UNIVERSITY OF TWENTE

MASTER THESIS

Artificial intelligence for assessing operation time and surgical difficulty in laparoscopic cholecystectomy

Author:
Vincent OOSTERHOFF

A thesis submitted in fulfillment of the requirements for the degree of Master of science

October 31, 2024

UNIVERSITY OF TWENTE

Abstract

Department of surgery Meander Medical Center

Improving operation time estimation in laparoscopic cholecystectomy using pre- and intraoperative parameters

Vincent OOSTERHOFF

The operating room (OR) is responsible for significant revenue and costs in hospitals. Therefore, carefully scheduling the operating time is of great importance. However, surgeries frequently exceed their planned duration due to unforeseen intraoperative challenges. To address this issue and enhance surgical scheduling, we proposed a method that utilizes both preoperative and intraoperative variables to estimate operating time dynamically. Our studies at Meander Medical Center (MMC) focused on laparoscopic cholecystectomy, a high-volume procedure characterized by substantial intraoperative variability. Our initial analysis examined surgical data from 2017 to 2023 to quantify the extent of operating time inaccuracies. Machine learning (ML) models were trained on available clinical data to predict operating time. A linear regression (LR) model achieved a root mean squared error (RMSE) of ± 14.18 minutes, marginally outperforming the conventional method's RMSE of ± 16.22 minutes. Despite this improvement, the predictive accuracy remains insufficient for clinical application. In a subsequent prospective analysis, additional patient-specific factors were incorporated to enhance predictive performance. Following data cleaning, 199 patients were retained from an initial cohort of 231. The best-performing LR model achieved an RMSE of ± 12.28 minutes, representing a modest improvement over existing methods but still falling short of the precision required for clinical implementation. To further advance the model, intraoperative variables were integrated through an assessment of surgical difficulty using the Nassar scale. The multi-scale vision transformer (MViTv2) was employed to classify surgical videos, with the best model achieving an accuracy of 36%. This level of performance remains inadequate for clinical use. While these studies have contributed to a deeper understanding of the factors influencing operating time, the predictive models developed thus far are not yet suitable for clinical deployment. Nonetheless, further refinement and integration of preoperative and intraoperative variables hold promise for more accurate operating time predictions in the future.

Preface

This thesis marks the end of my academic journey at the University of Twente. I am very grateful for the opportunity I had to learn not only about the human body but also to explore technical innovations in great depth. Moreover, I am particularly pleased with the personal growth I experienced throughout my internship. I learned a great deal about myself, especially in how I work and communicate with supervisors and colleagues.

I would like to extend my thanks to **Simon Baltus** and **Julian Abbing** for their daily support and supervision. I am also grateful for their patience during moments of stress when they took the time to calm me down and help me determine a new course of action. I would also like to thank **Beerend Gerats** for the in-depth discussions on deep learning and the valuable insights he provided throughout the internship. Furthermore, I wish to thank Dr. **Can Ozan Tan** for the valuable technical discussions and his no-nonsense approach, which always pushed me to make significant strides in the right direction.

For medical supervision, I would like to express my gratitude to Professor **Ivo Broeders**. Your calm and kind demeanor with patients truly inspired me during my internship. Additionally, you motivated me and showed me how technical physicians can make meaningful contributions within a hospital. You gave me the freedom to pursue my own projects and helped me get back on track when things did not go as planned. I would also like to thank Dr. **Frank Voskens** for making me feel at ease in the operating room and demonstrating how hard work can be combined with a positive atmosphere. Moreover, I am grateful to the entire Department of Surgery at Meander Medical Center. I learned so much over the past year, thanks to everyone who always took the time to explain and assist me.

I also wish to extend my appreciation to **Nicole Cramer Bornemann** for her guidance in my professional and personal development. I learned a great deal during our intervisions, and you consistently offered methods suited to the problem at hand. The sessions with **Veerle Michels** and **Nadine Boutkan** also helped me to put things into perspective and enjoy the graduation journey more.

I am also very grateful to **Inge Wijma** for being part of this graduation committee.

Additionally, I want to thank **Vincent Ribbens** for the countless walks around the Meander Medical Center gardens and our deep-learning discussions during those walks. I am also thankful to **Anna Florax**, **Lieke Arts**, and **Johanneke ten Broekce** for their constant love and support, which not only helped me during my final graduation internship but throughout my entire studies. I would also like to thank **Sander Barendsen**for always reminding me to have lunch at 12 o'clock sharp and all the M2 students I have met along the way.

Lastly, I want to express my heartfelt thanks to my family, who have always been a safety net for me and with whom I could discuss all my struggles and setbacks. **Janne Oosterhoff**, **Jonas Oosterhoff**, **Hinke Oosterhoff**, **Maartje Oosterhoff**, **Esther Oosterhoff**, and **Sikko Oosterhoff**, thank you for all your support over the past 25 years.

Vincent Pieter Sikke Oosterhoff
October 31, 2024

Graduation Committee

Chairman & medical supervisor:
Prof. Dr. I.A.M.J. Broeders
Department of Surgery, Meander Medical Center, Amersfoort
Robotics and Mechatronics, University of Twente, Enschede

Technical supervisor:
Dr. C.O. Tan
Robotics and Mechatronics, University of Twente, Enschede

Daily supervisor:
S.C. Baltus, Msc.
Department of Surgery, Meander Medical Center, Amersfoort
Robotics and Mechatronics, University of Twente, Enschede

Process supervisor:
Drs. N.S. Cramer Bornemann
Technical Medicine, University of Twente, Enschede

External member:
I.N. Wijma, Msc.
Department of Pulmonology, Radboud Medical Center, Nijmegen

Table of contents

1

General introduction

General introduction

Accurate estimation of operation time is crucial for optimizing the operating room (OR) efficiency, which is a significant source of revenue and costs for hospitals [1]. Inaccurate operation time estimations can lead to operational inefficiencies, such as delays, unplanned overtime, or underutilized OR time [2, 3]. The ability to accurately predict surgery times not only improves financial outcomes for hospitals but also enhances patient care by reducing wait times and minimizing cancellations.

In standard clinical practice, surgical scheduling depends on estimated case durations provided by the surgeons themselves, which are often unreliable [4, 5, 6]. Previous research indicates limited accuracy in these estimations [5, 7]. Alternatively, in some hospitals, historical averages of case-time durations for a specific surgeon have been used to schedule operation time. Nonetheless, these, too, lack the required accuracy due to variations in the preoperative situation and are, therefore, unable to effectively predict the operating time [8].

To improve these current estimations, several studies attempted to estimate the procedure time based on patient factors preoperatively, reporting promising results in different departments [6, 9, 10, 11, 12]. However, these models are not accurate enough because the intraoperative setting too often differs from the expectations set by the preoperative parameters [13]. When adjusting for unexpected intraoperative findings during surgery, additional resources can be allocated, schedules can be modified, and the flow of the OR can be maintained efficiently [14]. Moreover, this information can assist recovery room nurses in tracking and managing the progress of multiple ongoing surgeries in the OR [15].

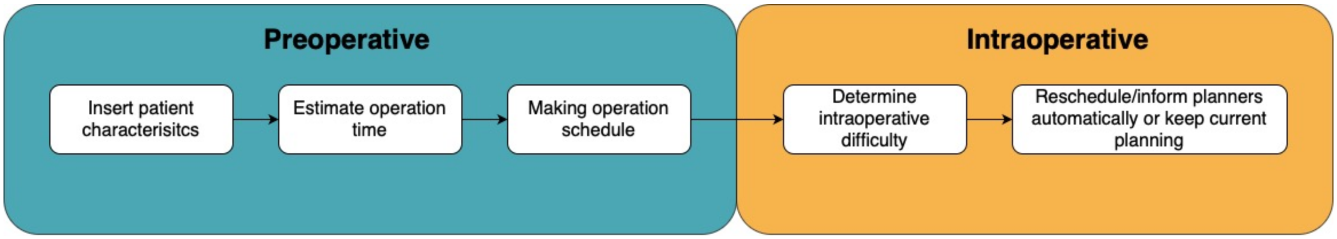

Preoperative
Insert patient characterisitcs
Estimate operation time
Making operation schedule
Intraoperative
Determine intraoperative difficulty
Reschedule/inform planners automatically or keep current planning

FIGURE 1: Overview of the individual components of the proposed model.

We hypothesize that combining preoperative and intraoperative information in a single model can significantly improve the estimation of operation time. To this end, we propose a model that begins with an initial estimation based on preoperative patient parameters, which can be linked to operation time using machine learning (ML) models. After generating this initial estimate, intraoperative findings—reflecting the difficulty of the operation, can be used to refine the estimate. Figure 1 shows an overview of our proposed model. We aim to achieve this refinement using the multiscale Vision Transformer version 2 (MViT2). To assess the feasibility of this approach, we focus on laparoscopic cholecystectomy (LC), a procedure commonly performed in peripheral hospitals with a wide range of intraoperative variations [16, 17]. We anticipate that incorporating such a scheduling model for LC could improve operating room planning, and with minor adjustments, the model could be adapted for other surgeries.

To investigate the feasibility of our proposed method, this thesis aims to answer the following research questions:

1. To what extent can we predict the operation time needed for a laparoscopic cholecystectomy based on preoperative variables?

2. To what extent can we determine intraoperative difficulty of a laparoscopic cholecystectomy using a multi-scale vision transformer?

To answer these questions, the thesis is divided into several chapters. **Chapter 2** provides medical and technical background. In **Chapter 3**, a retrospective study is presented to assess the planning accuracy of the current method, offering a preliminary exploration of operation time estimation. **Chapter 4** expands on this by incorporating additional parameters, aiming to capture a more comprehensive understanding of the intraoperative environment. In **Chapter 5**, a novel approach to evaluate intraoperative difficulty is proposed using a deep learning model. The thesis concludes in **Chapter 6** with a general discussion and summary of the applicability of our proposed dynamic operation time estimating model.

Bibliography

[1] Afnan Aljaffary, Fatimah AlAnsari, Abdulaleem Alatassi, Mohammed AlSuhaibani, and Ammar Alomran. Assessing the precision of surgery duration estimation: A retrospective study. *Journal of Multidisciplinary Healthcare*, pages 1565–1576, 2023.

[2] Faris A Alotaibi and Mohammed M Aljuaid. A comparison of surgeon estimated times and actual operative times in pediatric dental rehabilitation under general anesthesia. a retrospective study. *Journal of Clinical Medicine*, 12(13):4493, 2023.

[3] Vahid Riahi, Hamed Hassanzadeh, Sankalp Khanna, Justin Boyle, Faraz Syed, Barbara Biki, Ellen Borkwood, and Lianne Sweeney. Improving preoperative prediction of surgery duration. *BMC Health Services Research*, 23(1):1343, 2023.

[4] Daniel M Laskin, A Omar Abubaker, and Robert A Strauss. Accuracy of predicting the duration of a surgical operation. *Journal of Oral and Maxillofacial Surgery*, 71(2):446–447, 2013.

[5] Christopher H Stucky, Felichism W Kabo, Marla J De Jong, Sherita L House, Chandler H Moser, and Donald E Kimbler. Surgical control time estimation variability: Implications for medical systems and the future integration of ai and ml models. *Perioperative Care and Operating Room Management*, page 100432, 2024.

[6] Zhengli Wang and Franklin Dexter. More accurate, unbiased predictions of operating room times increase labor productivity with the same staff scheduling provided allocated hours are increased. *Perioperative Care and Operating Room Management*, 29:100286, 2022.

[7] Dario R Roque, Katina Robison, Christina A Raker, Gary G Wharton, and Gary N Frishman. The accuracy of surgeons' provided estimates for the duration of hysterectomies: a pilot study. *Journal of minimally invasive gynecology*, 22(1):57–65, 2015.

[8] Jinshi Zhou, Franklin Dexter, Alex Macario, and David A Lubarsky. Relying solely on historical surgical times to estimate accurately future surgical times is unlikely to reduce the average length of time cases finish late. *Journal of clinical anesthesia*, 11(7):601–605, 1999.

[9] Ingwon Yeo, Christian Klemt, Christopher M Melnic, Meghan H Pattavina, Bruna M Castro De Oliveira, and Young-Min Kwon. Predicting surgical operative time in primary total knee arthroplasty utilizing machine learning models. *Archives of Orthopaedic and Trauma Surgery*, 143(6):3299–3307, 2023.

[10] Jeremy M Lipman, Jeffrey A Claridge, Manjunath Haridas, Matthew D Martin, David C Yao, Kevin L Grimes, and Mark A Malangoni. Preoperative findings predict conversion from laparoscopic to open cholecystectomy. *Surgery*, 142(4):556–565, 2007.

[11] Reshma Bharamgoudar, Aniket Sonsale, James Hodson, and Ewen Griffiths. The development and validation of a scoring tool to predict the operative duration of elective laparoscopic cholecystectomy. *Surgical endoscopy*, 32:3149–3157, 2018.

[12] Ewen A Griffiths, James Hodson, Ravi S Vohra, Paul Marriott, Tarek Katbeh, Samer Zino, Ahmad HM Nassar, and West Midlands Research Collaborative. Utilisation of an operative difficulty grading scale for laparoscopic cholecystectomy. *Surgical endoscopy*, 33:110–121, 2019.

[13] Maria Vannucci, Giovanni Guglielmo Laracca, Paolo Mercantini, Silvana Perretta, Nicolas Padoy, Bernard Dallemagne, and Pietro Mascagni. Statistical models to preoperatively predict operative difficulty in laparoscopic cholecystectomy: a systematic review. *Surgery*, 171(5):1158–1167, 2022.

[14] Andru Putra Twinanda, Gaurav Yengera, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Rsdnet: Learning to predict remaining surgery duration from laparoscopic videos without manual annotations. *IEEE transactions on medical imaging*, 38(4):1069–1078, 2018.

[15] York Jiao, Bing Xue, Chenyang Lu, Michael S Avidan, and Thomas Kannampallil. Continuous real-time prediction of surgical case duration using a modular artificial neural network. *British journal of anaesthesia*, 128(5):829–837, 2022.

[16] Michael Sugrue, Federico Coccolini, Magda Bucholc, and Alison Johnston. Intra-operative gallbladder scoring predicts conversion of laparoscopic to open cholecystectomy: a wses prospective collaborative study. *World Journal of Emergency Surgery*, 14:1–8, 2019.

