IT UNIVERSITY OF COPENHAGEN

Towards Realistic Digital Twins: Segmentation and Reconstruction of Collapsed Colons from CT Imaging

IT University of Copenhagen
KISPECI1SE – Master Thesis

Niclas Claßen (niclc@itu.dk)
Ioana-Daria Vasile (iova@itu.dk)

Supervisor: Prof. Dr. Ir. Veronika Cheplygina
Co-Supervisor: Dr. Martina Finocchiaro (KU)

Date: 01.06.2025

Towards Realistic Digital Twins: Segmentation and Reconstruction of Collapsed Colons from CT Imaging

Ioana-Daria Vasile
IT-Univerity of Copenhagen
Copenhagen, Denmark
iova@itu.dk

Niclas Claßen
IT-Univerity of Copenhagen
Copenhagen, Denmark
niclc@itu.dk

Abstract

Digital twins of the colon are essential for advancing robotic colonoscopy systems aimed at improving colorectal cancer detection. Their effectiveness relies on accurate segmentation from CT scans, which is particularly challenging in cases of collapsed or fluid-filled colons. These cases are common but often excluded from prior work, reducing dataset diversity. We present a pipeline that segments both collapsed and non-collapsed colons by combining semi-automatic label generation, a novel method for quantifying the degree of collapse, and an U-Net model. For collapsed cases, we further explore the use of non-rigid image registration to infer missing anatomy. Our segmentation model shows strong performance across varying insufflation states, outperforming manual annotations in some collapsed cases. While registration results in continuous colon shapes, some anatomical inaccuracies remain.

Code: Available on GitHub [1].

1 Introduction

Colorectal cancer (CRC) is the third most commonly diagnosed cancer worldwide, accounting for approximately 10% of all cancer cases, according to a 2023 report by the World Health Organization. It is the third leading cause of cancer-related deaths among men and the fourth among women. When considering both sexes combined, CRC becomes the second most common cause of cancer mortality [2] [3]. Despite advances in prevention and early detection, CRC remains a major global health burden, even in high-income countries with modern healthcare systems. According to the World Cancer Research Fund, in 2022 Denmark recorded the highest age-standardised incidence rate of CRC at 48.1 per 100,000 people, followed by Norway (45.3), Hungary (44.2), the Netherlands (42.8), and Croatia (41.1) [4]. Given its high incidence and mortality, the impact of CRC can be substantially mitigated through prevention strategies such as adopting a healthy lifestyle, avoiding risk factors, and most importantly practicing early detection through screening, with colonoscopy considered the gold standard [2]. However, colonoscopy has noteworthy limitations: it can result in complications such as perforation or looping of the endoscope, which may cause discomfort and deter patients from undergoing the procedure [5, 6]. Moreover, a meta-analysis of tandem colonoscopies revealed miss rates for adenomas (precancerous lesions) as high as 33% in high-risk patients [7]. Importantly, successful CRC detection rates vary between physicians [8].

Given the limitations of current colonoscopy, researchers are continuously working to develop more accurate, reliable, and patient-friendly alternatives. Among these, the European project Intelligent Robotic Endoscope (IRE) aims to improve colon intubation, which refers to the process of advancing a colonoscope, a flexible tube with a camera, through the colon. The IRE project seeks to achieve this by combining artificial intelligence, robotics, biomechanical modeling, and simulation [9]. A key component of this initiative is the use of digital twins, which are comprehensive virtual models that integrate anatomical structure with physical properties such as elasticity and deformation [10]. They are necessary to train colon navigation systems, which are essential for developing automated colonoscopy technologies.

However, extracting the anatomical structure of the colon from medical images such as computed tomography (CT) is a challenging task, primarily due to the scarcity of high-quality, annotated data. Moreover, the colon often appears collapsed in these images, a condition we define as either insufficient insufflation, or visible fragmentation, where the colon is represented as discontinuous segments in the scan. As a result, many previous studies have limited their scope to non-collapsed cases in order to simplify the problem [11, 12]. Additionally, the complex anatomy of the colon makes it difficult even for state-of-the-art segmentation methods to capture its detailed structure accurately [13]. These simplifications may be acceptable for segmentation-only tasks, but it is insufficient for creating digital twins intended to train autonomous systems for colonoscopy. Excluding collapsed colons risks introducing bias and limits the utility of digital twins to only ideal cases, even though such conditions could be associated with patient-specific factors.

Therefore, it is essential to capture the anatomical structure of a wide range of cases, including both collapsed and non-collapsed colons. To support this, we use abdominal CT scans to segment the colon and, in cases where it appears collapsed, apply image registration techniques to reconstruct its geometry. To guide this work, the following two key research questions are addressed:

- **RQ-1:** How can the colon be segmented from abdominal CT scans with high anatomical detail, including cases where the colon is collapsed?
- **RQ-2:** Can image registration be used to reconstruct the shape of a collapsed colon from segmented CT data, in order to create anatomically realistic digital twins for robotic endoscopy applications?

Accurate segmentation is the foundation of effective colon reconstruction and, by extension, the creation of reliable digital twins. For this reason, addressing the first research question is the primary focus of our research. We aim to not only achieve high segmentation quality, but also to develop a method that generalizes across both collapsed and non-collapsed colons. This generalizability is essential, as prior knowledge of collapse is rarely available in CT scans, and a unified approach improves the overall applicability of digital twins for robotic endoscopy.

2 Background Knowledge

This section introduces key concepts relevant to our research, including the anatomy and function of the colon, as well as the characteristics of CRC. We also provide an overview of the current gold-standard colonoscopy procedure and discuss the fundamentals of CT imaging, which serves as the foundation for the data used in our study. Additionally, we cover medical image segmentation, registration techniques, and anomaly detection, all of which are crucial to our research methodology. Finally, we present specific challenges encountered when working with medical images.

2.1 Medical Background

2.1.1 ***Anatomy and Function of the Colon.*** The intestine, or bowel, is a muscular tube extending from the stomach to the anus and is divided into the *small* and *large intestine*. The large intestine is responsible for absorbing water and salts from liquid waste left by the small intestine, forming solid stool. It consists of three main parts: the *colon*, *rectum*, and *anus*, which together facilitate waste movement and elimination [14, 15]. Anatomically, the colon is about 1.5 meters long in adults and loops around the small intestine within the abdominal cavity [16] (see Figure 1). The shape and alignment of the colon can vary significantly between individuals (see Appendix 22–25 for examples), and the illustration shown is a simplified view of the overall shape. Furthermore, sex-based anatomical differences have been observed: in a study of 295 patients, the colon was significantly longer in females than in males (154.3 ± 18.1 cm vs. 147.1 ± 18.3 cm, $p = 0.022$) [17]. Such variations are relevant in clinical practice and computational modeling. For example, colonoscopy is often more technically challenging in women, possibly due to longer colon length [18].

Compared to the small intestine, the colon has a larger diameter, a more fixed position, and distinct features. It contains three longitudinal muscle bands that form *haustra*, pouches that give the colon its segmented appearance [15, 19]. The colon consists of five segments: *cecum, ascending colon, transverse colon, descending colon*, and *sigmoid colon*, as shown in Figure 1. Since colonoscopy involves navigating from the anus through the rectum to the colon, **we define colon segmentation and reconstruction as the process of segmenting and reconstructing the entire large intestine, including the colon, rectum, and anus.**

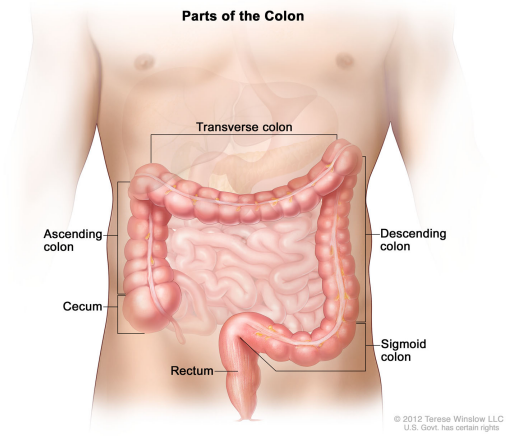

Figure 1: Illustration of the human colon, showing the cecum, ascending colon, transverse colon, descending colon, sigmoid colon, and rectum. (License to reproduce this image granted by Terese Winslow)

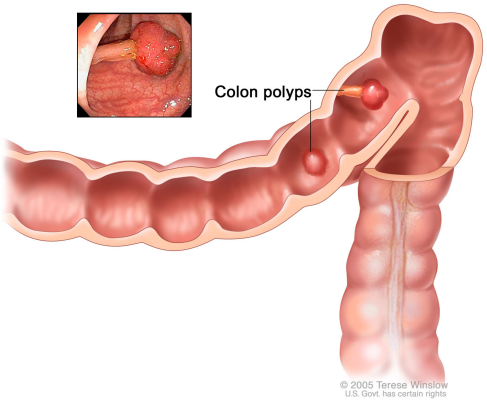

Figure 2: Illustration of polyps inside the colon. (License to reproduce this image granted by Terese Winslow)

2.1.2 ***Colorectal Cancer.*** CRC typically originates in the colon or rectum, often beginning as a polyp on the inner lining (see Figure 2). Adenomatous polyps (adenomas) are considered precancerous lesions and can develop into colorectal cancer over time. In contrast, other types, such as hyperplastic and inflammatory polyps, are usually noncancerous, though large hyperplastic polyps may require closer observation. The risk of malignant transformation increases with polyp size, number, and the presence of abnormal, but non-cancerous cellular changes [20].

Several factors increase the risk of developing CRC, with age being a primary contributor. Most cases occur in individuals aged 50 and older. Lifestyle factors also play a significant role, including high consumption of processed meats, low intake of fruits and vegetables, physical inactivity, obesity, smoking, and excessive alcohol use [2].

Although the exact duration for an adenoma to develop into cancer is not definitively known, evidence suggests that this process typically spans from ten to fifteen years. This slow progression provides a valuable window for early detection through regular screening and treatment [21]. Symptoms tend to appear only when the tumor becomes large enough to obstruct the bowel, leading to cramping, abdominal pain, rectal bleeding, sudden weight-loss or iron-deficiency anemia [22].

The impact of early detection is also reflected in survival statistics. For colon cancer, the five-year survival rate is 91% when diagnosed at a localized stage, but it drops to 14% when detected at a distant stage. Similarly, for rectal cancer, the five-year survival rate is 90% for localized cases and declines to 18% for distant-stage diagnoses. These statistics, reported by the American Cancer Society, are based on cases diagnosed between 2014 and 2020 [23].

2.1.3 ***Colonoscopy.*** Colonoscopy is an endoscopic procedure used to examine the inside of the large intestine. It involves the manual insertion of a colonoscope, a flexible tube with a lighted camera,

through the anus, rectum, and colon. The device transmits real-time images to a monitor, as illustrated in Figure 3. Colonoscopy is a key modality in preventive medicine, as it is primarily used to screen for precancerous polyps and early signs of CRC. It is considered the gold standard for CRC screening due to its ability to both detect and remove polyps during the procedure [24, Section 1].

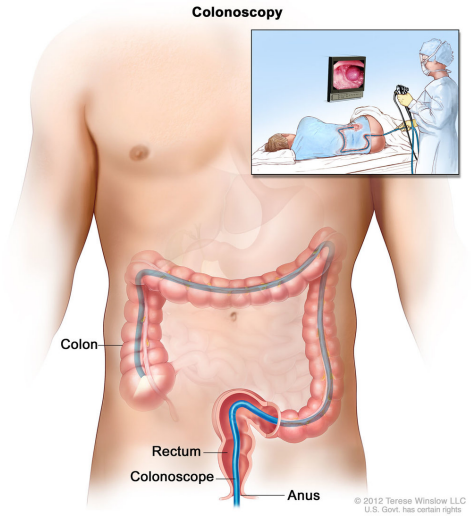

Figure 3: Illustration of a doctor performing colonoscopy. (License to reproduce this image granted by Terese Winslow)

2.1.4 ***CT Colonography*** . CT colonography (also known as virtual colonoscopy) is an advanced imaging technique that uses X-rays acquired from multiple angles to examine the large intestine. Unlike conventional X-rays, CT scans provide detailed 3D representations of the internal structures by rotating a narrow X-ray beam around the patient and capturing radiation intensities with opposing detectors. A computer then reconstructs the scanned area into volumetric images [25, 26].

