# MASTER THESIS

# Predicting complications and grading the difficulty of total mesorectal excision surgery using machine learning

Amersfoort, Meander Medisch Centrum, Surgery
Enschede, University of Twente, The Netherlands

**Author**
Sander Barendsen

**Chairman & Clinical Supervisor**
Prof. dr. I.A.M.J. Broeders

**Technological Supervisor**
Dr. C.O. Tan

**Day-to-Day Supervisor**
MSc S.C. Baltus

**Process Supervisor**
Dr. M. Groenier

**Date**
17-01-2025

meander MEDISCH CENTRUM

UNIVERSITY OF TWENTE.

# Contents

# Acknowledgments

I would like to express my heartfelt gratitude to everyone who has supported me throughout the development of this thesis.

Foremost, my supervisors at Meander Medical Center provided invaluable guidance, constructive feedback, and unwavering support during my research. Their expertise and encouragement were essential to the success of this project.

The technical supervisors at the University of Twente also deserve special recognition for their insights and mentorship, which significantly shaped my understanding of machine learning and its application in healthcare.

A special thanks goes to my family, friends, and partner for their continuous support and encouragement throughout this journey. Their belief in my abilities served as a constant source of motivation.

Lastly, gratitude is extended to the patients and clinicians who contributed to the data collection process, making this research possible.

# Introduction

Colorectal cancer is the second most common cancer in humans and the second leading cause of cancer-related deaths in the United States [1, 2]. Rectal cancer accounts for around 25% of colorectal cancer cases [3]. Diagnosis of rectal cancer is initially made through a digital rectal exam. To confirm rectal cancer, an endoscopy is performed to obtain a tissue biopsy and measure the distance from the lesion to the anal verge, defining rectal cancer as a tumor located less than 15cm from the anal verge[4]. Once cancer is pathologically established, Magnetic Resonance Imaging (MRI) or transrectal ultrasound is utilized to determine local tumor extension and nodal status. Additionally, computed tomography (CT) of the chest and abdomen is performed to detect metastases in the lungs or liver [5]. Surgical intervention utilizing the technique of Total Mesorectal Excision (TME) is widely recognized as the gold standard for treating locally advanced rectal carcinoma.

## 1.1 Total Mesorectal Excision

Total Mesorectal Excision is considered the gold standard technique for rectal cancer excision, as it has shown the lowest recurrence rates [6, 7]. The outcome improves further when surgical treatment is combined with preoperative chemo-radiotherapy. During TME, the pelvic autonomic nerves are identified and preserved, reducing the likelihood of sexual or bladder dysfunction [7]. TME can be performed using laparoscopic, open, robot-assisted, or transanal techniques [8].

TME involves the complete removal of the rectum and the surrounding mesorectum, as well as the pararectal lymph nodes, which are typically the initial site of metastases, as illustrated in Figure 5 [3]. The surgery is challenging due to the narrow and deep constraints of the pelvic cavity, resulting in a small workspace and limited vision, particularly when using rigid laparoscopic tools. Precision is critical during the resection of the mesorectum to ensure complete tumor removal. Despite its importance, there is currently no clear, objective definition of a "difficult pelvis"[9].

Other factors influencing the difficulty of TME surgery include tumor distance from the anal verge, tumor size, prior abdominal surgeries, neoadjuvant radiotherapy, male gender, and body mass index (BMI) [10].

## 1.2 Complications Associated with TME Surgery

TME surgery has a high rate of postoperative complications, reported to range between 36% and 59% [11]. Common complications include problematic anorectal dysfunction and dehiscence (re-opening) of the anastomosis. Anastomotic leakage (AL), is a complication in which a surgical anastomosis fails and intestinal contents leak into the body, is reported in 4% to 20% of cases [3, 12]. This can lead to further postoperative complications, often necessitating re-operation, which delays recovery, prolongs hospital stays, and increases mortality rates [13].

## 1.3 Research Objective

TME is a vital procedure for rectal surgery, yet it is technically challenging and associated with a high complication rate. Currently, there is no standardized method for surgeons to

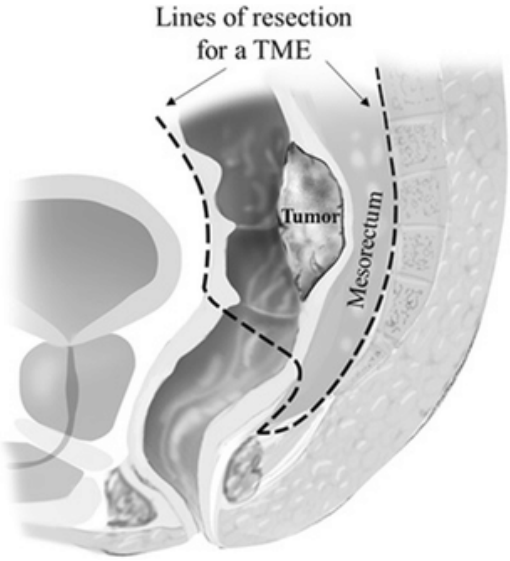



**Figure 2:** Illustration of the resection lines performed during TME to ensure complete tumor excision and address potential metastasis spread. The image highlights the tumor within the rectum. The dotted lines showcase the resection boundaries.

gauge the potential difficulty of TME surgery. Although the literature suggests that certain clinical and pelvimetry factors impact surgical outcomes, no objective definition exists for what constitutes a "difficult pelvis" or which combination of variables contributes to complications. Pelvic measurements, such as the pelvic inlet, pelvic outlet, or the length of the sacral curve, define the dimensions of the available working space for the surgeon during TME surgery. These pelvimetry measurements, along with other factors, could play a significant role in predicting complications and surgical difficulty.

This thesis aims to address this knowledge gap by developing a machine-learning model that integrates clinical and pelvimetry parameters to predict the complexity and complications of TME surgeries. Such a model could assist clinicians during outpatient clinic consultations by:

1. Providing patients with more accurate information regarding surgical outcomes and risks.
2. Estimating the duration of the surgery, improving the efficiency of the operation clinic roster.
3. Determining the level of surgical expertise required for each patient.

The research seeks to answer the following question:

*"Is it possible to create a machine learning model that accurately predicts complications and the duration of surgery using preoperative clinical and pelvimetry metrics of patients undergoing Total Mesorectal Excision surgery?"*

This study aims to provide a data-driven approach for surgical decision-making in TME surgery, combining clinical and pelvimetry parameters to deliver patient-tailored care.

## 2.1 Incidence and Prevalence

Colorectal cancer (CRC) is the third most common cancer worldwide, accounting for approximately 10% of all cancer cases. More than half of these cases occur in developed countries [14]. Furthermore, CRC is one of the leading causes of mortality in Western countries. Rectal cancer accounts for around 25% of all CRC cases [3]. In 2014, it was reported that each year approximately 13,000 patients are diagnosed with colorectal cancer in the Netherlands, and around 5,000 patients die from the disease annually [15]. A study examining data on colorectal cancer incidence from 1988 to 2007 showed a global increase in CRC among individuals under the age of 50, largely driven by a prominent rise in rectal cancer in many countries [16]. Despite this rising incidence, the 5-year survival rate for CRC has improved to 63%, and for rectal cancer, it has reached up to 67%. These survival rates vary based on the stage at diagnosis: localized (no spread outside the colon or rectum), regional (spread to nearby structures or lymph nodes), and distant (spread to distant lymph nodes or organs such as the liver or lungs) [17].

## 2.2 Symptoms and Risk Factors of colorectal cancer

There are several symptoms that are associated with colorectal cancer [18]:

- A change in bowel habits, such as constipation, diarrhea, or a change in stool consistency
- Rectal bleeding or blood in the stool
- A feeling that the bowel does not empty completely
- Fatigue or weakness
- Unexplained weight loss

There are several risk factors that increase the chance of CRC. These factors are displayed in Table 1.

| Category | Risk Factors Description |
|---|---|
| **Non-Modifiable Risk Factors** | • Age: Particularly after 50<br>• Family history: Increased risk with a family history of CRC<br>• Personal medical history: Inflammatory bowel diseases or previous radiation therapy<br>• Male gender: Higher risk in males |
| **Modifiable Risk Factors** | • Diet: High fats, low fiber, fruits, and vegetables, with red meat<br>• Alcohol consumption: Heavy drinking increases CRC risk [19]<br>• Obesity: High BMI<br>• Low physical activity: Sedentary lifestyle increases CRC risk<br>• Smoking: Smokers are 2.17 times more likely to develop CRC |

**Table 1:** Non-modifiable and modifiable risk factors for colorectal cancer [20]

## 2.3 Diagnosis and Staging of Rectal Cancer

Diagnosis of rectal cancer typically starts with a digital rectal exam, which allows the physician to palpate any abnormalities in the rectal area. This is often followed by an endoscopy, usually a flexible sigmoidoscopy, to detect colonic neoplasms in the rectum. A tissue biopsy is then performed to confirm rectal cancer and measure the distance from the lesion to the anal verge [21]. Rectal cancer is classified as a tumor is located withing 15 cm of the anal verge [22]. Once cancer is pathologically confirmed, sagital, axial and coronal T2-MRI scans are used to determine the depth of tumor invasion, lymph node involvement and circumferential resection margin [23]. Transrectal ultrasound can also be utilized to determine the depth of invasion and the absence of metastatic lymph nodes preoperatively [24]. Computed tomography of the chest and abdomen is performed to detect distant metastases in the lungs or liver [5]. While current diagnostic tools effectively stage rectal cancer, they do not assess the potential difficulty of TME surgery. This underscores the need for predictive tools incorporating both clinical and pelvimetry variables.

## 2.4 Advancements in Rectal Cancer Treatment

Rectal cancer treatment has evolved in the last few years, leading to a substantial reduction in mortality after surgery. Minimally invasive resection combined with neo-adjuvant chemotherapy has decreased mortality and improved patient outcomes. There are several techniques that are available for the treatment of rectal cancer. Among these advancements is Total Mesorectal Excision, which is considered the gold standard.

### Total Mesorectal Excision (TME): The Gold Standard

Total Mesorectal Excision is considered the gold standard due to its association with significantly lower recurrence rates [6, 7]. Recurrence rates dropped from 20.8% in patients undergoing conventional surgery to 5.9% in those treated with TME [25]. During TME, the pelvic autonomic nerves are identified and preserved, reducing the risk of sexual or bladder dysfunction. Additionally, TME helps preserve anal sphincter function, minimizing the need for a permanent stoma [7]. TME can be performed using laparoscopic, open, robot-assisted, or transanal techniques [8]. The procedure involves the complete removal of the rectum, including the surrounding mesorectum and pararectal lymph nodes, which are common sites of metastasis. This is performed along the visceral pelvic fascia, often referred to as the "holy plane", as it is crucial for ensuring complete tumor clearance [3].

### Low Anterior Resection

TME is applied in Low Anterior Resection. Due to advancements in the understanding of rectal cancer and improvements in surgical techniques, it was discovered that the traditional distal resection margin of 5 cm yielded similar patient survival and recurrence rates when compared to smaller distal margins. This made it increasingly feasible to preserve sphincter function during surgery [26, 27]. Furthermore, due to advances in stapling devices the ability to create a safe anastomosis at the distal rectum or the anal canal has been made possible, this is often performed using a circular stapler [28]. However, anastomotic leakage still a remains a common complication, with a higher chance of occurrence the more distal the anastomosis is. Failure of the anastomosis varies between 6 to 30% depending on other risk factors[29].

### Abdominoperineal Resection

Abdominoperineal resection (APR) is a procedure predominantly used to treat low-lying rectal cancer where the tumor is located close to the anal rectal junction, and sphincter preservation is not feasible. Like in other rectal surgeries, TME is a fundamental part of APR. It involves removing the sigmoid colon, rectum, and anus, resulting in a permanent colostomy. This radical approach ensures complete tumor removal and reduces the chances of residual cancer cells[30, 31].

### Role of Total Neo-adjuvant Therapy in Treatment of Rectal Cancer

Neo-adjuvant therapy is the standard of care and an important component of rectal cancer treatment. It has three primary goals: down-staging the tumor, eradicating distant micrometastases and preserving the rectum [32]. Traditionally, neoadjuvant therapy includes chemoradiotherapy, followed by Total Mesorectal Excision (TME) and systemic chemotherapy. Total neoadjuvant therapy combines (chemo)radiotherapy and chemotherapy before surgery to improve treatment outcomes and reduce recurrence.

## 2.5 Factors impacting TME surgical difficulty

Total Mesorectal Excision (TME) is a crucial yet challenging surgical procedure influenced by numerous clinical variables [10]. Another component contributing to surgical difficulty is pelvic anatomy. The narrow and constrained anatomy of the pelvis significantly limits the surgeon's working space, especially when using rigid instruments during laparoscopic surgery [9]. Research has shown that pelvimetry can be used to assess the surgical difficulty of rectal cancer. Table 2, displays an overview of key risk factors affecting the surgical difficulty in TME.

| Risk Factor | Description |
|---|---|
| **Tumor Location** | • Anterior tumors are more likely to have positive resection margins due to advanced staging.<br>• Tumors near the Anal Rectal Junction (ARJ) are more challenging to operate on. |
| **Tumor Size** | • Larger tumors increase surgical complexity and operation time.<br>• Tumor size raises the risk of positive circumferential resection margins. |
| **Previous Abdominal Surgery** | • Scar tissue from prior surgeries may lead to longer operation times, a higher risk of postoperative complications, and conversion to open surgery. |
| **Additional Risk Factors** | • Male gender and higher BMI increase surgical difficulty. |

**Table 2:** Factors influencing surgical difficulty during TME [10].

## 2.6 Pelvimetry in predicting surgical difficulty

In their systematic review, Hong et al. examined the role of MRI pelvimetry in predicting technical difficulty and outcomes of open and minimally invasive total mesorectal excision. Results of 11 studies suggested that a smaller intertubercular distance, interspinous distance, pelvic inlet, and a larger pubic tubercle height are correlated with increased surgical difficulty. Surgical difficulty was mentioned as a positive circumferential resection margin, incomplete TME, longer operative time, higher blood loss or anastomotic leak, conversion to open surgery, postoperative complications, or longer pelvic dissection time [9].

Raheem et al. performed a systematic review that aimed to evaluate the usefulness of pelvimetry data in assessing the surgical difficulty of rectal cancer and its correlation with complication outcomes. It researched the interspinous distance (IS), intertuberous distance (IT), mesorectal fat area (MFA), and pelvic inlet (PI). This review concludes that there is no standard definition of a difficult pelvis, therefore complicating the comparison of studies. Studies in this review mentioned that a smaller IS, narrow PI, and large MFA were indicators for a longer operation time and a higher chance of an anastomotic leakage [33].

Yamamoto et al. aimed to measure anatomical variables on MRI and analyze their predictive value in estimating the surgical difficulty of rectal surgery. Patients who underwent an LAR were included, with a distance to the ARJ of less than 10 cm. Eight measurements were included: one angle, the anorectal angle, and seven distances. These distances were pelvic inlet, pubococcygeal distance, sacral depth, pelvic length, pelvic outlet, intertuberous distance, and interspinous distance. This study identified four predictors that could signal a difficult surgery: BMI of 25+, tumor size above 45 mm, anorectal angle of 123+ degrees, and pelvic outlet <82.7 mm [34].

Chau et al. examined whether pelvic dimensions on preoperative MRI can predict poor-quality resections in laparoscopic low anterior resection (LAR) for rectal cancer. This study involved 92 patients with tumors within 10 cm from the anal verge and utilized pelvic measurements, such as the S1-S5-pubic symphysis angle. The findings showed that an S1-S5 angle above 74.3° was a significant predictor of poor-quality resections. These findings emphasize the potential of MRI pelvimetry in identifying patients who may require extra care or alternative surgical methods for improved outcomes [35].

These studies each show different results of what pelvimetry measurements have predictive value, again displaying that a difficult pelvis is hard to standardize. In this study, the definition of a difficult pelvis will be based on complication outcomes, as this is easily measurable and has a significant impact on the patient's quality of life.

The table below provides an overview of the findings from previous research, illustrating the complexity and diversity in identifying key anatomical predictors.

| Study | Significant Parameters | Surgical Outcomes |
|---|---|---|
| Hong et al. | • Smaller intertubercular distance<br>• Smaller interspinous distance<br>• Smaller pelvic inlet<br>• Larger pubic tubercle height | • Positive circumferential resection margin<br>• Incomplete TME<br>• Longer operative time<br>• Higher blood loss<br>• Anastomotic leakage<br>• Conversion to open surgery<br>• Postoperative complications<br>• Longer pelvic dissection time |
| Raheem et al. | • Smaller interspinous distance (IS)<br>• Narrow pelvic inlet (PI)<br>• Large mesorectal fat area (MFA) | • Longer operation time<br>• Higher chance of anastomotic leakage |
| Yamamoto et al. | • BMI $\geq 25$<br>• Tumor size $> 45$ mm<br>• Anorectal angle $\geq 123^{\circ}$<br>• Pelvic outlet $< 82.7$ mm | • Predictive value for surgical difficulty in rectal surgery |
| Chau et al. | • S1-S5 angle $> 74.3^{\circ}$ | • Poor-quality resections in laparoscopic low anterior resection (LAR) |

**Table 3:** Significant parameters and surgical outcomes studied in selected research.

## 2.7 Machine Learning in Total Mesorectal Excision

Miao Yu et al. developed multiple machine-learning models to predict the surgical difficulty of LaTME. This study had a dataset of 626 patients who each underwent LaTME. Grading of surgical difficulty was done using a modified Escal rating due to differences between eastern and western patients, where postoperative hospital stay changed from >15 days to >12 days and the duration of surgery from >300 min to >240 min. A total of 35 variables were used in this study, including 20 pelvimetry measurements. To reduce dimension and solve collinearity, the least absolute shrinkage and selection operator (LASSO) was used. A total of six machine learning models were developed, where XGBoost displayed the best results with an Area Under the Receiver Operating Characteristic curve (AUROC) of 0.855. A total of seven variables were identified as predictors of surgical difficulty. The variables include Tumor distance to the anal verge, prognostic nutritional index (PNI), pelvic inlet, pelvic outlet, sacrococcygeal distance, mesorectal fat area, and angle 5 (the angle formed by the apex of the sacral curvature and the lower boundary of the pubic bone)[36].

Liu et al. investigated the possibility of predicting the need for a permanent stoma using machine learning. A permanent stoma is often used for patients to reduce the pressure on the anastomosis when a patient has a preserved anus after LaTME. A total of 1163 patients were included in this study. Results from an XGBoost prediction model displayed an ROC as high as 0.963 for the validation set. Risk factors for a permanent stoma were age $> 65$, rectal stenosis,

history of hypertension, history of diabetes, history of chemo- or radiotherapy, and distance of >5 cm from the tumor to the dentate line[37].

## 2.8 Study Objective

This study addresses the clinical challenge of predicting complications and surgical difficulty in Total Mesorectal Excision (TME) surgeries, including Low Anterior Resection (LAR) and Abdominoperineal Resection (APR). By leveraging machine learning, the goal is to enhance patient outcomes through personalized care strategies, ultimately reducing postoperative complications and improving surgical success rates.

# Technical Background

MRI plays a crucial role in the preoperative assessment, diagnosis, and treatment of rectal cancer [38]. A growing field of interest is the use of pelvimetry measurements derived from MRI as predictive factors for surgical outcomes in rectal cancer procedures, such as Total Mesorectal Excision (TME). Automated pelvimetry measurements from T2-MRI scans could enhance a clinician's understanding of the complexity of a patient's case, improving the pre-operative planning phase [9]. Simultaneously, the application of artificial intelligence (AI) in healthcare is rapidly evolving, offering tools to improve surgical assessment of patients. This study aims to incorporate AI by developing automated pelvimetry measurement techniques and prediction models that assess both surgical difficulty and surgery duration [39]. Specifically, in the context of TME, these AI-driven models could support clinicians by combining pelvimetry measurements with clinical parameters to better forecast surgical challenges and potential complications, such as anastomotic leakage.

## 3.1 Machine Learning

Machine learning is a part of artificial intelligence, as seen in Figure 3. It is based on the ability to let computers learn from data, often used to make predictions. The machine, in this case, the computer, can learn and improve due to experience. An essential part of machine learning is data pre-processing, which involves the transformation, cleaning, and organization of data to improve the accuracy of an algorithm. Machine learning may be used when a human's expertise is unavailable or when the complexity of the data requires computers to find more complex relations between variables that are difficult for humans to detect alone [40, 41].

The application of machine learning in medicine is increasing, thanks to the improved data storage capacity, stronger computational power, and a large volume of data. Two of the main fields it is used for are the prediction of diagnosis and the outcome of, for example, a surgical operation such as TME. However, these are not the only subjects that machine learning is useful for in medicine [42]. Other machine learning tasks include the ability to reconstruct the mechanisms of a disease, find suitable patients for recruitment for clinical trials, predict a patient's prognosis, or continuously monitor a patient's health to detect arrhythmias [43].

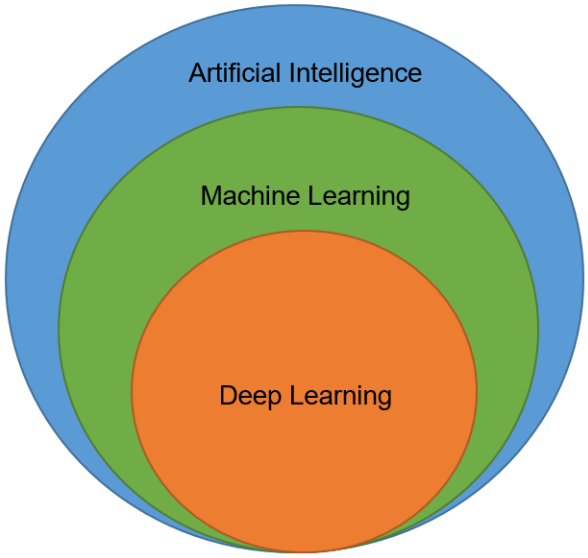



**Figure 3:** Venn Diagram Illustrating the Scope of Artificial Intelligence, Machine Learning, and Deep Learning

## 3.2 Machine Learning Models in study

### Logistic Regression

Logistic regression is a supervised machine learning algorithm used for binary classification. It estimates the probability of whether an instance belongs to one of two classes. The model applies a logistic function to the linear combination of input features, resulting in an output between 0 and 1. Using this probability, the instance is classified into one of the two classes, as shown in Figure 4a.

