UNIVERSITAT DE BARCELONA

End-to-End AI Solutions for Capsule Endoscopy: Enhancing Efficiency and Accuracy in Gastrointestinal Diagnostics

Pere Gilabert Roca

UNIVERSITAT DE BARCELONA

End-to-End AI Solutions for Capsule Endoscopy: Enhancing Efficiency and Accuracy in Gastrointestinal Diagnostics

Ph.D. THESIS

Pere Gilabert Roca

A thesis submitted in fulfillment of the requirements to the Ph.D. program in Mathematics and Computer Science

Facultat de Matemàtiques i Informàtica
Universitat de Barcelona

2024

Dr. Santi Seguí Mesquida
Facultat de Matemàtiques i Informàtica
Universitat de Barcelona

Director

Dr. Santi Seguí Mesquida

Associate Professor of the *Facultat de Matemàtiques i Informàtica* at the *Universitat de Barcelona*, Barcelona.

CERTIFICATE:

That **Pere Gilabert Roca** has completed his doctoral thesis titled **"End-to-End AI Solutions for Capsule Endoscopy: Enhancing Efficiency and Accuracy in Gastrointestinal Diagnostics"** under my expert guidance and supervision. The thesis fulfills all the necessary requirements and standards set forth by the *Universitat de Barcelona* for obtaining the Doctoral degree. It is now ready to be presented and defended before the relevant academic panel.

Barcelona, 2024

Dr. Santi Seguí Mesquida

SANTIAGO SEGUI MESQUIDA - DNI 41742934G

Firmado digitalmente por SANTIAGO SEGUI MESQUIDA - DNI 41742934G
Fecha: 2024.09.26 18:01:57 +02'00'

As with any great journey, the destination is only as meaningful as the companions who help you reach it.

Acknowledgements

Embarking on this thesis journey has felt like setting out on an ambitious and challenging hike, one that led me through steep climbs, dense forests, and breathtaking vistas. As I reach the summit, it is time to acknowledge those who have walked alongside me, providing guidance, support, and encouragement every step of the way.

First and foremost, I extend my deepest gratitude to Santi, who has been my trail guide throughout this expedition. Your wisdom and patience have illuminated the path ahead, helping me navigate even the most treacherous terrains. Without your insight and steady encouragement, I might have lost my way.

My fellow hikers, Jordi, Pablo, Paula, Mariona, Roger, Enrique, Álvaro, Guillem, Àxel and Àlex, thank you for sharing in the journey. Whether it was a word of encouragement during a steep climb or a moment of laughter by the campfire, your companionship made the journey not only bearable but also enjoyable.

To my collaborators beyond the university, Hagen, Carolina, Reece, Liz, thank you for your valuable contributions and support. Your insights and collaboration have been like finding a helpful guide on the trail, providing new perspectives and assisting me in navigating complex paths that I might not have been able to traverse alone.

To Anna, and our newborn child, Arià, you have been my unwavering companions on this hike. Anna, your love, patience, and belief in me have been the steadying hand I needed to keep moving forward, even when the path seemed impossible. And to Arià, your arrival has brought new joy and purpose to this journey, reminding me of the most important summit of all. I couldn't have made it to this peak without you both by my side.

To my parents and sister, Mireia, Jordi and Txell, thank you for being my constant support system. Your encouragement has been the foundation of my strength, providing me with the resilience to keep climbing. Whether it was a word of advice, a moment of laughter, or simply knowing you were there, your support has been like the safety ropes and shelter that kept me grounded and motivated throughout this journey.

Lastly, to all those who have supported me in ways both big and small, thank you for helping me reach this peak. This achievement is not mine alone; it belongs to all of us who walked this path together.

Abstract

Artificial Intelligence (AI) models are fundamentally transforming the way clinicians carry out their daily tasks. By streamlining various processes, AI offers a more robust and consistent method for reviewing medical procedures. This thesis is dedicated to the development of AI applications for Capsule Endoscopy (CE), a small device that patients swallow, which is equipped with both a light and a camera to traverse the digestive system, capturing detailed images of internal organs.

Once these images are captured, physicians are tasked with meticulously reviewing an extensive number of frames to identify potential pathologies, a process that is both time-consuming and tedious.

In this thesis, we aim to enhance the entire review pipeline from end to end, providing support to physicians at multiple stages of the process. These stages include data collection, data labeling, assessing the usability of the videos (particularly in determining whether intestinal residues may hinder the process), identifying the entry and exit points of the small and large intestines, and most crucially, detecting polyps as early indicators of Colorectal Cancer (CRC).

By employing advanced techniques such as Active Learning (AL) for data labeling and Vision Transformer (ViT) for polyp detection, we significantly improve upon existing systems in the literature, achieving state-of-the-art results.

Additionally, the integration of AI into CE holds the promise of not only improving diagnostic accuracy but also reducing the workload for clinicians, allowing them to focus on more complex cases. This technological advancement has the potential to revolutionize gastrointestinal diagnostics, leading to earlier detection of diseases and, ultimately, better patient outcomes.

Furthermore, this thesis led to the initiation of two clinical studies. The first was a controlled study that evaluated the performance of the polyp detection application. The second is a larger study involving over 600 patients, testing an enhanced version of the application, which is currently under development.

Resum

Els models d'Intel·ligència Artificial (AI) estan transformant la forma com els professionals sanitaris duen a terme les seves tasques diàries. Optimitzant diversos processos, la AI ofereix un mètode més robust i consistent per revisar diferents procediments mèdics. Aquesta tesi està dedicada al desenvolupament d'aplicacions d'AI per a la Càpsula Endoscòpica (CE), un petit dispositiu que els pacients empassen, equipat amb llum i càmera, que recorre el sistema digestiu capturant imatges detallades dels òrgans interns.

Un cop aquestes imatges són capturades, els professionals han de revisar meticulosament un gran nombre de fotogrames per identificar possibles patologies, un procés que és tant laboriós com tediós.

En aquesta tesi, ens proposem millorar tot el procés de revisió d'aquests vídeos, de principi a fi, proporcionant eines de suport als metges en diverses etapes del procés. Aquestes etapes inclouen la recopilació de dades, l'etiquetatge de dades, l'avaluació de la usabilitat dels vídeos (particularment per determinar si els vídeos són prou nets per a ser usats), la identificació dels punts d'entrada i sortida dels intestins prim i gruixut, i el més important, la detecció de pòlips com a indicadors precoços del càncer colorectal (CRC).

Mitjançant tècniques avançades com l'Aprenentatge Actiu (AL) per a l'etiquetatge de dades i models complexos com el Transformer de Visió (ViT) per a la detecció de pòlips, millorem significativament els sistemes existents, assolint resultats de referència.

A més, la integració de la AI en la CE té el potencial no només de millorar la precisió diagnòstica, sinó també de reduir la càrrega de treball dels professionals, permetent-los centrar-se en casos més complexos. Aquest avenç tecnològic pot revolucionar els diagnòstics gastrointestinals, facilitant una detecció precoç de malalties i, en última instància, a millors resultats per als pacients.

Aquesta tesi ha portat, a més, a l'inici de dos estudis clínics. El primer va consistir en un estudi controlat que va avaluar el rendiment de l'aplicació de detecció de pòlips. El segon és un estudi més gran que involucra més de 600 pacients, provant una versió millorada de l'aplicació. Aquest estudi està actualment en desenvolupament.

Declaration

I declare that this thesis was composed by myself, that the work contained here is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Pere Gilabert Roca

PERE GILABERT ROCA - DNI 48094220R

Digitally signed by PERE GILABERT ROCA - DNI 48094220R
Date: 2024.09.26 17:26:38 +02'00'

Contents

List of Figures

List of Tables

Acronyms

ACC	Accuracy
AGEM CSU	NHS Arden & GEM Commissioning Support Unit
AI	Artificial Intelligence
AL	Active Learning
AUC	Area Under the ROC Curve
BCE	Binary Cross-Entropy
CAD-CAP	Computer-Assisted Diagnosis for Capsule Endoscopy
CatCE	Categorical Cross-Entropy
CCE	Colon Capsule Endoscopy
CE	Capsule Endoscopy
CESCAIL	Capsule Endoscopy delivery at Scale through enhanced AI anaLysis
CHI	CorporateHealth International UK Limited
CMT	Context-Motion-Temporal
CNN	Convolutional Neural Network
CRC	Colorectal Cancer
CT	Computed Tomography
DL	Deep Learning
ESGE	European Society of Gastrointestinal Endoscopy
FDA	Food and Drug Administration
FIT	Fecal Immunochemical Test
GI	Gastrointestinal
GIANA	Gastrointestinal Image ANAlysis
HSI	Hue-Saturation-Intensity

IBD	Inflammatory Bowel Disease
IBS	Inflammatory Bowel Syndrome
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
KID	Knowledge-based Intelligent Digestive
KL	Kullback-Leibler
LSTM	Long Short-Term Memory
mACC	Mean Accuracy
MAE	Mean Absolute Error
MCC	Matthews Correlation Coefficient
MedAE	Median Absolute Error
MIL	Multiple Instance Learning
ML	Machine Learning
MLP	Multi Layer Perceptron
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
NIHR	National Institute for Health and Care Research
NLP	Natural Language Processing
NN	Deep Neural Network
PCA	Principal Component Analysis
PEG	Polyethylene Glycol
SAM	Sharpness-Aware Minimization
SENS	Sensitivity
SGD	Stochastic Gradient Descent
SPEC	Specificity
TL	Triplet Loss
ULBP	Uniform Local Binary Patterns
UMAP	Uniform Manifold Approximation and Projection for Dimension Reduction
ViT	Vision Transformer

Chapter 1

Introduction

Contents

This chapter sets the stage for the entire thesis. It provides essential background information to aid in understanding the content of this document. We will discuss the process of reviewing endoscopic videos and introduce Capsule Endoscopy (CE) as an alternative to traditional colonoscopies. Additionally, the contributions of this thesis will be summarized. Finally, the structure of the document will be outlined.

1.1 Motivation

Artificial Intelligence (AI) is revolutionizing the way clinicians perform their daily tasks, offering unprecedented support and efficiency in the medical field. These advanced AI systems are increasingly being integrated into medical routines, functioning as invaluable support tools for healthcare professionals. The applications of AI extend far beyond diagnostics, encompassing a wide array of routine tasks such as parsing complex medical reports, translating medical texts, and highlighting potential errors or anomalies in various clinical processes. This integration of AI not only enhances accuracy and efficiency but also allows clinicians to focus more on patient care and less on administrative burdens.

This thesis addresses a specific clinical process: the review of CE videos. CE (Iddan et al., 2000) is a medical technology that enables visualization of the interior of the digestive system, facilitating the detection of pathologies like Colorectal Cancer (CRC), which remains one of the deadliest forms of cancer today (Bray et al., 2024; Siegel et al., 2024). By intervening in this review process, we propose a set of AI-driven tools designed to assist physicians, thereby improving the efficiency and effectiveness of the diagnostic process. These tools aim to streamline the workflow, reduce the time required for video analysis, and ultimately enhance patient outcomes by enabling earlier and more accurate detection of Gastrointestinal (GI) issues.

The following chapter serves as an introduction to the full thesis. It begins by presenting essential anatomical concepts related to the intestine, which are crucial for understanding the terms frequently used throughout this work, such as small intestine, large intestine, and polyps. By grounding the reader in these foundational concepts, we set the stage for a detailed exploration of how AI can transform CE reviews. Subsequent sections will delve into the technical aspects of the proposed AI tools, their development, and their integration into clinical practice.

Furthermore, we will explore the current challenges faced by clinicians in reviewing CE videos, including the sheer volume of data and the meticulous attention to detail required. We will then demonstrate how AI can alleviate these challenges, offering innovative solutions that not only speed up the review process but also enhance the accuracy of diagnoses. The ultimate goal is to show how these advancements can lead to better patient care, reduced workload for healthcare providers, and more efficient use of medical resources.

In summary, this thesis aims to illustrate the profound impact of AI on medical practices, particularly in the CE field.

1.2 Artificial Intelligence in Medical Imaging

AI has revolutionized medical imaging and diagnostics by analyzing complex data from various modalities, including Computed Tomography (CT), Magnetic Resonance Imaging

(MRI), X-rays, mammography, electrocardiograms, and video-derived images from CE, among other sources. These screening techniques generate vast amounts of data requiring meticulous analysis by trained professionals.

The sheer volume of images produced by these devices necessitates careful review by medical experts, which can be time-consuming and costly. AI can automate certain non-critical processes, such as routine image assessments or preliminary screenings, thereby reducing the workload on healthcare professionals. By acting as a support tool, AI helps streamline the diagnostic process, allowing clinicians to focus their expertise on more complex cases.

Furthermore, AI can assist in prioritizing patients based on the urgency of their conditions. By analyzing imaging data and patient information, AI systems can identify those who need immediate attention, thus optimizing resource allocation and improving patient outcomes. This prioritization is especially crucial in high-pressure environments where timely intervention can make a significant difference.

AI also holds promise as a pre-screening tool for diseases that may present without symptoms, such as CRC, one of the deadliest cancers nowadays. By analyzing patterns in imaging data and identifying potential indicators of early-stage disease, AI can help detect these conditions before they become symptomatic, facilitating earlier intervention and treatment.

In addition to its diagnostic applications, AI-powered tools can be deployed remotely, extending diagnostic capabilities to underserved regions with limited access to medical facilities. During emergencies, such as the COVID-19 pandemic, AI provides immediate medical interpretation, helping to manage and respond to crises more effectively.

Finally, AI has the potential to personalize medicine by tailoring treatments to the individual characteristics of each patient. By analyzing comprehensive data sets, including genetic information and response patterns to previous treatments, AI can assist in developing customized treatment plans that enhance efficacy and minimize adverse effects. This personalization aims to improve overall patient care and treatment outcomes.

Here are several key areas where AI is making significant advancements in medical imaging:

Radiology

In radiology, AI enhances image analysis and interpretation by detecting subtle patterns in X-rays, MRIs, CT scans, and ultrasound images that may be challenging for human observation (Rajpurkar et al., 2018; Khalid et al., 2020; Zhang et al., 2021). Its ability to identify early signs of conditions like cancer enables more accurate and timely diagnoses compared to traditional methods.

AI's automated segmentation capabilities further boost efficiency by providing detailed

image analysis, ensuring that critical findings are quickly identified and addressed by radiologists (Lenchik et al., 2019). Additionally, AI facilitates quantitative imaging, which allows for precise measurements of tumor size and growth rates, essential for monitoring disease progression and evaluating treatment effectiveness (Bera et al., 2022).

Moreover, AI optimizes workflow by prioritizing urgent cases, thereby improving radiologists' efficiency and enhancing overall patient care delivery. This advanced technology not only streamlines the diagnostic process but also contributes to better management of patient outcomes.

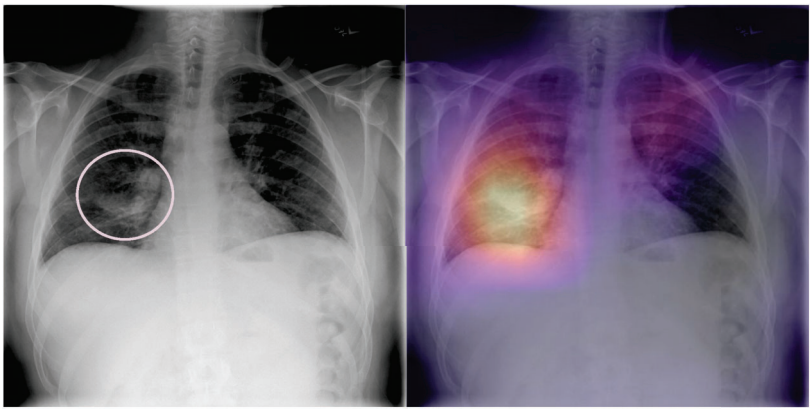


Figure 1.1: AI solution for detecting pulmonary tumors. Image from Rajpurkar et al. (2018).

Pathology

In pathology, AI is increasingly utilized to analyze digitized tissue samples, enhancing the identification of cancerous cells, tumor grading, and the detection of various pathological conditions (Bera et al., 2019; Hosseini et al., 2024). This application improves diagnostic accuracy and allows pathologists to focus on more complex cases, while ensuring consistent and reliable results.

AI also supports whole slide imaging by examining high-resolution images of complete tissue sections (Kumar et al., 2020; Chen et al., 2021a; Tan et al., 2023; Carcagnì et al., 2023). This capability enables AI to detect microscopic anomalies that might be missed by traditional methods, providing detailed insights essential for precise diagnostic assessments. Additionally, AI helps reduce inter-observer variability among pathologists, contributing to more consistent and accurate diagnoses (Hanna et al., 2019).

Cardiology

In cardiology, echocardiograms are the primary method for capturing cardiac images to evaluate cardiac anatomy and function during clinical routines. Echocardiography systems

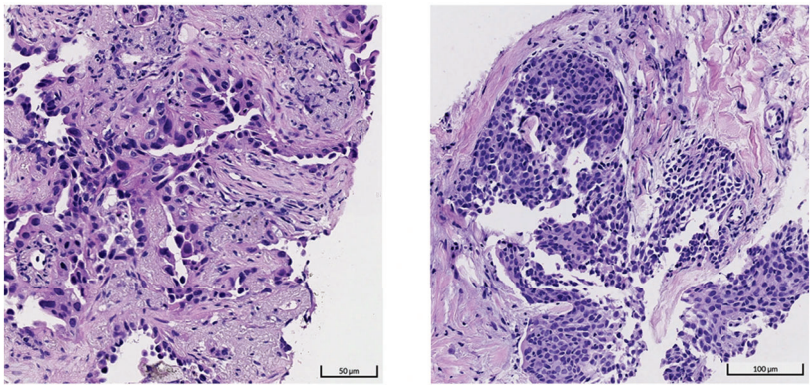


Figure 1.2: Two cropped sections from different whole slide images: one showing an adenocarcinoma (left) and the other depicting a squamous cell carcinoma (right). Image from Chen et al. (2021a).

that integrate AI solutions offer significant benefits for both cardiologists and patients. AI can assist in identifying conditions such as myocardial infarctions (Makimoto et al., 2020), arrhythmias (Petmezas et al., 2022), and aortic stenosis (Cohen-Shelly et al., 2021). It can also aid in measuring heart features, including the aortic valve (Thalappillil et al., 2020) and the overall heart structure.

Moreover, AI can automatically generate reports following a patient's visit (Lopez-Jimenez et al., 2020), helping cardiologists diagnose and manage various heart conditions more accurately and efficiently. This integration ensures comprehensive and timely patient care.

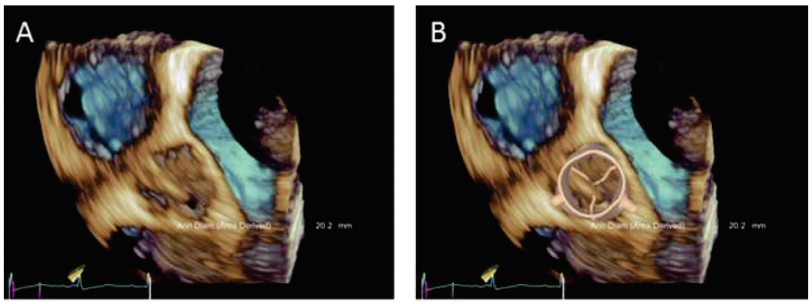


Figure 1.3: Aortic valve measurement. Image from Thalappillil et al. (2020).

Ophthalmology

AI excels at interpreting complex medical images, leading to early detection and accurate diagnosis of conditions such as diabetic retinopathy, glaucoma, and age-related macular degeneration. Studies have shown that AI models can identify diabetic retinopathy with greater accuracy than human experts (Vujosevic et al., 2020; Lakshminarayanan et al., 2021).

In addition to diagnostic applications, AI-powered tools are being developed to enhance surgical precision and efficiency. For instance, AI-assisted systems can offer real-time guidance during cataract surgery, improving outcomes and reducing complications (Gutierrez et al., 2022).

AI is also making strides in teleophthalmology by enabling remote screening and diagnosis of eye diseases in underserved areas. AI-powered platforms can analyze images taken with smartphone cameras, facilitating timely intervention and reducing the risk of vision loss (Al-Aswad et al., 2021). While the full potential of AI in ophthalmology is still unfolding, the field is experiencing rapid advancements with promising prospects for improving patient care and outcomes.

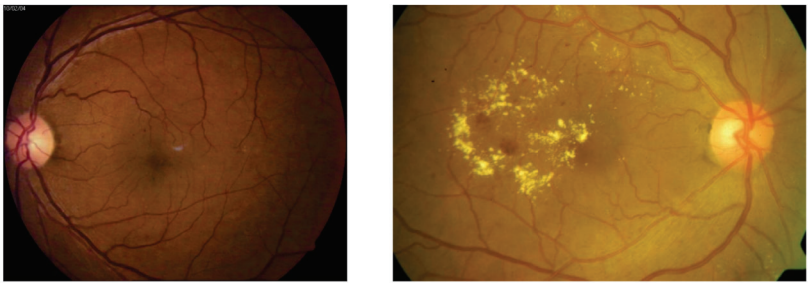


Figure 1.4: Diabetic retinopathy. Two different stages of the disease. Image from Lakshminarayanan et al. (2021).

Respiratory Medicine

AI is highly effective at analyzing chest X-rays and CT scans, facilitating the early detection of conditions such as lung cancer, pneumonia, and pulmonary embolism (Soffer et al., 2021). Research has shown that AI-powered systems can sometimes outperform human radiologists in specific diagnostic tasks (Wu et al., 2020). However, its primary role remains in assisting expert radiologists by improving their work and helping to avoid missing small pathologies.

The COVID-19 pandemic underscored AI's potential in respiratory medicine. AI-powered tools were quickly developed to assist with diagnosis, triage, and treatment planning for COVID-19 patients (Xu et al., 2021). For instance, AI algorithms were employed to analyze chest CT scans, identifying patterns associated with the disease and facilitating early detection and effective patient management (Abdulkareem and Petersen, 2021; Lasker et al., 2022).

Gastroenterology

In gastroenterology, AI has been successfully applied to detect liver fibrosis from serum markers (Hashem et al., 2018), pancreatic cancer from ultrasound images Ozkan et al. (2016), and colorectal polyps from colonoscopy procedures (Hassan et al., 2021), among

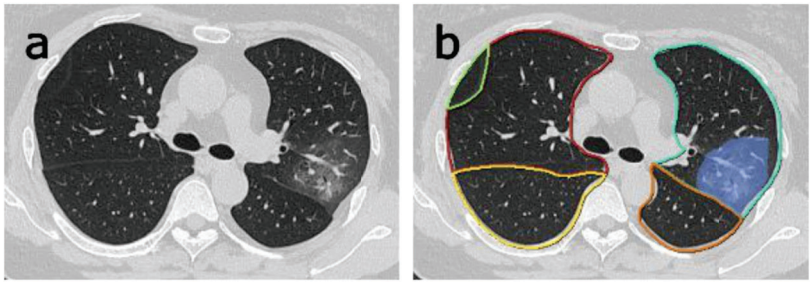


Figure 1.5: Occlusion area segmentation from CT image for COVID-19 detection. Image from Xu et al. (2021).

other applications.

In the specific field of CE, AI has been effectively utilized to detect lesions such as hemorrhages (Musha et al., 2023), angioectasias (Nennstiel et al., 2017), erosions and ulcers (Bang et al., 2021), Crohn's disease (Berre et al., 2019), and polyps (Laiz et al., 2020). Additionally, AI aids physicians in daily tasks such as capsule localization (Hanscom and Cave, 2022), assessing bowel preparation quality (Noorda et al., 2020), and improving the explainability of models, although this field is still in a very early stage (Varam et al., 2023).

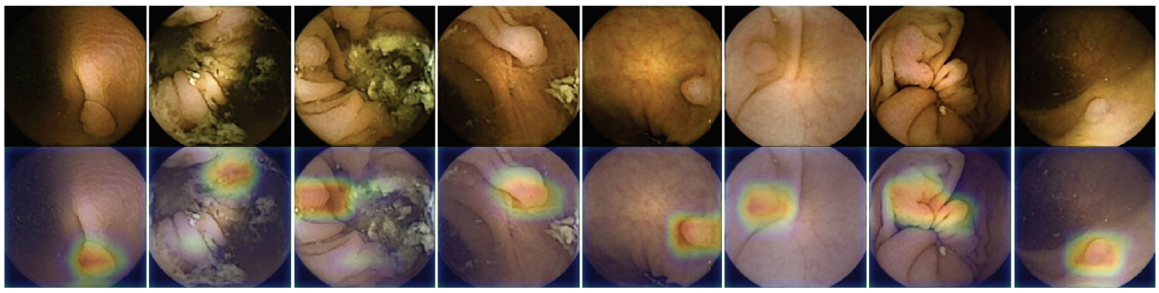


Figure 1.6: Polyp detection. Image from Laiz et al. (2020).

This thesis focuses on detecting pathologies in the digestive system, making it essential to understand some key concepts related to this area.

1.3 The Digestive System

The digestive system, also known as the GI system, is a complex network of organs and structures that work collaboratively to convert food into energy and essential nutrients. This intricate system begins at the mouth and extends through the esophagus, stomach, intestines, and ultimately ends at the anus. Alongside the main GI tract, accessory organs like the liver, pancreas, and gallbladder are essential in digestion.

At the heart of this system lie the small intestine and the large intestine, the main organs involved in digestion, nutrient absorption, and waste processing. Their sophisticated structures and functions are vital to maintaining overall health and will be the focal point of

this introduction. Figure 1.7 illustrates the main parts of the GI system, with a particular focus on the intestines.

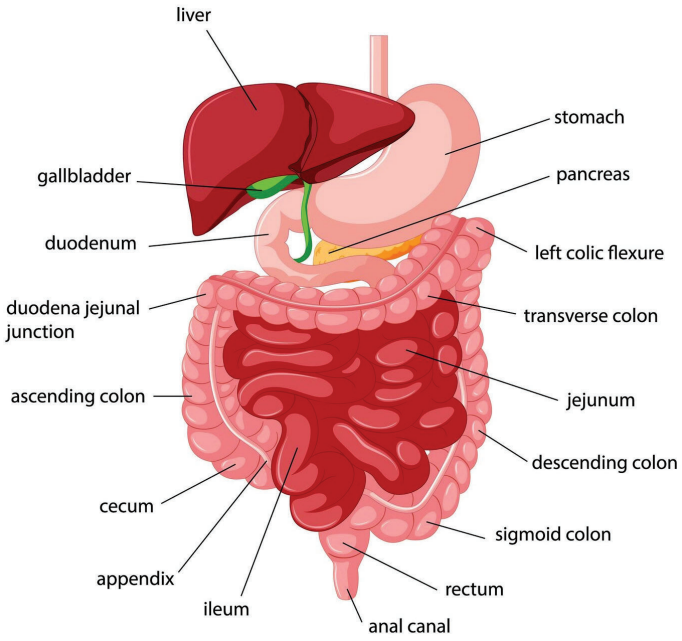


Figure 1.7: Main sections of the GI system, highlighting the intestines. Image from @brgfx (Freepik).

.

1.3.1 The Intestines

Small Intestine

The small intestine is the primary site for digestion and absorption of nutrients. It is a long, coiled tube approximately 6 meters in length and divided into three segments: the duodenum, the jejunum, and the ileum.

The breakdown of macronutrients into their smaller components (proteins into amino acids, carbohydrates into simple sugars, and fats into fatty acids and glycerol) is completed here. Pancreatic enzymes and bile work together to break down the fats ingested. Nutrients are absorbed through the epithelial lining of the small intestine into the bloodstream. The extensive surface area provided by villi and microvilli facilitates efficient nutrient uptake.

The small intestine can be divided into three segments:

- **Duodenum**: The initial segment, where chyme from the stomach is mixed with bile from the liver and digestive enzymes from the pancreas. This section is responsible for the emulsification of fats and the neutralization of stomach acids.

- **Jejunum**: The middle segment, where the majority of nutrient absorption occurs. It has a highly folded surface lined with villi and microvilli, which increase the surface area for absorption.
- **Ileum**: The final segment, responsible for absorbing remaining nutrients, particularly vitamin B12 and bile salts, and passing the residue into the large intestine.

Large Intestine

The large intestine, or colon, is about 1.5 meters long and is primarily involved in absorbing water and electrolytes from the remaining indigestible food matter, transforming it into solid waste (feces) to be excreted. It is divided into several regions: the cecum, ascending colon, transverse colon, descending colon, sigmoid colon, and rectum.

- **Cecum**: A pouch connected to the ileum that serves as a junction between the small and large intestines. The appendix is attached here.
- **Colon**: The majority of the large intestine, responsible for absorbing water and salts from the material that remains after nutrient absorption in the small intestine. It can be further divided into the ascending, transversal, descending and sigmoid colon.
- **Rectum**: The final section, where feces are stored before being expelled through the anus.

1.3.2 Pathophysiology of the Intestines

Given their central role in digestion and absorption, the intestines are susceptible to a range of disorders and diseases that can significantly impact health:

- **Inflammatory Bowel Disease (IBD)**: Includes Crohn's disease and ulcerative colitis, characterized by chronic inflammation of the GI tract, leading to symptoms like abdominal pain, diarrhea, and weight loss.
- **Inflammatory Bowel Syndrome (IBS)**: A functional disorder causing symptoms such as bloating, abdominal pain, and altered bowel habits without detectable structural abnormalities.
- **Celiac Disease**: An autoimmune disorder triggered by gluten, leading to damage in the small intestine and impaired nutrient absorption.
- **Diverticulitis**: Inflammation or infection of small pouches (diverticula) that can form in the walls of the large intestine.

- **Colorectal Cancer (CRC)**: Malignant growths in the colon or rectum, which may arise from benign polyps and require early detection and treatment for the best outcomes.

Each of these conditions involves complex interactions between genetics, immune responses, and environmental factors, highlighting the intestines' critical role in health and disease.

1.3.3 Polyps

Polyps are abnormal tissue growths that arise on the inner lining of the colon and rectum. While most polyps are benign and asymptomatic, some have the potential to become cancerous. Understanding the different types of polyps and their development provides crucial insights into their role in CRC, the second leading cause of cancer-related deaths worldwide (Bray et al., 2024).

The Paris Classification (Participants in the Paris Workshop, 2002) is a standardized system used to describe the macroscopic appearance of polyps in the GI tract, particularly the colon. It categorizes polyps based on their shape and growth pattern as seen during endoscopy. This classification helps in assessing the malignancy risk, guiding removal techniques, and predicting patient outcome (Johnson et al., 2023; Wilson et al., 2023).

According to this classification, as summarized in Figure 1.8, polyps are divided into two main groups: polypoid lesions, which includes polyps with an elevation larger than 2.5 mm, and non-polypoid lesions, which encompasses those with a smaller elevation. There is a third group, ulcerated lesions, that account for ulcerations that are eroded into the intestinal wall as can be seen in the figure.

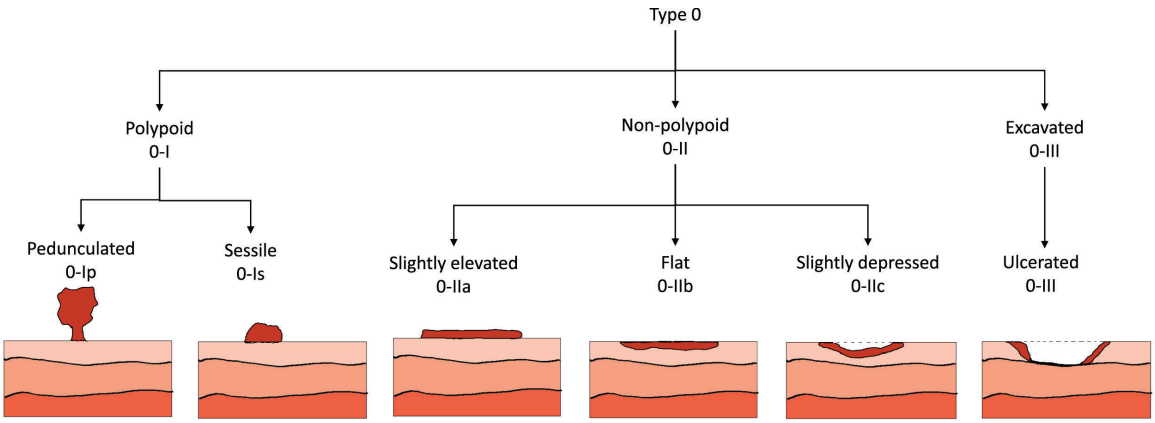


Figure 1.8: Classification of polyps according to Paris Classification. Image from Wilson et al. (2023).

Pedunculated polyps (Type 0-Ip)

Pedunculated polyps resemble mushrooms, featuring a head attached to the mucosal lining by a narrow stalk. This stalked appearance makes them relatively straightforward to identify and remove during endoscopic procedures. The presence of a stalk allows endoscopists to easily snare and remove these polyps during a polypectomy. Generally, pedunculated polyps are less likely to harbor invasive cancer compared to other types, particularly if they are not excessively large. Their distinct shape also means they tend to move slightly when touched by the endoscope, which helps in their identification.

Sessile polyps (Type 0-Is)

Sessile polyps are raised lesions with a broad base and no stalk, sitting directly on the mucosal surface and resembling a dome in shape. These polyps are more challenging to manage because their wide base makes complete removal during endoscopy more difficult. Sessile polyps can pose a higher risk of malignancy, especially if they are large or located on the right side of the colon, such as in the descending or sigmoid sections. Due to their broad attachment to the mucosa, sessile polyps do not move as easily as pedunculated polyps during removal, making them somewhat harder to detect and fully resect.

Non-polypoid lesions (Type 0-II)

Non-polypoid lesions, also known as flat polyps, are characterized by minimal elevation above the mucosal surface and are classified into three subtypes: slightly elevated (0-IIa), completely flat (0-IIb), and slightly depressed (0-IIc).

Slightly elevated polyps (0-IIa) have a minimal rise above the mucosa, appearing almost flat but with a slight bump. Completely flat polyps (0-IIb) lie flush with the mucosal surface, blending seamlessly with the surrounding tissue. Slightly depressed polyps (0-IIc) show a slight indentation below the mucosal level, which can make them particularly difficult to identify.

Flat polyps, especially the depressed type, pose a higher risk for containing cancerous cells. Their subtle appearance makes them challenging to detect during colonoscopy, increasing the importance of thorough endoscopic examination. These polyps require meticulous attention as they are more likely to be missed and can be harder to remove completely due to their minimal elevation or depression.

Excavated Polyps (Type 0-III)

Excavated, also known as depressed or ulcerated polyps are characterized by a distinctive morphology where part of the lesion is below the level of the surrounding mucosa. Unlike

the more common raised, pedunculated, sessile or flat types, these polyps have a sunken or crater-like appearance. They may appear as a shallow pit or a deeper excavation into the mucosal layer, creating a concave surface. This type of polyp is less frequently encountered in the colon.

1.4 Colorectal Cancer

CRC is the third most common type of cancer worldwide and ranks second on the list of most aggressive and deadly cancers (Siegel et al., 2024). According to the Global Cancer Observatory, out of an estimated total of 1.9 million cases in 2022, this disease has caused the death of more than 903,000 people worldwide (Bray et al., 2024).

Several factors contribute to the risk of developing CRC. Lifestyle is one of the most important factors; unhealthy habits such as smoking, heavy alcohol consumption, and sedentary behavior have been linked to an increased risk of this cancer. Diet is also a significant factor, with diets high in red and processed meats and low in fiber, fruits, and vegetables being associated with higher incidence rates (O'Sullivan et al., 2020; Ahmed, 2020). A clear relationship exists between obesity and CRC, supported by extensive research (Ye et al., 2020; Lauby-Secretan et al., 2016). Genetic predisposition is another critical factor; individuals with a family history of this type of cancer or with other inherited conditions are at higher risk (Valle, 2014).

One of the initial signs of the development of CRC is the appearance of polyps in the colon that grow in an uncontrolled manner (Bond, 2003). This process, illustrated in Figure 1.9, can take years, often starting without noticeable symptoms, which makes regular screening vital. The detection of polyps when they are still small is crucial to prevent their transformation into cancer. Screening programs are aimed to detect early-stage cancer, improving the patient's chances of survival (Levin et al., 2018; Loveday et al., 2021) because, as the cancer progresses, it can penetrate deeper layers of the bowel wall and spread to other parts of the body (a process known as metastasis). For individuals at average risk, screening is often recommended starting at age 50, but those with higher risk factors may need to begin earlier.

1.5 Endoscopy

Endoscopy is a medical technique that allows doctors to look inside the body without making large incisions. This is achieved using an instrument called an endoscope, which is a long, thin, flexible tube with a light and camera at the tip. The camera transmits images to a screen, providing a clear view of the internal organs or cavities. Endoscopy is versatile and used in many areas of medicine to diagnose, monitor, and sometimes treat various conditions.

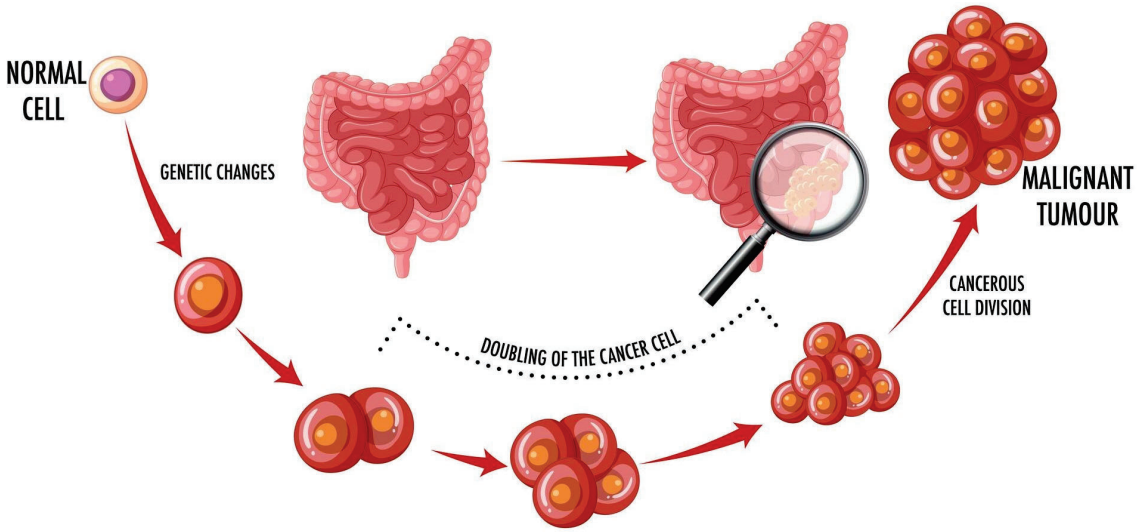


Figure 1.9: CRC developement. Due to genetic changes, normal cells mutate and replicate, forming polyps that can become malignant tumors. Image from @brgfx (Freepik).

.

The types of endoscopy vary depending on the area of the body being examined. For example, an upper endoscopy, also known as esophagogastroduodenoscopy, looks at the esophagus, stomach, and the beginning of the small intestine. Bronchoscopy inspects the airways and lungs, while cystoscopy examines the bladder. Other types include laparoscopy, which explores the abdominal or pelvic cavities, and arthroscopy, which looks into joints.

Colonoscopy is a specific type of endoscopy that focuses on the colon (large intestine) and rectum. This procedure is useful for detecting various conditions such as polyps, tumors, inflammation, and sources of bleeding within the colon. The detailed examination provided by a colonoscopy is particularly important for screening and early detection of CRC.

The colonoscopy procedure involves several key steps. Before the procedure, thorough preparation is necessary to ensure the colon is completely empty. This typically involves a special diet and the use of laxatives the day before the procedure, along with fasting from solid foods and consuming clear fluids (Hassan et al., 2013).

On the day of the colonoscopy, patients are usually given a sedative or anesthesia. The doctor then inserts the colonoscope into the rectum and advances it through the colon. The camera at the end of the colonoscope transmits images to a monitor, allowing the doctor to examine the entire length of the colon. During this process, the doctor can also perform interventions such as taking biopsies or removing polyps if necessary, as shown in Figure 1.10.

After the procedure, patients are monitored as the effects of sedation wear off. It is common to experience some mild cramping or bloating due to the air introduced into the colon during the examination. The doctor will discuss the findings with the patient, and any tissue samples taken will be sent to a laboratory for further analysis.

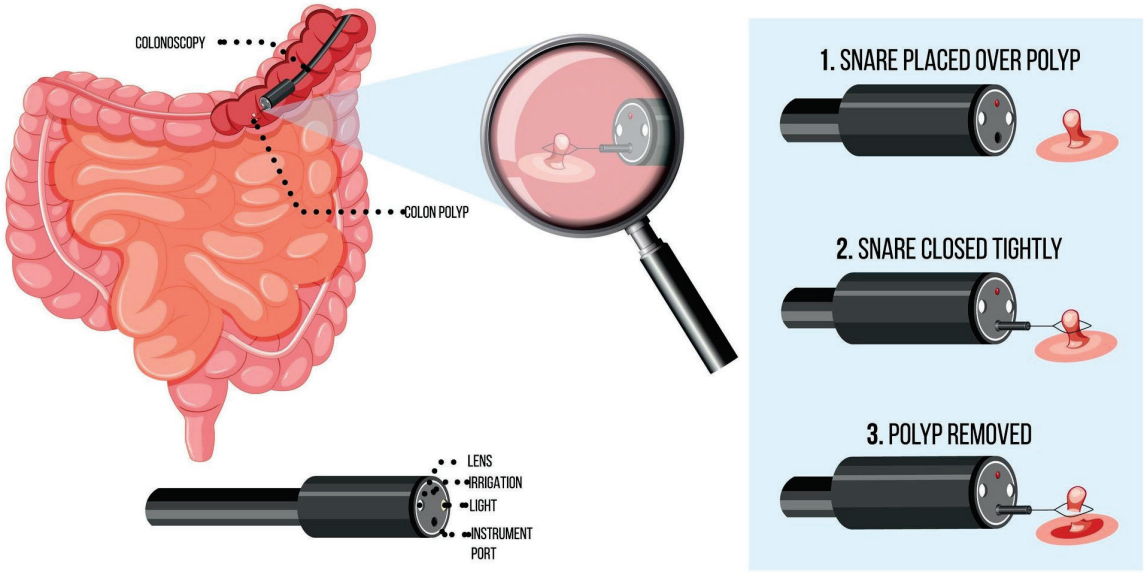


Figure 1.10: Colonoscopy with polyp removal. Image from @brgfx (Freepik).

.

1.6 Capsule Endoscopy

CE is a minimally invasive diagnostic procedure that allows doctors to examine the GI tract, particularly the small intestine and colon (Iddan et al., 2000). The small intestine is of special interest because is difficult to reach with traditional endoscopy methods. This technology utilizes a small, pill-sized camera that patients swallow, enabling the capture of thousands of images as it travels through the digestive system. These images provide detailed views of the mucosal lining, helping to diagnose and monitor various GI conditions. Figure 1.11 shows some of the steps involved in a CE procedure.

