster the video frames into the same number of clusters as there are unique categories of findings (which was 12 for *Dataset 2c*, as seen in Table 6.5).

The column titled 'Video + Category name' refers to the column of the same name in Table 6.3, where *Dataset 2c* is described. In the tables 7.12 and 7.13 the 'Cluster count' refers to the number of video frames of that category in that cluster, while the 'Parent count' refers to the number of images of that category in the Parent Dataset, i.e. *Dataset 2c*. The 'Comparison' column shows what percentage of a category from the 'Parent Dataset' is in that cluster. So for example, the first category in cluster 1 is 'ERAIend021 - Villous Mucosa 1'. The 'Cluster Count' of 2 here means that there are 2 video frames of this category in this cluster, while the 'Parent Count' of 92 means that there are a total of 92 video frames of the category 'ERAIend021 - Villous Mucosa 1' in the 'Parent Dataset' i.e. *Dataset 2c*. The comparison value of 2.17% means that 2.17% of 'ERAIend021 - Villous Mucosa 1' from *Dataset 2c* are in cluster 1.

7.2 Discussion

7.2.1 For Object Detection with *Dataset 1*

The results observed for the *Object detection task* with *Dataset 1* were very promising.

Figure 7.1 shows that the model that took the least amount of time for training on *Dataset 1* was YOLOv8n (Nano), which took 0.958 hours to finish training for 100 epochs. Meanwhile, YOLOv8x (Extra Large) took the highest training time of 4.949 hours for 100 epochs of training. However, this was also the best-performing model, in terms of the mAP50-95 value on the test split of the dataset. YOLOv8x also took the highest inference time per image of 8.7 ms. The YOLOv5n (nano) model took the least inference time per image, but it was also the worst performing of these models in terms of mAP50-95 value.

It was also seen that the nano model of YOLOv5 was worse performing than the nano model of YOLOv8, in terms of the mAP50-95 value. The same trend was observed for the other sized models of YOLOv5 and YOLOv8. This leads to the conclusion that YOLOv8 was a much better-performing model in terms of mAP50-95 values than YOLOv5. This was why for the other experiments associated with the *Object Detection Task* including those with the colour-channel separated and recombined datasets, YOLOv8x was the model used for comparison.

From the numerical results seen in tables 7.1 and 7.2, it was observed that when all YOLOv8 and YOLOv5 models were trained with the same dataset for a standard number of training epochs (which was 100, here) and then tested on the same test dataset with the default testing parameters, the **YOLOv8x** model was the best performing one in terms of the mAP50-95 value, which was 0.868 for it. But if there is a need to find a balance between high mAP50-95 value, low training time for the model, and low inference time for prediction on test images, then **YOLOv8l** is a more suitable option. It offers a mAP50-95 value of 0.867 while having a training time that was 40.59% less than that of the YOLOv8x model, while also having an inference time that was 36.7% less than that of the YOLOv8x model.

Figure 7.2 shows the plot-based results of training and validating a YOLOv8x model with *Dataset 1*, more commonly referred to by the term *learning curves*. It is observed that the validation loss, represented by *val/box_loss*, *val/cls_loss* and *val/dfl_loss*, seem to be approaching a plateau. Moreover, the metrics like precision, recall and mAP50 also seem to have reached a plateau too and are not giving any significant improvement with an increase in the number of epochs. The mAP50-95 value seems to be improving with the number of epochs, so there is a potential to make the model further improve with respect to this evaluation metric, with further training over a larger number of epochs. Based on the graphs here, it can be concluded that underfitting is certainly not an issue with the model here. A similar trend in the plateauing of these metrics and loss values was seen in most other models' training with different datasets.

From Table 7.3 and the confusion matrices in Figure 7.4, it was observed that if the need is for **more accurate class predictions**, YOLOv8x is a better choice for a model than YOLOv8l. When the confidence threshold parameter was set as 0.9 for testing in order to get highly accurate predictions, the YOLOv8x model gives more correct predictions of class labels than the YOLOv8l model in 5 out of the 7 classes in *Dataset 1*, as seen in the confusion matrices. The YOLOv8x model consistently gave a higher number of accurate predictions than the YOLOv8l model across all classes except *Dysplasia (Biopsy)* where both models had the same number of accurate predictions, and *Inflammation (Polyp)*, where the YOLOv8l model was more accurate.

In Table 7.3, a much more significant difference in the performance of the different sized models was seen, especially between the worst performing (YOLOv8n) and best performing (YOLOv8x) model, in terms of the mAP50-95 values. The performance of the YOLOv8s and YOLOv8m models were very close, and this was also the case for the YOLOv8l and YOLOv8x models, with the YOLOv8l model being just slightly better in terms of the mAP50 value, while the YOLOv8x model was slightly better in terms of the mAP50-95 value. However, the difference between the performance of the YOLOv8x and YOLOv8l models was more noticeable when analysing their confusion matrices, as seen in Figure 7.4 and discussed earlier.

Hence, for testing the performance of the colour channel separation and recombination pipeline with the object detection model, the YOLOv8x model was chosen to be the one to train and test these models on.

7.2.2 For Object Detection with colour-channel separation and recombination

The promising results observed on training the *object detection* models with *Dataset 1* were further improved, when the datasets formed after **colour channel separation and recombination** were used.

From Table 7.4 and Figure 7.5, it was seen that when the object detection model (more specifically, YOLOv8x) was trained and tested on the datasets produced after colour channel separation and recombination of the RGB images, just having the **Blue-channel images** of the original RGB images in the dataset was enough to significantly improve the performance of the YOLOv8x model on the test datasets, when the confidence threshold was 0.9. With *Dataset 1-B* (which had only the blue channel images), the mAP50-95 value increased to 0.828, when compared to the value of mAP50-95 value of 0.78 given by the same model on the same set of test images, but when trained and tested with *Dataset 1* which had RGB images. This was an improvement of **6.15%**. It was only *Dataset 1-RG* which gave a decrease in the mAP50-95 value, as seen in Table 7.4. This made sense, since *Dataset 1-RG* did not have the blue channel of the original RGB images, the same channel which was responsible for producing the best results.

Moreover, observing the confusion matrices in figures 7.6 and 7.7, it was seen that the colour channel separation and recombination was contributing towards producing much more accurate predictions (except in the case of *Dataset 1-RG*) across most classes, with ***Dataset 1-B*** outperforming every other dataset in terms of the number of accurate predictions for all classes except one, as seen in Figures 7.6g and 7.6h.