### Random Forest

A Random Forest is an ensemble machine learning model based on decision trees. During the training phase, multiple random decision trees are combined, and each tree outputs a prediction, displayed in Figure 4b For classification tasks, the final output is the class with the most instances, and for regression tasks, the mean of all outputs is taken.

### XGBoost (eXtreme Gradient Boosting)

XGBoost implements the gradient-boosting technique applied to decision trees. It constructs a series of decision trees where each new tree corrects errors from the previous one by minimizing residual error. The final prediction is calculated by taking the weighted sum of all trees and adjusting it by the learning rate, shown in Figure 4c. Each of these models has its own advantages and disadvantages, as outlined in Table 4.


**(a)** Logistic Regression[44] **(b)** Random Forest[45] **(c)** XGBoost[46]

**Figure 4:** Illustrations of the structure of Logistic Regression, Random Forest, and XGBoost

| Model | Advantages | Disadvantages |
|---|---|---|
| Logistic Regression | • Simple and interpretable<br>• Scales well with large datasets | • Assumes linear relationships<br>• Sensitive to missing data |
| Random Forest | • Resistant to noise and overfitting<br>• Provides feature importance<br>• Handles non-linear relationships | • Difficult to interpret<br>• High computational requirements |
| XGBoost | • Optimized for speed and scalability<br>• Supports various loss functions<br>• Automatically handles missing values | • Complex hyperparameter tuning<br>• Requires significant computational power<br>• Can overfit with complex models |

**Table 4:** Summary of advantages and disadvantages for Logistic Regression, Random Forest, and XGBoost.

## 3.3 Deep Learning

Deep learning is a subset of artificial intelligence and a specialized area of machine learning. It uses a neural network that, when properly trained, can handle large datasets and achieve high accuracy. A neural network is structured in multiple layers, each containing data transformations that make the data more abstract. Starting layers recognize simple features, while later layers recognize increasingly complex features. At the final layer, the model identifies patterns and creates a prediction [47, 48].

Each layer consists of multiple neurons, each containing a weight and a bias. The weight influences how input values propagate through the network, while the bias allows the neurons to have flexibility by changing the output, independent of the input. After the final layer, the model calculates the error (loss) for each prediction against the actual output (target value).

This loss is used as feedback for the model to learn and adjust weights and biases through a process called backpropagation, improving accuracy over time.

Deep learning has seen significant success in medicine, especially in the field of computer vision, which focuses on image and video understanding. Tasks such as object detection, classification, and segmentation are common. The use of convolutional neural networks (CNNs) leverages spatial invariance, ensuring features remain relevant regardless of their position in the image. Deep learning has excelled in medical imaging, with examples like melanoma identification [49] or breast lesion detection in mammograms [50].

# Three Pillars of this study

This chapter outlines the three main components of this thesis, illustrated in Figure 5:

1. **Machine Learning Model Development**: Creating a predictive model to assess the likelihood of complications during Total Mesorectal Excision (TME) surgery.
2. **Deep Learning for Sacral Curve Analysis**: Utilizing deep learning techniques to estimate the length of the sacral curve, aiming to identify novel parameters that may impact the complexity and outcomes of TME surgery.
3. **Dashboard Development**: Designing a proof-of-concept dashboard to support surgeons in decision-making, enabling personalized, patient-specific care planning.

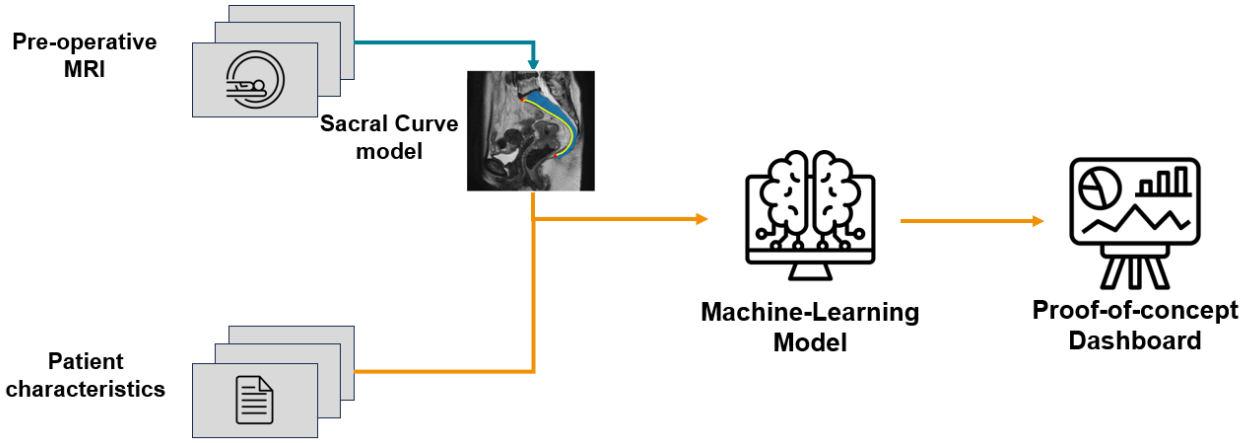



**Figure 5:** Workflow of the Three-Pillar Approach in This Thesis

## 5.1 Introduction

The main chapter of this thesis was the development of a machine learning model to predict post-operative complications, aiming to optimize patient-tailored care and outcomes. The following steps were taken to develop such a model:

1. **Data preprocessing:** Data preprocessing involved cleaning and organizing the dataframe for further analysis, filtering the dataset based on the in- and exclusion criteria.
2. **Univariate Analysis:** Initial exploration of individual relationships between features and the target outcome, aimed at identifying potential predictors of complications.
3. **Machine Learning Pipeline Development:** Target selection, univariate correlation, feature selection, and model training were implemented in a pipeline to build and evaluate machine learning models.
4. **Model Evaluation:** The different machine learning models were evaluated and optimized for clinical application.

The final models were evaluated for two post-operative target outcomes predicted following TME surgery:

1. **Clavien-Dindo grade 3 or higher (CDC3+):** This outcome predicted whether a patient would experience a Clavien-Dindo grade complication of 3 or higher after TME surgery, defined as requiring surgical, endoscopic, or radiological intervention. The detailed gradings of the classification are provided in Table 5.
2. **Anastomotic leakage:** This outcome predicts the probability of anastomotic leakage occurring after surgery.

| Grade | Definition |
|---|---|
| Grade I | Any deviation from the normal postoperative course without the need for pharmacological treatment or surgical, endoscopic, and radiological interventions. |
| Grade II | Requiring pharmacological treatment with drugs other than those allowed for Grade I complications. Includes blood transfusions and total parenteral nutrition. |
| Grade III | Requiring surgical, endoscopic, or radiological intervention. |
| Grade IIIa | Intervention not under general anesthesia. |
| Grade IIIb | Intervention under general anesthesia. |
| Grade IV | Life-threatening complication (including CNS complications) requiring IC/ICU management. |
| Grade IVa | Single organ dysfunction (including dialysis). |
| Grade IVb | Multi-organ dysfunction. |
| Grade V | Death of a patient. |
| Suffix "d" | If the patient suffers from a complication at the time of discharge (for "disability"), this is added to the respective grade of complication. |
| *Brain hemorrhage, ischemic stroke, subarachnoidal bleeding, but excluding transient ischemic attacks. CNS: central nervous system, IC: intermediate care, ICU: intensive care unit. | |

**Table 5:** Clavien-Dindo Classification of Postoperative Complications[51]

## 5.2 Clinical parameters

An overview of all the clinical features in the final dataset after preprocessing with description is displayed in Table 6. These features, combined with the pelvimetry measurements in Table 7 are utilized in the final dataset.

| Feature | Description | Feature type |
|---|---|---|
| abdomen | Earlier non-related abdomen surgery | Binary |
| age | Age | Continuous |
| asascore | ASA (American Society of Anesthesiologists) physical status score | Categorical |
| BMI | Body Mass Index (BMI) | Continuous |
| distance to tumor | Distance from tumor to ARJ | Continuous |
| intervention | Laparoscopic or Robotic performed | Binary |
| height | Patient's height | Continuous |
| chemotherapy | Indicates whether preoperative chemotherapy was administered | Categorical |
| therapy | Surgery, chemo- or radiotherapy before surgery | Categorical |
| radiotherapy | Radiotherapy applied | Continuous |
| procedure | LAR or APR | Binary |
| T-stage | T-staging before operation | Categorical |
| N-stage | N-staging before operation outcomes | Categorical |
| M-stage | M-staging before operation | Categorical |
| sex | Patient's gender | Binary |
| weight | Patient's weight | Continuous |

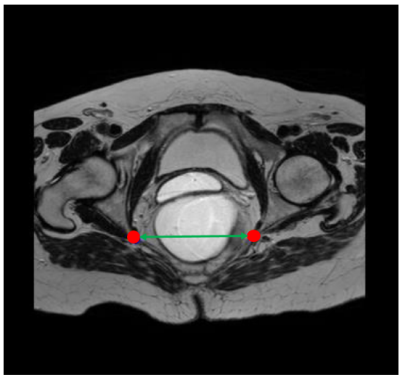
**Table 6:** Summary of clinical features of the dataset with descriptions and types.

## 5.3 Pelvimetry measurements

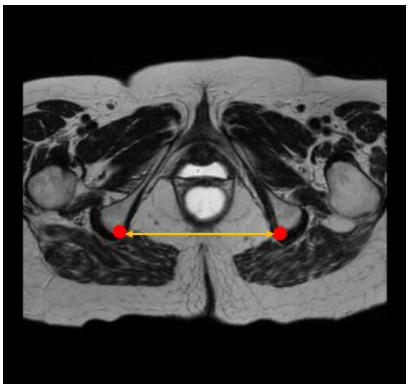
An overview of all pelvimetry measurement features used in the machine learning model are displayed in Table 7, including pelvimetry distances and angles calculated from points identified in the pelvis. Distances such as the Pelvic Inlet, Pelvic Outlet, pubic tubercle height, and sacrococcygeal distance were measured. Furthermore, the interspinous distance (IS) and intertuberous distance (IT) were measured, as displayed in Figure 6.

The anatomical points that these measurements are based on were automatically detected using an in-house deep learning model. This model identified five key points in the pelvis on a sagittal T2-MRI scan, as shown in Figure 7a. These points included the promontorium (A), S3 vertebra (B), coccyx (C), caudal part of the pubic symphysis (D), and cranial part of the pubic symphysis (E).

In addition to these measurements, the dataset included the distance between the promontorium (A) and G (the middle of the pelvic outlet) and the distance between point F (the middle of the pelvic inlet) and the coccyx, as shown in Figure 7b. Moreover, the pelvic depth (illustrated by the red line) and the sacral depth (depicted by the yellow line) were also measured, as seen in Figure 8a. Angles that were measured are displayed in Figure 8b.

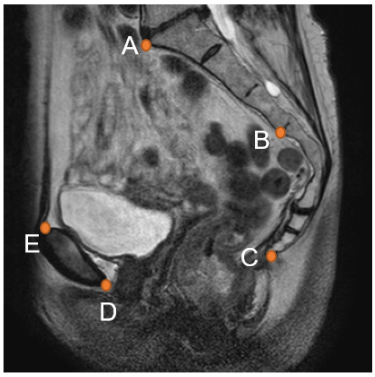


**(a)** Interspinous distance (green line) visualized on a transversal MRI.

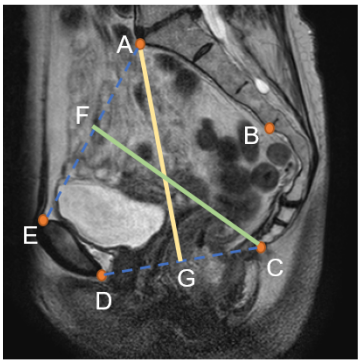


**(b)** Intertuberous distance (orange line) visualized on a transversal MRI.

**Figure 6:** Visualization of transversal measurements on MRI: (a) Interspinous distance and (b) Intertuberous distance.

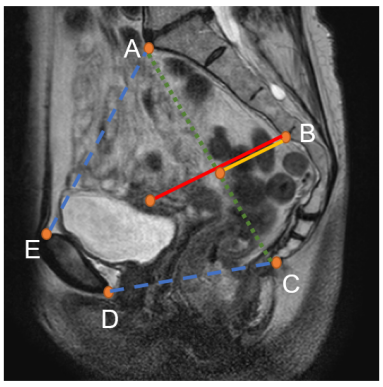



**(a)** Five automatically detected key points: promontorium (A), S3 vertebra (B), coccyx (C), lower part of the os pubis (D), and upper part of the os pubis (E).

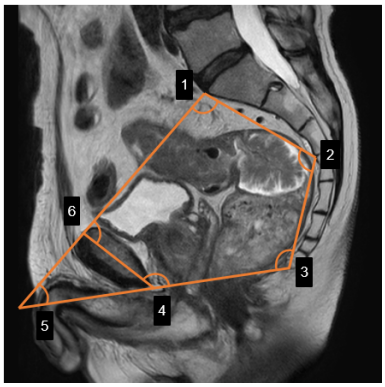



**(b)** Distances between key pelvic points: point A to G (middle of pelvic outlet) and point F to C (middle of pelvic inlet).

**Figure 7:** Illustration of automatically detected key pelvic points (a) and the distances measured between points F-C and A-G (b) on sagittal MRI




**(a)** Pelvic depth (red line) and sacral depth (yellow line) visualized on a sagittal MRI.




**(b)** Visualization of the six angles calculated for the machine learning model

**Figure 8:** Visualization of pelvic and sacral depth measurements (a) and six angles (b) on MRI.

### Frustum Volume of a Cone

Using the pelvic inlet (AE) and pelvic outlet (CD), a rough estimation of the pelvic volume (cm3) was calculated based on an inverted frustum-shaped cone visualized in Figure 9 calculated using Equation 1. The pelvic area resembles a cone, which is the region the surgeon operates in. For $R$, half the distance of the pelvic inlet was used and for the smaller radius $r$, half the distance of the pelvic outlet was chosen. The height $h$ was calculated as the mean of distances $AC$ and $DE$.

$$V = \frac{1}{3}\pi h \left(R^2 + Rr + r^2\right) \tag{1}$$

- $V$: Volume of the frustum
- $h$: Height of the frustum (distance between the two circular bases)
- $R$: Radius of the larger circular base
- $r$: Radius of the smaller circular base
- $\pi$: Pi, approximately 3.14

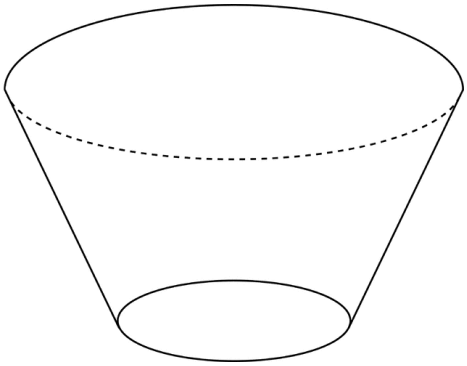


**Figure 9:** Visual representation of the inverted frustum used for volume calculation.

### Area Pentagon

To calculate the area of the pentagon (cm2), all five points $A$ through $E$ were used. To simplify the calculation, the points were projected onto a 2D plane on the xy-axis. The area was then calculated using the Shoelace formula shown in Equation 2:

$$A = \frac{1}{2}\left|\sum_{i=1}^{n}\left(x_i \cdot y_{i+1} - y_i \cdot x_{i+1}\right)\right| \tag{2}$$

- $A$: Area of the pentagon.
- $x_i, y_i$: Coordinates of the pentagon points in the 2D plane.
- $n$: Number of vertices in the pentagon.

**Table 7:** Pelvimetry measurements used in the machine learning model.

| Pelvimetry Measurement | Description |
|---|---|
| AB | Distance between promontorium and S3 vertebra. |
| AC | Sacrococcygeal distance. |
| AD | Diagonal conjugate. |
| AE | Pelvic inlet. |
| AG | Distance between promontorium and G (middle of pelvic outlet). |
| BD | Distance between S3 vertebra and caudal part of os pubis. |
| BE | Distance between S3 vertebra and cranial part of os pubis. |
| BC | Distance between S3 vertebra and coccyx. |
| CD | Pelvic outlet. |
| CE | Distance between coccyx and caudal part of os pubis. |
| DE | Pubic tubercle height. |
| FC | Distance between F (middle of pelvic inlet) and coccyx. |
| Angle 1 | Angle between points ABE. |
| Angle 2 | Sacral curve angle. |
| Angle 3 | Angle between points CDB. |
| Angle 4 | Angle between points DEC. |
| Angle 5 | Sacrococcygeal-pubic angle. |
| Angle 6 | Angle between points EAD. |
| IS | Distance between ischial spines. |
| IT | Distance between ischial tuberosities. |
| Frustum Volume | Volume of the frustum-shaped cone. |
| Sacral Depth | Distance from S3 vertebra to midline of AC. |
| Area Pentagon | Area of the pentagon formed by pelvic points (A-E). |
| Pelvic Depth | Distance from S3 vertebra to midline of AC and DE. |
| Frustum Volume/Pelvic Depth | Ratio of frustum volume to pelvic depth. |

## 5.4 Pre-processing

### Study Design

This is a retrospective study involving eight TME centers in the Netherlands, performing more than forty TME surgeries annually. For each patient involved, they underwent a primary tumor rectal resection. Furthermore, all patients had a preoperative MRI performed between January 1, 2013, and December 31, 2021.

### Inclusion Criteria

As shown in Figure 12, the following criteria were employed to include patients in the dataset:

- **Resection Treatment:** Only patients who underwent resection treatment were included.
- **Curative Operation:** The operation must have been intended as a curative procedure.
- **Primary tumor:** TME must have been performed for the primary tumor.
- **Elective Surgery:** Only elective (non-emergency) surgeries were considered.

- **Specific Surgical Procedures:** Only Low Anterior Resection (LAR) or Abdominoperineal Resection (APR) procedures were included.
- **Surgical Method:** The operation must have been performed laparoscopically or robotically.
- **MRI Annotation:** A preoperative pelvic MRI must be available containing complete visualization of the bony pelvis to ensure complete pelvimetry measurements.

### Exclusion Criteria

The following criteria were used to exclude patients from the dataset:

- **No Transanal Endoscopic Microsurgery (TEM):** Patients who had TEM prior to resection were excluded.
- **Tumor Proximity to Anal Rectal Junction (ARJ):** Tumors must be within 14 cm of the ARJ.

### Data Cleaning

After applying the inclusion and exclusion criteria, the remaining dataset underwent a thorough cleaning process to ensure its integrity and quality for analysis. This process involved merging datasets, handling missing values, removing outliers, and engineering new features.

*Data merging:* Data from several hospitals was combined to create a single clinical dataset, afterwards the pelvimetry dataset was incorporated.

*Handling missing values:* Columns with a high percentage of missing values (80% or more) were removed. For features with fewer missing values, the *K-nearest neighbor (KNN)* imputation algorithm was applied, as it preserves relationships in the data and works for both categorical and continuous variables. Missing values that represented unavailable information, such as radiotherapy, were imputed with a placeholder value of 1, corresponding to "no radiotherapy performed."

*Outlier removal:* Outliers were identified using boxplot visualizations and Z-scores analysis in which a threshold of 3 standard deviations was applied. In the Z-score formula (Equation 3), $x$ represents the data point. As a result, age values below 33 were excluded (see Figure 10). Anatomically impossible values, such as unrealistic distances between the promontorium and coccyx (AC), Pelvic inlet (AE) and Pelvic outlet (CD), were also removed

$$Z\text{-}score = \frac{x - \text{mean}}{\text{std. deviation}} \tag{3}$$

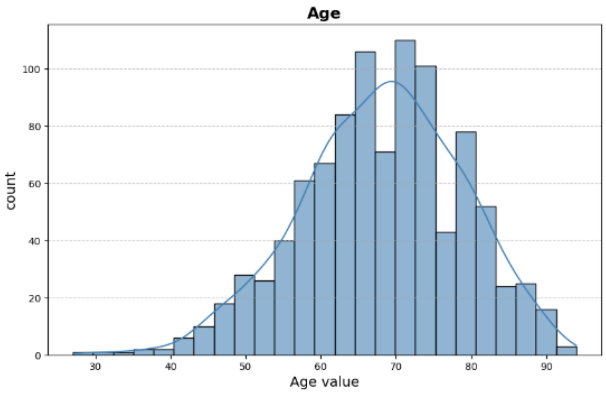
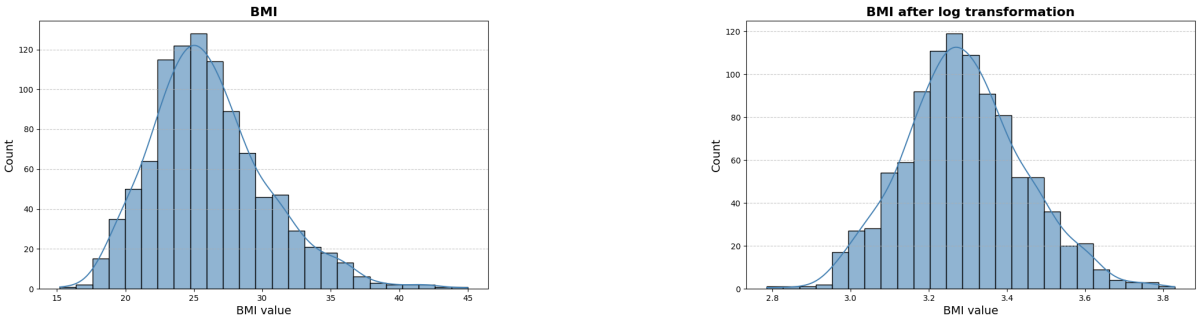



**Figure 10:** Histogram distribution of Age, displaying a small right skew

*Feature engineering:* To enhance model performance, new features were created: Body Mass Index (BMI), calculated from the patient's weight and height; and Clavien-Dindo 3+ grading, transformed into a binary outcome where grades 1–2 were classified as negative (0) and grades 3–5 as positive (1).