Mathematically, CT image reconstruction is based on the *Radon transform*, which models each projection as a line integral of the internal *attenuation*. CT imaging relies on attenuation, the reduction in X-ray intensity due to absorption or scattering by tissues. Attenuation varies by tissue density: denser tissues like bone absorb more X-rays and appear brighter, while less dense tissues such as lungs allow more radiation through and appear darker. These variations are quantified in *Hounsfield Units* (HU), which determine the grayscale values in CT images [27].

In CT colonography, the colon is insufflated with gas to enhance the visibility of its internal walls, resulting in dark regions corresponding to gas-filled sections. These grayscale patterns form the basis for tissue segmentation. However, accurate colon segmentation presents several main challenges [28]:

- The colon is not the only gas-filled structure in the image: lungs, stomach, and the small intestine also contain air resulting in similar intensities, making them difficult to distinguish from the colon.
- Obstructions within the colon can be present due to incomplete bowel preparation. This results in residual fluid of varying intensities, can complicate automated segmentation and hinder the extraction of a continuous colon segment.
- Suboptimal insufflation, or pressure from adjacent organs, can cause the colon to collapse, resulting in under-expanded or discontinuous regions in the scan. These fragmented or "broken" appearances further increase the complexity of accurate colon segmentation.

2.2 Technical Background

2.2.1 ***Medical Image Segmentation***. Medical image segmentation is the process of dividing an image into meaningful regions based on features such as intensity, texture, or contrast. In clinical contexts, this process is essential for analyzing anatomical structures, detecting abnormalities such as tumors or lesions, estimating tissue volume changes over time, and planning treatments like radiation therapy [29]. Overall, segmentation methods can be broadly categorized as follows [30, Chapter 10]:

- **Edge-Based:** these methods detect object boundaries by identifying discontinuities in pixel intensity using filters like Sobel or Laplacian. The resulting edges are refined to form closed regions corresponding to anatomical structures.
- **Pixel-Based:** this category classifies each pixel individually based on image histogram statistics. Techniques such as Gaussian mixture models and the Otsu method are commonly used for thresholding. However, they often ignore spatial context, making them vulnerable to noise and intensity variations that can distort segmentation accuracy.
- **Region-Based:** these methods group neighboring pixels with similar characteristics into coherent regions. Techniques include region-growing, which expands from initial seed points, and region-splitting, which adjusts regions based on local similarity criteria.

2.2.2 ***U-Net Architecture***. At the core of many deep learning algorithms is the use of multi-layered neural networks that progressively transform input data, such as medical images, into meaningful outputs. In medical imaging, deep learning models have achieved remarkable success across a range of tasks, including classification, detection, and segmentation [31].

One of the most widely adopted neural network architectures for image segmentation is U-Net, introduced by Ronneberger et al. (2015) [32]. Specifically designed for biomedical segmentation, U-Net delivers accurate results even with limited training data. Its architecture follows a symmetric U-shape, consisting of a contracting path (encoder) that captures context via convolution and downsampling, and an expanding path (decoder) that restores spatial resolution through upsampling. Crucially, skip connections between corresponding encoder and decoder layers enable precise localization by combining deep semantic features with detailed spatial information. In medical imaging, U-Net based models have demonstrated strong performance in tasks such as lung segmentation from CT scans [33] and multi-organ segmentation in abdominal CT scans [34, 35]. A simplified diagram of the U-Net architecture is shown in Figure 4.

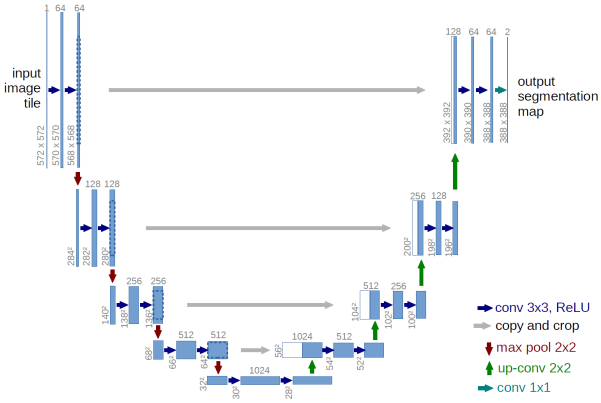

Figure 4: The U-Net architecture for biomedical image segmentation. Figure taken from the original U-Net publication [32].

2.2.3 ***Anomaly Detection.*** Anomaly detection is the task of identifying data patterns that deviate from what is normal or expected. These unusual instances (known as anomalies or outliers) stand apart from the rest of the data and can indicate rare or potentially important occurrences [36].

A key distinction within anomaly detection lies between outlier detection and novelty detection, although these terms are often used interchangeably in the literature. Outlier detection assumes that the training data may already contain anomalies. The goal is to model the dense regions of the data distribution while ignoring data points that are far from these regions. In contrast, novelty detection assumes the training data is clean and focuses on identifying whether new, unseen observations differ significantly from this normal distribution. Anomaly detection methods can also be categorized based on the number of classes used during training. In one-class classification, the model is trained solely or mostly on normal (inlier) data, learning a decision boundary that separates normal instances from any potential anomalies [37].

2.2.4 ***Medical Image Registration.*** Image registration is a well-established method in medical imaging, used to geometrically align images captured at different times, from different viewpoints, or across modalities. It plays a central role in tasks such as multi-modal image fusion, tracking anatomical changes, and supporting treatment planning [38]. Formally, it involves computing a spatial transformation that warps a moving image to align with a fixed reference. The success of registration largely depends on the choice of geometric transformation model, which generally falls into one of the following categories [30, Chapter 12], [39]:

- **Rigid transformations** preserve shape and size by allowing only translation and rotation. They are typically used in *intra-subject registration*, where anatomical structures remain largely unchanged.
- **Affine transformations** extend rigid ones by adding scaling and shearing, making them suitable for *inter-subject registration*, where there may be anatomical differences between subjects.
- **Non-rigid transformations** handle complex deformations like local stretching or bending. These are ideal for aligning images with significant anatomical differences, such as in non-linear registration of brain scans across multiple individuals or modalities.

In addition to the transformation model, the choice of similarity measure plays a critical role in image registration. It determines how alignment quality is quantified and guides the optimization process. Similarity measures are generally categorized as *intensity-based* or *feature-based*. Intensity-based methods operate on pixel or voxel (volumetric pixel) values and are well-suited for intra-modality registration, where intensity distributions are similar. A common example is *mean squared error* (MSE), which minimizes the average squared difference between corresponding pixels. Feature-based methods, by contrast, align extracted landmarks or structures such as edges, contours, or anatomical points, and are useful when intensity values vary. For multi-modal registration, where direct intensity comparison is unreliable (e.g., CT vs. MRI), *mutual information* (MI) is frequently used. It measures the statistical dependency between images and is maximized when shared information is preserved, making it robust across different modalities [40, Chapter 2].

2.2.5 ***Challenges in Medical Imaging.*** Working with medical imaging data for tasks like colon segmentation presents several major challenges, including the following:

- **Limited data availability and privacy constraints**: medical image datasets are often small and difficult to access due to strict privacy regulations and ethical concerns. Since they contain sensitive health information, securing their storage, processing and transmission is essential. Legal restrictions and risks of data breaches or tampering also limit data sharing and reuse [41].
- **Costly and time-consuming annotation**: annotating medical images requires expert clinicians, making the process expensive and slow. Despite the need for large, high-quality labeled datasets, availability remains limited. This bottleneck slows model training and may introduce bias, particularly when datasets are imbalanced or lack rare conditions [42].
- **Lack of diversity and standardization**: research datasets often fail to reflect the diversity seen in real-world clinical settings, despite hospitals maintaining large image archives. High variability in imaging protocols, contrast usage, dose levels, and image sizes, combined with incomplete or inconsistent metadata, makes it difficult to categorize or retrieve relevant data. In many cases, image analysis is needed just to identify usable subsets [43].

3 Related Work

Early research on automatic colon segmentation in CT images [12, 28, 44–48] primarily relies on traditional image processing techniques such as region growing and thresholding. These methods often utilize geometrical, voxel intensities, and anatomical features to distinguish the colon from other air-filled or soft tissue structures in CT scans.

Wyatt et al. (2000) [28] propose a method that combines automatic seed point detection with region growing, guided by prior knowledge of colon geometry. However, the resulting segmentation is not intended to be highly precise, but to produce a complete mask that excludes unrelated anatomical structures and could be further

refined. Masutani et al. (2001) [44] introduce an anatomy-oriented strategy for segmenting the colonic wall in CT colonography, incorporating prior knowledge of adjacent anatomical structures. Although they address the challenge of collapsed colons, their volume-based inclusion criterion (including regions based on their relative size to the largest component) lacks anatomical specificity. As a result, this method can lead to misclassifications of small bowel segments as colon and may exclude parts of the collapsed colon that fall below the threshold.

Building on this work, Näppi et al. (2002) [45] develop an automated technique aimed at minimizing extracolonic components such as the small intestine. Their two-stage approach begins with anatomy-based extraction to identify colon candidates, followed by a colon specific analysis to remove non-colonic regions. While effective for inflated colons, the method struggles with collapsed segments. Bert et al. (2009) [12] propose an adaptive 3D region-growing algorithm for segmenting air-inflated, cleansed colons. Their pipeline includes external air masking, lung segmentation, and colon extraction, with a self-adjusting growth condition based on local intensity variations.

Most of these early methods focus on well-insufflated and cleansed colons and face limitations when applied to images with collapsed segments. Chowdhury et al. (2011) [47] directly address this issue by introducing a method specifically designed for the automatic segmentation of collapsed colons in CT images. Their approach consists of two main steps: first removing surrounding air voxels and lung tissues and second, identifying and labeling the remaining air regions. Colon segmentation is then performed based on geometrical features such as volume-to-length ratio, orientation, segment length, endpoints and centerline gradient. Despite its effectiveness, the method has a notable limitation. The algorithm rejects datasets as unsuitable for automated analysis if the total segmented colon length falls below a certain threshold. This typically occurs in scans with poor insufflation, where up to 50% of the colon may be filled with residual material or fluid.

In a related effort, Yang et al. (2015) [48] propose a graph inference-based segmentation framework that is intended to be robust to extracolonic structures and collapse. However, a limitation of this approach is that false negatives introduced during their initial classification step cannot be recovered at a later stage. Furthermore, they address the presence of fluid through an automatic method based on thresholding and gradient features. This approach avoids manual intervention and improves robustness in poorly prepared scans, reason why we adopt this fluid identification strategy as part of our own pipeline.

More recently, segmentation methods have shifted toward the use of neural networks. Deep learning has gained increasing popularity in medical image analysis, including for colon segmentation tasks [11, 13].

One of the latest achievements is TotalSegmentator [13], a deep learning model trained on over 1,000 CT scans to segment 104 different anatomical structures. While it has demonstrated strong performance on external datasets and outperforms many publicly available models, its accuracy in colon segmentation remains limited. The segmentation masks produced for the colon are often coarse and lack the anatomical detail required for high-precision applications such as digital twin reconstruction. In fact, the authors

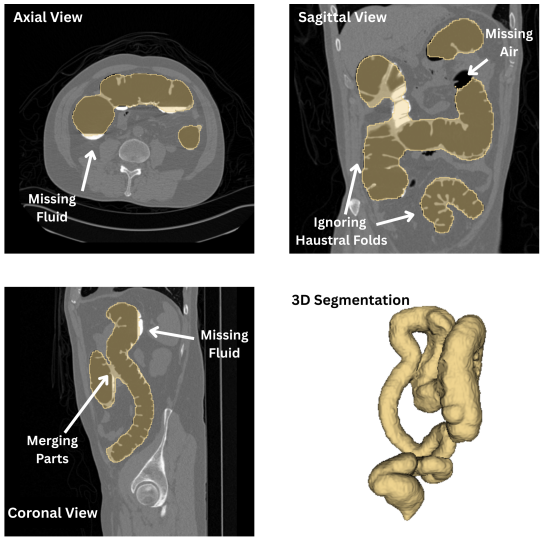

Figure 5: Example of colon segmentation by TotalSegmentator on a CT scan from the $\mathcal{D}_{\text{TCIA}}$ dataset. The segmentation misses parts of the colon and associated fluid regions, and fails to accurately capture haustral folds.

of TotalSegmentator explicitly identify missing parts of the colon as the most common segmentation error, occurring in approximately 35% of evaluated cases. An example of a segmented colon generated by TotalSegmentator is shown in Figure 5.