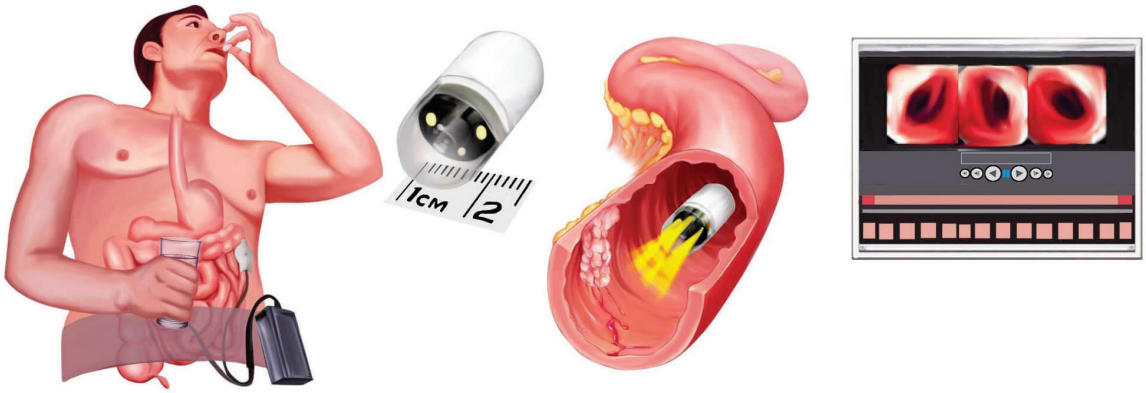


Figure 1.11: CE overview. The patient swallows the capsule, which records images of the digestive system. The data is transmitted wirelessly to a data recorder that can be reviewed later by medical personnel. Image from gutworks.

One of the primary benefits is its non-invasive nature. Unlike traditional endoscopy,

which requires sedation and the insertion of scopes into the body, CE involves simply swallowing a small, pill-sized camera. This approach is far more comfortable for patients and reduces anxiety associated with more invasive procedures (Ismail et al., 2022). Additionally, CE is highly convenient. Patients can carry out most of their daily activities while the capsule is capturing images, making it less disruptive to their routines.

Another advantage of CE is the comprehensive imaging it provides. The capsule captures thousands of high-resolution images as it travels through the digestive system, offering a detailed and extensive view of the entire small intestine. This level of thoroughness is difficult to achieve with conventional endoscopic techniques, which often struggle to access and visualize the full length of the small intestine due to its complex structure and significant length. However, the reading of CE videos is time-consuming and requires qualified medical personnel (Maieron et al., 2004; Rondonotti et al., 2020; Koulaouzidis et al., 2021).

The structure of a CE is quite simple: one or two cameras, a battery, LEDs to illuminate the scene, and a transmitter to send the images to the patient's recorder. Figure 1.12 shows some of the most important devices currently available.

(a) PillCam SB3. Image from Medtronic.

(b) PillCam COLON 2. Image from Medtronic.

(c) Mirocam. Image from Intromedic.

(d) Olympus EC-10. Image from Olympus.

Figure 1.12: Some of the available CE devices.

The most common types of CE procedures use the natural movement of the digestive system, known as peristalsis, for the capsule to advance through the digestive system. Recent methods currently in development use external magnetic fields to guide, orient, and power the capsule to stabilize it during its advance through the intestines (Shamsudhin et al., 2017; Jiang et al., 2024).

Since its initial conceptualization in 1980 by engineer Gavrial Iddan and gastroenterologist Eitan Scapa, and the first pill swallowed in 1997, the technology of these devices has advanced significantly. In 2001, the Food and Drug Administration (FDA) approved the capsule from Given Imaging, the company founded by Iddan, for patient use. In 2014, Medtronic acquired the technology from Given Imaging and has since sold over 4 million capsules. Table 1.1 summarizes some of the currently available capsules and their specifications.

Device	Size (mm)	Weight (g)	Field (degree)	Images/s	Battery life	Resolution (pixels)
PillCam SB 3, Medtronic	11 x 26	3	156	2-6	8h	320 x 320
PillCam SB 3 EX, Medtronic	11 x 26	3	156	2-6	12h	320 x 320
PillCam COLON 2, Medtronic	11 x 32	2.9	172	4-35	10h	256 x 256
PillCam Crohn's, Medtronic	11 x 32	2.9	168	4-35	10h	256 x 256
PillCam UGI, Medtronic	11 x 32	2.9	172	18-35	90min	256 x 256
EndoCapsule, Olympus	11 x 26	3.3	160	2	12h	-
CapsoCam Plus, CapsoVision, Inc	11 x 31	4	360	5	15h	-
Mirocam single-lens, Intromedic	10.8 x 24.5	3.2	170	3	12h	320 x 320
Mirocam dual-lens, Intromedic	10.8 x 23.1	3.5	340	3	12h	320 x 320
OMOM, Jlinshan	11 x 25.4	3	172	2-10	12h	512 x 512

Table 1.1: Specifications of different CE devices.

1.7 Analyzing a Capsule Endoscopy Video

A CE procedure does not conclude when the capsule exits the patient's body. Instead, the process continues with trained medical personnel reviewing hours of video footage at high speed to identify lesions and other abnormalities. This review procedure involves several key steps, summarized below.

1.7.1 Bowel Preparation

Before any exploration of the video from a CE begins, it is essential to ensure that the footage is usable. This means that the view of the mucosa must be clear and not obstructed by intestinal contents or residue. If the presence of these residues is suspected to obscure potential lesions, the video may be deemed unacceptable for diagnostic purposes. In such cases, the procedure must be repeated, which involves redoing the CE and possibly revising the bowel preparation methods used prior to the procedure. This may include optimizing the patient's diet, using more effective cleansing solutions, or refining pre-procedure instructions.

1.7.2 Landmark Identification

Once the video footage is deemed usable, the next step is to identify and examine the key anatomical landmarks of the organ being reviewed. For a comprehensive assessment of the small intestine, it is important to identify specific areas such as the pylorus, which is the junction between the stomach and the duodenum, and the ileocecal valve, where the small intestine connects to the large intestine. Identifying these landmarks ensures that the entire small intestine is thoroughly evaluated, allowing for the detection of any abnormalities.

Similarly, when analyzing the large intestine, also known as the colon, it is essential to locate the ileocecal valve, which marks the transition from the small intestine to the large intestine, and the anorectal valve, which separates the end of the colon from the anus, as depicted in Figure 1.7.

1.7.3 Pathology Identification

Once the key anatomical landmarks are accurately identified, the next step is to carefully review the video footage between these points to detect any potential pathologies. This process is crucial for ensuring that the entire section of the GI tract in question, whether it be the small intestine or the large intestine, is thoroughly examined. The goal is to identify any abnormalities, such as polyps, ulcers, or other lesions, that could indicate underlying health issues.

The European Society of Gastrointestinal Endoscopy (ESGE) recommends reviewing these videos at a speed of 10 frames per second (Rondonotti et al., 2018) to balance efficiency with accuracy. While this guideline is widely followed, it is important to note that there is no definitive evidence proving that this specific frame rate significantly reduces the risk of missing potential lesions. However, adhering to such standardized protocols helps maintain a consistent approach to video review, thereby improving the reliability of the diagnostic process.

1.8 Contributions

This thesis is presented as a copilot for physicians, designed to assist them in their daily tasks by offering support tools that simplify their work. Our goal is to enhance the efficiency and accuracy of medical processes, from data acquisition to pathology identification, through the integration of advanced AI techniques.

- **Dataset Compilation**: We developed a method for automatically generating video CE datasets. Additionally, we present HistoLungs700, a dataset of histological images specifically designed for detecting lung carcinomas.
- **Bowel Preparation Classification**: We introduce a method to classify the bowel preparation quality of CE videos using minimally annotated data. Current methods are expensive to train due to the large data requirements. Our system, however, achieves comparable agreement to real physicians with a low-effort annotation strategy.
- **Landmark Detection**: We enhance the review process by developing a model to automatically detect landmarks, identifying the entrance and exit points of the intestines. This improves upon state-of-the-art methods, streamlining the workflow for physicians.
- **Polyp Detection Tool**: We test a polyp detection application in a real clinical environment, comparing it to the current review method. Our system achieved higher accuracy in less time, demonstrating that the integration of AI in clinical routines can reduce the likelihood of missing polyps while being significantly faster than current

methods. This system is currently being evaluated in a larger clinical trial to fully assess its potential.

1.9 Document Structure

This thesis is organized into six chapters:

1. **Introduction**. This chapter lays the foundation for understanding the thesis. It reviews key concepts of digestive system anatomy relevant to the study and introduces artificial intelligence applications in medical imaging, specifically focusing on CE, which is the primary tool used throughout this document.

2. **Learning from Data**. This chapter emphasizes the importance of data. It presents a method for automatically creating a dataset of CE videos without the need for previously annotated data, highlighting the significance of data collection and preparation in AI applications.

3. **Bowel Preparation Assessment**. Before reviewing a CE video, it's crucial to assess its quality. This chapter introduces a method that uses minimally annotated data to evaluate bowel preparation using a well-known scale, ensuring the video is clean enough for analysis.

4. **Landmark Identification**. Once the video's quality is confirmed, the next step is to identify the organ of interest, such as the large or small bowel. This chapter presents a new method for accurately identifying anatomical landmarks, such as flexures, despite the camera's high speed and the large number of images captured.

5. **Polyp Detection**. This chapter introduces a method for detecting colonic polyps, which is critical for the early detection of CRC. The study compares the proposed method with traditional review systems, demonstrating improved accuracy, particularly for small polyps or those appearing in a limited number of frames. Additionally, this chapter discusses the method's use in a real clinical trial, the CESCAIL study.

6. **Conclusions and Future Work**. The final chapter of this thesis summarizes the key findings and outlines potential research directions that could enhance the results further. Additionally, it lists all the publications submitted during the course of the research.

Chapter 2

Learning from Data

Contents

Datasets are a fundamental component of any AI solution. Training AI methods requires a data source, whether derived from the real world, simulations, or artificial generation. This chapter will emphasize the critical role of data. First, it will discuss the significance of datasets and their impact on AI solutions. Next, it will present a method for automatically generating a dataset from unlabeled CE videos. Finally, the chapter will describe a collaboration with the Hospital Clínico Universitario de Valladolid, highlighting how assistance was provided in creating a dataset and validating an AI solution using it.

2.1 The learning process

Learning is the process by which a model improves its ability to make accurate predictions or classifications based on data. This process is grounded in the use of datasets, which consist of samples and their corresponding labels, $\mathcal{D} = \{x_i, y_i\}_{i=1}^{N}$. Each data point in a dataset can be denoted as x_i (a sample) and y_i (the label associated with that sample). The essence of learning revolves around approximating a function that links these samples to their labels.

Let's start with a ground-truth function g, which perfectly maps each sample x_i to its corresponding label y_i. In practice, however, the exact form of g is unknown. Instead, we work with a model function f, which is our attempt to approximate g. The goal of training the model is to make f as close as possible to g. This means that for each sample x_i, the model's prediction $\hat{y_i} = f(x_i)$ should closely match the actual label y_i.

During the learning process, the model is initially set up with some parameters, often denoted by θ. As a result, the model function f can be expressed as f_θ where θ represents the parameters that are adjusted during training. As it processes the data, the model's parameters are adjusted to minimize the difference between the predicted values $\hat{y_i}$ and the actual labels y_i. This difference is measured by a loss function, which quantifies the model's performance. By using optimization techniques, such as gradient descent, the model parameters are iteratively updated to reduce the loss.

The learning process involves training the model on a dataset and then evaluating its performance on new, unseen data. This evaluation checks how well the model generalizes beyond the training examples. A model that generalizes well will make accurate predictions on new data, not just on the data it has already seen. Overfitting occurs when a model learns the details and noise in the training data to the extent that it performs exceptionally well on this data but poorly on unseen data. In other words, the model becomes too specialized to the training set and fails to generalize to new examples. This happens because the model captures not only the underlying patterns but also the random fluctuations or anomalies present in the training data.

Figure 2.1 illustrates illustrates a collection of points with three different fitting functions, demonstrating the problems of overfitting and underfitting.

To detect and avoid overfitting, it is essential to partition the data into three distinct sets: training, validation, and test, each serving a specific purpose.

The training set is used to fit the model. During training, the model learns to map inputs to outputs by adjusting its parameters based on this data. Once the model has been trained, its performance is evaluated on the validation set. The validation set acts as a proxy for unseen data, providing a measure of how well the model performs when exposed to new, but not entirely unknown, data. It is also used to fine-tune hyperparameters and make decisions about the model's architecture.

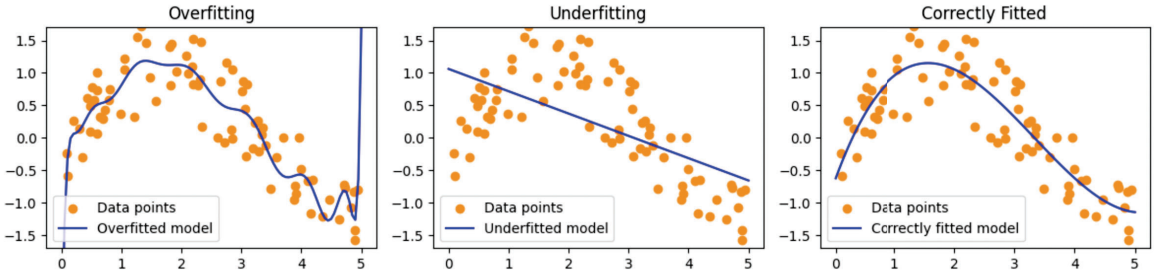


Figure 2.1: Overfitting, underfitting and correctly fitting a set of points. Points were generated using a sinus function, adding random normal noise to illustrate these concepts.

Finally, the model's performance is assessed on the test set. This set consists of data that the model has never seen before and is used to provide a final evaluation of its generalization ability. The test set is crucial for obtaining an unbiased estimate of how the model will perform in real-world scenarios, as it has been completely separate from the data used during training and validation.

To combat overfitting, various techniques such as regularization, cross-validation, and careful model selection are employed. Regularization methods add constraints to the model to prevent it from becoming overly complex, while cross-validation helps ensure that the model's performance is robust across different subsets of the data. By using these strategies, the aim is to build a model that generalizes well, providing accurate predictions not only on the training data but also on new, unseen data.

2.1.1 Deep Learning

Deep Learning (DL) is a subset of Machine Learning (ML) and AI that involves the use of neural networks with many layers, often referred to as deep neural networks. At the core of these networks are artificial neurons. Each artificial neuron receives input data, processes it, and passes the result to the next layer. By stacking many layers of these neurons, DL networks can automatically learn to recognize complex patterns and features directly from raw data, without requiring manual feature engineering. This capability makes DL particularly effective for tasks like image recognition, Natural Language Processing (NLP), and other complex decision-making processes.

An artificial neuron is a function $a : \mathbb{R}^n \longrightarrow \mathbb{R}$ designed to mimic the behavior of a biological neuron. It processes input data $x \in \mathbb{R}^n$, which can come from either the input layer or preceding neurons, and produces an output by transforming this data through the following formula:

$$a(x) = a(x_1, \ldots, x_n) = \sigma\left(\sum_{i=1}^{n} w_i x_i + b\right) = \sigma\big(w^T x + b\big) \tag{2.1}$$

In the expression, $w = (w_1, \ldots, w_n)$ represents the weights assigned to each input, which determine the significance of each input feature, $b \in \mathbb{R}$ is the bias, a constant that shifts the output function and $\sigma : \mathbb{R} \longrightarrow \mathbb{R}$ is the activation function, which introduces non-linearity into the model. This non-linearity is crucial because it allows the neuron to learn and model complex patterns beyond simple linear relationships. The most commonly used activation functions are *sigmoid*, *tanh*, *ReLU* or *Leaky ReLU* among others. This expression can be generalized into the multi-layered artificial neural network. For a neural network of L layers, the expression would be:

$$f_\theta(x) = (a_L \circ a_{L-1} \circ \cdots \circ a_2 \circ a_1)(x) = \sigma_L\Big(W_L\sigma_{L-1}\big(W_{L-1}\ldots\sigma(W_1x+b_1)+\cdots+b_{L-1}\big)+b_L\Big) \tag{2.2}$$

where θ is a collection of weights, $\{W_i\}_{i=1}^n$, and biases, $\{b_i\}_{i=1}^n$.

In broad terms, a DL model is represented as $f_\theta : X \longrightarrow Y$, where X denotes the input space and Y represents the output space. The function f_θ transforms input data from X to outputs in Y, with θ denoting the parameters of the model learned during training. The training process involves minimizing the error associated with a target objective that depends on a dataset $\mathcal{D} \subset X$ and the fixed model parameters. This error, also referred to as the loss function $l(\mathcal{D}, \theta) \in \mathbb{R}$, evaluates the disparity between the model's predictions and the actual labels provided in the dataset.

The output of a classifier, which is a neural network designed to assign samples to a finite number of classes, can be interpreted as $P_\theta(y|x)$. This notation represents the probability that a given sample x belongs to class y given the model parameters θ. For simplicity, the parameters θ are often omitted in the notation, and the probability is typically expressed as $P(y|x)$.

Through the backpropagation algorithm, the error is propagated backward through the layers, starting from the last layer and moving all the way to the first. This process adjusts the weights of each layer in order to minimize the overall error of the model. This iterative adjustment of weights based on the error propagation allows the neural network to learn and improve its predictions over time. By updating the weights in the direction that reduces the error, backpropagation enables the network to fine-tune its parameters to better match the desired outputs for the given inputs, thereby optimizing maximum likelihood. This is equivalent to minimizing the log-likelihood:

$$\arg\min_\theta \left(-\sum_{i=1}^{N} \log\left(P_\theta(y_i|x_i)\right) \right) \tag{2.3}$$

2.2 Embeddings

Embeddings are a crucial concept in the DL and ML field, particularly in the context of NLP and other high-dimensional data applications. Essentially, embeddings are dense vector representations of data, which capture the semantic relationships between items in a continuous vector space. This transformation from discrete to continuous space allows for the preservation of meaningful information and relationships, which can be leveraged by various ML models. For instance, in NLP, word embeddings like Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and BERT (Devlin et al., 2018) map words or phrases to vectors of real numbers, where semantically similar words are positioned closer together in the vector space. This facilitates the models to perform tasks such as sentiment analysis, translation, and question-answering more effectively.

In the image paradigm, embeddings play an equally significant role. Image embeddings involve transforming images into high-dimensional vectors that capture essential features and patterns within the images. Techniques such as Convolutional Neural Network (CNN) are commonly used to generate these embeddings. These vectors can then be used in various tasks, including image classification, object detection, and image retrieval. For example, in a CNN, the final layers produce an embedding that encapsulates the learned features of the image, which can then be used to compare and categorize images based on their content. The versatility and power of embeddings make them indispensable in both textual and visual data applications, enabling more nuanced and sophisticated analysis and interpretation of complex data.

Image embeddings have been significantly advanced by the development of DL architectures such as AlexNet (Krizhevsky et al., 2012) and ResNet (He et al., 2016). AlexNet, marked a breakthrough in image classification by utilizing deep CNNs to learn hierarchical features from raw image pixels. This architecture, consisting of multiple convolutional and fully connected layers, demonstrated the effectiveness of DL in extracting rich and informative embeddings from images. AlexNet's success in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) showcased the power of deep embeddings in achieving unprecedented accuracy in image recognition tasks.

ResNet, or Residual Networks, further revolutionized the field by addressing the problem of vanishing gradients in deep networks, which often hindered training of very deep models. Introduced in He et al. (2016), ResNet employs residual learning with shortcut connections to allow gradients to propagate more effectively through the network. This innovation enabled the construction of extremely deep networks with hundreds or even thousands of layers, significantly improving the quality of image embeddings. ResNet's architecture, featuring identity mappings (also known as skip connections) and batch normalization, has become fundamental in computer vision, driving state-of-the-art performance in tasks like image classification, object detection, and segmentation. The advancements brought by AlexNet and ResNet have cemented the importance of robust embeddings in

capturing complex visual patterns, driving forward the capabilities of AI in processing and understanding image data.

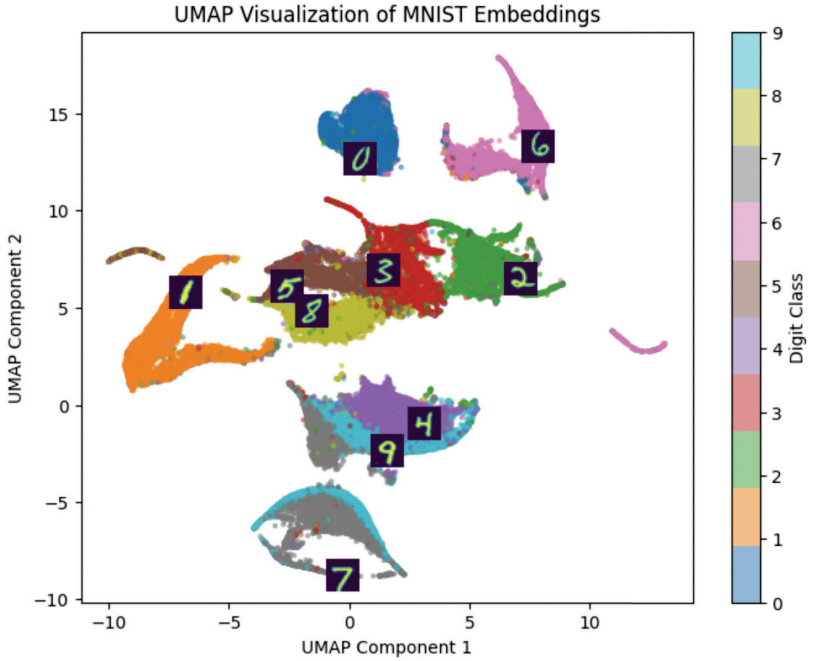


Figure 2.2: MNIST images projected into 2-D space using UMAP.

To fully understand the concept of embeddings and how they are created, consider the following formal definition: Let f be a function that transforms a sample $x \in X$ into a vector representation $f(x) = z \in \mathbb{R}^d$. This function f is known as an encoder, and the resulting vector z is called an embedding of x. Note that this vector representation is not unique; it is entirely dependent on how the function f is trained and the specific problem being addressed. Figure 2.2 shows a visual representation of image embeddings. Images from MNIST (Lecun et al., 1998) dataset, were flattened to a vector representation of 784 values and projected to a 2-dimensional embedding space using UMAP (McInnes et al., 2018).

Embeddings are typically not learned directly. Instead, they often emerge as feature vectors within the intermediate layers of a neural network architecture. One common approach to learning embeddings is through the use of an autoencoder.

An autoencoder (Rumelhart et al., 1986) is a type of neural network that does not require annotated data, which is why it is considered an unsupervised technique. Its primary characteristic is that the output is designed to have the same shape and domain as the input. The autoencoder model learns to compress the data from its original space into a smaller vector representation, the embedding, and then reconstructs the original sample from this encoded representation. This process helps the model capture and preserve essential features of the data in the embedding. The basic autoencoder architecture is presented in Figure 2.3.

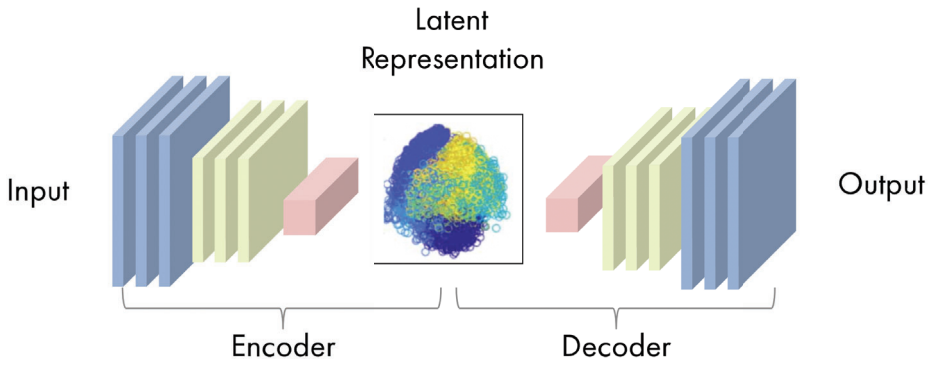


Figure 2.3: Basic autoencoder architecture. Images are passed through the encoder to obtain the embedding, which is the latent representation of the image. This vector is then passed through the decoder, which reconstructs the image. Image from MathWorks.

2.3 Capsule Endoscopy Datasets

This thesis focuses on CE solutions, making CE images or videos the most crucial resource for developing methods for physicians. Without these, it is impossible to accurately replicate a real clinical environment. A high-quality CE dataset must possess several key properties:

- **Proper data splitting:** Each image needs to be linked to a patient identifier, or at least ensure that a proper train, validation, and test split can be done. Since the camera can be stuck for several minutes (resulting in hundreds of frames) in the same location, a CE video contains a large number of similar frames. Ensuring that all these frames are not split between the training and validation sets ensures a proper evaluation. Otherwise, there is a problem of data leakage, one of the most common issues faced by AI beginners.

- **Large variety:** A large variety of patients should be considered. Some of the pathologies we will be dealing with later on, such as polyps, appear very infrequently in the videos. Ensuring there is a large number of patients in the dataset will make it richer and result in a better model.

- **High-quality annotations:** Accurate and detailed annotations are essential for training and evaluating AI models. Each frame in the dataset should be annotated with relevant clinical information, including the presence of pathologies, anatomical landmarks, and other significant findings. High-quality annotations help in developing precise and reliable models.

- **Temporal continuity:** Maintaining the temporal continuity of frames is important for understanding the progression of video content. The dataset should include sequences of frames that capture the dynamic changes within the GI tract. This allows models to learn patterns over time, which is essential for detecting anomalies that

may develop gradually or identifying specific changes in the video, such as the valves marking the entrance and exit of the small and large intestines.

There are very few large and comprehensive CE datasets, as most published studies use private data. Releasing such data is challenging due to stringent controls required to protect patient privacy. Additionally, competing interests from private companies further complicate the release of these datasets. However, some CE datasets are available, as shown in Table 2.1.

Dataset	Images	Labeled	Classes	Resolution	Device	Public	Organ	Published
KID (Set1)	77	77	9	360×360	MiroCam	❖	-	Koulaouzidis et al. (2017)
KID (Set2)	2,371	2,371	8	360×360	MiroCam	❖	-	Koulaouzidis et al. (2017)
KID (Videos)	3 videos	0	0	360×360	MiroCam	❖	-	Koulaouzidis et al. (2017)
RedLesions (Set1)	3,295	3,295	1	Multiple	MiroCam, PillCam SB1, SB2, SB3	✓	Small Bowel	Coelho et al. (2018)
RedLesions (Set2)	600	0	2	320×320	PillCam SB3	✓	Small Bowel	Coelho et al. (2018)
CAD-CAP	25,124	25,124	4	576×576	PillCam SB3	Under request	Small Bowel	Leenhardt et al. (2020)
ChronIPI	3,498	3484	7	320×320	PillCam SB3	Under request	Small Bowel	de Maissin et al. (2021)
Kvasir-Capsule	4,741,504	47,238	14	336×336	Olympus Endocapsule 10	✓	Small Bowel	Smedsrud et al. (2021)

Table 2.1: CE datasets. Datasets marked with the symbol ❖, were available at the moment we downloaded them but are not available anymore.

KID

The Knowledge-based Intelligent Digestive (KID) dataset (Koulaouzidis et al., 2017) is an open-access, non-profit repository that provides high-quality annotated CE images and videos. It aims to assist clinicians in diagnosis and to foster the development of automated medical decision support systems. This dataset includes contributions from an international community of CE researchers, ensuring that all data provided is anonymized and stripped of any identifiers. The annotations are verified by expert reviewers from the international scientific committee of KID, guaranteeing the quality and accuracy of the information.

The dataset is divided into three parts. Dataset 1 comprises 77 CE images from MiroCam capsule endoscopes, illustrating various GI abnormalities such as angioectasias, aphthae, chylous cysts, polypoid lesions, villous edema, bleeding, lymphangiectasias, ulcers, and stenoses (Figure 2.4). Dataset 2 contains 2,371 CE images from the same type of en-

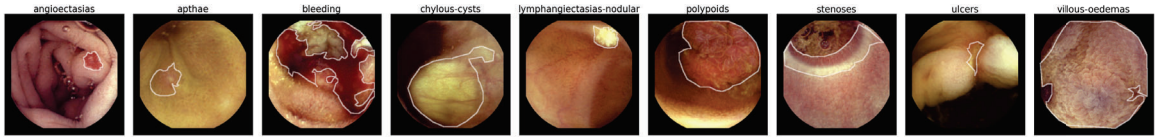


Figure 2.4: KID Set 1: 77 images organized in 9 pathological classes with annotated masks. In this figure only the border of the masks are displayed for better visualization.

doscopes, depicting a broad spectrum of small bowel findings, including polypoid, vascular, and inflammatory lesions, along with normal images from the esophagus, stomach, small bowel, and colon (Figure 2.5).

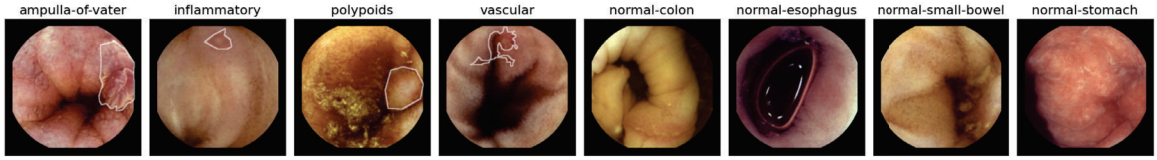


Figure 2.5: KID Set 2: 2,371 images organized in 8 classes: 4 pathological and 4 non-pathological. In this figure only the border of the masks for the pathological classes are displayed for better visualization.

Additionally, the dataset includes 3 CE videos obtained using MiroCam capsule endoscopes, further supporting research and development in GI diagnostics (Figure 2.6).

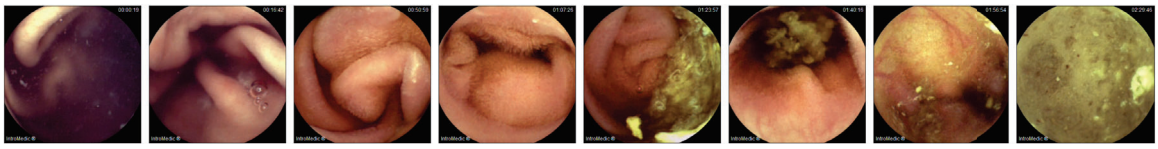


Figure 2.6: Images extracted from KID's dataset (Video 1), in sequential order.

CAD-CAP

The Computer-Assisted Diagnosis for Capsule Endoscopy (CAD-CAP) dataset is a comprehensive resource designed to support the development of AI tools for interpreting CE videos. This large, multicenter database encompasses a total of 4,174 third-generation small bowel CE videos, meticulously collected from twelve French endoscopic centers. Out of these, 1,480 videos (35%) were identified to contain at least one pathological finding, while the rest were deemed normal. This vast collection aims to address the tedious and time-consuming nature of CE, which involves scrutinizing an average of 50,000 frames per video, typically requiring 30 to 60 minutes per video.

The dataset is composed of 25,124 frames in total, providing a rich source of data for ML and automated diagnostic tool development. Specifically, it includes 5,184 frames with pathological findings, which are further categorized into 718 frames showing fresh blood, 3,097 frames with vascular lesions, and 1,369 frames displaying inflammatory and ulcerative lesions. Additionally, the dataset features 20,000 normal frames extracted from 206 normal CE videos, offering a robust control set. Each frame is paired with a short video sequence of 25 frames upstream and downstream to ensure contextual understanding. This database has already been part of international challenges (Endoscopic Vision Challenge, 2017) on medical computerized analysis, highlighting its significance and utility in advancing CE diagnostic technology. Particularly, reduced versions of this dataset were released in the Gastrointestinal Image ANAlysis (GIANA) challenges: GIANA 2017, GIANA 2018 and GIANA 2021.

(a) Inflammatory (b) Vascular Lesions (c) Normal

Figure 2.7: CAD-CAP dataset. The version from GIANA 2018 contains three classes: Inflammatory, Vascular lesions and Normal images. Some pathological images have acompaning masks highlighting the area of interest. Here the masks are replaced by outlines for better visualization.

RedLesions

The RedLesions dataset introduced in Coelho et al. (2018), is composed of annotated frames extracted from small bowel CE videos, specifically to train and evaluate algorithms for red lesion detection. The dataset includes two sets of images:

- Set 1: This set contains 3,295 frames from various devices, including MiroCam and PillCam models SB1, SB2, and SB3. Among these frames, 1,131 contain red lesions such as angioectasias, angiodysplasias, and bleeding. The frames have been manually annotated, with resolutions of either 320×320 or 512×512, all of them resampled to 512×512 for uniformity. Figure 2.8 shows random images from the dataset, highlighting the outline of the available masks.
- Set 2: This set consists of a sequence of 600 frames from a PillCam SB3 video, intended to provide a more clinically realistic evaluation. Approximately 73% of these frames include red lesions, labeled as Blood/Non-blood although this annotations are not publicly available. Figure 2.9 shows a sequence of the dataset containing blood.

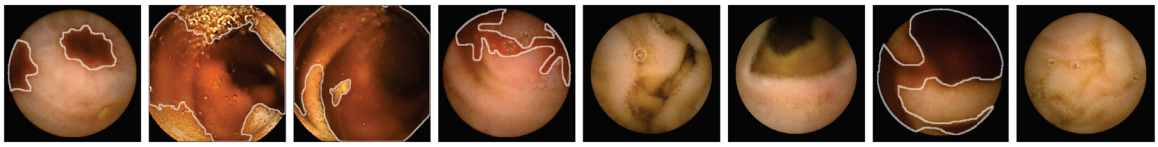


Figure 2.8: RedLesion Set1: Random images highlighting the available masks.

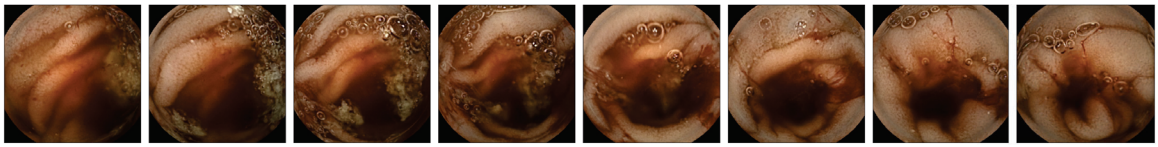


Figure 2.9: RedLesion Set2: Sequence containing blood.

The dataset was created to address the lack of publicly available annotated CE images large enough to train DL models effectively. However, Set 1 contains multiple images of the same patients without identifiers, making it impossible to perform a proper training split.

CrohnIPI

The CrohnIPI dataset (de Maissin et al., 2021) is s multicentric dataset for the development and testing of computer-aided diagnosis tools for small bowel CE. The dataset consists of 66 videos from 63 patients diagnosed with Crohn's disease, which were acquired using the Pillcam SB3 system between 2014 and 2018. From these videos, a total of 3,498 frames were extracted and annotated. These frames were categorized into three groups: 1,630 frames containing pathological lesions, 1,734 non-pathological frames, and 134 inconclusive frames. The annotation process involved multiple rounds, starting with an initial reading by one gastroenterology fellow and followed by reviews from three experts. This rigorous annotation process aimed to ensure the highest accuracy, resulting in a consensus annotation from the four experts.

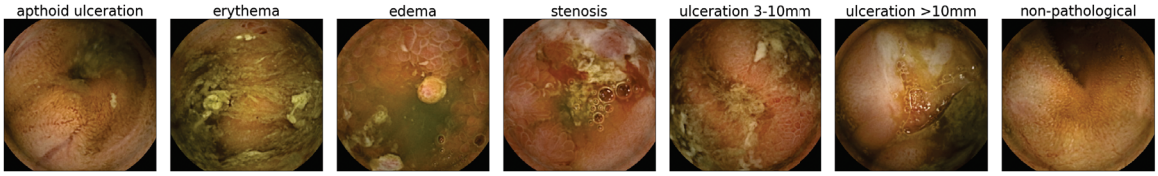


Figure 2.10: Images representing all the classes displayed in the CrohnIPI dataset.

The frames were carefully reviewed to ensure that only the most severe lesion was annotated when multiple lesions were present. This dataset not only facilitates the development of more accurate algorithms but also helps in understanding the inter-observer variability among experts in annotating CE images. The performance of the neural network classifiers tested on this dataset showed significant improvement, achieving a precision of 93.7%, sensitivity of 93%, and specificity of 95%.

Kvasir-Capsule

The Kvasir-Capsule dataset, presented in Smedsrud et al. (2021), is a comprehensive CE dataset collected from examinations at a Norwegian hospital. It includes 117 videos that together comprise 4,741,504 image frames. Out of these, 47,238 frames are labeled and medically verified, featuring bounding boxes around findings in 14 different categories, while the remaining 4,694,266 frames are unlabeled.

The labeled frames include various GI findings, which are categorized for analysis, classification, segmentation, and retrieval purposes. The dataset is publicly available under a Creative Commons Attribution 4.0 International License, encouraging its use, sharing, and adaptation in research and medical education.

Figure 2.11 shows an example image for each of the 14 classes.

The polyp class in the dataset contains only 55 images. In Jha et al. (2021), they released segmentation masks for the polyps, the Kvasir-Capsule-Seg subset. Images from

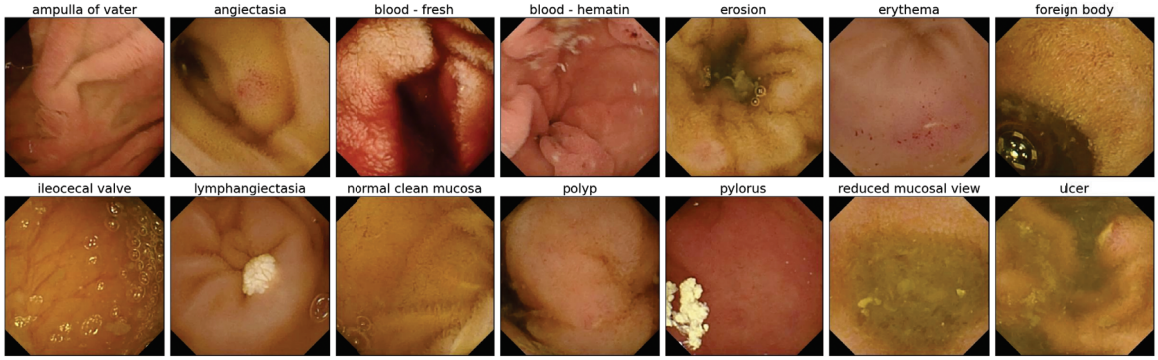


Figure 2.11: Kvasir-Capsule dataset. Shown here are 14 images, each representing a different category.

this subset are displayed in Figure 2.12. As can be seen, all of them contains the same polyp, a very big mass seen in different positions.

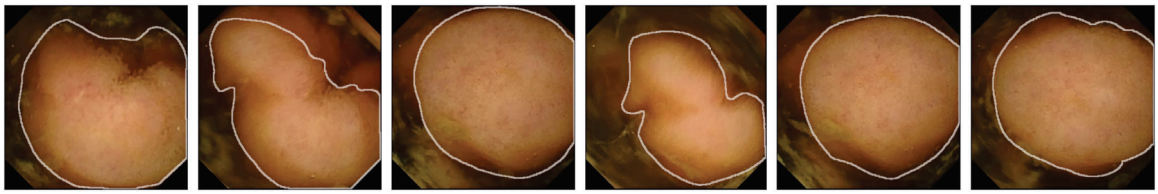


Figure 2.12: The 55 images depict various views of the same polyp, a very large mass. The area occupied by the polyp is labeled in the Kvasir-Capsule-Seg subset. Only the borders of the masks are shown here to enhance visualization.

2.4 Active Learning

Creating datasets in the medical domain is a complex and challenging task. Collecting data from patients often involves lengthy procedures, which can take hours and require strict controls and regulations to ensure patient privacy and protection. This process contrasts sharply with data acquisition in other industries, where data collection is often more straightforward. For example, in the mobile gaming industry, collecting data is relatively simple because millions of users are constantly generating data every second.

Active Learning (AL) is a technique designed to minimize the amount of labeled data needed to learn a task (Settles, 2009). The goal of an AL algorithm is to iteratively select the most informative samples for labeling, thereby reducing the overall labeling effort. By focusing on the most useful examples, AL aims to achieve efficient learning with fewer labeled data points.

2.4.1 The Active Learning Framework

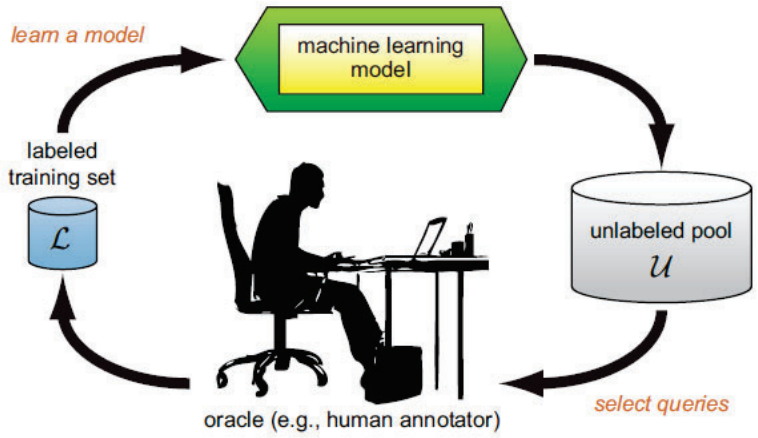


Figure 2.13: AL Framework. The annotator takes new data proposed by the method and adds it to the pool. Image from Settles (2009).

An AL process involves several steps, as depicted in Figure 2.13. Initially, all data is considered unlabeled. The method then proposes a subset of this data for annotation, which is provided to the annotator. The newly labeled data, along with the remaining unlabeled data, is fed back into the model, which then selects the next subset to be labeled. This iterative process continues until acceptable performance is achieved or the budget is exhausted.

Let $X = \{x_1, ..., x_n\}$ be a finite set of samples. Let L represent the set of labeled samples and U the set of unlabelled samples, such that $X = L \cup U$. Initially, $L = \emptyset$ and $U = X$. An AL algorithm A takes as input, at least, the labelled and unlabelled data and outputs a subset $U^* \subseteq U$ that optimizes a metric m subject to a constraint c:

$$U^* = A(L, U, \cdot) = \underset{V \subseteq U}{\arg\max} \{m(L, V) \text{ such that } c(V)\} \tag{2.4}$$

The constraints, like other parameters in this process, are problem-dependent. Typical constraints include a maximum size for the set V, such as $|V| \leq k$. In some cases, it is required to select exactly k samples in each iteration of the AL process, imposing the constraint $|V| = k$.