From subfigures 7.6a and 7.7a it was seen that even when it comes to the worst performing dataset (*Dataset 1-RG*), it gave more accurate predictions than *Dataset 1* for 1 class - 'Inflammation (Polyp)'. However, *Dataset 1* gave much more accurate predictions across all other classes when compared to *Dataset 1-RG*.

Finally, figure 7.8 shows a visual comparison between the predictions made by the YOLOv8x model that was trained and tested on images from *Dataset 1*, and the YOLOv8x model that was trained and tested on images from *Dataset 1-B* (for the confidence threshold set as 0.9). It was seen that for Image 'AI-1-4_70550.jpg', the model trained and tested on *Dataset 1* gave a lower confidence score (0.9) than the model trained and tested on *Dataset 1-B*, which gave a confidence score of 1.0 . Also in the case of Image 'AI-1-4_70562.jpg', the model trained and tested on *Dataset 1* did not detect any object and hence gave no predictions, while the model trained and tested with *Dataset 1-B* that had just the blue channel of the image, was able to detect the object present and give the correct prediction for it. This demonstrates the superiority of the YOLOv8x model that was trained and tested with just the blue channel of the original RGB images.

7.2.3 For DL-based Image Clustering Task

On *Dataset 2a*

Table 7.5 where the assessment of the clusters observed for the video *ERAI004* are mentioned, shows that there were 17 low-quality clusters with no noticeable observations. However, the clustering model was effective in clustering into separate clusters the images that were visually distinguishable and different from the rest of the video. The observation of 'Outside patient' refers to the frames of the video that were captured when the endoscope was outside the patient, and so were visually much different than the rest of the images. Similarly there is a separate cluster for 'Bubbles and stool' that was visually dissimilar to the other clusters since stool and bubbles had different visual features. The model was also successful in finding clusters with possible abnormalities as well as actual abnormalities and hyperplasia. Finally, the low-quality clusters only formed a minority (35.41%) of all the 48 clusters generated by the clustering model, hence signifying the model's effectiveness.

The assessment for video *ERAI006* was mentioned in Table 7.6. The doctor reviewing these clustering results gave an additional comment that it appeared that the clustering seemed mostly based on colours at that point in time. This was justified, since DL models tend to extract visual features of an image, which includes colour. It was also observed that some clusters showed a tendency to have several consecutive frames in the same cluster. This can be explained by the fact that consecutive frames in a colonoscopy video tended to show the same visual features and looked very similar visually if the endoscope was not moving too much. Hence, based on the visual similarity of consecutive frames, the model had a tendency to cluster them together. A positive pattern observed here was that just like in video *ERAI004*, the model tended to cluster the blurry video frames into a separate cluster.

The assessment for some of the clusters in the other videos in this dataset was summed up in Table 7.7. From these observations, it was again noticeable that the image clustering system worked well in clustering out visually different images, like the 'Video ERAI020, cluster 7' where all video frames were from outside the body. Similarly, the clustering system was also able to cluster out video frames where NBI was used, which can be explained by how NBI creates different lighting conditions and hence results in visually different images than those in the rest of the colonoscopy video.

On *Dataset 2b*

The assessment for the clusters observed for video *ERAI006* in this dataset was seen in Table 7.8. It was promising to see the clustering model being able to cluster the video frames having lesions into separate clusters. It was also noticeable that the lesions with NBI formed 3 separate clusters, while the lesions with Indigo were also forming different clusters. This also supports the observation seen from the clustering results on *Dataset 2a*, that the clustering model tends to separate out video frames into clusters on the basis of colour, since NBI and Indigo both add a distinct colour to the video frames. There were also a few clusters with blurred images, which can be explained by the fact that there was motion blur while moving the endoscope to conduct the colonoscopy, and so there are a lot of video frames that are blurry because of it. Moreover, since in *Dataset 2b* all the videos' frames were extracted at the original frame rate of 50 fps, there were a lot more blurry images than that seen in the previous dataset, *Dataset 2b*, where a lower frame rate was used for video frame extraction.

Table 7.9 mentions the assessment of the clusters for video *ERAI028* in this dataset. The clustering system was able to create separate clusters of the video frames that had inflammation as a finding in them. It was also able to create separate clusters for the video frames where instruments appeared, which have much different visual features than the rest of the colorectal tract.

From Table 7.10 which shows the assessment of clustering results on video *ERAI010*, it was seen that there were some clusters with possibly neoplastic lesions, while some with non-neoplastic lesions and 1 cluster with neoplastic lesion. So the clustering of lesions or potential lesions into separate clusters showed promise, again. Moreover, there were separate clusters where instruments appeared, sometimes also with lesions in them. This was natural, since sometimes during colonoscopies, instruments are used to conduct biopsies on lesions, which is why in some video frames instruments and a lesion can appear together. The video frames with Indigo also formed a separate set of clusters, since the indigo dye gave the video frames a distinguishable colour that was different from the rest of the video frames.

For video *ERAI005*, the observation of the resulting clusters showed how clustering just a small number of frames especially into a relatively high number (23) of clusters was not very productive. Since the histopathological findings for this video had just 2 separate findings (as seen in Table 6.2), in the clustering results there were mostly 2 different kinds of clusters observed - each with 1 type of finding. At least the ability of the clustering system to separate out the 2 different types of findings into different clusters was promising.

The clustering results for video *ERAI020* are mentioned in Table 7.11. The results obtained here were really promising, which might be because the original video had a very large number of findings which could have visually distinguishable features that the clustering system might have been able to use in order to better separate the video frames into clusters with common findings in them. There were 1-2 separate clusters for each of these observations- 'Blood', 'Dysplastic plaque', 'Chronic inflammation' and 'Mucosa with chronic inflammation'. The ability of the clustering system to create separate clusters for each of these findings was really useful. Moreover, there was a separate cluster for video frames that were blurry. Similarly, there was an individual cluster each for the video frames that appeared red coloured, those that appeared blue coloured, and those that appeared green coloured. However, there were 12 clusters where no pattern was observed and which showed no promising observations in them. This was to be expected, when there were so many video frames for this video. However, there is a possibility that these clusters with nothing in common could possibly have a few findings in them that were just mixed up with other frames with no-findings. This would be a potentially negative aspect of this clustering system.

Finally, for video *ERAI030*, the fact that the clustering system created different clusters for images with NBI and different clusters for images with white light for the same finding of Inflammation again demonstrates the clustering system's ability to distinguish between white light and NBI video frames, because of the different visual characteristics the video frames having these tend to display.

On *Dataset 2c*

The clustering results for this dataset were analysed numerically rather than visually, and these results are mentioned in Tables 7.12 and 7.13.