A more relevant study to our work is by Finocchiaro et al. [11], who propose a high-quality colon segmentation model that outperforms TotalSegmentator in several metrics. Their model architecture and dataset are closely aligned with ours, though their work does not specifically address collapsed colons or reconstruction.

The reconstruction of collapsed or disconnected colon segments remains an underexplored area. The only relevant work we identified is by Lu and Zhao (2011) [49], who propose a heuristic connection algorithm to link disconnected colon parts with straight lines for use in virtual colonoscopy. While this method provides basic anatomic continuity, it lacks anatomical plausibility and does not support downstream applications such as accurate colon unfolding or digital twin creation.

In contrast to previous methods, our work specifically addresses the challenges posed by poorly insufflated or fluid-filled colons. We propose a segmentation approach capable of handling colons regardless of their degree of insufflation or the presence of residual material. This is particularly important, as such conditions are common in clinical imaging and often limit the number of usable scans, thereby reducing dataset diversity and model generalizability. To support consistent evaluation, we introduce a method to quantify the degree of colonic collapse, enabling systematic comparison across varying insufflation states. In addition, we explore the reconstruction of collapsed or disconnected segments using image registration techniques, with the goal of restoring anatomical continuity in cases where conventional methods fall short.

Dataset	Collapse Level	Subjects (unique)	Scans (total)	Sex (F/M/U)	Positions (P/S/O)	Notes
$\mathcal{D}_{\textbf{TCIA}}$	M	825	3,451	401/353/71	850/855/1,746	Multicenter dataset from TCIA [50] of CT colonographies, including prone, supine, and other positions. Serves as the primary source for this work.
$\mathcal{D}_{\textbf{Finocchiaro}}$	NC	315	435	133/150/32	212/219/4	High-quality segmentations of non-collapsed colons derived from $\mathcal{D}_{\text{TCIA}}$ by Finocchiaro et al. (2025) [11]. Air-filled regions were extracted using region-growing, and fluid-filled regions were annotated using RootPainter [51]. All segmentations were subsequently quality-checked by an expert.
$\mathcal{D}_{\textbf{SA-Labels}}$	M	245	370	136/94/15	194/176/0	Semi-automatically labeled dataset developed in this work. Combines automated segmentation with expert correction. Includes both collapsed and non-collapsed cases.
$\mathcal{D}_{\textbf{Expert-NC}}$	NC	8	12	2/6/0	4/8/0	Expert-annotated non-collapsed colon masks used as gold-standard reference. Developed as part of the work by Finocchiaro et al. (2025) [11].
$\mathcal{D}_{\textbf{Expert-C}}$	C	9	10	6/3/0	4/6/0	Expert-annotated collapsed colon masks created specifically for this study to serve as gold-standard reference data. We use our method to quantify the degree of colon collapse to select varying collapse levels.
$\mathcal{D}_{\textbf{CNC-Test}}$	M	88	129	40/39/9	61/68/0	Combined test set assembled from $\mathcal{D}_{\text{Finocchiaro}}$ and $\mathcal{D}_{\text{SA-Labels}}$ for evaluating. Includes only scans and subjects not used during training.
$\mathcal{D}_{\textbf{CNC-Train}}$	M	347	513	174/142/31	253/260/0	Combined train set assembled from $\mathcal{D}_{\text{Finocchiaro}}$ and $\mathcal{D}_{\text{SA-Labels}}$. Used for model training.

Table 1: Summary of all datasets used in this work. The Collapse column uses the abbreviations: `C = Collapsed`, `NC = Non-Collapsed`, and `M = Mixed` (containing both collapsed and non-collapsed cases). Sex is reported as Female (F), Male (M), or Unknown (U). Patient positions are recorded as Prone (P), Supine (S), or Other (O). The table includes the number of unique subjects and total scans, along with brief notes on each dataset's origin and use.

4 Data

This project uses the CT Colonography dataset from The Cancer Imaging Archive (TCIA) [50], referred to in this work as $\mathcal{D}_{\text{TCIA}}$. The dataset originates from a multicenter clinical trial conducted at 14 study sites across the United States [52]. It includes 3,451 CT scans from 825 patients. In some cases, multiple scans were acquired per patient, including not only prone (lying face down) and supine (lying face up) positions but also additional orientations. Accompanying metadata includes patient demographics (e.g., age, sex) and scan-specific information such as patient position, acquisition date, and scanner model. In addition to the raw CT scans, we make use of several datasets derived from $\mathcal{D}_{\text{TCIA}}$ for different segmentation tasks and evaluations. Table 1 provides a summary of all datasets used in this work, including level of collapse, number of subjects and scans, sex distribution, and scan acquisition positions.

5 Methodology

We propose a robust pipeline for the segmentation and reconstruction of the large intestine from abdominal CT scans, designed to handle both collapsed and non-collapsed colons, as well as presence of fluid. Our method does not rely on prior knowledge of the colon's collapse degree and proceeds through the following key steps:

(1) **Data Preparation:** we filter the dataset to remove incomplete or low-quality scans and ensure a representative distribution of collapsed and non-collapsed cases.
(2) **Semi-Automatic Label Generation:** we generate initial segmentation masks using traditional, non-learning-based methods, including intensity thresholding and connected component analysis, addressing the absence of ground truth labels for collapsed colons.
(3) **Quantifying Colon Collapse:** we estimate the degree of collapse in each segmented colon to guide downstream processing.

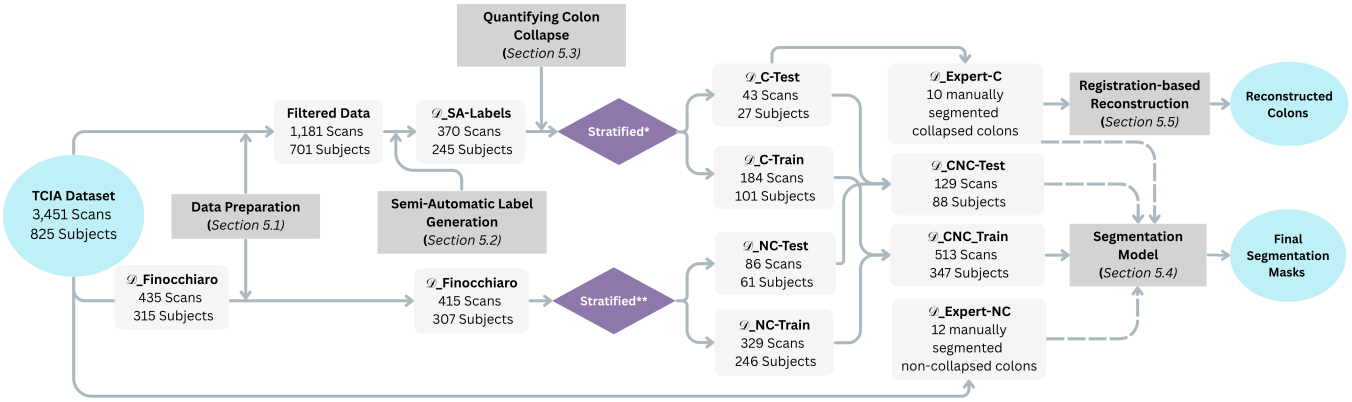

Figure 6: Data flow diagram of the proposed method for colon segmentation and reconstruction. *Data is stratified by degree of collapse, position, and subject. **Data is stratified by subject to match those in $\mathcal{D}_{\text{C-train}}$. Dashed arrows indicate components used solely for evaluation.

(4) **Colon Segmentation Model:** we train a U-Net based model to accurately segment any colon, regardless of state.
(5) **Registration-Based Reconstruction:** we register anatomically similar non-collapsed colons to collapsed ones in order to infer missing or distorted structures.

The complete data flow and methodological pipeline is summarized in Figure 6.

5.1 Data preparation

To prepare the dataset for our semi-automatic label generation method, we first exclude all cases from $\mathcal{D}_{\text{TCIA}}$ that are already annotated in $\mathcal{D}_{\text{Finocchiaro}}$, a dataset containing high-quality segmentations of non-collapsed colons from prior work [11]. Based on the findings of this research, we assume that the remaining cases in $\mathcal{D}_{\text{TCIA}}$ are predominantly, though not exclusively, collapsed. Inspired by their methodology, we apply a data filtering step to ensure consistency and quality across the remaining scans. The same filtering is also applied to $\mathcal{D}_{\text{Finocchiaro}}$, as it is used later in our pipeline and must meet the same preprocessing standards.The criteria for exclusion are as follows:

- **Position:** Scans are excluded if the patient position is neither prone nor supine, as these are the common positions for CT colonography.
- **Age:** Patients younger than 50 or older than 90 years are excluded to focus on the CRC target demographic range (see Appendix 13 for age distribution).
- **Z-Dimension:** Images with a slice count (Z-dimension) outside the range [350, 700] are considered low quality or non-standard and are excluded (see Appendix 14 for distribution).
- **Duplicates:** In cases of duplicate scans (e.g., multiple versions such as v1, v2) for the same subject-position pair, only version 1 is retained.

After applying these filters, $\mathcal{D}_{\text{TCIA}}$ contains 1,181 CT scans from 701 subjects and $\mathcal{D}_{\text{Finocchiaro}}$ 415 scans from 307 subjects.

5.2 Semi-Automatic Label Generation

Given the complexity of our task, including interference from the small intestine, collapsed colons, and fluid-filled regions, traditional methods alone are insufficient. Therefore, we adopt a semi-automatic pipeline that combines automated segmentation with manual correction on a subset of the generated labels. This hybrid approach not only reduces the time required compared to fully manual segmentation, but also allows for efficient correction of errors that traditional methods like thresholding and region growing often fail to resolve.

This pipeline is designed to produce reasonably accurate, colon masks for training purposes, rather than to serve as a standalone segmentation method. The approach relies on a sequence of heuristic steps, with parameters selected empirically based on visual inspection and domain knowledge. Key design choices are the following:

- **Priority on coverage:** parameters are tuned to maximize anatomical completeness, even at the cost of allowing small false positives.
- **No exhaustive tuning:** the aim is to generate labels of sufficient quality to support effective model training, rather than to achieve optimal segmentation performance.
- **Manual post-processing:** outputs are manually refined to correct obvious errors and ensure anatomical plausibility.

5.2.1 ***Anatomical Isolation.*** The first goal is to isolate air-filled gastrointestinal structures, while removing irrelevant air regions such as external air and lungs. We apply the following steps:

(1) **Air Identification with Thresholding:** we identify air regions by applying a fixed intensity threshold of −500 HU, empirically determined through histogram analysis of pixel intensities.
(2) **External Air Removal with Region Growing:** we use a 3D region growing operation, with seed points placed in the corners of the image, to isolate and remove external air surrounding the patient's body. This step helps eliminate the largest air region in the scan.

(3) **Connected Component Analysis for Bed Removal:** after removing external air, we perform connected component analysis on the remaining mask. The two largest components are assumed to correspond to the patient's body and the supporting bed. The second-largest component (typically the bed) is removed from the volume.
(4) **Lung Removal:** Finally, we use segmentation masks created with TotalSegmentator to remove lung regions.

After these steps, remaining air-filled regions in the CT images include the colon, small intestine, and occasionally the stomach.

5.2.2 ***Component-Level Colon Classification.*** After isolating gastrointestinal air-filled regions, the key challenge is to distinguish between components of the colon and those of the small intestine. Both appear as tubular structures and are anatomically connected, but differ in shape, location, and structure. To classify each air-filled connected component as colon or small intestine, we apply a voting system based on four independent criteria:

(1) **Anatomical Mask Overlap:**
- We use anatomical masks for both the small and large intestine provided by TotalSegmentator. Although these masks are not highly detailed, they are sufficient to provide a rough estimate of each organ's location and general path through the abdomen.
- To account for minor inaccuracies and boundary overlap, we apply morphological *dilation* to the small intestine mask (3×3×3 kernel) and *erosion* to the colon mask (2×2×2 kernel).
- For each air-filled component, we compute its degree of overlap with both anatomical masks. Components that overlap more with the colon mask are preliminarily marked as colon.

(2) **Shape-Based Classification:**
- We use geometric features to distinguish between the small intestine and the colon. The small intestine is generally thinner and more convoluted, whereas the colon is thicker and follows a more fixed anatomical course [16].
- *Elongation:* We compute the bounding box of each component within the air mask, which at this stage mainly includes the colon and small intestine. Specifically, we calculate the height-to-width ratio of the component's 2D bounding box in the axial plane. Components with a ratio > 5.5 are considered small intestine.
- *Solidity:* For components that are not already classified as small intestine based on elongation, we compute their average solidity. This is defined as the ratio between the actual area of the component and the area of its convex hull, computed across the axial, sagittal, and coronal planes. Components with an average solidity greater than 0.5 are considered colon.