[17] Samah Osailan, Muhanad Esailan, Abdulaziz M Alraddadi, Faisal M Almutairi, and Zaid Sayedalamin. The use of intraoperative cholangiography during cholecystectomy: a systematic review. *Cureus*, 15(10), 2023.

2

Background

Background

This chapter provides a brief overview of laparoscopic cholecystectomy (LC), the surgical procedure central to this research. Next, we provide the technical background needed to develop our model. This is done in two parts. The first part elaborates on the methods used to estimate operation time from surgical data. The second part discusses the approaches used to assess surgical difficulty objectively from laparoscopic videos.

2.1 Laparoscopic cholecystectomy

Approximately 6% of all men and 9 %of all women have gallstones [1]. Asymptomatic gallbladder stones found in a normal gallbladder and biliary tree do not need treatment unless symptoms develop. However, approximately 20% of these asymptomatic gallstones will develop symptoms over 15 years of follow-up [2]. These patients experience intense colic pains at random moments, impeding their quality of life. Moreover, this can lead to cholecystitis, choledocholithiasis and biliary pancreatitis [3].

The standard treatment for symptomatic gallbladder disease is a gallbladder resection (cholecystectomy), which is traditionally performed with an open approach. However, since the introduction of laparoscopic surgery, the golden standard of cholecystectomy is changed to a laparoscopic approach [4]. This change in surgical technique and a less invasive approach leads to an increase in the total amount of LCs done worldwide [5]. Despite the minimally invasive nature and average low post-operative in-hospital time, an LC is not without risk. Some possible complications are damage to the common bile duct (CDB), only partial removal of the gallbladder, internal bleeding, conversion to open surgery, and post-operative inflammation [6].

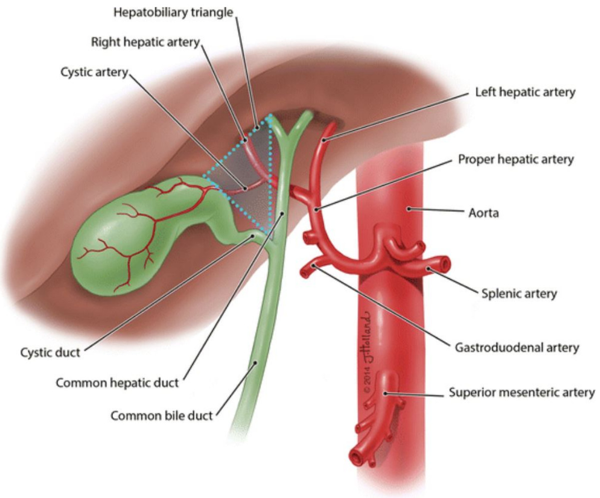


FIGURE 2.1: Anatomical overview of the gallbladder and the cystic artery [7].

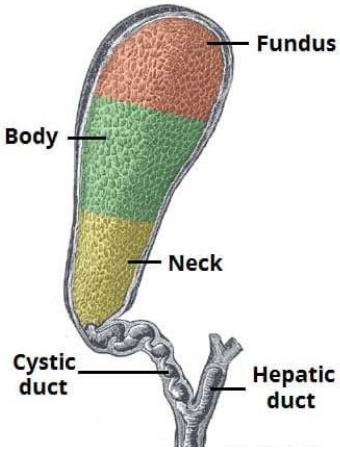


FIGURE 2.2: The different regions of the gallbladder [8].

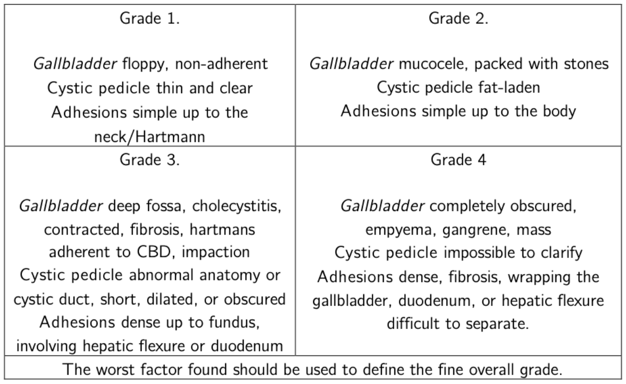
Knowledge of the anatomy of the gallbladder is of great importance to limit complications. In Figure 2.1 an overview of the relation of the gallbladder with the important arteries and the liver is shown. The gallbladder is connected with the cystic duct to the common bile duct [9]. Via the cystic duct, the gallbladder drains bile to the duodenum while also receiving the produced bile from the liver. The cystic artery supplies the blood to the gallbladder. The gallbladder is divided into a neck, body, and fundus, as shown in Figure 2.2. The neck of the gallbladder contains a mucosal fold known as Hartmann's pouch. This pouch is generally used to grasp the gallbladder and generate an overview.

A laparoscopic cholecystectomy consists of different stages [10]. The procedure begins with establishing access to the abdomen and creating a pneumoperitoneum to inflate the abdominal cavity. Once this is done, trocars are inserted to allow surgical instruments into the abdomen. The liver is first elevated to expose the gallbladder and surrounding structures. The surgeon then lifts the gallbladder's fundus to maintain this exposure while retracting Hartmann's pouch to improve the visibility of the bile ducts and arteries [11]. If adhesions are present, they are carefully separated before proceeding. Next, the surgeon dissects the peritoneum and surrounding fat from the cystic duct and artery within the hepatocystic triangle, visible in blue in Figure 2.1. Often done with blunt instruments, this dissection clears the field to expose the critical structures. The aim here is to achieve what is known as the Critical View of Safety (CVS), where both the cystic duct and cystic artery are identified and isolated before any further action [12]. Once the CVS is confirmed, the cystic duct and cystic artery are each clipped in two places, close to the gallbladder and away from it, to prevent bleeding and bile leakage. After the clips are securely placed, the cystic duct and artery are transected between the clips using scissors.

Following this, the gallbladder is carefully separated from the liver bed. Once freed, it is placed in a retrieval bag to prevent contamination or the spillage of bile and gallstones into the abdominal cavity. The gallbladder is then removed from the body through one of the incisions. Once the gallbladder is removed, the abdominal cavity is inspected, and the instruments are withdrawn, concluding the surgery.

2.1.1 Nassar grade

The difficulty of laparoscopic cholecystectomy (LC) can vary significantly due to factors such as patient anatomy, adhesions to neighbouring structures, and active inflammation [13]. Several scoring systems have been developed to assess these characteristics. In this thesis, we primarily utilize the Nassar grading system, as it is the only clinically validated method [14]. The Nassar grade consists of four levels: a Nassar grade of 1 indicates a relatively straightforward procedure, while a Nassar grade of 4 denotes a very challenging one. This grading system includes three subscores that evaluate the gallbladder, the presence of adhesions, and the cystic pedicle. The cystic pedicle is a triangular fold of peritoneum that contains the cystic duct, cystic artery, and a variable amount of fat [15]. Figure **??** illustrates the different Nassar grades, and Figure **??** displays laparoscopic images corresponding to each grade.



Grade 1. *Gallbladder* floppy, non-adherent Cystic pedicle thin and clear Adhesions simple up to the neck/Hartmann	Grade 2. *Gallbladder* mucocele, packed with stones Cystic pedicle fat-laden Adhesions simple up to the body
Grade 3. *Gallbladder* deep fossa, cholecystitis, contracted, fibrosis, hartmans adherent to CBD, impaction Cystic pedicle abnormal anatomy or cystic duct, short, dilated, or obscured Adhesions dense up to fundus, involving hepatic flexure or duodenum	Grade 4 *Gallbladder* completely obscured, empyema, gangrene, mass Cystic pedicle impossible to clarify Adhesions dense, fibrosis, wrapping the gallbladder, duodenum, or hepatic flexure difficult to separate.
The worst factor found should be used to define the fine overall grade.	

FIGURE 2.3: Description of the different subscores of the nassar grade corresponding to a specific grade [16].

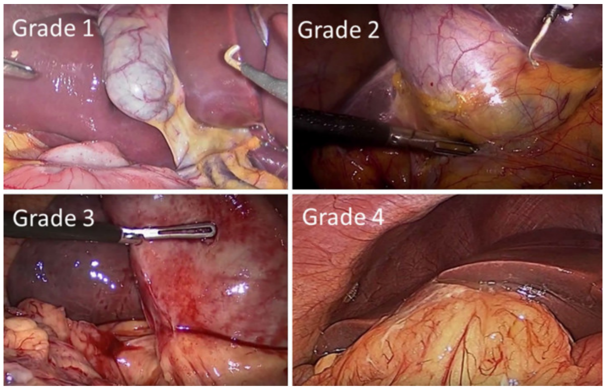


FIGURE 2.4: Visual example of the different Nassar grades [14]

2.2 Technical background

In this thesis, we work with two distinct types of data. The first type consists of patient characteristics, including age, body mass index (BMI), and history of prior surgeries. We employ a subset of machine learning (ML) methods to estimate operation times from this data. The two regression models are explained, and the evaluation metrics are briefly discussed in the "Regression Models" section. The second data type involves video recordings, from which difficulty ratings are derived. This is accomplished using a deep learning model. An overview of the fundamentals of Convolutional Neural Networks (CNNs) is included in the "Deep Learning" section to provide an understanding of the core mechanisms behind this approach. Finally, the specific framework used for video analysis is described.

2.2.1 Regression models

Machine learning is a powerful tool in predictive modelling, enabling computers to learn from data and make predictions or classifications. By leveraging algorithms that identify patterns in complex datasets, machine learning is now widely applied across various fields, including healthcare, finance, and engineering [17]. Regression algorithms are often employed in predictive tasks, such as estimating a numerical outcome or identifying trends. These algorithms aim to model the relationship between input variables (features) and continuous output variables (dependent variables). Among the diverse array of regression techniques, the Linear Regressor (LR) and the Random Forest Regressor (RFR) are two widely utilized models, each with distinct characteristics and advantages [18].

Linear regressor

LR is one of the most fundamental and interpretable algorithms in machine learning [19]. The core principle of LR is to establish a linear relationship between the independent and dependent variables. The equation of the LR is described in Equation 2.1, where $\hat{y}_i$ is the prediction of the model, β is the weight of each feature, X is the input for each of the features and ϵ is the error present.

$$\hat{Y}_i = \beta_0 + \beta_1 X_1 + \beta_n X_n + \epsilon \tag{2.1}$$

One of LR's key strengths is its simplicity and interpretability, which allow for a straightforward understanding of how each predictor affects the outcome. However, LR is based on several assumptions, such as linearity, constant variance of errors, and the absence of multicollinearity among predictors. This limits its effectiveness in capturing non-linear relationships.

Random forest regressor

A Random Forest Regressor (RFR) is built using multiple decision trees. A decision tree is a straightforward machine learning model composed of several nodes, where the model makes decisions based on the input features. After passing through each node, the data moves to the next one until it reaches a terminal node, which provides the final output. However, a single decision tree can be prone to overfitting, which is why RFR is employed. An RFR is constructed from multiple decision trees, and the final output is determined by a majority vote of the individual trees [20]. Figure 2.5 illustrates the difference between a single tree and multiple trees.

RFR offers several advantages over traditional regression models. It is highly flexible and can capture non-linear relationships between the features and the target variable. Additionally, aggregating predictions from multiple trees makes them more generalizable to unseen data. Moreover, RFRs can handle a large number of features. However, despite its robustness and flexibility, RFR can be computationally intensive, mainly when dealing with large datasets or many trees. Furthermore, the model's complexity can make it less interpretable than simpler models like LR.

Evaluation metrics

Two metrics are used to evaluate the regression models. The first metric is the mean squared error (MSE), and the second is the R^2-score. The MSE is an absolute measure to evaluate a model, which is presented in Equation 2.2,

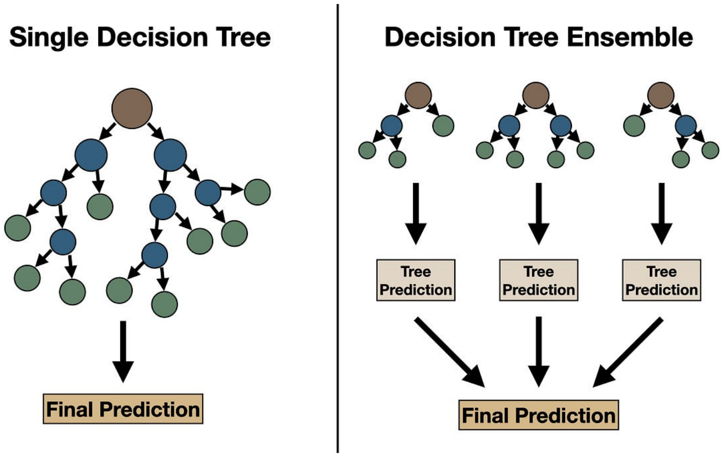


FIGURE 2.5: Visualisation of the difference between a single decision tree and an RFR[21].

where y represents the actual value and $\hat{y}$ the predicted value [22]. It determines the squared distance from the model prediction to the actual value. This measure can never be negative. The RMSE (root mean squared error) represents the same value but in actual units.

The R^2-score is a metric that indicates how well a model fits the data. The R^2-score is situated between 0 and 1, with 1 indicating the model can explain all variance of the model and 0 indicating the model cannot explain any of the variance in the data [23]. The R^2-score is calculated using the sum of squares of residuals (SSR) and the total sum of squares (TSS) as shown in 2.3. y denotes the actual value, $\hat{y}$ denotes the predicted value, and $\bar{y}$ presents the mean value. It is important to note that the R^2-score itself is not squared; only the separate parts of it are. If the SSR is bigger than the TSS, the R^2-score can be lower than zero. This happens in cases where using the mean as a predictor instead of the model results in better predictions. Based on these metrics, the models will be fine-tuned, and the best-performing model will be evaluated.

$$MSE = \frac{1}{n}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2.2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} = 1 - \frac{SSR}{TSS} \tag{2.3}$$

2.2.2 Deep-learning

We use a deep learning model called the multiscale visions transformer version 2 (MViTv2) to extract difficulty gradings from surgical videos. To provide a clear understanding of how the model works, we begin with a general overview of artificial neural networks (ANNs). Next, we focus on convolutional neural networks (CNNs), widely used for extracting information from images. Finally, we introduce the model used in this thesis, the Vision Transformer (ViT). A detailed explanation of the specific ViT employed in this research is provided in Chapter 5.