There were a few promising results/clusters seen here. For cluster 2 (in Table 7.12), there were only 'Mild Active Inflammation' video frames, and there were frames of 2 different videos having this same finding that was clustered together here. Similarly, in cluster 9 (seen in Table 7.13), there were only video frames of

1 finding - ‘Blurred Vascularity’. Cluster 4 had a majority (85.8%) of its images being of just 1 finding - ‘ERAIend038 - Normal Vascularity’, while a few other images made up the rest of the cluster.

However, some other clusters had a mixture of findings in them and did not show an ability to distinguish between the different findings in the dataset.

8

Future Work

Despite the promising results obtained over the course of this project, there is significant scope for further exploring this topic of deep learning in precancerous lesion detection of chronic IBD patients undergoing surveillance colonoscopies, and improving on the results obtained in this project.

The most promising outcome of this project was the improvement in the performance of the object detection model YOLOv8 when colour channel separation and recombination was done on the original dataset's images. This approach can also be applied and explored across other fields of medical research when working with data similar to the one used in this thesis. There is also a scope to explore whether this approach also works as effectively on other colonoscopy datasets, and with other object detection models.

The *Object Detection* models that were trained over the course of this project can be tested on a different test dataset having images of the same classes as *Dataset 1,* but which come from a different source. This will help determine how well the models trained here generalise across other datasets with the same classes of findings in them.

Dataset 1 (as seen in table 5.1) that was used in this project had an imbalanced number of images across different classes. In future works, more images from all classes can be collected in order to make this a more balanced dataset, when it comes to the number of images in each class. Moreover, DL-based techniques like Generative Adversarial Network (GAN) [121] can be used to artificially synthesise additional images for these classes, hence significantly increasing the size of the dataset while also solving any potential class-imbalance problems associated with the original dataset.

Dataset 2 was initially supposed to come with bounding-box annotations for the findings in it, but those annotations could not be generated before or over the course of this project. So, another scope for future work is to create a bounding box annotations based version of *Dataset 2*. This new dataset can be used to explore the performance of object detection models, as was the initial intention of this project.

Since *Dataset 2* got time-stamp-based labels of histopathological findings for many of its videos by the end of this project, these labels can now be used to train Image Classification models with *Dataset 2*.

There is also scope for doing a clinical study where the improved colour-channel separation and recombination based object detection system can be integrated with the endoscopy systems currently in use to create a Computer-Aided Diagnosis (CAD) tool, to see if this helps improve the detection of precancerous lesions. Seeing the implementation and potential improvement to precancerous lesion detection provided by this CAD tool during live colonoscopies will give an important evaluation of the real-life performance of this system, and its potential benefits and fallacies.

Finally, when it comes to the image-clustering system, even this system can be further improved. Instead of using a VGG-16 or VGG-19 model that was pre-trained on the ImageNet dataset as was done in this project, the VGG-16 or VGG-19 model can be pre-trained with a dataset of colonoscopy video frames with pre-cancerous lesions/findings in them. There is a potential for a DL model that has been trained on a dataset of precancerous lesions detection/classification to extract more useful features in the clustering pipeline, than a DL model that has just been trained on a generic image classification dataset like ImageNet. An additional thing that can be tried in this regard is using a DL model that has been pre-trained on ImageNet and then further training it on a precancerous-lesion dataset, hence making use of transfer learning. Other clustering approaches based on Deep Learning can also be explored for this system.

One last thing that can be explored when it comes to the image-clustering system is the colour-channel separation and recombination pipeline mentioned in this thesis, to see if that helps the DL model extract more useful features and hence help form better clusters.

9 Conclusion

The work done over the course of this project helps increase the understanding of how deep learning can be used to help precancerous lesion detection in chronic IBD patients undergoing surveillance colonoscopies.

The object detection system that was experimented with in this project showed how an object detection model like **YOLOv8x that uses the blue channel** of the original RGB images of a dataset of precancerous findings can give **significantly better performance** than a model that uses the RGB channels of the images for training and prediction. The improvement in performance of the object detection system associated with using just the blue channel of the images has significant implications not just in future systems associated with colonoscopies, but also those associated with other forms of medical research.

This thesis also provides a comparison of the performance of the different-sized YOLOv5 and YOLOv8 models when trained for the same number of epochs on the same dataset. It was seen that the YOLOv8x (extra large) model was the best-performing model, while YOLOv5n (nano) was the worst-performing one. It was also seen that the YOLOv8 models were all better performing than their YOLOv5 counterparts.

It was noted that the colour channel separation and recombination of the original RGB images in the dataset gave an improvement in the performance of the object-detection model in all cases but one - *Dataset 1-RG*, where the blue channel was missing. This makes sense, since the dataset having just the blue channel of the images, *Dataset 1-B*, was the best-performing one. So removing the data stored in the blue channel of the original RGB images had the potential to lead to a worse-performing model, as was observed here.

Dataset 1 that was used in this thesis for the object detection task had a few issues associated with it, which contributed to the models trained in this project not being ideal for clinical applications in its current state. The first issue is the class imbalance problem with this dataset, where a few classes had a very high number of instances in the dataset, while a few classes had a very few number of instances in the same dataset. Secondly, the test dataset on which the object detection models were tested was split from *Dataset 1*, which was also the source for the training and validation datasets for these models. This led to the testing and training splits of *Dataset 1* being visually similar, hence resulting in models that potentially over-fitted on *Dataset 1*, giving excellent performance metrics on the test dataset, but having an increased risk of not generalising well for generating predictions on images from other datasets that are visually different than *Dataset 1*. *Dataset 1* also happens to be an older dataset from 2021, and current endoscopy systems used for conducting colonoscopies generate images of much higher resolution than the images in *Dataset 1*, which is another reason why the models trained with this dataset are not ideal for clinical applications in their current form.

The DL-based image clustering system's results on the other hand were a bit more of a mixed bag. This system was able to cluster the video frames into separate clusters of visually distinguishable video frames, like the ones where there were instruments, the ones from outside the patient, or the ones containing stool. Moreover, video frames in which NBI was used instead of white light also tended to get segregated into separate clusters. Video frames with motion blur also tended to form separate clusters in most cases. When this system was applied on *Dataset 2b* which had video frames of only findings, the system's ability to group the precancerous findings into clusters was seen. However, the qualitative form of assessment done on these clustering results as well as the inability to analyse patterns observed in each individual cluster of all the clusters is a potential negative when it comes to the assessment of the clustering system. This was attempted to be rectified by using *Dataset 2c* where a numerical approach towards analysing the clustering results was done. With *Dataset 2c*, it was seen that a better approach to creating clusters of findings in a dataset is to run the clustering system on a single folder with each video's frames, so that similar findings from different videos can be potentially separated out into the same cluster by the DL-based clustering system. This ability of the system was noticed with *Dataset 2c* when it came to findings like 'Mild Active Inflammation' and 'Blurred Vascularity'.