*Scaling and transformation:* Features with skewed distributions, such as BMI (skewness: 0.77), were log-transformed to reduce extreme values. Continuous variables were scaled between 0 and 1 to standardize feature magnitudes, ensuring equal contribution during modelling.




**(a)** BMI distribution (right skew)

**(b)** BMI after log transformation

**Figure 11:** (a) Histogram of BMI distribution, and (b) Histogram of BMI distribution after log transformation.

*Encoding categorical variables:* Categorical variables were encoded. Dichotomous variables, such as sex (male, female) and surgery type (LAR, APR), were converted to binary values. Ordinal variables, including neo-adjuvant radiotherapy (none, short-course, long-course, chemoradiotherapy) and TNM classification (T-stadium: 1–4, N-stadium: 0–2), were scaled between 0 and 1. The ASA score was preserved in its original ordinal structure due to its clinical significance.

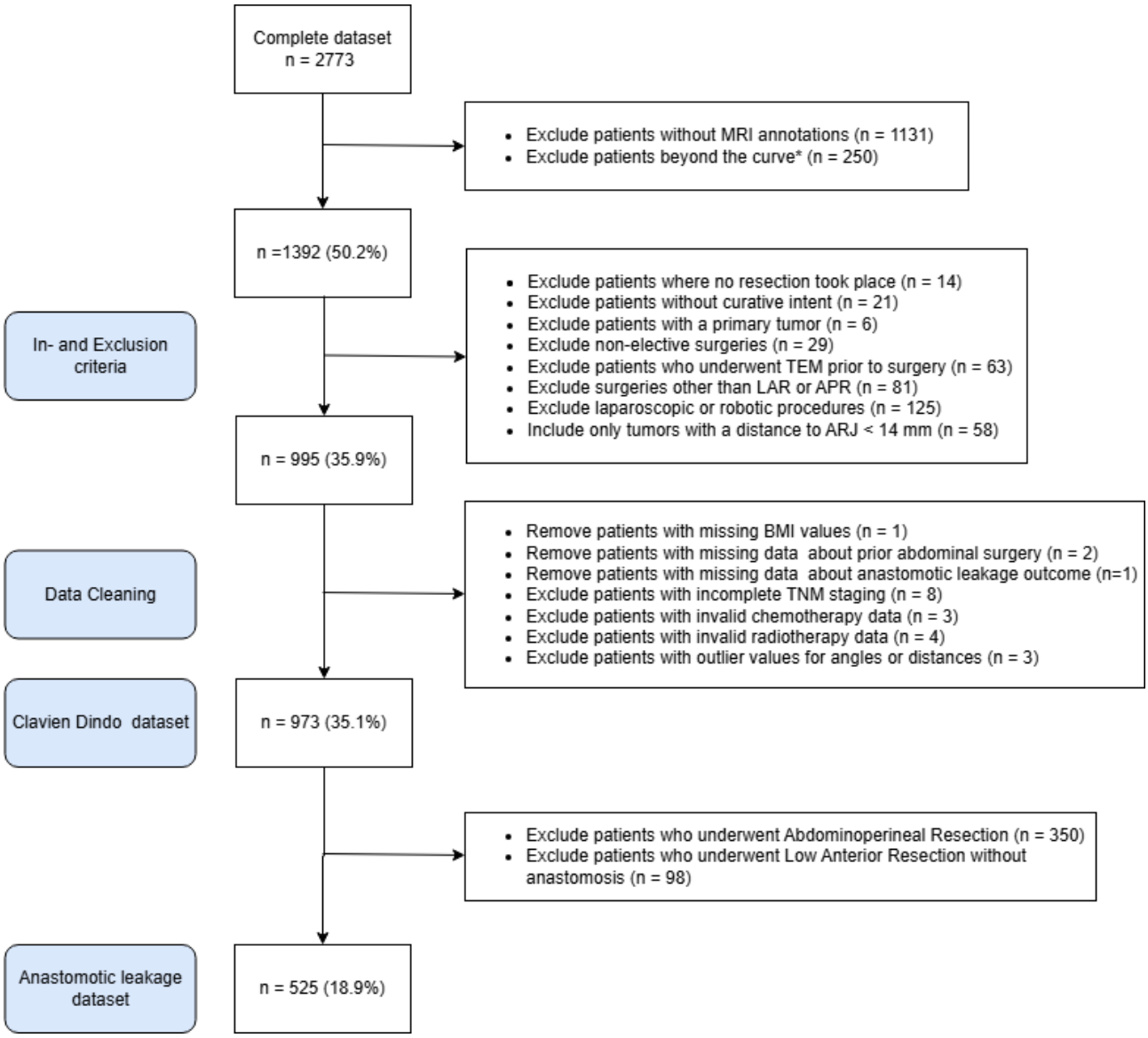



**Figure 12:** Flowchart illustrating the inclusion and exclusion criteria, along with additional data cleaning steps, used to shape the final dataset.

### Final Dataset

The final dataset of CDC3+ consisted of 973 data points with 42 predictive features selected based on clinical expertise and univariate correlation analysis. The dataset exhibited a class imbalance with an approximate 5.2:1 ratio between negative and positive classes.

For anastomotic leakage predictions, a subset of 525 data points was constructed by filtering for patients who underwent Low Anterior Resection (LAR) with a performed anastomosis. This subset displayed a class ratio of 5.5:1 between negative and positive classes. Addressing these imbalances is critical for model reliability, using techniques like class weighting or synthetic resampling (e.g., SMOTE).

### Univariate correlation

To evaluate the relationship between the predictor features and the target variable, two types of univariate analyses were performed: the Point Biserial correlation for acontinious predictor features and the Chi-square-test with Cramer's V for categorical and binary features. The

formulas for these tests are explained in the appendix. These methods allowed for measuring the associations between variables and provided insights into potential variables influencing surgical difficulty. The strength and interpretation of the association are displayed in Table 8.

| Estimated values | Interpretation of association |
|---|---|
| 0.00–0.09 | Negligible |
| 0.10–0.19 | Weak |
| 0.20–0.39 | Moderate |
| 0.40–0.59 | Relatively strong |
| 0.60–0.79 | Strong |
| 0.80–1.00 | Very strong |

**Table 8:** Interpretation of correlation based on features

## 5.5 Pipeline

This chapter explains the development of the pipeline design. A well-designed and structured pipeline ensures that no data-leakage occurs and due to the systematic approach, it has several benefits such as: consistency, reproducibility, efficiency, scalability and experimentation. The use of a pipeline reduces the chance of errors and improves the ease of logging results and models. An overview of the pipeline and its components are displayed in Figure 14

In this study, the best model was trained by systematically optimizing parameters for multiple models. This was conducted for both target outcomes. Each step involved testing multiple combinations of parameters to identify the optimal configuration based on performance metrics. All individual runs and their results were tracked and recorded in Weights and Biases for comprehensive analysis and comparison [52].

### Stratified Cross Validation

In this pipeline, cross-validation is applied at multiple components. Cross-validation is the process of splitting the data into k-folds where each fold has a different distribution of train and test sets[53]. By averaging the results of the final testing, this ensures that the model does not over fit for a certain fold. However, when using normal cross-validation, the distribution of classes across folds might not be guaranteed to match the distribution of the original dataset. This can lead to problems, especially if the dataset has imbalanced classes.

Stratified k-fold cross-validation tackles this problem by ensuring that the class distribution is similar in each fold ($k$), displayed in Figure 13. This is especially useful in datasets with a class imbalance. This method leads to more robust results and a more generalizable model[54]. In this method, a value of five folds is chosen as this strikes a good balance between bias and variance[55].

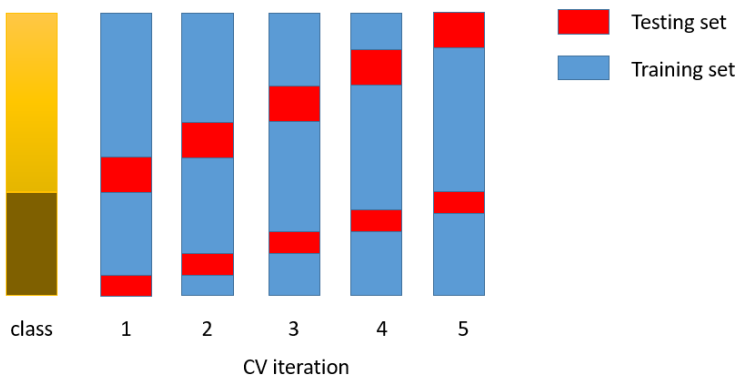



**Figure 13:** Visualization of Stratified K-Fold Cross-Validation with Training (Blue) and Testing (Red) Splits

## 5.6 Pipeline components

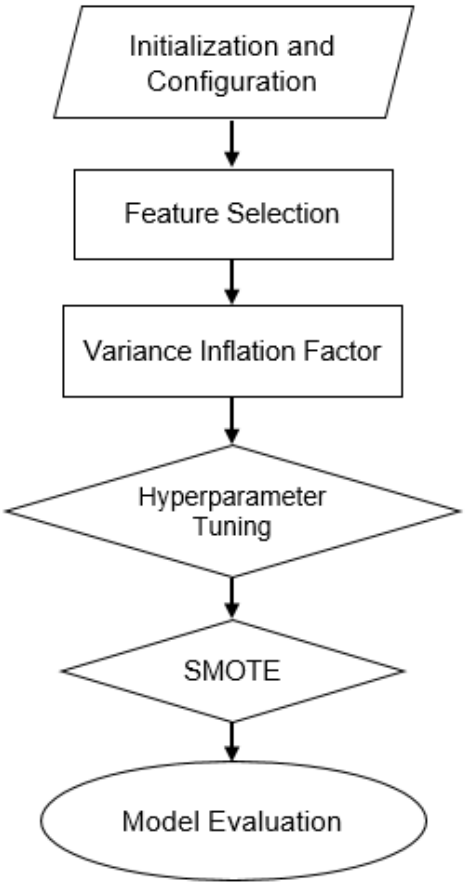



**Figure 14:** Pipeline structure illustrating key steps: initialization and configuration, feature selection, variance inflation factor (VIF) calculation, hyperparameter tuning, SMOTE application, and model evaluation

### Pipeline initialization and configuration

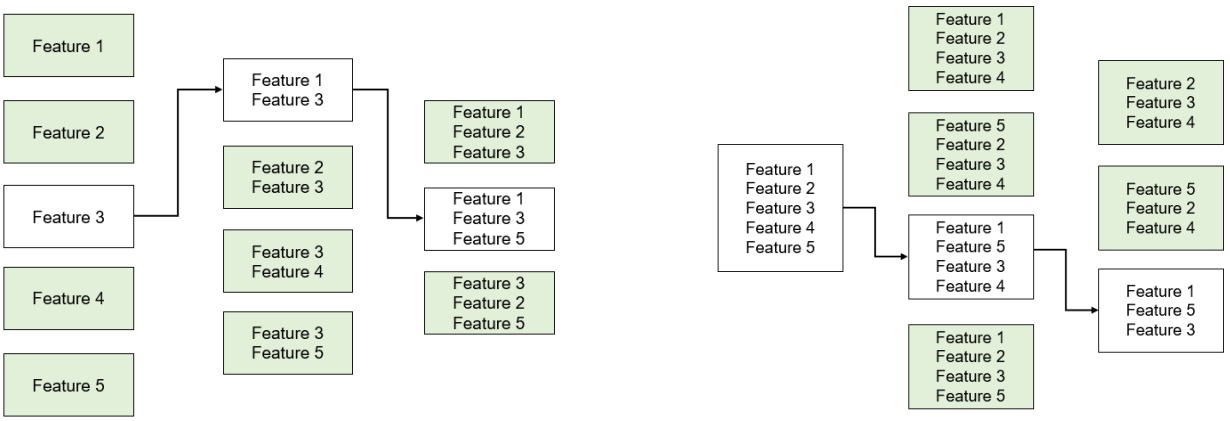
The first component was the loading of the preprocessed dataset and selection of one of the two classification target outcomes. Furthermore, multiple important decisions are made such as:

- **Machine learning model type:** In this pipeline the decision can be made between three classification machine learning models: Logistic regression, Random Forest Classifier, and XGBoost.
- **Class Weighting:** Class imbalance is when one of the classes is underrepresented in the dataset[56]. This can lead to incorrect and biased results in model training. Class weighting assigns higher weights to the minority class, which makes the model pay more attention to that class and reduces the bias towards the majority class.
- **Distance threshold to anorectal junction (ARJ):** Selecting the maximum distance between the tumor and the ARJ. This will influence the size of the dataset due to more cases being excluded when lowering the distance. In this method, a value of 14 cm was used.
- **Scoring metric:** Choose the evaluation metric that will be used to measure performance during feature selection. The options for this were accuracy, precision, recall (sensitivity), F1-score, and AUC.

### Feature Selection

A large dimensionality in a dataset is often not preferred as this often leads to overfitting, has a high computational cost, and makes it difficult to interpret the model. Feature selection is a method that is widely used to minimize the use of irrelevant features; it aims to choose a small subset of features that are deemed relevant for the training of the model while removing features that add noise to the dataset[57, 58, 59]. In this method, both forward and backward feature selection are applied to use the strengths of both methods. To ensure robustness in this process a pre-built python package from sci-kit learn is used called SequentialFeatureSelector to create the forward and backward feature selection.

During feature selection, the F1-score is used as the scoring metric for feature selection as it considers both precision and recall. If the focus is solely on recall (sensitivity), the model might choose features that lead to a perfect recall (1.0), but at the cost of poor precision. By choosing F1, the aim is to develop a balanced model.




**(a)** Forward Feature Selection **(b)** Backward Feature Selection

**Figure 15:** Illustration of (a) Forward and (b) Backward feature selection

Forward feature selection: The procedure iteratively determines the best new feature to add to the list of chosen features Figure 15a. It starts with zero features and identifies the feature that, when an estimator is trained on it, maximizes a cross-validated score using stratified k-fold cross-validation. The process continues by adding a new feature to the collection of selected features once the initial feature is chosen. When in the last 10 interactions no improvement was seen in the scoring metric the feature selection process ends and the feature set, with the highest score is used. An advantage of forward feature selection is that it starts with a small dataset, therefore it is less prone to collinearity. However, when a feature is added it can't be removed. It can make the feature added before this less relevant because it explains the same variance [59].

Backward feature selection: This procedure starts with all features, and then it iteratively determines the feature to remove after training on an estimator to minimize the negative impact on the stratified k-fold cross-validated score Figure 15b. The process continues with removing a feature until no improvement is seen. This continues until no improvement is displayed in the last ten iterations in the scoring metric. The final set of features that produces the best performance score is selected. An advantage of backward feature selection is due to starting with all features it is able to capture relationships between features. Sometimes a single feature might not contribute significantly, but in combination with other features have a meaningful impact on the performance of a model [59].

After forward and backward feature selection, the feature sets of both processes are combined and overlapping features will be selected displayed in Figure 16.

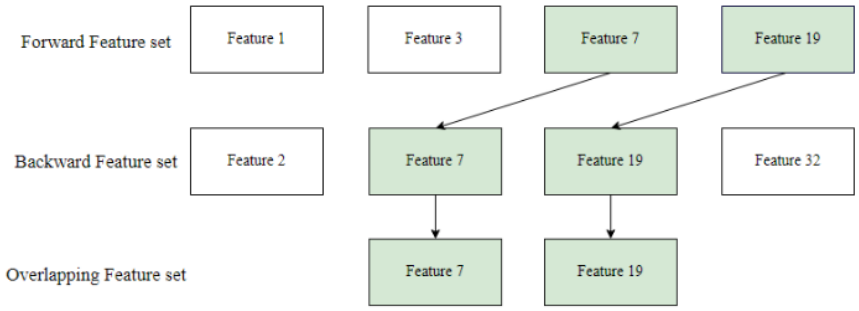



**Figure 16:** "Visualization of Features Selected by Forward and Backward Feature Selection, Highlighting Overlapping Features

### Variance Inflation Factor

The Variance Inflation Factor (VIF) is a statistical measure that is used to address multi-collinearity[60]. This happens when there is a high correlation between predictor variables, which can create instability in the training of a model and making it harder to interpret. The formula for the VIF in equation 4 [61].

$$VIF_i = \frac{1}{1 - R_i^2} \tag{4}$$

R2 is the coefficient of determination, which explains how well a model fits the data. It tells how well a predictor is explained by the other predictor variables. If the R2 is close to 0, the predictor is not closely correlated to other predictor variables. Whereas if it is closer to 1 it is highly correlated, resulting in a high VIF.

In this component of the pipeline the VIF is calculated for all predictor variables. Afterward, it removes the variable with the highest VIF and this iteration is repeated till all variables have a VIF below the selected threshold, The function is modified to ensure that overlapping features selected during the feature selection process cannot be removed. For this method a VIF of ten is selected, this value is somewhat high as it allows some multicollinearity. The reasoning behind this decision is that in our case it is preferred to have a greater confidence in the final prediction and less about the interpretability of the individual variables.

### Hyperparameter tuning

Hyperparameters are set manually in machine learning, this is conducted by the user before training. These types of parameters differ from internal parameters in such a way that hyperparameters affect the behavior and architecture of the algorithm. Tuning of hyperparameters is used to let a model be used optimally, and it can be performed manually but also done through hyperparameter tuning[62]. For this, stratified k-fold is used, to find the optimal parameters for five folds.

In this pipeline, the option used for hyperparameter tuning in the final models was Bayesian Optimization. This approach starts with random combinations and then builds a mathematical model to predict the hyperparameters that are most likely to improve the model. The benefit of Bayesian Optimization is its efficiency compared to Gridsearch. Instead of exhaustively searching all possible combinations, it focuses only on the most promising areas in the hyperparameter space, evaluating only the "good" combinations.

### SMOTE (Synthetic Minority Over-Sampling Technique)

In case class weighting is not applied to tackle the class imbalance, the second option is to apply SMOTE [63]. When SMOTE is applied, instead of duplicating existing samples, synthetic samples are generated by interpolating between a minority class instance and its nearest neighbour. A random point is created along the line segment between the feature vector and one of its $k$-nearest neighbours. $k$ can be adjusted based on the dataset. SMOTE is only performed for the minority class to balance the data set. An example of regular SMOTE is shown in Figure 17.

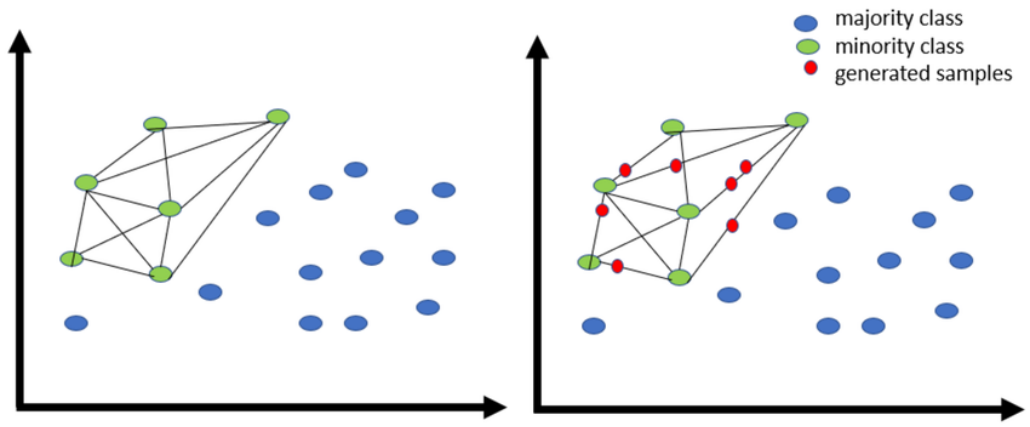



**Figure 17:** Illustration of Regular-SMOTE: Synthetic Minority Points (Red) Generated Between Original Minority Points (Blue) Amid Non-Minority Points (Grey) [64]

In this pipeline, multiple types of SMOTE can be applied, such as:

1. **Regular-SMOTE:** The standard implementation of SMOTE, where synthetic samples are generated by interpolating between a randomly chosen minority class instance and one of its k-nearest neighbors.
2. **Borderline-SMOTE:** Focuses on generating synthetic samples near the decision boundary, where minority class samples are more likely to be misclassified.
3. **ENN-SMOTE:** Combines SMOTE with Edited Nearest Neighbors (ENN), which removes noisy or borderline samples from the dataset after oversampling.
4. **Tomek-SMOTE:** Combines SMOTE with Tomek Links, which identify and remove overlapping samples from the majority class after oversampling.
5. **ADASYN (Adaptive Synthetic Sampling)-SMOTE:** Focuses on generating more synthetic samples in regions where the minority class is underrepresented, based on data density.
6. **SVM-SMOTE:** Uses Support Vector Machines (SVM) to identify critical boundary regions in the feature space where synthetic samples should be generated.
7. **Safe-Level-SMOTE:** Generates synthetic samples in "safe" regions of the minority class by considering the proximity of majority class samples.
8. **Means-SMOTE:** Applies K-Means clustering to group data into clusters and generates synthetic samples within each cluster.
9. **Cluster-SMOTE:** A variant of Means-SMOTE that generates synthetic samples specifically within clusters to respect the natural data structure.

### Model evaluation

To evaluate the results five fold stratified cross validation is applied as explained before, this approach helps to create a more robust model and prevents one of the five folds to produce unrealistic results. After training and fitting the data for each fold in the stratified k-fold, scoring metrics such as Accuracy, Precision, Recall, F1-score, Specificity, and Area Under the Curve (AUC) were calculated. By averaging the scores over the five folds, the results provide a realistic model performance. Furthermore, the final model was evaluated using the leave-one-out method on specific hospital test sets. For the evaluation, data from an individual hospital was used as a test set. The model was then trained on the rest of the dataset. Afterwards that model is evaluated for the hospital subset to assess the performance.

## 5.7 Results Anastomotic Leakage

### Top Features by Absolute Correlation

Among the top 10 features based on absolute correlation values, 50% of the features are pelvimetry measurements, highlighting their importance. All features displayed have a significant p-value. For continuous variables, the correlation can be either negative or positive. A negative correlation indicates that, as the feature value decreases, the chance of a positive case increases. For anastomotic leakage, the correlations are generally weak, indicating that they do not contribute significantly to prediction. However, the feature 'pelvic inlet' (AE) emerges as the largest predictor.