Artificial neural networks

Deep learning methods are based on artificial neural networks (ANNs). ANNs consist of neurons, which function similarly to those in the human brain [24]. These neurons are organized in layers and interconnected within a broad network. Each neuron contains a weight and a bias, as shown in Figure **??**. Neurons are either connected to previous neurons or directly to inputs. Upon receiving an input, a neuron multiplies it by a specific weight, which determines the signal's direction and strength. The product of the weight and input is then added to the bias, a constant term unique to each neuron. Note that a neuron can receive multiple inputs, each with its weight. The sum of all weighted inputs is combined with the bias and passed to the activation function, which decides whether the neuron will fire. The combination of weights and biases ultimately determines whether a neuron produces an output for a given input. Each distinct pattern of activated neurons is associated with a specific output. In this way, an ANN can make a classification or prediction for a particular input.

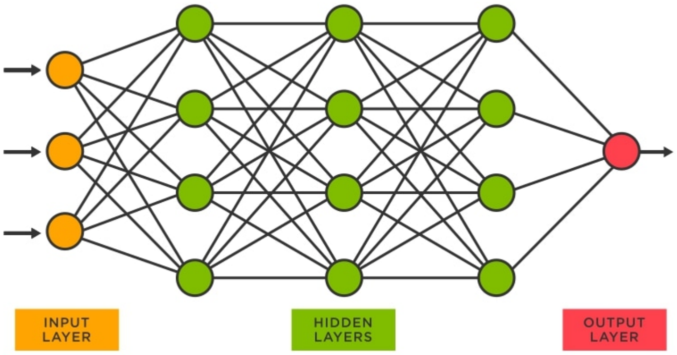


FIGURE 2.6: Overview of a neural network structure. In yellow, the input layer is defined. In green, the hidden layers are visualized. Within these layers, the computations are done to generate an output based on the weights and biases of the hidden layers. Red denotes the output layer [25].

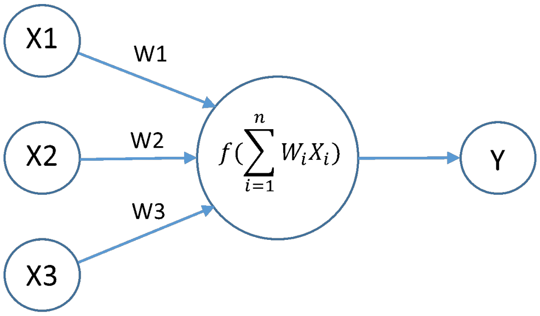


FIGURE 2.7: Representation of a single neuron. X represents the input, and W denotes the weight. In the neuron, the weights are summed and multiplied with the bias[26].

For a neural network to accurately produce predictions, it must tune the weights and biases. This is achieved through forward and backward propagation [27]. The input is presented to the model, and all necessary calculations are performed. Next, back-propagation is initiated. An error is calculated by comparing the model's output with the ground truth of the input images. Based on this error, a gradient is computed to update the weights and biases. After the update, the input is presented again, and the process is repeated until the error reaches the desired level. In this way, a neural network can make non-linear predictions for complex data. ANNs can also process image data. However, a specific version of an ANN is required to effectively capture information from an image. For this purpose, convolutional neural networks (CNNs) are commonly used.

2.2.3 Convolutional neural networks

CNNs are a type of deep neural network designed for image classification [29]. The goal of a CNN is to extract different levels of features from an image and output a prediction or a class, which is shown in Figure 2.8. This is done using different types of layers. Most novel models are based on the basic layers and are fine-tuned using new techniques or adding extra information. These basic layers essential to highlight and to understand the different models we use are the convolutional layer, the pooling layer, and the fully connected layer.

Convolutional layer

The first layer in a CNN Is the convolutional layer [30]. This layer is used to extract features from an image. The convolution is done using a filter. The size of this filter can vary. Typically, a 2 x 2 or 3 x 3 filter is used. These filters are placed on the image. The filter contains a subset of numbers used to calculate a convolution of the part of the image covered by the filter. This filter outputs one number representing a feature corresponding to that given portion of the image. The next step is to slide the filter over the image until all pixels are used. This leaves a smaller subset of the original image representing features. This can be repeated using different filters to extend the feature space. Figure 2.8 represents this using the small yellow square.

Pooling layer

A convolutional layer is often directly followed by a pooling layer [30]. A pooling layer reduces the number of features from the convolutional layer. This is also done by combining the information of multiple features into one, as represented in Figure 2.9. There are two main variants to conduct pooling: average pooling or max pooling.

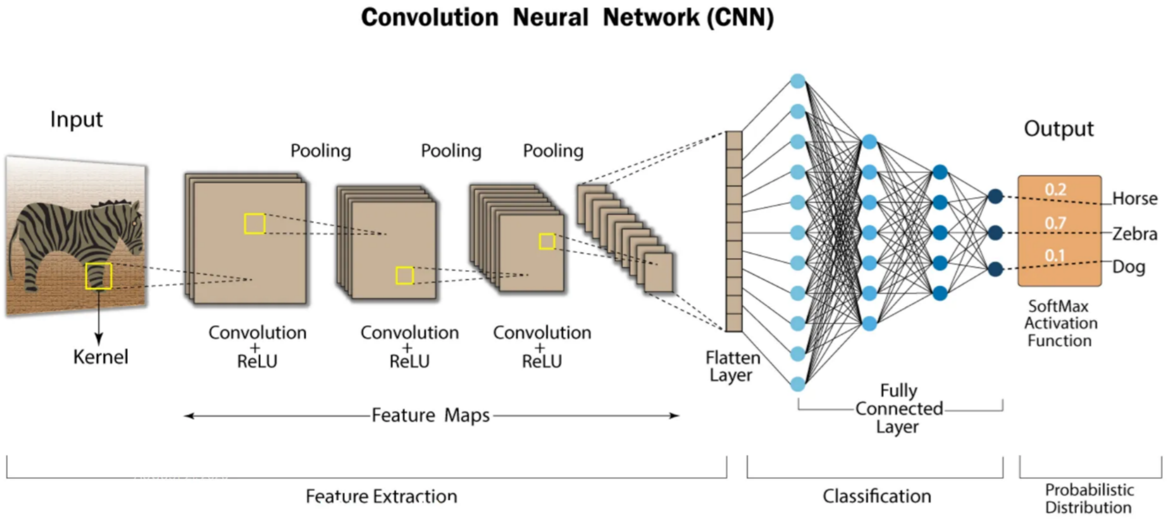
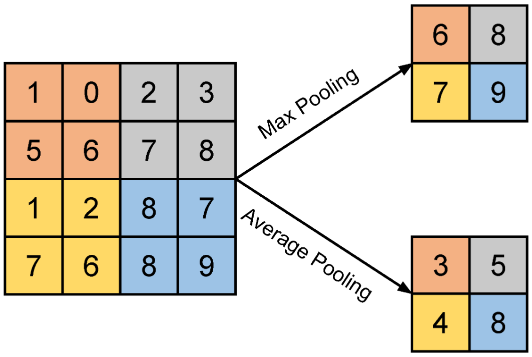


FIGURE 2.8: General overview of a convolutional neural network [28]

When applying the pooling layer, a lot of information is lost. However, it also has some main benefits: they help to reduce complexity, improve efficiency, and limit the risk of overfitting.



1	0	2	3
5	6	7	8
1	2	8	7
7	6	8	9

6	8
7	9

3	5
4	8

FIGURE 2.9: Use of pooling and the different types of pooling [31].

Fully connected layer

After all desired convolutions, we are left with a volume of features [30]. These features are used as input for an ANN. To achieve this, the feature space is flattened so it has 1XN dimensions. Next, a fully connected layer is introduced, in which all neurons of the flattened layer are connected to the neurons of the fully connected layer. The neurons of the fully connected layer generate an output that, with the help of an activation function, can be transformed into a probability for one of the classes. The softmax function is the most commonly used function, which maps the outputs to a probability between 0 and 1.

2.2.4 Vision transformers

A Vision Transformer (ViT) is derived from the transformer architecture used in networks like ChatGPT. ViTs use self-attention to focus on different parts of the input sequence when processing each element. Self-attention uses three matrices called the query, key, and value matrices [**?**]. Figure 2.10. These matrices are all three linear projections of the original embedding and are used to calculate attention scores. We compare the concept with a general information retrieval system to grasp it. If you want, for example, to search for a specific video on YouTube, you use the search bar. The search algorithm processes the query (the text entered in the search bar) by matching it against a set of keys (metadata such as video titles, descriptions, tags, etc.) associated with potential candidate videos in its database. The system returns the most relevant results (values), representing the best-matching videos based on the query and associated keys. So, in the context of images, we present specific images to the system, and within the database, the model identifies which keys best fit the presented image. After calculating this, the model defines which values to assign to this combination of queries and keys. This value matrix can finally give a final prediction of the image.

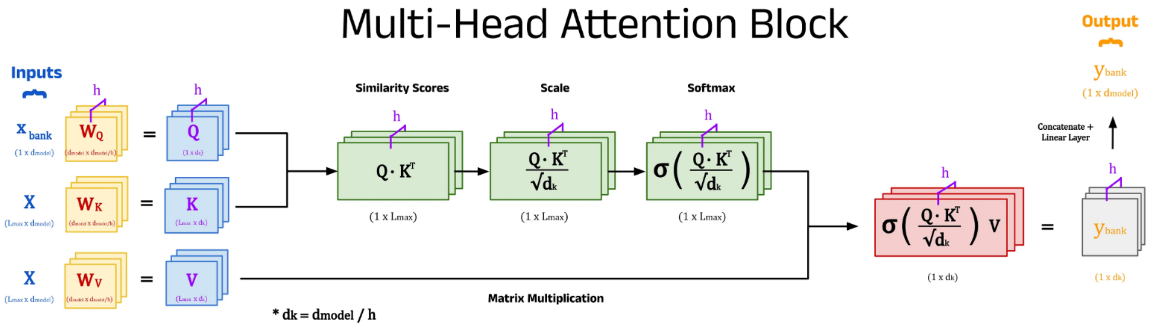


FIGURE 2.10: Example of the self attention [32]

The images need to be vectorised to use this self-attention system in the context of ViTs. This is done by dividing the image into different parts called patches. Next, these patches are fed into a ResNet50, a type of CNN. From this CNN, the fully connected layer is removed so that the model's output is a feature space that can serve as input for the ViT. These representations are labelled with a location vector so that the model can define the position of each of the individual patches. Next, these representations are fed into the self-attention mechanism called a transformer block. Multiple transformer blocks can be stacked together and comparable with the different layers of a CNN. Within these transformer blocks, different variations exist to introduce non-linearity and compute the different matrices.

The ViT can conceptually be seen as an extension of the CNN that uses the extracted features to add an additional layer of information [33]. Due to this extra layer of information, ViTs can link different parts of the input image to each other and, therefore, understand the context. A CNN is more focused on tracing more and more high-level features and has a more narrow focus. However, because the query, key, and value matrices all contain trainable weights, a ViT includes a lot of parameters. Therefore, ViTs need a lot more data than conventional CNNs. One significant advantage of the ViT is that it primarily focuses on matrix multiplications. Thus, a lot of the computing can be done in parallel. This is done by using multi-head attention instead of single-head attention. Multi-head attention improves performance, delivers a more complete representation because each head can focus on different aspects of the data, and improves flexibility.

Bibliography

[1] Xin Wang, Wenqian Yu, Guoheng Jiang, Hongyu Li, Shiyi Li, Linjun Xie, Xuan Bai, Ping Cui, Qi Chen, Yanmei Lou, et al. Global epidemiology of gallstones in the 21st century: a systematic review and meta-analysis. *Clinical Gastroenterology and Hepatology*, 2024.

[2] Jasmin Tanaja, Richard A Lopez, and Jehangir M Meer. Cholelithiasis. 2017.

[3] Mahendra Lodha, Anupam S Chauhan, Ashok Puranik, Satya Prakash Meena, Mayank Badkur, Ramkaran Chaudhary, Indra Singh Chaudhary, Metlapalli V Sairam, Vinod Kumar, and Rashi Lodha. Clinical profile and evaluation of outcomes of symptomatic gallstone disease in the senior citizen population. *Cureus*, 14(8), 2022.

[4] Raam Mannam, Rajagopal Sankara Narayanan, Arpit Bansal, Vishnu R Yanamaladoddi, Sai Suseel Sarvepalli, Shree Laya Vemula, and Saikumar Aramadaka. Laparoscopic cholecystectomy versus open cholecystectomy in acute cholecystitis: a literature review. *Cureus*, 15(9), 2023.

[5] Antonio P Legorreta, Jeffrey H Silber, George N Costantino, Richard W Kobylinski, and Steven L Zatz. Increased cholecystectomy rate after the introduction of laparoscopic cholecystectomy. *Jama*, 270(12):1429–1432, 1993.

[6] Miodrag Radunovic, Ranko Lazovic, Natasa Popovic, Milorad Magdelinic, Milutin Bulajic, Lenka Radunovic, Marko Vukovic, and Miroslav Radunovic. Complications of laparoscopic cholecystectomy: Experience from a retrospective analysis. *Open Access Macedonian Journal of Medical Sciences*, 4, 11 2016.

[7] RG Andall, Petru Matusz, Maira du Plessis, Robert Ward, RS Tubbs, and Marios Loukas. The clinical anatomy of cystic artery variations: a review of over 9800 cases. *Surgical and Radiologic Anatomy*, 38:529–539, 2016.

[8] Matt Baguley. The gallbladder, 2024. Image of the different regions of the gallbladder.

[9] Anne MR Agur and Arthur F Dalley II. *Grant's atlas of anatomy*. Lippincott Williams & Wilkins, 2023.

[10] Andrea T Fisher, Kovi E Bessoff, Rida I Khan, Gavin C Touponse, MK Maggie, Advait A Patil, Jeff Choi, Christopher D Stave, and Joseph D Forrester. Evidence-based surgery for laparoscopic cholecystectomy. *Surgery open science*, 10:116–134, 2022.

[11] Vishal Gupta and Gaurav Jain. Safe laparoscopic cholecystectomy: Adoption of universal culture of safety in cholecystectomy. *World journal of gastrointestinal surgery*, 11(2):62, 2019.