(3) **Position-Based Heuristics:** Spatial information is used to distinguish colon components from those of the small intestine, based on two key anatomical assumptions: (1) the small intestine is typically located near the center of the abdominal cavity, and (2) the colon tends to extend peripherally, especially toward the posterior [16]. We evaluate two features:
- *Centrality Score:* We define the center of the volume using the bounding box of the combined intestine mask (colon + small intestine from TotalSegmentator). A spherical region centered at this point, with a radius set to 25% of the image width, is used to assess centrality. For each component, we compute the proportion of voxels located within this central region. A high proportion suggests the component is likely part of the small intestine.
- *Spatial Spread:* For each dimension, we calculate the spread as the difference between the maximum and minimum voxel coordinates within the component, divided by the corresponding dimension of the image volume. If a component is highly central, but also exhibits broad spatial spread, is more likely to be colon. In contrast, compact and tightly clustered components are typically small intestine.

(4) **Distance to Skeleton:** This step is inspired by prior work using centerlines to characterize tubular structures such as the colon [53]. However, defining centerlines for collapsed colons is particularly challenging, as it becomes unclear how the centerline should be oriented, especially for smaller, rounded segments where directionality is ambiguous. To address this, we compute a 3D skeleton of the full (non-eroded) colon mask generated by TotalSegmentator. Although the segmentation may be coarse, the skeleton captures the colon's general medial structure and connectivity. The skeleton is a one-pixel-wide, topologically representative curve derived from a binary volume [54]. We then assess the proximity of each component to the skeleton:
- We compute the Euclidean Distance Transform (EDT) of the inverse skeleton, producing a 3D map where each voxel stores its distance to the nearest skeleton point.
- This distance map is masked with the current component to isolate relevant values.
- We compute the median of the component's distances to the skeleton and normalize it by the square root of the component's voxel count. This normalization ensures that distance comparisons are not biased by component size.

Components with a normalized median distance below an empirically defined threshold of 0.1 are considered likely to be part of the colon.

Final Decision Rule: Each component receives a binary vote from the four tests above. A component is classified as colon if it receives at least 3 out of 4 votes. This voting system prioritizes recall of colon regions, favoring the inclusion of potential colon parts even at the cost of false positives. All thresholds were tuned empirically, and we found that removing false positives manually is far easier and more reliable than recovering missing colon regions.

5.2.3 ***Fluid Detection.*** While the initial segmentation focuses on air-filled regions, portions of the colon may also be filled with fluid, which are missed in that process. To obtain a more complete colon mask, we extend our method to include fluid-filled regions. Fluid appears brighter than air in CT scans but has highly variable intensity, ranging from gray to bright white, which overlaps with other tissues such as muscle, soft tissue, and bone. This makes intensity-based fluid segmentation particularly challenging. Our two-stage method builds on the work of Yang et al. [48], using

air-fluid boundary analysis and region growing to detect fluid-filled parts of the colon.

Stage 1: Slice-by-Slice Gap Detection and Fluid Expansion

(1) **Fluid Candidate Masking:** for each axial slice, we generate a binary fluid mask by thresholding voxel intensities above 100 HU. This captures potential fluid regions, though it includes false positives such as bone.
(2) **Distance Transform and Gradient Analysis:** we begin by computing the EDT for both the air and fluid masks. Next, we calculate the vertical (Y-axis) gradient of each EDT to capture directional changes in distance. To identify regions where air and fluid components oppose each other vertically, we compute the dot product of the two gradients and detect locations where the vectors point in opposite directions.
(3) **Gap Identification:** we identify candidate "gap" regions that lie between air and fluid. A voxel is considered part of a gap if it satisfies the following:
   - Close to both air and fluid regions.
   - Located at an air-fluid boundary based on opposing vertical gradients.
(4) **Shape-Based Filtering:** we apply connected component analysis to the raw gap mask to eliminate false positives. To refine the results, we retain only regions that meet specific geometric criteria: a minimum area, high eccentricity, and low flatness. These thresholds are adopted from the original work by Yang et al. [48].
(5) **Fluid Region Growing (2D):** we combine the validated gap regions with the air mask to create an initial seed region. Using binary propagation (region growing), we then expand this seed within the fluid mask to extend the colon segmentation into adjacent fluid areas, while preventing leakage into unrelated bright areas like muscle or bone.

Stage 2: 3D Post-Processing for Continuity

(6) **3D Refinement and Smoothing:** since the 2D merging step can result in discontinuities across slices, we perform a second round of 3D region growing. First, we apply a threshold of 200 HU to define a binary high-intensity fluid mask. Starting from the current colon mask (air + initial fluid), we expand into neighboring high-intensity voxels. To improve anatomical continuity, we apply a 3D Gaussian filter to the resulting binary mask and then re-threshold at 0.5 to restore binary form.

5.2.4 ***Manual Postprocessing & Expert Feedback.*** To evaluate the quality of the preliminary segmentation results, we perform manual inspection and correction on the generated labels. The scans are manually inspected using *3D Slicer* [55], an open-source platform for medical image analysis. Segmentation masks are visualized in 3D to assess both completeness and anatomical correctness. During this review, we check that no obvious false positives are present (e.g., small intestine, bones, or lungs) and that all visible colon segments, both air-filled and fluid-filled, are included.

Following this assessment, each scan is assigned to one of three categories:

- **Accepted as is**: no errors are found; the segmentation is used without modification.
- **Manually corrected**: minor errors are corrected using 3D Slicer's editing tools, such as removing false positives.
- **Discarded**: the segmentation is deemed too inaccurate to be reliably corrected and is excluded from further use.

As a final quality check, we obtained expert feedback on the generated colon masks. While this review was informal and not accompanied by quantitative metrics, it provides a valuable practical assessment of segmentation quality prior to model training. The resulting dataset, including all accepted and manually corrected segmentations, is referred to as $\mathcal{D}_{\text{SA-Labels}}$ (Semi-Automatic Labels).

5.3 Quantifying Colon Collapse

Quantifying the degree of colon collapse serves two key purposes in our pipeline. First, it enables stratification of the dataset when creating train and test splits, ensuring the model is exposed to a diverse range of collapse scenarios. This is essential for generalizing across varying anatomical states. Second, the collapse detection mechanism is intended to be applied after model inference: once a segmentation mask has been generated from a CT scan, our method automatically assesses whether the colon is collapsed and to what degree. By selecting only colons that are not severely collapsed, we reduce uncertainty in the reconstruction process and improve anatomical plausibility.

To quantify collapse, we extract two geometric features from each segmentation mask: the colon's volume and length. These features serve as proxies for inflation:

- **Volume:** Calculated as the total number of voxels within the 3D binary colon mask.
- **Length:** Approximated by computing a skeleton of the colon mask, similar to the procedure used during semi-automatic label generation. While this does not reflect true anatomical length, it provides an estimate for comparing inflation across samples.

Although one could attempt to detect collapsed colons by simply counting disconnected components in the segmentation mask, this approach is insufficient on its own. A colon may appear as a single connected component, yet still represent only a partial or collapsed structure. Conversely, a mask with multiple components might not indicate a severe collapse if the gaps between the components are minimal and their spatial relationship is coherent. For this reason, we additionally rely on volume and length, which allow us to capture more nuanced forms of collapse, including cases where the colon remains connected but is not fully inflated. To further account for fragmentation, we introduce a constant penalty when the mask contains more than one connected component. This reflects the added difficulty in reconstructing anatomically plausible geometry from disjointed segments.

Given the availability of high-quality segmentation masks for non-collapsed colons in $\mathcal{D}_{\text{Finocchiaro}}$, we frame collapse classification as a problem of determining whether a new observation is likely to fall within this well-characterized reference distribution. Specifically, our approach combines two complementary anomaly detection strategies:

(1) **Outlier detection:** A visual inspection of $\mathcal{D}_{\text{Finocchiaro}}$ reveals that some masks, though seemingly complete in 3D, contain internal gaps or "black holes" (see Figure 7). These distort both volume and length estimates and make such masks unreliable as ground truth for inflation. To address this, we first apply an *Isolation Forest* to detect outliers within the dataset. This step removes masks with anomalous volume-length profiles.
(2) **Novelty detection:** The remaining masks, exhibiting consistent and anatomically plausible geometry, are then used to train a *One-Class Support Vector Machine (One-Class SVM)* [56]. The One-Class SVM performs novelty detection across the $\mathcal{D}_{\text{SA-Labels}}$ dataset, which includes collapsed and potentially non-collapsed cases.

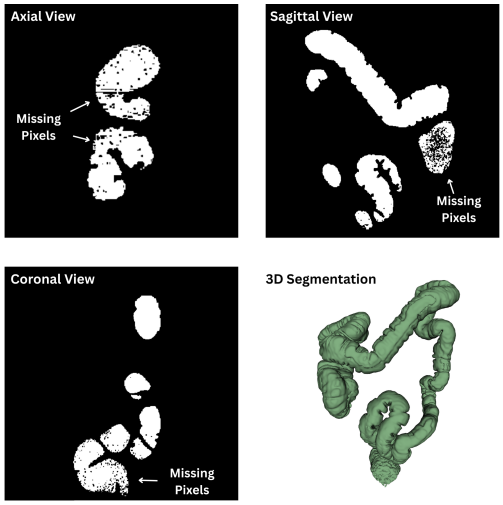

Figure 7: Examples of segmentation masks with internal inconsistencies ("black holes") caused by missing voxel values. These are visible in the axial, sagittal, and coronal slices, while the 3D rendering may still appear visually complete.

5.4 Colon Segmentation Model

For fully automatic colon segmentation, we adopt *nnU-Net* [57], a self-configuring deep learning framework designed for biomedical image segmentation. nnU-Net automatically adapts its network architecture, preprocessing pipeline, and training schedule based on the characteristics of the input dataset, eliminating the need for extensive manual tuning, which can introduce variability or lead to suboptimal design choices.

Our choice is motivated by nnU-Net's consistent performance across diverse medical segmentation tasks. It has been shown to match or outperform highly specialized methods on 23 public datasets, without requiring task-specific customization [57]. Notably, it has also demonstrated strong performance in segmenting non-collapsed colons [11], further supporting its suitability for our task. In addition to its accuracy, nnU-Net provides a reliable and reproducible baseline.

While nnU-Net handles many core preprocessing tasks, we perform additional preparatory steps to ensure anatomical consistency and compatibility across scans:

- **Alignment Normalization:** All CT images are anatomically aligned to a consistent orientation using the RAI (Right–Anterior–Inferior) coordinate system.
- **Spacing Normalization:** Voxel spacing is standardized to 1 mm in all directions, resulting in isotropic resolution across the dataset.

5.5 Registration-based Reconstruction

The main idea of our approach is to infer missing segments in a collapsed colon by aligning it with a similar non-collapsed colon. We treat the non-collapsed scan as the moving image and the collapsed scan as the fixed image, allowing the former to deform in order to approximate the missing geometry. This approach relies on the assumption that anatomically similar individuals have comparable colon morphology, and that the colon's flexibility allows for plausible deformation between collapsed and non-collapsed configurations. To account for subject-specific differences in colon shape, a non-rigid registration method is used. Even among anatomically similar individuals, the colon can vary significantly in its curvature, length, and folding pattern, making rigid and affine registration insufficient.

Specifically, we perform the registration on the raw CT scans rather than directly on the segmentation masks. This approach preserves anatomical integrity, as it aligns the entire abdominal cavity, not just the colon. Registering only the colon masks would lead to undesirable deformation, where the non-collapsed colon could be unnaturally eroded to match the collapsed form. By aligning the full CT volume, we ensure that the transformation respects the broader anatomical structure. After registration, the resulting deformation field is applied to the corresponding colon segmentation masks. The reconstruction pipeline is defined as follows:

(1) **Candidate Dataset Creation:** we use the segmentations from $\mathcal{D}_{\text{Finocchiaro}}$ and the corresponding CT scans from $\mathcal{D}_{\text{TCIA}}$ to form a pool of potential moving images.
(2) **Anatomical Alignment:** similar to the nnU-Net preprocessing, all scans are reoriented to the RAI coordinate system to ensure consistent spatial reference during registration.
(3) **Best Match Identification:** to select a suitable non-collapsed scan for each collapsed case, we compute MI between their corresponding CT volumes. MI is chosen because the data was acquired using different scanners, which can introduce variations in intensity distributions. This metric is robust to such differences and helps identify structurally similar scans. From the candidate dataset, the scan with the highest MI is selected as the best match and used as the moving image for registration.
(4) **B-Spline Registration on CT Volumes:** we perform non-rigid registration using a B-spline transformation. MI is again used as the similarity metric, which is part of the optimization objective, for its robustness to intensity variations across scans. Registration is carried out using a multi-resolution strategy, which improves alignment accuracy while maintaining computational efficiency [58]. The final deformation

field is applied to the colon mask of the moving subject to generate the reconstructed volume.