The core of each AL strategy is the metric m, which scores each subset of data. This metric often relies on the performance of a classifier to select samples with higher uncertainty, a method known as uncertainty sampling. Alternatively, diversity sampling seeks samples that not only improve classification directly but also increase the variability of the classifier, making it more robust to outliers and unseen data. We will now detail some of the primary techniques that employ both sampling strategies.

2.4.2 Uncertainty Sampling

Uncertainty sampling involves selecting samples for which the classifier is least certain about how to classify. The output of a binary classifier is typically a probability distribution between 0 and 1, which can be interpreted as the probability of belonging to the positive class, denoted here as class 1. We denote this probability as $P(y|x)$ where $x \in X$ is a sample and $y \in Y$ is a possible label for that sample, with Y representing all possible labels. In a binary classification scenario, we can assume $Y = \{0, 1\}$.

Classifier outputs in the range $[1 - \epsilon_1]$ are interpreted as belonging to the positive class, while outputs in the range $[0, \epsilon_0]$ are classified as negative. Here, $\epsilon_0, \epsilon_1 \in [0, 1]$ are threshold values that need to be tuned based on the problem, and they must satisfy $\epsilon_0 + \epsilon_1 \leq 1$. Often, ϵ is set such that $\epsilon \coloneqq \epsilon_0 = 1 - \epsilon_1$, ensuring that all samples are classified into one of the two classes. Otherwise, samples with probabilities in the range $(\epsilon_0, 1 - \epsilon_1)$ may remain unclassified.

Let us examine the various uncertainty strategies that can be employed. Initially, we will focus on selecting a single sample, which corresponds to the case where $U^* = \{x^*\},\ x^* \in U$ We will then generalize this approach to handle larger subsets. The notation used in this section follows that established in Settles (2009); Schröder et al. (2021); Gilabert et al. (2023); Bardají et al. (2024).

Least Confidence

The first and most direct approach is to take the predictions of the classifier as a measure of uncertainty. This approach uses the probability of the classifier aforementioned, $P(y|x)$, as the unique value to compute the metric. It measures the distance from the most confident prediction to the maximum score, 1. The formula for obtaining the least confident sample

is:

$$x^* = \arg\max_{x \in U} \left(1 - \max_{y \in Y} P(y|x)\right) = \arg\min_{x \in U} \left(\max_{y \in Y} P(y|x)\right) \tag{2.5}$$

Margin of Confidence

Following the same idea as the least confidence, this strategy measures the difference between the top two predictions for each sample. Samples with a small distance between these top two predictions are samples the classifier is unsure how to classify. The expression that satisfies the sample with the largest margin of confidence is:

$$x^* = \arg\max_{x \in U} \Big(1 - (P(y_m|x) - P(y_p|x))\Big) = \arg\min_{x \in U} \big(P(y_m|x) - P(y_p|x)\big) \tag{2.6}$$

where $y_m = \arg\max_{y \in Y} P(y|x)$ and $y_p = \arg\max_{y \in Y \setminus y_m} P(y|x)$.

Ratio of Confidence

Following the exact same idea of the previous metric, ratio of confidence measures the relation between the top two predictions. The expression is:

$$x^* = \arg\max_{x \in U} \frac{P(y_p|x)}{P(y_m|x)} \tag{2.7}$$

where $y_m = \arg\max_{y \in Y} P(y|x)$ and $y_p = \arg\max_{y \in Y \setminus y_m} P(y|x)$.

Entropy

Shannon Entropy (Shannon, 1948) is one of the most used metrics in AL. Entropy is a measure of uncertainty with an interesting property: the uniform distribution have maximum entropy. This means that a random classifier would exhibit maximum uncertainty and would have lower values when the classifier is improving classification. This method computes the sample with highest entropy, considering all possible labels.

$$x^* = \arg\max_{x \in U} \left(-\sum_{y \in Y} P(y|x) \log P(y|x)\right) \tag{2.8}$$

The following are some of the most important diversity sampling strategies.

2.4.3 Diversity Sampling

In the previous section, metrics were defined in an "exact" manner, meaning they rely on closed formulas that can be systematically applied to select the next sample. In contrast, diversity sampling aims to capture more complex properties such as coverage, diversity, or representativeness.

To achieve this, the metrics discussed in this section utilize embeddings produced by a model M trained on the labeled data. By analyzing the spatial distribution of these embeddings, the goal is to identify samples that are either the most dissimilar from the majority of other samples or that exhibit other desirable properties.

For embedding-based metrics, we assume that the classification model M can be expressed as $M = M_C \circ M_E$ where M_E is a function that transforms data from the original space into a vector representation, and M_C takes this vector and classifies it into one of the possible classes. This architecture is typical in many CNNs models, where M_E corresponds to the convolutional layers that extract features, and M_C includes the final dense layers that perform classification. Using this notation, $M_E(x)$ represents the embedding vector of a sample x.

Contrastive

This method, presented in Margatina et al. (2021), selects samples with the largest Kullback-Leibler (KL) divergenece between the predictions and its k nearest neighbors.

$$x^* = \underset{x_i \in U}{\arg\max} \left(\frac{1}{k} \sum_{j=1}^{k} \mathrm{KL}\big(P(\cdot|x_j^{nn}) \;||\; P(\cdot|x_i)\big) \right) \tag{2.9}$$

In the previous equation $x_j^{nn},\ j = 1, \ldots, k$ are the k nearest neighbors of sample x_i. Given two discrete probability distributions, A, B, that share the same sampling space X, the KL divergence is defined as:

$$\mathrm{KL}(A \;||\; B) = \sum_{x \in X} A(x) \log \left(\frac{A(x)}{B(x)} \right) \tag{2.10}$$

Although this strategy may appear to be an uncertainty strategy due to its closed-form nature, an important aspect that classifies it under diversity sampling is the computation of nearest neighbors. In the original paper, the nearest neighbors are retrieved by selecting the top k nearest samples using euclidean distance between their embeddings.

Coverage

The coverage strategy calculates the distances from each unlabelled sample to its k nearest neighbors and then computes the average distance. The sample selected by the coverage strategy is the one that is farthest from its neighbors.

$$x^* = \arg\max_{x \in U} \left(\frac{1}{k} \sum_{x' \in \mathrm{NN}^k_{d,L}(x)} d\left(M_E(x'), M_E(x)\right) \right) \tag{2.11}$$

where $\mathrm{NN}^k_{d,L}(x)$ denotes the k nearest neighbors of x among the already labeled samples, L, using the distance metric, d.

Cloud-Clustering

The problem with averaging of the coverage strategy is that it can loose some important information. The mean of a collection of numbers doesn't retain information of the original collection and this was addressed in this new metric. To this end, instead of considering the mean of the distances, the volume of the embeddings is considered. The Convex-Hull of a set of embeddings is defined as the smallest convex volume that encloses all the embeddings. This is why it is also called an enclosure of the embedding set.

To develop this method, it is first necessary to cluster the samples into k groups. This method will then select the cluster that overlaps the least with the other clusters, extending the idea presented in the coverage method. In the following formula, the cluster with the smallest intersection with the others is the one selected.

$$U^* = \arg\min_{V \in \mathrm{Clust}_k(U)} \sum_{\substack{W \in \mathrm{Clust}_k(U) \\ W \neq V}} \mathrm{Volume}\big(\text{Convex-Hull}(V) \cap \text{Convex-Hull}(W)\big) \tag{2.12}$$

In this equation, $\mathrm{Clust}_k(U)$ is a function that clusters all the data in the unlabelled set into k groups, and $\mathrm{Volume}(A)$ computes the volume the set A occupies.

2.5 Building colon capsule endoscopy datasets

While CE is a valuable diagnostic tool, reviewing the videos captured by the capsule can be a long and tedious process (Koulaouzidis et al., 2021; Maieron et al., 2004; Rondonotti et al., 2020). Each video contains thousands of images that must be carefully analyzed by trained personnel to detect any abnormalities or signs of disease. The process of reviewing these videos requires a great deal of concentration and attention to detail, as the images are often small and difficult to interpret. Additionally, it is worth noting that the process of labeling videos to create a dataset is a significant challenge, primarily due to the high cost of hiring annotators.

To fully leverage the potential of DL technology for CE, a vast amount of labeled data is required to train the algorithms to be robust and reliable in real-world clinical settings. It is essential that the videos used to train these models are diverse and representative of the full range of conditions encountered in clinical practice. By including a high diversity of images and capturing real-world scenarios, DL models that emerge from this training can be better equipped to predict and identify abnormalities that are present in actual patients.

In the studies by Gilabert et al. (2023); Bardají et al. (2024), AL frameworks were developed to build a comprehensive dataset of CE videos. The main objective was to create this dataset iteratively, starting with a small, labeled subset of videos. The framework then determined which additional videos required labeling in subsequent iterations.

This selection process focused on diversity scenarios, considering the embeddings of the video frames. Embeddings represent the key features of the video images and are used to measure the similarity between different videos. By computing the distances between these embeddings using specific strategies outlined in the study, the framework identified which new videos to label next, based on their potential to add valuable information to the dataset. Further details on the methodology will be provided in the following sections.

2.5.1 Dataset

The dataset used for the experiments consists of 23 Colon Capsule Endoscopy (CCE) videos from different patients obtained with PillCam SB3 at Raigmore Hospital, Inverness (Scotland). From each video, 2,000 images were randomly extracted and a medical doctor with more than 10 years of experience classified them into one of the following six classes of interest: Bubbles, Turbid, Large Blob, Wall, Wrinkles, and Undefined. As can be seen in Table 2.2, the dataset is very unbalanced. The “Undefined” class corresponds to all those images that have no medical relevance and, therefore, is the majority class with almost 42% of the images.

Figure 2.14 shows three images per class of a single patient, showing the diversity of the dataset.

	Bubbles	Turbid	LargeBlob	Wall	Wrinkles	Undefined
# images	3,073	12,497	2,957	3,020	5,226	19,078
(%)	6.70	27.26	6.45	6.59	11.40	41.61

Table 2.2: Number of images of each class in the dataset.

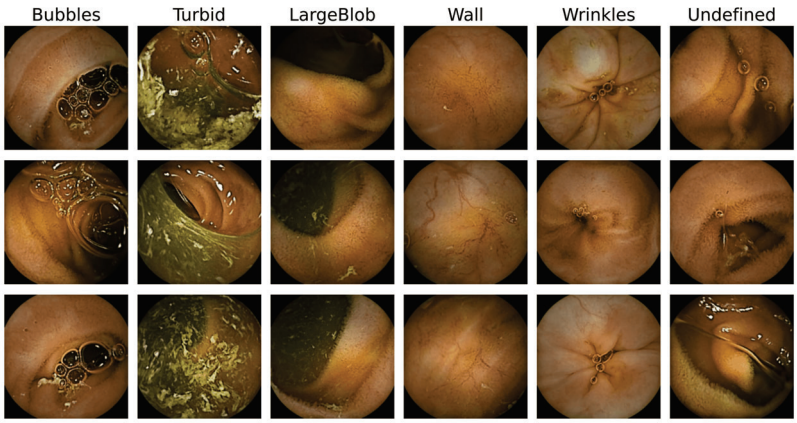


Figure 2.14: Random images of a patient with their labels grouped in the six classes.

2.5.2 Experimental Setup

Experiments were performed using an NVIDIA 3090 GTX running Tensorflow 2.8 and CUDA 11.6. An EfficientNetB3 (Tan and Le, 2019) pre-trained with ImageNet was selected as a backbone model for all the strategies. An additional dense layer was applied at the end of the backbone to generate an embedding space of $n = 64$ dimensions. All models were trained for 20 epochs with 64 images per batch. Adam optimizer with a learning rate of 1×10^{-5} and weighted cross-entropy as a loss function were used to train the model. The loss weights were computed using the balanced strategy of *sklearn* that computes the number of samples of each class in the training set (currently labeled set, L). All the models were trained using the same learning rate decay policy: during the 5 first epochs, the learning rate was fixed at 1×10^{-5} and linearly decreased after that until 1×10^{-7}. When computing nearest neighbors for some of the strategies, the number of neighbors was set to $k = 30$. Since the "Undefined" class is not medically relevant, it was not taken into account when computing the diversity metrics.

The video split was performed as follows: from the total of 23 videos, 5 videos were randomly selected as a validation test, 5 as the test set, 1 video as the initially labeled set of videos, and 12 videos as the initially unlabeled set of videos. A total of 8 videos were selected using the active learning algorithm with the three compared strategies. At each iteration of the process, the model was trained on the labeled set and the metrics of the 20th epoch were reported.

This same process was repeated three times to ensure its robustness. The results presented are the average of these three trials.

2.5.3 Active learning strategies

Let M be a model trained with some labeled videos, L. Let U be a set of unlabeled videos. The goal is to find, at each step, the unlabeled video $V^* \in U$ that adds the most diversity possible to the training set. That is:

$$V^* = \arg\max_{V \in U} d(M, V, L) \tag{2.13}$$

where $d(M, V, L)$ is a function that scores a video V using the model M, based on those already present in the training set, L. This is the video that is sent to be labeled by the expert and added to the labeled pool of videos. The way we define $d(\cdot)$ is crucial since it is the metric that changes the way videos are selected.

The active learning process can be implemented following the algorithm presented in Algorithm 1. We start with a set (small) of labeled videos and a set (large) of unlabeled videos. Then, the video that maximizes $d(\cdot)$ is selected. This video is then added to the labeled pool of videos and removed from the pool of unlabeled videos. We repeat this operation until there are no more unlabeled videos.

Algorithm 1 Active learning video selection

Require: Unlabeled videos, U_0; labeled videos, L_0; model, M
  $U \leftarrow U_0$
  $L \leftarrow L_0$
  while $U \neq \emptyset$ **do**
    $M \leftarrow \text{train}(M, L)$ ▷ Train the model on labeled data.
    $V^* \leftarrow \arg\max_{V \in U} d(M, V, L)$
    $L \leftarrow L \cup V^*$
    $U \leftarrow U \setminus V^*$
  end while

As we have seen in Sections 2.4.2 and 2.4.3, the function $d(\cdot)$ relies on different properties of the classifier or the embedding space to compute the score for each of the samples to be selected and the one with the highest value is, indeed, sent to be labeled to the experts. In the following sections we present different approaches to tackle this problem, providing results for the CE dataset. All the strategies are reformulated to adapt them to the specific case. P_M is used to express that the model M induces a probability on the classification.

Random Strategy

To compare any method, the random strategy is defined as a discrete uniform distribution on the videos. Given $|U| = n$ with $V_1, \ldots, V_n \in U$ being the collection of unlabeled videos, we

define the random distance of each video as $d_{\text{rand}}(M, V, L) = \frac{1}{n}$. In other words, the random strategy assigns each video an equal probability of being selected, effectively following a uniform distribution:

$$V^* \sim \text{Uniform}\{V_1, \ldots, V_n\}, \; n = |U| \tag{2.14}$$

Greedy Strategy

An upper bound on the results must be established, which is done using a brute-force approach to approximate an almost-optimal strategy. This method involves selecting the video that yields the maximum improvement in each iteration. While this approach does not guarantee optimality, since all permutations would need to be considered, it provides a good upper bound for comparison. The greedy strategy, as outlined in Algorithm 1, operates similarly but differs in that it trains a classifier for each potential video and selects the one that maximizes improvement in a chosen metric, such as the Area Under the ROC Curve (AUC) in our case.

Figure 2.15 illustrates the comparison between the random and greedy strategies. The figure shows how various metrics on the test set evolve as new videos are introduced. Initially, the system starts with a single video, and then eight additional videos are added sequentially. The strategies under consideration should ideally perform better than the random baseline (indicated by the blue line) and approach, but not exceed, the performance of the greedy strategy (indicated by the orange line). The objective is to develop methods that closely align with the performance of the greedy curve.

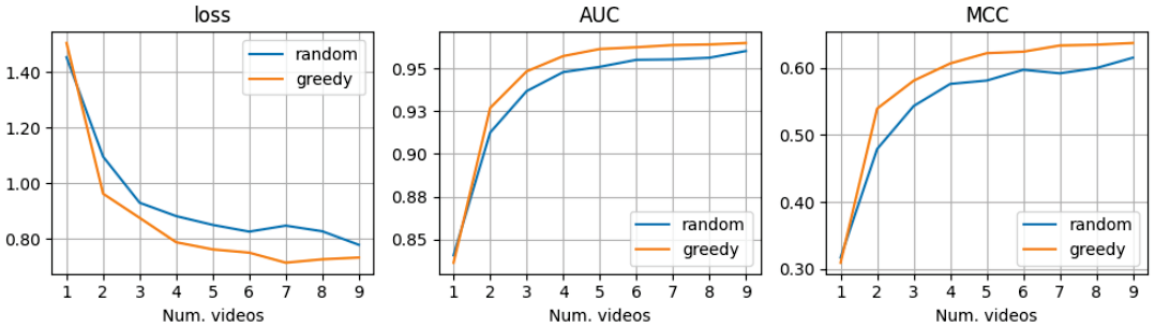


Figure 2.15: Greedy vs Random strategies evaluated on the test set. Average of three trials.

Least Confidence Strategy

The least confidence or the discriminative score for an image is the highest value after the classification layer. Since we are not selecting individual frames but entire videos, to compute this metric for a video, d_{disc} is defined as the average of the discriminative score for each frame.

This is summarized in Eq. 2.15.

$$d_{\text{disc}}(M, V, L) = -\frac{1}{|V|} \sum_{x \in V} \max_{y \in Y} P_M(y|x) \tag{2.15}$$

As can be seen in the equation, this function does not use the labeled set of videos when computing the score of one video. The video selected using this strategy is presented in Equation 2.16.

$$V^* = \arg\max_{V \in U} \left(-\frac{1}{|V|} \sum_{x \in V} \max_{y \in Y} P_M(y|x) \right) \tag{2.16}$$

Margin Strategy

The margin strategy takes into account the top two predictions for each sample. This is summarized in Eq. 2.17. As introduced in Equation 2.6, y_m and y_p represent the top-1 and top-2 prediction, respectively, that is $y_m = \arg\max_{y \in Y} P_M(y|x)$ and $y_p = \arg\max_{y \in Y \setminus y_m} P_M(y|x)$.

$$d_{\text{marg}}(M, V, L) = \frac{1}{|V|} \sum_{x \in V} \max_{y \in Y} \left(1 - \left(P_M(y_m|x) - P_M(y_p|x)\right)\right) \tag{2.17}$$

The video selected using this strategy is presented in Equation 2.18.

$$V^* = \arg\max_{V \in U} \left(\frac{1}{|V|} \sum_{x \in V} \max_{y \in Y} \left(1 - \left(P_M(y_m|x) - P_M(y_p|x)\right)\right) \right) \tag{2.18}$$

Entropy Strategy

Shannon's Entropy can be used to select the most informative video. The distance that this strategy uses is presented in Equation 2.19.

$$d_{\text{entrop}}(M, V, L) = \frac{1}{|V|} \sum_{x \in V} \left(-\sum_{y \in Y} P_M(y|x) \log P_M(y|x) \right) \tag{2.19}$$

Then, choosing the video with the highest entropy leads to the formula presented in Equation 2.20.

$$V^* = \arg\max_{V \in U} \left(\frac{1}{|V|} \sum_{x \in V} \left(-\sum_{y \in Y} P_M(y|x) \log P_M(y|x) \right) \right) \tag{2.20}$$

The Least Confidence, Margin, and Entropy strategies do not always yield the best results, as demonstrated for this particular dataset. Figure 2.16 illustrates the evolution of

loss, AUC, and Matthews Correlation Coefficient (MCC) as new videos are incorporated into the dataset. For this dataset, these strategies either perform worse than or are comparable to the random strategy.

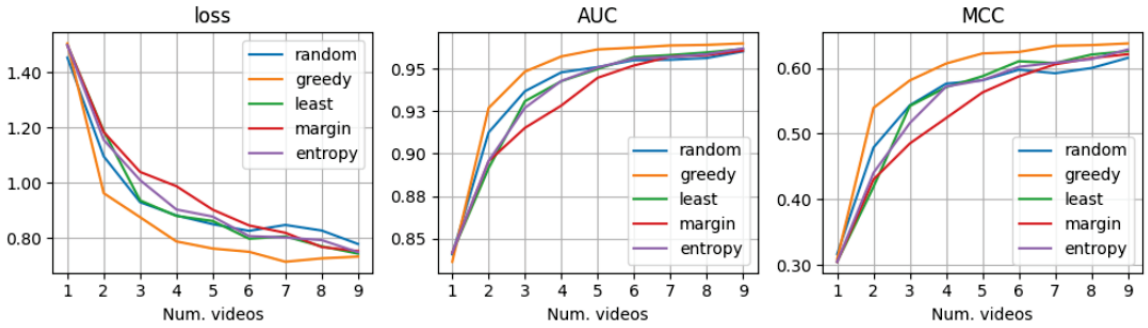


Figure 2.16: Uncertainty strategies evaluated on the test set. The curves show that these strategies perform worse than the random strategy for this specific dataset.

Maximal Coverage Strategy

To better leverage the information on the latent space of the model M, i.e, the embedding before the last classification layer, we designed a metric that computes the distance from one video candidate to the current labeled set of videos. In this manner, the video with the highest distance is expected to be the one that contributes the most to global diversity. As we did with the previous metrics, we compute the distance of a video to the rest of the labeled videos by averaging the value calculated per each individual frame.

Our model M can be decomposed into two sub-models $M = M_C \circ M_E$ where M_E gets an image and projects it to an embedding space $\mathbb{R}^n$ and M_C is the sub-model responsible of classifying this embedding into the C classes of the problem. This means applying a last layer and an activation function, usually a softmax to transform the output into a probability in the 0-1 range.

To compute the distance of each frame x of the candidate video V to the already labeled videos L, we first compute the k nearest embeddings per frame. That is, $\{x' \mid x' \in \mathrm{NN}_L^k(x)\}$. We then compute the euclidean distance from each neighbor x' to the current frame x and take the mean of these values. This number provides a distance from a video candidate to the set of currently labeled videos.

Figure 2.17 and Eq. 2.21 summarize the procedure explained above.

$$d_{\text{cover}}(M, V, L) = \frac{1}{|V|} \sum_{x \in V} \left(\frac{1}{k} \sum_{x' \in \mathrm{NN}_L^k(x)} ||M_E(x') - M_E(x)||_2 \right) \tag{2.21}$$

Then, to select the video:

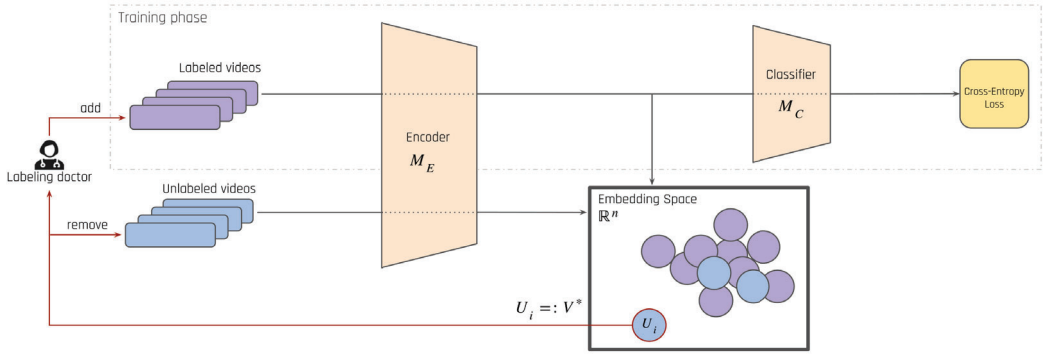


Figure 2.17: Active learning framework using the maximal coverage distance. The algorithm iterates through two steps: initially training the model on labeled data using cross-entropy loss, then selecting the optimal video for addition to the training set using the maximal coverage distance function.

$$V^* = \arg\max_{V \in U} \left(\frac{1}{|V|} \sum_{x \in V} \left(\frac{1}{k} \sum_{x' \in \mathrm{NN}_L^k(x)} ||M_E(x') - M_E(x)||_2 \right) \right) \tag{2.22}$$

Figure 2.18 shows an example of how the system works. First, the labeled set of data is considered. Then, for each video candidate, the embeddings of their images are computed. The video with the highest distance to the labeled set is selected as the new video to be labeled. In the figure we can observe how the first video depicted, V_1, contains some images "far" from the distribution of the labeled data. This is the reason why this video has a higher distance and would be selected among these two videos.

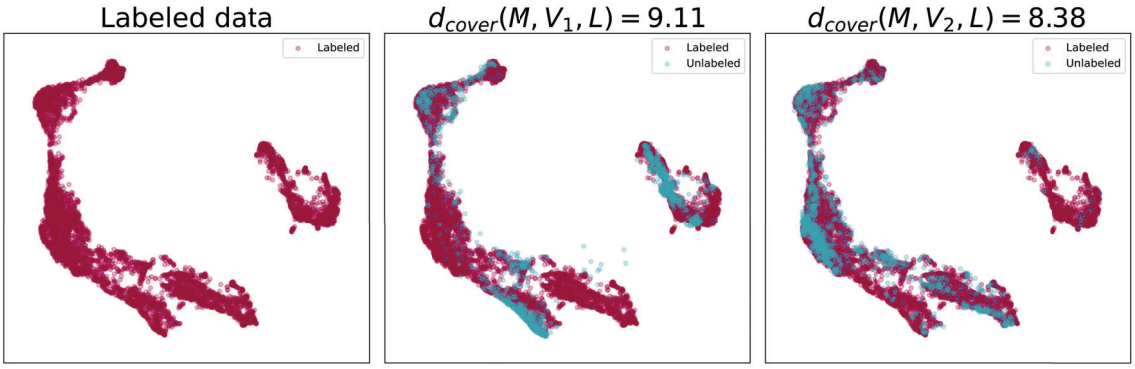


Figure 2.18: UMAP projection to 2-coordinates of the embeddings. Left: Embeddings of the labeled set. Center, Right: Embeddings of two unlabeled videos. The distance to the labeled set is computed to choose among the two videos, choosing V_1 in this example since it has a larger distance of 9.11.

An additional experiment was conducted to stabilize the embedding space, as the embeddings of the frames tend to change when a new video is introduced. In the initial stages

of the process, when only one or two videos are used for training, these embeddings cannot be fully trusted because they can be highly biased by the content of the frames from these initial videos.

To enhance the embedding representation of the videos, an autoencoder was employed as an unsupervised method, trained with all the available training data without labels. The autoencoder consists of an encoder and a decoder. The encoder processes data from the original space (images, in this case) and outputs an embedding, which in our case is a vector of 256 values. This vector is then processed by the decoder, which applies upsampling and deconvolutions to restore the image to its original size. These networks are trained using reconstruction losses that compare the similarity between input and output, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), or Categorical Cross-Entropy (CatCE). In this experiment, cross-entropy was used.

Figure 2.19 depicts the loss, AUC and MCC metrics of the model at each iteration of the active learning algorithm, evaluated on the test set. It is evident that enhancing the cover strategy with the autoencoder, incorporates significant videos at each step, approximating the greedy curve.

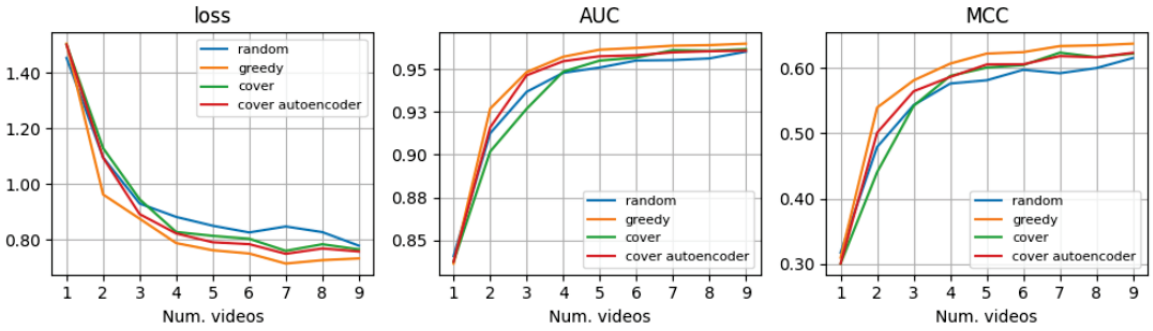


Figure 2.19: Cover strategies evaluated on the test set. By using an autoencoder as the embedding space, the system improves classification.

Selecting the appropriate video is what enables the system to continue improving without stalling. Figure 2.20 displays four of the top 10 most diverse images from the first three selected videos. These images contribute the most diversity to the embedding space and are responsible for the selection of this particular video over other candidates.

As depicted in the figure, the system identifies various types of outliers at each stage, i.e., images that are far from the labeled set. This is vital information that cannot be obtained through conventional methods such as the *disc* strategy, which solely relies on the classifiers' confidence without considering the similarity (or dissimilarity) of the images. In summary, our approach not only selects the most informative videos for experts to review but also identifies the critical images within those videos that contribute to their selection.

One of the main problems of this method is the embedding aggregation. That is, the video is represented by the mean embedding of all its frames. This leads to problems when

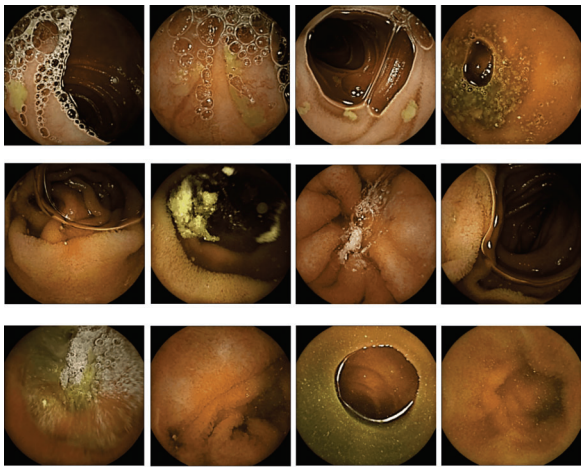


Figure 2.20: These images are located far from the distribution of the labeled set. The first row represents the first video that was selected, while the second and third rows correspond to the second and third videos, respectively.

the data is not uniform among the embedding space as it is almost always the case. Next proposed strategies tackle these problems to achieve better results.

Cloud Strategy

Instead of considering the average embedding to represent a video, the Convex Hull of all its frames was considered. Following the definition from Equation 2.12, this strategy can be defined as:

$$V^* = \arg\min_{V \in U} \left(\sum_{W \in U \setminus V} \text{Volume}\big(\text{Convex-Hull}(V) \cap \text{Convex-Hull}(W)\big) \right) \tag{2.23}$$

Note that in the original equation, Equation 2.12, images needed to be previously clusterized. Now we use the label of the video as a cluster identifier so, instead of selecting one cluster we are selecting one video.

The most important part of this strategy is the embedding computation. Since embeddings were large vectors, to compute the Convex-Hull they were previously projected into a lower dimension space using two distinct strategies: Principal Component Analysis (PCA) (Pearson, 1901) and UMAP (McInnes et al., 2018).

Moreover, two distinct approaches were implemented to compute embeddings. First, the strategy inherited by the cover strategy where the embedding of a frame is the vector representation output by the first part of the classifier itself. Again, if the classifier is the model $M = M_C \circ M_E$, the embedding of a sample $x \in X$ is $M_E(x)$. The reliability of these embeddings is questionable at the first steps of the process, that is, when the model

is trained with very few examples.

Second, to improve this method, an autoencoder was used to learn the distribution of all the data in an unsupervised manner and generate embeddings that do not change during the full process, following the exact same procedure as for the cover strategy.

Figure 2.21 shows the results of this strategy, displaying four of the possible combinations. Curves with the word "cloud" refers to the strategy using embeddings from the encoder, M_E. In "cloud_autoencoder", the embeddings are replaced with those from the pretrained autoencoder model. For these two strategies, the Convex-Hull is computed with the original embedding size. The other two strategies follow the same configuration, but before computing the Convex-Hull, the embeddings are projected into a 2-D representation vector using UMAP.

Although the idea seems promising, the results for this dataset did not meet expectations. Projecting the embedding space into a lower dimension resulted in the loss of some important information, and if the Convex-Hull was computed in high dimensions, the distances would not hold the expected significance. To refine this strategy, an alternative approach was considered.

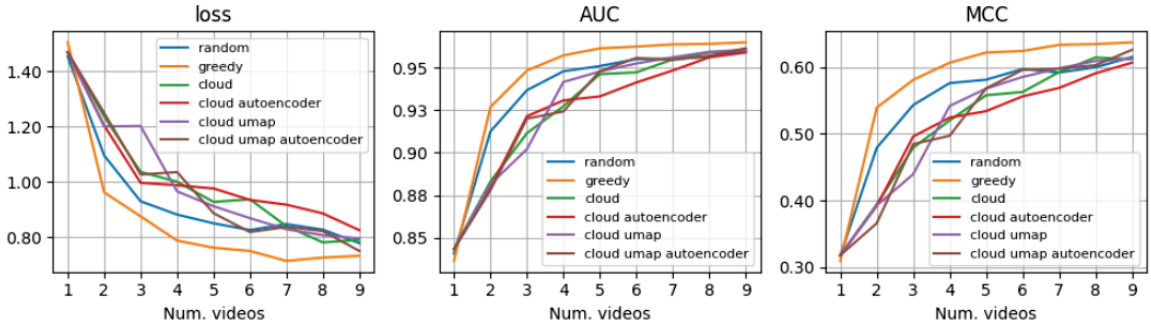


Figure 2.21: Cloud strategies evaluated on the test set. All of them produced strategies that did not meet the expected results.

Clustering Strategy

Lastly, multiple clustering strategies were tested. Following with the autoencoder setting, an autoencoder was learnt with all the unlabeled data and the embdeding of each frame was fixed. These embeddings were used during the full active-learning procedure to represent each frames.

Then, the embeddings of all frames were clusterized into k clusters, $C = \{C_1, \ldots, C_k\}$ using two methods: k-means and gaussian mixture. For each of them, the following three strategies were considered.

- **Entropy**: Following the idea of the entropy strategy introduced in the beginning of this section, the entropy of each cluster is computed. $V_{|C_j}$ represents the restriction

of video V that belongs to cluster C_j. It can also be viewed as the intersection $V_{|C_j} = V \cap C_j$.

$$V^* = \arg\max_{V \in U} \left(- \sum_{C_j \in C} \frac{|V_{|C_j}|}{|V|} \log \left(\frac{|V_{|C_j}|}{|V|} \right) \right) \tag{2.24}$$

- **Gini**: Gini impurity is a measure of the quality of the cluster distribution. It considers the probability that a random sample in the dataset is classified incorrectly, acording to the dataset's class distribution. Therefore, a perfect classification would raise 0 gini impurity and a dataset with just one class would raise 1.

$$V^* = \arg\max_{V \in U} \left(\sum_{C_j \in C} \left(\frac{|V_{|C_j}|}{|V|} \right) \left(1 - \frac{|V_{|C_j}|}{|V|} \right) \right) \tag{2.25}$$

- **Weighted Log**: Inspired by the previous strategies, a new method was developed.

$$V^* = \arg\max_{V \in U} \left(\sum_{C_j \in C} w_j \log \left(\frac{1 + |V_{|C_j}|}{|C_j|} \right) \right), \text{ where } w_j = 1 - \frac{|C_j \cap L|}{|C_j|}. \tag{2.26}$$

These strategies performed the best for this specific dataset. The use of the autoencoder and the predefined clusters produced three stable strategies that performed better during the first iterations were is more difficult to choose videos. Figure 2.22 shows the metrics for these three strategies. It can be observed how these three curves are the closest to the greedy strategy, comparing them with the other strategies from previous sections.

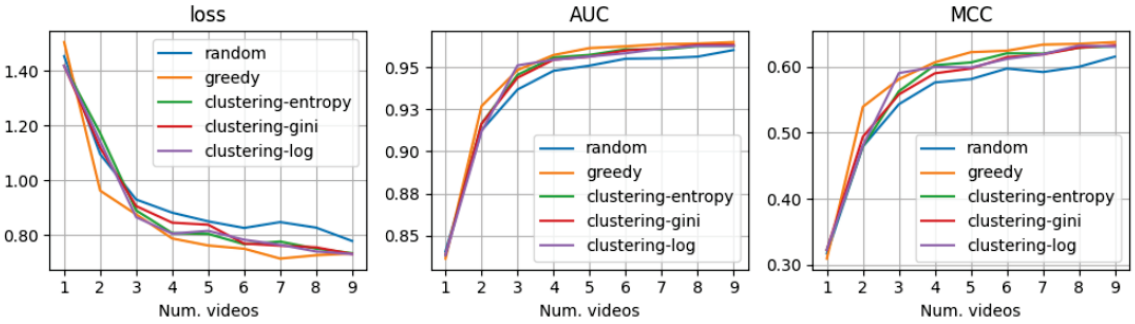


Figure 2.22: Clustering strategies evaluated on the test set using the k-means algorithm. These three strategies are very close to the greedy strategy, indicating that they are among the best performers.

Table 2.3 summarizes all the implemented strategies. It can be observed that clustering strategies perform better than all the other ones.

Although the primary focus of this thesis is on solutions for CE, the methods and algorithms developed have broad applicability across various medical domains.

	loss	AUC	F1	MCC	precision	recall	P@R 90	P@R 95	P@R 99	accuracy	top-3-accuracy
cloud autoencoder	0.894	0.824	0.490	0.458	0.595	0.477	0.374	0.303	0.181	0.647	0.846
cloud	0.878	0.826	0.498	0.465	0.595	0.477	0.383	0.316	0.203	0.652	0.848
cloud umap autoencoder	0.868	0.826	0.493	0.463	0.597	0.481	0.390	0.321	0.210	0.649	0.850
cloud umap	0.871	0.828	0.497	0.469	0.597	0.487	0.388	0.320	0.206	0.659	0.850
margin	0.849	0.829	0.516	0.476	0.600	0.482	0.395	0.328	0.222	0.660	0.852
entropy	0.825	0.833	0.530	0.491	0.612	0.492	0.406	0.335	0.217	0.677	0.856
least confidence	0.815	0.833	0.529	0.492	0.613	0.499	0.412	0.343	0.222	0.676	0.857
random	0.815	0.836	0.532	0.495	0.615	0.499	0.412	0.336	0.211	0.687	0.861
cover	0.796	0.835	0.535	0.498	0.624	0.506	0.420	0.347	0.232	0.680	0.857
cover autoencoder	0.778	0.840	0.541	0.509	0.623	0.511	0.434	0.362	0.238	0.691	0.864
clustering gini	0.782	0.841	0.545	0.512	0.621	0.509	0.431	0.359	0.237	0.699	0.865
clustering entropy	0.781	0.841	0.547	0.513	0.624	0.509	0.431	0.362	0.237	0.700	0.865
clustering log	0.775	0.840	0.548	0.514	0.624	0.511	0.436	0.365	0.235	0.699	**0.866**
greedy	**0.735**	**0.844**	**0.559**	**0.527**	**0.642**	**0.536**	**0.450**	**0.375**	**0.250**	**0.707**	0.865

Table 2.3: Metrics comparison: Area under the curves of test metrics during training. Rows are sorted by the average of all values. Bold values highlight the best result in each column, while underscored values indicate the second-best.

As this chapter delves into data preparation and validation, it's important to explore the intricacies involved in creating a medical dataset. This includes understanding the specifications of the images, identifying and tagging detected pathologies, and addressing the technical challenges that may arise during data collection, which can either complicate or facilitate its use in an AI model.

Once the dataset is created, properly tagged, and its value clearly understood, it must be presented to the AI community in a structured manner. This involves detailing its characteristics, demonstrating its suitability for training, testing, and validation, and providing examples of its use in predicting the relevant pathology or pathologies.

In the following pages, you will find an example of this process within the medical imaging field of respiratory diseases.

2.6 LungHist700

In this section, a novel dataset, LungHist700 (Diosdado et al., 2024a), is introduced, comprising 691 images sized 1200×1600 pixels, depicting both normal lung tissue and primary lung carcinomas. The carcinomas are classified into adenocarcinomas and squamous cell carcinomas, each further subclassified based on the degree of carcinoma differentiation into three levels: well differentiated, moderately differentiated, and poorly differentiated.

Data collection occurred at Hospital Clínico de Valladolid in 2023 as part of routine diagnostic procedures, involving 45 patients. The dataset encompasses images of hematoxylin and eosin stained samples extracted from pathology glass slides utilizing a Leica DM 2000 microscope and Leica ICC50 W microscope camera at two distinct magnifications: 20x and 40x. Pathologists meticulously selected the field of view to encompass representative tissue categories, which are typically discernible in all four quadrants of the image.

All individuals included in the study underwent surgical procedures, hence all images pertain to patients with malignancies. Images categorized as depicting normal lung areas represent regions where tumors have not proliferated.

For each patient, two concurrent evaluations were conducted to determine tumor type (adenocarcinoma or squamous cell carcinoma) and level of differentiation (well, moderately, or poorly differentiated). The first analysis involved morphological examination of tissue based on hematoxylin and eosin stained samples, establishing the classification of well and moderately differentiated samples. The second evaluation comprised immunohistochemical tests to discern tumor type (adenocarcinoma or squamous). These tests, supplemented by contextual information, contributed to accurate classification of poorly differentiated categories. The performed tests encompassed TTF1, CK7, Napsin A, P40, and CK5/6. Based on the results obtained from all tests, a specialist pathologist classified images into the 7 classes of the dataset.

For adenocarcinomas, the differentiation grading system recommended by (Butnor et al., 2009), the College of American Pathologists, was employed. According to their guidelines, there are three differentiation levels:

1. Well-differentiated: Tumors primarily exhibiting a lepidic pattern, with no high-grade components or less than 20% high-grade features (such as solid, micropapillary, or complex glandular patterns).

2. Moderately differentiated adenocarcinoma: Tumors mainly showing acinar or papillary patterns, with less than 20% high-grade features.

3. Poorly differentiated adenocarcinoma: Tumors that have 20% or more high-grade features.

Pulmonary squamous cell carcinoma has also traditionally been divided into well differentiated, moderately differentiated, and poorly differentiated, similar to squamous cell carcinomas of other organ systems. The degree of differentiation is generally dependent on a combination of features, such as the presence or absence of keratinization and intercellular bridges, as well as cellular pleomorphism and mitotic activity Weissferdt (2020). Following these guidelines, squamous cell carcinoma has been divided into the following three categories:

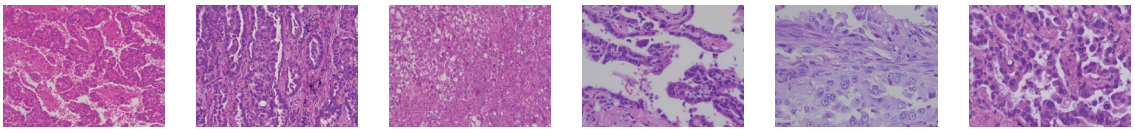
1. Well differentiated: These tumors exhibit keratinization, such as keratin pearls and intercellular bridges. They typically grow in sheets or nests, with polygonal cells that have round to oval nuclei, vesicular features, and eosinophilic cytoplasm. Additionally, mitotic figures and focal areas of hemorrhage or necrosis may be present.