[12] Steven M Strasberg and L Michael Brunt. The critical view of safety: why it is not the only method of ductal identification within the standard of care in laparoscopic cholecystectomy. *Annals of surgery*, 265(3):464–465, 2017.

[13] Giulia Missori, Francesco Serra, Roberta Gelmini, et al. A narrative review about difficult laparoscopic cholecystectomy: technical tips. *Laparoscopic Surgery*, 6:24–24, 2022.

[14] Ewen A Griffiths, James Hodson, Ravi S Vohra, Paul Marriott, Tarek Katbeh, Samer Zino, Ahmad HM Nassar, and West Midlands Research Collaborative. Utilisation of an operative difficulty grading scale for laparoscopic cholecystectomy. *Surgical endoscopy*, 33:110–121, 2019.

[15] E Suliman and RȘ Palade. Importance of cystic pedicle dissection in laparoscopic cholecystectomy in order to avoid the common bile duct injuries. *Journal of medicine and life*, 9(1):44, 2016.

[16] Ruby Egging. Ai-based prediction model of surgical difficulty in laparoscopic cholecystectomy, 10 2023.

[17] Diyana Kyuchukova, G.V. Hristov, Petko Kyuchukov, Georgiev, and Rosen Daskalov. Machine learning algorithms for regression analysis and predictions of numerical data. pages 1–6, 06 2021.

[18] Paul F Smith, Siva Ganesh, and Ping Liu. A comparison of random forest regression and multiple linear regression for prediction in neuroscience. *Journal of neuroscience methods*, 220(1):85–91, 2013.

[19] Dastan Maulud and Adnan Abdulazeez. A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, 1:140–147, 12 2020.

[20] Raphael Couronné, Philipp Probst, and Anne-Laure Boulesteix. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC bioinformatics*, 19:1–14, 2018.

[21] Shaw Talebi. 10 decision trees are better than 1, 2024. Difference between one decision tree and random forest.

[22] Timothy Hodson, Thomas Over, and Sydney Foks. Mean squared error, deconstructed. *Journal of Advances in Modeling Earth Systems*, 13, 12 2021.

[23] Xue Ying. An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, 1168:022022, 02 2019.

[24] Iqbal H Sarker. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN computer science*, 2(6):420, 2021.

[25] spotfire. What is a neural network?, 2024. Image representing the different layers in a ANN.

[26] Parminder Singh. Neuron explained using simple algebra, 2024. Image representing a neuron.

[27] Saeed Mouloodi, Hadi Rahmanpanah, Soheil Gohari, Colin Burvill, and Helen MS Davies. Feedforward backpropagation artificial neural networks for predicting mechanical responses in complex nonlinear structures: A study on a long bone. *Journal of the Mechanical Behavior of Biomedical Materials*, 128:105079, 2022.

[28] Pratham Modi. Convolutional neural networks for dummies, 2024. Overview of CNN.

[29] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, 9:611–629, 2018.

[30] Purwono Purwono, Alfian Ma'arif, Wahyu Rahmaniar, Haris Imam, Haris Imam Karim Fathurrahman, Aufaclav Frisky, and Qazi Mazhar Ul Haq. Understanding of convolutional neural network (cnn): A review. 2:739–748, 01 2023.

[31] Muhamad Yani, S Irawan, and Casi Setianingsih. Application of transfer learning using convolutional neural network method for early detection of terry's nail. *Journal of Physics: Conference Series*, 1201:012052, 05 2019.

[32] Bradney Smith. Self-attention explained with code, 2024. Overview of transformerblock.

[33] Shuoxi Zhang, Hanpeng Liu, Stephen Lin, and Kun He. You only need less attention at each stage in vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6057–6066, 2024.

3

Applicability of operation time prediction using preoperative patient factors

Applicability of operation time prediction using preoperative patient factors

V.P.S. Oosterhoff, BSc.[1], S.C. Baltus, MSc.[1,2], Dr. C.O. Tan[1], Prof. Dr. I.A.M.J. Broeders[2]

Abstract

Background: Accurate estimation of surgery duration can lead to cost-effective utilization of surgical staff and operating rooms and decrease patients' waiting time. This study aimed to determine the accuracy with which the operation room (OR) time is scheduled within the Meander Medical Center (MMC). Moreover, we attempted to improve OR planning using two different machine learning (ML) approaches.

Methods: Surgical data of elective cholecystectomies from 2017 to 2023 were used to determine whether the planned operation time matched the actual operation time. Moreover, a univariate analysis was done to determine the correlation between age, body mass index (BMI), sex, American Standards Association (ASA) score, and the level of expertise with the actual operation time. Next, the importance of each of these features was determined using forward and backward feature selection. Lastly, the selected features were used to train two ML algorithms: Linear regression (LR) and Random Forest Regression (RFR). The predictions of both models were compared with the conventional planning method.

Results: 848 elective laparoscopic cholecystectomies were included in the model. In 21%, the absolute error between actual and planned operation time was greater than 15 minutes. The features BMI, sex, and level of expertise were statistically significant with a correlation coefficient of respectively 0.21, -0.083, and 0.096. The model that fitted the data best was the LR model using the features BMI, sex, and level of expertise, resulting in a root mean squared error (RMSE) of 14.18 (±0.58) minutes. The conventional planning method achieved a RMSE of 16.22 (±0.50) minutes. The mean R^2-score of the best model was 0.040, which indicates a poor explanation of the variance in the data

Conclusion: Much OR time is lost using the current planning method. Improving operational planning could lead to a more efficient use of OR time. The use of ML models can be promising; however, the features currently used in the model were not sufficient to describe the variance in the data.

Keywords
Operation time — Cholecystectomy — Linear regression — Random forest regression

[1] *Technical medicine, University Twente, Enschede, The Netherlands*
[2] *Department of Surgery, Meander Medical Center, Amersfoort, Netherlands*

1. Introduction

In modern hospitals, operating rooms (ORs) are both a significant source of costs and revenue [1]. Due to delayed surgeries, the surgery schedule can deviate significantly [2]. This can affect hospital costs and care delivery due to surgery cancellation, nursing staff turnovers due to conflict in planning, staff overtime, and employee dissatisfaction [3, 4, 5]. Optimizing ORs' scheduling and use is crucial for effective hospital management.

Several attempts are made to improve OR planning using simple historical data for a given procedure and then calculating an average duration [6]. This can involve using the surgeon's estimate for the given operation or taking the average duration of only the most recent records of the same procedure [7]. However, these simple methods do not accurately capture the high variance between patients.

A possible solution for this problem could be using patient-specific and surgery-specific factors [8, 9]. The relationship between these factors and operation varies across different studies. Additionally, the methods described in the literature show considerable diversity, ranging from basic machine learning (ML) algorithms to advanced deep learning (DL) techniques [8, 10, 11, 12]. Despite these variations, most articles reported promising results for improving operational planning.

This study aims to determine the accuracy of the current planning method in the Meander Medical Center (MMC) and explore the possibility of improving surgery scheduling by using ML models. Our study focuses on two key objectives. First, we will evaluate the current method used in clinical practice by comparing the actual operation duration with the planned time. Second, we will explore the feasibility of using ML algorithms on data from the MMC.

2. Methods

To research the possibilities for operation time assessment, we did a retrospective analysis on a dataset containing surgical information of all patients within the MMC of the past 5 years. After acquiring the dataset, outliers were filtered out, and the relevant parameters were determined. Next, the univariate correlation of the parameters with the actual operation time was determined. A multivariate analysis was done using forward and backward feature selection. Lastly, two ML models were trained and compared with the planned operation time by the OR schedulers.

2.1 Dataset

After the internal review board's approval, a retrospective dataset was used. This dataset contains all operations done in the MMC over the past five years. Emergency procedures, patients with concurrent procedures, and patients under 18 were excluded. Moreover, outliers within the actual surgery time were filtered out based on the interquartile range (IQR).

2.2 Outcome variable

The primary outcome variable was the actual operation time, which was defined as the time from the first incision to skin closure. After model training, the predictions of the best models were compared with the originally planned operation time to assess whether the models improved the current OR planning.

2.3 Independent variables

Independent factors for each case were examined to determine their association with operative duration. Independent variables included in this study were the patient factors age, sex, body mass index (BMI), and American Society of Anesthesiologists (ASA) class) and the non-patient factor level of expertise of the performing surgeon. The level of expertise existed in two classes: resident and fully trained surgeon. The continuous variables age and BMI were linearly normalized between zero and one. The ordinal variable ASA score was one hot encoded, creating three separate ASA classes. The variable's level of expertise and sex were binarized.

2.4 Feature selection

Feature selection was applied to optimize the model and filter out possible redundant features. This was done in two steps:

1. The variance inflation factor (VIF) was calculated for each feature. Features with a VIF bigger than ten were excluded from the dataset.
2. A forward and backward approach was used to filter out redundant features. This was done for five folds to determine the stability of the selected feature sets. The forward and backward feature selection performance was evaluated using the R^2-score. After five-folds, the features present three times in either the forward or backward feature selection were selected.

2.5 Model fine-tuning and evaluation

The final step was to fine-tune the model on the selected parameters. Due to the simplicity of the LR, this was done only for the RFR. Table 1 shows the selection of hyperparameters used. The selected features were used to train both models using five-fold cross-validation. In addition, the influence of the feature selection was determined by training the models on the complete feature set. Finally, the trained models were compared with the planned operation time. This was done by comparing the RMSE and the R^2-score.

Table 1. Hyperparameters used for RFR

Hyperparameter	Values
Max depth	3, 5, 7
Minimal samples per leaf	1, 2, 4
Minimal samples per split	2, 5, 10
Number of estimators	100, 200, 300

3. Results

After data cleaning, 848 cases were included in the analysis. The patient demographics are presented in Table 2. All continuous independent variables, along with the actual and planned operation times, were not normally distributed. The median actual operation time was 50 minutes (19 [60:41]), and the median planned operation time was 45 minutes (0 [45:45]). Figure 1 shows that, in most cases, a planned operation time of 45 minutes was used, which does not match the distribution of the actual operation times. The absolute error between the actual and planned operation times was illustrated in Appendix A.1. In Table 3, the individual correlation of the dependent variable with the independent variables is shown. The Variables BMI, sex, and level of expertise were significantly correlated with the actual operation time.

3.1 Feature selection

After calculating the VIF, only the feature ASA score 1 was excluded. Both the forward and backward feature selection process is done using five-fold cross-validation. In Appendix B.1 & B.2 the influence of the number of features on the R^2-score is visualized. Figure 2 shows the features' prevalence. On the y-axis, the frequency of the features is plotted, and on the x-axis features are plotted. The selected features for the LR model were age, BMI, sex, and level of expertise. The features chosen for the RFR were age, sex, ASA Score 2, ASA score 3, and Level of expertise.

3.2 Model fine-tuning and evaluation

The optimal hyperparameters for the RFR were a maximum depth of 10, a minimum sample per leaf of 4, a minimum sample per split of 10, and 200 estimators. The models were successfully trained using the selected feature set. Moreover, both models were also trained using all features. The results are shown in Table 4. The LR performed the best. In Appendix C.1, the best fold for both models is presented in a Bland–Altman plot.

Table 2. Patient characteristics

Characteristics	N(%)[1]	Shapiro-Wilk	P-value
Sex			
Males	561 (66.2%)		
Females	287 (33.8%)		
Age	54.0 (24 [42.0: 66.0])	0.98	9.25e-9
Level of expertise			
Surgeon	284 (33.5%)		
Resident	564 (66.5%)		
ASA			
1	240 (28.3%)		
2	492 (58.0%)		
3	116 (13.7%)		
BMI	27.45 (6.32 [24.49: 30.81])	0.97	4.83e-13

[1] Median is denoted as Median (IQR [Q1:Q3]).

Table 3. Correlation between the different features and the outcome variable

Predictor	Correlation	P-value	Correlation Type
Age	-0.036	0.29	Pearson
BMI	0.21	1.3e-8	Pearson
Sex	-0.083	0.015	Point-Biserial
Level of expertise	0.096	0.0052	Point-Biserial
ASA 1	-0.059	0.086	Point-Biserial
ASA 2	0.029	0.39	Point-Biserial
ASA 3	0.035	0.30	Point-Biserial

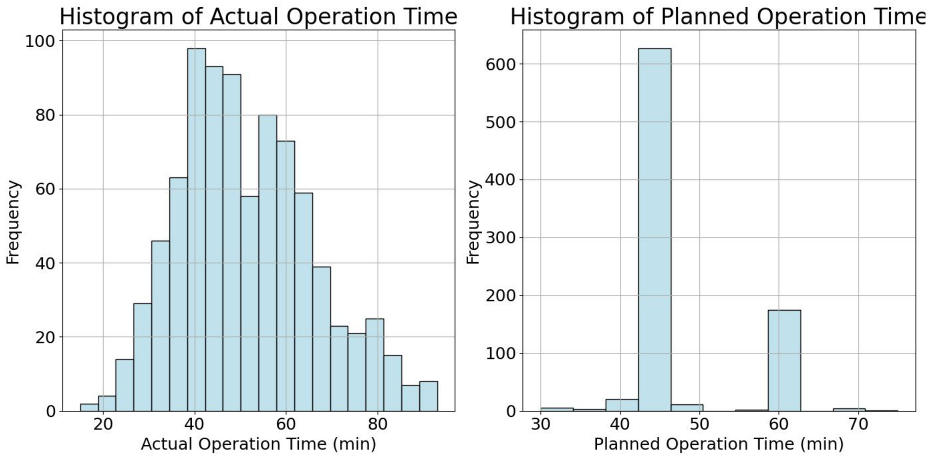


Figure 1. The distribution of the actual operation time and the planned operation time

4. Discussion

Operating room (OR) time is expensive and a limited resource. Much research is done to optimize operative workflow and minimize costs [13, 14, 15, 16]. In this study, we trained two basic ML models to investigate the possibility of predicting operation time using five preoperative features in a dataset of 848 patients. The models were cross-validated 5 times to increase model stability. Our best model was the LR model trained with a selection of the available features with a R^2-score of 0.040 ±0.040 and a root mean squared error (RMSE) of 14.18 minutes ±0.58, which makes it outperform the current planning method, which has a RMSE of 16.22 ±0.50 minutes. As seen in Appendix C.1 the model predictions were mainly around the mean operation time. This is caused because, as visualized by the R^2-score, the model cannot yet describe the variance in the data set.