5.6 Evaluation

5.6.1 ***Semi-automatic Label Generation.*** Initial label quality is assessed by the authors, followed by expert visual inspection focusing on anatomical completeness and plausibility. No formal scoring system or inter-rater agreement metrics are recorded. It is important to note that this pipeline is not optimized for segmentation metrics. Instead, the method is designed to prioritize inclusion of colon regions, even at the cost of increased false positives. This approach ensures that most or all relevant colon structures are captured, which is especially important in cases involving collapsed or fluid-filled colons where certain regions may be difficult to reconstruct later.

5.6.2 ***Colon Segmentation Model.*** For evaluating model performance, we follow guidance from the Metrics Reloaded framework [59], which emphasizes complementing an *overlap-based* metric with a *boundary-based* metric to assess both volumetric agreement and boundary accuracy. We use the Dice Similarity Coefficient (DSC) to measure the overlap between predicted and ground truth volumes. In addition, we include the 95th percentile Hausdorff Distance (HD_{95}) and the Average Symmetric Surface Distance (ASSD) to quantify how well the predicted boundaries align with the ground truth. Although HD_{95} and ASSD are both boundary-based metrics and are expected to be strongly correlated, we report both to maintain consistency with existing literature and to provide a more complete picture of surface-level performance.

To complement quantitative evaluation metrics, we perform a component-wise analysis of false positive predictions. We extract connected components in the predicted segmentation that do not correspond to any annotated ground truth region. Each component is quantified by voxel count and degree of overlap with ground truth. Components larger than 10,000 voxels with less than 95% overlap are flagged for visual inspection and further uncertainty analysis.

5.6.3 ***Registration-based Reconstruction.*** As no ground truth reconstructions are available, evaluation is limited to qualitative visual assessment. Reconstructed colons are reviewed by the authors for continuity, anatomical plausibility, and the absence of major defects. The goal of this step is not to produce perfect or fully accurate reconstructions, but rather to explore the feasibility and potential of registration-based methods for recovering collapsed colon geometry.

6 Experimental Setup

6.1 Semi-Automatic Label Generation

We apply our semi-automatic label generation pipeline to the $\mathcal{D}_{\text{TCIA}}$ dataset obtained after initial data preparation. Due to time constraints, manual verification was performed on a stratified subset of the data. Specifically, we sample 300 subjects (501 scans) from the automatically labeled set, using stratification based on patient sex, age, and scanner model to ensure diversity. The sample size was chosen to approximately match the number of non-collapsed segmentations available in $\mathcal{D}_{\text{Finocchiaro}}$, while accounting for the fact that some segmentations produced by our method would likely be excluded during quality control.

6.2 Quantifying Colon Collapse

We use the *scikit-learn* implementations of both the Isolation Forest and the One-Class SVM to identify anomalous cases in $\mathcal{D}_{\text{Finocchiaro}}$ [60]. Since we lack prior knowledge about the proportion of anomalies in $\mathcal{D}_{\text{Finocchiaro}}$, we set the `contamination` parameter to `auto`. This allows the threshold to be determined according to the method described in the original paper by Liu et al. [61]. The One-Class SVM is configured with an radial basis function (RBF) kernel ({kernel='rbf'}), which is commonly used in anomaly detection tasks due to its ability to capture non-linear decision boundaries in high-dimensional spaces. Furthermore, we set `nu=0.05` and `gamma=0.1`. This configuration assumes a low proportion of anomalies in the dataset, following initial filtering with Isolation Forest. To account for uncertainty and the limitations of our feature set, we define the inlier threshold based on the median value of the training data used by the One-Class SVM. This conservative approach helps mitigate misclassification due to anatomical variability and the fact that our features may not fully capture all relevant aspects of colon collapse. Additionally, we introduce a penalty to masks consisting of more than one connected component, with a value of -5 (used as a practical substitute for $-\infty$) to ensure they are classified as collapsed.

6.3 Segmentation Model Configuration

We use the 3D full-resolution variant of nnU-Net and all hyperparameters and training settings follow nnU-Net's default configuration. To ensure robustness and improve generalization, the model is trained using five-fold cross-validation. Final predictions are obtained by averaging the voxelwise probabilities from all five models. This ensembling provides a more reliable estimate of how the model will generalize to unseen data and helps prevent overfitting. All models are trained over 1,000 epochs using nnU-Net's built-in training pipeline, which includes dynamic learning rate scheduling (starting at 0.01) and a combined DSC and cross-entropy loss function. Training curves for all five cross-validation fold, including loss, pseudo Dice, and epoch duration can be found in the Appendix (Figures 17-21). Training is conducted on a high-performance computing cluster due to GPU and memory requirements.

We construct a balanced training set using $\mathcal{D}_{\text{SA-Labels}}$ (varying degrees of colon collapse) and $\mathcal{D}_{\text{Finocchiaro}}$ (non-collapsed colons only). The splitting procedure is as follows:

- Before splitting, instances in $\mathcal{D}_{\text{SA-Labels}}$ that were classified as inliers (likely non-collapsed) by the outlier detection method are removed. The remaining dataset is then split into $\mathcal{D}_{\text{C-Train}}$ and $\mathcal{D}_{\text{C-Test}}$, using stratification to ensure an even distribution across degree of collapse, subject, and position. Additionally, instances in the test set that were marked by the expert as exhibiting missing dark fluid are removed.
- $\mathcal{D}_{\text{Finocchiaro}}$ is split into $\mathcal{D}_{\text{NC-Train}}$ and $\mathcal{D}_{\text{NC-Test}}$, also stratified, with the constraint that any subject included in $\mathcal{D}_{\text{C-Train}}$ must also be included in $\mathcal{D}_{\text{NC-Train}}$. This prevents data leakage.

Final sets are constructed by combining the respective splits, resulting in the datasets $\mathcal{D}_{\text{CNC-Train}}$ and $\mathcal{D}_{\text{CNC-Test}}$ (where CNC stands for Collapsed and Non-Collapsed). This setup ensures a representative mix of collapsed and non-collapsed colons across both training and test sets, reflecting the primary focus of this work on robust segmentation under varying inflation states, while also maintaining diversity in subject demographics.

6.4 Reconstruction Configuration

Non-rigid image registration is performed using a B-spline transformation implemented via the *SimpleITK* library [62]. The registration process uses Mattes Mutual Information with 50 histogram bins as the similarity metric, which is well-suited for multi-modal alignment. Optimization is carried out using the Limited-memory BFGS algorithm (L-BFGS-B), with a convergence tolerance of 1×10^{-5} and a maximum of 50 iterations. A multi-resolution strategy is applied with shrink factors of [4, 2, 1] and corresponding smoothing sigmas of [2, 1, 0], specified in physical units. The B-spline transform is initialized over the fixed image using a control point grid with uniform spacing of 4 in each spatial dimension. For fixed images, we utilize $\mathcal{D}_{\text{Expert-C}}$ and for the candidate moving images $\mathcal{D}_{\text{Finocchiaro}}$.

7 Results

7.1 Semi-Automatic Label Generation

A total of 501 CT scans were processed using our semi-automatic label generation pipeline. Figure 8 presents an example of a resulting colon segmentation. Visual inspection confirms that, in the majority of cases, both air-and fluid-filled regions are well segmented, with fine anatomical details such as haustral folds and fluid pockets clearly outlined. Compared to the TotalSegmentator output shown in Figure 5, the semi-automatic mask demonstrates improved anatomical accuracy and higher structural detail.

Following manual review, each scan was categorized based on segmentation quality:

- **103 scans** – Accepted without modification.
- **267 scans** – Accepted after manual correction.
- **131 scans** – Discarded due to poor segmentation quality.

Manual corrections typically involved removing small false positives such as parts of the small intestine, spine or lungs, which were occasionally included in the segmentation, especially in cases where these structures are very close to the colon boundaries. An additional 58 scans were flagged as missing fluid-filled regions based on informal expert feedback. Although these cases passed the general quality check, they were later excluded from evaluation due to incomplete segmentations, but retained for training. The accepted and corrected masks were combined to create the final labeled dataset, $\mathcal{D}_{\text{SA-Labels}}$, consisting of 370 high-quality colon segmentations used for training and evaluation of our U-Net model.

7.2 Quantifying Colon Collapse

An Isolation Forest was applied to $\mathcal{D}_{\text{Finocchiaro}}$ with the goal to remove segmentation masks with anomalous volume-length profiles. This reduced the dataset from 415 to 346 high-confidence, non-collapsed masks (see Figure 9). The resulting subset was then used to define a reference distribution.

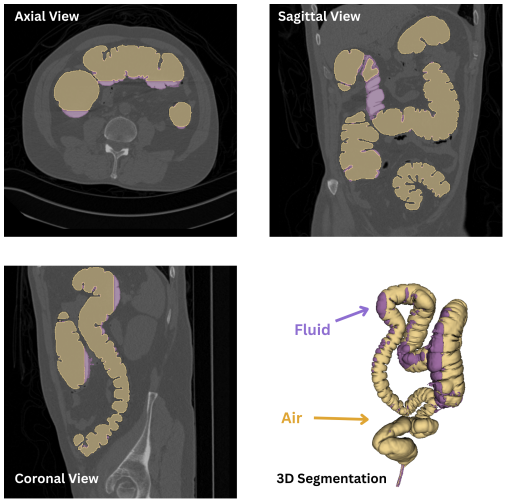

Figure 8: Example of colon segmentation produced by our semi-automatic pipeline. The air-filled region is shown in yellow, and the fluid-filled region is shown in purple.

A comparison with the $\mathcal{D}_{\text{SA-Labels}}$ highlights clear differences in volume and skeleton size distributions (see Appendix Figure 15). The filtered $\mathcal{D}_{\text{Finocchiaro}}$ exhibits a higher average colon volume of 1.87 L (standard deviation, SD = 0.40 L), whereas $\mathcal{D}_{\text{SA-Labels}}$ has a lower mean of 1.70 L with greater variability (SD = 0.66 L). Similarly, the average skeleton size in $\mathcal{D}_{\text{Finocchiaro}}$ is 29.41k voxels (SD = 26.33k), compared to 13.41k voxels (SD = 9.18k) in $\mathcal{D}_{\text{SA-Labels}}$.

Building on this reference distribution, a One-Class SVM was trained on the filtered non-collapsed masks and applied to $\mathcal{D}_{\text{SA-Labels}}$ for collapse classification. The resulting distribution of scan categories is summarized below and visually illustrated in Figure 10:

- **Inliers (likely non-collapsed):** 119 scans
- **Outliers (one component):** 102 scans
- **Outliers (multiple components):** 149 scans (of which 49 were flagged solely due to multi-component penalty)

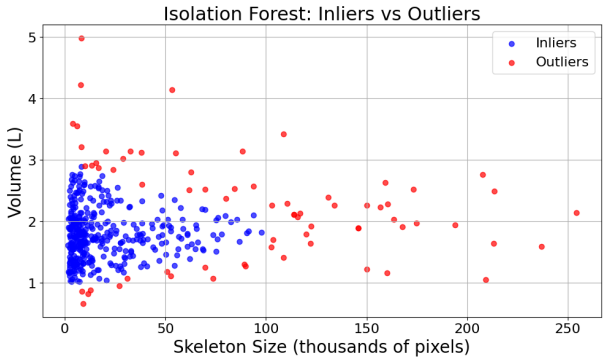

Figure 9: Isolation Forest classification of colon segmentations based on volume and skeleton size. Each point represents a scan from $\mathcal{D}_{\text{Finocchiaro}}$, with inliers shown in blue and outliers in red.