2. Moderately differentiated: These tumors show increased cytologic atypia and mitotic activity. Although keratinization and intercellular bridges are still present, they are

less prominent compared to well-differentiated tumors. Moreover, areas of hemorrhage or necrosis are more common.

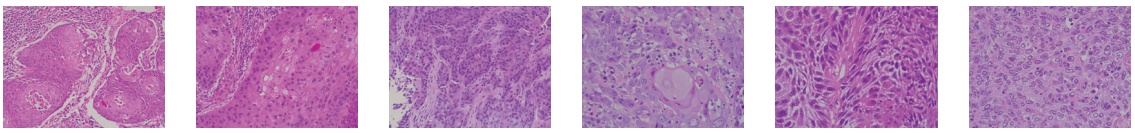
3. Poorly differentiated: These tumors grow in sheets and are often unrecognizable as squamous type without immunohistochemistry. They display significant cellular pleomorphism, high mitotic activity, and extensive areas of necrosis.

Figure 2.23 shows adenocarcinoma samples, Figure 2.24 shows squamous cell carcinoma samples at varying levels of differentiation and resolution. Figure 2.25 shows images of normal lung tissue at two different resolution.



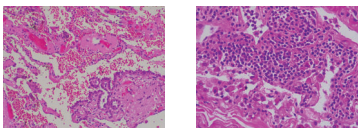
(a) Well diff. at 20x. (b) Mod. diff. at 20x (c) High diff. at 20x (d) Well diff. at 40x. (e) Mod. diff. at 40x (f) High diff. at 40x

Figure 2.23: Images displaying adenocarcinoma at varying levels of differentiation and resolution.



(a) Well diff. at 20x. (b) Mod. diff. at 20x (c) High diff. at 20x (d) Well diff. at 40x. (e) Mod. diff. at 40x (f) High diff. at 40x

Figure 2.24: Images displaying squamous cell carcinoma at varying levels of differentiation and resolution.



(a) Normal at 20x. (b) Normal at 40x

Figure 2.25: Normal lung images at different resolution.

The dataset is available at figshare Diosdado et al. (2024b). It consists of 691 images from 45 patients, with each image having a resolution of 1200×1600 pixels and stored in *.jpg* format. These images are captured at either 20x or 40x magnification levels and are categorized into seven classes (see Table 2.4). An accompanying *.csv* file links each image to the associated patient ID. All patients have been anonymized, and the file includes an identifier to match images from the same patient.

Description	Id.	20x	40x	Subclass total	Superclass total
Well differentiated adenocarcinoma	aca_bd	57	46	103	
Moderately differentiated adenocarcinoma	aca_md	44	46	90	280
Poorly differentiated adenocarcinoma	aca_pd	45	42	87	
Normal lung	nor	85	66	151	151
Well differentiated squamous cell carcinoma	scc_bd	50	49	99	
Moderately differentiated squamous cell carcinoma	scc_md	30	36	66	260
Poorly differentiated squamous cell carcinoma	scc_pd	48	47	95	
Total		**359**	**332**	**691**	**691**

Table 2.4: The dataset comprises three classes: adenocarcinoma (aca), squamous cell carcinoma (scc), and normal (nor). Images showing malignant tissue are further categorized based on their differentiation level.

Here, we present two baseline methods for classifying the dataset into the three major superclasses. First, a classic approach was employed where images were resized, and a Deep Neural Network (NN) was trained. The second method involves a Multiple Instance Learning (MIL) strategy, where patches of the images were extracted, and the same NN was used to obtain multiple embeddings, one for each patch. An attention Vaswani et al. (2017) layer was then applied to relate and aggregate these embeddings for image classification.

All the experiments used the same training configuration: the networks were implemented using Keras and executed on an NVIDIA RTX 3090 with CUDA 11.0. The NN model used in both methods was a ResNet50 network pretrained on ImageNet. The Adam optimizer was employed with an initial learning rate of 1e-5, which was reduced by a factor of 0.1 if the model began to overfit. CatCE was used as the loss function in both experiments. The Albumentations library Buslaev et al. (2020) was utilized to generate augmentations on the fly during training.

Images were classified into their superclasses: "aca" (adenocarcinoma), "scc" (squamous cell carcinoma), and "nor" (normal). The data was divided into three sets: 80% for training, 10% for validation, and the remaining 10% for testing. A patient-wise strategy was employed, ensuring that images from the same patient were placed in the same set to ensure fair evaluation and prevent data leakage.

DNN Baseline

To train the ResNet50 model, images were resized to 300×400 pixels to better fit this architecture. The published dataset, however, contains images at their original resolution (1200×1600 pixels). Figure 2.26 illustrates the learning curves on the training and validation splits, as well as the classification confusion matrix of the experiment on the test set for the 20x resolution. The model achieved an accuracy of 90%, a ROC-AUC of 98%, a precision of 92%, and a recall of 87%.

The experiment was then repeated with the same configuration but using images at 40x

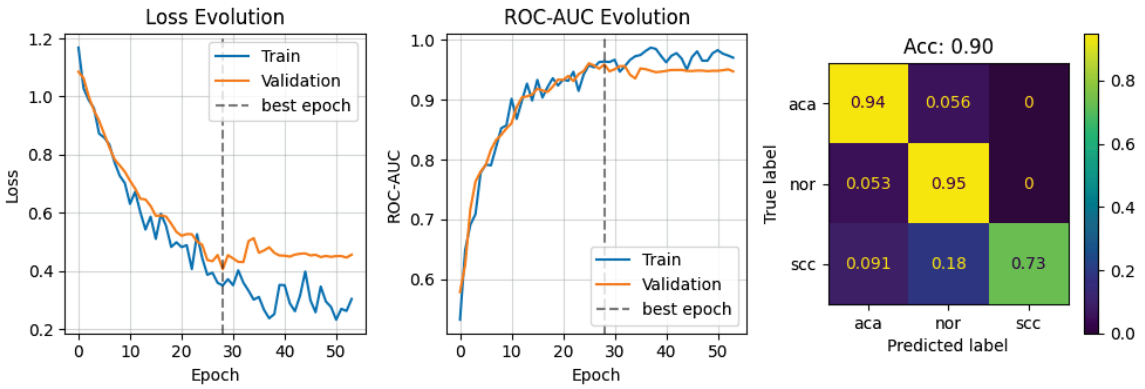
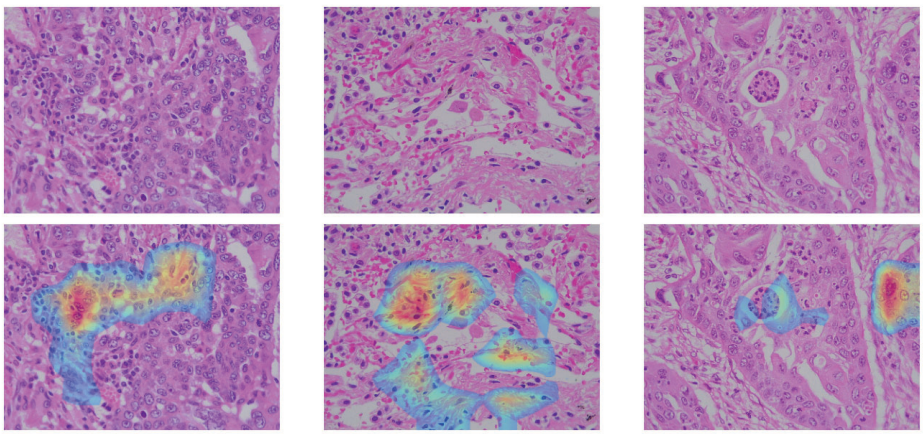


Figure 2.26: Classification results of the proposed baseline for 20x resolution. Early stopping was triggered at epoch 28, based on the validation set. After that, the best weights were loaded. The confusion matrix shows the correctly classified percentage of samples and the classification errors on the test set. The results are normalized by rows (True label).

resolution. The model achieved an accuracy of 82%, a ROC-AUC of 94%, a precision of 82%, and a recall of 84%.

To assess the validity and explainability of the results, we used Grad-CAM Selvaraju et al. (2020) on the last convolutional layer of the ResNet50 model. The threshold was set to 0.25 to visualize the Grad-CAM activations. Figure 2.27 shows the explanation masks generated by the algorithm on some test images, each representing a distinct histopathological class: adenocarcinoma, normal tissue, and squamous cell carcinoma. The masks illustrate how the model highlights specific areas relevant to image classification. The results were cross-checked with the medical team to validate the model's output.



(a) Adenocarcinoma (b) Normal lung (c) Squamous cell carcinoma

Figure 2.27: Masks generated by the Grad-CAM algorithm on some test images.

MIL Baseline

A second strategy based on ResNet50 was also tested. We trained a MIL algorithm that consisted of a ResNet50 followed by a Multi-Head Attention layer. During training, we extracted 20 random patches of size 224×224 and used the ResNet architecture to obtain embeddings for each patch. An attention layer with four heads was then applied, followed by average pooling to obtain a single embedding for classification. All the training parameters remained the same, though the batch size was reduced to three to fit within the GPU's memory constraints. The results of the MIL algorithm for images at 20x resolution are shown in Figure 2.28. This baseline model achieved an accuracy of 81%, a ROC-AUC of 89%, a precision of 80%, and a recall of 81% on the test set.

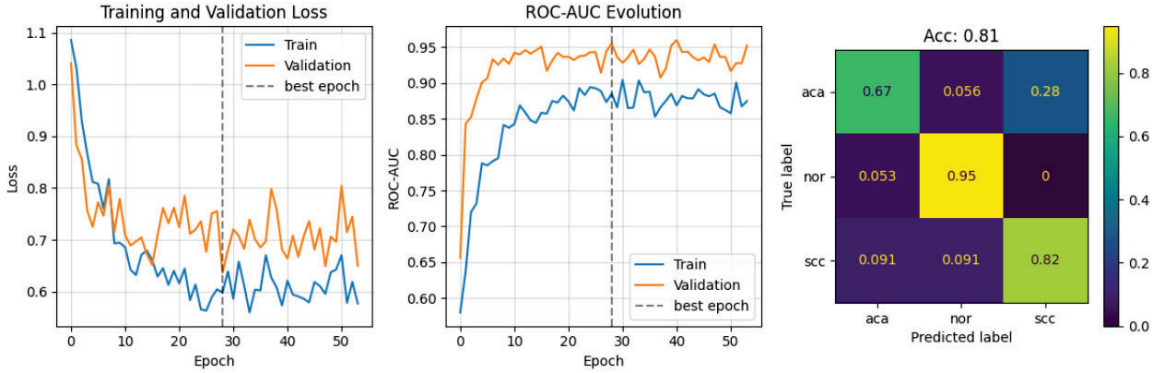


Figure 2.28: Classification performance of the MIL algorithm (ResNet50 + Multi-Head Attention layer) for 20x resolution. Early stopping was triggered at epoch 28.

2.7 Conclusions

In this chapter, we explored the critical role of datasets in the development and success of AI solutions, particularly within the field of medical imaging through CE. We began by discussing the learning process, emphasizing that high-quality datasets are indispensable for effective model training, validation, and testing. The importance of meticulous dataset design was highlighted, including strategic data splitting and ensuring diversity to prevent issues such as overfitting and data leakage.

We then delved into the fundamentals of DL, illustrating how deep neural networks, with their layered architectures, autonomously learn and extract complex patterns from raw data. This section also addressed the role of embeddings, which transform high-dimensional data into more manageable and insightful vector representations—an essential technique in applications such as NLP and image recognition.

A significant portion of the chapter focused on the integration of AL into the dataset creation and model training processes. Active learning was presented as a strategy to enhance the efficiency and effectiveness of model training by selectively querying the most

informative data points for labeling. This approach is particularly valuable in the context of CE, where acquiring labeled data is often labor-intensive and costly. By concentrating on uncertain or ambiguous cases, active learning allows the model to improve its performance with fewer labeled examples, accelerating the training process and helping to create more balanced and representative datasets—crucial for ensuring the model's ability to generalize to real-world clinical scenarios.

The chapter also discussed a collaboration with the Hospital Clínico Universitario de Valladolid, where we assisted in the creation of a medical imaging dataset, particularly focusing on lung carcinomas. An AI application was developed alongside the dataset to demonstrate its practical utility in clinical settings.

In summary, this chapter has underscored the essential role that data plays in any AI solution. By recognizing the importance of robust datasets, advanced learning techniques, and efficient labeling strategies, we are better equipped to develop AI models that can aid healthcare professionals in making accurate and timely diagnoses. The insights gained here lay a strong foundation for the more specialized discussions in the chapters to follow.

Chapter 3

Bowel Preparation Assessment

Contents

Analyzing a CE video is a complex task. The initial step, before utilizing the video for any diagnostic or research purposes, is to evaluate its admissibility. Videos may be deemed invalid if patients retain GI content, as this can obscure the view of the intestinal mucosa. In this chapter, we present a contribution from this thesis that introduces a method for determining an admissibility score for such videos.

3.1 Bowel Preparation

Performing a CCE involves several important steps to ensure a successful examination. The process begins with bowel preparation, which is crucial for optimal visualization of the GI tract. Bowel preparation typically involves a liquid diet before the procedure and an 8-hour fast (Song et al., 2013). In addition, patients may be required to take other medications, such as purgative agents, most commonly Polyethylene Glycol (PEG)(Xavier et al., 2019; Koornstra, 2009; Bjoersum-Meyer et al., 2021), to cleanse the intestines and remove any residual debris or stool.

Despite this preparation some individuals may not effectively purge their GI tract however, leading to the presence of residual intraluminal content. This residue can limit visibility within the intestines, concealing potentially significant findings, which may be overlooked by physicians reviewing the procedure. Videos from such procedures become ineffective for patient diagnosis. Consequently, a repeat often becomes necessary, which involves adjustments in bowel preparation to improve cleansing or the switch to a different diagnostic test such as colonoscopy or CT colonography.

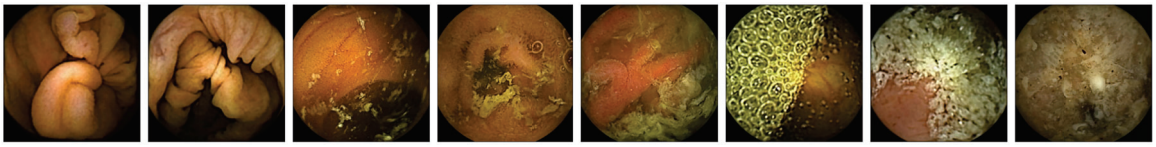


Figure 3.1: Images with varying levels of GI content. Bubbles and debris can hide the mucosa and, therefore, make the video unusable.

An accurate assessment of the cleanliness score immediately after the procedure can accelerate this process. Establishing the level of cleanliness of these videos enables the prompt identification of procedures that require repeating, thus alleviating the workload of physicians and ensuring a more efficient diagnostic workflow.

As stated in Chapter 2, training an AI model usually requires a substantial amount of data to achieve the desired results. Obtaining and curating these data is a process that can be quite laborious, demanding significant time and dedication to build a solid training database. Specifically in the context of image segmentation, there is an added layer of complexity. Beyond merely categorizing individual images, one needs to manually outline the area of interest in each frame, making the overall process slower. In the problem at hand, if the aim is to create an AI model that computes the visibility of an image, there are two potential approaches: treating it as a classification task, where each image is mapped to a predefined scale of cleansing, or as a segmentation task, predicting the visible mucosa area within each frame.

The method presented is based on the TransUNet (Chen et al., 2021b), one method that combines the U-Net (Ronneberger et al., 2015) architecture with Vision Transformer (ViT) (Dosovitskiy et al., 2021) blocks. To fully understand this architecture, let's revisit

the most important aspects of image segmentation and the architectures that exist in this domain.

3.2 Image Segmentation

Image segmentation is one of the most important techniques in computer vision. It involves partitioning images into regions to simplify their representation and make them more meaningful for analysis. This process transforms images into rich sources of data, enabling various applications. By dividing an image into distinct regions, each representing different objects or parts of objects, segmentation enhances the interpretability and utility of image data for subsequent processing.

In the medical domain, image segmentation offers critical support for diagnostics, treatment planning, and research. For instance, in radiology, segmentation of medical images such as MRI, CT, and X-rays is essential for identifying and delineating various anatomical structures. This capability allows radiologists to accurately diagnose and monitor medical conditions. In brain MRI, segmentation helps differentiate between gray matter, white matter, and cerebrospinal fluid, which is vital for diagnosing neurological disorders like multiple sclerosis, Alzheimer's or detecting brain tumors (Withey and Koles, 2007).

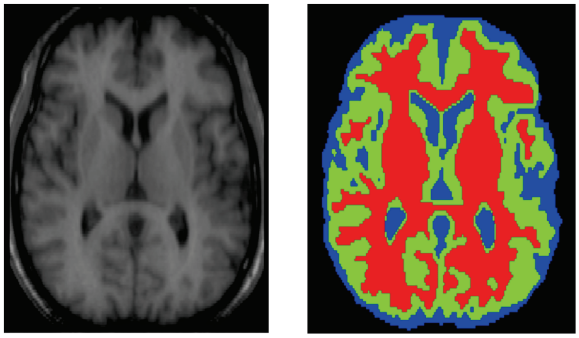


Figure 3.2: Example of image segmentation. Relevant regions are manually highlighted, creating a mask (right). In this example, different types of tissue are differentiated: white matter (red), gray matter (green) and cerebrospinal fluid (blue). Image from Withey and Koles (2007).

From a data science perspective, advances in deep learning have revolutionized image segmentation. Convolutional Neural Network (CNN) have markedly improved the accuracy and efficiency of segmentation tasks. Architectures such as U-Net (Ronneberger et al., 2015), designed initially specifically for biomedical image segmentation, have become widely adopted due to their ability to produce precise segmentations with limited training data and extend to other computer vision fields. These models leverage large annotated datasets to learn intricate patterns and structures within medical images, enabling automatic segmentation with high accuracy.

Figure 3.3 shows examples of different anatomical organs and their segmented structures. As can be seen, this task is quite challenging for most of them.

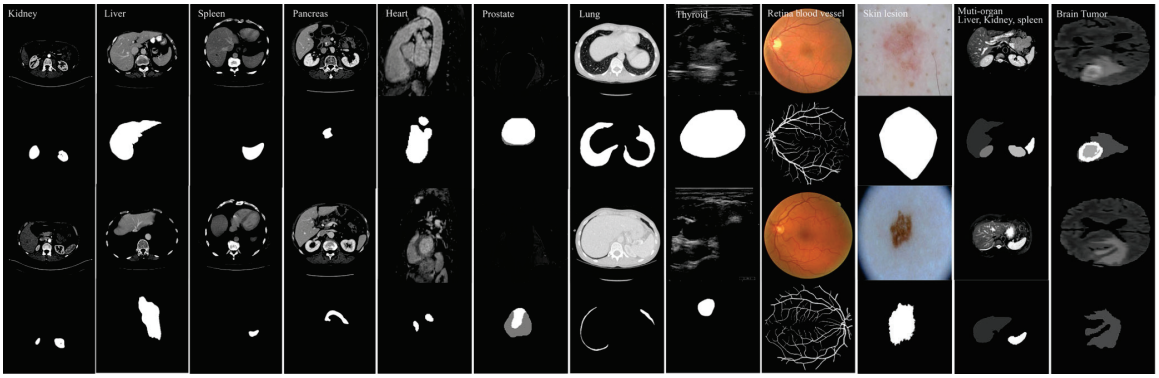


Figure 3.3: Segmentation examples on medical images. Image from Wang et al. (2022).

3.2.1 U-Net architecture

The U-Net is a CNN architecture that has become a benchmark in the field of biomedical image segmentation. Introduced by Ronneberger et al. (2015), the U-Net is specifically designed to handle the complex and often minute details present in medical images. Its architecture consists of a contracting path that aggregates pixels into a lower dimensional representation and a symmetric expanding path that reconstructs the image, forming a U-shaped structure. The contracting path or encoder applies successive convolutions and max-pooling operations to downsample the input image, while the expanding path or decoder uses upsampling and concatenation with high-resolution features, the horizontal paths between the encoder and the decoder to restore spatial information. This design allows the U-Net to produce high-resolution segmentations.

The U-Net and its variants, have proven particularly effective in segmenting various medical images, including those of the brain (M. Gab Allah et al., 2023), liver (Manjunath and Kwadiki, 2022), and retina (Zhao et al., 2020), making it an invaluable tool for diagnostics, treatment planning, and medical research. Figure 3.4 shows the architecture of the network.

3.2.2 Transformer

The Transformer architecture introduced in Vaswani et al. (2017) revolutionized the NLP field and is now being applied to other domains, such as medical imaging. Figure 3.5 displays its main architecture. The most important part of the Transformer, and the part that is crucial to understand for the following sections, is the Transformer block with its attention mechanism. The attention operation is the core of the Transformer. It consists of a dot product of vectors and a non-linear function that relates all the inputs together,

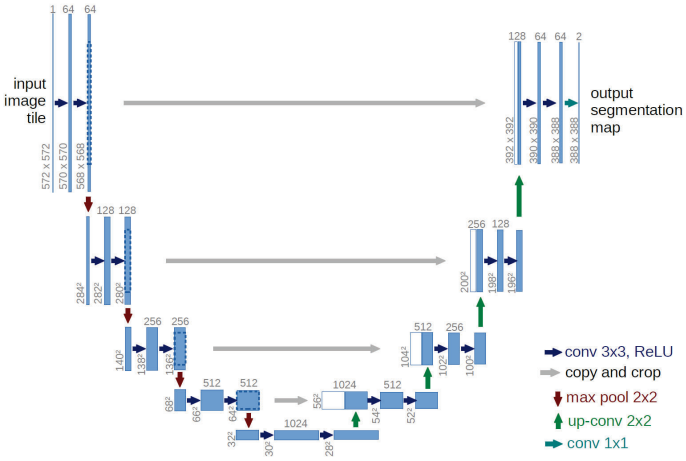


Figure 3.4: U-Net architecture. Image from Ronneberger et al. (2015).

providing scores that weight their importance. In the following paragraph, this mechanism is briefly introduced.

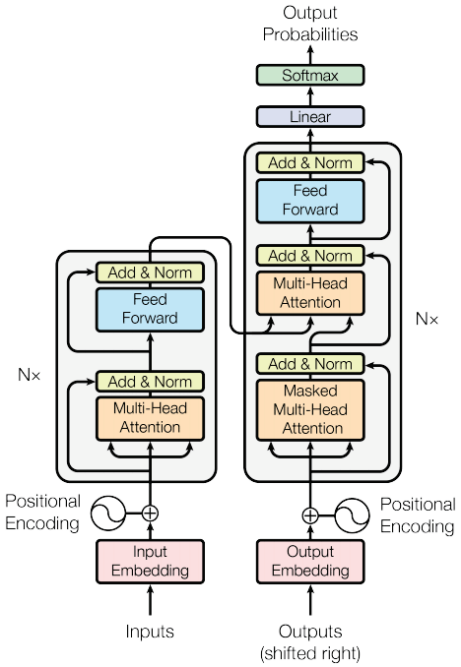


Figure 3.5: Transformer architecture. Image from Vaswani et al. (2017).

A Transformer block is a function $f : \mathbb{R}^{n \times d} \longrightarrow \mathbb{R}^{n \times d}$, that is, given a collection of n samples of size d, it outputs another collection of the same number of samples of the same size. A Transformer block learns three matrices of weights $W_q, W_k, W_v \in \mathbb{R}^{d \times k}$ to project an input sample $x \in \mathbb{R}^{n \times d}$ into three new spaces: $Q = W_q^T x$, $K = W_k^T x$, $V = V_v^T x$. These spaces receive the names of Queries, Keys, and Values respectively. Then, the attention operation, presented in Equation 3.1, is computed to obtain the attention weights.

$$\alpha := \text{Attention}(K, Q, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{k}}\right)V \tag{3.1}$$

The attention weights are applied to the values to obtain the desired output: $z = W_c^T \alpha^T V$ where $W_c \in \mathbb{R}^{k \times d}$ is also a learned matrix. This process is shown in the left part of Figure 3.6. Authors of Vaswani et al. (2017) also propose to parallelize the process and compute h independent transformations, called heads. The output of all the heads is concatenated and a last fully-connected layer between two layer normalizations (Ba et al., 2016) is applied to reshape the data into the final shape. Heads can be seen in the right part of Figure 3.6. Moreover, these blocks are usually stacked together to create a larger network.

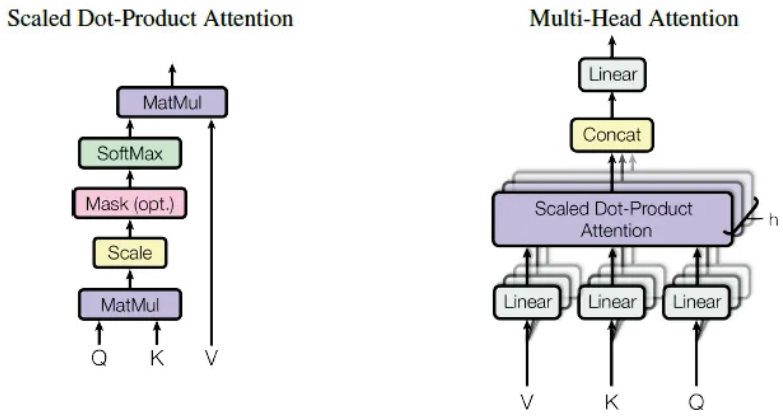


Figure 3.6: Attention and Multi-Headed attention. Image from Vaswani et al. (2017).

As mentioned, what the Transformer architecture does is, indeed, transform the data from samples $x \in \mathbb{R}^{n \times d}$ to $z \in \mathbb{R}^{n \times d}$. It is important to note that these transformations are applied to samples, not sequences. This means that all the samples are transformed in parallel without any temporal information; the model doesn't know which sample comes before which other, and so this information is not used to produce the output. This is why positional encoding is necessary. Positional encoding is a mechanism that gives the model information about the order of the samples in a sequence. The first approach to positional encoding was presented in the original paper (Vaswani et al., 2017) and used sinusoidal functions to encode the position. Later works (Gehring et al., 2017) established that these positional representations could be learned during the training of the network.

3.2.3 Vision Transformer

The ViT (Dosovitskiy et al., 2021) takes all the relevant parts of the Transformer and applies them to images. Although the Transformer was specifically developed for NLP problems, it has been successfully applied to the image domain with some twists. The idea behind the ViT is quite simple, to maintain as much as possible from the original Transformer architecture but to change the type of the input and output data. While the

original Transformer architecture acts as a sequence to sequence model, the ViT acts as a classifier. What the authors proposed is to divide the images in patches of fixed size (see Figure 3.7) and project these patches into an embedding of size d so that it mimics the input of the Transformer network. Now the positional encoding needs 2-dimensional

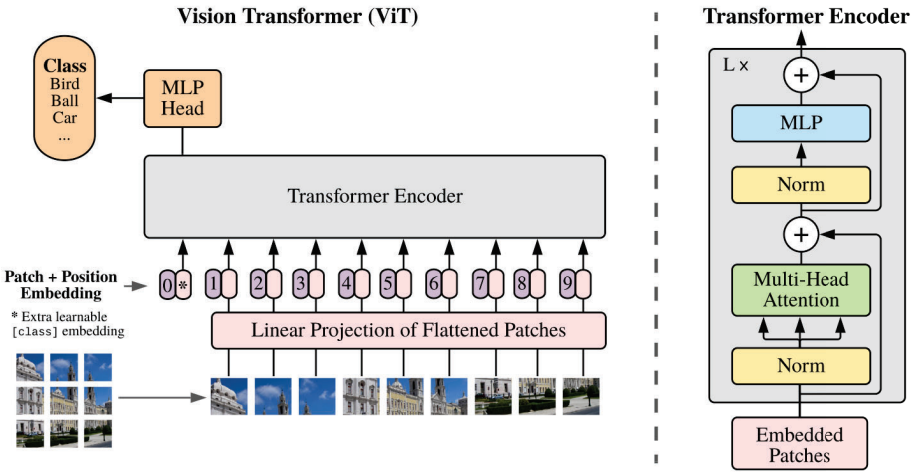


Figure 3.7: ViT architecture. Image from Dosovitskiy et al. (2021).

information and they also introduce an extra learnable token, the [class] token. The output of the Transformer encoder uses the modified version of this token for a final Multi Layer Perceptron (MLP) layer that transforms it into a probability between the target classes. An implementation of the ViT architecture is available on GitHub (Gilabert, 2024).

3.2.4 TransUNet

The TransUNet architecture (Chen et al., 2021b) uses the strengths of the applications previously introduced. It uses a U-Net-like architecture, that is, maintaining the skip connections of the original architecture but also use ViT blocks to encode the data in deeper parts of the network. Figure 3.8 shows its global structure.

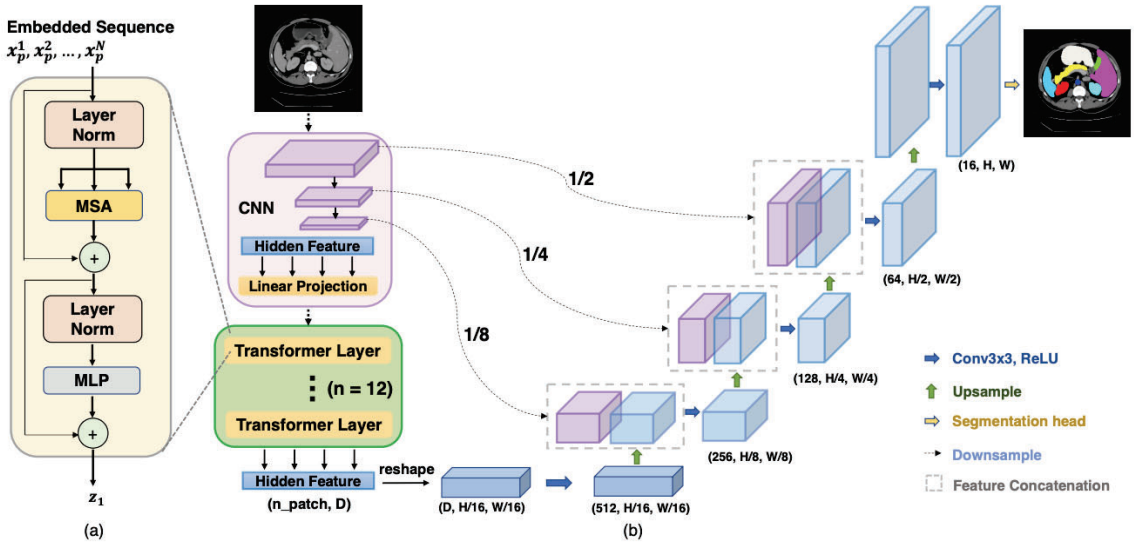


Figure 3.8: TransUNet architecture. Image from Chen et al. (2021b).

3.3 Cleansing Score

Efforts have been made to standardize the review process for CE videos (Koulaouzidis et al., 2021), but consensus on the reproducibility of the readings remains elusive (Cortegoso Valdivia et al., 2022; Lee et al., 2022; Leenhardt et al., 2021).

A thorough and safe evaluation of CE videos requires clean visualization of the GI tract (Vuik et al., 2021). Residual debris can obscure crucial pathology or lesions, complicating the task of reviewing the video and risking missing these findings. Assessing how clean a CE video is can be difficult, as it involves looking at both the overall video quality and individual frames. This makes it even harder to agree on a standard way to score them.

Several scales have been proposed to standardize CE readings over the years. The introduction of the CAC Score in 2018 (Becq et al., 2018) quantified cleanliness by calculating the percentage of red over green in each frame. The KODA Score (Alageeli et al., 2020) used a two-scale system, evaluating the percentage of visible mucosa and obstructed view, using predefined scores ranging from 0 to 3. Simplifying this approach, the CC-Clear Score (de Sousa Magalhães et al., 2021) assigns cleanliness scores also from 0 to 3 based on thresholds of visible colonic mucosa percentages, later adapted into the SB-Clear Score (Macedo Silva et al., 2022) to assess cleanliness for the small bowel. The CC-Clear score has a higher degree of consensus in practice (Tabone et al., 2021) compared to the previous scales, and so was selected for our evaluation.

Various solutions have been presented to assess the cleanliness of CE videos using various models and methodologies. Early approaches, such as those of Buijs et al. (2018), utilized support vector machines to classify images into binary categories of dirty and clean. Progressing into the era of deep learning, subsequent works such as that of Noorda et al. (2020)

employed neural networks to classify patches of size 64×64 pixels into these same categories.

More recent efforts, such as those of Nam et al. (2021a,b), explored systems that classify images into five categories and compared them with cleanliness scores assigned by physicians. Mascarenhas Saraiva et al. (2023) introduced a system to classify images into three categories: excellent, satisfactory, and unsatisfactory. All of these previous works attempted to assign scores to CE videos by classifying either images or patches of images. Ju et al. (2022) introduced a system for automatically segmenting intraluminal content and dark areas in images. While the method is interesting, it has a significant drawback: the cost of annotation. It takes a considerable amount of time to effectively segment the GI content in a set of images.

3.4 Assessing a Cleanliness Score

The method we present consists of three steps. Figure 3.9 shows a visual representation of all of them.

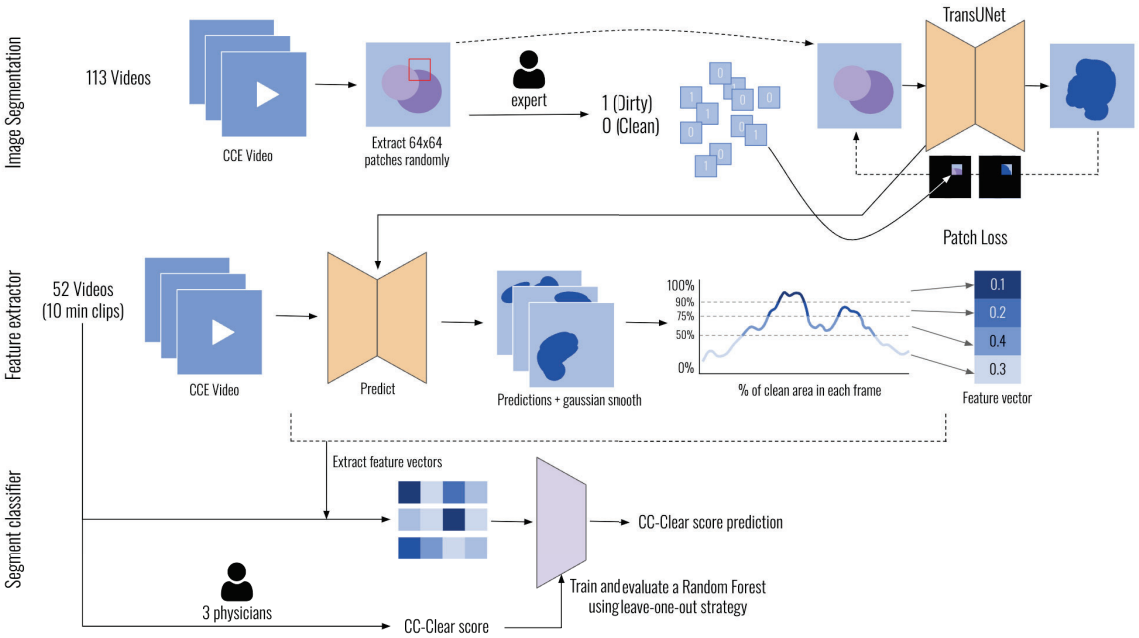


Figure 3.9: Overview of the method to assess the cleanliness score. The method consists of three steps: segmentation, feature extraction and classification.

1. **Image segmentation**. A TransUNet neural network is trained to segment images using patch labels, that is, without using fully annotated segmentation masks. We implemented a custom loss function we have named Patch Loss.

2. **Feature extractor**. Using the area of intestinal content in each image, we extract features to represent the cleanliness of a video.

3. **Segment classification**. Using the features extracted for each video, we predict the CC-Clear Score by training a RandomForest classifier, taking as ground truth the scores provided by expert physicians.

In the following sections, we explain the full method in further detail.

3.4.1 Intraluminal Content Segmentation

To create a segmentation model, labeled masks for the regions containing intraluminal content are usually required, serving as ground truth for the models. These segmentation masks consist of images that are the same size as the original ones and that highlight all the area obscured by intraluminal content, and that the model will learn how to generate. However, this traditional labeling process is slow and expensive. To address this, we suggest optimizing the annotation procedure by redefining image labeling. We propose a binary classification task for image patches, where a label of 1 indicates the presence of intraluminal content that obscures part of the image, and 0 signifies the absence of such content. The labeling criteria are straightforward: A patch receives a label of 1 if a physician believes that the intraluminal content present in the patch could conceal a pathology and 0 otherwise.

Traditionally, the task of classifying clean and dirty patches has been performed using only the patches themselves (Noorda et al., 2020), without considering the content of the entire image. We propose to maintain this straightforward approach of annotating patches alone but leverage this information within the loss function during segmentation model training. By doing so, we condition the segmentation model on the patch-level information we have.

To achieve this, we use a TransUNet that takes an input of (256,256,3) and produces a mask of the same size. Given an image X_k and a patch P_k of the same image, we define the *Patch Loss* as a cross-entropy loss restricted to each patch. For a batch of B patches, it can be expressed as:

$$\begin{aligned} \text{Patch Loss} &= -\frac{1}{B}\sum_{k=1}^{B}\mathcal{L}_k \\ \mathcal{L}_k &= \sum_{i=1}^{H}\sum_{j=1}^{W}\Big(y_k \log(P(X_k^{(i,j)}|_{P_k}, y_k) + (1-y_k)\log(1-P(X_k^{(i,j)}|_{P_k}, y_k)\Big) \end{aligned} \tag{3.2}$$

where $X_k^{(i,j)}$ indicates the pixel with coordinates (i,j), y_k is the patch label assigned by the expert and H, W are patch height and width, respectively. In our setting, both dimensions are 64.

To smooth the result, we apply a gaussian kernel of size 0.4. Final segmentation masks

are then obtained as those pixels with an activation higher than 0.5, obtaining binary masks for all the frames.

3.4.2 Feature extractor

Using the masks from the previous step, we compute the area that the segmentation occupies. That is, the number of pixels in the segmentation mask over the total number of pixels in the image, ignoring the black area in each corner.

Having this score per image, we create a plot depicting the visibility score per frame of the entire segment. This score represents the evolution of visibility across the entire segment and it is relevant for understanding areas where the capsule is stuck in a zone of poor visibility.

From the visibility plot we extract features following the CC-Clear Score. We compute the number of frames with visibility <50%, 50-75%, 75-90% and >90%, respectively. These four features are used in the last step of the process.

3.4.3 Segment Classification

Since achieving reproducibility is challenging due to the inter-interpreter variability of assigning a cleanliness score to an entire video (Cortegoso Valdivia et al., 2022), we propose to validate the system at a segment level. That is, to extract 10-minute clips and assess their cleanliness using the system proposed.

We randomly extracted 10min clips (real capsule moving time) that were evaluated by three experienced physicians. Each physician of them assigned a CC-Clear Score to each clip ranging from 0 to 3, assessing the level of cleanliness of the segment.

Using the features extracted in the previous steps, we trained a Random Forest classifier with a leave-one-out cross-validation strategy (Hastie et al., 2009). The physicians' scores served as the ground truth for training. We explored two approaches:

- **Individual Model Training**: We trained separate models to replicate the scoring patterns of each physician.
- **Consensus Model Training**: We trained a single model using the consensus score derived from the three physicians as the ground truth. This consensus score was calculated by averaging the individual scores given by the physicians and rounding the result to the nearest integer.

This dual approach allows us to capture both individual scoring nuances and a consensus view of video cleanliness. For each method, we assessed the agreement between the physicians' scores and the model's predictions.

3.5 Experimental Setup

3.5.1 Dataset

This study utilizes 165 CCE videos, sourced from two retrospective studies conducted at the facilities of the NHS Highlands Raigmore Hospital in Inverness. Both studies included patients who were referred with symptoms or for surveillance within the Highlands and Islands area of Scotland. All these patients had a positive Fecal Immunochemical Test (FIT).

All patients underwent bowel preparation following a standardized protocol that involved a split-dose of PEG.

The videos were captured using a PillCam^{TM} COLON 2, which contains two cameras (front and rear). To ensure patient confidentiality, all videos were anonymized removing relevant information stamped on the images.

3.5.2 Data Splits

To ensure a fair train, evaluation and test, we split the videos into two main groups to avoid data leakage between the steps. The total set of 165 videos is split in the following way:

- We used 113 videos to train, validate, and test the image segmentation model. From these videos, 8,492 patches of size 64×64 were randomly extracted. This set of 113 patients was split into three groups: 69 patients for training (5,306 patches), 22 for validation (1,539 patches), and 22 for testing (1,647 patches) to evaluate the model's performance.

- The remaining videos, 52, were used to train and evaluate the performance of the segment classifier using a leave-one-out strategy.

3.5.3 Training configuration

All the deep learning code was implemented in Python and executed in an NVIDIA 3090 RTX GPU.

We used Keras as framework to reproduce (Noorda et al., 2020) strategy and used a pretrained TransUNet as a core of our method.

To evaluate the effectiveness of our solution, we not only performed the segmentation task but also compared our method against other classifiers for patch classification. Although our primary focus is on segmentation rather than classification, we assigned labels to patches based on the model's segmentation output. Specifically, a patch was labeled

as positive (Dirty, 1) if the predicted segmented area covered 50% or more of the patch. Conversely, it was labeled as negative (Clean, 0) if less than 50% of the patch was segmented.

For the CC-Clear score classifier, we employed a leave-one-out cross-validation strategy using the remaining 52 videos. This approach involves training a model on a subset of 51 videos and testing it on the one remaining video, ensuring that every video is used for testing exactly once. We trained a Random Forest classifier with 100 estimators and set the maximum depth to 2 to prevent overfitting. Standard algorithms from the sklearn library were utilized throughout the process.

3.6 Experiments and Results

The results we present are organized in the following way. Firstly, we present an evaluation of the image segmentation. Secondly, we compare the patch classification performance. Lastly, we present results on the segment classification.

3.6.1 Segmentation Results

We initially demonstrate the model's ability to segment intraluminal content by requesting an expert to manually segment a small, random set of 32 images. Table 3.1 presents the

Strategy	**mIoU**
Noorda et al. (2020)	0.43
ResNet50	0.48
ViT-B16	0.55
TransUNet + Patch Loss (Ours)	**0.73**

Table 3.1: Mean intersection over union score evaluated on 32 images manually segmented by an expert annotator.

mean Intersection over Union (mIoU) results comparing the predicted masks to the ground truth. Our model, shown in the final row, exhibits improved segmentation masks with the introduction of this conditioned loss.