Despite these positive results, it was observed that the DL-based clustering system's performance in grouping together video frames with the same findings into the same cluster was erratic and unreliable. With *Dataset 2a* and *Dataset 2b*, it was seen that the system often showed a tendency to create clusters of frames showing the same colour or visual characteristics. It was also observed that the system had a tendency to group together consecutive frames of a video into the same cluster. With *Dataset 2c*, it was seen that despite a few clusters succeeding in having similar findings together, most clusters had a mixed number of different findings.

However, despite these downfalls, this project helped provide an important insight into the potential for using DL in the field of precancerous lesion detection in chronic IBD patients - both with a supervised learning based approach that uses object detection, and an unsupervised learning based approach for image clustering. This will act as the foundation for further research into this topic which has the potential for significantly improving the current systems in place for detecting precancerous lesions during surveillance colonoscopies of IBD patients.

Bibliography

[1] *2018 Turing Award*. Jan. 9, 2024. URL: https://awards.acm.org/about/2018-turing (visited on 01/09/2024).

[2] Hervé Abdi and Lynne J Williams. "Principal component analysis." In: *Wiley interdisciplinary reviews: computational statistics* 2.4 (2010), pp. 433–459.

[3] *About the functionality of DFL_loss · Issue 4219 · ultralytics/ultralytics*. Jan. 4, 2024. URL: https://github.com/ultralytics/ultralytics/issues/4219 (visited on 01/04/2024).

[4] Ahsan Adeel, Mandar Gogate, and Amir Hussain. "Contextual deep learning-based audio-visual switching for speech enhancement in real-world environments." In: *Information Fusion* 59 (2020), pp. 163–170.

[5] Fahim Irfan Alam et al. "Conditional random field and deep feature learning for hyperspectral image classification." In: *IEEE Transactions on Geoscience and Remote Sensing* 57.3 (2018), pp. 1612–1628.

[6] Elie Aljalbout et al. "Clustering with deep learning: Taxonomy and new methods." In: *arXiv preprint arXiv:1801.07648* (2018).

[7] Laith Alzubaidi et al. "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions." In: *Journal of big Data* 8 (2021), pp. 1–74.

[8] Vito Annese et al. "European evidence based consensus for endoscopy in inflammatory bowel disease." In: *Journal of Crohn's and Colitis* 7.12 (2013), pp. 982–1018.

[9] Diego Ardila et al. "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography." In: *Nature medicine* 25.6 (2019), pp. 954–961.

[10] *Are class and box losses calculated the same in YoloV8 and YoloV5? · Issue 2789 · ultralytics/ultralytics*. Jan. 3, 2024. URL: https://github.com/ultralytics/ultralytics/issues/2789 (visited on 01/03/2024).

[11] A Asuntha and Andy Srinivasan. "Deep learning for lung Cancer detection and classification." In: *Multimedia Tools and Applications* 79 (2020), pp. 7731–7762.

[12] *Binary Cross Entropy: Where To Use Log Loss In Model Monitoring - Arize AI*. Jan. 3, 2024. URL: https://arize.com/blog-course/binary-cross-entropy-log-loss/ (visited on 01/03/2024).

[13] Chris M Bishop. “Neural networks and their applications.” In: *Review of scientific instruments* 65.6 (1994), pp. 1803–1832.

[14] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. “Yolov4: Optimal speed and accuracy of object detection.” In: *arXiv preprint arXiv:2004.10934* (2020).

[15] Gaudenz Boesch. *VGG Very Deep Convolutional Networks (VGGNet) - What you need to know - viso.ai*. Oct. 7, 2021. URL: https://viso.ai/deep-learning/vgg-very-deep-convolutional-networks/ (visited on 12/09/2023).

[16] Aurélien Bour et al. “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.

[17] Alison T Brenner, Michael Dougherty, and Daniel S Reuland. “Colorectal cancer screening in average risk patients.” In: *Medical Clinics* 101.4 (2017), pp. 755–767.

[18] Hermann Brenner et al. “Risk of progression of advanced adenomas to colorectal cancer by age and sex: estimates based on 840 149 screening colonoscopies.” In: *Gut* 56.11 (2007), pp. 1585–1589.

[19] *Brief summary of YOLOv8 model structure · Issue 189 · ultralytics/ultralytics*. Jan. 2, 2024. URL: https://github.com/ultralytics/ultralytics/issues/189 (visited on 01/02/2024).

[20] Jason Brownlee. *How to choose loss functions when Training Deep Learning Neural Networks*. Aug. 2020. URL: https://machinelearningmastery.com/how-to-choose-loss-functions-when-training-deep-learning-neural-networks/.

[21] Anna M Buchner. “The role of chromoendoscopy in evaluating colorectal dysplasia.” In: *Gastroenterology & Hepatology* 13.6 (2017), p. 336.

[22] Mathilde Caron et al. “Deep Clustering for Unsupervised Learning of Visual Features.” In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Sept. 2018.

[23] Mathilde Caron et al. “Deep clustering for unsupervised learning of visual features.” In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 132–149.

[24] Antoni Castells. “Choosing the optimal method in programmatic colorectal cancer screening: current evidence and controversies.” In: *Therapeutic advances in gastroenterology* 8.4 (2015), pp. 221–233.

[25] Hongming Chen et al. “The rise of deep learning in drug discovery.” In: *Drug discovery today* 23.6 (2018), pp. 1241–1250.

[26] *Colorectal Cancer Awareness Month 2022 – IARC*. Feb. 6, 2023. URL: https://www.iarc.who.int/featured-news/colorectal-cancer-awareness-month-2022/ (visited on 02/06/2023).

[27] *Colorectal cancer statistics | WCRF International*. Feb. 6, 2023. URL: https://www.wcrf.org/cancer-trends/colorectal-cancer-statistics/ (visited on 02/06/2023).

[28] Philomena M Colucci, Steven H Yale, and Christopher J Rall. "Colorectal polyps." In: *Clinical medicine & research* 1.3 (2003), pp. 261–262.

[29] Vincenza Conteduca et al. "Precancerous colorectal lesions." In: *International journal of oncology* 43.4 (2013), pp. 973–984.

[30] Solemane Coulibaly et al. "Deep neural networks with transfer learning in millet crop images." In: *Computers in industry* 108 (2019), pp. 115–120.

[31] Angel Alfonso Cruz-Roa et al. "A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection." In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II 16*. Springer. 2013, pp. 403–410.