A study by Thiels et al. also attempted to predict chole-

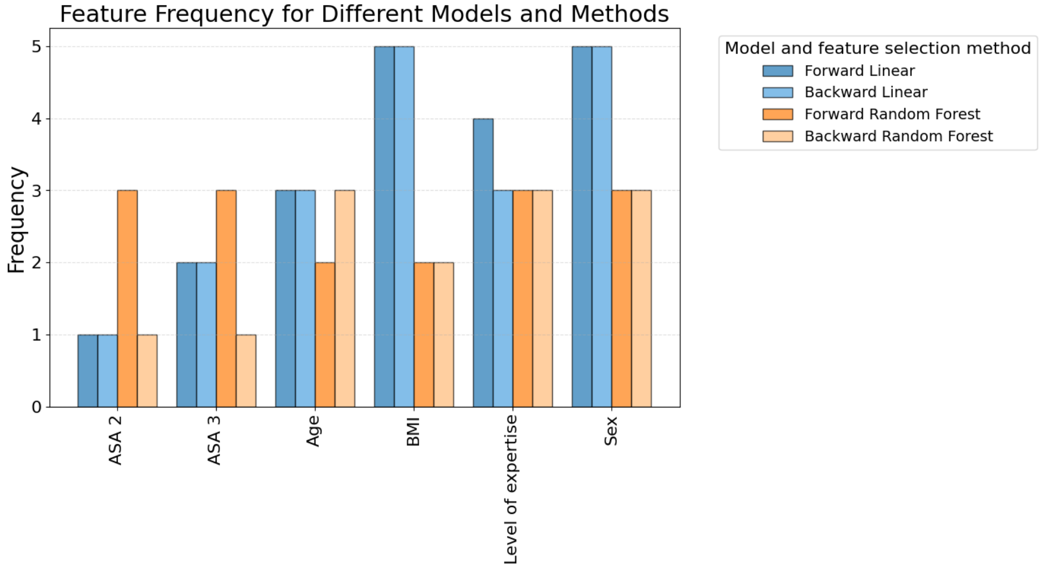


Figure 2. Influence of amount of features on model performance after five-fold cross-validation for backward feature selection

Table 4. Results of the five-fold cross-validation

Model	R^2-score	RMSE
LR model		
Selected features	0.040 ±0.040	14.18 ±0.58
All features	0.036 ±0.039	14.21 ±0.057
RFR		
Selected features	-0.087 ±0.045	15.01 ±0.56
All features	-0.029 ±0.064	14.68 ±0.70
Planned	-0.26 ±0.088	16.22 ±0.50

cystectomy operation time [8]. They created a stable model incorporating patient factors and surgical expertise. However, their mean operative time was significantly higher than in our dataset, resulting in a standard deviation of 32 minutes, which is, for clinical practice, not yet workable. Their model achieved a R^2-score of 0.18, which is relatively low for ML purposes [17]. It should be noted that their model could filter out outliers present in the dataset. Outliers impact surgical scheduling the most, therefore a model that can effectively predict outliers is desirable. The reason Thiels et al. achieved higher R^2-scores is probably due to the use of more parameters. In their study, more parameters were available that described the level of expertise. Moreover, they also used laboratory results of the patients. In our study, the amount of parameters and information they represent were more limited.

Other studies researched the applicability of operation time prediction for all operations within the OR complex using LR and RFR models. They report a RMSE between 26.09 minutes and 45.18 minutes and estimation errors between 15% and 40% [16, 18, 19]. In these studies, the developed model had more trouble predicting outliers and was more accurate in predicting cases that were more situated around the mean. This behaviour indicates that also in these studies, a limited amount of the variance can be predicted using patient-specific and surgery-specific factors.

In the present study features, Sex, Level of expertise, and BMI were found to be statistically significant, with BMI having the strongest correlation of 0.21. The finding of the correlated values is similar to other studies found in literature [20, 21]. However, other studies also indicated that gender was associated with a longer operation time. Our study did not show the feature ASA-score as a significant outcome variable. A study by Master et al. found that the ASA score had low importance within their models, suggesting this is because important information within the ASA score may already be coded more clearly within other variables, such as the patient's weight [22].

One of the limitations of our study is the limited amount of machine learning models fitted compared to other studies. In our study, we restricted ourselves to two traditional regression models. This is because, for the goal of operation time prediction, decision tree models often perform equally and sometimes even better than deep learning networks [23]. Moreover, the acquired dataset is relatively simple and does not contain many features and complex correlations. It could be discussed that using more complicated models and other types of feature selection could increase the model performance. However, it is not expected that it would increase the R^2-score drastically [24]

Another limitation of our study is the limited number of features available for model training. As visible in the model results, the selected features do not capture all the information needed to predict operation duration accurately. More specific features, such as preoperative medical imaging, are expected to improve the model significantly.

In conclusion, our current study emphasizes the operation scheduling problem in the OR. We found that in 21% of the cases, the prediction error is bigger than 15 minutes. Moreover, we fitted two ML models to determine the feasibility of preoperative operation time prediction. In the current setup, we were able to outperform the current scheduling method slightly. However, It is expected that if we incorporate more patient-specific features, we can further increase the model predictions and improve OR scheduling.

References

[1] C. P. Childers and M. Maggard-Gibbons, "Understanding costs of care in the operating room," *JAMA surgery*, vol. 153, no. 4, pp. e176233–e176233, 2018.

[2] R. Kaddoum, S. Tarraf, F. M. Shebbo, A. B. Ali, C. Karam, C. Abi Shadid, J. Bouez, and M. T. Aouad, "Reduction of nonoperative time using the induction room, parallel processing, and sugammadex: a randomized clinical trial," *Anesthesia & Analgesia*, vol. 135, no. 2, pp. 406–413, 2022.

[3] W. N. Schofield, G. L. Rubin, M. Piza, Y. Y. Lai, D. Sindhusake, M. R. Fearnside, and P. L. Klineberg, "Cancellation of operations on the day of intended surgery at a major australian referral hospital," *Medical Journal of Australia*, vol. 182, no. 12, pp. 612–615, 2005.

[4] T. P. Thompson and H. N. Brown, "Turnover of licensed nurses in skilled nursing facilities.," *Nursing Economics*, vol. 20, no. 2, 2002.

[5] E. Strachota, P. Normandin, N. O'brien, M. Clary, and B. Krukow, "Reasons registered nurses leave or change employment status," *JONA: The Journal of Nursing Administration*, vol. 33, no. 2, pp. 111–117, 2003.

[6] Z. ShahabiKargar, S. Khanna, N. Good, A. Sattar, J. Lind, and J. O'Dwyer, "Predicting procedure duration to improve scheduling of elective surgery," in *PRICAI 2014: Trends in Artificial Intelligence: 13th Pacific Rim International Conference on Artificial Intelligence, Gold Coast, QLD, Australia, December 1-5, 2014. Proceedings 13*, pp. 998–1009, Springer, 2014.

[7] J. Zhou, F. Dexter, A. Macario, and D. A. Lubarsky, "Relying solely on historical surgical times to estimate accurately future surgical times is unlikely to reduce the average length of time cases finish late," *Journal of clinical anesthesia*, vol. 11, no. 7, pp. 601–605, 1999.

[8] C. A. Thiels, D. Yu, A. M. Abdelrahman, E. B. Habermann, S. Hallbeck, K. S. Pasupathy, and J. Bingener, "The use of patient factors to improve the prediction of operative duration using laparoscopic cholecystectomy," *Surgical endoscopy*, vol. 31, pp. 333–340, 2017.

[9] V. Riahi, H. Hassanzadeh, S. Khanna, J. Boyle, F. Syed, B. Biki, E. Borkwood, and L. Sweeney, "Improving preoperative prediction of surgery duration," *BMC Health Services Research*, vol. 23, no. 1, p. 1343, 2023.

[10] C. Spence, O. A. Shah, A. Cebula, K. Tucker, D. Sochart, D. Kader, and V. Asopa, "Machine learning models to predict surgical case duration compared to current industry standards: scoping review," *BJS open*, vol. 7, no. 6, p. zrad113, 2023.

[11] O. Babayoff, O. Shehory, M. Shahoha, R. Sasportas, and A. Weiss-Meilik, "Surgery duration: optimized prediction and causality analysis," *Plos one*, vol. 17, no. 8, p. e0273831, 2022.

[12] M. Vannucci, G. G. Laracca, P. Mercantini, S. Perretta, N. Padoy, B. Dallemagne, and P. Mascagni, "Statistical models to preoperatively predict operative difficulty in laparoscopic cholecystectomy: a systematic review," *Surgery*, vol. 171, no. 5, pp. 1158–1167, 2022.

[13] W. C. Levine and P. F. Dunn, "Optimizing operating room scheduling," *Anesthesiology clinics*, vol. 33, no. 4, pp. 697–711, 2015.

[14] T. M. Ward, D. A. Hashimoto, Y. Ban, G. Rosman, and O. R. Meireles, "Artificial intelligence prediction of cholecystectomy operative course from automated identification of gallbladder inflammation," *Surgical Endoscopy*, vol. 36, no. 9, pp. 6832–6840, 2022.

[15] Y. Jiao, A. Sharma, A. Ben Abdallah, T. M. Maddox, and T. Kannampallil, "Probabilistic forecasting of surgical case duration using machine learning: model development and validation," *Journal of the American Medical Informatics Association*, vol. 27, no. 12, pp. 1885–1893, 2020.

[16] O. Martinez, C. Martinez, C. Parra, S. Rugeles, and D. Suarez, "Machine learning for surgical time prediction," *Computer Methods and Programs in Biomedicine*, vol. 208, p. 106220, 06 2021.

[17] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation," *Peerj computer science*, vol. 7, p. e623, 2021.

[18] S. Lam, H. Zaribafzadeh, B. Ang, W. Webster, D. Buckland, C. Mantyh, and H. Tan, "Estimation of surgery durations using machine learning methods-a cross-country multi-site collaborative study," *Healthcare*, vol. 10, p. 1191, 06 2022.

[19] M. Fairley, D. Scheinker, and M. L. Brandeau, "Improving the efficiency of the operating room environment with an optimization and machine learning model," *Health care management science*, vol. 22, pp. 756–767, 2019.

[20] T. MD *et al.*, "Validation of a scoring system to predict difficult laparoscopic cholecystectomy: a one-year cross-sectional study," *Journal of the West African College of Surgeons*, vol. 8, no. 1, p. 23, 2018.

[21] A. Ary Wibowo, O. Tri Joko Putra, Z. Noor Helmi, H. Poerwosusanta, T. Kelono Utomo, and K. Marwan Sikumbang, “A scoring system to predict difficult laparoscopic cholecystectomy: A five-year cross-sectional study,” *Minimally Invasive Surgery*, vol. 2022, no. 1, p. 3530568, 2022.

[22] N. Master, Z. Zhou, D. Miller, D. Scheinker, N. Bambos, and P. Glynn, “Improving predictions of pediatric surgical durations with supervised learning,” *International Journal of Data Science and Analytics*, vol. 4, pp. 35–52, 2017.

[23] B. Zhao, R. S. Waterman, R. D. Urman, and R. A. Gabriel, “A machine learning approach to predicting case duration for robot-assisted surgery,” *Journal of medical systems*, vol. 43, no. 2, p. 32, 2019.

[24] M. Nasr, A. Abdelmegaly, and D. Abdo, “Performance evaluation of different regression models: application in a breast cancer patient data,” *Scientific Reports*, vol. 14, p. 12986, 06 2024.

Appendix

A. Distribution of the error of the planned operation time

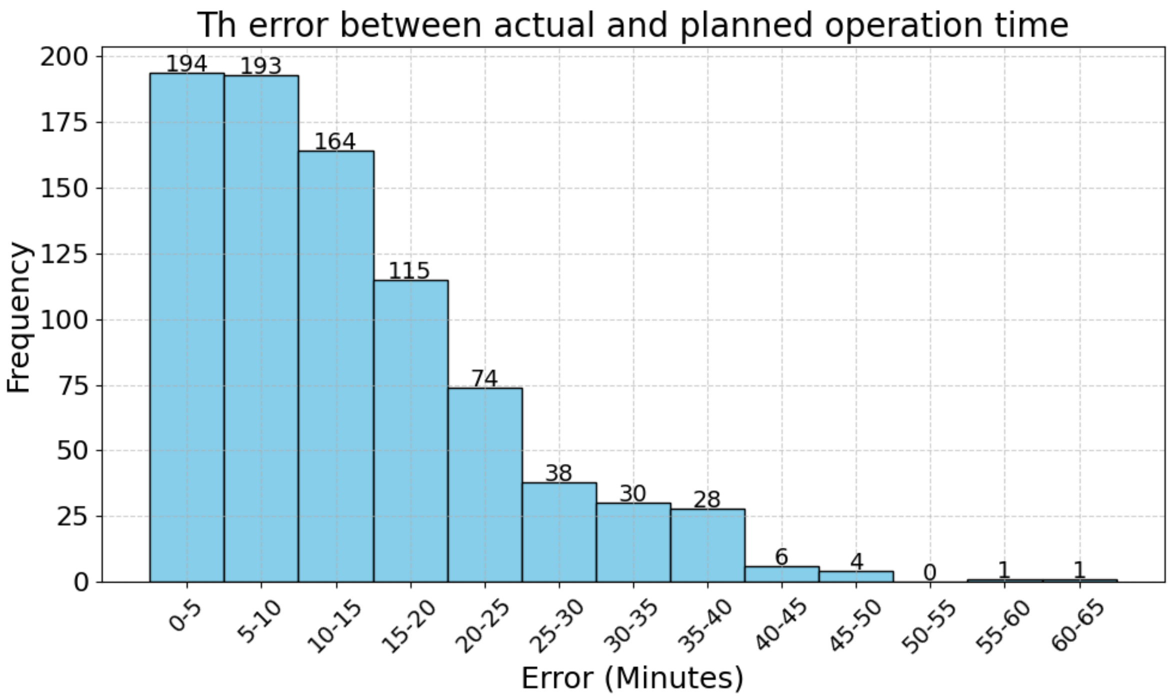


Figure A.1. The error between the actual operation time and the planned operation time