**Table 9:** Top 10 features Based on Correlation

| Variable | P-value | Correlation | CI Lower | CI Upper |
|---|---|---|---|---|
| AE (Pelvic Inlet) | 0.001 | -0.139 | -0.222 | -0.054 |
| distance to tumor | 0.002 | -0.137 | -0.220 | -0.052 |
| Frustum Volume / Pelvic depth | 0.002 | -0.137 | -0.220 | -0.052 |
| sex | 0.004 | 0.126 | 0.045 | 0.201 |
| T-stage | 0.003 | 0.118 | 0.048 | 0.192 |
| AD | 0.010 | -0.113 | -0.197 | -0.028 |
| IS (Interspinous) | 0.010 | -0.103 | -0.196 | -0.027 |
| Frustum Volume Cone | 0.019 | -0.103 | -0.184 | -0.017 |
| chemotherapy | 0.020 | 0.092 | 0.044 | 0.200 |
| therapy | 0.052 | 0.085 | 0.009 | 0.160 |

### 5.7.1 Machine Learning Model Results

### Best Performing Model

The best performing model was a Logistic Regression model trained using a Stratified 5-fold cross-validation approach, ensuring balanced representation of classes in each fold. The training set size was 516 samples, generated using SMOTE-ENN to address class imbalance. The test set size was 105 samples, consisting of 89 negative cases and 16 positive cases, with a class imbalance ratio of 5.5:1. The final hyperparameters used are a regularization strength of C=1 leading to a balanced model. An L2 penalty to prevent overfitting and the solver liblinear as this works well with L2 regularization. The features used in the final model are displayed in table 10 below:

**Table 10:** Overview of Features included the final anastomotic leakage model

| **Features** | |
|---|---|
| Angle 1 | distance to tumor |
| Angle 4 | weight |
| Angle 6 | chemotherapy |
| procedure | radiotherapy |
| intervention | sex |
| M-stage | therapy |

### Evaluation Metrics

The main performance metrics, based on stratified 5-fold cross-validation, are displayed below. The objective was partially met for the anastomotic leakage model, which reached a sensitivity score of 0.765 ± 0.140. The accuracy for the anastomotic leakage model was 0.528 ± 0.034. Furthermore, the model displayed a low specificity of 0.484 ± 0.057. The F1-score, representing the harmonic mean of precision and recall, reached 0.332 ± 0.042 for the anastomotic leakage model.

**Table 11:** Performance Metrics for Anastomotic Leakage Model

| **Metric** | **Accuracy** | **Sensitivity** | **Specificity** | **F1** | **AUC** |
|---|---|---|---|---|---|
| **Value** | 0.528 ± 0.034 | 0.765 ± 0.140 | 0.484 ± 0.057 | 0.332 ± 0.042 | 0.667 ± 0.075 |

### Confusion matrix

Figure 18 shows the average confusion matrix of the model's stratified cross-validation on a test set of 105 cases visualizing the class imbalance.

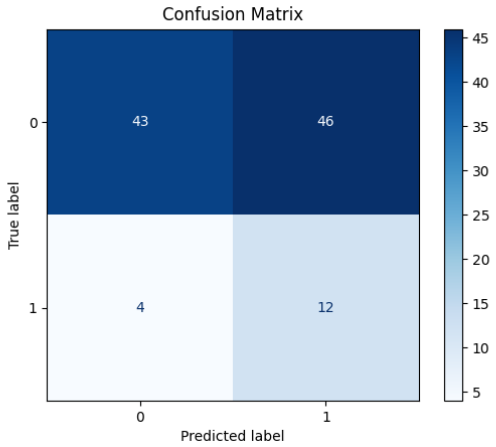



**Figure 18:** Confusion matrix showing the performance of the model. The values represent True Negatives (43), False Positives (46), False Negatives (4) and True Positives (12)

### Prediction Probability Distribution

Figure 19 displays the prediction probability distribution.

**Class 0 (Red):** A significant portion of Class 0 predictions falls within the 0.6 to 0.9 probability range, indicating that the model often classifies these true negative cases as positive with moderate confidence. This results in a large number of false positives, as true negative cases are frequently misclassified when their predicted probabilities exceed the 0.5 threshold.

**Class 1 (Green):** The majority of Class 1 cases are also concentrated in the 0.6 to 0.9 range, suggesting that the model has some discriminative ability in identifying positive cases but with only moderate confidence. This moderate range indicates that while the model can separate positive cases from negative ones to some extent, it lacks strong certainty.

Overall, there are no positive cases with predicted probabilities in the high-confidence range (0.9–1), indicating that the model does not demonstrate strong confidence in its classifications. This reinforces the interpretation that while the model is capable of identifying positive cases to some degree, its predictive confidence is limited, impacting the accuracy.

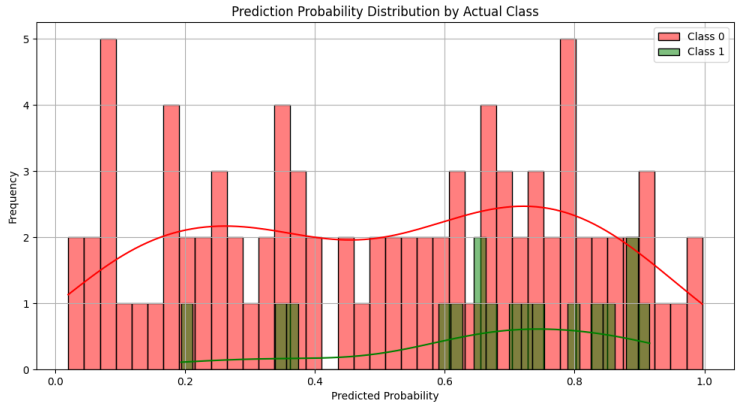



**Figure 19:** Prediction probability distribution shows the models prediction probability for each actual class, where Class 0 represents negative cases and Class 1 represents positive cases.

### Receiver-Operating Characteristic curve

The model's average AUC of 0.667 indicates that it is only slightly better than random chance. The curve does not display a prominent spike, indicating that the model struggles to achieve high sensitivity without generating a significant number of false positives. This pattern is typical of models that face difficulty distinguishing between classes effectively. Looking at the objective of a sensitivity (True Positive Rate) of 0.8 at a False Positive Rate of 0.5, the models interval is relatively small displaying consistency in the folds at this sensitivity.

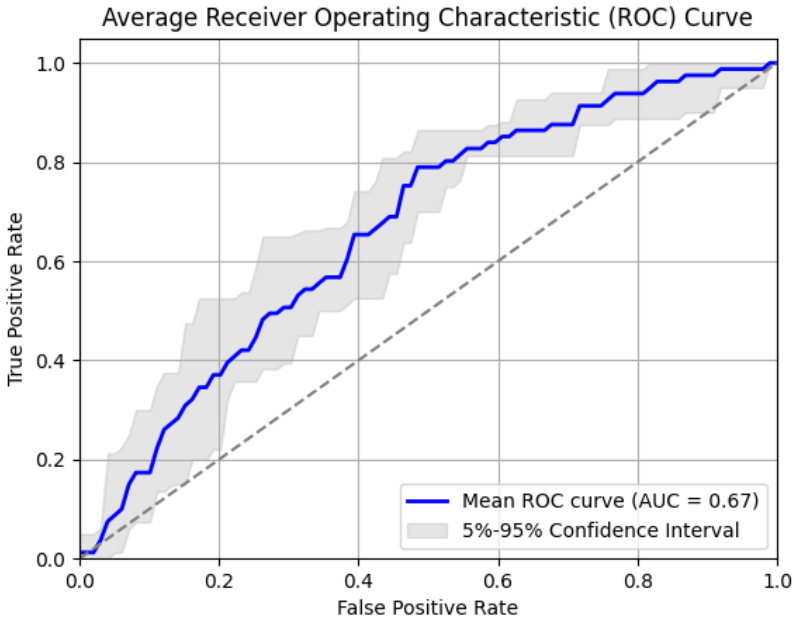



**Figure 20:** ROC curve with False Positive Rate (FPR) on the x-axis and True Positive Rate (TPR) on the y-axis

### Sensitivity vs Specificity

Figure 21 represents the trade-off between sensitivity and specificity at different thresholds during 5-fold cross-validation. This visualization helps identify the threshold that achieves the desired balance between sensitivity and specificity. In our case, the threshold should be set at approximately 0.50 to reach the target sensitivity of 0.8. While the intersection point of the sensitivity and specificity curves typically indicates an optimal threshold for balanced performance, our goal is to prioritize sensitivity for the model. Both the sensitivity and specificity are fairly consistent in their increase and decrease, with sensitivity having a slightly stronger decrease after the intersection point.

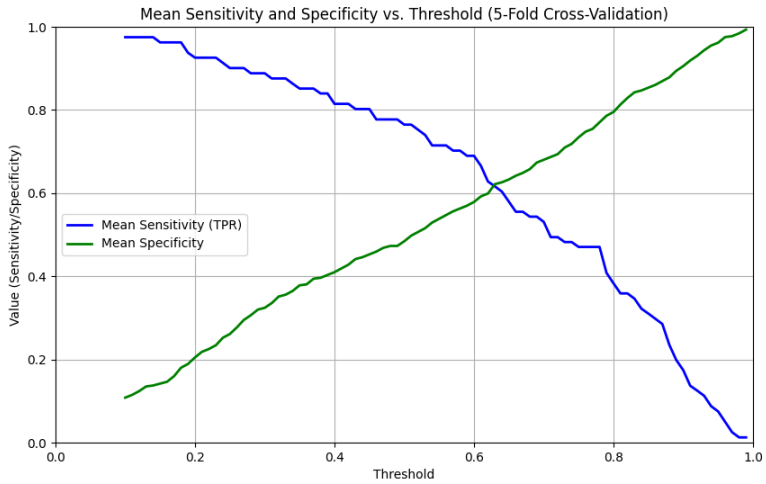



**Figure 21:** Graph visualizing the the mean sensitivity (blue) and specificity (green) at different thresholds

### Feature Importance

Compared to univariate correlation, feature importance reflects the role of a feature in the model's decision-making process. In LR, it is determined by the magnitude of the model's coefficients. Larger coefficients, whether positive or negative, indicate a stronger influence on the target outcome. A positive coefficient means that as the variable value increases, the log-odds of the positive class increase, making the positive outcome more likely. Conversely, a negative coefficient implies that as the feature value decreases, the positive class becomes more likely. If a feature has a coefficient of 1, a one-unit increase (similar to one standard deviation due to scaling) increases the log odds of the positive outcome by 1.

Figure 22 displays the feature importance for the top 10 features based on coefficient ($\beta$). For each feature, the average feature coefficient is calculated across all five folds. *Angle 1* displays the largest positive feature coefficient, as the angle of 1 increases. Furthermore, the largest negative coefficient is *distance to tumor*. This indicates that as the distance from the tumor to the ARJ decreases, the likelihood of anastomotic leakage increases. *sex* has a significant negative contribution to the outcome, displaying that being a female reduces the chance of anastomotic leakage. Most feature coefficient values are below 2. For anastomotic leakage, two out of the top 10 features are related to pelvimetry measurements.

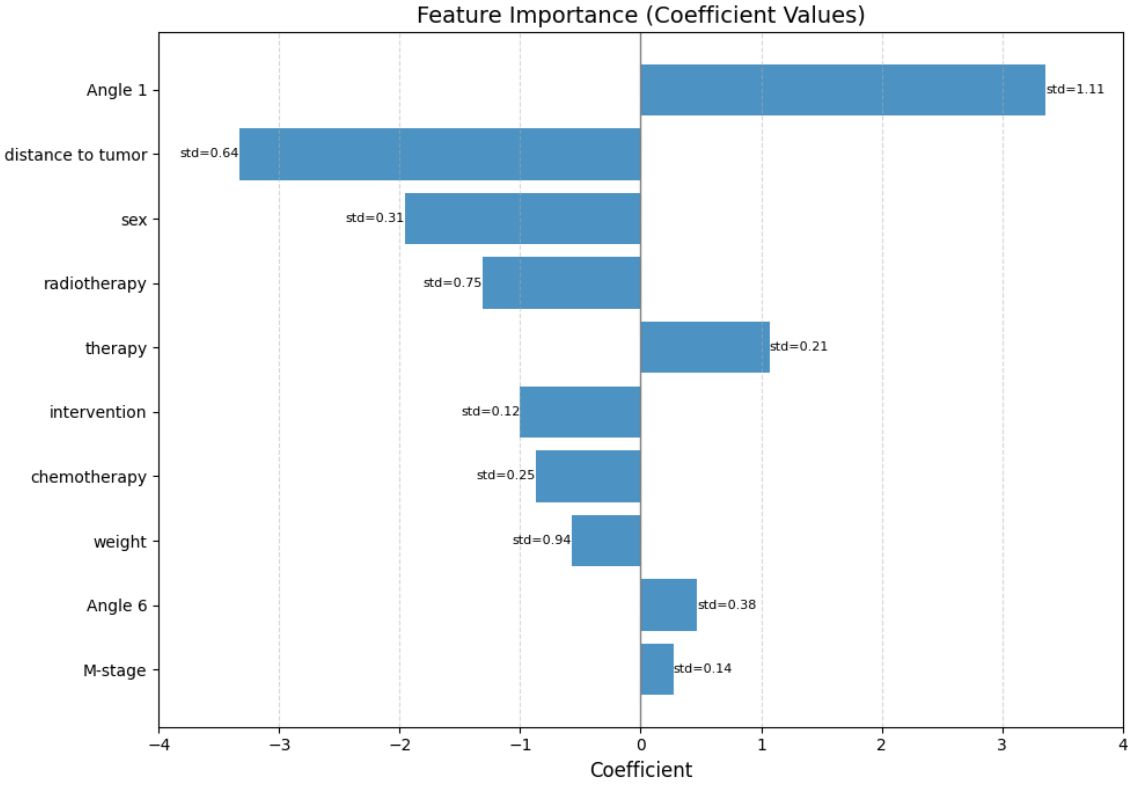



**Figure 22:** Ranked Feature Importance in the Models predictions (Highest to Lowest)

### External Validation

To test the models consistency across various subsets of the main dataset. For each dataset, only data from a certain hospital was chosen, therefore there is a strong variation in test set size. Sensitivity varies ranging from 0.0 to 1.0, and specificity 0.38 to 0.8. Results display that only three hospitals display the target sensitivity of 0.8. Hospital 7 displays a high AUC of 0.83 and a F1-score of 0.6, however due to the low sample size this result is likely unreliable.

**Table 12:** Performance Metrics per Hospital

| Hospital | n | Accuracy | Sensitivity | Specificity | F1 | AUC |
|---|---|---|---|---|---|---|
| 1 | 199 | 0.48 | 0.60 | 0.46 | 0.26 | 0.57 |
| 2 | 89 | 0.46 | 1.00 | 0.44 | 0.14 | 0.79 |
| 3 | 107 | 0.56 | 0.84 | 0.48 | 0.47 | 0.63 |
| 4 | 46 | 0.63 | 0.56 | 0.65 | 0.37 | 0.64 |
| 5 | 28 | 0.71 | 0.00 | 0.80 | 0.00 | 0.33 |
| 6 | 31 | 0.68 | 0.80 | 0.65 | 0.44 | 0.73 |
| 7 | 17 | 0.76 | 0.60 | 0.83 | 0.60 | 0.83 |
| 8 | 8 | 0.38 | 0.00 | 0.38 | 0.00 | NaN |

## 5.8 Results CDC3+

### Top Features by Absolute Correlation

Table 13 displays the top 10 features for the outcome of CDC3+ a confidence interval of 95% is used. The variable *sex* demonstrates the strongest correlation with the outcome, with a value of 0.104. However, this correlation is still weak and only marginally above the threshold for being considered negligible. All other features exhibit correlations below 0.1, indicating that their relationships with the outcome are weak. *sex*, the ratio between *Frustum Volume and Pelvic Depth*, *Pelvic inlet*, *Interspinous* distance, *AD* and *Frustum Volume Cone* all display a significant p-value.

**Table 13:** Top 10 features Based on Correlation

| Variable | P-value | Correlation | CI Lower | CI Upper |
|---|---|---|---|---|
| sex | 0.001 | 0.104 | 0.061 | 0.163 |
| Frustum Volume/Pelvic depth | 0.004 | -0.093 | -0.155 | -0.033 |
| AE (Pelvic Inlet) | 0.004 | -0.093 | -0.155 | -0.033 |
| IS (Interspinous) | 0.026 | -0.072 | -0.130 | -0.008 |
| AD | 0.026 | -0.072 | -0.130 | -0.008 |
| Frustum Volume Cone | 0.026 | -0.072 | -0.130 | -0.008 |
| height | 0.290 | 0.066 | 0.024 | 0.123 |
| asascore | 0.296 | 0.066 | 0.024 | 0.123 |
| weight | 0.296 | 0.066 | 0.024 | 0.123 |
| T-score | 0.338 | 0.059 | 0.026 | 0.118 |

### 5.8.1 Machine Learning Model Results

### Final model parameters

The model was trained using a 5-fold stratified K-fold cross-validation approach to ensure balanced representation of classes in each fold. The training set consisted of 778 samples, and to address the class imbalance ratio of 5.2:1, class weights of 1:9 were applied to make the model more sensitive to the minority class. For each fold, the test set comprised 185 data points, with 163 negative cases and 32 positive cases, maintaining the original class proportions. The final hyperparameters used are a regularization strength of C=1 leading to a balanced model. An L2 penalty to prevent overfitting and the solver liblinear as this works well with L2 regularization. Table 14 displays the features included in the final model:

**Table 14:** Overview of Features included the final model

| **Feature** | |
|---|---|
| AB | AD |
| AE | AG |
| BE | BMI |
| CE | IS |
| IT | Frustum Volume Cone |
| distance to tumor | Angle 2 |
| Angle 3 | Angle 5 |
| Area pentagon | asascore |
| sex | weight |
| intervention | length |
| age | chemotherapy |
| radiotherapy | therapy |
| procedure | scorecm |

### Evaluation metrics

The main performance metrics are displayed in table 15 based on stratified 5-fold cross-validation. The primary aim of this study was to develop sensitive models, prioritizing the identification of positive cases over minimizing false positives. The target sensitivity score was set at 0.8. The objective was met for the CDC model, which achieved an average sensitivity score of $0.835 \pm 0.102$. The accuracy for the CDC model was $0.392 \pm 0.075$. The CDC model's low specificity of $0.306 \pm 0.097$ reflects the trade-off made to prioritize sensitivity, which typically comes at the cost of reduced specificity. The F1-score reached a value $0.309 \pm 0.025$. Finally, the AUC score of $0.608 \pm 0.03$ for the CDC model suggests it performs only slightly better than random chance when distinguishing between positive and negative cases. Similar to the anastomotic leakage, the model displays a large standard deviation for sensitivity, highlighting inconsistency across folds. On the other hand, the smaller standard deviations for the other metrics indicate a more consistent performance across folds.

**Table 15:** Performance Metrics for the CDC3+ model

| **Measurement** | **Accuracy** | **Sensitivity** | **Specificity** | **F1** | **AUC** |
|---|---|---|---|---|---|
| **Value** | $0.392 \pm 0.075$ | $0.835 \pm 0.102$ | $0.306 \pm 0.097$ | $0.309 \pm 0.025$ | $0.608 \pm 0.031$ |

### Confusion matrix

Figure 23 displays the average confusion matrix of the five folds.

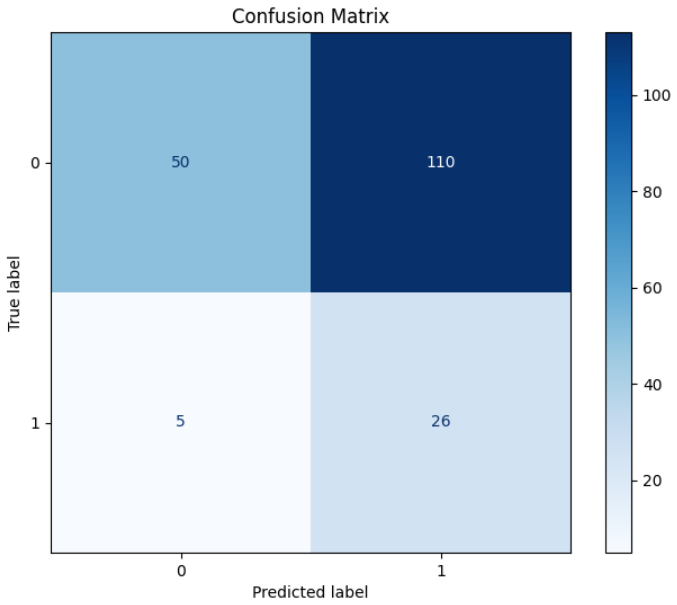



**Figure 23:** Confusion matrix showing the performance of the model. The values represent True Negatives (50), False Positives (110), False Negatives (5) and True Positives (26)

### Prediction Probability Distribution

Figure 24 shows the prediction probability distribution for CDC3+ model.

**Class 0 (Red):** The largest portion of Class 0 predictions falls within the 0.55 to 0.75 probability range, indicating that the model falsely classifies negative cases as positive. However, not with high confidence, demonstrating a lot of uncertainty in the model.

**Class 1 (Green):** The majority of Class 1 cases are also concentrated in the 0.58 to 0.75 range, suggesting that the model has some discriminative ability in identifying positive cases but only with moderate confidence. This moderate range indicates that while the model can separate positive cases from negative ones to some extent, it lacks strong certainty.