Figure 3.10 showcases the segmentation performance of the TransUNet with the Patch Loss strategy. Each row in the figure sequentially displays the original images, the ground truth masks manually annotated by an expert, the masks predicted by our model, and the final binary masks obtained by applying a 0.5 threshold to the model's predictions. This visual comparison illustrates the accuracy and effectiveness of our approach in segmenting intraluminal content with a low annotation strategy. It is important to highlight the inherent difficulty in segmenting intraluminal content accurately. As evident in the images, the boundaries of the areas occluded are often not clearly defined. This lack of clear demarcation

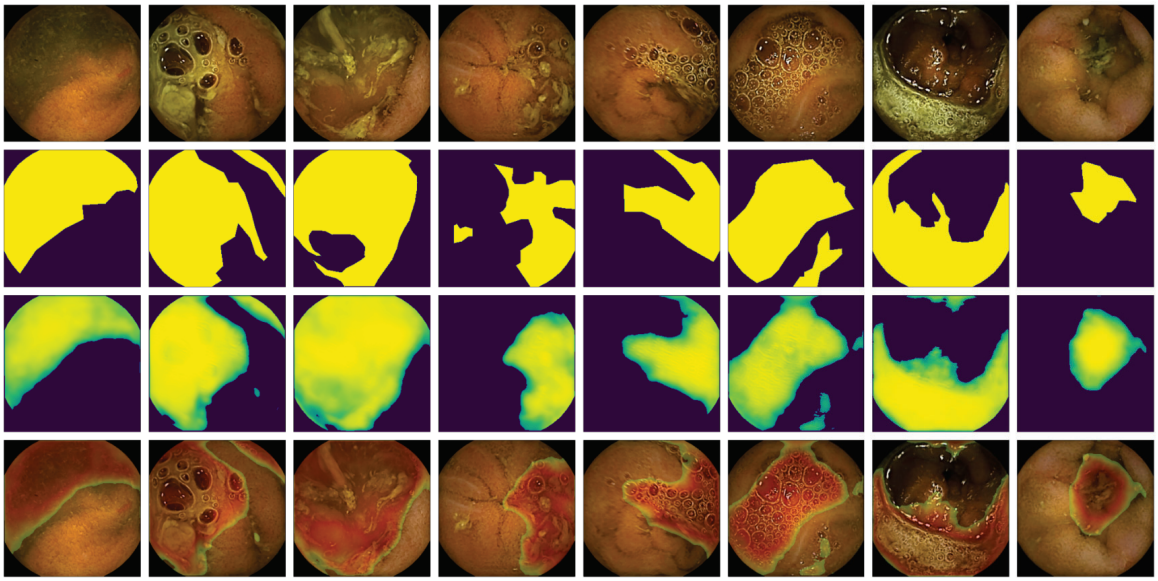


Figure 3.10: Segmentation results for the TransUNet + Patch Loss strategy. Random images annotated by an expert. The rows are ordered as follows: the original image, the ground truth mask as annotated by an expert, the predicted mask from the model, and the thresholded mask with a 0.5 cutoff.

makes segmentation particularly challenging and leads to significant variability in the masks depending on the expertise and interpretation of the annotator.

3.6.2 Patch Classification Results

We evaluated patch classification using the method previously explained, to compare our proposed strategy with other baseline methods.

Strategy	Acc.	AUC	Prec.	Rec.	F1
Noorda et al. (2020)	0.89	0.82	0.92	0.67	0.78
ResNet50	0.89	0.87	0.75	0.84	0.79
ViT-B16	0.90	0.88	0.84	0.82	0.83
TransUNet + Patch Loss (Ours)	**0.97**	**0.96**	**0.93**	**0.93**	**0.93**

Table 3.2: Results of the four strategies evaluated on the test set. Results show that the proposed strategy, TransUNet + Patch Loss, improves patch classification.

Table 3.2 compares four strategies: Noorda et al. Noorda et al. (2020), ResNet50, ViT-B16 and TransUNet + Patch Loss, evaluated on the test set of 22 videos. The results clearly demonstrate that our proposed strategy surpasses the previous methods across all metrics presented.

3.6.3 Segment Classification

We randomly extracted 10-minute clips from 52 new videos, one per video, and three expert physicians evaluated their cleansing. Following the CC-Clear score, each clip received a score between 0 and 3. 0 being a video almost without visible mucosa, and therefore unusable, and 3 a video with a very clean mucosa without any doubt of missed pathology because of the presence of intraluminal content. The scores that the physicians reported are shown in Table 3.3. A moderate agreement based on the Cohen's kappa score was

CC-Clear Score	0	1	2	3	Mean Score
Physician #1	4	12	27	9	1.79 ± 0.82
Physician #2	1	10	23	18	2.12 ± 0.78
Physician #3	2	10	27	13	1.98 ± 0.78

Table 3.3: Video clip scores. Number of videos each physician scored for each different score.

found between the experts: $k_{12} = 0.537$, $k_{23} = 0.459$, and $k_{13} = 0.643$, where k_{ij} represents the score between physician i and physician j. The average score among the physicians is $\bar{k}_{\text{orig}} = \frac{k_{12}+k_{23}+k_{13}}{3} = 0.546$. These results highlight the significant inter-observer variability previously noted in existing literature.

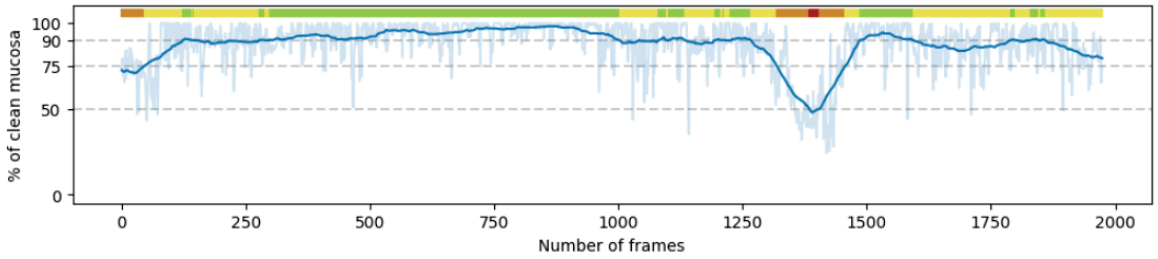


Figure 3.11: Example of a test procedure. Clean mucosa prediction for each frame in the clip. A centered moving average is applied to smooth the results. At the top of the plot, the predicted CC-Clear score for each part of the video clip is shown using a color scale. Red (<50%), Orange (50-75%), Yellow (75-90%), and Green (>90%) which matches with the thresholds set by the horizontal dashed lines.

Figure 3.11 shows a test procedure with the predicted cleanliness of each frame. Horizontal dashed lines show the different thresholds set for the CC-Clear Score. These thresholds are better visualized in the colored bar at the top of the plot, summarizing the cleanliness score of each area of the video. Depending on the level of cleanliness of the neighboring frames, the scale is Red (<50%), Orange (50-75%), Yellow (75-90%), and Green (>90%). The plot shows the predicted percentage of clean mucosa and a centered moving average for better visualization.

For each clip, we extracted four features based on the number of frames in each of the four regions of visibility, obtaining a 4-dimensional vector representing the video. Figure 3.12

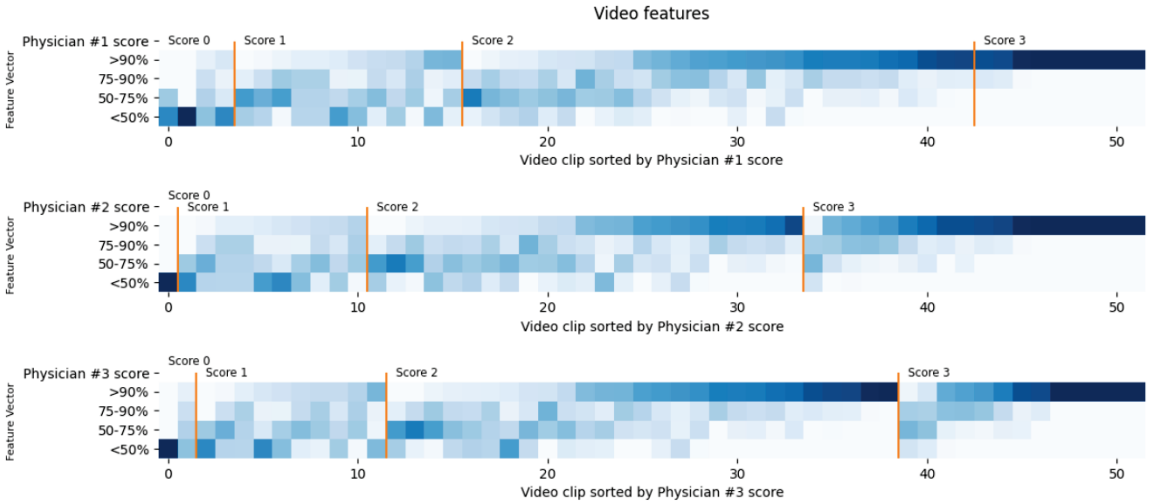


Figure 3.12: Feature vector for each video. Each video is a column showing its 4 values. A darker color means a higher value. Values are sorted first by ground truth (physician score) and then by the first component of the vector (first row of each plot).

shows a visual correlation between the features and the ground truth established by the physicians. We can observe that, while the extremes seem more homogeneous, the central area of the plots is more ambiguous.

Following the process explained in the methodology, we evaluated the system using two strategies: Individual Model Training and Consensus Model Training:

Individual Model Training

We independently trained a model to replicate each of the physicians' scoring.

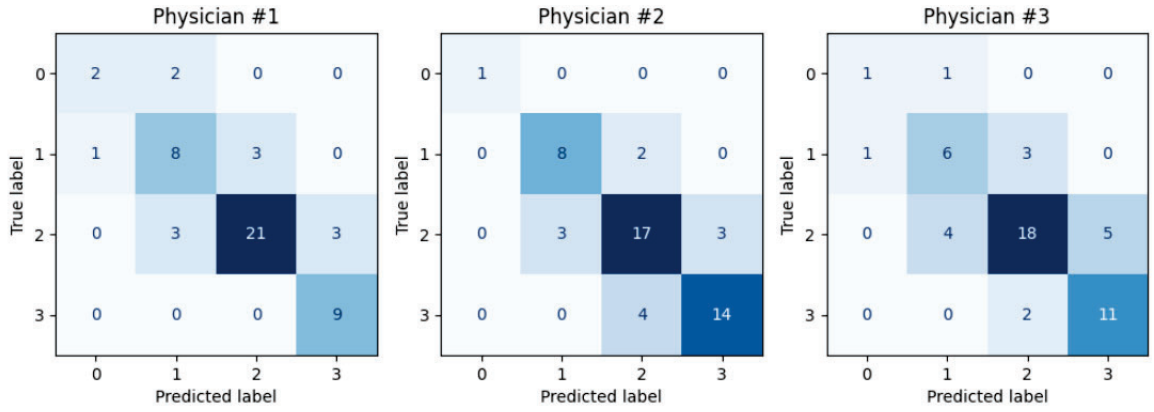


Figure 3.13: Confusion matrices for the three models. Each model is trained on the scores of a single physician using a leave-one-out strategy.

Figure 3.13 presents the confusion matrices for these three models. The first model,

trained to mimic the first physician, achieved a Cohen's kappa agreement of $k_1 = 0.649$ with an accuracy of 76.9%. The second model, aligned with the second physician, reached an agreement of $k_2 = 0.645$ and also an accuracy of 76.9%. The third model, corresponding to the third physician, attained an agreement of $k_3 = 0.528$ with an accuracy of 69.2%. The average agreement across these individual models is $\bar{k}_{\text{indiv}} = 0.607$. Notably, both the mean agreement and each individual model's agreement with their respective physician exceed the average original inter-observer agreement between the physicians, which was $\bar{k}_{\text{orig}} = 0.546$.

Consensus Model Training

We also developed a Random Forest classifier model using the average of the physicians' scores, rounded to the nearest integer, as the consensus ground truth. This approach simulates the combined consensus scoring of each video by the three physicians. Figure 3.14 shows the confusion matrix for the consensus model's results.

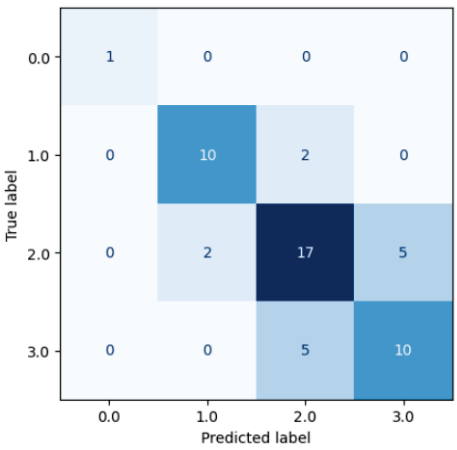


Figure 3.14: Results of a regressor model trained using the average of physicians' scores as ground truth.

The consensus model achieved a Cohen's kappa agreement of $k_{\text{cons}} = 0.586$, which is an improvement over the average original inter-observer agreement between the experts, $\bar{k}_{\text{orig}} = 0.546$.

	Physician			**Individual Model**			**Consensus Model**
	P_1	P_2	P_3	M_1	M_2	M_3	M_{cons}
P_1	-	$k_{12} = 0.537$	$k_{13} = 0.643$	$k_1 = 0.649$	-	-	
P_2	-	-	$k_{23} = 0.459$	-	$k_2 = 0.645$	-	$k_{\text{cons}} = 0.586$
P_3	-	-	-	-	-	$k_3 = 0.528$	
avg.	$\bar{k}_{\text{orig}} = 0.546$			$\bar{k}_{\text{indiv}} = 0.607$			-

Table 3.4: Summary of agreement scores: physicians, individual models, and consensus approach.

Table 3.4 summarizes all the numbers mentioned in the previous sections.

3.7 Conclusions

We propose a novel method to enhance the classification and segmentation of intraluminal content with minimal labeling effort. By leveraging our model's training approach, we achieve more accurate segmentation masks when compared with existing low-effort annotating methods. Our process involves several stages, each validated with different CCE videos to ensure robustness and prevent data leakage.

To support our claims, we conducted a series of experiments. Initially, we evaluated the segmentation performance without relying on annotated masks. We achieved a high mIoU score by introducing a conditioned loss on annotated patches. Subsequently, we assessed the patch classification accuracy across all models, with our newly introduced model demonstrating the highest accuracy.

Given the significant variability in determining cleanliness scores for entire videos, as noted in the literature, we focused on evaluating our method using 10-minute clips from various CCE videos. Three physicians provided cleanliness scores on a scale from 0 to 3 for each clip. Using these scores, we trained Random Forest classifiers –one for each physician– to replicate their scoring patterns. The models exhibit stronger agreement in replicating physician assessments than the individual scores provided by the physicians. Specifically, the mean achieved agreement of the models was $\bar{k}_{\text{indiv}} = 0.607$, surpassing the original inter-physician agreement, which was $\bar{k}_{\text{orig}} = 0.546$.

Furthermore, we developed a jointly trained Random Forest classifier model using the physicians' average scores as consensual ground truth. This model, which simulates a consensued score by three physicians, achieved an agreement of $k_{\text{cons}} = 0.586$, also an improvement when compared to the original inter-physician agreement.

Overall, our approach demonstrates significant advancements in both segmentation and classification tasks in the context of CCE, with minimal manual annotation. By effectively combining individual physician assessments into a consensus model, we provide a robust framework that enhances the reliability and accuracy of intraluminal content analysis. These findings suggest promising applications for our method in clinical practice, potentially reducing the burden on medical professionals and improving patient outcomes through more accurate diagnostics.

Chapter 4

Landmark Identification

Contents

In the process of reviewing a CE video, the next step is to correctly identify the organ to be analyzed. This chapter presents existing solutions for organ segmentation and presents a state-of-the-art method to automatically detect organ boundaries. Specifically, the presented solution is able to automatically delimitate the small and large intestines in CE videos.

4.1 Organ Segmentation

Organ segmentation is the task of identifying the entrance and exit of an organ in a video. As displayed in Figure 4.1, in CE videos this task consists of detecting the entrance and exit of the small intestine, that is, the last pylorus image and the ileocecal valve image; and the entrance and exit of the large intestine which are the first cecal image and the last rectal image.

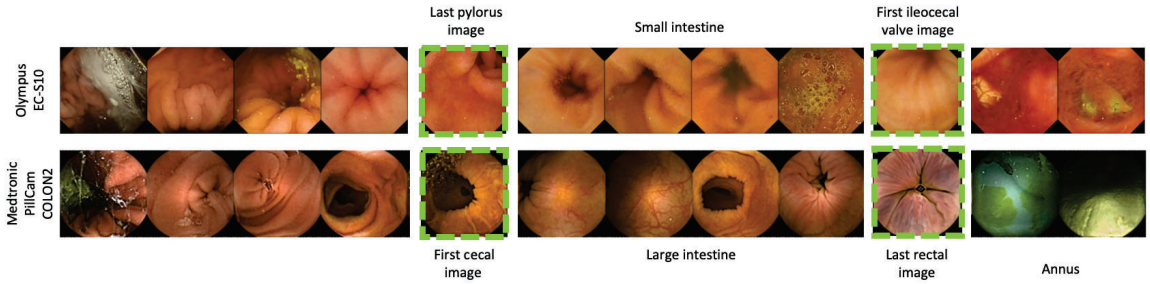


Figure 4.1: Landmark detection.

The organ segmentation task has been extensively studied by various researchers, with several significant solutions emerging over the years. Before the advent of deep learning, traditional image analysis methods dominated this field. One of the pioneering efforts was by Berens et al. (2005), who compared the performance of a Support Vector Classifier and a Nearest Neighbor Classifier to detect the last pylorus image and the ileocecal valve using hue-saturation chromaticity histograms. Similarly, Lee et al. (2007) focused on contraction patterns of different organs to automatically detect their entrance and exit points, extracting these patterns using color features in the Hue-Saturation-Intensity (HSI) range. Continuing in this vein, Mackiewicz et al. (2008) presented a solution that utilized a feature vector containing color information from HSI histograms and motion information from selected image patches. Haji-Maghsoudi et al. (2012) approached the problem from a classification perspective, extracting image features, including color, to create a representative vector for each image. They then employed two different methods: Fuzzy K-Means and a MLP. Lastly, Li et al. (2015) distinguished between the small and large intestines using HSI histograms and color Uniform Local Binary Patterns (ULBP).

Deep learning came to improve all these methods. The latest and most important works in the field include CNNs and the use of Transformers (Vaswani et al., 2017), following the global tendency in the deep learning field. Zou et al. (2015) and Chen et al. (2017) presented CNNs solutions to classify frames into three and four classes respectively. The approach from Zhao et al. (2021) used an encoder step, a ResNet, to generate an embedding of each frame in the image. These embeddings were then introduced into a Transformer module that compared each frame with its neighbors and classified them into three categories: esophagus/stomach, small intestine and colorectum. Moreover, they apply a search algorithm to assess the exact frame of entrance and exit of the organ. More recent works include the

method presented by Son et al. (2022) that use a combination of CNNs, a Savitzky-Golay filter and a median filter to find the frame of transition between organs.

It is evident that deep learning methods outperformed all the previous works in terms of classification approaches. Moreover, only Zhao et al. (2021) and Son et al. (2022) apply thresholding techniques to identify the boundaries of the small bowel. To the best of our knowledge, all the studies were performed using private datasets and with only one type of capsule.

4.2 Methodology

The method presented in Figure 4.2 is specifically designed to segment organs in CE videos. Following the approaches of Zhao et al. (2021) and Son et al. (2022), the method consists of three main steps:

1. Train a deep learning model to assess the probability of each frame belonging to a specific organ, such as the small or large intestine.

2. Smooth the probability signal of the full video by introducing two new features: temporal features that help to identify nearby frames and motion features that use the similarity between images to account for the speed of the camera.

3. Use the smoothed signal to identify the entrance and exit of the desired organ by fitting a rectangular pulse to the signal.

These steps will be explained in detail in the following sections. As mentioned, the method is based on three important features that can be observed in Figure 4.2. If i is the index of the i^{th} frame of the video, these features are: frame temporal information, $s_t(i)$, frame probability of belonging to the organ $s_p(i)$ and frame motion, $s_m(i)$.

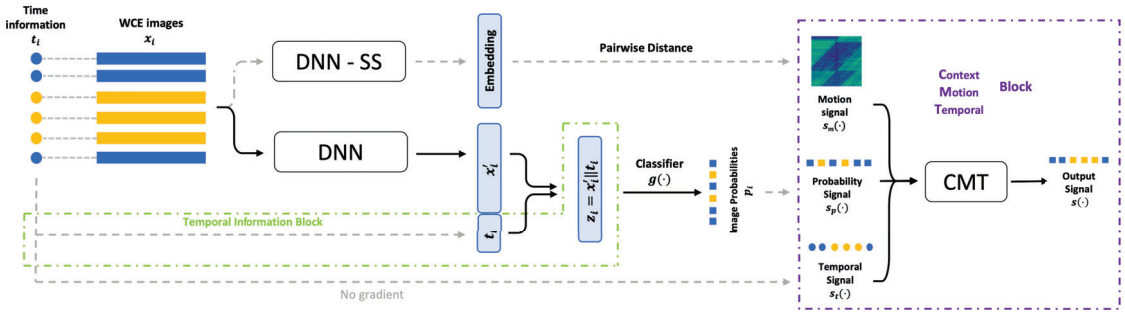


Figure 4.2: Overview of the method. Blue: frames outside the organ. Yellow: frames inside the organ.

4.2.1 Image Classification

Let $x_i \in V$ be an image, where x_i is the i^{th} frame of a CE video V with k frames so that $|V| = k$. Let $f(\cdot)$ be an encoder architecture that, given an image, returns an embedding. This low-dimensional representation of the image x_i can be defined as $x'_i = f(x_i) \in \mathbb{R}^n$.

Let t_i be the time, in seconds, that has elapsed from the start of the video, normalized by the length of the video, that is:

$$s_t(i) := t_i = \frac{\text{timestamp}_i}{\text{video length}} \in [0, 1],\ \forall i = 1, \ldots, k \tag{4.1}$$

This value can be concatenated to the embedding vector to obtain a larger embedding that contains this temporal information. We can express it as $z_i = x'_i \parallel t_i \in \mathbb{R}^{n+1}$ where $\parallel$ represents the concatenation operation.

The capsule advances through the GI tract, recording all organs in a continuous manner. It is worth noting that although the camera might move back and forth, it remains within the same organ. This allows the model to establish a relationship between time and the organs. The temporal feature added by our system enables the model to discard erroneous predictions in different sections of the video.

To extract the probability of each frame, a linear classifier is used. The classifier, $g(\cdot)$, receives the extended embedding and classifies it as inside or outside of the organ of interest. Equation 4.2 summarizes the process explained above.

$$s_p(i) := g(z_i) = g(x'_i \parallel t_i) = g(f(x_i) \parallel t_i),\ \forall i = 1, \ldots, k \tag{4.2}$$

4.2.2 Signal Smoothing

The second step in the pipeline involves assessing the movement of the capsule. When the camera is moving at high speed, the images will exhibit more differences compared to when the camera is stationary in the same area for an extended period. This information helps relate nearby frames not only from a temporal perspective, as in the previous step, but also from a spatial perspective. This feature considers two frames to be similar if they are physically located near each other. Another critical aspect to consider is the consistency of the signal. As the camera moves through the digestive system, it may move back and forth due to natural intestinal movements, but it cannot switch organs more than once, because the valves separating the organs prevent the camera from returning to a previous organ. Therefore, the signal obtained in this step must be consistent and should not account for more than one entrance and one exit.

The movement signal is defined as the function $s_m(\cdot)$. This signal is computed by considering the similarity of frames, particularly the distance between frame embeddings. A self-supervised method from Pascual et al. (2022) was used to compute these embeddings.

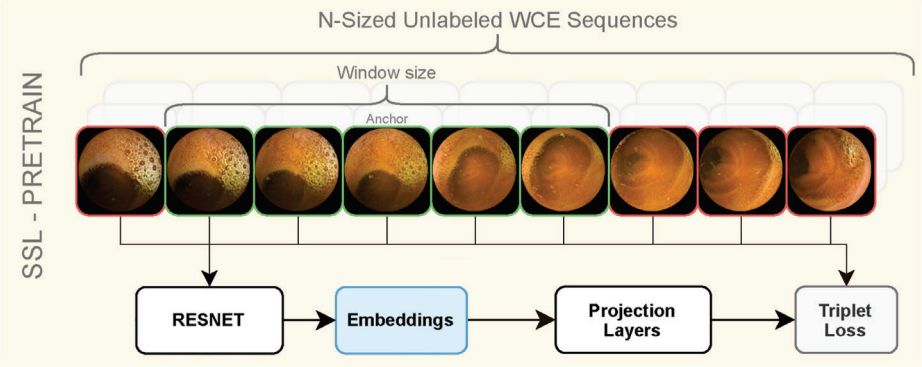


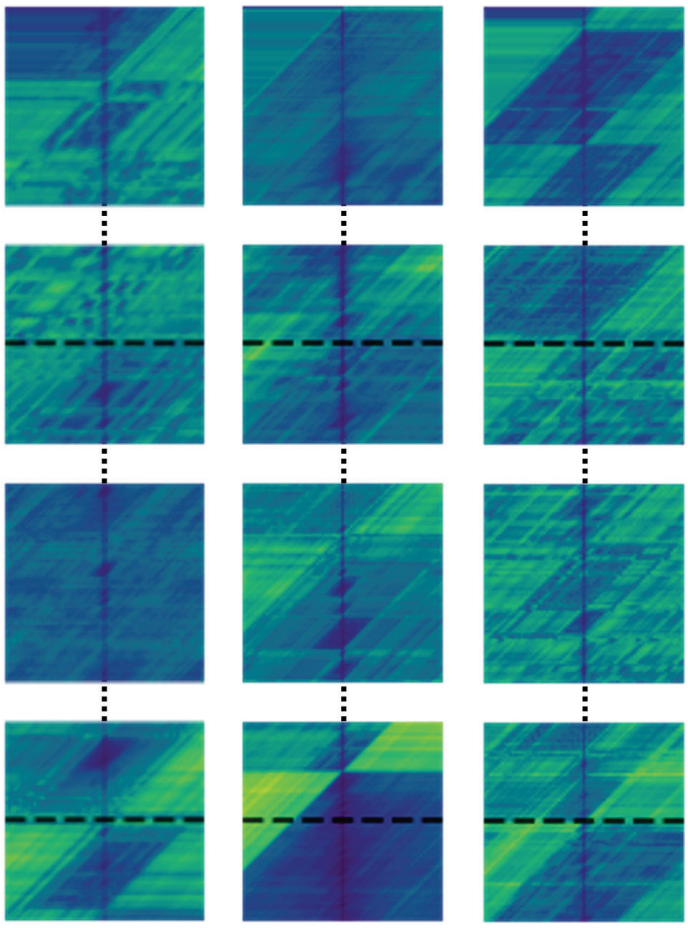
Figure 4.3: Self-Supervised learning strategy to generate similarity embeddings. Image adapted from Pascual et al. (2022).

This method considers an anchor frame, the central frame of a sequence, and all its neighbors within a specified window size as positive, while frames outside this window are considered negative. Using the Triplet Loss (TL) function, it learns meaningful embeddings that capture the visual characteristics of the frames. Following this strategy, neighboring frames will have similar embeddings.

Let $e_i = f_{ss}(x_i)$ be the temporal embedding of an image using the self-supervised encoder, f_{ss}. Let M be the matrix of euclidean distances between embeddings, $M_{ij} = ||e_i - e_j||_2$. The motion signal for the i^{th} frame is the vector computed by taking the normalized i^{th} row of M, that is:

$$s_m(i) = \frac{M_i}{\sum_{j=1}^{k} M_{ij}}, \ \forall i = 1, \ldots, k \tag{4.3}$$

Figure 4.4 presents a representation of the values in matrix M. Due to the intractable size of the complete matrix, only four moments of a video are illustrated here. Each column showcases one video from three different datasets.



(a) Kvasir-Capsule dataset (b) VH dataset (c) Capri dataset

Figure 4.4: Camera movement visualization on four key moments: the beginning of the video, the first landmark, the interior of an organ, and the second landmark. Each row represents a single video of three different datasets. Brighter colors represent larger Euclidean distances between frames.

Each of the twelve subplots contains a blue line in the middle, indicating that the distance from a frame to itself is zero; in the plots, brighter colors represent higher values. The videos are divided into four key moments: the first row represents the beginning of the video, the second row corresponds to the first landmark annotated by experts, the third row captures a random moment within the organ, and the fourth row presents the second landmark annotated by experts. The second and fourth rows display the transition frame from one organ to another, marked by a black dashed line. In these plots, only the nearest 500 temporal frames are displayed for each central frame.

With this, the three relevant signals for the method are computed: $s_p(i)$ representing the probability signal, $s_t(i)$ representing the temporal signal and $s_m(i)$ representing the motion signal. To combine them, the Context-Motion-Temporal (CMT) block is applied.

The CMT block is composed by two bi-directional Long Short-Term Memory (LSTM) layers and one dense layer that receives the three signals and outputs a single smoothed one, $s(i)$.

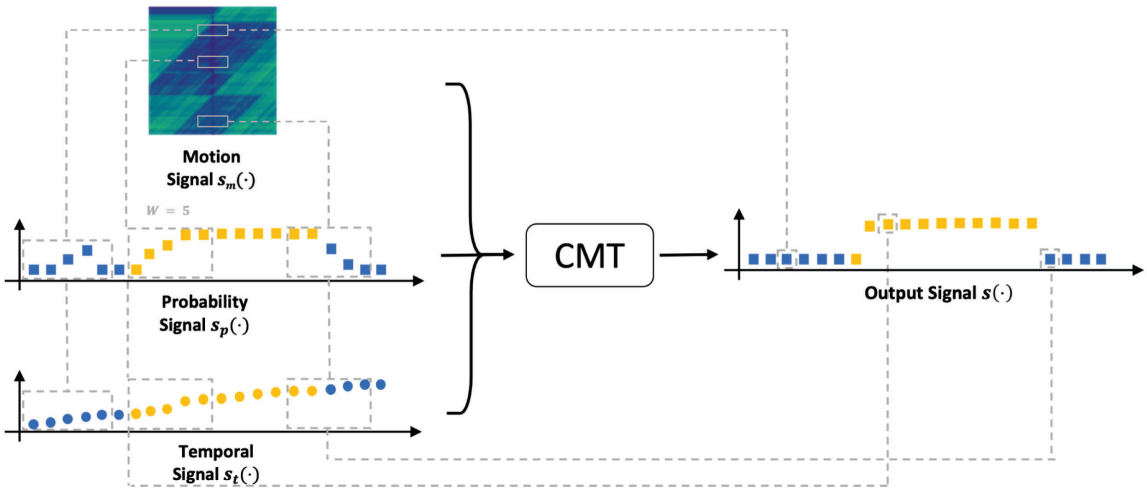


Figure 4.5: Overview of the CMT block with a window size of $w = 5$. The block recieves the three feature signals, i.e., the probability s_p, the motion, s_m and the time s_t and produces a single smoothed signal s by combining them.

As can be seen in Figure 4.5, a window size, w is considered as an hyperparameter so that, to compute value $s(i)$, the CMT block receives the values with indexes $i - \frac{w-1}{2}, \ldots, i + \frac{w-1}{2}$ of the three signals as a concatenated vector. The exact formulation of the CMT block is:

$$
\begin{aligned}
s(i) &= \mathrm{CMT}_w\left(s_t^w(i) \mid\mid s_p^w(i) \mid\mid s_m^w(i)\right) \quad \text{with} \\
s_t^w(i) &= \left(s_t\left(i - \frac{w-1}{2}\right), \ldots, s_t\left(i + \frac{w-1}{2}\right)\right) \\
s_p^w(i) &= \left(s_p\left(i - \frac{w-1}{2}\right), \ldots, s_p\left(i + \frac{w-1}{2}\right)\right) \\
s_m^w(i) &= \left(s_m\left(i - \frac{w-1}{2}\right), \ldots, s_m\left(i + \frac{w-1}{2}\right)\right)
\end{aligned}
\tag{4.4}
$$

4.2.3 Landmark Identification

From the output signal $s(\cdot)$, the last step is to identify the entrance and exit of the organ. To do so, the best frames for the entrance and exit of the organ need to be found. From the plot in Figure 4.5, one can see when this transition occurs. Given a rectangular pulse, that is, an expression with closed form:

$$
V(t) = u(t-a) - u(t-b), \quad u(t) = \begin{cases} 0 & \text{if } t < 0 \\ 1 & \text{if } t \geq 0 \end{cases}
\tag{4.5}
$$

a minimization problem between this pulse and the signal needs to be solved. This means finding the best values for a and b that satisfy that the distance between V and s is minimal. Figure 4.6 represents the minimization problem.

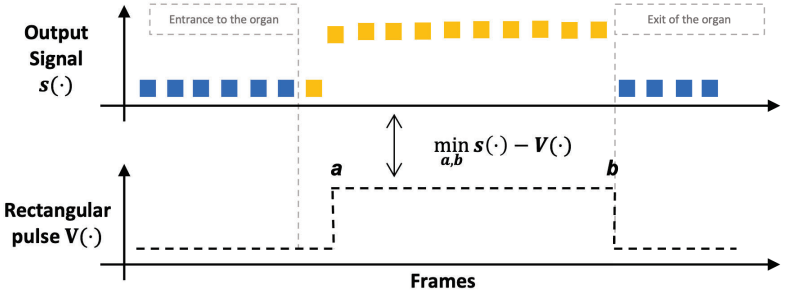


Figure 4.6: Rectangular pulse minimization. The signal that the model outputs is minimized against the rectangular pulse by finding the best a and b parameters that adapts the best to the function.

Formally, this means solving this constrained expression:

$$a^*, b^* = \arg\min_{a,b} \sum_{i=0}^{a-1} s(i) + \sum_{i=a}^{b-1} (1 - s(i)) + \sum_{i=b}^{n} s(i) \quad \text{s.t.} \quad a < b \tag{4.6}$$

4.3 Experimental Setup

4.3.1 Datasets

The proposed method was evaluated on three datasets: a public one, *Kvasir-Capsule* (Smedsrud et al., 2021) and two private datasets: *VH* and *Capri*.

Kvasir-Capsule dataset

The Kvasir-Capsule dataset comprises 117 videos recorded at a Norwegian hospital using an Olympus EC-S10 capsule. Out of these, only 24 videos include anatomical landmarks, specifically the entrance and exit of the small intestine. Images of the intestine make up approximately 75% of the dataset. Each video averages around 44,000 frames, but does not include temporal information for these frames.

VH dataset

The VH dataset consists of 48 videos recorded at Vall d'Hebrón Hospital in Barcelona using a PillCam SB3. Expert physicians annotated the entrance and exit of the small bowel in all

videos. Each video contains an average of 35,000 frames, with a mean duration of 4 hours and 36 minutes. Intestine images comprise approximately 71% of the dataset.

Capri dataset

The Capri dataset comprises 68 videos recorded at Raigmore Hospital - NHS Highland in Inverness, using a PillCam COLON 2. This device captures images with two cameras, front and rear, both of which are utilized in the study. All the videos contain expert annotations marking the entrance and exit of the colon. They have an average duration of 8 hours and 19 minutes, containing around 14,000 frames each. Colonic images make up approximately 75% of the dataset.

4.3.2 Evaluation Criteria

All models developed for the datasets were evaluated using the methodology described in the original *Kvasir-Capsule* paper by Smedsrud et al. (2021). This approach involves a two-fold cross-validation strategy based on patients. This ensures that frames are split into two groups, with frames from the same patient belonging to only one group. Table 4.1 presents the number of patients and frames of each group.

Dataset	Partition	#Patients	#Inside	#Outside	Total
Kvasir-Capsule	**Fold 0**	12	400K	160K	560K
	Fold 1	12	384K	97K	481K
	Total	**24**	**784K**	**257K**	**1M**
VH	**Fold 0**	24	602K	246K	848K
	Fold 1	24	592K	249K	841K
	Total	**48**	**1.2M**	**495K**	**1.6M**
Capri	**Fold 0**	34	347K	148K	495K
	Fold 1	34	393K	97K	490K
	Total	**68**	**740K**	**245K**	**985K**

Table 4.1: Number of patients and frames in each of the folds used to train the models. Patients were divided into two folds to ensure that all their frames belong to the same group, preventing data leakage. The "Inside" and "Outside" columns display the number of frames within and outside the corresponding organ, depending on the dataset.

Following the validation presented in previous works (Zou et al., 2015; Chen et al., 2017; Adewole et al., 2020; Zhao et al., 2021; Son et al., 2022), the methods were evaluated using the following metrics: AUC, Accuracy (ACC), Mean Accuracy (mACC), Specificity (SPEC), and Sensitivity (SENS). AUC and mACC are particularly suitable for assessing the performance of binary classification models on imbalanced datasets. AUC quantifies the

model's ability to distinguish between images belonging to the target organ and those that do not, while mACC and ACC measure the overall accuracy of predictions. It is important to note that mACC is computed by calculating accuracy per video and then averaging across the number of videos.

To report the anatomical landmark localization performance, MAE and Median Absolute Error (MedAE) are used, following the results presented in Mackiewicz et al. (2008); Li et al. (2015); Zhao et al. (2021); Son et al. (2022). These metrics measure the distance in number of frames and time between the anatomical landmark annotated by the expert physician and the one predicted by the method. The Kvasir-Capsule dataset does not contain time information, so only the error in the number of frames is presented. To account for the differences in the lengths of the videos, all the presented metrics are computed per video independently and then averaged to report a single value.

4.3.3 Implementation Details

The core of the model is a ResNet50 (He et al., 2016) pretrained with ImageNet weights. As previously mentioned, the CMT block consists of two bidirectional LSTM layers with 200 and 100 units, respectively, followed by a dense layer with two neurons and softmax activation for classification.

All experiments were conducted on a single NVIDIA GeForce RTX 2080 TI using TensorFlow 2.4 with CUDA 11.0. The training methodology comprised two steps: first, training the ResNet50 network, and second, freezing its weights to train the CMT block.

For ResNet50, the optimization process utilized Stochastic Gradient Descent with a batch size of 256. All networks were trained for 10,000 iterations, starting with a learning rate of 0.1, which was decreased by a factor of 0.1 every 2,000 steps. All images were resized to 128×128 pixels, with black edges masked to ensure the removal of any artifacts. Data augmentation techniques such as 90, 180, and 270-degree rotations, vertical and horizontal flips, and brightness adjustments were randomly applied during training.

For the CMT block, optimization was performed using the RMSprop algorithm with a fixed learning rate of 10^{-3} over 4,000 iterations and a batch size of 512. A grid search was employed independently for each dataset to determine the optimal value for the window hyperparameter, w.

4.4 Results

The experiments conducted to evaluate the proposed method were two-fold. First, experiments were performed to classify each of the three datasets as either within or outside the studied organ. Second, another set of experiments focused on identifying the anatomical landmarks, specifically the exact frame marking the transition between organs. To conclude

this section, some qualitative results displaying specific videos are presented.

Before detailing these two types of experiments, a grid search was applied to determine the best window parameter for each dataset. Table 4.2 shows various window size parameters, the AUC results obtained, the error in the number of frames for each evaluated dataset, and the total error as the sum of the entrance and exit values. As seen in the table, the optimal values are around 151 and 251 in most cases, so $w = 201$ was selected as the window size for the *Kvasir-Capsule* and *Capri* datasets, and $w = 151$ was selected as the window size for the *VH* dataset. These parameters correspond to the lowest total error value for each dataset.

Window Size	**Datasets**											
	Kvasir-Capsule				**VH**				**Capri**			
	AUC	**Median Error**			**AUC**	**Median Error**			**AUC**	**Median Error**		
		Entrance	Exit	Total		Entrance	Exit	Total		Entrance	Exit	Total
11	95.66	58.25	982.00	1040.25	98.09	46.50	266.75	313.25	99.61	4.00	1.50	5.50
51	95.47	**55.75**	693.25	749.00	98.66	**31.25**	276.75	308.00	99.79	**2.75**	1.50	4.25
75	95.53	82.75	954.25	1037.00	98.59	35.75	260.00	295.75	99.74	3.00	1.50	4.50
101	94.60	75.75	1540.50	1616.25	98.55	37.25	259.00	296.25	99.76	3.50	1.75	5.25
151	95.41	111.75	1082.75	1194.50	98.54	41.50	**210.75**	**252.25**	99.78	6.25	1.50	7.75
201	**96.00**	76.50	**487.25**	**563.75**	**98.68**	53.75	260.00	313.75	**99.79**	**2.75**	**1.00**	**3.75**
251	93.71	92.75	758.00	850.75	98.43	43.50	218.00	261.50	99.62	3.50	2.25	5.75
301	95.63	76.00	777.00	853.00	98.41	42.50	218.00	260.50	99.70	**2.75**	**1.00**	**3.75**

Table 4.2: Grid search to find the best window size for each dataset. The proposed method was trained using different window size for the CMT block and the parameter that produces the best value, i.e., as the lowest total error, is individually selected for each dataset.

To demonstrate the effectiveness of the presented methodology, for each of the two experiments, two types of results are showcased. First, comparisons of the proposed strategy with state-of-the-art methods, followed by ablation studies illustrating the impact of each specific addition to the method. In particular, each added feature is individually tested. Table 4.3 introduces the nomenclature used in these ablation experiments.

For comparison purposes, the baseline ResNet50 model, which lacks any of the new features proposed, is trained and evaluated. The "Temp. Block" refers to the concatenation of the frame embedding with the temporal signal for each frame. The other three settings relate to the three signals detailed in the Methodology section: s_p, s_m, and s_t. Note that the proposed method incorporates all the features, namely, *ResNet + Time + CMT*. Henceforth, the term "Proposed Method" will be used to refer to this model.

4.4.1 Image Classification

The first set of experiments focused on classifying images as either inside or outside of the organ. For the *Kvasir-Capsule* and *VH* datasets, the organ was the small intestine, while for the *Capri* dataset, the organ was the colon. In Table 4.4, the proposed method is compared with those from Zou et al. (2015), Chen et al. (2017), Zhao et al. (2021), and Son

Method	Temp. block	Prob. sig., s_p	Mot. sig., s_m	Temp. sig., s_t
ResNet				
ResNet + C		✓		
ResNet + CM		✓	✓	
ResNet + CT		✓		✓
ResNet + CMT		✓	✓	✓
ResNet + Time	✓			
ResNet + Time + C	✓	✓		
ResNet + Time + CM	✓	✓	✓	
ResNet + Time + CT	✓	✓		✓
ResNet + Time + CMT (Proposed Method)	✓	✓	✓	✓

Table 4.3: Ablation study nomenclature. Each feature is introduced incrementally. This notation is used in the subsequent result tables. The first column indicates the temporal feature concatenated to the embedding, while the last three columns represent the probability, motion, and time signals, respectively.

et al. (2022). The results demonstrate that the proposed method outperforms all others in almost all metrics and is comparable to the second-best in those metrics where it does not rank first.

Following the notation in Table 4.3, Table 4.5 presents the ablation experiments. The introduction of the different components of the method significantly improves all metrics. Across all datasets, the temporal block enhances the performance compared to the baseline *ResNet*. Similarly, models incorporating the context block outperform the baselines (*ResNet* and *ResNet + Time*). Generally, adding time or motion to the context block leads to even better results, indicating that the combination of visual, temporal, and contextual information creates a powerful discriminative model. The highest performance of the proposed model is an AUC value of 99.79% on the *Capri* dataset. For *Kvasir-Capsule* and *VH*, the obtained AUC values are 96.00% and 98.54%, respectively.