B. Feature selection plots of forward and backward feature selection

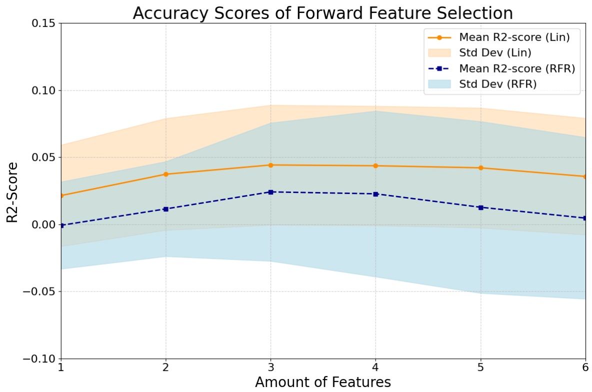


Figure B.1. Influence of amount of features on model performance after five-fold cross-validation for forward feature selection

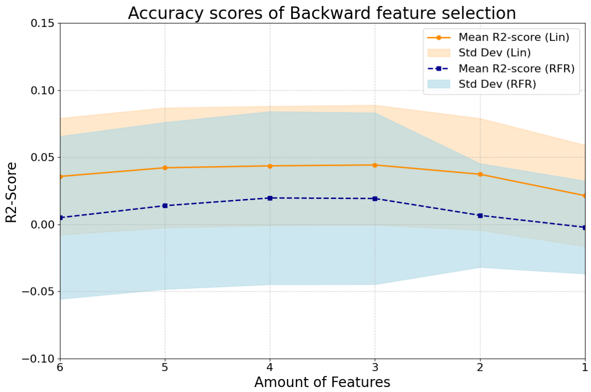


Figure B.2. Influence of amount of features on model performance after five-fold cross-validation for backward feature selection

C. Bland-Altmann plot of the best prediction of the LR and RFR

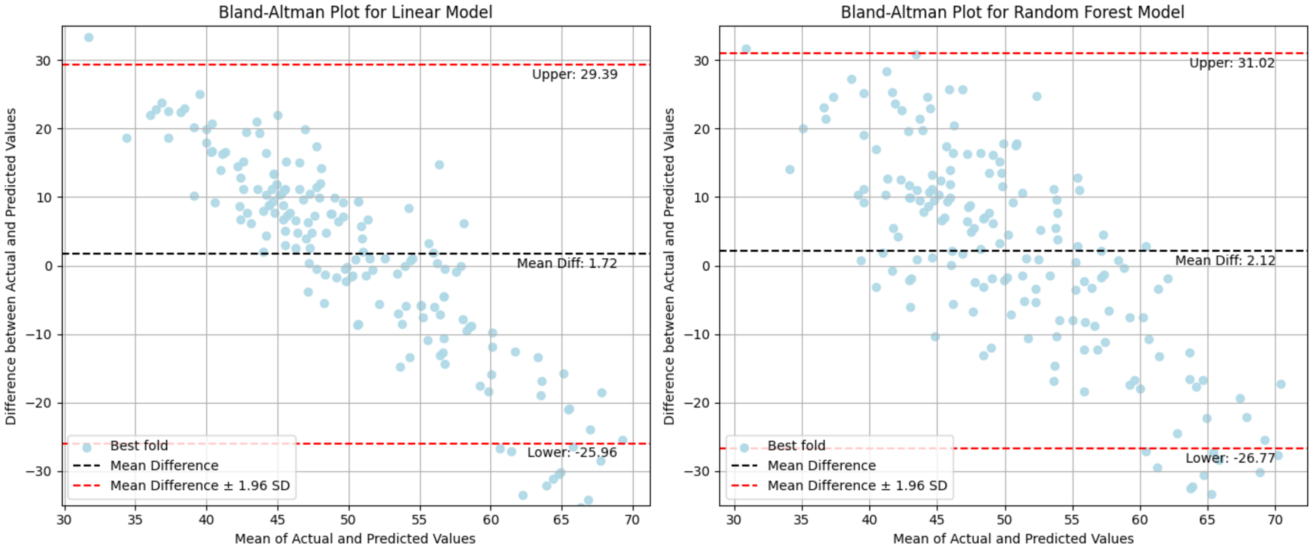


Figure C.1. Bland–Altman plot of the best predictions of the LR and RFR

4

Leveraging patient, imaging, and surgical factors to enhance operation time prediction in laparoscopic cholecystectomy

Leveraging patient, imaging, and surgical factors to enhance operation time prediction in laparoscopic cholecystectomy

V.P.S. Oosterhoff, BSc.[1], S.C. Baltus, MSc.[1,2], Dr. C.O. Tan.[1], Prof. Dr. I.A.M.J. Broeders[2]

Abstract

Background: Accurately estimating surgery duration can improve the cost-effective use of surgical staff and operating rooms while also reducing patient wait times. In a previous retrospective study, we attempted to predict operation time using the features of age, body mass index (BMI), sex, American Standards Association (ASA) score, and surgical level of expertise. The resulting models were not clinically applicable. To further increase model accuracy, this study incorporates a broader subset of patient factors and added imaging factors to improve the model further.

Methods: Patient data from 199 elective cholecystectomy procedures performed between 2021 and 2024 was extracted from the electronic health record (EHR). From this data, patient, imaging, and surgical factors were obtained. The variance inflation factor (VIF) and forward/backward feature selection were conducted to reduce the number of used features for operation time estimation. Lastly, a linear regressor (LR) and a random forest regressor (RFR) were used to predict operating time. The predictions of both models were compared with the conventional planning method.

Results: The inflammation by indication, thickened gallbladder wall, and surgical level of expertise were significantly correlated to the operating time. The combined features of inflammation by indication with a thickened gallbladder wall and thickened gallbladder wall with the operator's level of expertise were also significant. The best model achieved a R^2-score of -0.083 ± 0.10 and Root mean squared error (RMSE) of 12.28 ± 1.20. This produced similar results to the planned operation time, which had a R^2-score of -0.20 ± 0.17 and RMSE of 12.88 ± 0.05.

Conclusion: The generated models perform similarly to the current planning method. To enhance their suitability for clinical application, further refinement is necessary. This could be accomplished by expanding the dataset and integrating intraoperative data, which may provide additional predictive accuracy and robustness in real-world settings.

Keywords

Operation time — Cholecystectomy — Linear regression — Random forest regression

[1] *Technical medicine, University Twente, Enschede, The Netherlands*
[2] *Department of Surgery, Meander Medical Center, Amersfoort, Netherlands*

1. Introduction

In our previous retrospective analysis, we evaluated the accuracy of predicting the required operation time for laparoscopic cholecystectomy (LC) based on patient characteristics such as age, sex, body mass index (BMI), American Society of Anesthesiologists (ASA) score, and the surgeon's level of expertise. The study demonstrated an average planning error of 16.22 ± 0.58 minutes, and the model achieved a root mean squared error (RMSE) of 14.18 (± 0.58) minutes. Despite these results showing a slight improvement over manual planning, the model's performance remained insufficient for clinical application, as indicated by an R^2 score of 0.040. This low R^2 value suggests that the model explained only a tiny fraction of the variance in operation time, likely due to the limited set of parameters considered in the initial analysis [1].

In addition to the factors examined in our previous study, the literature identifies several other patient-related variables associated with prolonged operation times. These include smoking, diabetes mellitus (DM), prior abdominal surgery, indication for surgery, anticoagulant, and Endoscopic Retrograde Cholangiopancreatography (ERCP) [2, 3, 4, 5, 6, 7, 8]. Furthermore, imaging factors may also enhance the prediction of operating time in LC [9]. Common imaging factors include increased gallbladder wall thickness, the presence of stones in the gallbladder neck, and dilatation of the common bile duct (CBD) [3, 10, 11].

To improve our model's predictive capability, we are conducting a prospective study that incorporates additional factors identified in the literature. By integrating these variables, we aim to enhance the model's accuracy and develop a tool that can be effectively implemented in clinical practice. This prospective analysis will build upon the findings of our previous work and provide a more comprehensive understanding of the factors influencing surgical duration.

2. Methods

This prospective study aimed to predict the operation time required for laparoscopic cholecystectomy (LC). We utilised three distinct subsets of factors: patient factors, imaging factors, and surgical factors. Collectively, these factors are referred to as the features used to train the model. The workflow, as depicted in Figure 1, consists of five key steps:

1. **Data collection & cleaning:** Preoperative data from patients undergoing an elective LC in the Meander Medical Center (MMC) is collected. Outliers were removed based on the interquartile range (IQR). Moreover, patients with incomplete data were also removed.
2. **Univariate analysis:** For all variables, the correlation with the outcome variable is determined without considering other variables' influence.
3. **Feature engineering:** Polynomial features were engineered based on clinical expertise to ensure non-linear data relations are accounted for [12].
4. **Feature selection:** With the variance inflation factor (VIF), multicollinear features were removed, and using forward and backward feature selection, the final feature set was selected.
5. **Model development:** Two machine learning (ML) models were used to make operation time estimation: a linear regression model (LR) and a random forest regression (RFR). The stability and reproducibility of the proposed method is tested using K-fold cross-validation. Finally, a comparison is made between the different models' predictions and the planners' planned time.

2.1 Data

2.1.1 Data source

We collected patient data from 231 patients undergoing an elective LC in the MMC, Amersfoort, The Netherlands. All procedures took place between 1 January 2021 and 1 August 2024. The study was approved by the local Institutional Review Board of the MMC (Protocol No: TWO 21–007. Patients with non-elective procedures, those who refrained from treatment, individuals with missing medical images, patients under 18, and those undergoing concurrent procedures were excluded.

2.1.2 Outcome variable

The outcome variable in this study was the actual operation time, which was defined as the time between the first incision and skin closing. The anaesthesia staff recorded the actual operation time during surgery.

2.1.3 Independent variables

Nineteen independent patients, medical imaging, and surgical factors were selected in total. Eighteen factors were directly extracted from the electronic health record (EHR). The factor time between indication and operation >6 weeks after diagnosis was computed based on the operation date and the date of indication. This factor was included because the time of surgery can influence the intraoperative findings.

2.1.4 Patient factors

Selected patient factors were BMI, age, ASA score, sex, smoking, hypertension, indication, previous jaundice, prior abdominal surgery, diabetes, and anticoagulant use. These factors were extracted from the EHR. BMI and age were continuous. Sex was categorical, and the other factors were binary.

2.1.5 Medical imaging factors

Imaging was performed for each patient before surgery using abdominal ultrasound (US), computed tomography (CT), or magnetic resonance cholangiopancreatography (MRCP). The extracted factors from the medical images were gallbladder volume, thickening of the common bile duct, thickening of the gallbladder wall, presence of stones in the gallbladder neck, and presence of stones in the gallbladder, as well as whether Endoscopic Retrograde Cholangiopancreatography (ERCP) was performed before surgery. Gallbladder volume was categorised as shrunken, hydropic, or normal, while the other factors were binary variables.

2.1.6 Surgical factors

The variables of the surgeon's level of expertise and time between indication and operation >6 weeks after diagnosis were included. Both were binary variables.

2.2 Univariant analysis

The correlation with the outcome variable is determined using Pearson correlation for continuous variables and point-biseral correlation for binary variables. A correlation was deemed significant if $P < 0.05$. A correlation was considered very strong if higher than 0.7 and weak if lower than 0.2 [13].

2.3 Feature engineering

New features were created based on clinical knowledge and the variables that had a significant correlation with the dependent variable. For example, in the case of gallbladder inflammation as an indication, a longer operative time is expected based on clinical experience. Typically, six weeks is considered necessary to control the inflammation. Therefore, if these six weeks have passed, the surgery is expected to be less difficult and thus shorter than if the surgery took place within the six weeks. The same applies when a stone in the gallbladder neck was not removed using ERCP. This can make it more difficult for the surgeon to tension the gallbladder, negatively impacting the operative time.

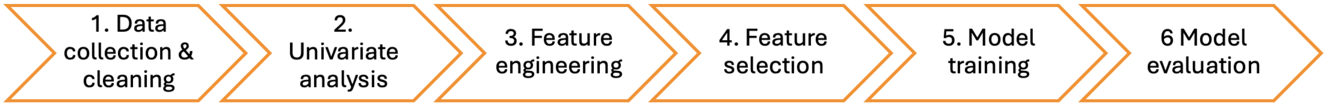


Figure 1. Used workflow for training ML models for operation time estimation

2.4 Feature Selection

After feature engineering, multicollinearity was checked using the Variance Inflation Factor (VIF). Any feature with a VIF over ten was removed, starting with the one with the highest VIF if multiple exceeds this threshold. If both a polynomial feature and its base feature have VIFs over ten, the polynomial was removed.

Next, the best set of features was chosen through forward-backward selection, performed with 5-fold cross-validation to ensure stable feature selection. Features appearing in at least three of the five folds in either forward selection or backward selection were used for the final model.

For any features unique to either set, VIF was recalculated. If there was no more collinearity, those features were added to the model. If collinearity was still present, the feature with the highest VIF was removed. This process was repeated until the collinearity issue was resolved.

2.5 Model development

Lastly, the models were trained using the selected features. Therefore, an LR model and an RFR model were selected. The LR model was chosen to see how well the data can be fitted assuming linearity [14]. In the case of a high-performing linear model, there was no need to add extra layers of complexity to predict operation time. The RFR was used to indicate the usability of a model accounting for non-linear relationships [15]. If the RFR outperforms the LR model, which is expected, it could be favourable to fit more complex models. Moreover, both LR and RFR models were used in other studies aiming to predict operation time [16, 17].

To optimise the performance of the RFR, hyperparameter tuning is applied using the hyperparameter grid shown in Table 1. No additional hyperparameters were selected for the LR model due to the limited effect of hyperparameter tuning for LR [18]. Lastly, the models were fitted with the previously selected features. A 5-fold cross-validation was conducted to ensure the final model's stability.

To assess the effectiveness of the feature selection process, the models were also trained using manually selected features, selected based on significant correlation. Additionally, the models were evaluated using the complete set of features to determine the impact of feature selection on overall model performance.

3. Results

Table 1. Grid used for RFR hyperparameter tuning.

Hyperparameter	Values
Max Depth	3, 5, 7
Minimal samples per leaf	1, 2, 4
Minimal samples per split	2, 5, 10
Number of estimators	100, 200, 300

3.1 Data cleaning

32 patients were excluded due to missing medical imaging and due to patients refraining from treatment. Three data points were identified as outliers based on the IQR, leaving a dataset of 199 patients used for model development. The patient characteristics are shown in Table 2.