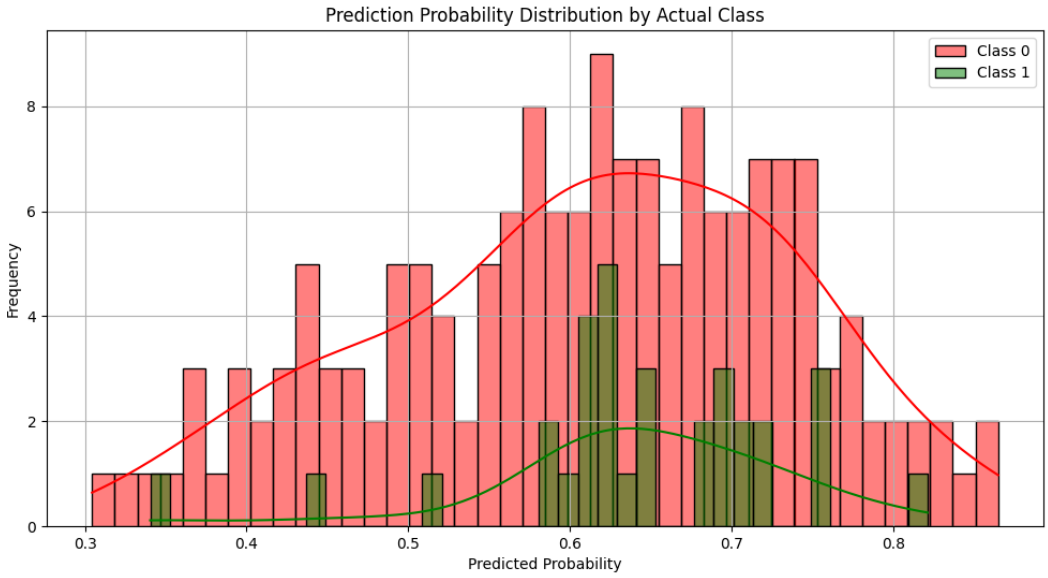



**Figure 24:** Prediction probability distribution with negative cases shown in red (0) and positive cases shown in green (1)

### Receiver-Operating characteristic Curve

Figure 25 displays the ROC curve For the CDC3+ model, the AUC is 0.61, displaying that the models predictive ability is only slightly higher than chance. At a threshold of 0.8 True Positive Rate (TPR), the model exhibits relatively wide confidence interval, indicating a lack of consistency across the folds.

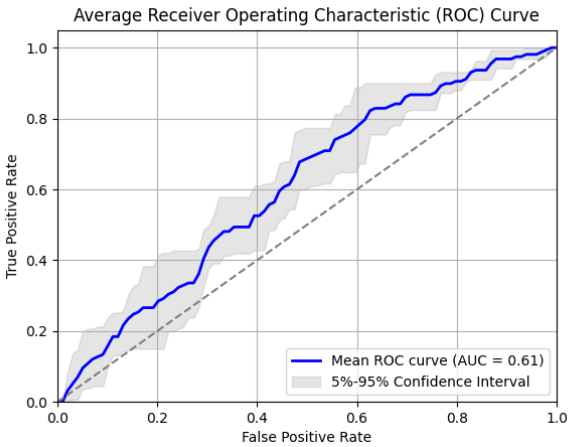



**Figure 25:** ROC curve with False Positive Rate (FPR) on the x-axis and True Positive Rate (TPR) on the y-axis

### Sensitivity vs Specificity

Figure 26 displays the mean sensitivity vs mean specificity of all folds at various thresholds. The target of 0.8 sensitivity, comes down to a threshold of 0.53, resulting in a specificity of 0.4. The intersection of sensitivity and specificity is at a threshold of 0.63, where both sensitivity and specificity have a value of 0.57.

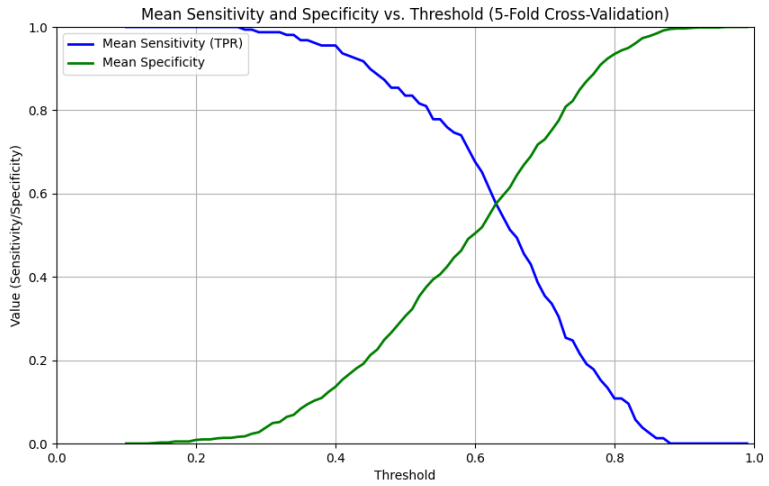



**Figure 26:** Graph visualizing the the mean sensitivity (blue) and specificity (green) at different thresholds

### Feature Importance

Figure 27 displays the feature coefficients for the CDC model, seven of the top 10 features are pelvimetry measurements. All features have a mean coefficient below 2, but some show a high standard deviation (up to 1,73).

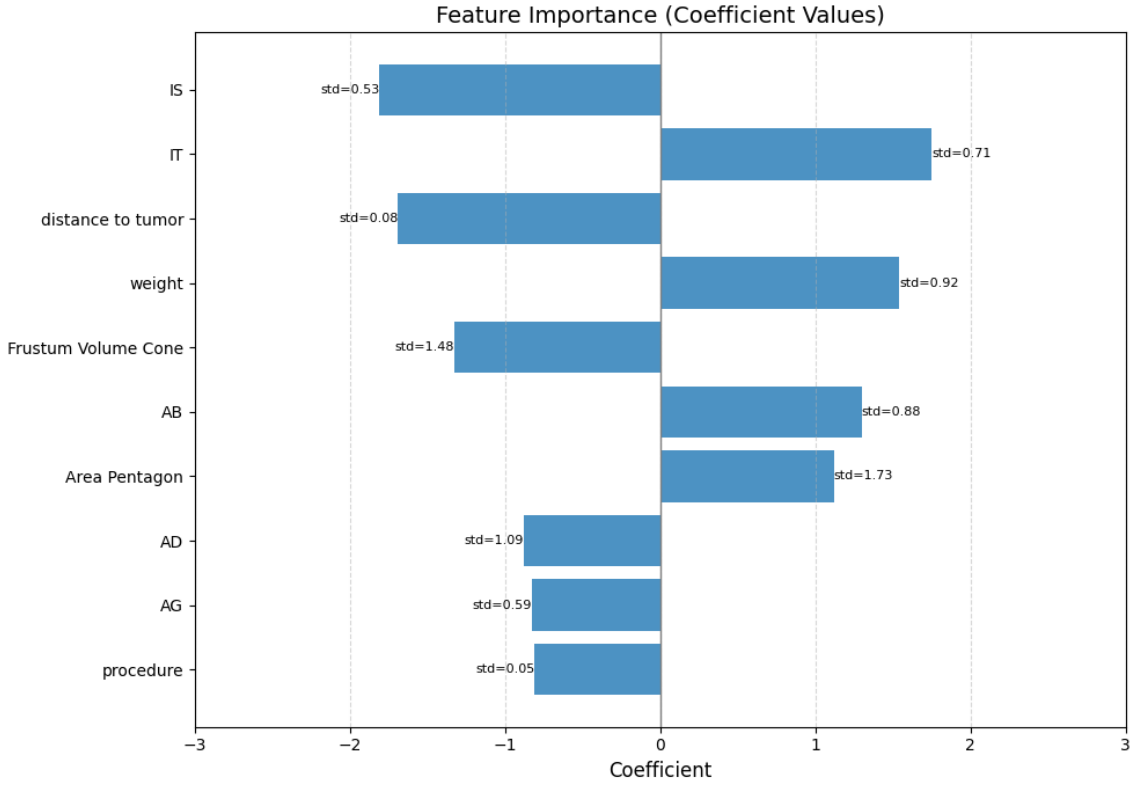



**Figure 27:** Feature Importance in the Models predictions

### External Validation

Table 16 displays the performance of the model on various subsets of the participating hospitals. These vary in sample size and in class balance therefore results vary. The sensitivity varies between 0.67-1.0 and specificity between 0.24-0.44.

**Table 16:** Performance Metrics per Hospital

| Hospital | n | Accuracy | Sensitivity | Specificity | F1 | AUC |
|---|---|---|---|---|---|---|
| 1 | 309 | 0.35 | 0.76 | 0.25 | 0.30 | 0.53 |
| 2 | 230 | 0.37 | 0.81 | 0.31 | 0.22 | 0.51 |
| 3 | 174 | 0.39 | 0.82 | 0.28 | 0.34 | 0.61 |
| 4 | 81 | 0.48 | 0.67 | 0.44 | 0.32 | 0.56 |
| 5 | 70 | 0.43 | 0.71 | 0.40 | 0.20 | 0.54 |
| 6 | 50 | 0.36 | 1.00 | 0.24 | 0.33 | 0.66 |
| 7 | 30 | 0.47 | 0.89 | 0.29 | 0.50 | 0.62 |
| 8 | 29 | 0.31 | 1.00 | 0.26 | 0.17 | 0.67 |

## 5.9 Discussion

The primary aim of this study was to develop a machine learning model to predict complications in patients undergoing Total Mesorectal Excision. By incorporating both clinical and pelvimetry input variables, the model aims to assist surgeons in pre-operative decision-making, enabling more tailored and patient-specific care. Since positive cases are the most critical in clinical settings, a sensitivity of 0.8 was chosen as the target for model development.

### Anastomotic leakage

Our findings showed that we were able to develop a model that achieved 80% sensitivity, demonstrating that the model is moderately effective at detecting positive cases, albeit at the cost of low specificity, as displayed in Figure 18. The low accuracy of $0.528 \pm 0.034$ is a poor indicator of the model's performance due to class imbalance and should therefore be interpreted with caution. With an average AUC of 0.667, the model demonstrates predictive power slightly better than chance. Clinically, an AUC of 0.667 would have limited utility [65]. Moreover, the use of ROC AUC in a model with a strong class imbalance may provide a misleading impression, as a high AUC could be achieved by correctly classifying the majority class while neglecting the minority class. Due to this limitation, the use of Precision-Recall AUC would be more informative, as it prioritizes the model's performance for the minority class, providing better insights into its applicability in a clinical setting [66]. The model exhibited a large standard deviation for sensitivity ($\pm 0.140$), indicating inconsistency across folds and challenges in reliably identifying true positives, which might be attributed to the class imbalance in the dataset. In contrast, the smaller standard deviations for the other metrics suggest more consistent performance across folds.

The univariate correlation identified five of the top 10 features as pelvimetry features. With *Pelvic Inlet* having the largest negative correlation. This is logical as a decrease would probably lead to less operating space for the surgeon, resulting in a higher chance of anastomotic leakage. However, after feature selection, only two of the top 10 features based on highest feature coefficient were pelvimetry features, indicating that when features are combined in a model, clinical parameters hold greater significance. The most important features in the machine learning model included *Angle 1*, which had a large positive coefficient, contradicting literature findings where a smaller *Angle 1* was associated with more difficult surgeries [67]. While these findings are not directly comparable, they are related, making this contradiction relevant. The second important feature is the *distance to tumor* containing a strong negative coefficient. Indicating that as the distance to the tumor reduces, the likelihood of a positive outcome increases. This outcome corresponds with the literature, which mentions that a lower anastomosis corresponds with an increased chance of anastomotic leakage [14, 68].

The dataset was relatively small, consisting of only 525 data points. For a prediction model to perform well and capture meaningful relationships within the data, a larger dataset is generally required. However, such a small dataset might limit generalizability and comparing this with clinically applied prediction models in the literature, the size of our dataset may be too small for clinical implementation. Expanding the dataset will require a lot of time unless more hospitals contribute their data to improve model training.

However, in the literature, there are studies that develop machine learning models that predict surgical complications with smaller datasets yet achieve higher AUCs. For example, Li et al. used a dataset of 322 patients to develop a machine learning model that predicts incisional infection following a right hemi-colectomy, with the best performing model reaching an AUC of 0.885[69]. This suggests that a larger dataset might not necessarily be the answer and that

the predictive features required for optimal performance are not yet included in the dataset or that the outcome we are attempting to predict is to challenging.

The dataset has a class imbalance ratio of 5.5:1, making it highly imbalanced, favoring the majority class (negative cases, 0). To account for the small dataset and class imbalance, SMOTE-ENN was applied to create synthetic samples for the minority class (positive cases, 1). In our anastomotic leakage model, the class distribution for the training set folds change from [negative:355, positive:65] to [negative:178, positive:286], resulting in a new class imbalance of 0.62:1 (negative:positive). This creates a minority-class-dominated balance, which is requested to create a sensitive model. The significant reduction in Class 0 samples is due to the removal of borderline samples close to the minority class. With a $k$ of 3, any Class 0 sample not surrounded by at least three other Class 0 samples is removed. A default value of $k = 3$ is chosen, as this strikes a balance between effectively removing the majority class border samples while preserving its structure and avoiding oversmoothing, which could lead to overfitting for the minority class.

The results of Table 12 display the usage of the model on small hospital subsets. These results demonstrated the models robustness, focusing on sensitivity opposed to specificity. However the results display that the model is fairly robust in terms of sensitivity prioritization over specificity. However, due to the small sample size, these hospitals show extreme variability in their performance.

The results demonstrate, that accurately predicting if a patient will develop anastomotic leakage following TME surgery may not be feasible. However, the models prediction can serve to stimulate surgeons to take additional precautions before surgery. Furthermore, it is important to have mentioned that both univariate correlation and feature importance display that pelvimetry measurements have a predictive factor on the outcome of a patient. Therefore, future practice could include pelvimetry measurements as a standard protocol for each patient, based on MRI.

### Clavien Dindo 3+

The main findings of the model training for Clavien-Dindo grade 3+ were that it is possible to train a model for 80% sensitivity. However, at a cost of low specificity, resulting in a large number of false positives. The standard deviation for the model is low for all metrics except sensitivity $\pm$ 0.102 demonstrating that the model is effective at predicting positive cases and but lacks consistency over each fold. The F1-score of 0.309 indicates that the model has difficulty discriminating between true and false positives, which is often the case in imbalanced datasets.

The prediction probability distribution displayed in Figure 24 follows a Gaussian-like shape, centered slightly to the right of the threshold at 0.5, with a peak around 0.65. The significant overlap between the two classes in the 0.5–0.7 range indicates that the model assigns similar confidence to both classes, highlighting its limited ability to discriminate effectively between them.

In the univariate correlation, the correlation of the features is very low, indicating that there is little correlation between the predictor variables and CDC3+. This is confirmed again in the feature coefficients of the feature predictions in the logistic regression model, where the coefficients are really low again, sometimes varying with high standard deviations, displaying that over multiple folds, a consistent set of features have a strong influence on the prediction, while in others there impact is low. Displaying that its predictive power is dependent of the variance in the subset.

For this model, a class weighting of 1:9 is applied, while the class imbalance is 5.2:1. This weighting is implemented to create a more sensitivity-focused model. However, overweighting beyond the natural imbalance might lead to reduced generalizability, as it can cause the model to overfit on the minority class. Class weighting of 1:9 was selected during the model sweep to create a slightly larger class weighting than the natural ratio. In future research, the option to explore more specific class imbalance ratios, such as 1:8 or 1:8.5, could be considered to determine if this leads to improved results.

The challenge in predicting Clavien-Dindo grade 3+ lies in its nature as a universal classification system. It serves as a collective term encompassing diverse complications, such as a pulmonary embolism, which can be classified as grade 3+ even in the context of TME surgery without a direct causal link. This broad scope complicates the establishment of clear causal relationships and makes accurate prediction more difficult. This difficulty is also reflected in the performance metrics. Although the model achieves 80% sensitivity, it incorrectly predicts 70% of the negative cases as false positives, demonstrating very little discriminative power. What these results will add to the future is difficult to assess, probably that it is very difficult to predict a universal grading in a surgery. Therefore, creating surgery specific grading for TME might be relevant. A difficult surgery can be defined by the Escal grading [70]. This however also includes the Clavien Dindo grading, thus a different form of grading needs to be invented.

The results of Table 16 display the usage of the model on small hospital subsets. As mentioned for the anastomotic leakage subsets, the sample sizes are again small. However, the robustness of the model is demonstrated by the sensitivity for each model, which remains at the target level of around 0.67-1.0. Nevertheless, due to the low specificity (0.24–0.44), this model is not clinically applicable, as it offers almost no distinction between false positives and true positives.

### Pre-processing

Patients with a tumor distance of 14 cm or less are included in the dataset, as this distance includes only tumors defined as rectal cancer. However, literature states that tumors at a distance of 15 cm or less are classified as a rectal cancer [71]. If a distance of 15 cm or less was chosen, the final dataset for CDC would include 1006 patients instead of 973 and for anastomotic leakage, the dataset would increase to 556 from 525. For both outcomes, this expansion of the dataset could potentially have led to improved performance.

During pre-processing, the distances of all possible options between the five points were calculated as input for the training of the model and feature selection. However, distances such as *AD* and *CE* were also calculated, which are not clinically defined distances. This could make the interpretability of the model more difficult. Furthermore, the variable *prethp* was added to the model to indicate whether earlier surgery, chemoradiation, or radiotherapy took place before the surgery. This binary variable adds extra noise to the model, as this information is already described in *chemotherapy*, *radiotherapy*, and the *abdomen* variable. Therefore, the feature *prethp* is unnecessary and could potentially lead to multicollinearity.

One factor that might have influenced the results and feature selection is that the feature *scorecm* (M-staging of the tumor pre-operative) had three options instead of two "No metastases (0)" or "Metastases (1)". Among the 973 patients, 42 contained a value of "unknown (9)". This might have added noise to the model or potentially hindered the training process. Additionally, this feature is not scaled to have values between 0 and 1, the magnitude of 9 might influence the training, compared to the other scaled features

During pre-processing, the intertuberous distance had 22.5% missing values, therefore to combat this we used KNN-imputation using 5 nearest neighbours to keep the results robust, however

22.5% is a large percentage but still feasible based on literature, as in general imputation up to 50% improves the classification[72].

### Pipeline

During feature selection, we employed forward and backward feature selection using the F1 score as the scoring metric. While this approach creates the most balanced model, our ultimate goal is not to achieve balance but to develop a model with a sensitivity of 0.8. Therefore, a custom scoring metric would be useful for determining when the feature selection process should stop iterating, specifically when the sensitivity of 0.8 is reached. This approach would result in a more sensitive model, which is critical because identifying positive cases is of the most important. Another option would be to use Fbeta, which is used in imbalanced binary classification, where a custom weight can be applied to either precision or recall, thus improving the sensitivity of the model [73]. Although false positives are not ideal, they are acceptable in this context: if a patient is predicted to have complications but ends up having none, this is preferable to missing a patient who does have complications. Moreover, it is not realistic to expect a model to perfectly predict complications. The primary goal is a high sensitivity to prompt the surgeon to pay extra attention to certain patients, thereby improving patient tailored care.

At the moment, the value for early stopping during feature selection is set at 10. This value is currently chosen based on intuition. If there is no improvement within 10 iterations, the stopping criterion could be set to a higher value; however, doing so would significantly increase computation time, but this could improve the results of a higher sensitivity.

Currently, a value of 10 is chosen for VIF. However, lowering this value to five is common practice, resulting in fewer features to train the final model, reducing multicollinearity between predictor features [74]. This also improves the interpretability of the model, as having fewer features makes the model simpler and easier to interpret. However, this could potentially result in the loss of relevant information. In our case, this is not the most critical aspect, as our primary goal is not to have a highly interpretable model but rather to focus on the probability of the model predicting a complication. A high probability on positive cases would prompt the surgeon to pay closer attention to additional patient characteristics and base its decision on that.

During hyper parameter tuning, an important step that is missed is that, for the logistic regression model, only one solver 'liblinear' is available during hyper parameter tuning. This is a mistake, as the choice of the solver chosen could influence the performance metrics. Therefore ensuring a broader spectrum of solvers could improve these metrics and the model.

### General

One limitation of the univariate correlation analysis is that it combines two different correlation measurements—Point-Biserial correlation for continuous features and Cramér's V for categorical features—to create a ranked list based on correlation values. This approach provides an indication of the strength of the relationship between each feature and the target variable, but these two metrics have inherently different scales and interpretations. Therefore, this should be taken into account when analyzing the univariate correlations.

### Future works

*Feature engineering and Refinement*
Currently, no polynomial features were applied, as this would strongly increase the number of features and potentially capture new relationships between the predictor features and the target variable. However, this would come at the cost of significantly increased computation time due to both forward and backward feature selection. Alternatively, this approach could be applied to the selected features and limited to the pelvimetry measurements.

The *Frustum Volume/Pelvic Depth ratio*, together with the *Pelvic Depth*, had a significant p-value for both anastomotic leakage and CDC3+, ranking both in the top 10 of highest correlation for both outcomes. To further investigate this feature, the Frustum can be refined by segmenting the volume of the pelvic cavity to create a more accurate prediction.

*New Features*
The addition of new features, such as surgical expertise, could add significant value. However, this poses a challenge, as institutions like the Meander Medical Center are teaching hospitals, where an attending surgeon might perform certain parts of the procedure while a resident handles simpler tasks. This situation can also occur in other institutions. Therefore, this feature would only be valid if a single surgeon performed the entire operation.

Furthermore, it might be interesting to predict the operation time using regression. However, it is crucial to ensure that the start and end points of the surgery, such as the closure of the body, are clearly defined. Additionally, it is essential that the start times are followed with precision to maintain accuracy in the predictions.

*Model calibration*
Something important to consider, as probabilities are crucial in our analysis, is the calibration of the model. Calibration can potentially improve the accuracy of the predicted probabilities. However, due to the model's unbalanced performance, it is possible that the calibration process may result in a specificity-focused model.

## 5.10 Conclusion

This study aimed to develop a machine learning model capable of accurately predicting complications for patients undergoing Total Mesorectal Excision surgery. Two targets were selected: anastomotic leakage and Clavien-Dindo grade 3+. The aim was to create sensitive models capable of detecting at least 80% of positive cases. This target was nearly achieved for both models, with sensitivity values of 0.765 for anastomotic leakage and 0.835 for Clavien-Dindo grade III+, ensuring that high-risk patients could be effectively identified.