Several factors contribute to the performance differences observed among datasets. The primary distinctions include: 1) the specific organ under study (colon for *Capri* versus small bowel for *Kvasir-Capsule* and *VH*); 2) the type of capsule device used (Olympus EC-S10, Medtronic PillCam SB3, and Medtronic PillCam COLON 2); and 3) variations in intestinal content volume. Consequently, differences in capsule characteristics such as optics, illumination, and resolution, as well as variations in intestinal mucosa and content, contribute to these disparities. Additionally, the statistical characteristics of each dataset, as summarized in Table 4.1, further underscore these distinctions. Despite these variations, the results exhibit coherence across the different datasets.

Dataset	Methods	AUC (%)	ACC (%)	MACC (%)	SPEC (%)	SENS (%)
Kvasir-Capsule	ResNet	91.48 ± 4.96	87.13 ± 7.00	82.10 ± 7.78	71.75 ± 15.43	92.45 ± 7.22
	Zou et al. (2015)	75.37 ± 9.42	69.51 ± 11.67	69.11 ± 8.68	70.19 ± 16.35	68.03 ± 14.60
	Chen et al. (2017)	83.65 ± 10.37	82.38 ± 9.20	76.90 ± 10.28	67.95 ± 16.18	85.84 ± 10.96
	Zhao et al. (2021)	94.05 ± 4.50	89.46 ± 7.72	85.09 ± 7.87	76.29 ± 14.40	93.89 ± 7.46
	Son et al. (2022)	95.75 ± 4.85	90.96 ± 6.56	81.03 ± 12.55	64.42 ± 26.40	$\mathbf{97.64 \pm 4.40}$
	Proposed Method	$\mathbf{96.00 \pm 4.57}$	$\mathbf{91.36 \pm 5.75}$	$\mathbf{87.47 \pm 7.49}$	$\mathbf{78.91 \pm 16.28}$	96.03 ± 4.29
VH	ResNet	94.42 ± 6.70	84.60 ± 9.59	86.26 ± 8.36	88.26 ± 13.96	84.25 ± 10.27
	Zou et al. (2015)	90.05 ± 9.91	84.56 ± 11.00	74.78 ± 12.22	56.87 ± 24.96	92.68 ± 9.98
	Chen et al. (2017)	95.86 ± 5.46	90.29 ± 8.12	87.69 ± 8.28	82.53 ± 15.62	92.84 ± 10.02
	Zhao et al. (2021)	97.81 ± 4.24	93.56 ± 7.12	91.95 ± 8.04	87.54 ± 14.89	96.37 ± 5.07
	Son et al. (2022)	96.46 ± 6.65	89.27 ± 9.35	90.46 ± 8.73	$\mathbf{91.05 \pm 14.59}$	89.88 ± 9.21
	Proposed Method	$\mathbf{98.54 \pm 2.36}$	$\mathbf{94.58 \pm 5.17}$	$\mathbf{92.26 \pm 7.74}$	87.25 ± 15.78	$\mathbf{97.27 \pm 3.42}$
Capri	ResNet	99.09 ± 1.41	95.71 ± 3.67	92.36 ± 4.62	85.70 ± 9.06	99.00 ± 3.16
	Zou et al. (2015)	86.06 ± 7.93	80.64 ± 12.24	65.93 ± 6.29	33.50 ± 12.51	98.35 ± 2.02
	Chen et al. (2017)	95.28 ± 4.37	88.42 ± 7.84	88.69 ± 6.49	88.31 ± 10.25	89.07 ± 9.62
	Zhao et al. (2021)	99.85 ± 0.47	98.59 ± 2.23	98.17 ± 2.94	97.76 ± 3.74	98.58 ± 4.14
	Son et al. (2022)	$\mathbf{99.93 \pm 0.21}$	97.94 ± 2.74	96.06 ± 4.30	92.57 ± 8.22	$\mathbf{99.57 \pm 2.58}$
	Proposed Method	99.79 ± 0.82	$\mathbf{99.07 \pm 2.12}$	$\mathbf{98.96 \pm 2.21}$	$\mathbf{98.58 \pm 3.48}$	99.35 ± 2.97

Table 4.4: Comparison of the proposed method with the previous ones. The results clearly show that the proposed method outperforms the others in almost all the metrics.

4.4.2 Landmark Localization

The second set of experiments aimed to locate the exact transition frame between organs by comparing the predicted values with annotations provided by expert physicians. For the *Kvasir-Capsule* and *VH* datasets, this involved identifying the anatomical landmarks at the entrance and exit of the small bowel: specifically, the last pylorus image as the entrance and the ileocecal valve as the exit. In the *Capri* dataset, the task focused on identifying the boundaries of the large bowel (colon), i.e., the first cecal image as the entrance and the last rectal image as the exit.

As detailed in the methodology section, the boundary identification method consisted of minimizing a rectangular pulse against the signal outputted by the method. Table 4.6 presents a comparison of the proposed method with previous methods described in the literature. Error is reported in terms of the number of frames and time, where available. It is evident that the proposed method significantly outperforms all previous strategies.

Table 4.7 shows the ablation study of the different features introduced in the method for this second task. The introduction of each of the different features clearly improves the baseline method as can be seen in the last row of each dataset.

Dataset	Methods	AUC (%)	ACC (%)	MACC (%)	SPEC (%)	SENS (%)
Kvasir-Capsule	ResNet	91.48 ± 4.96	87.13 ± 7.00	82.10 ± 7.78	71.75 ± 15.43	92.45 ± 7.22
	ResNet + C	93.53 ± 5.51	87.24 ± 13.04	82.78 ± 10.18	73.05 ± 16.82	92.50 ± 11.90
	ResNet + CM	92.70 ± 6.22	88.47 ± 11.18	83.95 ± 9.71	74.41 ± 16.75	93.49 ± 10.66
	ResNet + CT	94.40 ± 6.22	87.65 ± 11.59	83.72 ± 11.06	75.45 ± 18.98	91.98 ± 11.32
	ResNet + CMT	95.47 ± 5.39	90.07 ± 7.25	85.18 ± 8.86	75.23 ± 17.76	95.12 ± 6.54
	ResNet + Time	92.40 ± 5.06	87.88 ± 6.04	81.16 ± 7.66	67.92 ± 15.42	94.40 ± 5.27
	ResNet + Time + C	94.91 ± 4.29	89.53 ± 6.73	**87.69 ± 7.31**	**82.11 ± 15.57**	93.27 ± 6.96
	ResNet + Time + CM	94.67 ± 5.24	89.87 ± 6.54	87.39 ± 7.48	80.83 ± 15.88	93.95 ± 6.30
	ResNet + Time + CT	**96.36 ± 3.98**	90.80 ± 5.80	87.62 ± 7.77	80.21 ± 16.73	95.03 ± 4.92
	Proposed Method	96.00 ± 4.57	**91.36 ± 5.75**	87.47 ± 7.49	78.91 ± 16.28	**96.03 ± 4.29**
VH	ResNet	94.42 ± 6.70	84.60 ± 9.59	86.26 ± 8.36	88.26 ± 13.96	84.25 ± 10.27
	ResNet + C	96.28 ± 6.93	93.44 ± 6.27	91.89 ± 8.01	87.31 ± 15.88	96.47 ± 3.97
	ResNet + CM	97.70 ± 4.05	93.78 ± 5.94	91.97 ± 7.86	87.15 ± 15.69	96.79 ± 3.93
	ResNet + CT	97.81 ± 3.69	92.59 ± 6.93	91.12 ± 8.37	86.10 ± 17.00	96.15 ± 4.68
	ResNet + CMT	97.98 ± 3.31	93.49 ± 6.52	92.02 ± 8.06	87.55 ± 15.86	96.49 ± 4.19
	ResNet + Time	95.97 ± 6.28	88.64 ± 8.16	89.24 ± 7.57	**89.94 ± 12.64**	88.56 ± 9.00
	ResNet + Time + C	97.17 ± 6.35	94.63 ± 5.84	**92.58 ± 8.06**	88.00 ± 16.06	97.16 ± 3.85
	ResNet + Time + CM	98.13 ± 3.49	94.55 ± 5.84	92.41 ± 8.07	87.56 ± 15.96	97.26 ± 3.81
	ResNet + Time + CT	98.20 ± 3.79	**94.69 ± 5.79**	92.55 ± 7.94	87.25 ± 15.89	**97.84 ± 3.28**
	Proposed Method	**98.54 ± 2.36**	94.58 ± 5.17	92.26 ± 7.74	87.25 ± 15.78	97.27 ± 3.42
Capri	ResNet	99.09 ± 1.41	95.71 ± 3.67	92.36 ± 4.62	85.70 ± 9.06	99.00 ± 3.16
	ResNet + C	99.83 ± 0.81	98.63 ± 3.40	98.51 ± 3.30	98.50 ± 3.63	98.52 ± 5.59
	ResNet + CM	99.82 ± 0.74	98.70 ± 3.21	98.37 ± 3.56	97.83 ± 5.01	98.90 ± 4.94
	ResNet + CT	99.79 ± 0.92	98.54 ± 3.51	98.51 ± 3.13	98.54 ± 3.03	98.47 ± 5.69
	ResNet + CMT	99.86 ± 0.54	98.73 ± 3.18	98.42 ± 3.54	98.01 ± 4.90	98.84 ± 5.07
	ResNet + Time	99.66 ± 0.76	97.18 ± 3.50	96.51 ± 3.37	93.82 ± 6.46	99.20 ± 2.28
	ResNet + Time + C	99.88 ± 0.54	98.97 ± 2.22	98.79 ± 2.36	98.08 ± 4.19	99.51 ± 2.28
	ResNet + Time + CM	99.59 ± 1.52	98.91 ± 2.55	98.76 ± 2.51	97.92 ± 4.66	**99.60 ± 2.11**
	ResNet + Time + CT	**99.90 ± 0.47**	99.02 ± 2.01	98.90 ± 2.16	98.43 ± 3.49	99.36 ± 2.81
	Proposed Method	99.79 ± 0.82	**99.07 ± 2.12**	**98.96 ± 2.21**	**98.58 ± 3.48**	99.35 ± 2.97

Table 4.5: Ablation study to show that the introduction of the different components of the proposed method makes the model to improve gradually.

The minimization of the rectangular pulse in the final step (Step 3) of our pipeline adds significant value by enabling the method to accurately identify the best transition frames between organs. To conclude the quantitative analysis, this final step of identifying the boundaries was incorporated into the methods proposed by Zhao et al. (2021) and Son et al. (2022). The aim is to demonstrate that this step enhances the accuracy of landmark identification. Table 4.8 supports this hypothesis, showing that the inclusion of this final process reduces the error in identifying landmarks. This improvement underscores the importance of the rectangular pulse minimization in achieving precise boundary detection between organs.

4.4.3 Qualitative Results

Figure 4.7 presents a qualitative analysis of one video from each dataset. Each plot represents the probability of a specific frame belonging to the inside of an organ, denoted as

Dataset	Methods	Entrance				Exit			
		MAE		Median		MAE		Median	
		Frame	Time	Frame	Time	Frame	Time	Frame	Time
Kvasir-Capsule	Zhao et al. (2021)	2644.16 ± 4637.65	-	1251.00 ± 115.25	-	4603.58 ± 1545.95	-	1669.00 ± 185.26	-
	Son et al. (2022)	2711.00 ± 3435.83	-	1786.00 ± 231.93	-	2409.00 ± 3106.30	-	1506.75 ± 1161.42	-
	Proposed Method	$\mathbf{465.88 \pm 918.13}$	-	$\mathbf{76.50 \pm 46.50}$	-	$\mathbf{1679.67 \pm 2775.72}$	-	$\mathbf{487.25 \pm 163.75}$	-
VH	Zhao et al. (2021)	1304.23 ± 1394.26	00 : 08 : 20	915.25 ± 22.98	00 : 04 : 34	3308.58 ± 583.28	00 : 31 : 44	1765.25 ± 461.03	00 : 16 : 59
	Son et al. (2022)	1390.47 ± 3487.30	00 : 07 : 12	304.75 ± 220.97	00 : 02 : 14	1552.00 ± 520.78	00 : 16 : 47	627.75 ± 1469.01	00 : 08 : 09
	Proposed Method	$\mathbf{443.69 \pm 1064.05}$	**00 : 02 : 38**	$\mathbf{41.50 \pm 11.00}$	**00 : 00 : 15**	$\mathbf{837.77 \pm 1485.79}$	**00 : 09 : 46**	$\mathbf{210.75 \pm 164.75}$	**00 : 03 : 14**
Capri	Zhao et al. (2021)	214.24 ± 437.6	00 : 07 : 28	85.0 ± 23.33	00 : 01 : 11	524.94 ± 1258.34	00 : 35 : 23	23.5 ± 2.12	00 : 01 : 02
	Son et al. (2022)	56.39 ± 161.94	00 : 04 : 00	14.50 ± 0.70	00 : 00 : 08	32.25 ± 111.27	00 : 07 : 29	7.50 ± 0.70	00 : 00 : 43
	Proposed Method	$\mathbf{29.40 \pm 83.94}$	**00 : 00 : 55**	$\mathbf{2.75 \pm 0.25}$	**00 : 00 : 01**	$\mathbf{7.91 \pm 37.82}$	**00 : 01 : 47**	$\mathbf{1.00 \pm 0.00}$	**00 : 00 : 01**

Table 4.6: This figure compares the proposed method with two prior approaches from the literature. The results show that the proposed method significantly outperforms the others, achieving lower errors in the number of frames and time relative to actual landmarks annotated by expert physicians.

signal $s(\cdot)$. The four rows depicted show the results of training the proposed model with varying feature sets described in the methodology section. In the first row, a plain *ResNet* is trained. The second row adds the "Temporal" block. The third row includes the "CMT" block, which incorporates the three signals of Probability, Time, and Motion. Finally, the last row combines all features: "Temporal" and the "CMT" block.

These plots are color-coded: blue corresponds to frames identified by experts as outside the organ, and yellow to those inside. The green dashed line indicates the expert-identified transition frames from outside to inside, while the purple dashed line shows the transition detected by the proposed method. In the *VH* dataset, it is particularly evident that the addition of these extra features to the model results in more accurate landmark identification.

The sequence near the expert annotations and the frames predicted by the method are presented using the same color palette for the landmarks: green for the expert-annotated transitions and purple for those predicted by the method. Finding the exact transition frame is extremely complex and requires reviewing the full sequence to account for the camera's movement.

Some false positive and false negative frames, those incorrectly predicted as outside or inside the organ, are displayed. Many of these frames are very similar to others in the opposite class or contain intestinal content that obscures part of the mucosa, complicating accurate identification.

Nevertheless, the plots demonstrate that the predicted boundaries are close to the real ones, highlighting the good performance of the proposed system.

Dataset	Methods	Entrance				Exit			
		MAE		Median		MAE		Median	
		Frame	Time	Frame	Time	Frame	Time	Frame	Time
Kvasir-Capsule	ResNet	668.34 ± 1091.84	-	111.75 ± 28.25	-	1875.96 ± 2747.54	-	1124.00 ± 493.00	-
	ResNet + C	1505.83 ± 3316.56	-	207.50 ± 153.00	-	2147.42 ± 2967.07	-	908.50 ± 757.50	-
	ResNet + CM	1504.54 ± 3272.05	-	217.25 ± 131.75	-	1877.62 ± 2736.03	-	928.00 ± 684.50	-
	ResNet + CT	1677.79 ± 3509.05	-	139.00 ± 51.50	-	1902.29 ± 2683.68	-	1220.50 ± 391.50	-
	ResNet + CMT	830.08 ± 1341.51	-	127.50 ± 31.00	-	1770.79 ± 2771.27	-	743.75 ± 580.25	-
	ResNet + Time	785.88 ± 1182.88	-	93.00 ± 15.50	-	2002.38 ± 2959.43	-	1077.00 ± 539.00	-
	ResNet + Time + C	535.50 ± 1089.54	-	**26.75 ± 2.25**	-	1710.71 ± 2769.83	-	559.00 ± 169.00	-
	ResNet + Time + CM	606.08 ± 1104.42	-	61.25 ± 9.25	-	1727.67 ± 2767.06	-	651.50 ± 383.00	-
	ResNet + Time + CT	556.38 ± 951.79	-	116.00 ± 43.00	-	1730.50 ± 2756.90	-	663.25 ± 83.25	-
	Proposed Method	**465.88 ± 918.13**	-	76.50 ± 46.50	-	**1679.67 ± 2775.72**	-	**487.25 ± 163.75**	-
VH	ResNet	667.70 ± 1070.13	00 : 04 : 00	220.75 ± 191.75	00 : 01 : 45	2710.98 ± 4505.58	00 : 22 : 02	1235.75 ± 1126.75	00 : 12 : 20
	ResNet + C	559.98 ± 1007.92	00 : 03 : 11	78.50 ± 17.00	00 : 00 : 34	1290.33 ± 2117.18	00 : 14 : 52	198.50 ± 82.00	00 : 04 : 02
	ResNet + CM	512.38 ± 870.20	00 : 02 : 59	91.75 ± 48.75	00 : 00 : 39	1028.02 ± 1704.22	00 : 12 : 14	199.50 ± 85.50	00 : 03 : 24
	ResNet + CT	557.83 ± 1181.40	00 : 03 : 10	78.25 ± 16.75	00 : 00 : 32	1279.73 ± 1774.48	00 : 14 : 55	369.75 ± 3.75	00 : 05 : 53
	ResNet + CMT	500.90 ± 1166.62	00 : 02 : 45	50.50 ± 17.0	00 : 00 : 25	1077.35 ± 1605.62	00 : 12 : 05	259.50 ± 109.00	00 : 05 : 40
	ResNet + Time	731.71 ± 1451.30	00 : 03 : 57	103.50 ± 59.50	00 : 01 : 14	2050.94 ± 4142.68	00 : 17 : 01	397.00 ± 350.00	00 : 06 : 26
	ResNet + Time + C	502.62 ± 877.55	00 : 02 : 47	59.50 ± 4.00	00 : 00 : 25	911.23 ± 1619.69	00 : 11 : 38	167.25 ± 106.25	00 : 03 : 03
	ResNet + Time + CM	**417.79 ± 814.67**	**00 : 02 : 18**	44.00 ± 7.00	**00 : 00 : 10**	1110.48 ± 1992.44	00 : 13 : 02	**166.50 ± 103.00**	**00 : 03 : 00**
	ResNet + Time + CT	443.46 ± 1092.63	00 : 02 : 31	50.25 ± 8.25	00 : 00 : 23	1051.17 ± 1933.66	00 : 11 : 37	225.25 ± 170.25	00 : 03 : 29
	Proposed Method	443.69 ± 1064.05	00 : 02 : 38	**41.50 ± 11.00**	00 : 00 : 15	**837.77 ± 1485.79**	**00 : 09 : 46**	210.75 ± 164.75	00 : 03 : 14
Capri	ResNet	53.70 ± 110.44	00 : 01 : 19	14.50 ± 4.00	00 : 00 : 06	13.80 ± 48.35	00 : 02 : 19	1.00 ± 0.00	**00 : 00 : 01**
	ResNet + C	28.94 ± 85.33	00 : 00 : 59	4.50 ± 0.50	00 : 00 : 02	41.76 ± 215.63	00 : 04 : 44	3.00 ± 0.00	**00 : 00 : 01**
	ResNet + CM	32.43 ± 97.72	00 : 01 : 00	2.75 ± 0.25	**00 : 00 : 01**	38.58 ± 210.08	00 : 03 : 01	1.75 ± 0.25	**00 : 00 : 01**
	ResNet + CT	32.62 ± 91.01	00 : 01 : 02	5.00 ± 1.00	00 : 00 : 02	48.45 ± 240.70	00 : 07 : 08	2.00 ± 0.50	00 : 00 : 02
	ResNet + CMT	32.35 ± 100.58	00 : 01 : 01	2.50 ± 0.50	**00 : 00 : 01**	37.71 ± 209.38	00 : 03 : 06	2.00 ± 0.00	**00 : 00 : 01**
	ResNet + Time	38.40 ± 86.46	00 : 01 : 11	5.00 ± 2.00	00 : 00 : 02	19.80 ± 114.86	00 : 05 : 37	**1.00 ± 0.00**	**00 : 00 : 01**
	ResNet + Time + C	29.57 ± 81.18	00 : 00 : 51	4.50 ± 1.50	**00 : 00 : 01**	8.89 ± 38.02	00 : 02 : 13	**1.00 ± 0.00**	**00 : 00 : 01**
	ResNet + Time + CM	**23.58 ± 69.27**	**00 : 00 : 41**	3.50 ± 0.50	**00 : 00 : 01**	8.26 ± 38.02	00 : 02 : 04	**1.00 ± 0.00**	**00 : 00 : 01**
	ResNet + Time + CT	30.40 ± 79.19	00 : 00 : 58	5.25 ± 3.25	**00 : 00 : 01**	8.88 ± 38.15	00 : 02 : 09	**1.00 ± 0.00**	**00 : 00 : 01**
	Proposed Method	29.40 ± 83.94	00 : 00 : 55	**2.75 ± 0.25**	**00 : 00 : 01**	**7.91 ± 37.82**	**00 : 01 : 47**	**1.00 ± 0.00**	**00 : 00 : 01**

Table 4.7: This figure illustrates the incremental introduction of each feature in the proposed method, demonstrating the value added by each component.

4.5 Conclusions

This chapter introduces and validates a deep learning solution for organ segmentation, a crucial initial step for physicians when reviewing a CCE video.

Identifying anatomical landmarks is a complex task due to its reliance on neighboring frames. To address this challenge, the proposed method incorporates multiple modules designed to enhance the accuracy of landmark identification.

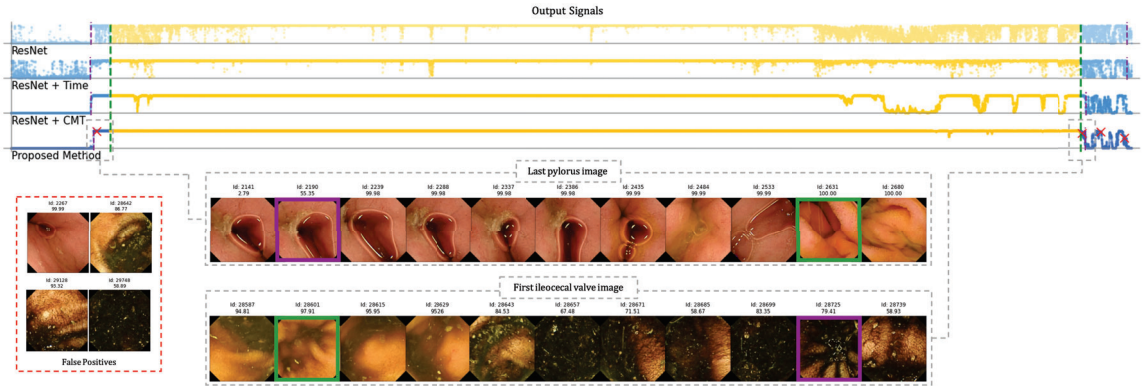
The system first integrates temporal information for each frame, which helps to accurately position the frame within the sequence of the video. Next, it introduces motion information, a feature that measures the similarity or dissimilarity between neighboring frames, providing insights into camera movement. Similar frames indicate slow camera movement, whereas significant differences between adjacent frames suggest rapid movement.

These two features, temporal and motion information, are then utilized to smooth the probability signal that determines whether a frame is inside or outside of an organ, thereby significantly improving boundary detection accuracy.

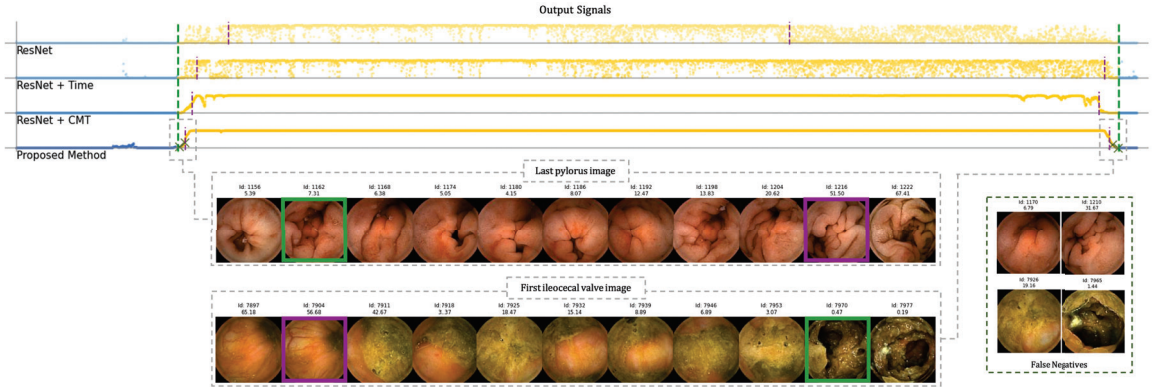
We conducted extensive experiments to compare the proposed method with existing approaches in the literature, achieving superior results. The inclusion of motion and temporal features proved to be highly beneficial, as demonstrated by incorporating these features into existing methods, which led to enhanced performance in all experiments."

Method	Dataset					
	Kvasir-Capsule		VH		Capri	
	Entrance	Exit	Entrance	Exit	Entrance	Exit
Zhao et al. (2021)	1251.00	1669.00	915.25	1765.25	85.00	23.50
Zhao et al. (2021) + Step 3	93.00	702.50	65.00	478.00	3.50	1.00
Son et al. (2022)	1786.00	1506.00	304.75	627.75	14.50	7.50
Son et al. (2022) + Step 3	686.25	1683.25	214.50	1236.00	35.50	6.00
Proposed Method	**76.50**	**487.25**	**41.50**	**210.75**	**2.75**	**1.00**

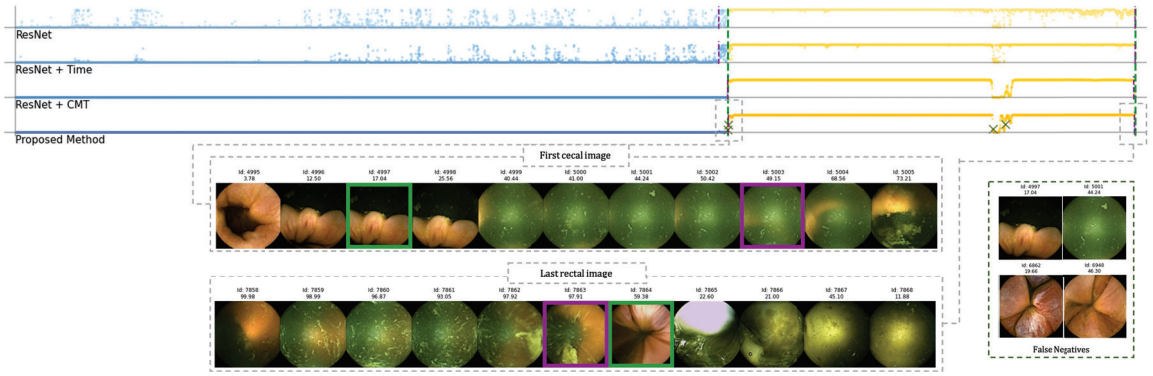
Table 4.8: Adding the final step of the presented system to existing methods helps improve the detection of anatomical landmarks.



(a) Kvasir-Capsule dataset



(b) VH dataset



(c) Capri dataset

Figure 4.7: Visual comparison of the methods. Each subfigure displays a random video from one of the datasets. For each dataset, four models are shown with the gradual addition of features introduced in the methodology. Real annotated landmarks and predicted ones are overlaid on the plot in green and purple dashed lines, respectively. The green lines indicate the transition between outside (blue dots) and inside the organ (yellow dots). To better understand these transition frames, the sequence is presented visually. Some misclassified frames are also shown in a side figure.

Chapter 5

Polyp Detection

Contents

The last step in the video review process is to identify pathologies. Colorectal cancer is the third most common type of cancer worldwide and ranks second on the list of most aggressive and deadly cancers. One of the initial signs of the development of colon cancer is the appearance of polyps in the colon that grow in an uncontrolled manner. The detection of polyps when they are still small is crucial to prevent their transformation into cancer. Screening programs are aimed at detecting early-stage cancer, improving the patient's chances of survival. This chapter presents existing AI solutions for polyp detection.

5.1 AI for polyp detection

Since the public release of CE technology in Iddan et al. (2000), extensive research has been conducted on polyp detection. Early studies focused on shape descriptors (Li et al., 2009), geometric features (Figueiredo et al., 2010), and custom-defined filters based on color and illumination (Hwang and Celebi, 2010).

With the advent of ML, more sophisticated algorithms were introduced, including those based on support vector machines (Li and Meng, 2012; Yuan and Meng, 2014) and custom-designed algorithms for polyp detection and segmentation (Mamonov et al., 2014).

As deep learning emerged, the focus shifted to detecting polyps using CNNs (Yuan and Meng, 2017; Yuan et al., 2020; Guo and Yuan, 2019; Laiz et al., 2020).

With the appearance of the ViT, we tested a new method for polyp detection. Following the same methodology and data as in Laiz et al. (2020), the ViT architecture was evaluated for polyp detection using 120 videos in a 5-fold cross-validation setup to ensure robust results. The images were resized to $224 \times 224 \times 3$ to match the architecture's default input shape. A patch size of 16×16 was selected, resulting in 196 patches per image $(224/16)^2$. The network was initialized with pre-trained ImageNet weights, including both the network weights and the learned positional embeddings of the patches. The Transformer was configured with 12 ViT blocks, each comprising an attention module with 12 heads. The internal embedding dimension was set to 768, and the MLP was defined with 3,072 units. The code from Gilabert (2024) can be adjusted to match the specified configuration.

To train the network, Binary Cross-Entropy (BCE) was used with Stochastic Gradient Descent (SGD) as optimizer. The learning rate was initialy set to 2.5×10^{-3} and reduced by a factor of 0.2 on training step 8k and again by a factor of 0.02 on step 16k.

The addition of the Sharpness-Aware Minimization (SAM) (Foret et al., 2020) algorithm to the base SGD optimizer was also tested. This addition boosted the results of the polyp detection method, achieving the best results. Table 5.1 shows the comparison of these methods. Most research papers evaluate network performance using private test datasets

Method	AUC	Spec@99%	Spec@95%	Spec@90%	Spec@80%
SSAEIM (Yuan and Meng, 2017)	57.76	-	6.98	13.29	27.82
UDCS (Yuan et al., 2020)	88.64	-	70.44	78.22	83.31
ANET (Guo and Yuan, 2019)	90.44	-	72.02	78.92	85.23
ResNet50 + TL (Laiz et al., 2020)	93.65	53.56	74.17	81.77	88.41
EfficientNetB3 + TL	94.14	61.83	77.90	83.60	88.89
Vision Transformer + BCE	97.13	64.89	84.64	91.82	96.46
Vision Transformer + BCE + SAM	**97.45**	**66.94**	**86.01**	**92.67**	**97.07**

Table 5.1: Polyp detection results. Some results are extracted from Laiz et al. (2020).

and report the results. In the final section of this thesis, two studies are presented, each deploying a polyp detection algorithm in real clinical practice. The first study evaluates

the method from Laiz et al. (2020), where a tool incorporating this network was developed and tested by physicians who assessed its output and usefulness. The second study, still in a preliminary stage, introduces a more advanced architecture, which was deployed and tested on a larger population.

5.2 Reviewing videos with the RAPID tool

Before going into detail in the specific configuration of the studies for polyp detection, let's revise how the clinicians are currently revising CE videos to search for pathologies.

RAPID Reader v9.0 is a proprietary software developed by Medtronic for reviewing and interpreting CE videos. Widely used by professionals specializing in CCE, this software is designed for comfortable and efficient video review. The software displays the video in temporal order and includes features such as image labeling, commenting, and measuring with an integrated ruler function. Users can adjust the replay speed of the video at any time, and the application indicates the approximate area of the abdomen where the capsule is located. Additionally, it alerts users if the capsule is moving through an area at high speed, signaling that special attention may be needed.

RAPID offers several modes for video review: single view with the frontal camera, twin head mode displaying both cameras simultaneously (Figure 5.1), or Top 100 mode, which shows a batch of frames selected using traditional computer vision techniques (Figure 5.2).

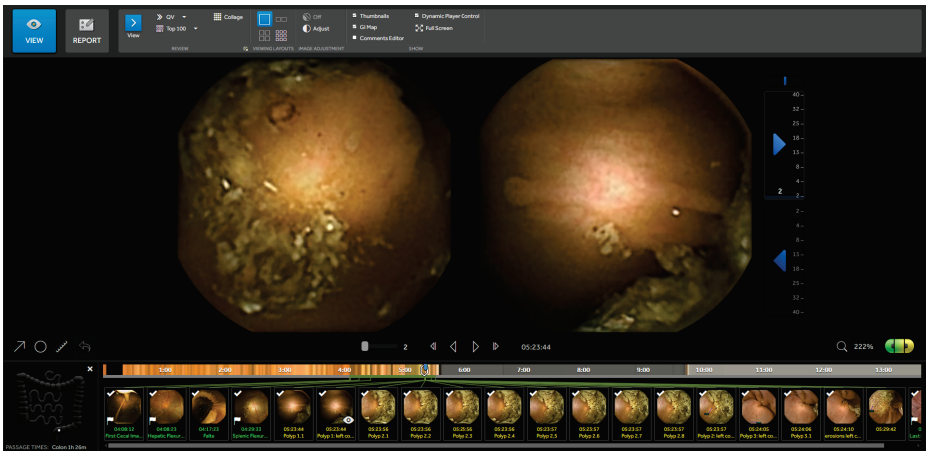


Figure 5.1: RAPID Reader Software v9.0: screen with images from both camera heads (green/yellow) and marked thumbnails.

The first part of a clinician's job is to gather patient information, including medical history, current medications, and, most importantly, the reason for referral for a CE. This information is crucial for analyzing the video with the necessary detail and focus, especially when looking for a specific pathology or condition (Yamamoto et al., 2017; Rondonotti et al., 2020; Koulaouzidis et al., 2021; Pennazio et al., 2023).

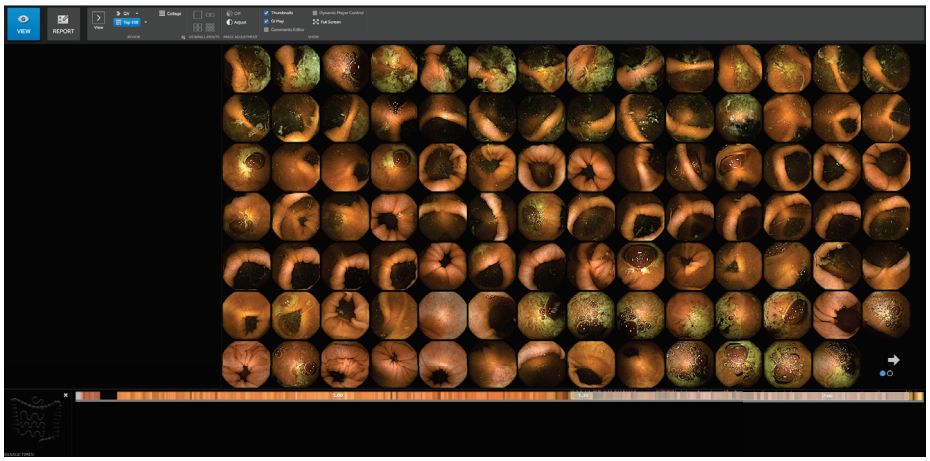


Figure 5.2: The Top 100 mode displays the most important frames the physician need to review, summarizing the video.

The next step is to open the video in the RAPID software and begin the analysis. After confirming that the video is admissible, ensuring that intraluminal content does not obstruct a complete view of the intestinal walls, the first task is to perform a quick overview of the entire video to identify key anatomical landmarks for detailed review (Koulaouzidis et al., 2021). The frame rate is then slowed to 8-15 frames per second (Koulaouzidis et al., 2021; Rondonotti et al., 2018) to carefully detect all potential pathologies or significant frames that will be included in the final report. The RAPID software's Top100 mode can assist in highlighting important frames. Although the CE may take up to 5 hours to traverse the small bowel, a physician should dedicate 40-50 minutes to thoroughly review the video. This process is complex and requires a high level of concentration (Rondonotti et al., 2020).

Finally, a detailed report must be written, including comprehensive information on the detected pathologies, measurements of polyps if found, the number of different polyps identified, and the duration of the video within each organ to identify any potential motility disorders. A more detailed specification of a quality CCE report can be found in Koulaouzidis et al. (2021).

5.3 Improving the polyp detection with the AI-Tool

The AI-Tool is a software designed to assist clinicians in the detection of polyps, by complementing any proprietary video reviewing software, such as RAPID. It embeds a CNN into a web tool that presents images with potential polyps to the user in a sequence of declining certainty. Therefore, images that are very likely to contain a polyp will appear first. At the time this study was started, the model from Laiz et al. (2020) was the state of the art for polyp detection so it was chosen among other candidates (Yuan and Meng, 2017; Yuan et al., 2020; Guo and Yuan, 2019) as the core of the AI-Tool. The hyperparameters of the model were fixed after a 5-Fold cross validation process using 120 CCE videos (2,080 polyp

images and 246k negative images). Different hyperparameters were tested (the same for the five models) and those that gave the best results in the validation sets were selected. A single model was then trained using the 120 CCE and embedded into the AI-Tool. The experimental validation of the network has shown a sensitivity over 90% at a specificity of 95% when evaluated in a fully automatic setting (when no expert is involved) using full videos. All 120 videos used in the training of the CNN were excluded for this experiment.

The AI-Tool computes two outputs using a CNN: a probability score per frame to contain a polyp and a heatmap to visualize the reasoning behind the score using CAM (Zhou et al., 2016), an algorithm that uses the values of the latest CNN layers to display the image areas most relevant for classification. In this particular case, this method presents the most relevant image zones that allow the CNN to classify an image as polyp.

Each potential polyp image is displayed along with eight context frames, the four preceding and the four following it (Figure 5.3). For each frame, a colored square is shown to indicate the probability of it being a polyp using a colorblind friendly palette that can be customized when the application starts. Each image can be enlarged by clicking on it, then further information is presented such as the probability or the timestamp of the sequence. The image sequence can be also displayed as a video using the left and right keyboard arrows. Heatmaps can always be activated showing the most likely area to contain a polyp (Figure 5.4). A further benefit of the heatmap is that it helps readers to understand the reasoning behind the polyp probability and as a consequence, increase their trust in the system.

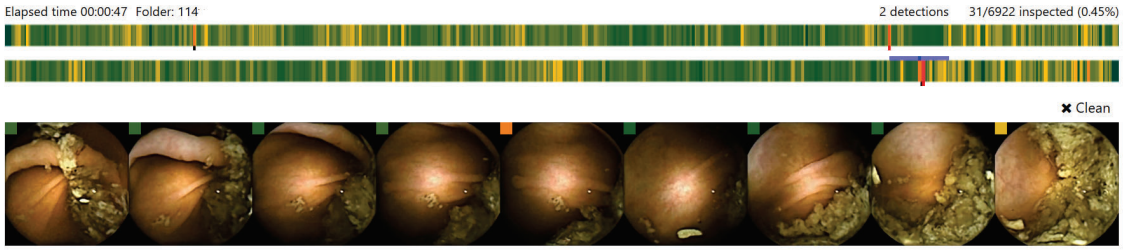
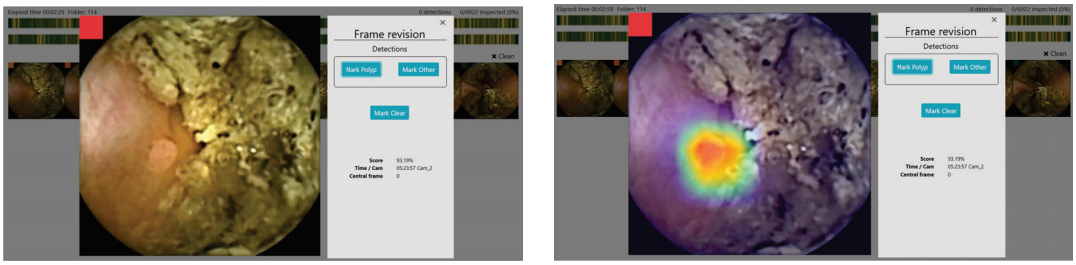


Figure 5.3: Candidate polyp sequence displayed in the AI-Tool. Each colored bar shows the probabilities for a polyp in one head of the capsule. The proposed image is presented in the center frame and 4 context images are placed by each side.

5.3.1 Study Population

Eighteen videos of patients with at least one colon polyp obtained using the PillCam COLON 2 capsule were randomly selected for this experiment following a Simple Randomization strategy. The data used in this study are retrospective CCE videos from patients that were conducted on behalf of the NHS Highland Raigmore Hospital in Inverness. All patients from this study came from referrals for symptoms or were on surveillance lists within the Highlands and Islands area of Scotland and had a positive FIT. Referrals and



(a) Images can be enlarged for better visualization

(b) Heatmap activated.

Figure 5.4: A heatmap layer can be activated at any moment to visualize where the tool is focusing when classifying an image as polyp.

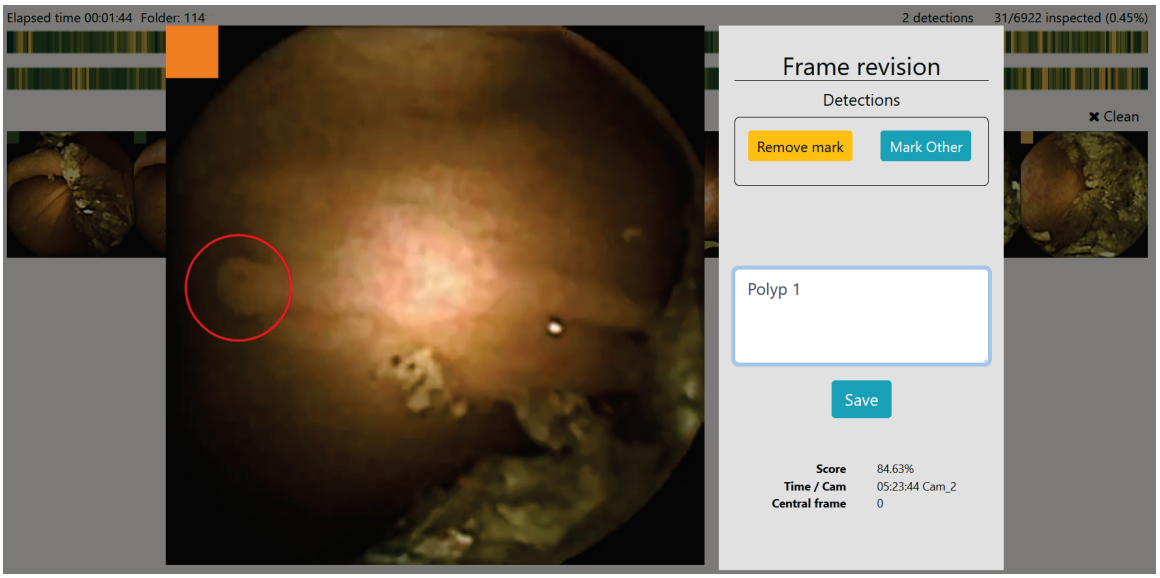


Figure 5.5: Detailed view of the tool. The user is able to highlight anything relevant in the image.

final diagnoses were made based on local considerations, outside the influence of the teams conducting CCE procedures. Bowel preparation in accordance with a standardized, PEG-based, split-dose cleansing protocol was performed in all patients. All videos were obtained using PillCam COLON 2 which has two heads (front and rear). They were anonymized to protect patient information. Patients' mean age was 58.1 ± 18.7 years (range, 18–92 years) and mean colon transit time was 4h 10 min (range, 0.17–14.2 hours).