Figure 2 displays the distribution of the actual and planned operation times. The actual operation time was not normally distributed, with a median of 46.61 (16.25 [38.0:54.25]). The planned operation time is also not normally distributed with a median of 45 (0.0[45.0: 45.0]). The error between actual operation time and planned operation time is shown in Appendix A.1

3.2 Univariant analysis

Based on the univariant analysis, the patient factors inflammation by indication, the imaging factor thickened gallbladder wall, and the surgical factor level of expertise showed significant correlations with the outcome variable. These significant correlations are highlighted in Table 2.

3.3 Feature engineering

The constructed features inflammation by indication and stone in gallbladder neck, inflammation by indication with gallbladder wall thickened, and gallbladder wall thickened with operator level of expertise were significant (see Appendix A Table A.1). In most of the engineered features, the distribution of the classes was imbalanced (see Appendix Figure B.1).

3.4 Feature selection

Given that the variables, a stone in the gallbladder neck in combination with previous ERCP performed, ASA 1, and a shrunken gallbladder exhibited a VIF greater than 10, they were excluded from the analysis. Subsequently, forward and backward feature selection was conducted using 5-fold cross-validation. The outcomes of these selections are shown in Appendix D.1 and D.2, respectively. Figure 3 depicts the distribution of the selected features across the five folds. For the LR model, the features chosen for the final model were:

- Anticoagulation

Table 2. Patient characteristics and correlation with the outcome variable

Characteristics	**N(%)**[1]	**Correlation**	**P value**
	PATIENT FACTORS		
Age	53 (22 [41; 63])	0.046	0.54
BMI	27.05 (5.9 [24.6; 30.5])	0.035	0.63
Sex		-0.070	0.33
Males	61 (31.12%)		
Females	135 (68.88%)		
Previous surgery		0.0039	0.96
Yes	68 (34.69%)		
No	128 (65.31%)		
Smoking		0.071	0.32
Yes	26 (13.27%)		
No	170 (86.73%)		
Hypertension		0.020	0.79
Yes	45 (22.96%)		
No	151 (77.04%)		
Previous Jaundice		0.010	0.17
Yes	16 (8.16%)		
No	180 (91.84%)		
Diabetes		0.12	0.093
Yes	15 (7.65%)		
No	181 (92.35%)		
Anticoagulant		0.057	0.43
Yes	18 (9.18%)		
No	178 (90.82%)		
Inflammation by indication[2]		**0.20**	**0.0044**
Yes	9 (4.57%)		
No	187 (95.41%)		
ASA score 1	43 (21.94%)	-0.030	0.70
ASA score 2	119 (60.71%)	-0.053	0.46
ASA score 3	30 (15.31%)	0.096	0.18
ASA score 4	4 (2.04%)	0.020	0.79
	IMAGING FACTORS		
Shrunken gallbladder	15 (7.65%)	-0.11	0.13
Normal gallbladder	170 (86.73%)	0.060	0.41
Hydropic gallbladder	11 (5.61%)	0.037	0.60
Stones		0.025	0.73
Yes	184 (93.88%)		
No	12 (6.12%)		
ERCP Performed		0.056	0.44
Yes	22 (11.22%)		
No	174 (88.87%)		
Stone in gallbladder neck		0.12	0.085
Yes	19 (9.69%)		
No	177 (90.31%)		
Dilated CBD		0.10	0.16
Yes	28 (14.29%)		
No	168 (85.71%)		
Thickened gallbladder wall[2]		**0.16**	**0.029**
Yes	33 (16.84%)		
No	163 (83.16%)		
	SURGICAL FACTORS		
Level of expertise[2]		**0.16**	**0.014**
Surgeon	79 (40.31%)		
Resident	117 (59.69%)		
Operation >6 Weeks After Diagnosis		-0.062	0.39
Yes	158 (80.61%)		
No	38 (19.39%)		

[1] Median is denoted as Median (IQR [Q1:Q3]).

[2] The emphasised correlations are statistically significant.

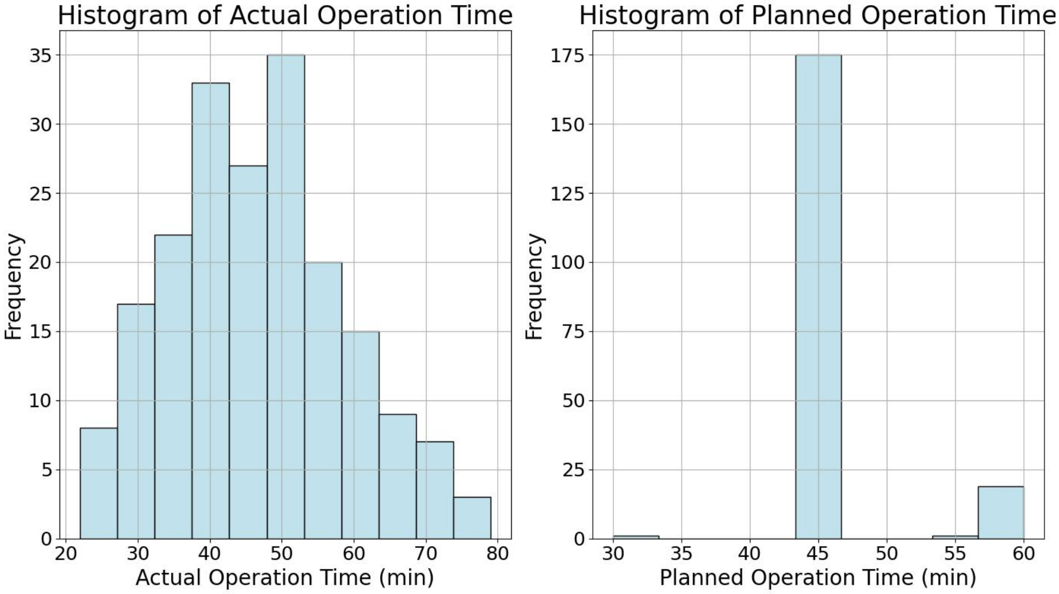


Figure 2. Distribution of the actual and planned operation times.

- Smoking
- Gallbladder Wall Thickened
- Level of Expertise
- Hypertension
- BMI
- CBD dilation
- Inflammation by Indication
- Normal gallbladder
- tone in Gallbladder Neck
- Operation 6 Weeks After Diagnosis

For the RFR model, the features used to train the final model were:

- Age
- Sex
- Diabetes
- BMI
- Previous Abdominal Surgery

3.5 Model evaluation

The optimal hyperparameters for the RFR include a maximum depth of 10, a minimum of 2 samples per leaf, 10 samples per split, and 100 estimators. The results of the 5-fold cross-validation are presented in Table 3. The model with the lowest R^2-score is the RFR model with features selected using forward and backward feature selection.

4. Discussion

This study aimed to evaluate the efficacy of preoperative factors in predicting the duration of elective laparoscopic cholecystectomy (LC) procedures. Our analysis used both an LR and RFR model to predict operation times. The linear regression model demonstrated a mean R^2-score of -0.17 ±0.14 minutes across 5-fold cross-validation. The RFR model yielded a mean R^2-score of -0.0853 ±0.091. When comparing the predicted operation times from our model to the scheduled times estimated by planners, our model showed similar accuracy in predicting actual operation duration. Interestingly, our model did not outperform the models in the retrospective study. This could be due to several reasons.

Table 3. Results of the 5-fold cross-validation

Model		R^2-score	RMSE
LR model			
	Selected features	-0.17 ±0.21	12.72 ±1.75
	Manual Selected features	-0.17 ±0.27	12.73 ±2.044
	All features	-0.63 ±0.14	15.11 ±1.69
RFR			
	Selected features	-0.083 ±0.10	12.28 ±1.20
	Manual Selected features	-0.087 ±0.16	12.12 ±1.65
	All features	-0.13 ±0.18	12.56 ±1.58
Planned		-0.20 ±0.17	12.88 ±0.05

Firstly, the dependent variable, operation time, may be subject to measurement errors [19]. The anesthesiology staff records These times manually, and factors such as operational stress, distractions, and individual variations in work habits can lead to inconsistencies. Given the small error margins in our study, even a minor deviation could significantly affect model performance. Adopting objective time-tracking systems or surgical phase detection mechanisms would be necessary to mitigate this issue. This could be achieved by deriving the absolute operation time from the laparoscopic videos.

Secondly, the univariate analysis revealed fewer statistically significant predictors of operation time than anticipated. Literature commonly identifies patient factors such as BMI, age, gender, and ASA score as significant predictors of op-

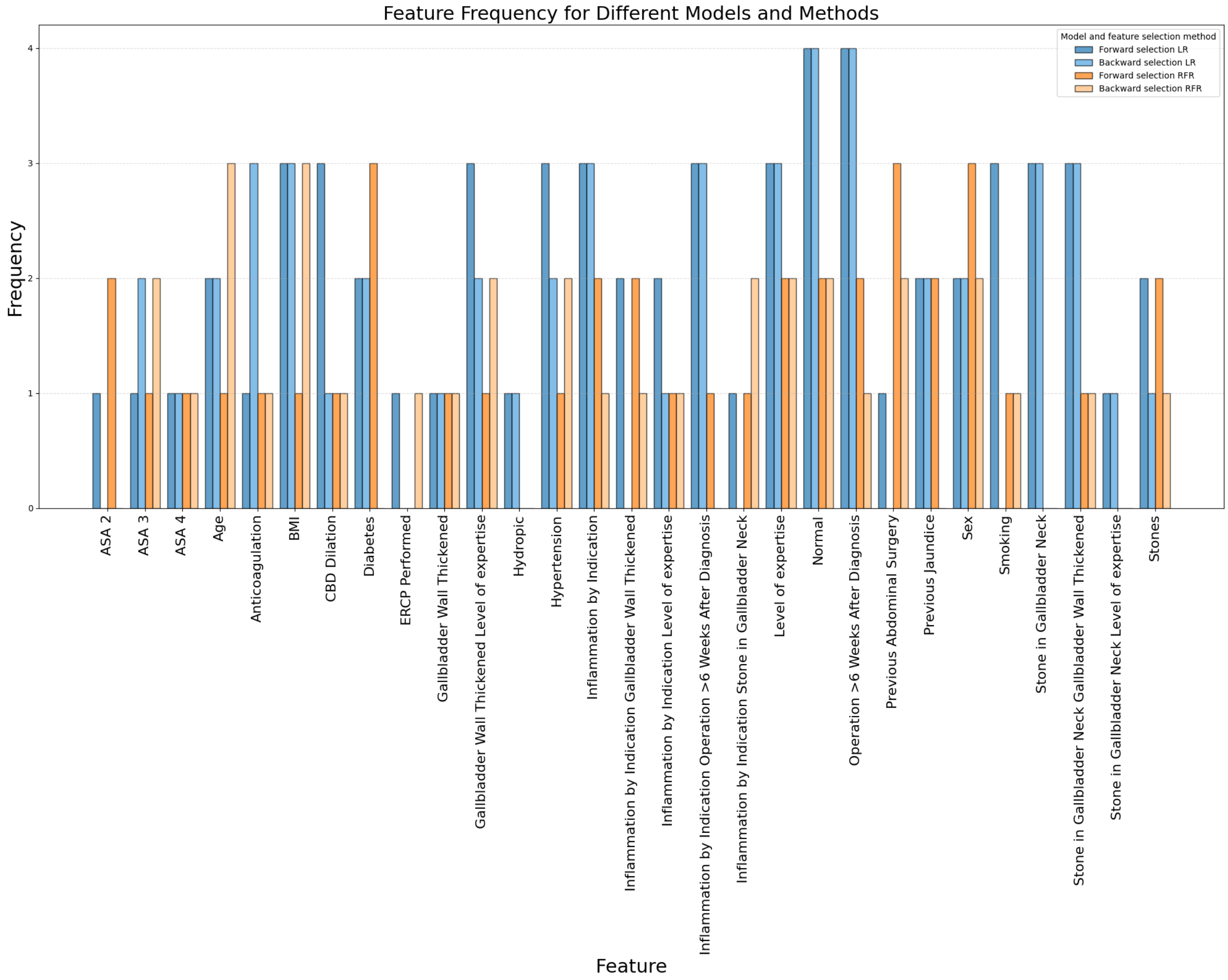


Figure 3. Results of the cross-validated feature selection process for forward and backward feature selection.

eration time [20, 21, 22, 23, 24]. However, our analysis did not find these variables to be significant predictors. Potential reasons for this discrepancy include the relatively small size of our dataset compared to other studies, which typically involve patient populations ranging from 323 to 7,227 [8, 25]. This limitation and measurement error in the outcome variable may have influenced the observed correlations. Additionally, the current study found relatively low correlations, ranging from 0.14 to 0.20 [13].

Moreover, the selected imaging factors were also less correlated with the operation time than reported in other studies [3, 26, 27]. In our study, only the thickened gallbladder wall was shown to be correlated with the operation time. However, in a study by Siddiqui et al., stones in the gallbladder neck and a dilated common bile duct were significant in predicting operating time [27]. This difference may be due to the manual extraction of imaging factors from radiology reports in the current study, which introduces inter-observer variability. Different radiologists at MMC handle the reporting, and while some specify the exact gallbladder wall thickness, others use broader terms like ”thickened” or ”normal.” Since we only extracted binary values, and the criteria for these classifications were not always clearly defined, this could affect the results.

It is also important to note that factors correlated with the outcome were relatively imbalanced. The patient factor inflammation by indication was only present 9 times in the dataset but showed a strong correlation. The Same accounts for the imaging factor of thickened gallbladder wall, which was present in only 33 cases and all three significant constructed features. This underrepresentation of correlated factors could partially declare the high standard deviation and the low score of the regression models, which is visualised in Appendix D.1 and D.2 [28]. Expanding the dataset to include more underrepresented cases will enhance the model's performance.

The predictions could potentially be optimised by exploring alternative modelling approaches. This study focussed on two general machine learning algorithms, but numerous other models and their variations exist [29, 30]. A systematic review of predictive models across various surgical domains has highlighted the efficacy of neural networks in estimating operation time [17]. However, it is only favourable to investigate more complex models if a certain relation in the data exists [31].

Lastly, in both the retrospective and prospective studies, not all the variance can be explained by the preoperative variables. Additional information is necessary to determine the operation time accurately. One method to achieve this could be by incorporating intraoperative information based on laparoscopic videos. Due to the fact that the patient and imaging factors try to predict the intraoperative situation, this information would likely improve the model's performance. One way to quantify this is by using the difficulty grading [32, 33]. We hypothesise that adding the difficulty grade to the model could lead to a more accurate and dynamic operation time prediction.