[32] *CSP-DarkNet*. Dec. 12, 2023. URL: https://huggingface.co/docs/timm/models/csp-darknet (visited on 12/12/2023).

[33] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. "Supervised learning." In: *Machine learning techniques for multimedia: case studies on organization and retrieval* (2008), pp. 21–49.

[34] J. Deng et al. "ImageNet: A Large-Scale Hierarchical Image Database." In: *CVPR09*. 2009.

[35] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." In: *arXiv preprint arXiv:1810.04805* (2018).

[36] Mehwish Dildar et al. "Skin cancer detection: a review using deep learning techniques." In: *International journal of environmental research and public health* 18.10 (2021), p. 5479.

[37] Geoff Easton. "Critical realism in case study research." In: *Industrial marketing management* 39.1 (2010), pp. 118–128.

[38] *Endoscopy Campus - Polyp Classification: NICE*. Mar. 30, 2023. URL: https://www.endoscopy-campus.com/en/classifications/polyp-classification-nice/ (visited on 03/30/2023).

[39] Rasool Fakoor et al. "Using deep learning to enhance cancer diagnosis and classification." In: *Proceedings of the international conference on machine learning*. Vol. 28. ACM New York, NY, USA. 2013, pp. 3937–3949.

[40] Claudio Fiocchi. "Inflammatory bowel disease: etiology and pathogenesis." In: *Gastroenterology* 115.1 (1998), pp. 182–205.

[41] Gabe Flomo. *How to cluster images based on visual similarity | by Gabe Flomo | Towards Data Science*. Oct. 13, 2020. URL: https://towardsdatascience.com/how-to-cluster-images-based-on-visual-similarity-cd6e7209fe34 (visited on 10/07/2023).

[42] Alexander L. Fradkov. “Early History of Machine Learning.” In: *IFAC-PapersOnLine* 53.2 (2020). 21st IFAC World Congress, pp. 1385–1390. ISSN: 2405-8963. DOI: https://doi.org/10.1016/j.ifacol.2020.12.1888. URL: https://www.sciencedirect.com/science/article/pii/S2405896320325027.

[43] Ross Girshick. “Fast r-cnn.” In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448.

[44] *GitHub - ultralytics/yolov5: YOLOv5 in PyTorch > ONNX > CoreML > TFLite*. Oct. 5, 2023. URL: https://github.com/ultralytics/yolov5 (visited on 10/05/2023).

[45] *glenn-jocher (Glenn Jocher)*. Dec. 6, 2001. URL: https://github.com/glenn-jocher (visited on 12/12/2023).

[46] Ian Goodfellow et al. “Generative adversarial networks.” In: *Communications of the ACM* 63.11 (2020), pp. 139–144.

[47] Antonio Gulli and Sujit Pal. *Deep learning with Keras*. Packt Publishing Ltd, 2017.

[48] Rachit Kumar Gupta, Jatinder Manhas, et al. “Improved classification of cancerous histopathology images using color channel separation and deep learning.” In: *Journal of Multimedia Information System* 8.3 (2021), pp. 175–182.

[49] Anne Håkansson. “Portal of research methods and methodologies for research projects and degree projects.” In: *The 2013 World Congress in Computer Science, Computer Engineering, and Applied Computing WORLDCOMP 2013; Las Vegas, Nevada, USA, 22-25 July*. CSREA Press USA. 2013, pp. 67–73.

[50] Asmaa Halbouni et al. “Machine learning and deep learning approaches for cybersecurity: A review.” In: *IEEE Access* 10 (2022), pp. 19572–19585.

[51] Karin Hammarberg, Maggie Kirkman, and Sheryl de Lacey. “Qualitative research methods: when to use them and how to judge them.” In: *Human reproduction* 31.3 (2016), pp. 498–501.

[52] John A Hartigan and Manchek A Wong. “Algorithm AS 136: A k-means clustering algorithm.” In: *Journal of the royal statistical society. series c (applied statistics)* 28.1 (1979), pp. 100–108.

[53] Kaiming He et al. “Deep residual learning for image recognition.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[54] Roberta Heale and Alison Twycross. “What is a case study?” In: *Evidence-Based Nursing* 21.1 (2018), pp. 7–8. ISSN: 1367-6539. DOI: 10.1136/eb-2017-102845. eprint: https://ebn.bmj.com/content/21/1/7.full.pdf. URL: https://ebn.bmj.com/content/21/1/7.

[55] Dan Hendrycks and Kevin Gimpel. “Gaussian error linear units (gelus).” In: *arXiv preprint arXiv:1606.08415* (2016).

[56] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. “A fast learning algorithm for deep belief nets.” In: *Neural computation* 18.7 (2006), pp. 1527–1554.

[57] Yaoshiang Ho and Samuel Wookey. “The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling.” In: *IEEE access* 8 (2019), pp. 4806–4813.

[58] *Home - Ultralytics YOLOv8 Docs*. May 2, 2023. URL: https://docs.ultralytics.com/ (visited on 05/02/2023).

[59] *How is confidence calculated by YOLOv8? · Issue 4149 · ultralytics/ultralytics*. Dec. 13, 2023. URL: https://github.com/ultralytics/ultralytics/issues/4149 (visited on 12/13/2023).

[60] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift.” In: *International conference on machine learning*. pmlr. 2015, pp. 448–456.

[61] *Irritable bowel syndrome (IBS) | NHS inform*. May 2, 2023. URL: https://www.nhsinform.scot/illnesses-and-conditions/stomach-liver-and-gastrointestinal-tract/irritable-bowel-syndrome-ibs (visited on 05/02/2023).

[62] Humayun Irshad, Ludovic Roux, and Daniel Racoceanu. “Multi-channels statistical and morphological features based mitosis detection in breast cancer histopathology.” In: *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2013, pp. 6091–6094.

[63] A G Ivakhnenko and V G Lapa. *Cybernetics and forecasting techniques*. Modern analytic and computational methods in science and mathematics. Trans. from the Russian, Kiev, Naukova Dumka, 1965. New York, NY: North-Holland, 1967. URL: https://cds.cern.ch/record/209675.

[64] Anil K Jain, Jianchang Mao, and K Moidin Mohiuddin. “Artificial neural networks: A tutorial.” In: *Computer* 29.3 (1996), pp. 31–44.

[65] Debesh Jha et al. “A comprehensive study on colorectal polyp segmentation with ResUNet++, conditional random field and test-time augmentation.” In: *IEEE journal of biomedical and health informatics* 25.6 (2021), pp. 2029–2040.

[66] Debesh Jha 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.

[67] Debesh Jha et al. “Real-time polyp detection, localization and segmentation in colonoscopy using deep learning.” In: *Ieee Access* 9 (2021), pp. 40496–40510.