Before the experiments began, the two videos obtained from each of the heads of the capsule were meticulously reviewed by four independent CCE readers, experts from now on, in order to create the ground truth for the experiments (gold standard). Each detected polyp was assigned a unique identifier, the timestamp of the first and the last image where the polyp was visible, and from which head it was reported. The independent analysis of the experts was then shared with all experts, reaching a consensus in case of discrepancies. As

experts, we have arbitrarily considered CCE readers with at least three months of experience in CCE. All of them have formal training in reviewing CCE videos and they analyze about 5-20 videos a week. On a daily basis, they follow the standard review protocol to ensure that each video is reviewed in a consistent, repeatable and well-documented manner. The results of all of them were validated by a medical doctor with two years of experience who created a final report about the results. In no case were any concerns reported back by that clinician either from the review nor from any possible follow-up procedure about the quality of the report. Both the final report and the gold standard used as ground-truth for the experiment in this study are the responsibility of the medical doctor that approved the results.

During this process, a total of 52 unique polyps were found. The video with the most polyps had 7, while there were 5 videos with only one polyp. The polyps' size was estimated with RAPID. A total of 23 polyps were identified as large ($\geq$ 6mm) and 29 polyps as small ($<$ 6mm). There were 5 polyps larger than 10mm and only one polyp smaller than 3mm. The characterization of morphologies was done in accordance with the requirements from the referring clinicians, matching standard Paris classifications where possible. In no case polyps were selected or discarded based on size or morphology, all polyps reported by the physicians were included in the study.

5.3.2 Experimental Design

Three experienced CCE readers reviewed the videos selected for this experiment. Each reader reviewed half of the videos using the standard RAPID Reader Software v9.0 (Medtronic) and the other half using the AI-Tool. Results obtained by the experts using each of the tools are reported in terms of number of polyps detected (sensitivity) as well as time needed to complete the reviews (screening time).

The experiment conducted in this study was restricted to the images of the colon. Identification of the entrance and exit of the colon was previously provided to the readers. When they used RAPID software, they were asked to perform the standard screening procedure without any screening time limitation. No information other than a single video identifier was provided to the reviewers during the analysis of the videos. For the AI-Tool, the review time was limited to 30 minutes regardless of the video's length.

For both tools, readers were required to review the videos without pauses or external stimuli that could lead to distractions. During the review, the readers labelled the images that they identified as a polyp using the tools provided within each of the applications. In the case of RAPID, experts were asked to tag all the unique polyps they found. When using the AI-Tool, experts were asked to make a decision, *polyp*, *clear* or *other*, for each sequence that was presented to them.

5.3.3 CCE Readers

The expert readers are endoscopy nurses with at least 2 years of experience with CCE. They have a formal CCE training and conduct between 5-20 video analyses per week. On a daily basis, they follow a standard operating procedure to ensure that each video is analyzed in a consistent, repeatable way and documented according to common standards. None of these three readers were part of the gold standard creation process.

5.3.4 Results

Polyp detection

The overall sensitivity of polyp detection using RAPID as the screening procedure was 81.08% while using the AI-Tool the sensitivity increased to 87.80%. Table 5.2 shows the percentage of polyps found using both tools distinguishing between three categories: polyp size (in millimeters), visibility (in number of frames) and morphology. The sensitivity using RAPID turned out to be 76.92% for polyps smaller than 6mm and 85.71% for larger polyps. Both numbers increased when the AI-Tool was used (85.42% and 91.18% for small and large polyps respectively).

		#Polyps	***RAPID***	***AI-Tool***
Size	*Small (<6mm)*	29	76.92%	85.42%
	Large ($\geq$ 6mm)	23	85.71%	91.18%
Visibility	*Low (< 4 frames)*	9	58.33%	80.00%
	Normal (4 – 10 frames)	15	73.91%	81.82%
	High (> 10 frames)	28	92.31%	93.33%
Morphology	*Pedunculated*	4	100.00%	100.00%
	Sessile	25	81.82%	92.86%
	Flat	23	75.76%	80.56%
		52	81.08%	87.80%

Table 5.2: Detection of polyps distinguishing by size, visibility and morphology.

The biggest difference between both tools was observed in the visibility of the polyp. For polyps appearing in a few frames (low visibility), the table shows a significant improvement when using the AI-Tool. While RAPID achieved an accuracy of 58.33%, the AI-Tool reached 80.00%. This represents a 37.15% increase in this category. Smaller improvements using the AI-Tool were also reported for polyps appearing in a larger number of frames.

These results show that small polyps and polyps that appear for only a few frames are more likely to be detected using the AI-Tool than using the RAPID application.

CCE Screening Time

One of the aims of this study was to compare the time needed for the detection of polyps using RAPID and our AI-Tool. The average time required for the experiments performed with RAPID was 47.11 minutes (11.6 minutes for each hour of CCE video reviewed) with a maximum of 126 minutes. Let us recall that the time for analysis using the AI-Tool was fixed at 30 minutes for all the experiments.

Figure 5.6 shows the average sensitivity curve as a function of time for both applications. AI-Tool took only 8.00 minutes to reach the same accuracy as RAPID (point A). Since the mean of RAPID experiments was 47.11 minutes we can state that the AI-Tool reduces the time needed to reach the same accuracy as RAPID by a factor of $47.11/8.00 \approx 6$. We can also see that RAPID experiments reached a 55.41% of sensitivity (point B) by minute 30 when the AI-Tool experiments finished.

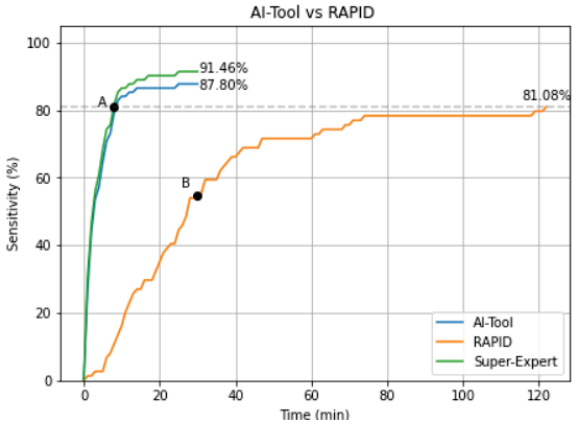


Figure 5.6: Mean sensitivity curve of the experiments using both applications. In green the Super-Expert curve (gold standard) that represents the maximum value that the blue line could reach. This curve has been calculated simulating an expert who never makes mistakes when identifying a polyp while using the AI-Tool.

The shape of the curves is also an aspect worth considering. While the RAPID curve has an almost linear behavior, the AI-Tool curve shows an initial steep slope and, after minute 16, it is almost flat. In the first 10 minutes of the analysis 84.14% of the polyps are detected. In the next 10 minutes, this number rises to 86.59%, which represents an increase of 2.91% in this period. Finally, only 1.21% of the polyps are detected in the last 10 minutes of the experiment. This indicates that our application is proposing the relevant images in the first minutes of visualization. It is also worth mentioning that without limiting the time to 30 minutes as we did, the detection of polyps may slightly increase because those with a very low score would have been presented to the reviewers. However, this was not the objective of our study, since we aimed to see if the video review could be done better and in a shorter period of time.

Qualitative Results

Heatmaps are a key element of our tool. They allow the medical staff to trust the system and make an informed decision on each image sequence. Figure 5.7 shows the heatmaps activation for three different categories. First, in the image on the left, we see high-scoring polyps. We observe how the heatmaps generated by the CAM algorithm are well defined and show the area where the polyp is located. In the center, we see polyps with a very low score, which the network erroneously classifies as negative. In this case, the heatmaps are not activated. Finally, in the image on the right we can see images that do not contain any polyp but to which the system assigns a high score.

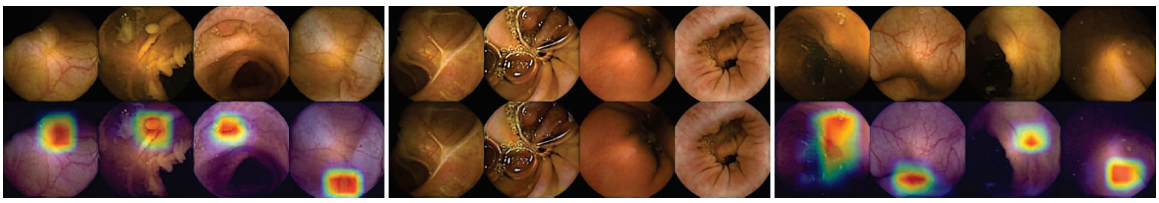


Figure 5.7: Left: Images of correctly identified polyps with their respective heatmaps (True Positives). Center: Images of polyps with a very low score (False Negatives). Right: Images that do not contain a polyp but still have a very high score (False Positives).

It can also be seen that the false positive patterns correspond to textures and morphologies compatible with polyps. In almost all of them a rounded area is displayed which, without the context of the other images, can be difficult to classify as a polyp image or not. In addition, this is a valuable information for the reviewers of the video, as the heatmap shows the area on which it is important to focus their attention.

We now focus on showing the images of polyps that have not been detected with either of the two applications (specificity). Figure 5.8 shows those polyps not detected with the AI-Tool because of the imposed time restriction (30 minutes). The score given to these image does not exceed 15%, therefore, the experts never reviewed them. In fact, these four images are the polyps with the lowest score of this study. The polyps of these images are difficult to find since they are partially occluded or do not present a regular morphology.

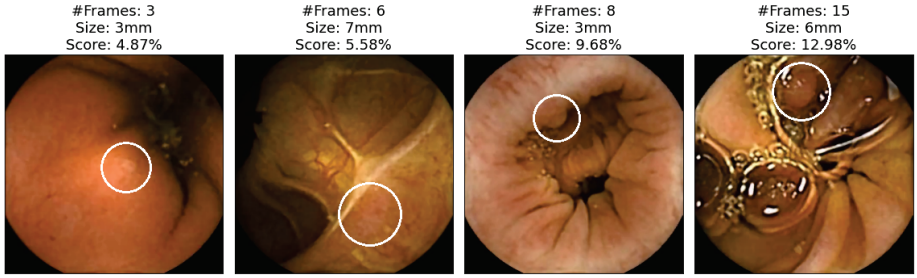


Figure 5.8: Polyp frames to which the app has attributed a small score and, therefore, none of the experts have been able to review in the first 30 minutes. Polyps are circled in white.

Figure 5.9 shows images of polyps reviewed and discarded by all the experts even though

the AI-Tool assigned them a remarkably high probability. These missed polyps are the result of human error or discrepancies between the video reviewer and the experts who generated the ground truth.

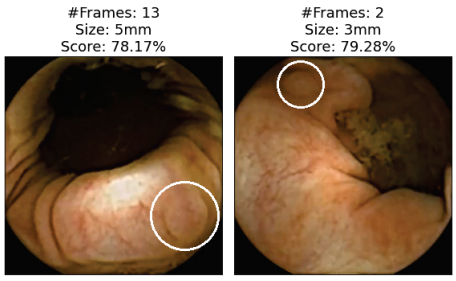


Figure 5.9: Images that all experts have reviewed and found not to be polyps. Polyps are circled in white.

Finally, Figure 5.10 shows some examples of polyp images missed in RAPID experiments but correctly detected using the AI-Tool. Due to the size of the polyp and the fast movement of the capsule that took few images, these polyps are especially difficult to find using RAPID. In contrast, they are easy to find for AI-Tool users as they are presented with the clearest image.

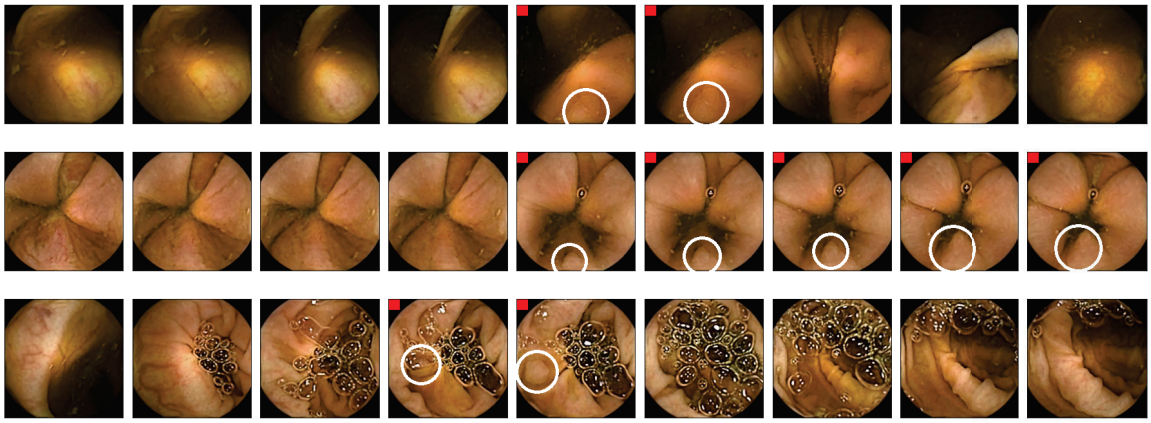


Figure 5.10: Example of polyps missed in RAPID experiments. Polyp frames are tagged with a red square. Polyps are circled in white.

5.4 Reordering sequences

The Super-Expert is a key feature for evaluating whether the system is improving. Although it was not previously discussed in detail, the Super-Expert represents the best performance achievable by the physician given the model's limitations. If a polyp is not shown to a physician because the model fails to predict certain images as polyps, it is impossible for the physician to identify them. Conversely, if there is a very clear and evident polyp that

appears in many frames, the model lacks the ability to prioritize more diverse images over redundant ones.

The solution to handle the first of these two problems is clear: to improve the latent model. The second one can be tackled by improving the way the frames are shown to the physicians, ensuring a higher diversity of frames. In Gilabert and Seguí (2022) a very simple, yet powerfull approach was presented to tackle this problem. The idea behind the method is to obtain a representation of the frames that can be used to find similar frames and discard them in favor of different ones. This process is done relatively to the prior order of the images that were sorted by probability. The following explanation contains a more formal definition of the problem and the solution proposed.

Let V be a CE video with n frames: $f_*^1, \ldots, f_*^n$ temporally ordered. Let M be a polyp detector model that assigns to each frame a score, $M(f_*^i)$. M induces a new ordering of the video, $\hat{f}_*^1, \ldots, \hat{f}_*^n$ where each frame has the same or less score of being a polyp than the previous one, i.e., $M(\hat{f}_*^i) \geq M(\hat{f}_*^{i+1})\ \forall i = 1, \ldots, n-1$.

Moreover, let $P_V = \{p_1, \ldots, p_k\}$ be the set of different polyps in video V. Each frame of the video has a P_V label if it contains a polyp, or a non-polyp label label, $p_\emptyset$. This label is indicated in the sub-index, e.g., $f_{p_\emptyset}^1$. Let $\hat{P_V} = P_V \cup \{p_\emptyset\}$ be the set of all possible labels of a frame.

The goal is to find a new ordering $\bar{f}_{q_1}^1, \ldots, \bar{f}_{q_n}^n,\ q_i \in \hat{P_V}\ \forall i = 1, \ldots, n$ such that the number of frames required to display all polyp labels is the minimum possible, i.e, we want to find an ordering such that the value m is minimal:

$$\bar{f}_{q_1}^1, \ldots, \bar{f}_{q_m}^m \ \Big|\ \bigcup_{i=1}^{m} q_i \supset P_V,\ m \leq n \tag{5.1}$$

To transform the initial ordering $\hat{f}_*^1, \ldots, \hat{f}_*^n$ to the new ordering $\bar{f}_*^1, \ldots, \bar{f}_*^n$ we use a similarity distance metric between frames computed using the image content. Let S be a model that extracts an embedding from each image, $S(f^i) = e_{f^i}$, we compute the similarity between two frames f^i and f^j as:

$$d_s(f^i, f^j) = ||S(f^i) - S(f^j)||_2 = ||e_{f^i} - e_{f^j}||_2 \tag{5.2}$$

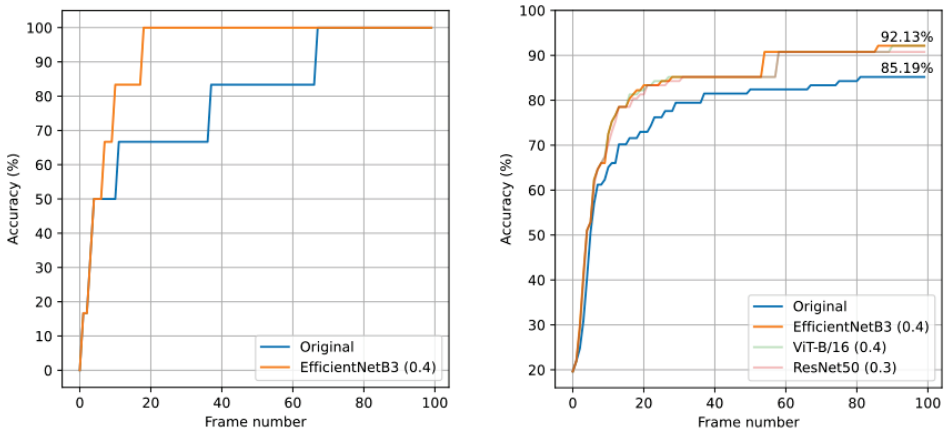
Then, to reorder the sequence of frames $\hat{f}_*^1, \ldots, \hat{f}_*^n$ we compute the similarity distance between each frame and all the next ones and we modify its score if it is below a threshold, μ_S, i.e.,

$$\begin{aligned} &\bar{f}_*^1, \ldots, \bar{f}_*^n \leftarrow \hat{f}_*^1, \ldots, \hat{f}_*^n \\ &\text{score}(\bar{f}_*^j) \leftarrow \text{score}(\bar{f}_*^j) d_s(\bar{f}_*^i, \bar{f}_*^j) \quad \forall i \geq 1 \quad \forall j > i \quad (\text{ if } \quad d_s(\bar{f}_*^i, \bar{f}_*^j) < \mu_S\ ) \end{aligned} \tag{5.3}$$

At each step i we reorder the sequence of frames $\bar{f}_*^{i+1}, \ldots, \bar{f}_*^n$ according to this new score. To avoid underflow problems we use logarithms in Equation 5.3.

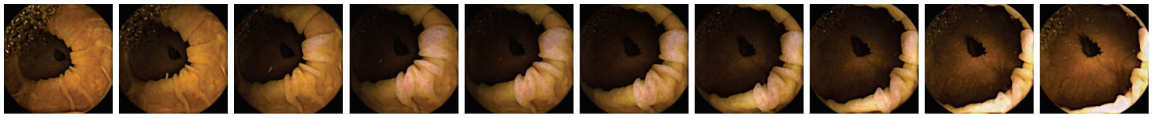
5.4.1 Results

Figure 5.11 shows the result of applying this process using three different similarity models, S, pretrained using ImageNet: ResNet50(0.3), EfficientNetB3(0.4), ViT-B/16(0.4). Inside the parenthesis we indicate the value of μ obtained after a gridsearch process. We used the polyp detector model, M, from Laiz et al. (2020). All of them present a similar behaviour since no process of fine-tuning was done, we simply used the default ImageNet weights. EfficientNetB3(0.4) achieved the highest score in both accuracy at frame 100 and area under the curve from Figure 5.11.

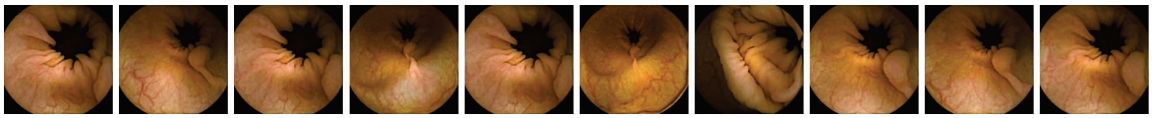


(a) Temporal Sorting (b) Probability Sorting

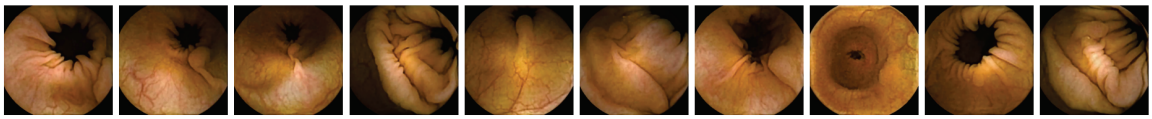
Figure 5.11: These plots compare the accuracy of the Super-Expert method when videos are sorted by probability alone (blue) versus after applying the reordering algorithm (orange). The left plot shows an example from a single video, while the right plot displays the average across the entire dataset. It can be observed that the orange line requires fewer frames to detect more polyps in the videos.



(a) Temporal Sorting



(b) Probability Sorting



(c) Probability + Similar discarding

Figure 5.12: Comparison of different sorting methods. The last row sorts the frames by probability and utilizes frame embeddings to discard similar frames, thereby avoiding redundancy.

5.5 The CESCAIL Study

The Capsule Endoscopy delivery at Scale through enhanced AI anaLysis (CESCAIL) study investigates the application of the AI-Tool developed in this work to enhance the accuracy and efficiency of diagnosing colonic polyps and colorectal cancer (Lei et al., 2023). Sponsored by University Hospitals Coventry and Warwickshire and funded by the National Institute for Health and Care Research (NIHR) AI in Health and Care Award, the study involves multiple institutions, including CorporateHealth International UK Limited (CHI), NHS Highland, NHS Arden & GEM Commissioning Support Unit (AGEM CSU) and the University of Barcelona. The CESCAIL study leverages AiSPEED™, whose development was led by CHI, to automate the video analysis of CCE and uses the AI-Tool developed in this work as a core system. This AI-enabled analysis is compared to conventional clinician analysis, aiming to validate its diagnostic accuracy and productivity.

The study design is a combined retrospective and prospective multicenter diagnostic study, with a targeted recruitment of 674 participants. These participants are either referred routinely for CCE or urgently for lower GI symptoms as part of their standard care pathway. The study includes both retrospective patients who have previously undergone CCE and prospective patients newly referred for the procedure. Demographic details, past medical history, and procedural data are collected and stored in a GCP-compliant electronic data capture system. The primary outcomes focus on comparing the sensitivity and specificity of the AI-enabled analysis with the gold standard of clinician analysis, ensuring a comprehensive assessment of AiSPEED™'s effectiveness.

The CESCAIL study's robust methodology involves detailed data collection and man-

agement practices to ensure the integrity and security of patient data. Pseudonymization techniques are applied to maintain confidentiality, with data securely stored and accessible only to authorized study team members. The study not only seeks to validate the AI tool but also aims to observe its performance in real-world settings, potentially revolutionizing lower GI diagnostics. By addressing the demand for more efficient diagnostic tools, CESCAIL has the potential to significantly impact clinical practices and patient outcomes in colorectal care.

Althought the study is currently in developement, preliminary results (Lei et al., 2024) demonstrate a comparable sensitivity and positive predicted value to the standard polyp review method while reducing the total revision time.

5.6 Conclusions

This chapter explored the advancements and efficacy of AI-based tools in enhancing polyp detection during CCE procedures, a critical aspect of CRC prevention. The study highlighted the evolution of detection methodologies, from early shape descriptors and geometric features to the more advanced deep learning techniques, particularly the application of CNNs and variants of the Transformer architecture. The implementation of innovative approaches like the ViT+SAM algorithm further optimized the performance of AI models, showcasing the potential for ongoing improvements in AI-assisted diagnostics. The findings highlight the importance of integrating AI-driven solutions in clinical practice to enhance patient outcomes.

The integration of AI models, such as the AI-Tool examined in this chapter, significantly improved the accuracy and efficiency of polyp detection. Specifically, the AI-Tool demonstrated a notable increase in sensitivity compared to traditional methods, achieving a sensitivity rate of 87.80% compared to 81.08% when using the RAPID Reader software alone. This enhancement is particularly crucial in CE, where accurate detection of polyps is vital given the procedure's non-invasive nature and the vast amount of video data generated.

Additionally, the successful application of our AI-Tool in the CESCAIL study underscores its practical value and effectiveness in a real-world clinical setting. The involvement of this tool in the CESCAIL study not only validates its efficacy but also positions it as a critical component in the future of CCE and CRC screening.

Chapter 6

Conclusions and Future Work

Contents

This chapter concludes the thesis, summarizing the key topics explored and the major contributions made to CE, especially in advancing AI integration in the video review pipeline.

It also highlights new challenges and areas for future research, such as exploring emerging technologies and interdisciplinary approaches, to further enhance the impact and development of CE.

6.1 Summary of Findings

Dataset Compilation

A key objective of this thesis was to understand, optimize, and improve the data-gathering pipeline. While data is crucial for any AI solution, its importance is often overshadowed by the focus on advanced model architectures.

To emphasize the significance of data, in Chapter 2, we conducted several studies. We developed an AI-based solution for automatically creating CE datasets using AL. AL is an AI paradigm where the model selects the most informative data points to label from a pool of unlabeled data. By doing so, the model is trained on the most diverse and representative samples, optimizing the use of the labeling budget. This approach is especially useful in cases where labeling is expensive or time-consuming, such as in medical imaging or video datasets, as it minimizes the amount of labeled data required to achieve high performance.

The method aimed to gather a diverse range of data to maximize the overall diversity of the dataset. We evaluated various methods from the literature and proposed new approaches, achieving the best results in our CE dataset. The most effective system employed a hybrid strategy: precomputing clusters of frame embeddings from all unlabeled data and utilizing these embeddings throughout the process without further modification as a representation of the frames. Videos that appeared in the most clusters were selected as the most diverse, labeled, and added to the training set. This process was repeated until the budget was exhausted, which, in our experiments, involved selecting a total of eight videos.

In this same chapter we also presented a collaboration with the Hospital Clínico de Valladolid. We released a dataset of 691 images of lung carcinomas from 45 patients and validated an AI solution to classify them, establishing a baseline for the research community.

Bowel Preparation Assessment

In CE, videos with significant intraluminal content, meaning a large amount of residue remaining in the intestines after bowel preparation, cannot be effectively analyzed for pathologies. Classifying videos as usable or unusable helps speed up the review process, as unusable videos can be flagged early and redone if necessary.

In Chapter 3, we present a method to automatically assign a CC-Clear score, a scale from 0 to 3 that evaluates the quality of bowel cleansing in the videos. Our method calculates the cleanliness of each video by first computing the cleanliness of each frame, segmenting the intraluminal content. Traditional approaches would require full image segmentation for model training, as seen in conventional U-Net (Ronneberger et al., 2015) and similar architectures, which typically use MSE or BCE as loss functions.

However, our system introduces a custom-designed loss function, Patch Loss, which

evaluates segmentation performance only in specific regions of the image, small patches, previously classified by expert physicians. This allows us to train the system using binary labels (clear/no clear) for these patches, significantly reducing the annotation workload for physicians.

Landmark Identification

Once videos are classified as usable, the next challenge is to accurately identify the organ of interest. These videos are typically long, and the transitions between organs are not always clearly defined. Even for physicians, pinpointing the exact frame where the transition occurs can be difficult.

While this might seem like a straightforward classification task, the problem is complicated by the fact that there is only one landmark (a positive example) among hundreds of frames showing the organ's interior (negative examples). This imbalance, combined with the importance of temporal information in video analysis, requires a more sophisticated approach.

In Chapter 4, we introduce a method for automatically finding the boundaries of large and small intestines, achieving higher precision in identifying the entrance and exit points than existing methods in the literature. Our approach utilizes a custom-designed architecture that incorporates both temporal information from the video sequence and frame similarity to track the motion of the capsule, classifying frames as either inside or outside the organ of interest. Additionally, we applied a rectangular pulse fitting to the method's output, allowing us to precisely identify the entrance and exit frames. This additional step was also tested on other methods from the literature, improving their results.

Polyp Detection

All the preparatory work with the videos culminates in the detection of GI pathologies. In Chapter 5, we focus on the early detection of CRC by identifying polyps. We first demonstrate the effectiveness of ViT and SAM in improving polyp classification, comparing the results with other state-of-the-art architectures.

Next, we present two clinical studies. The first compares the performance of the traditional review method using the RAPID app for CCE with our newly developed AI-Tool. The AI-Tool is a web-based application that incorporates a polyp detector and presents the frames to physicians by sorting them according to the output of the polyp detector model. This allows videos to be reviewed more quickly by focusing on the most relevant frames.

The experiments showed that the AI-powered method detects more polyps in less time. Notably, the highest improvement in accuracy was observed in detecting small polyps or polyps appearing in only a few frames, which are often missed by physicians during video

reviews. The second study, currently in progress, focuses on deploying one of these AI tools in a clinical setting, testing it on over 600 patients.

We also present a preliminary approach to avoid duplicate images during the review process. Experiments demonstrate that by using a simple method to identify similar frames proposed by the AI-Tool, the review process can be expedited even further.

6.2 Future Work

While the results presented in this thesis demonstrate significant advancements in AI-based solutions for improving the CE video review process, there are several opportunities for further exploration and enhancement.

Datasets

Access to public data is essential for advancing medical research, especially for teams that may not have access to large private datasets. Unfortunately, despite our efforts to release a large, well-annotated public dataset of CE videos, we were ultimately unsuccessful. We believe that significant effort is needed in this area, with a focus on making as much data publicly available as possible to foster collaboration and innovation in the field.

In addition, it is crucial to develop systems that can effectively handle data from various CE devices. Different manufacturers often produce devices with distinct specifications, such as resolution, frame rate, and video format. To maximize the utilization of available data, AI systems must be designed to standardize and integrate these diverse data sources. By creating models that are device-agnostic and adaptable to a variety of data formats, researchers can tap into a much larger pool of data. This would not only enhance the training and validation of AI models but also improve their robustness and generalizability, making them more suitable for widespread clinical adoption across different institutions and regions.

A promising approach to addressing the challenges of integrating data from various CE devices is federated learning. Federated learning allows AI models to be trained across decentralized datasets from different devices and institutions without requiring the transfer of sensitive data to a central location. This method is particularly advantageous in healthcare, where data privacy and regulatory constraints often limit the sharing of patient information. By training models locally on each institution's data, federated learning enables the development of robust, device-agnostic AI systems. These systems can generalize across different device specifications, improving diagnostic accuracy and facilitating widespread clinical adoption without compromising patient privacy.

Improving Polyp Detection

One of the primary limitations identified during this thesis is the scarcity of annotated data for polyps. However, the potential for improvement is significant if more finely-grained data were available. A key area that requires further attention is the characterization of the morphology and size of polyps. Current tools lack the precision needed to measure polyp size accurately, as the size can appear to fluctuate based on the camera's field of view. Developing a robust, well-annotated dataset of polyps would enable a more reliable training for AI systems for size and morphology estimation.

Another challenge arises from the movement of the capsule within the digestive tract. As the capsule moves back and forth due to the natural contractions of the intestines, it constantly changes its field of view of the intestines. This makes the re-identification of polyps particularly difficult, as the same polyp might appear different in subsequent frames. Reliable labeled datasets that account for these shifts in perspective are critical to address this challenge. Training models to recognize polyps despite these changes would enhance the reliability of polyp detection and tracking over time.

Additionally, a promising direction for future research lies in patient-oriented screening programs. CE is often used to monitor patients over time, and as such, these programs typically involve multiple endoscopy sessions. With a comprehensive dataset of a patient's endoscopic images collected across time, it would be possible to develop personalized systems capable of tracking the growth and evolution of polyps. This would significantly enhance the prediction of CRC, as AI models could be trained to detect subtle changes in polyp size or morphology across sessions, offering more precise and early predictions of cancer development.

Finally, this thesis focused on solutions primarily based on individual images, as that was the data available during the research. However, with the adoption of more advanced architectures, such as the Transformer and other sequence-based models, future research could leverage full sequences of images or even complete video data for training. This shift would allow for more contextual understanding, leading to improvements in polyp detection and overall diagnostic performance. The exploration of sequence-based data could unlock new possibilities in CE, pushing the boundaries of what AI models can achieve in GI diagnostics.

System Integration and Real-World Application

Throughout this thesis, we developed a comprehensive suite of applications designed to enhance the CE video review process. While each of these modules has proven effective individually, we believe that integrating them into a unified system could significantly streamline the overall workflow and improve the user experience. A fully integrated platform would allow for smoother transitions between different stages of the review process, from data

preprocessing and landmark identification to polyp detection and pathology classification.

Most of these systems were tested on controlled datasets, which may not fully represent the variability and challenges encountered in real-world clinical settings. To ensure the robustness and generalizability of the proposed solutions, larger-scale trials, such as those conducted in the CESCAIL project (Lei et al., 2023), should be performed before moving these systems into production. These trials would help assess performance across diverse patient populations and real-world conditions, addressing potential limitations and ensuring that the integrated system can be effectively deployed in clinical environments.

Much of the research is centered around comparing objective metrics to existing methods in the literature. While this approach is essential, we believe it is equally important to ensure that these systems are understandable to physicians. Developing interpretability modules that explain the decisions made by AI models, along with the associated confidence levels, is crucial. Physicians must recognize that these systems are not intended to replace their expertise but rather to enhance and accelerate their work. By providing a more consistent and repeatable method of evaluation, AI tools can help improve diagnostic accuracy and efficiency, ultimately benefiting patient care.

6.3 Research Outcome

Journal Publications

- **P. Gilabert**, J. Vitrià, P. Laiz, C. Malagelada, A. Watson, H. Wenzek, and S. Seguí. Artificial intelligence to improve polyp detection and screening time in colon capsule endoscopy. *Frontiers in Medicine*, 9, 1 2022. ISSN 2296-858X. doi: 10.3389/fmed.2022.1000726.

- P. Laiz, J. Vitrià, **P. Gilabert**, H. Wenzek, C. Malagelada, A. Watson, and S. Seguí. Anatomical landmarks localization for capsule endoscopy studies. *Computerized Medical Imaging and Graphics*, 108, 2023. ISSN 0895-6111. doi: 10.1016/j.compmedimag.2023.102243.

- PG. Duran, **P. Gilabert**, S. Seguí and J. Vitrià. Overcoming Diverse Undesired Effects in Recommender Systems: A Deontological Approach. *ACM Transactions on Intelligent Systems and Technology*, 2 2024. ISSN 2157-6904. doi: 10.1145/3643857.

- J. Diosdado, **P. Gilabert**, S. Seguí and H. Borrego. LungHist700: A Dataset of Histological Images for Deep Learning in Pulmonary Pathology. ***Submitted***.

- **P. Gilabert**, C. Malagelada, H. Wenzek, Á. Finta, A. Watson, J. Vitrià and S. Seguí. AI-Assisted Evaluation of Colon Cleanliness in Capsule Endoscopy Videos. ***Submitted***.

International Conferences

- **P. Gilabert** and S. Seguí. Gradient boosting and language model ensemble for tweet recommendation. *2020 Proceedings of the Recommender Systems Challenge (RecSys)*, pages 24-28, 9 2020. ISBN 978-1-45038-835-1. doi: 10.1145/3415959.3415997.

- **P. Gilabert** and S. Seguí. Addressing the cold-start problem with a two-branch architecture for fair tweet recommendation. *2021 Proceedings of the Recommender Systems Challenge (RecSys)*, pages 34-38, 11 2021. ISBN 978-145038-693-7. doi: 10.1145/3487572.3487598.

- **P. Gilabert** and S. Seguí. Improving CCE video review time with a model based on frame similarity. *2022 Medical Imaging with Deep Learning (MIDL)*.

- **P. Gilabert**, C. Malagelada, H. Wenzek, J. Vitrià and S. Seguí. Leveraging Embedding Information to Create Video Capsule Endoscopy Datasets. *2023 18th International Conference on Machine Vision and Applications (MVA)*, pages 1-5, 8 2023. ISBN 978-4-88552-343-4. doi: 10.23919/MVA57639.2023.10215919.

- **P. Gilabert**, C. Malagelada, H. Wenzek, J. Vitrià and S. Seguí. Automated Cleanliness Scoring and Digestive Content Segmentation for Capsule Endoscopy. *Frontiers in Artificial Intelligence and Applications 375, Proceedings of the 25th International Conference of the Catalan Association for Artificial Intelligence (CCIA)*, pages 134-135, 10 2023. ISBN 978-1-64368-449-9. doi: 10.3233/FAIA230673.

Datasets

- J. Diosdado; **P. Gilabert**, S. Seguí and H. Borrego. LungHist700: A Dataset of Histological Images for Deep Learning in Pulmonary Pathology. *figshare*. Dataset. doi: 10.6084/m9.figshare.25459174.

Book Chapters

- **P. Gilabert**, P. Laiz and S. Seguí. Artificial Intelligence for Vascular Lesions. In: M. Mascarenhas, H. Cardoso and G. Macedo, (eds) Artificial Intelligence in Capsule Endoscopy. 2023. *Springer*. ISBN: 9780323996471. doi: 10.1016/B978-0-323-99647-1.00012-5

Supervised Undergraduate and Master Projects

- **S. Bardají**. Learning contextual information via Deep Learning. Final Degree Project, 2021. Supervisors: S. Seguí and P. Gilabert.

- **A. Torralba and M. Moreno**. Twitter engagement model for RecSys Challenge. Final Masters Project, 2021. Supervisors: S. Seguí and P. Gilabert.
- **M. Tabarner**. Recommender Systems for 2022 Recsys Competition. Final Masters Project, 2022. Supervisors: S. Seguí and P. Gilabert.
- **S. Bardají**. Active Learning strategies for WCE images classification. Final Masters Project, 2023. Supervisors: S. Seguí and P. Gilabert.

Bibliography

M. Abdulkareem and S. E. Petersen. The Promise of AI in Detection, Diagnosis, and Epidemiology for Combating COVID-19: Beyond the Hype. *Frontiers in Artificial Intelligence*, 4, 5 2021. ISSN 2624-8212. doi: 10.3389/frai.2021.652669.

S. Adewole, M. Yeghyayan, D. Hyatt, L. Ehsan, J. Jablonski, A. Copland, S. Syed, and D. Brown. Deep Learning Methods for Anatomical Landmark Detection in Video Capsule Endoscopy Images. In *Proceedings of the Future Technologies Conference (FTC) 2020, Volume 1*, pages 426–434. Springer, 2020. doi: 10.1007/978-3-030-63128-4_32.

M. Ahmed. Colon Cancer: A Clinician's Perspective in 2019. *Gastroenterology Research*, 13(1):1–10, 2020. ISSN 1918-2805. doi: 10.14740/gr1239.

L. A. Al-Aswad, C. Y. Elgin, V. Patel, D. Popplewell, K. Gopal, D. Gong, Z. Thomas, D. Joiner, C.-K. Chu, S. Walters, M. Ramachandran, R. Kapoor, M. Rodriguez, J. Alcantara-Castillo, G. E. Maestre, J. H. Lee, and G. Moazami. Real-Time Mobile Teleophthalmology for the Detection of Eye Disease in Minorities and Low Socioeconomics At-Risk Populations. *Asia-Pacific Journal of Ophthalmology*, 10(5):461–472, 9 2021. ISSN 21620989. doi: 10.1097/APO.0000000000000416.

M. Alageeli, B. Yan, S. Alshankiti, M. Al-Zahrani, Z. Bahreini, T. T. Dang, J. Friedland, S. Gilani, R. Homenauth, J. Houle, M. Kloc, J. Luhoway, L. Merotto, R. Rofaiel, C. Singh, A. Smith, B. Thomas, C. Townsend, D. Yoo, S. Zepeda-Gomez, L. Stitt, V. Jairath, and M. S. L. Sey. KODA score: an updated and validated bowel preparation scale for patients undergoing small bowel capsule endoscopy. *Endoscopy International Open*, 08 (08):E1011–E1017, 8 2020. ISSN 2364-3722. doi: 10.1055/a-1176-9889.

J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer Normalization, 7 2016.

C. S. Bang, J. J. Lee, and G. H. Baik. Computer-Aided Diagnosis of Gastrointestinal Ulcer and Hemorrhage Using Wireless Capsule Endoscopy: Systematic Review and Diagnostic Test Accuracy Meta-analysis. *Journal of Medical Internet Research*, 23(12):e33267, 12 2021. ISSN 1438-8871. doi: 10.2196/33267.

S. Bardají, S. Seguí, and P. Gilabert. *Active Learning strategies for WCE images classification*. Masters thesis, Universitat de Barcelona, 2024. URL http://hdl.handle.net/2445/212901.

A. Becq, A. Histace, M. Camus, I. Nion-Larmurier, E. Abou Ali, O. Pietri, O. Romain, U. Chaput, C. Li, P. Marteau, C. Florent, and X. Dray. Development of a computed cleansing score to assess quality of bowel preparation in colon capsule endoscopy. *Endoscopy International Open*, 06(07):E844–E850, 7 2018. ISSN 2364-3722. doi: 10.1055/a-0577-2897.

K. Bera, K. A. Schalper, D. L. Rimm, V. Velcheti, and A. Madabhushi. Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology. *Nature Reviews Clinical Oncology*, 16(11):703–715, 11 2019. ISSN 1759-4774. doi: 10.1038/s41571-019-0252-y.

K. Bera, N. Braman, A. Gupta, V. Velcheti, and A. Madabhushi. Predicting cancer outcomes with radiomics and artificial intelligence in radiology. *Nature Reviews Clinical Oncology*, 19(2):132–146, 2 2022. ISSN 1759-4774. doi: 10.1038/s41571-021-00560-7.

J. Berens, M. Mackiewicz, and D. Bell. Stomach, intestine, and colon tissue discriminators for wireless capsule endoscopy images. In J. M. Fitzpatrick and J. M. Reinhardt, editors, *Proceedings of SPIE - The International Society for Optical Engineering*, page 283, 4 2005. doi: 10.1117/12.594799.

C. L. Berre, C. Trang-Poisson, and A. Bourreille. Small bowel capsule endoscopy and treat-to-target in Crohn's disease: A systematic review. *World Journal of Gastroenterology*, 25 (31):4534–4554, 8 2019. ISSN 1007-9327. doi: 10.3748/wjg.v25.i31.4534.