Feature importance analysis revealed that, for anastomotic leakage, both an increased Angle 1 and smaller distance to tumor had strong coefficients. Notably, Angle 1 differs from existing literature, while distance to tumor corresponds with previous findings. For Clavien-Dindo grade III+, the feature coefficients were low for variables with large standard deviations, highlighting the difficulty in assessing certain predictors to establish a universal grading system. Overall, this study advances the application of machine learning in predicting surgical complications in Total Mesorectal Excision surgery. In the future, achieving a prediction for CDC3+ may prove unattainable, and if a grading system for surgical difficulty were to be developed, it would need to be specific to TME surgery. Research on predicting anastomotic leakage should focus on developing new features, such as surgical expertise, and expanding the dataset by including more hospitals, potentially extending beyond The Netherlands

Despite these results, the study has several limitations. First, the size of the dataset was limited, with 973 data points for Clavien-Dindo grade III+ and 525 for anastomotic leakage. This was further hindered by the incorrect selection of the filter for tumor distance, set at 14 cm instead of 15 cm. Additionally, no custom scoring metric was applied for feature selection, resulting in the selection of features based on a balanced model rather than a sensitivity-focused model. Finally, due to class imbalance, model training was hindered, necessitating the use of methods such as SMOTE and class weighting to address these challenges. Nevertheless, this study underscores the potential of applying machine learning to predict surgical outcomes by combining pelvimetry measurements and clinical variables. It lays the foundation for developing patient-tailored care plans to be implemented by surgeons.

# Sacral Curve Model

## 6.1 Introduction

During Total Mesorectal Excision, the pelvic cavity presents a narrow constraint, which influences the surgeon's ability to properly operate and create a good anastomosis when performing a Low Anterior Resection [6, 75]. Furthermore, there are problems where the sacral curve obstructs the surgeon's movement due to the curvature of the sacrum, especially in the deep part of the pelvis where the rectum is present. This obstruction can create difficulty in the precise dissection along the "holy plane" – the optimal dissection plane, where the rectum is removed along with the mesorectum [76]. Min Soo et al. mentioned that when there is a deep sacral curve, surgeons tend to cut more in an oblique manner, departing from the ideal dissection plane, possibly leaving behind cancerous tissue [77]. However, other literature contradicts this statement and mentions that the sacral length and depth do not significantly influence the surgical difficulty [78]. Displaying that there has been a lack of research about the influence of the sacral curve on surgical difficulty and complications without a consensus being reached.

Therefore, the aim of this study was to develop an algorithm that automatically calculates the distance of the sacral curve based on coordinates and sagittal T2-MRI scans. These results are then quantified and combined with clinical parameters to predict the influence of the sacral curve on the difficulty of a patient after TME.

## 6.2 Method

During this study, two deep learning models were initially trained using the Dice score. The best-performing model was then applied to develop sacral curves based on sacrum segmentations. Following this, parameter extraction was performed and a statistical analysis was conducted to determine the correlation between individual parameters and anastomotic leakage.

### Study Population

The dataset consists of 1,707 preoperative sagittal T2-MRI scans, sourced from eight TME centers that perform more than 40 TME surgeries annually. All MRIs were acquired using 1.5 or 3.0-Tesla MRI scanners, where the complete bony pelvis was visible and the scans were artifact-free. The final statistical analysis included data from a total of 390 patients after applying all in- and exclusion criteria. The dataset included patients based on specific criteria: only those who underwent resection treatment intended as a curative operation were considered. The resection must have involved TME for the primary tumor and had to be an elective, non-emergency surgery. Only patients who had LAR or APR procedures performed laparoscopically or robotically were included. Additionally, a preoperative pelvic MRI, with complete visualization of the bony pelvis, was necessary for comprehensive pelvimetry measurements. Patients were excluded based on the following criteria: any prior Transanal Endoscopic Microsurgery (TEM) before resection. Tumors were required to be within 14 cm of the Anal Rectal Junction (ARJ).

### Pre-Processing

Variations in MRI parameters across the scans required for pre-processing to be performed ahead of model training. Therefore, all MRI scans were resampled to a uniform voxel spacing. The dataset for model training was created by manually labeling the sacrum in the MRI scans using 3D Slicer [79], resulting in a binary segmentation of the sacrum. This labeled dataset served as a ground truth dataset for a MONAI-based 3D U-Net, which is designed for medical image segmentation tasks [80]. First, the MRI scan and its corresponding label underwent transformations, including normalization where voxel values were scaled between 0 and 1. This was done to combat different intensities across varying MRI scans. Furthermore, random cropping was applied to select only the region of interest of the label. This reduced computational load and improved the algorithm's robustness, preventing overfitting. The final step involved converting the MRI and the label into tensors for compatibility with the 3D U-Net.

### Model Architecture

The input data for the model consists of 3D MRIs with voxels resampled to a uniform spacing to ensure model functionality. The input layers has one input channel, indicating the scan requires to be in greyscale. Following this are five feature map layers, where each layer has an increasing number of feature maps to capture progressively complex features. The model output consists of two channels, resulting in a binary image. In this image, background voxels are black (0), and foreground voxels, representing the sacrum segmentation, are white (1). Batch normalization is used throughout the network.

### Model Training

The model was trained using an 80-20 data split, with 80% of the data allocated for training and 20% for validation. Training was conducted over a maximum of 600 epochs, with early stopping applied if there was no improvement in validation loss for 50 consecutive epochs. A minimum improvement of 0.001 was required. The training process stopped at epoch 221, achieving a final training loss and validation loss of 0.22. The model utilized a batch size of 32, a learning rate of 0.0001, maximum channels set at 256, and a patch size of $224 \times 224 \times 16$. Training was performed on a GPU. For the optimizer, an Adam optimizer was used. The best model was selected based on the lowest validation loss.

### Model Evaluation

The model was then tested on an additional test set of 52 patients, with the results shown in Table 17. The Dice coefficient was used as the primary evaluation metric to assess the model's performance. The Dice coefficient is used to compare a predicted segmentation and its corresponding ground truth label—in our case, the manual segmentation. In the Dice score, 0 indicates no overlap and 1 indicates complete overlap. The Dice coefficient is calculated using Equation 5:

$$D = \frac{2|A \cap B|}{|A| + |B|} \tag{5}$$

Due to the elongated structure of the sacrum, an extra evaluation metric was applied, called the centerline Dice (clDice) [81], to quantify the segmentation quality. For this, the skeletons $S_P$ and $S_L$ were acquired from the ground truth ($V_L$) and predicted segmentation ($V_P$) using morphological operations. Using these, the fraction of the predicted segmentation skeleton ($S_P$) that lies within the ground truth segmentation ($V_L$) was calculated, referred to as the

topology precision (Tprec). Similarly, the topology sensitivity (Tsens) was calculated as the fraction of the ground truth segmentation skeleton ($S_L$) lying within the predicted segmentation ($V_P$). Using these values, the clDice is defined as the harmonic mean of these measurements, as displayed in Equation 7:

$$\text{Tprec}(S_P, V_L) = \frac{|S_P \cap V_L|}{|S_P|} \qquad \text{Tsens}(S_L, V_P) = \frac{|S_L \cap V_P|}{|S_L|} \tag{6}$$

$$\text{clDice}(V_P, V_L) = 2 \times \frac{\text{Tprec}(S_P, V_L) \times \text{Tsens}(S_L, V_P)}{\text{Tprec}(S_P, V_L) + \text{Tsens}(S_L, V_P)} \tag{7}$$

These measurements are used to evaluate the segmentation. However, to evaluate the final sacral curve, the Hausdorff distance is utilized [82]. This calculates the maximum distance between the predicted line segment and the manually drawn line, which serves as the ground truth. Lastly, the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) of the curve length are calculated [83].

### Parameter Extraction

The trained model was applied to 1707 patients to extract various parameters, with the sacral curve length being the most important. The coordinates of points A to C were loaded and resampled from uniform spacing to their original spacing. Additionally, the binary segmentation output of the model was resampled to the original spacing. Using the coordinates of points A and C as the start and end points, the Dijkstra 3D algorithm was employed to calculate the shortest distance along the sacral curve [84].

Before calculation of the distance, the coordinates A and C are verified if they are located in the same z-slice, as this could influence the path of the distance calculation. The input for this distance calculation was a distance map of the sacrum, where each voxel value represented its distance to the nearest boundary of the segmentation, serving as weights for the algorithm. To prevent zig-zagging, line smoothing is applied.

Additional parameters were computed, including the length of the sacral curve and the lengths of its segments (AB, BC). Furthermore, the ratio between the length of the sacral curve ($L_{AC}$) and the direct distance ($D_{AC}$) are calculated using equation 8, these lines are displayed in Figure 28. This is also performed for the segments AB and BC. Additionally, the bending energy (the amount of bending a curve is subjected to) is calculated using equation 9 [85], the maximum derivative in the curve and segments, and the position of the maximum derivative along the sacral curve and segments were analyzed.

$$\text{Ratio}_{AC} = \frac{L_{AC}}{D_{AC}}, \quad \text{Ratio}_{AB} = \frac{L_{AB}}{D_{AB}}, \quad \text{Ratio}_{BC} = \frac{L_{BC}}{D_{BC}} \tag{8}$$

$$E_b = \int \kappa^2 \, ds \tag{9}$$

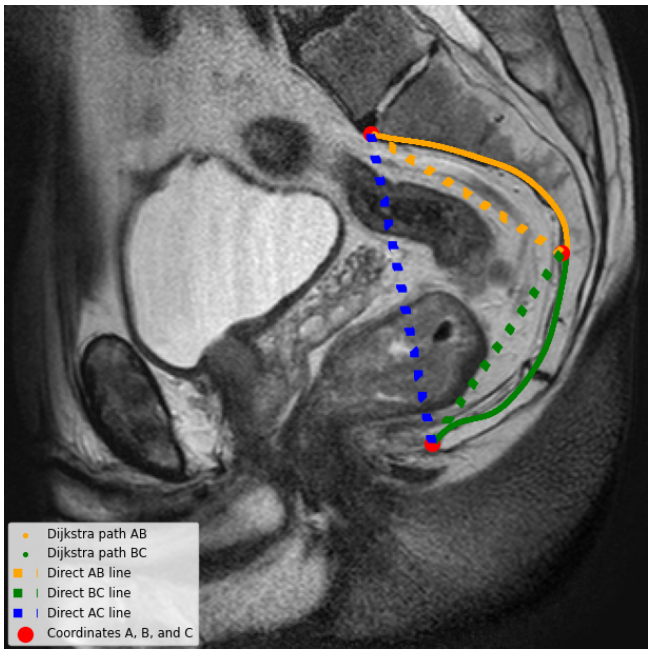
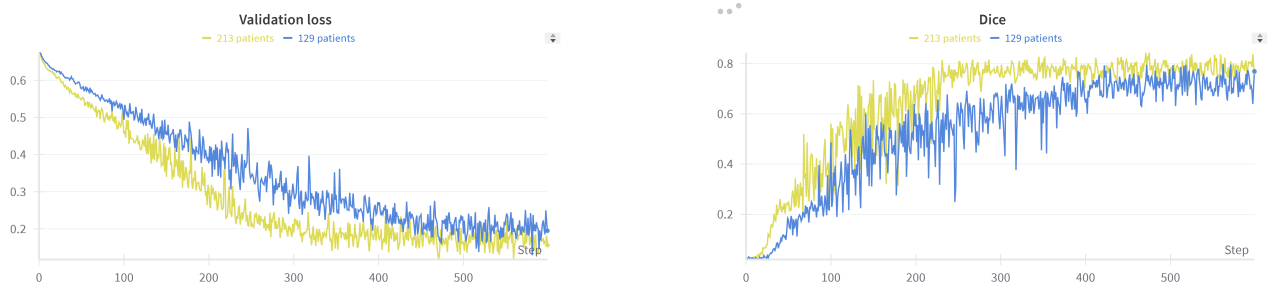



**Figure 28:** Illustration of line segments (AB, BC, AC) and Dijkstra paths of AB and BC

## 6.3 Results

### Model evaluation

Two models were trained, one with 129 patients and the other on 213 patients. Both models were trained for 600 epochs. The validation and training loss and dice are displayed in Figure 29. The model with the lowest validation loss was selected. Model 1 is denoted as the model with 129 patients and model 2 is noted as the model with 213 patients.




**(a)** Validation Loss

**(b)** Dice Score

**Figure 29:** Comparison of (a) Validation Loss and (b) Dice Score over training steps for datasets with 213 patients (yellow) and 129 patients (blue).

The performance metrics of both models are displayed in Table 17. Model 1 reached a Dice score of 0.82 (± 0.22), while Model 2 achieved a Dice score of 0.85 (± 0.19) with a smaller standard deviation. Similarly, Model 2 also achieved a higher clDice score of 0.89 (± 0.17), compared to Model 1's score of 0.87 (± 0.19). For the mean Hausdorff distance, the results of Model 2 (13.13 mm) compared to Model 1 (13.98 mm). Model 2 also outperformed Model 1 in both MSE and RMSE. Displaying more reliable and accurate predictions. Therefore model 2 is used for sacral curve determination.

**Table 17:** Comparison of Metrics for Model 1 and Model 2

| Metric | Model 1 | Model 2 |
|---|---|---|
| Dice | 0.82 (± 0.22) | 0.85 (± 0.19) |
| clDice | 0.87 (± 0.19) | 0.89 (± 0.17) |
| Mean Hausdorff Distance (curve) (mm) | 13.98 | 13.13 |
| Mean Squared Error (curvelengths) ($mm^2$) | 195.63 | 166.67 |
| Root Mean Squared Error (curvelengths) (mm) | 13.99 | 12.91 |

### Sacral Curves

Model 2 is applied to create sacral curves for 1707 patients. 1205 out of the 1707 sacral curves are included based on the accuracy of the segmentation verified through visual inspection. Two examples of sacral curves are displayed in Figure 30.

**(a)** Correct segmentation and sacral curve detection

**(b)** Incorrect segmentation resulting in an abnormal sacral curve

**Figure 30:** Comparison of sacral curve detection results: (a) shows a correct segmentation and sacral curve detection, while (b) demonstrates incorrect segmentation leading to an abnormal sacral curve.

### Univariate correlation

For the statistical analysis of the parameters with anastomotic leakage, a total of 390 patients out of the 1205 patients with an accurate segmentation were included after applying inclusion and exclusion criteria. For each feature, the p-value, correlation, and 95% confidence interval (CI) for the correlation were calculated. The results are displayed in Table 18. The analysis showed that only the ratio AB had a significant p-value. Furthermore, the correlations were all extremely low, with only the ratio AB (0.14) and BC (-0.10) showing the highest, although weak, correlations. All other parameters exhibited negligible correlations with anastomotic leakage.

**Table 18:** Univariate Correlation Analysis for Anastomotic Leakage

| Variable | Mean | Std Dev | P-value | Correlation | CI Lower | CI Upper |
|---|---|---|---|---|---|---|
| Curve length *AC* | 166.69 | 15.44 | 0.51 | -0.04 | -0.148 | 0.074 |
| Curve length *AB* | 90.05 | 8.09 | 0.58 | -0.03 | -0.142 | 0.080 |
| Curve length *BC* | 76.44 | 10.03 | 0.51 | -0.04 | -0.148 | 0.074 |
| Ratio *AC* | 1.33 | 0.09 | 0.35 | 0.05 | 0.032 | 0.250 |
| Ratio *AB* | 1.07 | 0.05 | **0.01** | **0.14** | -0.211 | 0.009 |
| Ratio *BC* | 1.13 | 0.07 | 0.07 | **-0.10** | -0.115 | 0.107 |
| Area *AC* | 3805.84 | 747.44 | 0.94 | -0.00 | -0.081 | 0.141 |
| Area *AB* | 1221.30 | 349.84 | 0.59 | 0.03 | -0.088 | 0.134 |
| Area *BC* | 1215.99 | 349.26 | 0.68 | 0.02 | -0.136 | 0.086 |
| Max. der. *AC* | 10915.51 | 44263.09 | 0.66 | -0.03 | -0.144 | 0.078 |
| Max. der. *AB* | 178.24 | 1096.50 | 0.56 | -0.03 | -0.135 | 0.087 |
| Max. der. *BC* | 10813.89 | 44276.20 | 0.67 | -0.02 | -0.163 | 0.059 |
| %Pos. max. der. *AC* | 70.11 | 9.40 | 0.35 | -0.05 | -0.061 | 0.161 |
| %Pos. max. der. *AB* | 89.92 | 24.32 | 0.37 | 0.05 | -0.170 | 0.051 |
| %Pos. max. der. *BC* | 36.29 | 17.77 | 0.29 | -0.06 | -0.110 | 0.113 |
| Bending En. *AC* | 0.49 | 0.67 | 0.98 | 0.00 | -0.077 | 0.145 |
| Bending En. *AB* | 0.16 | 0.33 | 0.54 | 0.03 | -0.132 | 0.090 |
| Bending En. *BC* | 0.33 | 0.49 | 0.71 | -0.02 | -0.109 | 0.114 |
| Mean Bending En. *AC* | 0.00 | 0.00 | 0.96 | 0.00 | -0.109 | 0.114 |
| Mean Bending En. *AB* | 0.00 | 0.00 | 0.72 | 0.02 | -0.091 | 0.131 |
| Mean Bending En. *BC* | 0.00 | 0.00 | 0.81 | -0.01 | -0.125 | 0.097 |

## 6.4 Discussion

The aim of this study was to develop a deep learning model to measure the distance of the sacral curve and evaluate its correlation with the surgical outcome, anastomotic leakage, following Total Mesorectal Excision during Low Anterior Resection surgery. After evaluating the various parameters, it became clear that there is a lack of evidence supporting a direct link between anastomotic leakage and the sacral curve. The discovered parameter, Ratio *AB*, although it has a significant p-value, displays a low correlation with the target outcome.

### Evaluation of results

During the training of the models, it is evident that neither model overfits, as the validation loss consistently decreases. Both the Dice score and validation loss are better for Model 2 compared to Model 1. Model 2 achieves a Dice score of 0.85, which is 0.03 higher than that of Model 1. While this is an acceptable result, the Dice score does not directly represent the accuracy of the sacral curve distance. Model 2 outperformed Model 1 across all metrics, including clDice, Mean Hausdorff Distance, MSE, and RMSE. Therefore, it can be concluded that Model 2 is the best-performing model. Increasing the amount of data the model is trained on would likely improve line detection, as demonstrated by the improvement in sacrum segmentation observed when the training dataset increased from 123 to 213 patients.

The expectation for the parameters was that a strongly curved sacral curve would prove difficulty and therefore that have both a high correlation with anastomotic leakage, especially

the lower part of the sacral curve, due to the surgeons inability to manouver the laparoscopic tools. The univariate correlation analysis based on Model 2 reveals that only 1 out of 21 features has a significant p-value (Ratio $AB$). Additionally, only Ratio $AB$ (0.14) and Ratio $BC$ (-0.10) demonstrate correlations above 0.1 or below -0.1, which are still weak. An increase in Ratio $AB$ would mean that the sacral curve segment AB increases, opposed to the direct line distance $AB$. This would result in extra curvature of the upper part of the sacrum. Ratio $BC$ has a negative correlation, meaning the opposite—that if the distance of $L_{BC}$ decreases, a higher chance of anastomotic leakage is happening, meaning that a flatter end of the sacral curve results in increased chance of anastomotic leakage. Displaying the complete opposite of the expectations.

The goal of creating an automatic determination of the sacral curve has been achieved. However, for 502 patients, the model was unable to generate a realistic sacral curve, indicating that the model is not yet robust enough to create reliable sacral curves in all cases. Therefore, achieving a Dice score higher than 0.85, preferably with a standard deviation of less than 0.19, is necessary to improve the model's robustness and reliability.

### Limitations

This study has several limitations. First, the model was trained on only 123 and 213 patients. Increasing the dataset would likely result in a higher Dice score, as is clearly shown by the improvement observed when training on 213 patients instead of 123. To expand the dataset, manual labeling could be employed, or visually confirmed sacral curves could be used as input labels for further training, focusing on direct sacral curve segmentation instead of sacrum segmentation and calculating the sacral curve using the Dijkstra algorithm.

Currently, the Dice coefficient is used as a loss function during training as well as when evaluating the models. Although the Dice score is feasible for training and validation, it does not capture all aspects when assessing the sacrum. It is most important that the sacrum is completely connected from the promontory to the coccyx. A good sacral curve can still be created even when the Dice score is not optimal. A penalty should be introduced to optimize results when a segmentation has gaps, which can lead to incorrect sacral curves. For all patients, validation of the line took place in 2D. There are a few patients where the line was created in multiple planes. Ideally, the validation for these sacral curve lines would also take place in 3D.

In this study, patients with rectal cancer are defined as those with tumors 14 cm or less from the ARJ. However, the literature indicates that most surgeons consider this distance to be 15 cm. In the clinical dataset of 2,773 patients used for the machine learning model, this discrepancy accounts for a difference of 265 patients. Including this additional criterion in the sacral curve study would result in 20 more patients being included in the univariate correlation analysis.

### Future works

In future research, a different loss function needs to be implemented. Currently, Dice loss is used for optimal segmentation. However, since we are dealing with a longitudinal structure, it may be more beneficial to use SoftclDice loss, which is specifically designed for small and elongated structures, such as vessels or nerves. This change would likely ensure that the segmentation remains continuous and intact. To increase the dataset, both manual labeling can be expanded, and data augmentation can be applied to make the model more robust and to further expand the dataset. Furthermore, more post-processing steps could be incorporated. Even with an incomplete segmentation, where a gap in the sacrum is present, the sacral curve could still be calculated, leading to a higher availability of patients. Currently, the analysis is limited to

univariate correlation. However, utilizing polynomial features in the future might uncover new relationships and enhance correlations, potentially stronger than those currently identified with anastomotic leakage.

## 6.5 Conclusion

This study aimed to develop a deep learning model to automatically measure the length of the sacral curve and analyze its relationship with anastomotic leakage in patients following TME surgery. A 3D U-net was trained on two datasets containing T2-sagittal MRI scans; one model was trained on 123 patients, and the other on 213. The larger dataset of 213 patients produced the best results, with a Dice score of 0.85, a mean Hausdorff distance of 13.13, a MSE of 166.67 and a RMSE of 12.91. The Dice score of 0.85 indicates acceptable performance. Combined with a 70.5% hit rate for accurate line creation, these results suggest that the model development was successful.

Statistical analysis, based on 390 patients across 21 variables, revealed that only Ratio $AB$ had a significant p-value of 0.01. While Ratio $AB$ (0.14) and Ratio $BC$ (-0.10) displayed the strongest correlations with anastomotic leakage, there is a lack of correlation between the sacral curve and anastomotic leakage. These results display that it is difficult to create an interpretable measure for the curvature of the sacral curve.