[68] Debesh Jha et al. “Resunet++: An advanced architecture for medical image segmentation.” In: *2019 IEEE International Symposium on Multimedia (ISM)*. IEEE. 2019, pp. 225–2255.

[69] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. *YOLO by Ultralytics*. Version 8.0.0. https://orcid.org/0000-0001-5950-6979 (Glenn Jocher), https://orcid.org/0000-0002-7603-6750 (Ayush Chaurasia), https://orcid.org/0000-0003-3783-7069 (Jing Qiu). Jan. 10, 2023. URL: https://github.com/ultralytics/ultralytics.

[70] *K means Clustering - Introduction - GeeksforGeeks*. May 2, 2017. URL: https://www.geeksforgeeks.org/k-means-clustering-introduction/ (visited on 01/03/2024).

[71] Md Rezaul Karim et al. "Deep learning-based clustering approaches for bioinformatics." In: *Briefings in bioinformatics* 22.1 (2021), pp. 393–415.

[72] Ralf Kiesslich et al. "Methylene blue-aided chromoendoscopy for the detection of intraepithelial neoplasia and colon cancer in ulcerative colitis." In: *Gastroenterology* 124.4 (2003), pp. 880–888.

[73] Yoon Ji Kim et al. "RGB Channel Superposition Algorithm with Acetowhite Mask Images in a Cervical Cancer Classification Deep Learning Model." In: *Sensors* 22.9 (2022), p. 3564.

[74] Georgia Koppe, Andreas Meyer-Lindenberg, and Daniel Durstewitz. "Deep learning for small and big data in psychiatry." In: *Neuropsychopharmacology* 46.1 (2021), pp. 176–190.

[75] Anders Krogh. "What are artificial neural networks?" In: *Nature biotechnology* 26.2 (2008), pp. 195–197.

[76] K Kuznetsov, R Lambert, and J-F Rey. "Narrow-band imaging: potential and limitations." In: *Endoscopy* 38.01 (2006), pp. 76–81.

[77] Roberto Labianca and Barbara Merelli. "Screening and diagnosis for colorectal cancer: present and future." In: *Tumori Journal* 96.6 (2010), pp. 889–901.

[78] Lily L Lai et al. "Separation of color channels from conventional colonoscopy images improves deep neural network detection of polyps." In: *Journal of Biomedical Optics* 26.1 (2021), pp. 015001–015001.

[79] Peter Laszlo Lakatos and Laszlo Lakatos. "Risk for colorectal cancer in ulcerative colitis: changes, causes and management strategies." In: *World journal of gastroenterology: WJG* 14.25 (2008), p. 3937.

[80] R Lambert. "The Paris endoscopic classification of superficial neoplastic lesions: esophagus, stomach, and colon: November 30 to December 1, 2002." In: *Gastrointest Endosc* 58 (2003), S3–S43.

[81] Y. Lecun et al. "Gradient-based learning applied to document recognition." In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324. DOI: 10.1109/5.726791.

[82] Jeonghun Lee et al. "Risk factors of missed colorectal lesions after colonoscopy." In: *Medicine* 96.27 (2017).

[83] Ji Young Lee et al. "Real-time detection of colon polyps during colonoscopy using deep learning: systematic validation with four independent datasets." In: *Scientific reports* 10.1 (2020), p. 8379.

[84] Sid Lee. *Precancerous conditions of the colon or rectum | Canadian Cancer Society*. May 31, 2023. URL: https://cancer.ca/en/cancer-information/cancer-types/colorectal/what-is-colorectal-cancer/precancerous-conditions (visited on 05/31/2023).

[85] Barbara Leggett and Vicki Whitehall. "Role of the serrated pathway in colorectal cancer pathogenesis." In: *Gastroenterology* 138.6 (2010), pp. 2088–2100.

[86] Joel S Levine and Dennis J Ahnen. "Adenomatous polyps of the colon." In: *New England Journal of Medicine* 355.24 (2006), pp. 2551–2557.

[87] Kaidong Li et al. "Colonoscopy polyp detection and classification: Dataset creation and comparative evaluations." In: *Plos one* 16.8 (2021), e0255809.

[88] Xiang Li et al. "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection." In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 21002–21012.

[89] Yi Li et al. "Medical image fusion method by deep learning." In: *International Journal of Cognitive Computing in Engineering* 2 (2021), pp. 21–29.

[90] Ming Liu, Jue Jiang, and Zenan Wang. "Colonic polyp detection in endoscopic videos with single shot detection based deep convolutional neural network." In: *IEEE Access* 7 (2019), pp. 75058–75066.

[91] Sebastien Jean Mambou et al. "Breast cancer detection using infrared thermal imaging and a deep learning model." In: *Sensors* 18.9 (2018), p. 2799.

[92] Michael A Manfredi et al. "Electronic chromoendoscopy." In: *Gastrointestinal endoscopy* 81.2 (2015), pp. 249–261.

[93] Olle Mannheimer. *Artificial Intelligence for improved detection of dysplastic lesions in inflammatory bowel disease*. Degree Project. Gothenburg, Sweden, 2021.

[94] Mohammed A Al-Masni et al. "Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system." In: *Computer methods and programs in biomedicine* 157 (2018), pp. 85–94.

[95] Martino Mezzapesa et al. "Serrated colorectal lesions: an up-to-date review from histological pattern to molecular pathogenesis." In: *International Journal of Molecular Sciences* 23.8 (2022), p. 4461.

[96] Erxue Min et al. "A survey of clustering with deep learning: From the perspective of network architecture." In: *IEEE Access* 6 (2018), pp. 39501–39514.

[97] P1 Munkholm. "The incidence and prevalence of colorectal cancer in inflammatory bowel disease." In: *Alimentary pharmacology & therapeutics* 18 (2003), pp. 1–5.

[98] KSV Muralidhar. *Learning Curve to identify Overfitting and Underfitting in Machine Learning | by KSV Muralidhar | Towards Data Science*. July 7, 2023. URL: https://towardsdatascience.com/learning-curve-to-

identify-overfitting-underfitting-problems-133177f38df5 (visited on 01/05/2024).

[99] Sameer A Nene, Shree K Nayar, Hiroshi Murase, et al. "Columbia object image library (coil-20)." In: (1996).

[100] William S Noble. "What is a support vector machine?" In: *Nature biotechnology* 24.12 (2006), pp. 1565–1567.

[101] Keiron O'Shea and Ryan Nash. "An introduction to convolutional neural networks." In: *arXiv preprint arXiv:1511.08458* (2015).