T. Bjoersum-Meyer, K. Skonieczna-Zydecka, P. Cortegoso Valdivia, I. Stenfors, I. Lyutakov, E. Rondonotti, M. Pennazio, W. Marlicz, G. Baatrup, A. Koulaouzidis, and E. Toth. Efficacy of bowel preparation regimens for colon capsule endoscopy: a systematic review and meta-analysis. *Endoscopy International Open*, 09(11):E1658–E1673, 11 2021. ISSN 2364-3722. doi: 10.1055/a-1529-5814.

J. H. Bond. Colon Polyps and Cancer. *Endoscopy*, 35(1):27–35, 1 2003. ISSN 0013-726X. doi: 10.1055/s-2003-36410.

F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74(3): 229–263, 5 2024. ISSN 0007-9235. doi: 10.3322/caac.21834.

M. M. Buijs, M. H. Ramezani, J. Herp, R. Kroijer, M. Kobaek-Larsen, G. Baatrup, and E. S. Nadimi. Assessment of bowel cleansing quality in colon capsule endoscopy using machine learning: a pilot study. *Endoscopy International Open*, 06(08):E1044–E1050, 8 2018. ISSN 2364-3722. doi: 10.1055/a-0627-7136.

A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 11(2):125, 2 2020. ISSN 20782489. doi: 10.3390/info11020125.

K. J. Butnor, M. B. Beasley, P. T. Cagle, S. M. Grunberg, F.-M. Kong, A. Marchevsky, N. T. Okby, V. L. Roggli, S. Suster, H. D. Tazelaar, W. D. Travis, and A. Arbor. Protocol for the Examination of Specimens From Patients With Primary Non–Small Cell Carcinoma, Small Cell Carcinoma, or Carcinoid Tumor of the Lung. *Arch Pathol Lab Med*, 133(10): 1552–1559, 10 2009. doi: 10.5858/133.10.1552.

P. Carcagnì, M. Leo, L. Signore, and C. Distante. Medical Transformers for Boosting Automatic Grading of Colon Carcinoma in Histological Images. In *Image Analysis and Processing – ICIAP*, pages 135–146. Springer, 2023. doi: 10.1007/978-3-031-43148-7_12.

C.-L. Chen, C.-C. Chen, W.-H. Yu, S.-H. Chen, Y.-C. Chang, T.-I. Hsu, M. Hsiao, C.-Y. Yeh, and C.-Y. Chen. An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning. *Nature Communications*, 12(1): 1193, 2 2021a. ISSN 2041-1723. doi: 10.1038/s41467-021-21467-y.

H. Chen, X. Wu, G. Tao, and Q. Peng. Automatic content understanding with cascaded spatial–temporal deep framework for capsule endoscopy videos. *Neurocomputing*, 229: 77–87, 3 2017. ISSN 09252312. doi: 10.1016/j.neucom.2016.06.077.

J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation, 2 2021b.

P. Coelho, A. Pereira, A. Leite, M. Salgado, and A. Cunha. A Deep Learning Approach for Red Lesions Detection in Video Capsule Endoscopies. *Lecture Notes in Computer Science*, 10882 LNCS:553–561, 2018. ISSN 16113349. doi: 10.1007/978-3-319-93000-8_63.

M. Cohen-Shelly, Z. I. Attia, P. A. Friedman, S. Ito, B. A. Essayagh, W.-Y. Ko, D. H. Murphree, H. I. Michelena, M. Enriquez-Sarano, R. E. Carter, P. W. Johnson, P. A. Noseworthy, F. Lopez-Jimenez, and J. K. Oh. Electrocardiogram screening for aortic valve stenosis using artificial intelligence. *European Heart Journal*, 42(30):2885–2896, 8 2021. ISSN 0195-668X. doi: 10.1093/eurheartj/ehab153.

P. Cortegoso Valdivia, U. Deding, T. Bjørsum-Meyer, G. Baatrup, I. Fernández-Urién, X. Dray, P. Boal-Carvalho, P. Ellul, E. Toth, E. Rondonotti, L. Kaalby, M. Pennazio, and A. Koulaouzidis. Inter/Intra-Observer Agreement in Video-Capsule Endoscopy: Are We Getting It All Wrong? A Systematic Review and Meta-Analysis. *Diagnostics*, 12(10): 2400, 10 2022. doi: 10.3390/diagnostics12102400.

A. de Maissin, R. Vallée, M. Flamant, M. Fondain-Bossiere, C. L. Berre, A. Coutrot, N. Normand, H. Mouchère, S. Coudol, C. Trang, and A. Bourreille. Multi-expert annotation of Crohn's disease images of the small bowel for automatic detection using a convolutional recurrent attention neural network. *Endoscopy International Open*, 09(07):E1136–E1144, 7 2021. ISSN 2364-3722. doi: 10.1055/a-1468-3964.

R. de Sousa Magalhães, C. Arieira, P. Boal Carvalho, B. Rosa, M. J. Moreira, and J. Cotter. Colon Capsule CLEansing Assessment and Report (CC-CLEAR): a new approach for evaluation of the quality of bowel preparation in capsule colonoscopy. *Gastrointestinal Endoscopy*, 93(1):212–223, 1 2021. ISSN 10976779. doi: 10.1016/j.gie.2020.05.062.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018.

J. Diosdado, P. Gilabert, S. Seguí, and H. Borrego. LungHist700: A dataset of histological images for deep learning in pulmonary pathology. *Scientific Data*, 2024a.

J. Diosdado, P. Gilabert, S. Seguí, and H. Borrego. LungHist700: A dataset of histological images for deep learning in pulmonary pathology. *figshare*, 2024b. doi: 10.6084/m9.figshare.25459174.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*, 10 2021. doi: 10.48550/arXiv.2010.11929.

Endoscopic Vision Challenge. Endoscopic Vision Challenge, 2017. URL https://opencas.dkfz.de/endovis/.

I. N. Figueiredo, S. Prasath, Y.-H. R. Tsai, and P. N. Figueiredo. Automatic detection and segmentation of colonic polyps in wireless capsule images. Technical report, The Institute for Computational Engineering and Sciences. The University of Texas at Austin, 9 2010.

P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur. Sharpness-Aware Minimization for Efficiently Improving Generalization, 10 2020.

J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin. Convolutional Sequence to Sequence Learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, pages 1243–1252, Sydney, NSW, Australia, 5 2017. JMLR.org. doi: 10.48550/arXiv.1705.03122.

P. Gilabert. Vision Transformer implementation, 6 2024. URL https://github.com/perecasxiru/VisionTransformer_minimal.

P. Gilabert and S. Seguí. Improving CCE video review time with a model based on frame similarity. In *Medical Imaging with Deep Learning*, 2022. URL https://openreview.net/forum?id=3lLvO8-a3EE.

P. Gilabert, C. Malagelada, H. Wenzek, J. Vitrià, and S. Seguí. Leveraging Embedding Information to Create Video Capsule Endoscopy Datasets. In *2023 18th International Conference on Machine Vision and Applications (MVA)*, pages 1–5. Universitat de Barcelona, IEEE, 7 2023. ISBN 978-4-88552-343-4. doi: 10.23919/MVA57639.2023.10215919.

X. Guo and Y. Yuan. Triple ANet: Adaptive Abnormal-aware Attention Network for WCE Image Classification. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 293–301, 2019. doi: 10.1007/978-3-030-32239-7_33.

L. Gutierrez, J. S. Lim, L. L. Foo, W. Y. Ng, M. Yip, G. Y. S. Lim, M. H. Y. Wong, A. Fong, M. Rosman, J. S. Mehta, H. Lin, D. S. J. Ting, and D. S. W. Ting. Application of artificial intelligence in cataract management: current and future directions. *Eye and Vision*, 9(1):3, 12 2022. ISSN 2326-0254. doi: 10.1186/s40662-021-00273-z.

O. Haji-Maghsoudi, A. Talebpour, H. Soltanian-Zadeh, and N. Haji-maghsoodi. Automatic organs' detection in WCE. In *The 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012)*, pages 116–121. IEEE, 5 2012. ISBN 978-1-4673-1479-4. doi: 10.1109/AISP.2012.6313729.

M. G. Hanna, V. E. Reuter, M. R. Hameed, L. K. Tan, S. Chiang, C. Sigel, T. Hollmann, D. Giri, J. Samboy, C. Moradel, A. Rosado, J. R. Otilano, C. England, L. Corsale, E. Stamelos, Y. Yagi, P. J. Schüffler, T. Fuchs, D. S. Klimstra, and S. Sirintrapun. Whole slide imaging equivalency and efficiency study: experience at a large academic center. *Modern Pathology*, 32(7):916–928, 7 2019. ISSN 08933952. doi: 10.1038/s41379-019-0205-0.

M. Hanscom and D. R. Cave. Endoscopic capsule robot-based diagnosis, navigation and localization in the gastrointestinal tract. *Frontiers in Robotics and AI*, 9, 9 2022. ISSN 2296-9144. doi: 10.3389/frobt.2022.896028.

S. Hashem, G. Esmat, W. Elakel, S. Habashy, S. A. Raouf, M. Elhefnawi, M. I. Eladawy, and M. ElHefnawi. Comparison of Machine Learning Approaches for Prediction of Advanced Liver Fibrosis in Chronic Hepatitis C Patients. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(3):861–868, 5 2018. ISSN 1545-5963. doi: 10.1109/TCBB.2017.2690848.

C. Hassan, M. Bretthauer, M. Kaminski, M. Polkowski, B. Rembacken, B. Saunders, R. Benamouzig, O. Holme, S. Green, T. Kuiper, R. Marmo, M. Omar, L. Petruzziello, C. Spada, A. Zullo, and J. Dumonceau. Bowel preparation for colonoscopy: European Society of Gastrointestinal Endoscopy (ESGE) Guideline. *Endoscopy*, 45(02):142–155, 1 2013. ISSN 0013-726X. doi: 10.1055/s-0032-1326186.

C. Hassan, M. Spadaccini, A. Iannone, R. Maselli, M. Jovani, V. T. Chandrasekar, G. Antonelli, H. Yu, M. Areia, M. Dinis-Ribeiro, P. Bhandari, P. Sharma, D. K. Rex, T. Rösch, M. Wallace, and A. Repici. Performance of artificial intelligence in colonoscopy for adenoma and polyp detection: a systematic review and meta-analysis. *Gastrointestinal Endoscopy*, 93(1):77–85, 1 2021. ISSN 00165107. doi: 10.1016/j.gie.2020.06.059.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer New York, New York, NY, 2009. ISBN 978-0-387-84857-0. doi: 10.1007/978-0-387-84858-7.

K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, 6 2016. IEEE. doi: 10.1109/CVPR.2016.90.

M. S. Hosseini, B. E. Bejnordi, V. Q.-H. Trinh, L. Chan, D. Hasan, X. Li, S. Yang, T. Kim, H. Zhang, T. Wu, K. Chinniah, S. Maghsoudlou, R. Zhang, J. Zhu, S. Khaki, A. Buin, F. Chaji, A. Salehi, B. N. Nguyen, D. Samaras, and K. N. Plataniotis. Computational

Pathology: A Survey Review and The Way Forward. *Journal of Pathology Informatics*, page 100357, 1 2024. ISSN 21533539. doi: 10.1016/j.jpi.2023.100357.

S. Hwang and M. E. Celebi. Polyp detection in wireless capsule endoscopy videos based on image segmentation and geometric feature. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. I E E E, 2010. ISBN 9781424442966. doi: 10.1109/ICASSP.2010.5495103.

G. Iddan, G. Meron, A. Glukhovsky, and P. Swain. Wireless capsule endoscopy. *Nature*, 405(417), 2000. doi: 10.1038/35013140.

M. S. Ismail, G. Murphy, S. Semenov, and D. McNamara. Comparing Colon Capsule Endoscopy to colonoscopy; a symptomatic patient's perspective. *BMC Gastroenterology*, 22(1):31, 12 2022. ISSN 1471-230X. doi: 10.1186/s12876-021-02081-0.

D. Jha, N. K. Tomar, S. Ali, M. A. Riegler, H. D. Johansen, D. Johansen, T. de Lange, and P. Halvorsen. NanoNet: Real-Time Polyp Segmentation in Video Capsule Endoscopy and Colonoscopy. In *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 37–43, 4 2021. ISBN 978-1-6654-4121-6. doi: 10.1109/CBMS52027.2021.00014.

X. Jiang, J. Pan, Q. Xu, Y.-H. Song, H.-H. Sun, C. Peng, X.-L. Qi, Y.-Y. Qian, W.-B. Zou, Y. Yang, S.-Q. Jin, B.-S. Duan, S. Wu, Y. Chu, D.-H. Xiao, L.-J. Hu, J.-Z. Cao, J.-F. Dai, X. Liu, T. Xia, W. Zhou, T. Chen, C.-H. Zhou, W. Wu, S.-J. Liu, Z.-Y. Yang, F. Wang, L. Zhang, C.-Z. Li, H. Xu, J.-X. Wang, B. Wei, Y. Lin, X. Deng, L.-H. Qu, Y.-Q. Shen, H. Wang, Y.-F. Huang, H.-B. Bao, S. Zhang, L. Li, Y.-H. Shi, X.-Y. Wang, D.-W. Zou, X.-J. Wan, M.-D. Xu, H. Mao, C.-H. He, Z. Li, X.-L. Zuo, S.-X. He, X.-P. Xie, J. Liu, C.-Q. Yang, C. Spada, Z.-S. Li, and Z. Liao. Diagnostic accuracy of magnetically guided capsule endoscopy with a detachable string for detecting oesophagogastric varices in adults with cirrhosis: prospective multicentre study. *BMJ*, page e078581, 3 2024. ISSN 1756-1833. doi: 10.1136/bmj-2023-078581.

G. G. Johnson, R. Helewa, D. C. Moffatt, J. G. Coneys, J. Park, and E. Hyun. Colorectal polyp classification and management of complex polyps for surgeon endoscopists. *Canadian Journal of Surgery*, 66(5):E491–E498, 9 2023. ISSN 14882310. doi: 10.1503/cjs.011422.

J. W. Ju, H. Jung, Y. J. Lee, S. W. Mun, and J. H. Lee. Semantic Segmentation Dataset for AI-Based Quantification of Clean Mucosa in Capsule Endoscopy. *Medicina (Lithuania)*, 58(3), 3 2022. ISSN 16489144. doi: 10.3390/medicina58030397.

H. Khalid, M. Hussain, M. A. Al Ghamdi, T. Khalid, K. Khalid, M. A. Khan, K. Fatima, K. Masood, S. H. Almotiri, M. S. Farooq, and A. Ahmed. A Comparative Systematic Literature Review on Knee Bone Reports from MRI, X-Rays and CT Scans Using Deep Learning and Machine Learning Methodologies. *Diagnostics*, 10(8):518, 7 2020. ISSN 2075-4418. doi: 10.3390/diagnostics10080518.

J. J. Koornstra. Bowel preparation before small bowel capsule endoscopy: What is the optimal approach?, 2009. ISSN 0954691X.

A. Koulaouzidis, D. Iakovidis, D. Yung, E. Rondonotti, U. Kopylov, J. Plevris, E. Toth, A. Eliakim, G. Wurm Johansson, W. Marlicz, G. Mavrogenis, A. Nemeth, H. Thorlacius, and G. Tontini. KID Project: an internet-based digital video atlas of capsule endoscopy for research purposes. *Endoscopy International Open*, 05(06):E477–E483, 6 2017. ISSN 2364-3722. doi: 10.1055/s-0043-105488.

A. Koulaouzidis, K. Dabos, M. Philipper, E. Toth, and M. Keuchel. How should we do colon capsule endoscopy reading: a practical guide, 2021. ISSN 26317745.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS'12: Proceedings of the 25th International Conference on Neural Information Processing Systems*, New York, 12 2012. Curran Associates Inc. ISBN 9781627480031. doi: 10.1145/3065386.

N. Kumar, R. Gupta, and S. Gupta. Whole Slide Imaging (WSI) in Pathology: Current Perspectives and Future Directions. *Journal of Digital Imaging*, 33(4):1034–1040, 8 2020. ISSN 0897-1889. doi: 10.1007/s10278-020-00351-z.

P. Laiz, J. Vitrià, H. Wenzek, C. Malagelada, F. Azpiroz, and S. Seguí. WCE polyp detection with triplet based embeddings. *Computerized Medical Imaging and Graphics*, 86:101794, 12 2020. ISSN 08956111. doi: 10.1016/j.compmedimag.2020.101794.

V. Lakshminarayanan, H. Kheradfallah, A. Sarkar, and J. Jothi Balaji. Automated Detection and Diagnosis of Diabetic Retinopathy: A Comprehensive Survey. *Journal of Imaging*, 7(9):165, 8 2021. ISSN 2313-433X. doi: 10.3390/jimaging7090165.

A. Lasker, S. M. Obaidullah, C. Chakraborty, and K. Roy. Application of Machine Learning and Deep Learning Techniques for COVID-19 Screening Using Radiological Imaging: A Comprehensive Review. *SN Computer Science*, 4(1):65, 11 2022. ISSN 2661-8907. doi: 10.1007/s42979-022-01464-8.

B. Lauby-Secretan, C. Scoccianti, D. Loomis, Y. Grosse, F. Bianchini, and K. Straif. Body Fatness and Cancer — Viewpoint of the IARC Working Group. *New England Journal of Medicine*, 375(8):794–798, 8 2016. ISSN 0028-4793. doi: 10.1056/NEJMsr1606602.

Y. Lecun, E. Bottou, Y. Bengio, and P. Haffner. Gradient-Based Learning Applied to Document Recognition. In *Proceedings of the IEEE*, pages 2278–2324. IEEE, 11 1998. doi: 10.1109/5.726791.

J. Lee, J. Oh, S. K. Shah, X. Yuan, and S. J. Tang. Automatic classification of digestive organs in wireless capsule endoscopy videos. In *Proceedings of the 2007 ACM symposium on Applied computing*, pages 1041–1045, New York, NY, USA, 3 2007. ACM. ISBN 1595934804. doi: 10.1145/1244002.1244230.

J. Y. Lee, A. H. Calderwood, W. Karnes, J. Requa, B. C. Jacobson, and M. B. Wallace. Artificial intelligence for the assessment of bowel preparation. *Gastrointestinal Endoscopy*, 95(3):512–518, 3 2022. ISSN 10976779. doi: 10.1016/j.gie.2021.11.041.

R. Leenhardt, C. Li, J.-P. Le Mouel, G. Rahmi, J. C. Saurin, F. Cholet, A. Boureille, X. Amiot, M. Delvaux, C. Duburque, C. Leandri, R. Gérard, S. Lecleire, F. Mesli, I. Nion-Larmurier, O. Romain, S. Sacher-Huvelin, C. Simon-Shane, G. Vanbiervliet, P. Marteau, A. Histace, and X. Dray. CAD-CAP: a 25,000-image database serving the development of artificial intelligence for capsule endoscopy. *Endoscopy International Open*, 08(03): E415–E420, 3 2020. ISSN 2364-3722. doi: 10.1055/a-1035-9088.

R. Leenhardt, M. Souchaud, G. Houist, J. P. Le Mouel, J. C. Saurin, F. Cholet, G. Rahmi, C. Leandri, A. Histace, and X. Dray. A neural network-based algorithm for assessing the cleanliness of small bowel during capsule endoscopy. *Endoscopy*, 53(9):932–936, 9 2021. ISSN 14388812. doi: 10.1055/a-1301-3841.

I. I. Lei, K. Tompkins, E. White, A. Watson, N. Parsons, A. Noufaily, S. Segui, H. Wenzek, R. Badreldin, A. Conlin, and R. P. Arasaradnam. Study of capsule endoscopy delivery at scale through enhanced artificial intelligence-enabled analysis (the CESCAIL study). *Colorectal Disease*, 25(7):1498–1505, 7 2023. ISSN 1462-8910. doi: 10.1111/codi.16575.

I. I. Lei, N. Parsons, A. Koulaouzidis, R. Alexander, H. Wenzek, P. Laiz, A. Watson, E. White, and R. Arasaradnam. Accelerating cutting edge breakthroughs: capsule endoscopy delivery at scale through enhanced AI analysis (CESCAIL) study – A comprehensive interim analysis. In *Gut*, pages A151–A151. BMJ Publishing Group Ltd and British Society of Gastroenterology, 6 2024. doi: 10.1136/gutjnl-2024-BSG.247.

L. Lenchik, L. Heacock, A. A. Weaver, R. D. Boutin, T. S. Cook, J. Itri, C. G. Filippi, R. P. Gullapalli, J. Lee, M. Zagurovskaya, T. Retson, K. Godwin, J. Nicholson, and P. A. Narayana. Automated Segmentation of Tissues Using CT and MRI: A Systematic Review. *Academic Radiology*, 26(12):1695–1706, 12 2019. ISSN 10766332. doi: 10.1016/j.acra.2019.07.006.

T. R. Levin, D. A. Corley, C. D. Jensen, J. E. Schottinger, V. P. Quinn, A. G. Zauber, J. K. Lee, W. K. Zhao, N. Udaltsova, N. R. Ghai, A. T. Lee, C. P. Quesenberry, B. H. Fireman, and C. A. Doubeni. Effects of Organized Colorectal Cancer Screening on Cancer Incidence and Mortality in a Large Community-Based Population. *Gastroenterology*, 155 (5):1383–1391, 11 2018. ISSN 00165085. doi: 10.1053/j.gastro.2018.07.017.

B. Li and M. Q. Meng. Automatic polyp detection for wireless capsule endoscopy images. *Expert Systems with Applications*, 39(12):10952–10958, 9 2012. ISSN 09574174. doi: 10.1016/j.eswa.2012.03.029.

B. Li, M. Q.-H. Meng, and L. Xu. A Comparative Study of Shape Features for Polyp Detection in Wireless Capsule Endoscopy Images. In *2009 Annual International Conference*

of the IEEE Engineering in Medicine and Biology Society, 2009. ISBN 9781424432967. doi: 10.1109/IEMBS.2009.5334875.

B. Li, G. Xu, R. Zhou, and T. Wang. Computer aided wireless capsule endoscopy video segmentation. *Medical Physics*, 42(2):645–652, 2 2015. ISSN 0094-2405. doi: 10.1118/1.4905164.

F. Lopez-Jimenez, Z. Attia, A. M. Arruda-Olson, R. Carter, P. Chareonthaitawee, H. Jouni, S. Kapa, A. Lerman, C. Luong, J. R. Medina-Inojosa, P. A. Noseworthy, P. A. Pellikka, M. M. Redfield, V. L. Roger, G. S. Sandhu, C. Senecal, and P. A. Friedman. Artificial Intelligence in Cardiology: Present and Future. *Mayo Clinic Proceedings*, 95(5):1015–1039, 5 2020. ISSN 00256196. doi: 10.1016/j.mayocp.2020.01.038.

C. Loveday, A. Sud, M. E. Jones, J. Broggio, S. Scott, F. Gronthound, B. Torr, A. Garrett, D. L. Nicol, S. Jhanji, S. A. Boyce, M. Williams, C. Barry, E. Riboli, E. Kipps, E. McFerran, D. C. Muller, G. Lyratzopoulos, M. Lawler, M. Abulafi, R. S. Houlston, and C. Turnbull. Prioritisation by FIT to mitigate the impact of delays in the 2-week wait colorectal cancer referral pathway during the COVID-19 pandemic: a UK modelling study. *Gut*, 70(6):1053–1060, 6 2021. ISSN 0017-5749. doi: 10.1136/gutjnl-2020-321650.

A. M. Gab Allah, A. M. Sarhan, and N. M. Elshennawy. Edge U-Net: Brain tumor segmentation using MRI based on deep U-Net model with boundary information. *Expert Systems with Applications*, 213:118833, 3 2023. ISSN 09574174. doi: 10.1016/j.eswa.2022.118833.

V. Macedo Silva, T. Lima Capela, M. Freitas, R. Sousa Magalhães, C. Arieira, S. Xavier, P. Boal Carvalho, B. Rosa, M. J. Moreira, and J. Cotter. Small Bowel CLEansing Assessment and Report (SB-CLEAR): Standardizing bowel preparation report in capsule endoscopy. *Journal of Gastroenterology and Hepatology (Australia)*, 2022. ISSN 14401746. doi: 10.1111/jgh.16086.

M. Mackiewicz, J. Berens, and M. Fisher. Wireless Capsule Endoscopy Color Video Segmentation. *IEEE Transactions on Medical Imaging*, 27(12):1769–1781, 12 2008. ISSN 0278-0062. doi: 10.1109/TMI.2008.926061.

A. Maieron, D. Hubner, B. Blaha, C. Deutsch, T. Schickmair, A. Ziachehabi, E. Kerstan, P. Knoflach, and R. Schoefl. Multicenter Retrospective Evaluation of Capsule Endoscopy in Clinical Routine. *Endoscopy*, 36(10):864–868, 9 2004. ISSN 0013-726X. doi: 10.1055/s-2004-825852.

H. Makimoto, M. Höckmann, T. Lin, D. Glöckner, S. Gerguri, L. Clasen, J. Schmidt, A. Assadi-Schmidt, A. Bejinariu, P. Müller, S. Angendohr, M. Babady, C. Brinkmeyer, A. Makimoto, and M. Kelm. Performance of a convolutional neural network derived from an ECG database in recognizing myocardial infarction. *Scientific Reports*, 10(1):8445, 5 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-65105-x.

A. V. Mamonov, I. N. Figueiredo, P. N. Figueiredo, and Y. H. Richard Tsai. Automated polyp detection in colon capsule endoscopy. *IEEE Transactions on Medical Imaging*, 33 (7):1488–1502, 2014. ISSN 1558254X. doi: 10.1109/TMI.2014.2314959.

R. Manjunath and K. Kwadiki. Modified U-NET on CT images for automatic segmentation of liver and its tumor. *Biomedical Engineering Advances*, 4:100043, 12 2022. ISSN 26670992. doi: 10.1016/j.bea.2022.100043.

K. Margatina, G. Vernikos, L. Barrault, and N. Aletras. Active Learning by Acquiring Contrastive Examples. In *Association for Computational Linguistics*, pages 650–663, 11 2021. doi: 10.18653/v1/2021.emnlp-main.51.

M. J. Mascarenhas Saraiva, J. Afonso, T. Ribeiro, P. Cardoso, F. Mendes, M. Martins, A. P. Andrade, H. Cardoso, M. Mascarenhas Saraiva, J. Ferreira, and G. Macedo. AI-Driven Colon Cleansing Evaluation in Capsule Endoscopy: A Deep Learning Approach. *Diagnostics*, 13(23), 12 2023. ISSN 20754418. doi: 10.3390/diagnostics13233494.

L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, 2 2018.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space, 1 2013.

A. Musha, R. Hasnat, A. A. Mamun, E. P. Ping, and T. Ghosh. Computer-Aided Bleeding Detection Algorithms for Capsule Endoscopy: A Systematic Review. *Sensors*, 23(16): 7170, 8 2023. ISSN 1424-8220. doi: 10.3390/s23167170.

J. H. Nam, Y. Hwang, D. J. Oh, J. Park, K. B. Kim, M. K. Jung, and Y. J. Lim. Development of a deep learning-based software for calculating cleansing score in small bowel capsule endoscopy. *Scientific Reports*, 11(1), 12 2021a. ISSN 20452322. doi: 10.1038/s41598-021-81686-7.

J. H. Nam, D. J. Oh, S. Lee, H. J. Song, and Y. J. Lim. Development and verification of a deep learning algorithm to evaluate small-bowel preparation quality. *Diagnostics*, 11(6), 6 2021b. ISSN 20754418. doi: 10.3390/diagnostics11061127.

S. Nennstiel, A. Machanek, S. von Delius, B. Neu, B. Haller, M. Abdelhafez, R. M. Schmid, and C. Schlag. Predictors and characteristics of angioectasias in patients with obscure gastrointestinal bleeding identified by video capsule endoscopy. *United European Gastroenterology Journal*, 5(8):1129–1135, 12 2017. ISSN 2050-6406. doi: 10.1177/2050640617704366.

R. Noorda, A. Nevárez, A. Colomer, V. Pons Beltrán, and V. Naranjo. Automatic evaluation of degree of cleanliness in capsule endoscopy based on a novel CNN architecture. *Scientific Reports*, 10(1), 12 2020. ISSN 20452322. doi: 10.1038/s41598-020-74668-8.

M. Ozkan, M. Cakiroglu, O. Kocaman, M. Kurt, B. Yilmaz, G. Can, U. Korkmaz, E. Dandil, and Z. Eksi. Age-based computer-aided diagnosis approach for pancreatic cancer on endoscopic ultrasound images. *Endoscopic Ultrasound*, 5(2):101, 2016. ISSN 2303-9027. doi: 10.4103/2303-9027.180473.

D. E. O'Sullivan, A. Metcalfe, T. W. R. Hillier, W. D. King, S. Lee, J. Pader, and D. R. Brenner. Combinations of modifiable lifestyle behaviours in relation to colorectal cancer risk in Alberta's Tomorrow Project. *Scientific Reports*, 10(1):20561, 11 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-76294-w.

Participants in the Paris Workshop. The Paris endoscopic classification of superficial neoplastic lesions: esophagus, stomach, and colon. *Gastrointestinal Endoscopy*, 58(6), 12 2002. ISSN 00165107. doi: 10.1016/S0016-5107(03)02159-X.

G. Pascual, P. Laiz, A. García, H. Wenzek, J. Vitrià, and S. Seguí. Time-based self-supervised learning for Wireless Capsule Endoscopy. *Computers in Biology and Medicine*, 146:105631, 7 2022. ISSN 00104825. doi: 10.1016/j.compbiomed.2022.105631.

K. Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. doi: 10.1080/14786440109462720.

M. Pennazio, E. Rondonotti, E. J. Despott, X. Dray, M. Keuchel, T. Moreels, D. S. Sanders, C. Spada, C. Carretero, P. Cortegoso Valdivia, L. Elli, L. Fuccio, B. Gonzalez Suarez, A. Koulaouzidis, L. Kunovsky, D. McNamara, H. Neumann, E. Perez-Cuadrado-Martinez, E. Perez-Cuadrado-Robles, S. Piccirelli, B. Rosa, J.-C. Saurin, R. Sidhu, I. Tacheci, E. Vlachou, and K. Triantafyllou. Small-bowel capsule endoscopy and device-assisted enteroscopy for diagnosis and treatment of small-bowel disorders: European Society of Gastrointestinal Endoscopy (ESGE) Guideline – Update 2022. *Endoscopy*, 55(01):58–95, 1 2023. ISSN 0013-726X. doi: 10.1055/a-1973-3796.

J. Pennington, R. Socher, and C. D. Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 1532–1543, Doha, Qatar, 10 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162.

G. Petmezas, L. Stefanopoulos, V. Kilintzis, A. Tzavelis, J. A. Rogers, A. K. Katsaggelos, and N. Maglaveras. State-of-the-Art Deep Learning Methods on Electrocardiogram Data: Systematic Review. *JMIR Medical Informatics*, 10(8):e38454, 8 2022. ISSN 2291-9694. doi: 10.2196/38454.

P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. P. Langlotz, B. N. Patel, K. W. Yeom, K. Shpanskaya, F. G. Blankenberg, J. Seekins, T. J. Amrhein, D. A. Mong, S. S. Halabi, E. J. Zucker, A. Y. Ng, and M. P. Lungren. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt

algorithm to practicing radiologists. *PLOS Medicine*, 15(11):e1002686, 11 2018. ISSN 1549-1676. doi: 10.1371/journal.pmed.1002686.

E. Rondonotti, C. Spada, S. Adler, A. May, E. Despott, A. Koulaouzidis, S. Panter, D. Domagk, I. Fernandez-Urien, G. Rahmi, M. Riccioni, J. van Hooft, C. Hassan, and M. Pennazio. Small-bowel capsule endoscopy and device-assisted enteroscopy for diagnosis and treatment of small-bowel disorders: European Society of Gastrointestinal Endoscopy (ESGE) Technical Review. *Endoscopy*, 50(04):423–446, 4 2018. ISSN 0013-726X. doi: 10.1055/a-0576-0566.

E. Rondonotti, M. Pennazio, E. Toth, and A. Koulaouzidis. How to read small bowel capsule endoscopy: a practical guide for everyday use. *Endoscopy International Open*, 08(10): E1220–E1224, 10 2020. ISSN 2364-3722. doi: 10.1055/a-1210-4830.

O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science*. Springer, 11 2015. ISBN 978-3-319-24574-4. doi: 10.1007/978-3-319-24574-4_28.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning Internal Representations by Error Propagation. In *Parallel Distributed Processing*, volume 1. The MIT Press, 1986. doi: 10.7551/mitpress/5236.003.0012.

C. Schröder, A. Niekler, and M. Potthast. Revisiting Uncertainty-based Query Strategies for Active Learning with Transformers, 7 2021.

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128(2):336–359, 2 2020. ISSN 0920-5691. doi: 10.1007/s11263-019-01228-7.

B. Settles. Active Learning Literature Survey. Technical report, University of Wisconsin–Madison, 2009. URL http://digital.library.wisc.edu/1793/60660.

N. Shamsudhin, V. I. Zverev, H. Keller, S. Pane, P. W. Egolf, B. J. Nelson, and A. M. Tishin. Magnetically guided capsule endoscopy. *Medical Physics*, 44(8), 8 2017. ISSN 0094-2405. doi: 10.1002/mp.12299.

C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423, 7 1948. ISSN 00058580. doi: 10.1002/j.1538-7305.1948.tb01338.x.

R. L. Siegel, A. N. Giaquinto, and A. Jemal. Cancer statistics, 2024. *CA: A Cancer Journal for Clinicians*, 74(1):12–49, 1 2024. ISSN 0007-9235. doi: 10.3322/caac.21820.

P. H. Smedsrud, V. Thambawita, S. A. Hicks, H. Gjestang, O. O. Nedrejord, E. Næss, H. Borgli, D. Jha, T. J. D. Berstad, S. L. Eskeland, M. Lux, H. Espeland, A. Petlund,

D. T. D. Nguyen, E. Garcia-Ceja, D. Johansen, P. T. Schmidt, E. Toth, H. L. Hammer, T. de Lange, M. A. Riegler, and P. Halvorsen. Kvasir-Capsule, a video capsule endoscopy dataset. *Scientific Data*, 8(1), 12 2021. ISSN 20524463. doi: 10.1038/s41597-021-00920-z.

S. Soffer, E. Klang, O. Shimon, Y. Barash, N. Cahan, H. Greenspana, and E. Konen. Deep learning for pulmonary embolism detection on computed tomography pulmonary angiogram: a systematic review and meta-analysis. *Scientific Reports*, 11(1):15814, 8 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-95249-3.

G. Son, T. Eo, J. An, D. Oh, Y. Shin, H. Rha, Y. Kim, Y. Lim, and D. Hwang. Small Bowel Detection for Wireless Capsule Endoscopy Using Convolutional Neural Networks with Temporal Filtering. *Diagnostics*, 12(8):1858, 7 2022. ISSN 2075-4418. doi: 10.3390/diagnostics12081858.

H. J. Song, J. S. Moon, J. H. Do, I. H. Cha, C. H. Yang, M. G. Choi, Y. T. Jeen, and H. J. Kim. Guidelines for bowel preparation before video capsule endoscopy. *Clinical Endoscopy*, 46(2):147–154, 2013. ISSN 22342400. doi: 10.5946/ce.2013.46.2.147.

T. Tabone, A. Koulaouzidis, and P. Ellul. Scoring Systems for Clinical Colon Capsule Endoscopy—All You Need to Know. *Journal of Clinical Medicine*, 10(11):2372, 5 2021. ISSN 2077-0383. doi: 10.3390/jcm10112372.

L. Tan, H. Li, J. Yu, H. Zhou, Z. Wang, Z. Niu, J. Li, and Z. Li. Colorectal cancer lymph node metastasis prediction with weakly supervised transformer-based multi-instance learning. *Medical & Biological Engineering & Computing*, 61(6):1565–1580, 6 2023. ISSN 0140-0118. doi: 10.1007/s11517-023-02799-x.

M. Tan and Q. V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, 5 2019.

R. Thalappillil, P. Datta, S. Datta, Y. Zhan, S. Wells, F. Mahmood, and F. C. Cobey. Artificial Intelligence for the Measurement of the Aortic Valve Annulus. *Journal of Cardiothoracic and Vascular Anesthesia*, 34(1):65–71, 1 2020. ISSN 10530770. doi: 10.1053/j.jvca.2019.06.017.

L. Valle. Genetic predisposition to colorectal cancer: Where we stand and future perspectives. *World Journal of Gastroenterology*, 20(29):9828, 2014. ISSN 1007-9327. doi: 10.3748/wjg.v20.i29.9828.

D. Varam, R. Mitra, M. Mkadmi, R. A. Riyas, D. A. Abuhani, S. Dhou, and A. Alzaatreh. Wireless Capsule Endoscopy Image Classification: An Explainable AI Approach. *IEEE Access*, 11:105262–105280, 2023. ISSN 2169-3536. doi: 10.1109/ACCESS.2023.3319068.

A. Vaswani, G. Brain, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems 30 (NIPS)*, 2017. doi: 10.48550/arXiv.1706.03762.

F. E. Vuik, S. A. Nieuwenburg, S. Moen, C. Spada, C. Senore, C. Hassan, M. Pennazio, E. Rondonotti, S. Pecere, E. J. Kuipers, and M. C. Spaander. Colon capsule endoscopy in colorectal cancer screening: A systematic review. *Endoscopy*, 53(8):815–824, 8 2021. ISSN 14388812. doi: 10.1055/a-1308-1297.

S. Vujosevic, S. J. Aldington, P. Silva, C. Hernández, P. Scanlon, T. Peto, and R. Simó. Screening for diabetic retinopathy: new perspectives and challenges. *The Lancet Diabetes & Endocrinology*, 8(4):337–347, 4 2020. ISSN 22138587. doi: 10.1016/S2213-8587(19)30411-5.

R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi. Medical image segmentation using deep learning: A survey. *IET Image Processing*, 16(5):1243–1267, 4 2022. ISSN 1751-9659. doi: 10.1049/ipr2.12419.

A. Weissferdt. *Diagnostic Thoracic Pathology*. Springer International Publishing, Cham, 2020. ISBN 978-3-030-36437-3. doi: 10.1007/978-3-030-36438-0.

N. Wilson, M. Gabr, and M. Bilal. Endoscopic Recognition and Resection of Malignant Colorectal Polyps. *Techniques and Innovations in Gastrointestinal Endoscopy*, 25(4): 385–398, 1 2023. ISSN 25900307. doi: 10.1016/j.tige.2023.03.001.

D. Withey and Z. Koles. Medical Image Segmentation: Methods and Software. In *2007 Joint Meeting of the 6th International Symposium on Noninvasive Functional Source Imaging of the Brain and Heart and the International Conference on Functional Biomedical Imaging*, pages 140–143. IEEE, 10 2007. ISBN 978-1-4244-0948-8. doi: 10.1109/NFSI-ICFBI.2007.4387709.

J. T. Wu, K. C. L. Wong, Y. Gur, N. Ansari, A. Karargyris, A. Sharma, M. Morris, B. Saboury, H. Ahmad, O. Boyko, A. Syed, A. Jadhav, H. Wang, A. Pillai, S. Kashyap, M. Moradi, and T. Syeda-Mahmood. Comparison of Chest Radiograph Interpretations by Artificial Intelligence Algorithm vs Radiology Residents. *JAMA Network Open*, 3(10): e2022779, 10 2020. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2020.22779.

S. Xavier, B. Rosa, S. Monteiro, C. Arieira, R. Magalhães, T. Cúrdia Gonçalves, P. Boal Carvalho, J. Magalhães, M. J. Moreira, and J. Cotter. Bowel preparation for small bowel capsule endoscopy – The later, the better! *Digestive and Liver Disease*, 51(10):1388–1391, 10 2019. ISSN 18783562. doi: 10.1016/j.dld.2019.04.014.

Q. Xu, X. Zhan, Z. Zhou, Y. Li, P. Xie, S. Zhang, X. Li, Y. Yu, C. Zhou, L. Zhang, O. Gevaert, and G. Lu. AI-based analysis of CT images for rapid triage of COVID-19 patients. *npj Digital Medicine*, 4(1):75, 4 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00446-z.

H. Yamamoto, H. Ogata, T. Matsumoto, N. Ohmiya, K. Ohtsuka, K. Watanabe, T. Yano, T. Matsui, K. Higuchi, T. Nakamura, and K. Fujimoto. Clinical Practice Guideline

for Enteroscopy. *Digestive Endoscopy*, 29(5):519–546, 7 2017. ISSN 0915-5635. doi: 10.1111/den.12883.

P. Ye, Y. Xi, Z. Huang, and P. Xu. Linking Obesity with Colorectal Cancer: Epidemiology and Mechanistic Insights. *Cancers*, 12(6):1408, 5 2020. ISSN 2072-6694. doi: 10.3390/cancers12061408.

Y. Yuan and M. Q.-H. Meng. A Novel Feature for Polyp Detection in Wireless Capsule Endoscopy images. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014. ISBN 9781479969340. doi: 10.1109/IROS.2014.6943274.

Y. Yuan and M. Q.-H. Meng. Deep learning for polyp recognition in wireless capsule endoscopy images. *Medical Physics*, 44(4):1379–1389, 4 2017. ISSN 00942405. doi: 10.1002/mp.12147.

Y. Yuan, W. Qin, B. Ibragimov, G. Zhang, B. Han, M. Q.-H. Meng, and L. Xing. Densely Connected Neural Network With Unbalanced Discriminant and Category Sensitive Constraints for Polyp Recognition. *IEEE Transactions on Automation Science and Engineering*, 17(2):574–583, 4 2020. ISSN 1545-5955. doi: 10.1109/TASE.2019.2936645.

J. Zhang, N. Boora, S. Melendez, A. Rakkunedeth Hareendranathan, and J. Jaremko. Diagnostic Accuracy of 3D Ultrasound and Artificial Intelligence for Detection of Pediatric Wrist Injuries. *Children*, 8(6):431, 5 2021. ISSN 2227-9067. doi: 10.3390/children8060431.

S. Zhao, T. Liu, B. Liu, and K. Ruan. Attention residual convolution neural network based on U-net (AttentionResU-Net) for retina vessel segmentation. *IOP Conference Series: Earth and Environmental Science*, 440(3):032138, 2 2020. ISSN 1755-1307. doi: 10.1088/1755-1315/440/3/032138.

X. Zhao, C. Fang, F. Gao, D.-J. FAN, X. Lin, and G. Li. Deep Transformers For Fast Small Intestine Grounding In Capsule Endoscope Video. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 150–154. IEEE, 4 2021. ISBN 978-1-6654-1246-9. doi: 10.1109/ISBI48211.2021.9433921.

B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929. IEEE, 6 2016. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.319.

Y. Zou, L. Li, Y. Wang, J. Yu, Y. Li, and W. J. Deng. Classifying digestive organs in wireless capsule endoscopy images based on deep convolutional neural network. In *2015 IEEE International Conference on Digital Signal Processing (DSP)*, pages 1274–1278. IEEE, 7 2015. ISBN 978-1-4799-8058-1. doi: 10.1109/ICDSP.2015.7252086.