A key limitation of this study was the lack of data for model training and statistical analysis. Future research should prioritize developing new features for the machine learning model, unrelated to the sacral curve, which has shown limited usefulness in predicting anastomotic leakage.

# Dashboard

The prediction model will be deployed in a real-time dashboard to enhance surgical decision-making. A pipeline is developed that integrates multiple artificial intelligence models, data processing steps, and real-time visualization in an easy-to-use, proof-of-concept dashboard. This dashboard aims to reduce post-operative complications and enable patient-tailored care. The final product features a drag-and-drop system for importing MRI scans, a standardized input dashboard for the surgeon to enter clinical parameters, automatic integration of clinical parameters with pelvimetry data and a final output visualization that presents a risk analysis for the patient.

## 7.1 Materials

### Dash

The dashboard is developed using Dash, a Python-based open-source framework for building interactive dashboards [86]. Dash was chosen due to several advantages: Ease of implementation: Integration with Python-based AI models is intuitive, Compatibility: Works with popular Python libraries such as Pandas, Plotly, and NumPy, Interactivity: Allows interactive features, such as buttons, for an enhanced user experience, Customization: Supports easy customization using HTML and CSS and Local deployment: Operates on a local host, eliminating the need for external databases.

### Flask

Flask, a lightweight Web Server Gateway Interface (WSGI) web application framework, is used in combination with Dash[87]. Within this pipeline, Flask acts as a web server hosting each application while continuously monitoring folders for new data, enabling real-time processing. The pipeline operates with five parallel web servers, tracking multiple folders.

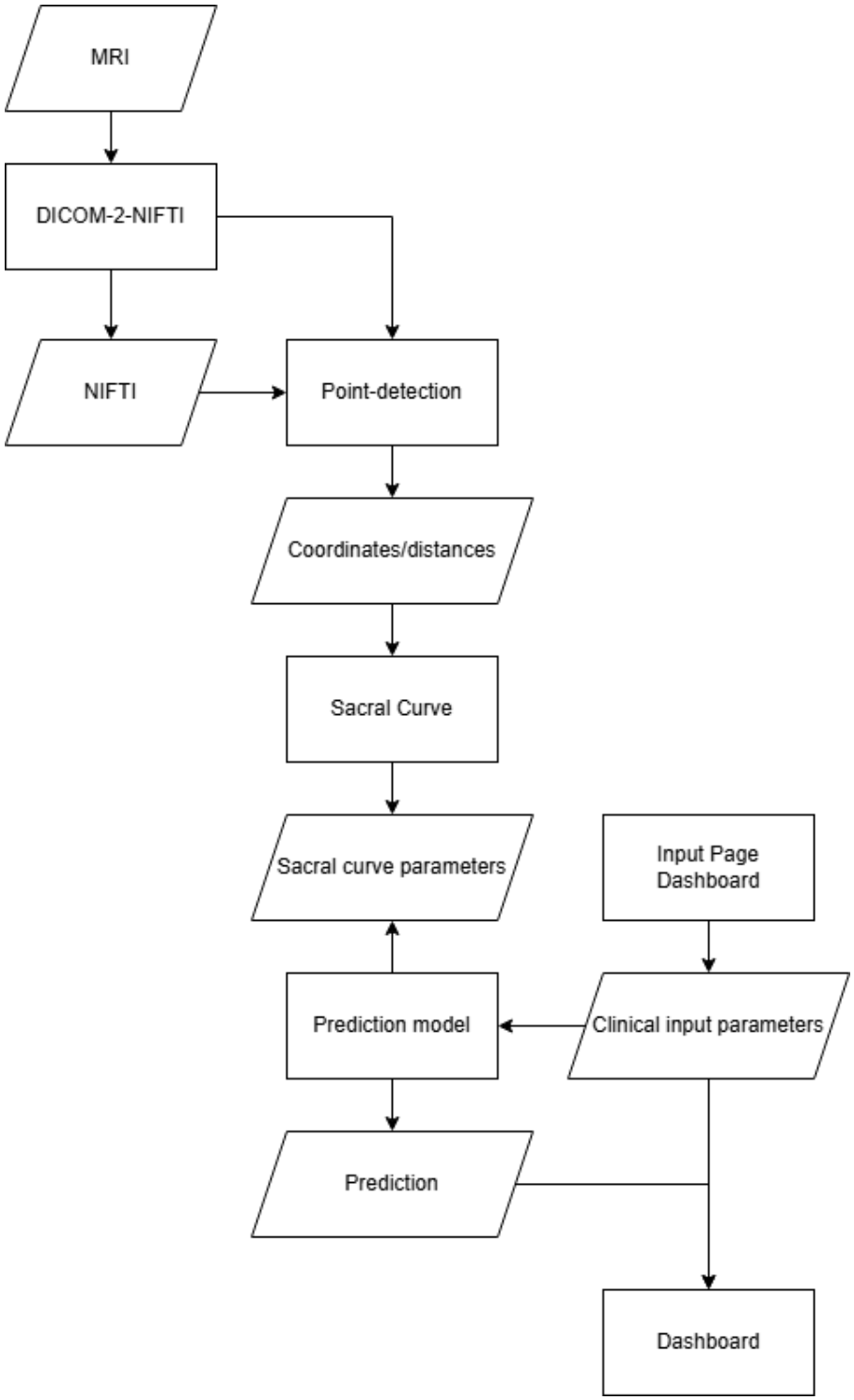



**Figure 31:** Flowchart of the pipeline illustrating all components and input values utilized within the process

## 7.2 Pipeline Components

A pipeline is created to streamline the process of importing medical images and the implementation of the different AI models. In each component Flask is used for the constant monitoring of input folders to streamline the process. An overview of the pipeline's structure is illustrated in the flowchart shown in Figure 31. The pipeline consists of the following components:

1. **DICOM to NIfTI Conversion:** DICOM files are converted to NIfTI format to meet the input requirements of the first deep learning model.
2. **Point Detection Model:** This step uses a pre-trained deep learning model to detect five points and calculates distances between these points. The model resamples images to meet spacing requirements and restores real-world dimensions for accurate measurements. The outputs include coordinates, distances and the resampled NIfTI file.
3. **Sacral Curve Model:** Detected points A and C are imported and the sacral curve model of chapter 6 is utilized to calculate the sacral curve distance using sacrum segmentation and the Dijkstra algorithm. Other pelvic parameters are extracted as well explained in chapter 6 additional pelvic parameters. These parameters are stored in a data frame for subsequent steps.
4. **Dashboard Input Page:** Dash is used to create the first dashboard page, where surgeons input patient clinical variables. The data is exported as a data frame for use in the prediction models. The input page is visualized in Figure 32.

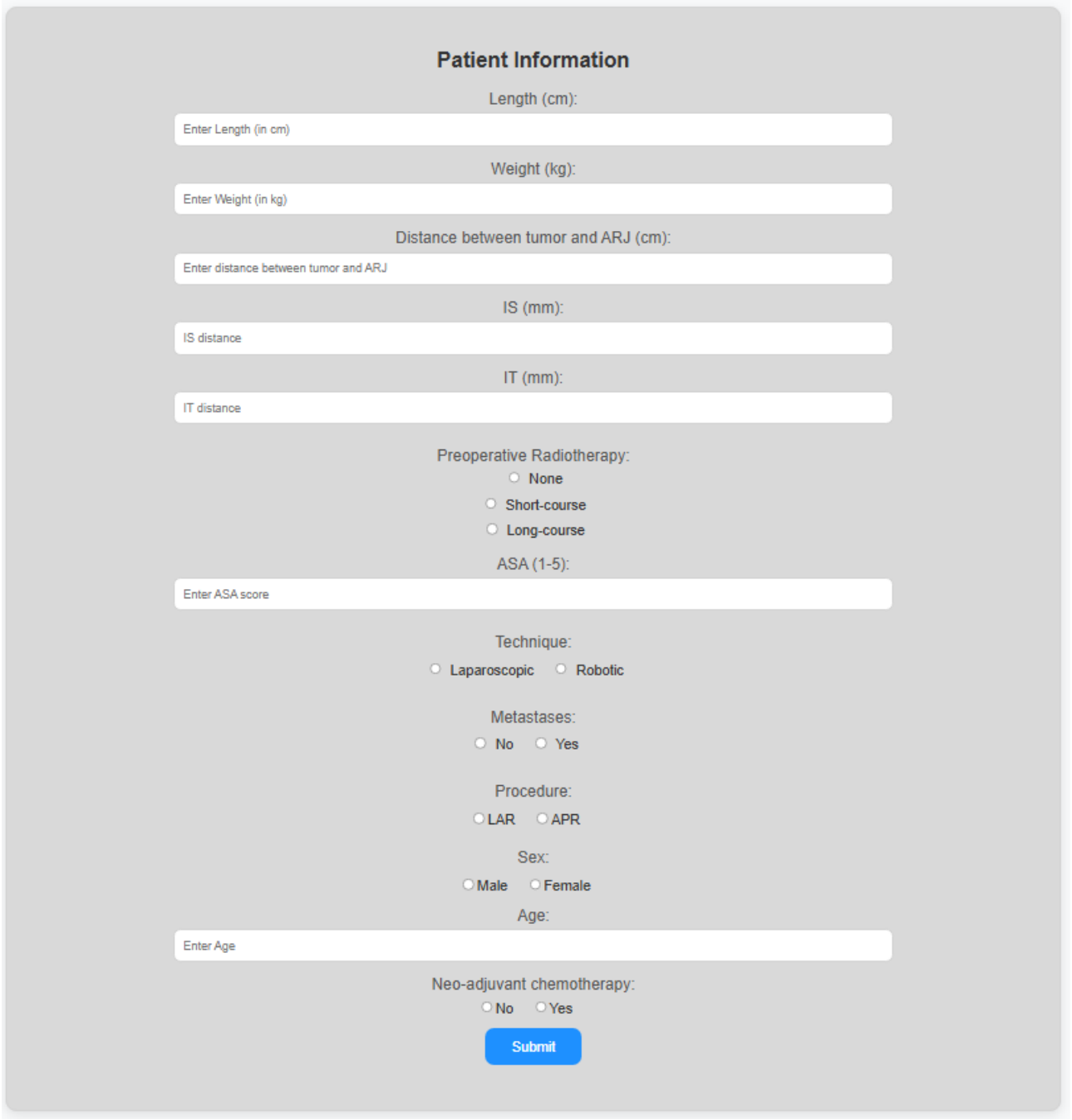



**Figure 32:** Dashboard input page

5. **Prediction Models:** Two pre-trained machine learning models are used for prediction:
   (a) Anastomotic leakage.
   (b) Clavien-Dindo grade 3+.

   Data frames from previous steps are merged and scaled to match the input requirements of the prediction models. The outputs include predictions and their associated probabilities.

6. **Dashboard output page:** The final dashboard page presents clinical parameters, predictions, and probabilities. Parameters indicating risks are highlighted in red, while those with positive implications are shown in green. An overview of the output page is displayed in Figure 33.

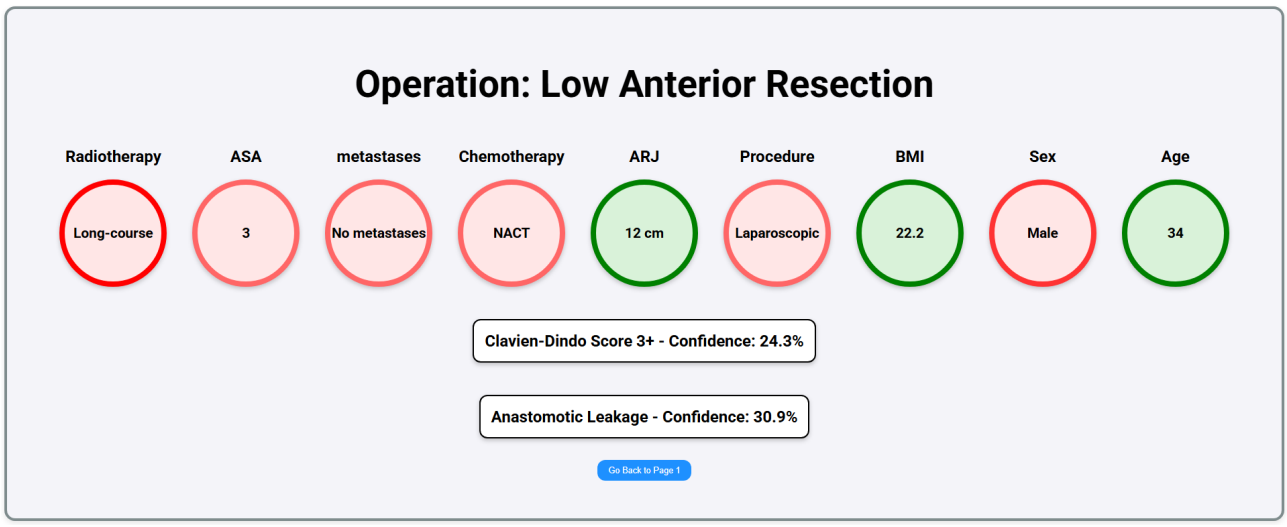



**Figure 33:** Dashboard output page highlights parameters, with red indicating risks and green representing positive values

## 7.3 Usability

The dashboard is envisioned as a tool to enhance clinical decision-making by providing surgeons and clinicians and relevant patient-specific information. The primary use case involves clinicians inputting patient parameters before the patient arrives, allowing for a quick and comprehensive overview of the patient's condition. For example, during preoperative consultations, the clinician can review MRI-derived measurements (e.g., IS, IT, and tumor-to-ano-rectal junction distance) provided by the radiologist, alongside clinical variables, to assess surgical risks.

The dashboard can also support multidisciplinary discussions by centralizing relevant data in an easy-to-read format. Surgeons could quickly visualize patient-specific risk factors during case presentations or surgical planning meetings, reducing preparation time and improving communication.

In the future, integration with Electronic Health Record (EHR) systems such as Easycare, HiX, or Epic and imaging systems like PACS could automate data transfer, further reducing manual input and potential errors. Additionally, customizable parameter visualization would enable surgeons to focus on the most critical information, tailored to their preferences and the specific case.

## 7.4 Discussion

*Dashboard Functionality*
The dashboard demonstrates the potential of a clinical decision support tool (CDST) for surgeons by providing patient-specific risk-factor insights. This proof-of-concept tool is designed to align with clinical workflows, enabling surgeons to identify at-risk patients and tailor their approaches accordingly. For example, during preoperative consultations, the dashboard centralizes clinical and imaging data, streamlining the decision-making process. Role-based access ensures that each user, from radiologists to assistants, interacts with the dashboard in a way that enhances efficiency and reduces redundant tasks.

Future usability improvements could include detailed visualizations of pelvimetry measurements. These visualizations would help surgeons identify parameters that deviate from expected ranges, offering an intuitive understanding of potential surgical challenges.

*Integration and Usability*
A critical factor in the dashboard's adoption is seamless integration into clinical workflows and electronic systems. Future iterations will integrate with EHR systems such as Easycare, HiX, or Epic, and imaging systems like PACS, automating data transfer and minimizing manual input. Customizable parameter visualization will further enhance usability by allowing surgeons to focus on the most critical information tailored to each case.

Challenges, such as resistance from clinicians or difficulties integrating with third-party systems, must be addressed through intuitive design and user feedback. For example, incorporating mechanisms like the System Usability Scale (SUS) will ensure the dashboard evolves based on clinician needs and preferences.

*Technical Development*
Role-based functionality is a feature that should be integrated in future iterations. For instance:

- **Radiologists:** Responsible for uploading MRI scans and facilitating measurements of IS, IT and distance-to-tumor.
- **Assistants:** Tasked with inputting clinical parameters before the surgeon's review to streamline the workflow.
- **Surgeons:** Able to customize the dashboard to display only the parameters most relevant to their decision-making process, ensuring a tailored and efficient interface.

Transitioning from Dash to Django will enhance the dashboard's security, scalability, and role-based functionality. By incorporating features such as user authentication and access control, the dashboard can operate securely in clinical environments. Additionally, future developments will include SHAP-based interpretability for transparent predictions, further aligning the tool with clinical usability standards.

As a proof of concept, the dashboard demonstrates significant potential to streamline clinical workflows, enhance decision-making, and improve patient outcomes. With continued development and a focus on usability, it can become an integral part of surgical decision-making in outpatient clinics.

# Final Conclusion and Future Works

The aim of this thesis was to develop a machine learning model to predict Anastomotic Leakage and Clavien Dindo Grade 3+ in patients who underwent total mesorectal excision. To accomplish this, a pipeline has been developed to train several machine learning models, and a selection is based on choosing a model with around 80% sensitivity, yielding the following results: For the anastomotic leakage model, the performance metrics were as follows — an accuracy of $0.528 \pm 0.034$, a sensitivity of $0.765 \pm 0.140$, a specificity of $0.484 \pm 0.057$, an F1 score of $0.332 \pm 0.042$, and an AUC of $0.667 \pm 0.075$. The CDC3+ model achieved an accuracy of $0.392 \pm 0.075$, a sensitivity of $0.835 \pm 0.102$, a specificity of $0.306 \pm 0.097$, an F1 score of $0.309 \pm 0.025$, and an AUC of $0.608 \pm 0.031$.

The second study performed is the development of a deep learning model to accurately predict the length of the sacral curve and extra parameters extracted from this result. This model displayed a Dice score of 0.85, clDice of 0.89, and a Hausdorff distance of 13.13, indicating decent performance. The trained model was utilized to segment sacral curves for 1,707 patients, with 1,205 patients having an accurate segmentation. Following in- and exclusion criteria, statistical analysis is performed for 390 patients. Results displayed a significant p-value for Ratio $AB$ and the highest correlation for Ratio $AB$ (0.14) and Ratio $BC$ (-0.10).

At last, utilizing both models, a pipeline is developed to predict both target outcomes based on MRI volumes and clinical metrics. This pipeline is used to develop an interactive dashboard using Flask and Dash that highlights patient characteristics and gives a probability of the chance of anastomotic leakage or CDC3+ happening. This information can assist surgeons in the decision-making process for improving patient-tailored care.

Future works should include increased data gathering for the machine learning model, with the main focus being on anastomotic leakage as this has more potential than CDC3+. Investigating additional features, such as surgical expertise, might strongly improve the model. The Sacral Curve model displayed little usage in terms of statistical analysis; therefore, the model needs to improve to show significant performance and as a result develop useful parameters that can prove to be an addition to the machine learning model. At last, the developed dashboard has to be tested in clinical practice, with the future of being integrated into the Electronic Health Records. Developing the pipeline using Django can add to scalability, security, and improve the user interface to make the dashboard more intuitive.

Looking at the complete study, the aim was to accurately predict anastomotic leakage with clinically acceptable accuracy, potentially assisting in decisions to not operate on a patient. However, this is unlikely to be achievable, even with a significant increase in data in the next years. Adding extra features might improve the model, but there is too much variability in human factors to predict it accurately. Predicting CDC3+ is off-limits due to the inability to model a direct causal link with the current features.

By combining machine learning, deep learning, and dashboard development, this thesis provides a foundation for advancing patient-tailored care for individuals undergoing Total Mesorectal Excision surgery, with the potential to significantly enhance decision-making processes in personalized medicine.

# Appendix

### Point Biserial correlation

This is a t-test used when you want to evaluate the relationship between a dichotomous variable and a continuous variable, such as weight or height. It calculates the correlation, which can range between -1 and 1. A negative value indicates that if the continuous variable increases, the chance of being positive decreases. A positive value indicates the opposite. The closer the value is to -1 or 1, the stronger the correlation.

$$r_{pb} = \frac{(Y_1 - Y_0)}{s_y} \cdot \sqrt{\frac{N_0 \cdot N_1}{N^2}}$$

Where:

- $Y_0$: Mean of the metric observations coded as 0.
- $Y_1$: Mean of the metric observations coded as 1.
- $N_0$: Number of observations coded as 0.
- $N_1$: Number of observations coded as 1.
- $N = N_0 + N_1$: Total number of observations.
- $s_y$: Standard deviation of all metric observations.

### Chi-square test

The chi-square test is used to test if there is a significant association between two categorical variables. It assesses the independence of the variables by comparing the observed frequencies in each category to the frequencies expected under the assumption of independence. Cramer's V can be calculated after conducting the Chi-square test to measure the strength of the correlation. Cramer's V ranges from 0 to 1. The interpretation of the correlation is illustrated in Figure X.

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Where:

- $O$: Observed frequency.
- $E$: Expected frequency.

### Metrics

The following metrics are used to evaluate the performance of classification models. These metrics help assess how well the model distinguishes between classes.

- **Definitions:**
  - **TP (True Positive):** The number of positive instances correctly identified by the model.
  - **TN (True Negative):** The number of negative instances correctly identified by the model.
  - **FP (False Positive):** The number of negative instances incorrectly classified as positive by the model.
  - **FN (False Negative):** The number of positive instances incorrectly classified as negative by the model.

  These terms form the foundation for the following metrics:

- **Accuracy:** Accuracy measures the overall correctness of the model by calculating the proportion of correctly classified instances.

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

- **Specificity (True Negative Rate):** Specificity measures the model's ability to correctly identify negative cases. It is the proportion of true negatives out of all actual negative instances.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

- **Sensitivity (Recall or True Positive Rate):** Sensitivity measures the model's ability to correctly identify positive cases. It is the proportion of true positives out of all actual positive instances.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- **Precision:** Precision measures the proportion of correctly predicted positive cases out of all instances predicted as positive.

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

- **F1-score:** The F1-score is the harmonic mean of precision and recall. It provides a balanced measure that considers both false positives and false negatives.

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

# References


[1] M S Fazeli and M R Keramati. *Rectal cancer: a review*. Tech. rep. 2015, p. 171. URL: http://mjiri.iums.ac.ir.

[2] Wolfgang B. Gaertner et al. "Rectal cancer: An evidence-based update for primary care providers". In: *World Journal of Gastroenterology* 21.25 (July 2015), pp. 7659–7671. ISSN: 22192840. DOI: 10.3748/wjg.v21.i25.7659.