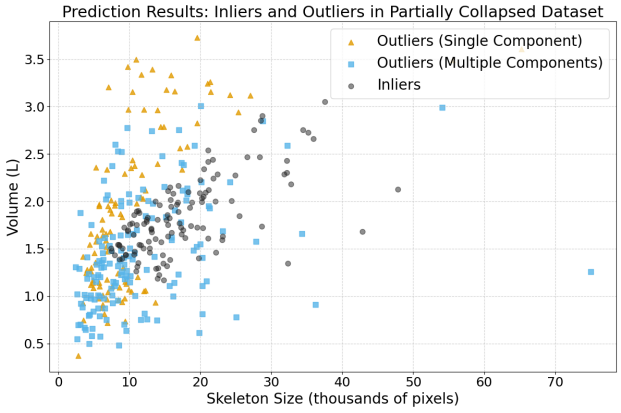

Figure 10: Classification of segmentations in the partially collapsed dataset ($\mathcal{D}_{\text{SA-Labels}}$) using a One-Class SVM trained on non-collapsed examples. Points are colored by prediction outcome: inliers (blue), outliers with a single component (yellow), and outliers with multiple components (gray). Outliers tend to have lower skeleton size or atypical volume–length combinations.

	Metric	All	Sex			Position	
			F	M	U	P	S
$\mathcal{D}_{\text{Expert-NC}}$	HD_{95}*	**1.339**	1.000	1.407	—	1.457	1.280
	ASSD*	**0.709**	0.542	0.742	—	0.738	0.694
	DSC	**0.953**	0.970	0.950	—	0.950	0.955
$\mathcal{D}_{\text{Expert-C}}$	HD_{95}*	**2.586**	1.261	5.678	—	4.508	1.305
	ASSD*	**0.771**	0.688	0.965	—	1.075	0.569
	DSC	**0.949**	0.955	0.934	—	0.940	0.954
$\mathcal{D}_{\text{CNC-Test}}$	HD_{95}*	**3.255**	4.060	2.536	2.522	3.598	2.873
	ASSD*	**0.586**	0.607	0.574	0.528	0.621	0.547
	DSC	**0.971**	0.972	0.971	0.968	0.969	0.973

Table 2: Segmentation performance metrics across three test sets, grouped by overall average (All), sex (F: female, M: male, U: unknown), and patient position (P: prone, S: supine). Reported metrics include the 95th percentile Hausdorff Distance (HD_{95}*), Average Symmetric Surface Distance (ASSD*), and Dice Similarity Coefficient (DSC). Metrics marked with * are reported in millimeters (mm). All values are rounded to three decimal places.

7.3 Colon Segmentation Model

We evaluate our colon segmentation model across three test sets: $\mathcal{D}_{\text{CNC-Test}}$, $\mathcal{D}_{\text{Expert-NC}}$, and $\mathcal{D}_{\text{Expert-C}}$, using both overlap-based (DSC) and boundary-based (HD_{95}, ASSD) metrics. The results are summarized in Table 2. A strong correlation between the boundary-based metrics is confirmed by our analysis (see Appendix 16), indicating they capture similar aspects of segmentation quality.

The model shows strong overall performance on $\mathcal{D}_{\text{CNC-Test}}$, with consistent DSC across sexes and scan positions. However, boundary metrics reveal increased surface error in female and prone position scans. This suggests that although volumetric overlap is similar, predictions in these subgroups may be less spatially precise, potentially due to greater anatomical variability, differences in colon collapse, or positional effects.

On $\mathcal{D}_{\text{Expert-NC}}$, which includes only fully inflated colons, the model achieves robust and consistent results, with minimal variation across sexes or scan orientations. This indicates reliability in segmenting well-distended colons. In contrast, segmentation performance on $\mathcal{D}_{\text{Expert-C}}$ declines, likely due to increased ambiguity associated with collapsed colons. Here, the boundary errors are more pronounced, particularly in male and prone cases. However, given the small size of both expert-annotated datasets and potential confounding between sex and position, these values should be interpreted with caution.

For qualitative assessment, we include visual comparisons between the model output and expert annotations for all instances in both $\mathcal{D}_{\text{Expert-NC}}$ (see Appendix 22,23) and $\mathcal{D}_{\text{Expert-C}}$ (see Appendix 24, 25).

To better understand the reduced performance in cases involving collapsed colons, we analyze segmented components across all test sets by size and their overlap with the ground truth (Table 3). While many of these components appear to be false positives, some may in fact represent valid colon segments that were missed by our semi-automatic or expert generated labels. We observe that the majority of low-overlap components are very small (fewer than 100 voxels), especially when compared to a non-collapsed colon, which typically consists of 1.5 to 2 million voxels. This suggests that most discrepancies are minor and likely result from oversegmenting low-contrast regions, such as those caused by collapse or fluid. As expected, the best performance is observed in $\mathcal{D}_{\text{Expert-NC}}$, which contains only non-collapsed colons. This dataset shows both fewer and smaller low-overlap components, indicating that the model is more confident and accurate when segmenting well-inflated colons. On the other side, $\mathcal{D}_{\text{Expert-C}}$ and $\mathcal{D}_{\text{CNC-Test}}$ show a higher number of components with lower overlap. This suggest that the model is more prone to uncertainty in the presence of collapsed structures, where anatomical features are less distinct and segmentation is inherently more challenging.

To assess the significance of these additional components, we conducted a targeted visual inspection of all cases containing components larger than 10,000 voxels with less than 95% overlap with the ground truth. These represent the most prominent and potentially consequential segmentation discrepancies. In $\mathcal{D}_{\text{Expert-C}}$, no components meeting this threshold were judged to be incorrect. The largest unmatched component in this set has a size of 4,680 voxels (see Case 4 in Figure 11). By contrast, in the combined test set $\mathcal{D}_{\text{CNC-Test}}$, we identified four such components across four scans that were confirmed to be genuine false positives. To further examine these cases, we analyzed voxel-wise probability scores associated with non-colonic segments. As illustrated in Figure 11, regions corresponding to false positives exhibit lower confidence compared to correctly segmented areas. This suggests that model uncertainty could serve as a valuable signal for filtering unreliable predictions in downstream tasks.

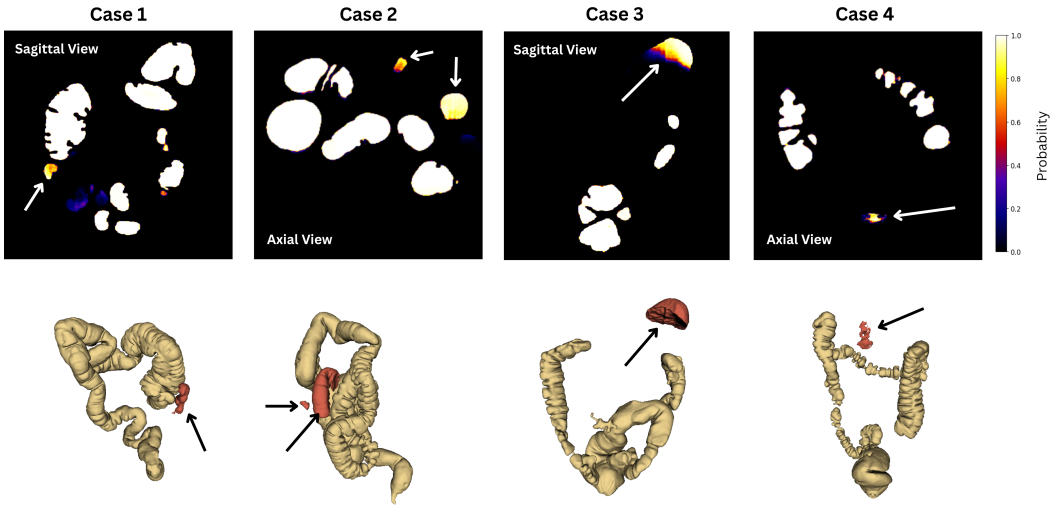

Figure 11: Four examples of false positives from the model's predictions on the test sets. The top row shows the model's voxel-wise probability maps, while the bottom row presents the 3D segmentations, with false positive components highlighted in red.

Finally, a qualitative comparison between model predictions and expert annotations (Figure 12) reveals several instances where the model accurately captures regions that are omitted in the ground truth. This is particularly evident in collapsed colons, where anatomical boundaries are often ambiguous. These observations underscore that some apparent "false positives" may not indicate true model errors, but rather limitations in human annotations.

7.4 Registration-based Reconstruction

Based on visual inspection of 10 registered image pairs (see Appendix Figure 26 and Figure 27), we found that all registered images retained the appearance of non-collapsed colons, suggesting that the deformation was controlled and did not result in excessive warping. In several cases (e.g., cases 1, 5, and 10 from Figure 26) the registered images closely matched the curvature of the corresponding fixed images, indicating successful alignment. In contrast, cases such as 2, 6, and 9 (Figure 27) showed only minor adjustments. Although moving images were selected based on MI, this criterion alone did not always ensure anatomical similarity, and in some instances, the mismatch in curvature between fixed and moving images likely limited the effectiveness of the registration.

8 Discussion

8.1 Data Preparation

The filtering steps applied were designed to enhance consistency, reduce variability, and improve the reliability of downstream tasks. These decisions promote standardized inputs and higher data quality, which is key when developing robust machine learning models. At the same time, we recognize that such filtering can introduce certain biases that may limit broader applicability.

For example, we included only prone and supine scans, the standard positions in CT colonography, to ensure consistent anatomical presentation. However, this may bias the dataset toward poorly insufflated cases, since lateral decubitus scans, shown to improve distension and reduce collapse [63], were excluded. Similarly, we restricted the dataset to patients aged 50-90 to align with the typical CRC screening population and to focus on the model's intended purpose. This excludes pediatric cases, which differ anatomically,

	Size (#Vox.)	Overlap (%)			
		<70	**70–89**	**90–95**	**>95**
$\mathcal{D}_{\text{Expert-NC}}$	<100	130 (11)	0	0	**131 (10)**
	100–999	**3 (2)**	0	0	0
	1,000–9,999	0	0	0	0
	≥10,000	0	0	0	**12 (12)**
$\mathcal{D}_{\text{Expert-C}}$	<100	**57 (10)**	0	0	23 (6)
	100–999	**9 (6)**	1 (1)	1 (1)	0
	1,000–9,999	3 (3)	3 (3)	0	**4 (4)**
	≥10,000	1 (1)	7 (3)	6 (5)	**19 (7)**
$\mathcal{D}_{\text{CNC-Test}}$	<100	**546 (110)**	41 (31)	21 (20)	251 (75)
	100–999	**27 (19)**	11 (9)	3 (3)	8 (8)
	1,000–9,999	13 (11)	1 (1)	1 (1)	**14 (9)**
	≥10,000	9 (8)	2 (2)	19 (18)	**161 (116)**

Table 3: Distribution of segmented components across three test sets, grouped by component size (in number of voxels) and overlap thresholds with ground truth labels (in percent). Each cell shows the number of components, with the number of affected images in parentheses. Component size is grouped into four bins: <100, 100–999, 1,000–9,999, and ≥ 10,000 voxels. Overlap with ground truth segmentations is grouped into four intervals: <70%, 70–89%, 90–95%, and >95%. This representation provides insight into the prevalence and characteristics of additional segmented regions not matching annotated structures. The highest value in each row (per component size bin) is highlighted in bold.

and patients over 90, who were treated as outliers. When multiple scan versions were available, we retained only version 1 to avoid duplication. While this likely removed some low-quality or incomplete scans, it may also have excluded higher-quality alternatives. Due to time constraints, we did not compare scan versions for quality, but future work could incorporate automated or manual quality assessment to refine this step.

8.2 Semi-Automatic Label Generation

We developed a semi-automatic pipeline to efficiently generate high-quality training labels, combining traditional segmentation methods with manual correction. Parameters were selected using information from the full dataset prior to defining the train/test split. While this could raise concerns about potential biases, the final split was stratified by age, sex, scanner model, and position, helping preserve key data characteristics across subsets. Importantly, no parameter tuning was performed to optimize quantitative metrics. As such, the selected parameters would likely have remained unchanged even if chosen after the split. Nonetheless, future work should address this to align more closely with best practices.

To detect fluid regions, the pipeline applies intensity thresholding followed by 3D region growing to improve connectivity at fluid-air interfaces (see Stage 2 in 5.2.3). This approach is effective across many cases but can be affected by inter-scan intensity variability. Fluid with attenuation below 100 HU is missed, while 3D region growing can occasionally leak into bone structures. To refine colon segmentation, we apply a four-check voting system, which relies on anatomical context derived from TotalSegmentator. While generally robust, inaccuracies in the TotalSegmentator mask can occasionally affect the final output. Still, the pipeline performs reliably across a wide range of cases and offers a strong foundation for label generation.