[102] Niall O'Mahony et al. "Deep learning vs. traditional computer vision." In: *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1*. Springer. 2020, pp. 128–144.

[103] *Overfitting vs Underfitting in Machine Learning [Differences]*. Jan. 5, 2024. URL: https://www.v7labs.com/blog/overfitting-vs-underfitting (visited on 01/05/2024).

[104] Ishak Pacal and Dervis Karaboga. "A robust real-time deep learning based automatic polyp detection system." In: *Computers in Biology and Medicine* 134 (2021), p. 104519.

[105] Bo Pang, Erik Nijkamp, and Ying Nian Wu. "Deep learning with tensorflow: A review." In: *Journal of Educational and Behavioral Statistics* 45.2 (2020), pp. 227–248.

[106] Participants in the Paris Workshop. "The Paris endoscopic classification of superficial neoplastic lesions: esophagus, stomach, and colon: November 30 to December 1, 2002." In: *Gastrointestinal Endoscopy* 58.6, Supplement (2003), S3–S43. ISSN: 0016-5107. DOI: https://doi.org/10.1016/S0016-5107(03)02159-X. URL: https://www.sciencedirect.com/science/article/pii/S001651070302159X.

[107] Christophe Pere. *What are loss functions?* June 2020. URL: https://towardsdatascience.com/what-is-loss-function-1e2605aeb904.

[108] Joanna Picot et al. "Virtual chromoendoscopy for the real-time assessment of colorectal polyps in vivo: a systematic review and economic evaluation." In: *Health Technology Assessment* 21.79 (2017).

[109] Anand B Pithadia and Sunita Jain. "Treatment of inflammatory bowel disease (IBD)." In: *Pharmacological Reports* 63.3 (2011), pp. 629–642.

[110] *Principal Component Analysis(PCA) - GeeksforGeeks*. July 7, 2018. URL: https://www.geeksforgeeks.org/principal-component-analysis-pca/ (visited on 01/03/2024).

[111] *RangeKing (Range King)*. Dec. 8, 2023. URL: https://github.com/RangeKing (visited on 12/08/2023).

[112] Jillella Sai Charan Reddy et al. "Real time Automatic Polyp Detection in White light Endoscopy videos using a combination of YOLO and DeepSORT." In: *2022 1st International Conference on the Paradigm Shifts in Communication, Embedded Systems, Machine Learning and Signal Processing (PCEMS)*. IEEE. 2022, pp. 104–106.

[113] Joseph Redmon et al. "You Only Look Once: Unified, Real-Time Object Detection." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016.

[114] Joseph Redmon et al. "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.

[115] Florian Rieder and Claudio Fiocchi. "Intestinal fibrosis in IBD—a dynamic, multifactorial process." In: *Nature reviews Gastroenterology & hepatology* 6.4 (2009), pp. 228–235.

[116] MD Rutter et al. "Pancolonic indigo carmine dye spraying for the detection of dysplasia in ulcerative colitis." In: *Gut* 53.2 (2004), pp. 256–260.

[117] Sameer D Saini, Hyungjin Myra Kim, and Philip Schoenfeld. "Incidence of advanced adenomas at surveillance colonoscopy in patients with a personal history of colon adenomas: a meta-analysis and systematic review." In: *Gastrointestinal endoscopy* 64.4 (2006), pp. 614–626.

[118] Fayez Sandouk, Feras Al Jerf, and MHD Al-Halabi. "Precancerous lesions in colorectal cancer." In: *Gastroenterology Research and Practice* 2013 (2013).

[119] Claudio Filipi Gonçalves Dos Santos and João Paulo Papa. "Avoiding overfitting: A survey on regularization methods for convolutional neural networks." In: *ACM Computing Surveys (CSUR)* 54.10s (2022), pp. 1–25.

[120] Iqbal H Sarker. "Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions." In: *SN Computer Science* 2.6 (2021), p. 420.

[121] Pooyan Sedigh, Rasoul Sadeghian, and Mehdi Tale Masouleh. "Generating synthetic medical images by using GAN to improve CNN performance in skin cancer classification." In: *2019 7th International Conference on Robotics and Mechatronics (ICRoM)*. IEEE. 2019, pp. 497–502.

[122] *sergiuwaxmann (Sergiu Waxmann)*. Dec. 11, 2001. URL: `https://github.com/sergiuwaxmann` (visited on 12/12/2023).

[123] Sagar Sharma, Simone Sharma, and Anidhya Athaiya. "Activation functions in neural networks." In: *Towards Data Sci* 6.12 (2017), pp. 310–316.

[124] Li Shen et al. "Deep learning to improve breast cancer detection on screening mammography." In: *Scientific reports* 9.1 (2019), p. 12495.

[125] Connor Shorten and Taghi M Khoshgoftaar. "A survey on image data augmentation for deep learning." In: *Journal of big data* 6.1 (2019), pp. 1–48.

[126] Juan Silva et al. "Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer." In: *International journal of computer assisted radiology and surgery* 9 (2014), pp. 283–293.

[127] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." In: *arXiv preprint arXiv:1409.1556* (2014).

[128] Elizabeth A Sturgiss and Alexander M Clark. "Using critical realism in primary care research: an overview of methods." In: *Family Practice* 37.1 (Dec. 2019), pp. 143–145. ISSN: 1460-2229. DOI: 10.1093/fampra/cmz084. eprint: https://academic.oup.com/fampra/article-pdf/37/1/143/32525913/cmz084.pdf. URL: https://doi.org/10.1093/fampra/cmz084.

[129] Venkataraman Subramanian et al. "Meta-analysis: the diagnostic yield of chromoendoscopy for detecting dysplasia in patients with colonic inflammatory bowel disease." In: *Alimentary pharmacology & therapeutics* 33.3 (2011), pp. 304–312.

[130] Christian Szegedy et al. "Going deeper with convolutions." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[131] Christian Szegedy et al. "Rethinking the inception architecture for computer vision." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826.

[132] Bilal Taha, Naoufel Werghi, and Jorge Dias. "Automatic polyp detection in endoscopy videos: A survey." In: *2017 13th IASTED International Conference on Biomedical Engineering (BioMed)*. 2017, pp. 233–240. DOI: 10.2316/P.2017.852-031.

[133] Nima Tajbakhsh, Suryakanth R. Gurudu, and Jianming Liang. "Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information." In: *IEEE Transactions on Medical Imaging* 35.2 (2016), pp. 630–644. DOI: 10.1109/TMI.2015.2487997.

[134] Nicholas J Talley and Robin Spiller. "Irritable bowel syndrome: a little understood organic bowel disease?" In: *The Lancet* 360.9332 (2002), pp. 555–564.