[3] "Introduction to Total Mesorectal Excision Samir Delibegovic". In: (2017). DOI: 10.5455/medarh.2017.71.434-438. URL: www.orcid.org/0000-0003-0525-3288..

[4] Saran Lotfollahzadeh et al. *Rectal Cancer*. 2024.

[5] Lotfollahzadeh S et al. "Rectal Cancer". In: *StatPearls* (2024). URL: https://pubmed.ncbi.nlm.nih.gov/29630254/.

[6] Joep Knol and Deborah S. Keller. "Total Mesorectal Excision Technique-Past, Present, and Future". In: *Clinics in colon and rectal surgery* 33.3 (May 2020), pp. 134–143. ISSN: 1531-0043. DOI: 10.1055/S-0039-3402776. URL: https://pubmed.ncbi.nlm.nih.gov/32351336/.

[7] P. Terry Phang. "Total mesorectal excision: technical aspects". In: *Canadian Journal of Surgery* 47.2 (Apr. 2004), p. 130. ISSN: 0008428X. URL: /pmc/articles/PMC3211930/%20/pmc/articles/PMC3211930/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3211930/.

[8] Thijs Adriaan Burghgraef et al. "Robot-Assisted Total Mesorectal Excision Versus Laparoscopic Total Mesorectal Excision: A Retrospective Propensity Score-Matched Cohort Analysis in Experienced Centers". In: *Diseases of the Colon and Rectum* 65.2 (Feb. 2022), pp. 218–227. ISSN: 15300358. DOI: 10.1097/DCR.0000000000002031. URL: https://journals.lww.com/dcrjournal/fulltext/2022/02000/robot_assisted_total_mesorectal_excision_versus.13.aspx.

[9] J. S.Y. Hong et al. "The role of MRI pelvimetry in predicting technical difficulty and outcomes of open and minimally invasive total mesorectal excision: a systematic review". In: *Techniques in Coloproctology* 24.10 (Oct. 2020), pp. 991–1000. ISSN: 1128045X. DOI: 10.1007/S10151-020-02274-X/FIGURES/5. URL: https://link.springer.com/article/10.1007/s10151-020-02274-x.

[10] Zhen Sun et al. "Establishment of Surgical Difficulty Grading System and Application of MRI-Based Artificial Intelligence to Stratify Difficulty in Laparoscopic Rectal Surgery". In: *Bioengineering* 10.4 (Apr. 2023). ISSN: 23065354. DOI: 10.3390/bioengineering10040468.

[11] Robert T. van Kooten et al. "The Impact of Postoperative Complications on Short- and Long-Term Health-Related Quality of Life After Total Mesorectal Excision for Rectal Cancer". In: *Clinical Colorectal Cancer* 21.4 (Dec. 2022), pp. 325–338. ISSN: 19380674. DOI: 10.1016/j.clcc.2022.07.004.

[12] F. D. McDermott et al. *Systematic review of preoperative, intraoperative and postoperative risk factors for colorectal anastomotic leaks*. Apr. 2015. DOI: 10.1002/bjs.9697.

[13] Bodil Gessler, Olle Eriksson, and Eva Angenete. "Diagnosis, treatment, and consequences of anastomotic leakage in colorectal surgery". In: *International Journal of Colorectal Disease* 32.4 (Apr. 2017), pp. 549–556. ISSN: 14321262. DOI: 10.1007/s00384-016-2744-x.

[14] Yufei Jiang et al. *Global pattern and trends of colorectal cancer survival: a systematic review of population-based registration data*. Feb. 2022. DOI: 10.20892/j.issn.2095-3941.2020.0634.

[15] "a7699". In: ().

[16] Rashid N. Lui et al. "Global Increasing Incidence of Young-Onset Colorectal Cancer Across 5 Continents: A Joinpoint Regression Analysis of 1,922,167 Cases". In: *Cancer Epidemiology, Biomarkers & Prevention* 28.8 (Aug. 2019), pp. 1275–1282. ISSN: 1055-9965. DOI: 10.1158/1055-9965.EPI-18-1111.

[17] American Cancer Society (Content team and contributors). "Colorectal Cancer Early Detection, Diagnosis, and Staging". In: *American Cancer Society* (2024).

[18] The Netherlands Cancer Institute (Antoni van Leeuwenhoek). *Colon Cancer - Population Screening (Bevolkingsonderzoek)*.

[19] Sarah McNabb et al. "Meta-analysis of 16 studies of the association of alcohol with colorectal cancer". In: *International Journal of Cancer* 146.3 (Feb. 2020), pp. 861–873. ISSN: 10970215. DOI: 10.1002/ijc.32377.

[20] Anna Lewandowska et al. "Title: Risk Factors for the Diagnosis of Colorectal Cancer". In: *Cancer Control* 29 (Jan. 2022). ISSN: 15262359. DOI: 10.1177/10732748211056692.

[21] Tobias Niedermaier et al. "Flexible sigmoidoscopy in colorectal cancer screening: implications of different colonoscopy referral strategies". In: *European Journal of Epidemiology* 33.5 (May 2018), pp. 473–484. ISSN: 15737284. DOI: 10.1007/s10654-018-0404-x.

[22] Muhammad Amir Saeed Khan et al. "The Impact of Tumour Distance From the Anal Verge on Clinical Management and Outcomes in Patients Having a Curative Resection for Rectal Cancer". In: *Journal of Gastrointestinal Surgery* 21.12 (Dec. 2017), pp. 2056–2065. ISSN: 1091255X. DOI: 10.1007/s11605-017-3581-0.

[23] Courtney C. Moreno, Patrick S. Sullivan, and Pardeep K. Mittal. "MRI Evaluation of Rectal Cancer: Staging and Restaging". In: *Current Problems in Diagnostic Radiology* 46.3 (May 2017), pp. 234–241. ISSN: 03630188. DOI: 10.1067/j.cpradiol.2016.11.011.

[24] Surendra Patel et al. "Role of Transrectal Ultrasound in Preoperative Local Staging of Carcinoma Rectum and It's Histopathological Correlation". In: *Indian Journal of Surgery* 76.1 (Feb. 2014), pp. 21–25. ISSN: 09739793. DOI: 10.1007/s12262-012-0613-6.

[25] C. A. Maurer et al. "The Impact of the Introduction of Total Mesorectal Excision on Local Recurrence Rate and Survival in Rectal Cancer: Long-Term Results". In: *Annals of Surgical Oncology* 18.7 (July 2011), pp. 1899–1906. ISSN: 1068-9265. DOI: 10.1245/s10434-011-1571-0.

[26] N S Williams, M F Dixon, and D Johnston. "Reappraisal of the 5 centimetre rule of distal excision for carcinoma of the rectum: A study of distal intramural spread and of patients' survival". In: *Journal of British Surgery* 70.3 (Mar. 1983), pp. 150–154. ISSN: 0007-1323. DOI: 10.1002/bjs.1800700305.

[27] Rakesh Kumar Gupta et al. "Anterior Resection for Rectal Cancer with Mesorectal Excision: Institutional Review". In: *Indian Journal of Surgery* 75.1 (Feb. 2013), pp. 10–16. ISSN: 0972-2068. DOI: 10.1007/s12262-012-0445-4.

[28] Behnam Behboudi et al. "The impact of circular stapler size on the risk of anastomotic stricture following total mesorectal excision in rectal cancer patients: A retrospective cross-sectional study." In: *Health science reports* 6.10 (Oct. 2023), e1658. ISSN: 2398-8835. DOI: 10.1002/hsr2.1658.

[29] Michael S Thomas and David A Margolin. "Management of Colorectal Anastomotic Leak." In: *Clinics in colon and rectal surgery* 29.2 (June 2016), pp. 138–44. ISSN: 1531-0043. DOI: 10.1055/s-0036-1580630.

[30] W. Brian Perry and J. Christopher Connaughton. *Abdominoperineal resection: How is it done and what are the results?* Aug. 2007. DOI: 10.1055/s-2007-984865.

[31] Richard Bamford Gopal Menon Rui Wei. "Abdominoperineal Resection". In: *StatPearls [Internet]* (2024).

[32] Ludmila Boublikova et al. "Total neoadjuvant therapy in rectal cancer: the evidence and expectations". In: *Critical Reviews in Oncology/Hematology* 192 (Dec. 2023), p. 104196. ISSN: 10408428. DOI: 10.1016/j.critrevonc.2023.104196.

[33] Mohammed Faisal Bin Abdur Raheem, Zi Qin Ng, and Mary Theophilus. "The impact of pelvimetry data on rectal cancer surgery—a systematic review". In: *Annals of Laparoscopic and Endoscopic Surgery* 0 (Oct. 2023), pp. 0–0. ISSN: 25186973. DOI: 10.21037/ales-24-12.

[34] T Yamamoto et al. "Prediction of surgical difficulty in minimally invasive surgery for rectal cancer by use of MRI pelvimetry". In: (2020). DOI: 10.1002/bjs5.50292. URL: www.bjsopen.com.

[35] Johnny Chau et al. "Pelvic dimensions on preoperative imaging can identify poor-quality resections after laparoscopic low anterior resection for mid- and low rectal cancer". In: *Surgical Endoscopy* 34.10 (Oct. 2020), pp. 4609–4615. ISSN: 0930-2794. DOI: 10.1007/s00464-019-07209-8.

[36] Miao Yu et al. "Interpretable machine learning model to predict surgical difficulty in laparoscopic resection for rectal cancer". In: *Frontiers in Oncology* 14 (Feb. 2024). ISSN: 2234943X. DOI: 10.3389/fonc.2024.1337219.

[37] Yuan Liu et al. "Applying interpretable machine learning algorithms to predict risk factors for permanent stoma in patients after TME". In: *Frontiers in Surgery* 10 (Mar. 2023), p. 1125875. ISSN: 2296875X. DOI: 10.3389/FSURG.2023.1125875/BIBTEX.

[38] Maria Clara Fernandes, Marc J Gollub, and Gina Brown. "The importance of MRI for rectal cancer evaluation". In: (). DOI: 10.1016/j.suronc.2022.101739.

[39] Christopher Spence et al. "Machine learning models to predict surgical case duration compared to current industry standards: scoping review". In: *BJS Open* 7.6 (Dec. 2023). ISSN: 24749842. DOI: 10.1093/BJSOPEN/ZRAD113. URL: /pmc/articles/PMC10630142/%20/pmc/articles/PMC10630142/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10630142/.

[40] Thomas G Dietterich. *Machine Learning*. Tech. rep.

[41] Iqbal H Sarker. "Machine Learning: Algorithms, Real-World Applications and Research Directions." In: *SN computer science* 2.3 (2021), p. 160. ISSN: 2661-8907. DOI: 10.1007/s42979-021-00592-x.

[42] Jenni A.M. Sidey-Gibbons and Chris J. Sidey-Gibbons. "Machine learning in medicine: a practical introduction". In: *BMC Medical Research Methodology* 19.1 (Mar. 2019). ISSN: 14712288. DOI: 10.1186/s12874-019-0681-4.

[43] Mike May. "Eight ways machine learning is assisting medicine". In: *Nature Medicine* 27.1 (Jan. 2021), pp. 2–3. ISSN: 1078-8956. DOI: 10.1038/s41591-020-01197-2.

[44] Jan Mendling et al. "Thresholds for error probability measures of business process models". In: *Journal of Systems and Software* 85.5 (May 2012), pp. 1188–1197. ISSN: 01641212. DOI: 10.1016/j.jss.2012.01.017.

[45] Lin Hao et al. "Deep Learning-Based Survival Analysis for High-Dimensional Survival Data". In: *Mathematics* 9.11 (May 2021), p. 1244. ISSN: 2227-7390. DOI: 10.3390/math9111244.

[46] Fatma M. Talaat and Rana Mohamed El-Balka. "Stress monitoring using wearable sensors: IoT techniques in medical field". In: *Neural Computing and Applications* 35.25 (Sept. 2023), pp. 18571–18584. ISSN: 0941-0643. DOI: 10.1007/s00521-023-08681-z.

[47] Andre Esteva et al. "A guide to deep learning in healthcare". In: *Nature Medicine* 25.1 (Jan. 2019), pp. 24–29. ISSN: 1078-8956. DOI: 10.1038/s41591-018-0316-z.

[48] Charu C. Aggarwal. *Neural Networks and Deep Learning*. Cham: Springer International Publishing, 2018. ISBN: 978-3-319-94462-3. DOI: 10.1007/978-3-319-94463-0.

[49] Andre Esteva et al. "Dermatologist-level classification of skin cancer with deep neural networks". In: *Nature* 542.7639 (Feb. 2017), pp. 115–118. ISSN: 0028-0836. DOI: 10.1038/nature21056.

[50] Thijs Kooi et al. "Large scale deep learning for computer aided detection of mammographic lesions". In: *Medical Image Analysis* 35 (Jan. 2017), pp. 303–312. ISSN: 13618415. DOI: 10.1016/j.media.2016.07.007.

[51] Keisuke Kazama et al. "Evaluation of short-term outcomes of laparoscopic-assisted surgery for colorectal cancer in elderly patients aged over 75 years old: a multi-institutional study (YSURG1401)". In: *BMC Surgery* 17.1 (Dec. 2017), p. 29. ISSN: 1471-2482. DOI: 10.1186/s12893-017-0229-7.

[52] Lukas Biewald. *Experiment Tracking with Weights and Biases*. 2020.

[53] Tyler J Bradshaw et al. "A Guide to Cross-Validation for Artificial Intelligence in Medical Imaging." In: *Radiology. Artificial intelligence* 5.4 (July 2023), e220232. ISSN: 2638-6100. DOI: 10.1148/ryai.220232.

[54] N.A. Diamantidis, D. Karlis, and E.A. Giakoumakis. "Unsupervised stratification of cross-validation for accuracy estimation". In: *Artificial Intelligence* 116.1-2 (Jan. 2000), pp. 1–16. ISSN: 00043702. DOI: 10.1016/S0004-3702(99)00094-6.

[55] Bruce G. Marcot and Anca M. Hanea. "What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis?" In: *Computational Statistics* 36.3 (Sept. 2021), pp. 2009–2031. ISSN: 0943-4062. DOI: 10.1007/s00180-020-00999-9.

[56] Azal Ahmad Khan, Omkar Chaudhari, and Rohitash Chandra. "A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation". In: *Expert Systems with Applications* 244 (June 2024), p. 122778. ISSN: 09574174. DOI: 10.1016/j.eswa.2023.122778.

[57] Nicholas Pudjihartono et al. *A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction*. 2022. DOI: 10.3389/fbinf.2022.927312.

[58] Suhang Wang, Jiliang Tang, and Huan Liu. "Feature Selection". In: *Encyclopedia of Machine Learning and Data Mining*. Boston, MA: Springer US, 2016, pp. 1–9. DOI: 10.1007/978-1-4899-7502-7{_}101-1.

[59] Mohammad Ziaul Islam Chowdhury and Tanvir C Turin. "Variable selection strategies and its importance in clinical prediction modelling." In: *Family medicine and community health* 8.1 (2020), e000262. ISSN: 2305-6983. DOI: 10.1136/fmch-2019-000262.

[60] Robert M. O'brien. "A Caution Regarding Rules of Thumb for Variance Inflation Factors". In: *Quality & Quantity* 41.5 (Sept. 2007), pp. 673–690. ISSN: 0033-5177. DOI: 10.1007/s11135-006-9018-6.

[61] Christopher Glen Thompson et al. "Extracting the Variance Inflation Factor and Other Multicollinearity Diagnostics from Typical Regression Results". In: *Basic and Applied Social Psychology* 39.2 (Mar. 2017), pp. 81–90. ISSN: 0197-3533. DOI: 10.1080/01973533.2016.1277529.

[62] Jenna Wong et al. "Can Hyperparameter Tuning Improve the Performance of a Super Learner?: A Case Study." In: *Epidemiology (Cambridge, Mass.)* 30.4 (July 2019), pp. 521–531. ISSN: 1531-5487. DOI: 10.1097/EDE.0000000000001027.

[63] Nitesh V Chawla et al. *SMOTE: Synthetic Minority Over-sampling Technique*. Tech. rep. 2002, pp. 321–357.

[64] Golshid Ranjbaran et al. "Leveraging augmentation techniques for tasks with unbalancedness within the financial domain: a two-level ensemble approach". In: *EPJ Data Science* 12.1 (July 2023), p. 24. ISSN: 2193-1127. DOI: 10.1140/epjds/s13688-023-00402-9.

[65] Şeref Kerem Çorbacıoğlu and Gökhan Aksel. "Receiver operating characteristic curve analysis in diagnostic accuracy studies". In: *Turkish Journal of Emergency Medicine* 23.4 (Oct. 2023), pp. 195–198. ISSN: 2452-2473. DOI: 10.4103/tjem.tjem{_}182{_}23.

[66] Takaya Saito and Marc Rehmsmeier. "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets". In: *PLOS ONE* 10.3 (Mar. 2015), e0118432. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0118432.

[67] Jianhua Chen et al. "MRI pelvimetry-based evaluation of surgical difficulty in laparoscopic total mesorectal excision after neoadjuvant chemoradiation for male rectal cancer". In: *Surgery Today* 51 (1234), pp. 1144–1151. DOI: 10.1007/s00595-020-02211-3. URL: https://doi.org/10.1007/s00595-020-02211-3.

[68] Won-Suk Lee et al. "Risk Factors and Clinical Outcome for Anastomotic Leakage After Total Mesorectal Excision for Rectal Cancer". In: *World Journal of Surgery* 32.6 (June 2008), pp. 1124–1129. ISSN: 0364-2313. DOI: 10.1007/s00268-007-9451-2.

[69] Jiatong Li and Zhaopeng Yan. "Machine learning model predicting factors for incisional infection following right hemicolectomy for colon cancer". In: *BMC Surgery* 24.1 (Oct. 2024), p. 279. ISSN: 1471-2482. DOI: 10.1186/s12893-024-02543-8.

[70] L. Escal et al. "MRI-based score to predict surgical difficulty in patients with rectal cancer". In: *The British journal of surgery* 105.1 (Jan. 2018), pp. 140–146. ISSN: 1365-2168. DOI: 10.1002/BJS.10642. URL: https://pubmed.ncbi.nlm.nih.gov/29088504/.

[71] N. Bagla and J. B. Schofield. "Rectosigmoid tumours: should we continue sitting on the fence?" In: *Colorectal Disease* 9.7 (Sept. 2007), pp. 606–608. ISSN: 1462-8910. DOI: 10.1111/j.1463-1318.2007.01329.x.

[72] Alireza Farhangfar, Lukasz Kurgan, and Jennifer Dy. "Impact of imputation of missing values on classification error for discrete data". In: *Pattern Recognition* 41.12 (Dec. 2008), pp. 3692–3705. ISSN: 00313203. DOI: 10.1016/j.patcog.2008.05.019.

[73] Namgil Lee, Heejung Yang, and Hojin Yoo. "A surrogate loss function for optimization of $F_\beta$ score in binary classification with imbalanced data". In: (Apr. 2021). URL: http://arxiv.org/abs/2104.01459.

[74] Jong Hae Kim. "Multicollinearity and misleading statistical results". In: *Korean Journal of Anesthesiology* 72.6 (Dec. 2019), pp. 558–569. ISSN: 2005-6419. DOI: 10.4097/kja.19087.

[75] Sonia Fernández-Ananín et al. "Reply to: doi:10.1007/s00464-010-1485-0: Evaluation of factors affecting the difficulty of laparoscopic anterior resection for rectal cancer: "narrow pelvis" is not a contradiction". In: *Surgical Endoscopy* 26.9 (Sept. 2012), pp. 2698–2699. ISSN: 0930-2794. DOI: 10.1007/s00464-012-2231-6.

[76] R J Heald. "The 'Holy Plane' of Rectal Surgery". In: *Journal of the Royal Society of Medicine* 81.9 (Sept. 1988), pp. 503–508. ISSN: 0141-0768. DOI: 10.1177/014107688808100904.

[77] Min Soo Cho, Hyeon Woo Bae, and Nam Kyu Kim. "Essential knowledge and technical tips for total mesorectal excision and related procedures for rectal cancer". In: *Annals of Coloproctology* 40.4 (Aug. 2024), pp. 384–411. ISSN: 2287-9714. DOI: 10.3393/ac.2024.00388.0055.

[78] Jonathan S. Y. Hong et al. "Can MRI pelvimetry predict the technical difficulty of laparoscopic rectal cancer surgery?" In: *International Journal of Colorectal Disease* 36.12 (Dec. 2021), pp. 2613–2620. ISSN: 0179-1958. DOI: 10.1007/s00384-021-04000-x.

[79] Andriy Fedorov et al. "3D Slicer as an image computing platform for the Quantitative Imaging Network". In: *Magnetic Resonance Imaging* 30.9 (Nov. 2012), pp. 1323–1341. ISSN: 0730725X. DOI: 10.1016/j.mri.2012.05.001.

[80] Özgün Çiçek et al. "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation". In: (June 2016). URL: http://arxiv.org/abs/1606.06650.

[81] Suprosanna Shit et al. "clDice – A Novel Topology-Preserving Loss Function for Tubular Structure Segmentation". In: (Mar. 2020). DOI: 10.1109/CVPR46437.2021.01629. URL: http://arxiv.org/abs/2003.07311%20http://dx.doi.org/10.1109/CVPR46437.2021.01629.

[82] D.P. Huttenlocher, G.A. Klanderman, and W.J. Rucklidge. "Comparing images using the Hausdorff distance". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15.9 (1993), pp. 850–863. ISSN: 01628828. DOI: 10.1109/34.232073.

[83] Davide Chicco, Matthijs J. Warrens, and Giuseppe Jurman. "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation". In: *PeerJ Computer Science* 7 (July 2021), e623. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.623.

[84] Edsger W Dijkstra. "A note on two problems in connexion with graphs". In: *Numerische mathematik* 1 (1959), pp. 269–271.

[85] Lei Xu and Jinhui Xu. "Approximating minimum bending energy path in a simple corridor". In: *Computational Geometry* 47.3 (Apr. 2014), pp. 349–366. ISSN: 09257721. DOI: 10.1016/j.comgeo.2013.09.001.

[86] Plotly Technologies Inc. *Dash Layout*. 2024.

[87] Miguel Grinberg. *Flask web development: developing web applications with python*. O'Reilly Media, Inc., 2018.