Manual post-processing was initially performed by two annotators, allowing for efficient review, but introducing potential variability in labeling style. Although inter-annotator consistency was not formally assessed, a third expert reviewed all masks to standardize the final outputs and reduce potential noise.

Overall, this method offers a practical and scalable trade-off. Full manual segmentation can take 5-10 minutes in well-distended cases and up to 45 minutes in collapsed ones, where even experts may miss subtle regions. In contrast, our pipeline produces initial segmentations automatically, with typical correction times in 3D Slicer reduced to just a few minutes.

8.3 Outlier detection

Prior work has typically defined collapsed colons based solely on the number of connected components. In contrast, our approach captures a broader range of collapse levels, including single-component colons that are not fully insufflated. As a first attempt, we use two simple and interpretable features: colon volume and skeleton length. While these features are effective in identifying more extreme cases, they may fail to detect subtler forms of collapse. As illustrated in Figure 10, inliers overlap with cases that are only excluded when multi-component segmentations are penalized. This indicates that single-component segmentations with similar volume and skeleton length may be misclassified. This could be due to anatomical variability and the fact that skeleton tends to overestimate true colon length due to branching. However, we take the uncertainty for single components into account by using a more conservative threshold for the inlier classification. Future work could incorporate additional features such as sex, weight or height that have an effect on colon anatomy.

8.4 Colon Segmentation Model

The performance and generalizability of our segmentation model are influenced by two main factors: the heterogeneity of label sources and the exclusion of certain scan types during training. The model was trained on a mix of segmentations from our semi-automatic pipeline and the $\mathcal{D}_{\text{Finocchiaro}}$ dataset. While both provide high-quality annotations, differences in segmentation techniques and labeling choices may introduce small inconsistencies. Future work could focus on harmonizing annotation styles.

As discussed earlier, scans containing low-attenuation ("dark") fluid were not excluded from the train data, but their fluid regions were not included in the segmentation labels (see Appendix 28 for an example). These regions had intensity values close to surrounding soft tissue, making them difficult to isolate using thresholding. Including them risked over-segmentation into adjacent structures, thus compromising the label quality. Nevertheless, we retained these scans in the training set under the assumption that the model could still benefit from their anatomical context. However, due to the lack of direct supervision for dark fluid, the model may struggle to such cases in unseen data. While our semi-automatic labels excluded them, $\mathcal{D}_{\text{Finocchiaro}}$ may include some examples of dark fluid, though this was not systematically verified. In future work, we plan to identify and annotate low-contrast fluid cases to improve model robustness.

A strength of our evaluation strategy is the use of both overlap-based and boundary-based metrics. While DSC captures overall volumetric agreement, it can hide boundary errors that may be clinically important. Metrics such as (HD_{95}) and ASSD address this by measuring surface alignment. For instance, although DSC scores were similar across male and female cases in $\mathcal{D}_{\text{CNC-Test}}$, ($HD_{95}$) was notably higher in females, suggesting lower boundary precision, potentially due to greater anatomical variability. Such observations highlight the importance of using boundary-based metrics in conjunction with overlap measures to obtain a more complete assessment of segmentation quality.

Quantitative metrics, while essential for benchmarking, also have limitations in complex tasks like collapsed colon segmentation. On our manually annotated collapsed-colon dataset, the model appeared to underperform according to standard metrics. Yet visual inspection revealed that in several cases, the model actually segmented correctly parts that were entirely omitted in expert annotations. This highlights the model's strength in recovering collapsed areas that are difficult even for experienced annotators. These findings emphasize the need to treat manual annotations as a practical gold standard, rather than an absolute ground truth.

8.5 Registration-based Reconstruction

One limitation of our current registration approach lies in the choice of B-spline transformation parameters. While strong regularization

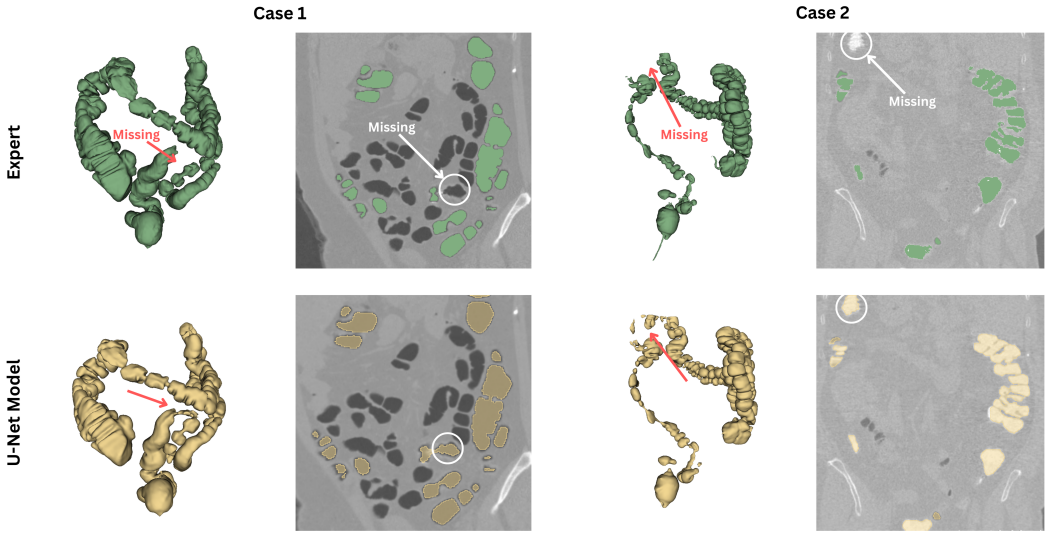

Figure 12: Qualitative comparison between model and expert annotations. Two examples of expert-segmented collapsed colons from $\mathcal{D}_{\text{Expert-C}}$ are shown in green, with corresponding nnU-Net predictions in yellow (bottom row). The figures highlight regions correctly segmented by the model but missing from the expert annotations.

helps prevent unrealistic deformations, it may also limit the flexibility needed to properly align colons with significant anatomical differences. However, relaxing this constraint risks over-warping the colon, potentially leading to anatomically implausible results.

A second challenge is the reliance on MI to select suitable moving images. While effective in some cases, MI does not guarantee anatomical similarity, and a poor initial match can degrade registration quality. To address this, we already leverage a relatively large pool of candidate images from $\mathcal{D}_{\text{Finocchiaro}}$, increasing the chances of identifying a reasonable starting point. However, expanding the pool of candidate images and incorporating anatomical landmarks or feature-based matching could further improve both selection and alignment.

Unlike more rigid organs such as the brain or lungs, the colon poses unique challenges for registration due to its high flexibility, inter-subject variability, and sensitivity to physiological changes. These factors limit the effectiveness of classical deformable registration techniques. These challenges suggest that classical registration techniques may be insufficient for colon shape reconstruction. As future work, learning-based methods should be explored to better handle the complex, non-linear variability of colon anatomy. Importantly, even when the registered image is not perfectly aligned with the fixed one, the resulting anatomy often remains plausible. Therefore, our registration method could potentially be used to generate synthetic data for training and augmentation.

8.6 Additional Considerations

Currently, our model uses only imaging data for segmentation. However, features such as patient sex could potentially enhance performance, especially given known anatomical differences in colon structure between males and females. Future work could explore multi-modal approaches that integrate both image features and auxiliary patient data to improve segmentation robustness across diverse populations.

In parallel with improving technical performance, it is increasingly important to consider the environmental footprint of machine learning workflows. To that end, we used CarbonTracker to monitor the energy usage of our training process [64]. Due to a high-performance computing system crash, emission data were recorded for only two of the five cross-validation folds, which consumed 12.37 kWh and 11.89 kWh, respectively. Assuming similar usage for the remaining folds, we estimate a total energy consumption of approximately 60 kWh, resulting in around 9.2 kg CO_2eq based on Denmark's 2023 average carbon intensity (151.65 gCO_2/kWh). This is roughly equivalent to 86 km of car travel or 8 CT scans (excluding standby energy consumption) [65].

9 Conclusion

In this work, we address the challenge of creating high-quality 3D colon segmentations suitable for constructing digital twins to support advanced robotic endoscopy navigation. Extracting accurate anatomical representations from CT scans is difficult due to limited annotated data and the frequent appearance of collapsed or fluid-filled colons. Prior work often excludes these challenging cases, reducing dataset diversity and limiting real-world applicability. To address this, we present a pipeline capable of handling both collapsed and non-collapsed colons. Our approach combines semi-automatic label generation, a robust U-Net segmentation model, a novel method for quantifying the degree of collapse, and a non-rigid registration approach for anatomical reconstruction.

In response to RQ-1, we demonstrate that a deep learning model trained on carefully curated labels can segment colons across varying insufflation states with high anatomical detail. The model even outperforms manual annotations in some collapsed cases, though

some false positives remain and could be reduced through thresholding or post-processing. For RQ-2, we show that non-rigid image registration can infer plausible colon geometries in collapsed or fragmented cases. While this produces continuous shapes, the anatomical fidelity is not perfect, and highly depended on the selection of the moving image. Nonetheless, registration appears promising for data augmentation, where perfect anatomical fidelity is less critical.

Future work should investigate the robustness of segmentation in dark, fluid-filled regions and explore improved collapse quantification using geometric or patient-specific features that have an effect on the colons anatomy. Reducing the number of false positives remains important, as these errors can negatively impact reconstruction quality. Additionally, developing a reliable ground truth dataset for collapsed colons is essential to enable consistent model evaluation. Reconstruction performance could benefit from neural network-based methods, particularly if paired training data is created through synthetic collapse simulation.

References

[1] Niclas Classen and Ioana-Daria Vasile. Colon digital twin. https://github.com/niclasclassen/colon-digital-twin, 2024. GitHub repository.

[2] World Health Organization. Colorectal cancer. https://www.who.int/news-room/fact-sheets/detail/colorectal-cancer, 2023. Accessed: 2025-02-17.

[3] American Cancer Society. Key statistics for colorectal cancer. https://www.cancer.org/cancer/types/colon-rectal-cancer/about/key-statistics.html, 2025. Accessed: 2025-02-17.

[4] World Cancer Research Fund International. Colorectal cancer statistics. https://www.wcrf.org/preventing-cancer/cancer-statistics/colorectal-cancer-statistics/, 2024. Accessed: 2025-04-24.

[5] ASGE Standards of Practice Committee, Douglas A. Fisher, John T. Maple, et al. Complications of colonoscopy. *Gastrointestinal Endoscopy*, 74(4):745–752, 2011.

[6] Richard M. Jones, Kelly J. Devers, Anton J. Kuzel, and Steven H. Woolf. Patient-reported barriers to colorectal cancer screening: A mixed-methods analysis. *American Journal of Preventive Medicine*, 38(5):508–516, 2010.

[7] Shengbing Zhao, Shuling Wang, Peng Pan, Tian Xia, Xin Chang, Xia Yang, Liliangzi Guo, Qianqian Meng, Fan Yang, Wei Qian, Zhichao Xu, Yuanqiong Wang, Zhijie Wang, Lun Gu, Rundong Wang, Fangzhou Jia, Jun Yao, Zhaoshen Li, and Yu Bai. Magnitude, risk factors, and factors associated with adenoma miss rate of tandem colonoscopy: A systematic review and meta-analysis. *Gastroenterology*, 156(6):1661–1674.e11, 2019.

[8] N. D. Pilonis, P. Spychalski, M. Kalager, et al. Adenoma detection rates by physicians and subsequent colorectal cancer risk. *JAMA*, 333(5):400–407, 2025.

[9] Horizon Europe 2023 IRE Consortium. Intelligent robotic endoscopes (ire) for improved healthcare services. https://ire4health.eu, 2025. Accessed: 2025-04-24.

[10] Jorge Corral-Acero, Francesco Margara, Maciej Marciniak, et al. The 'digital twin' to enable the vision of precision cardiology. *Nature Reviews Cardiology*, 17:543–554, 2020.

[11] Martina Finocchiaro, Ronja Stern, Abraham George Smith, Jens Petersen, Kenny Erleben, and Melanie Ganz. Hqcolon: A hybrid interactive machine learning pipeline for high quality colon labeling and segmentation, 2025.

[12] Alberto Bert, Ivan Dmitriev, Silvano Agliozzo, Natalia Pietrosemoli, Mark Mandelkern, Teresa Gallo, and Daniele Regge. An automatic method for colon segmentation in ct colonography. *Computerized Medical Imaging and Graphics*, 33(4):325–331, 2009.