[135] Chuanqi Tan et al. "A survey on deep transfer learning." In: *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*. Springer. 2018, pp. 270–279.

[136] Mingxing Tan and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." In: *International conference on machine learning*. PMLR. 2019, pp. 6105–6114.

[137] Shinji Tanaka and Yasushi Sano. "Aim to unify the narrow band imaging (NBI) magnifying classification for colorectal tumors: current status in Japan from a summary of the consensus symposium in the 79th Annual Meeting of the Japan Gastroenterological Endoscopy Society." In: *Digestive Endoscopy* 23 (2011), pp. 131–139.

[138] Keras Team. *VGG16 and VGG19*. Dec. 6, 2023. URL: https://keras.io/api/applications/vgg/ (visited on 12/07/2023).

[139] Juan Terven and Diana Cordova-Esparza. “A comprehensive review of YOLO: From YOLOv1 to YOLOv8 and beyond.” In: *arXiv preprint arXiv:2304.00501* (2023).

[140] Haiman Tian, Shu-Ching Chen, and Mei-Ling Shyu. “Evolutionary programming based deep learning feature selection and network construction for visual data classification.” In: *Information systems frontiers* 22 (2020), pp. 1053–1066.

[141] Suramya Tomar. “Converting video formats with FFmpeg.” In: *Linux journal* 2006.146 (2006), p. 10.

[142] Dai Quoc Tran et al. “Damage-map estimation using UAV images and deep learning algorithms for disaster management system.” In: *Remote Sensing* 12.24 (2020), p. 4169.

[143] *Transfer Learning - AI Wiki*. Jan. 9, 2024. URL: `https://machine-learning.paperspace.com/wiki/transfer-learning` (visited on 01/09/2024).

[144] John K Triantafillidis, Georgios Nasioulas, and Paris A Kosmidis. “Colorectal cancer and inflammatory bowel disease: epidemiology, risk factors, mechanisms of carcinogenesis and prevention strategies.” In: *Anticancer research* 29.7 (2009), pp. 2727–2737.

[145] Alexander Meining Ulm and Thomas Rösch. “Paris Classification Early Cancer.” In: ().

[146] Ultralytics. *Architecture Summary - Ultralytics YOLOv8 Docs*. Dec. 12, 2023. URL: `https://docs.ultralytics.com/yolov5/tutorials/architecture_description/#conclusion` (visited on 12/12/2023).

[147] Ultralytics. *YOLO Performance Metrics - Ultralytics YOLOv8 Docs*. Dec. 12, 2023. URL: `https://docs.ultralytics.com/guides/yolo-performance-metrics/#connect-and-collaborate` (visited on 12/13/2023).

[148] Alfred Ultsch. “Self-organizing neural networks for visualisation and classification.” In: *Information and Classification: Concepts, Methods and Applications Proceedings of the 16th Annual Conference of the “Gesellschaft für Klassifikation eV” University of Dortmund, April 1–3, 1992*. Springer. 1993, pp. 307–313.

[149] *Understanding Neurons in Deep Learning | Nick McCullum*. May 31, 2023. URL: `https://www.nickmccullum.com/python-deep-learning/understanding-neurons-deep-learning/` (visited on 05/31/2023).

[150] Jesper E Van Engelen and Holger H Hoos. “A survey on semi-supervised learning.” In: *Machine learning* 109.2 (2020), pp. 373–440.

[151] Shyam Varadarajulu et al. “GI endoscopes.” In: *Gastrointestinal endoscopy* 74.1 (2011), pp. 1–6.

[152] Paul Voigt and Axel Von dem Bussche. “The eu general data protection regulation (gdpr).” In: *A Practical Guide, 1st Ed., Cham: Springer International Publishing* 10.3152676 (2017), pp. 10–5555.

[153] Chien-Yao Wang et al. “CSPNet: A new backbone that can enhance learning capability of CNN.” In: *Proceedings of the IEEE/CVF conference*

on computer vision and pattern recognition workshops. 2020, pp. 390–391.

[154] DeLiang Wang and Jitong Chen. “Supervised speech separation based on deep learning: An overview.” In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.10 (2018), pp. 1702–1726.

[155] Fei Wang et al. “Residual Attention Network for Image Classification.” In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 2017.

[156] Jan Wehkamp et al. “Inflammatory bowel disease: Crohn’s disease and ulcerative colitis.” In: *Deutsches Ärzteblatt International* 113.5 (2016), p. 72.

[157] *What Are the Risk Factors for Colorectal Cancer? | CDC*. Feb. 6, 2023. URL: `https://www.cdc.gov/cancer/colorectal/basic_info/risk_factors.htm` (visited on 02/06/2023).

[158] *What Is Colorectal Cancer? | CDC*. May 1, 2023. URL: `https://www.cdc.gov/cancer/colorectal/basic_info/what-is-colorectal-cancer.htm` (visited on 05/01/2023).

[159] *What is confidence threshold · Issue 9679 · ultralytics/yolov5*. Dec. 13, 2023. URL: `https://github.com/ultralytics/yolov5/issues/9679` (visited on 12/13/2023).

[160] *What is the YOLO Darknet TXT Annotation Format?* Oct. 2, 2023. URL: `https://roboflow.com/formats/yolo-darknet-txt` (visited on 10/02/2023).

[161] Wikipedia contributors. *Deep learning — Wikipedia, The Free Encyclopedia*. [Online; accessed 9-January-2024]. 2024. URL: `https://en.wikipedia.org/w/index.php?title=Deep_learning&oldid=1193884121`.

[162] Yang Xin et al. “Machine learning and deep learning methods for cybersecurity.” In: *Ieee access* 6 (2018), pp. 35365–35381.

[163] Jianwei Yang, Devi Parikh, and Dhruv Batra. “Joint Unsupervised Learning of Deep Representations and Image Clusters.” In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016.

[164] Tom Young et al. “Recent trends in deep learning based natural language processing.” In: *ieee Computational intelligenCe magazine* 13.3 (2018), pp. 55–75.

[165] Lequan Yu et al. “Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos.” In: *IEEE journal of biomedical and health informatics* 21.1 (2016), pp. 65–75.

[166] Shifeng Zhang et al. “Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection.” In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9759–9768.

[167] Shifeng Zhang et al. “Single-shot refinement neural network for object detection.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4203–4212.

[168] Zhong-Qiu Zhao et al. “Object detection with deep learning: A review.” In: *IEEE transactions on neural networks and learning systems* 30.11 (2019), pp. 3212–3232.

[169] Zhaohui Zheng et al. “Distance-IoU loss: Faster and better learning for bounding box regression.” In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 07. 2020, pp. 12993–13000.