[13] Jakob Wasserthal, Hanns-Christian Breit, Manfred T. Meyer, Maurice Pradella, Daniel Hinck, Alexander W. Sauter, Tobias Heye, Daniel T. Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5), September 2023.

[14] A. E. Omole, P. Mandiga, P. Kahai, and et al. *Anatomy, Abdomen and Pelvis: Large Intestine*. StatPearls Publishing, Treasure Island (FL), 2025. [Updated 2025 Apr 6].

[15] Vishy Mahadevan. Anatomy of the caecum, appendix and colon. *Surgery (Oxford)*, 35(3):115–120, 2017.

[16] R. Bowcutt, R. Forman, M. Glymenaki, S. R. Carding, K. J. Else, and S. M. Cruickshank. Heterogeneity across the murine small and large intestine. *World Journal of Gastroenterology*, 20(41):15216–15232, 2014.

[17] Koichi Utano, Koichiro Nagata, Takashi Honda, Tetsuya Kato, Alan K. Lefor, and Kaori Togashi. Bowel habits and gender correlate with colon length measured by ct colonography. *Japanese Journal of Radiology*, 40(3):298–307, Mar 2022. Epub 2021 Oct 11.

[18] Brian P. Saunders, Manabu Fukumoto, Steve Halligan, Craig Jobling, Mohammed E. Moussa, Clive I. Bartram, and Christopher B. Williams. Why is colonoscopy more difficult in women? *Gastrointestinal Endoscopy*, 43(2, Part 1):124–126, 1996.

[19] Lee M. Bass and Barry K. Wershil. Anatomy, histology, embryology, and developmental anomalies of the small and large intestine. In Mark Feldman, Lawrence S. Friedman, and Lawrence J. Brandt, editors, *Sleisenger and Fordtran's Gastrointestinal and Liver Disease: Pathophysiology, Diagnosis, Management*, chapter 98. Elsevier, 11th edition, 2020.

[20] Noam Shussman and Steven D Wexner. Colorectal polyps and polyposis syndromes. *World journal of gastrointestinal surgery*, 6(3):91–100, 2014.

[21] Mohammad S. Hossain, Hany Karuniawati, Abdelrahman A. Jairoun, Zinat Urbi, Darren J. Ooi, Arockia John, Yee Chan Lim, Kazi M. Khalid Kibria, A. K. M. Mohiuddin, Long Chiau Ming, et al. Colorectal cancer: A review of carcinogenesis, global epidemiology, current challenges, risk factors, preventive and treatment strategies. *Cancers*, 14(7):1732, 2022.

[22] Tomasz Sawicki, Marta Ruszkowska, Andrzej Danielewicz, Elżbieta Niedźwiedzka, Tomasz Arłukowicz, and Katarzyna E. Przybyłowicz. A review of colorectal cancer in terms of epidemiology, risk factors, development, symptoms and diagnosis. *Cancers*, 13(9):2025, 2021.

[23] American Cancer Society. Survival rates for colorectal cancer. https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/survival-rates.html, 2025. Accessed May 2025.

[24] Jerome D. Waye, Douglas K. Rex, and Christopher B. Williams. *Colonoscopy: principles and practice*. Wiley-Blackwell, Chichester, 2nd ed edition, 2009.

[25] National Institute of Biomedical Imaging and Bioengineering. Computed tomography (ct). https://www.nibib.nih.gov/science-education/science-topics/computed-tomography-ct. Accessed: 2025-03-28.

[26] Atam P. Dhawan. *Medical Imaging Modalities: X-Ray Imaging*, chapter 4, pages 79–97. John Wiley & Sons, Ltd, 2011.

[27] Atam P. Dhawan. *Image Formation*, chapter 2, pages 23–63. John Wiley & Sons, Ltd, 2011.

[28] C. L. Wyatt, Y. Ge, and D. J. Vining. Automatic segmentation of the colon for virtual colonoscopy. *Computerized Medical Imaging and Graphics*, 24(1):1–9, 2000.

[29] Neeraj Sharma and Lalit M. Aggarwal. Automated medical image segmentation techniques. *Journal of Medical Physics*, 35(1):3–14, 2010.

[30] Atam P. Dhawan. *Medical image analysis*. IEEE Press series on biomedical engineering. Wiley-Blackwell, Oxford, 2nd ed edition, 2011.

[31] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen AWM van der Laak, Bram van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. 2015.

[33] M. H. Asnawi, A. A. Pravitasari, G. Darmawan, T. Hendrawati, I. N. Yulita, J. Suprijadi, and F. A. L. Nugraha. Lung and infection ct-scan-based segmentation with 3d unet architecture and its modification. *Healthcare*, 11(2):213, 2023.

[34] Changjian Yu, Chima P. Anakwenze, Yuxin Zhao, and et al. Multi-organ segmentation of abdominal structures from non-contrast and contrast enhanced ct images. *Scientific Reports*, 12:19093, 2022.

[35] Xiangyi Yan, Hao Tang, Shanlin Sun, Haoyu Ma, Deying Kong, and Xiaohui Xie. After-unet: Axial fusion transformer unet for medical image segmentation, 2021.

[36] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3), July 2009.

[37] Marco A.F. Pimentel, David A. Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, June 2014.

[38] Barbara Zitová and Jan Flusser. Image registration methods: A survey. *Image and Vision Computing*, 21(11):977–1000, 2003.

[39] Francisco P. M. Oliveira and João Manuel R. S. Tavares. Medical image registration: A review. *Computer Methods in Biomechanics and Biomedical Engineering*, 17(2):73–93, 2014.

[40] A. Ardeshir Goshtasby. *Image Registration*. Advances in Pattern Recognition. Springer London, London, 2012.

[41] Noor Shaik, Danial Moulavi, and Reza Samavi. Deep learning in medical imaging: Current trends, issues, and challenges. *ACM Computing Surveys*, 55(9):1–38, 2023.

[42] Tribikram Dhar, Nilanjan Dey, Surekha Borra, and R. Simon Sherratt. Challenges of deep learning in medical image analysis—improving explainability and trust. *IEEE Transactions on Technology and Society*, 4(1):68–75, 2023.

[43] S.K. Zhou, H. Greenspan, and D. Shen. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Medical Image Analysis*, 42:60–88, 2017.

[44] Yoshitaka Masutani, Hiroyuki Yoshida, Peter M. MacEneaney, and Abraham H. Dachman. Automated segmentation of colonic walls for computerized detection of polyps in ct colonography. *Journal of Computer Assisted Tomography*, 25(4):629–638, July 2001.

[45] Janne Näppi, Abraham H. Dachman, Peter MacEneaney, and Hiroyuki Yoshida. Automated knowledge-guided segmentation of colonic walls for computerized detection of polyps in ct colonography. *Journal of Computer Assisted Tomography*, 26(4):493–504, July 2002.

[46] D. Chen, M.R. Wax, L. Li, Z. Liang, B. Li, and A.E. Kaufman. A novel approach to extract colon lumen from ct images for virtual colonoscopy. *IEEE Transactions on Medical Imaging*, 19(12):1220–1226, 2000.

[47] Tarik A. Chowdhury and Paul F. Whelan. A fast and accurate method for automatic segmentation of colons at ct colonography based on colon geometrical features. In *Proceedings of the 15th Irish Machine Vision and Image Processing Conference (IMVIP)*, pages 94–100, 2011.

[48] Xiaoyun Yang, Xujiong Ye, and Greg Slabaugh. Multilabel region classification and semantic linking for colon segmentation in ct colonography. *IEEE Transactions on Medical Imaging*, 34(6):1236–1247, 2015. Senior Member, IEEE.

[49] Lin Lu and Jun Zhao. An automatic method for colon segmentation in virtual colonoscopy. In *2011 4th International Conference on Biomedical Engineering and Informatics (BMEI)*, volume 1, pages 105–108, 2011.

[50] K. Smith, K. Clark, W. Bennett, T. Nolan, J. Kirby, M. Wolfsberger, J. Moulton, B. Vendt, and J. Freymann. Data from ct colonography. The Cancer Imaging Archive, 2015.

[51] A. G. Smith, E. Han, J. Petersen, N. A. F. Olsen, C. Giese, M. Athmann, D. B. Dresbøll, and K. Thorup-Kristensen. Rootpainter: deep learning segmentation of biological images with corrective annotation. *New Phytologist*, 236(2):774–791, October 2022. Epub 2022 Aug 10.

[52] National Cancer Institute (NCI). The national ct colonography trial: Multicenter assessment of accuracy for detection of large adenomas and cancers in a healthy screening population. https://clinicaltrials.gov/study/NCT00084929, 2010. ClinicalTrials.gov Identifier: NCT00084929.

[53] Tarik A. Chowdhury, Paul F. Whelan, and Ovidiu Ghita. A method for automatic segmentation of collapsed colons at ct colonography. In *Proceedings of the Indian International Conference on Artificial Intelligence*, 2005.

[54] scikit-image contributors. Skeletonize example. https://scikit-image.org/docs/stable/auto_examples/edges/plot_skeleton.html, 2025. Accessed: 2025-04-28.

[55] Andriy Fedorov, Reinhard Beichel, Jayashree Kalpathy-Cramer, Julien Finet, Jean-Christophe Fillion-Robin, Sonia Pujol, Christian Bauer, David Jennings, Fiona Fennessy, Milan Sonka, John Buatti, Stephen R. Aylward, James V. Miller, Steve Pieper, and Ron Kikinis. 3d slicer as an image computing platform for the quantitative imaging network. *Magnetic Resonance Imaging*, 30(9):1323–1341, 2012.

[56] Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

[57] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, February 2021.

[58] Kadavigere V. Rajagopal Siddeshappa Nandish, Gopalakrishna Prabhu. Multiresolution image registration for multimodal brain images and fusion for better neurosurgical planning. *Biomedical Journal*, 40(6):329–338, 2017.

[59] Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D. Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, Mauricio Reyes, Michael A. Riegler, Manuel Wiesenfarth, A. Emre Kavur, Carole H. Sudre, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, Tim Rädsch, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, Arriel Benis, Matthew B. Blaschko, M. Jorge Cardoso, Veronika Cheplygina, Beth A. Cimini, Gary S. Collins, Keyvan Farahani, Luciana Ferrer, Adrian Galdran, Bram Van Ginneken, Robert Haase, Daniel A. Hashimoto, Michael M. Hoffman, Merel Huisman, Pierre Jannin, Charles E. Kahn, Dagmar Kainmueller, Bernhard Kainz, Alexandros Karargyris, Alan Karthikesalingam, Florian Kofler, Annette Kopp-Schneider, Anna Kreshuk, Tahsin Kurc, Bennett A. Landman, Geert Litjens, Amin Madani, Klaus Maier-Hein, Anne L. Martel, Peter Mattson, Erik Meijering, Bjoern Menze, Karel G. M. Moons, Henning Müller, Brennan Nichyporuk, Felix Nickel, Jens Petersen, Nasir Rajpoot, Nicola Rieke, Julio Saez-Rodriguez, Clara I. Sánchez, Shravya Shetty, Maarten Van Smeden, Ronald M. Summers, Abdel A. Taha, Aleksei Tiulpin, Sotirios A. Tsaftaris, Ben Van Calster, Gaël Varoquaux, and Paul F. Jäger. Metrics reloaded: recommendations for image analysis validation. *Nature Methods*, 21(2):195–212, February 2024.

[60] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[61] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422, 2008.

[62] Z. Yaniv, B. C. Lowekamp, H. J. Johnson, and R. Beare. Simpleitk image-analysis notebooks: A collaborative environment for education and reproducible research. *Journal of Digital Imaging*, 31(3):290–303, 2018.

[63] Perry J. Pickhardt, Joshua Bakke, Jarret Kuo, Jessica B. Robbins, Meghan G. Lubner, Alejandro Muñoz del Rio, and David H. Kim. Volumetric analysis of colonic distention according to patient position at ct colonography: Diagnostic value of the right lateral decubitus series. *AJR. American Journal of Roentgenology*, 203(6):W623–W628, 2014.

[64] Lasse F. Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models, 2020.

[65] Scott McAlister, Forbes McGain, Matilde Breth-Petersen, David Story, Kate Charlesworth, Glenn Ison, and Alexandra Barratt. The carbon footprint of hospital diagnostic imaging in Australia. *The Lancet Regional Health – Western Pacific*, 24, July 2022. Publisher: Elsevier.