In conclusion, the current study shows that the factors of thickened gallbladder wall, inflammation by indication, and surgical level of expertise significantly correlate to the operation time. The current models are not yet usable in clinical practice due to the lack of factors accurately describing the variance of the operation time and due to the relatively small dataset. We hypothesise that increasing the dataset and adding operational difficulty as an additional predictor could improve the model performance and lead to a more accurate operation time prediction.

References

[1] L. Plonsky and H. Ghanbar, "Multiple regression in l2 research: A methodological synthesis and guide to interpreting r2 values," *The Modern Language Journal*, vol. 102, no. 4, pp. 713–731, 2018.

[2] J. M. Lipman, J. A. Claridge, M. Haridas, M. D. Martin, D. C. Yao, K. L. Grimes, and M. A. Malangoni, "Preoperative findings predict conversion from laparoscopic to open cholecystectomy," *Surgery*, vol. 142, no. 4, pp. 556–565, 2007.

[3] A. H. Nassar, J. Hodson, H. J. Ng, R. S. Vohra, T. Katbeh, S. Zino, and E. A. Griffiths, "Predicting the difficult laparoscopic cholecystectomy: development and validation of a pre-operative risk score using an objective operative difficulty grading system," *Surgical endoscopy*, vol. 34, pp. 4549–4561, 2020.

[4] H. M. Kaafarani, T. S. Smith, L. Neumayer, D. H. Berger, R. G. DePalma, and K. M. Itani, "Trends, outcomes, and predictors of open and conversion to open cholecystectomy in veterans health administration hospitals," *The American Journal of Surgery*, vol. 200, no. 1, pp. 32–40, 2010.

[5] M. Sugrue, F. Coccolini, M. Bucholc, and A. Johnston, "Intra-operative gallbladder scoring predicts conversion of laparoscopic to open cholecystectomy: a wses prospective collaborative study," *World Journal of Emergency Surgery*, vol. 14, pp. 1–8, 2019.

[6] J. Lucocq and A. H. Nassar, "The effects of previous abdominal surgery and the utilisation of modified access techniques on the operative difficulty and outcomes of laparoscopic cholecystectomy and bile duct exploration," *Surgical Endoscopy*, pp. 1–12, 2024.

[7] B. Lowndes, C. A. Thiels, E. B. Habermann, J. Bingener, S. Hallbeck, and D. Yu, "Impact of patient factors on operative duration during laparoscopic cholecystectomy: evaluation from the national surgical quality improvement program database," *The American Journal of Surgery*, vol. 212, no. 2, pp. 289–296, 2016.

[8] R. Bharamgoudar, A. Sonsale, J. Hodson, and E. Griffiths, "The development and validation of a scoring tool to predict the operative duration of elective laparoscopic cholecystectomy," *Surgical endoscopy*, vol. 32, pp. 3149–3157, 2018.

[9] J. Goonawardena, R. Gunnarsson, and A. De Costa, "Predicting conversion from laparoscopic to open cholecystectomy presented as a probability nomogram based on preoperative patient risk factors," *The American Journal of Surgery*, vol. 210, no. 3, pp. 492–500, 2015.

[10] M. Ercan, E. B. Bostanci, Z. Teke, K. Karaman, T. Dalgic, M. Ulas, I. Ozer, Y. B. Ozogul, F. Atalay, and M. Akoglu, "Predictive factors for conversion to open surgery in patients undergoing elective laparoscopic cholecystectomy," *Journal of Laparoendoscopic & Advanced Surgical Techniques*, vol. 20, no. 5, pp. 427–434, 2010.

[11] C. Gholipour, M. B. A. Fakhree, R. A. Shalchi, and M. Abbasi, "Prediction of conversion of laparoscopic cholecystectomy to open surgery with artificial neural networks," *BMC surgery*, vol. 9, pp. 1–6, 2009.

[12] T. Shanableh and K. Assaleh, "Feature modeling using polynomial classifiers and stepwise regression," *Neurocomputing*, vol. 73, pp. 1752–1759, 06 2010.

[13] P. Schober, C. Boer, and L. A. Schwarte, "Correlation coefficients: appropriate use and interpretation," *Anesthesia & analgesia*, vol. 126, no. 5, pp. 1763–1768, 2018.

[14] Z. Zhou, C. Qiu, and Y. Zhang, "A comparative analysis of linear regression, neural networks and random forest regression for predicting air ozone employing soft sensor models," *Scientific Reports*, vol. 13, no. 1, p. 22420, 2023.

[15] G. Louppe, "Understanding random forests: From theory to practice," *arXiv preprint arXiv:1407.7502*, 2014.

[16] C. A. Thiels, D. Yu, A. M. Abdelrahman, E. B. Habermann, S. Hallbeck, K. S. Pasupathy, and J. Bingener, "The use of patient factors to improve the prediction of operative duration using laparoscopic cholecystectomy," *Surgical endoscopy*, vol. 31, pp. 333–340, 2017.

[17] C. Spence, O. A. Shah, A. Cebula, K. Tucker, D. Sochart, D. Kader, and V. Asopa, "Machine learning models to predict surgical case duration compared to current industry standards: scoping review," *BJS open*, vol. 7, no. 6, p. zrad113, 2023.

[18] M. Nevendra and P. Singh, "Empirical investigation of hyperparameter optimization for software defect count prediction," *Expert Systems with Applications*, vol. 191, p. 116217, 2022.

[19] A. C. Guédon, M. Paalvast, F. Meeuwsen, D. M. Tax, A. van Dijke, L. Wauben, M. van der Elst, J. Dankelman, and J. van den Dobbelsteen, "'it is time to prepare the next patient'real-time prediction of procedure duration in laparoscopic cholecystectomies," *Journal of Medical Systems*, vol. 40, pp. 1–6, 2016.

[20] A. Tongyoo, A. Liwattanakun, E. Sriussadaporn, P. Limpavitayaporn, and C. Mingmalairak, "The modification of a preoperative scoring system to predict difficult elective laparoscopic cholecystectomy," *Journal of Laparoendoscopic & Advanced Surgical Techniques*, vol. 33, no. 3, pp. 269–275, 2023.

[21] M. Eshghali, D. Kannan, N. Salmanzadeh-Meydani, and A. M. Esmaieeli Sikaroudi, "Machine learning based integrated scheduling and rescheduling for elective and emergency patients in the operating theatre," *Annals of Operations Research*, vol. 332, no. 1, pp. 989–1012, 2024.

[22] H. Kar and A. Atay, "Predictive factors and importance of critical view of safety in difficult elective laparoscopic cholecystectomy," *Journal of Experimental and Clinical Medicine*, vol. 39, no. 3, pp. 874–878, 2022.

[23] J. S. Randhawa and A. K. Pujahari, "Preoperative prediction of difficult lap chole: a scoring method," *Indian Journal of Surgery*, vol. 71, pp. 198–201, 2009.

[24] C. Teerawiwatchai, J. Polprative, K. Rattanachueskul, and R. Thomtong, "Pre-operative score development: Predicting difficulty in elective laparoscopic cholecystectomy," *Journal of Health Science and Medical Research*, vol. 42, no. 2, p. 2023994, 2024.

[25] M. A. K. M. Vivek, A. J. Augustine, and R. Rao, "A comprehensive predictive scoring method for difficult laparoscopic cholecystectomy," *Journal of minimal access surgery*, vol. 10, no. 2, pp. 62–67, 2014.

[26] L. Bouarfa, A. Schneider, H. Feussner, N. Navab, H. U. Lemke, P. P. Jonker, and J. Dankelman, "Prediction of intraoperative complexity from preoperative patient data for laparoscopic cholecystectomy," *Artificial intelligence in medicine*, vol. 52, no. 3, pp. 169–176, 2011.

[27] M. A. Siddiqui, S. A. A. Rizvi, S. Sartaj, I. Ahmad, and S. W. A. Rizvi, "A standardized ultrasound scoring system for preoperative prediction of difficult laparoscopic cholecystectomy," *Journal of medical ultrasound*, vol. 25, no. 4, pp. 227–231, 2017.

[28] I. Ramos-Pérez, Á. Arnaiz-González, J. J. Rodríguez, and C. García-Osorio, "When is resampling beneficial for feature selection with imbalanced wide data?," *Expert Systems with Applications*, vol. 188, p. 116015, 2022.

[29] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," *Journal of clinical epidemiology*, vol. 110, pp. 12–22, 2019.

[30] R. Núñez, I. Doña, and J. A. Cornejo-García, "Predictive models and applicability of artificial intelligence-based approaches in drug allergy," *Current Opinion in Allergy and Clinical Immunology*, pp. 10–1097, 2024.

[31] L. T. Rose and K. W. Fischer, "Garbage in, garbage out: Having useful data is everything," *Measurement: Interdisciplinary Research & Perspective*, vol. 9, no. 4, pp. 222–226, 2011.

[32] T. M. Ward, D. A. Hashimoto, Y. Ban, G. Rosman, and O. R. Meireles, "Artificial intelligence prediction of cholecystectomy operative course from automated identification of gallbladder inflammation," *Surgical Endoscopy*, vol. 36, no. 9, pp. 6832–6840, 2022.

[33] J. R. Abbing, F. J. Voskens, B. G. Gerats, R. M. Egging, F. Milletari, and I. A. Broeders, "Towards an ai-based assessment model of surgical difficulty during early phase laparoscopic cholecystectomy," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 11, no. 4, pp. 1299–1306, 2023.

Appendix

A. Error distribution

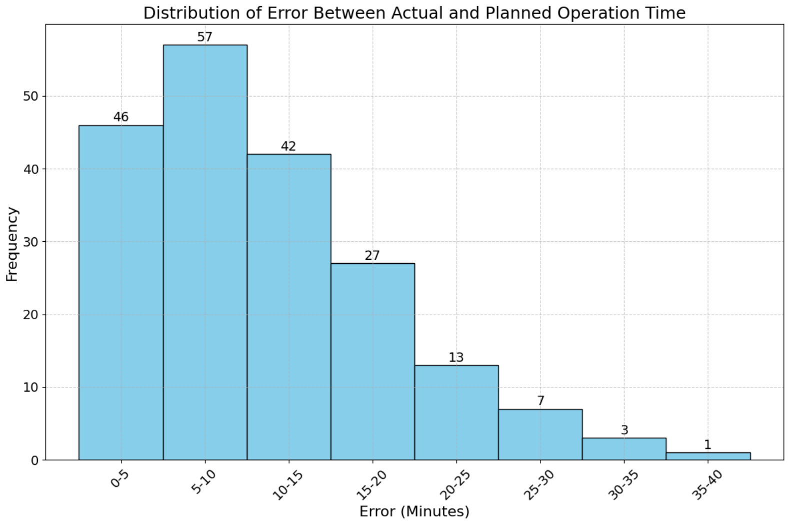


Figure A.1. Distribution of the error between the actual and planned operation times.

B. Violin plots of constructed features

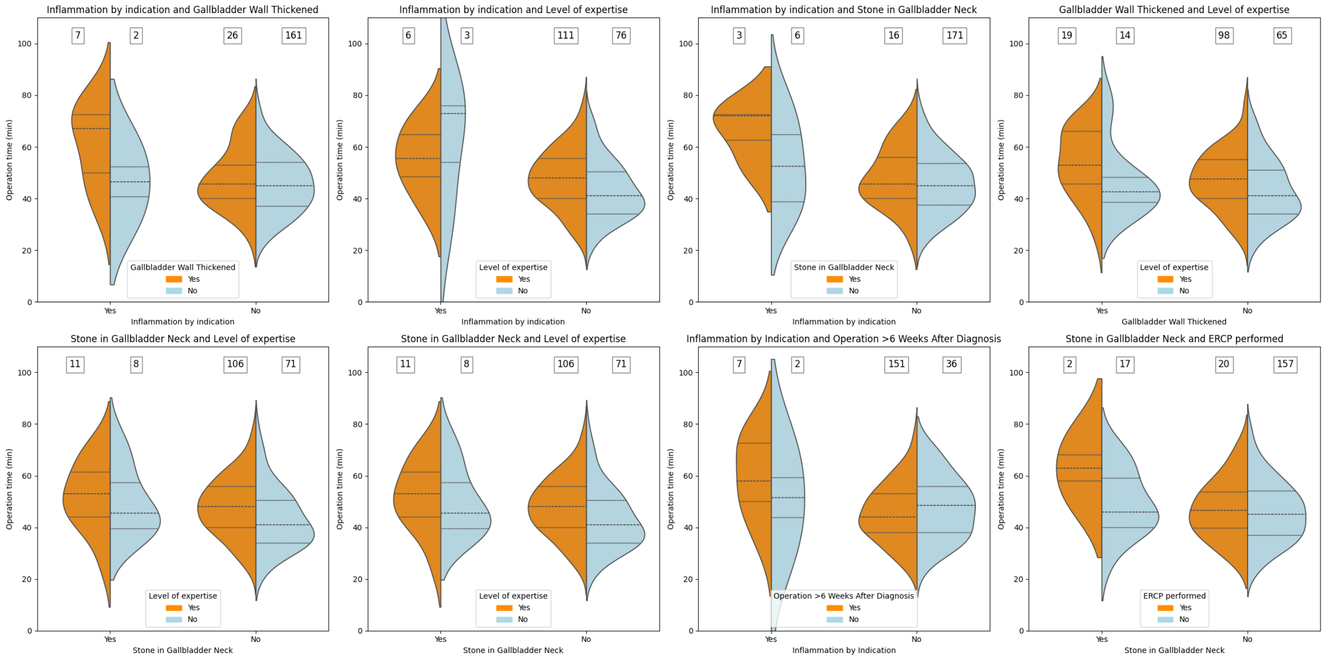


Figure B.1. The violin plots of the selected polynomial features display the distribution of the new features.

C. Correlation of the constructed features with the outcome variable

Table A.1. Results of feature engineering and correlation with the outcome variable.

Characteristics	N(%)	Correlation	P value
Inflammation by indication and stone in gallbladder neck		0.20	0.0049 [1]
Yes	3 (1.53%)		
No	193 (98.47%)		
Inflammation by indication and gallbladder wall thickened		0.23	0.0012 [1]
Yes	7 (3.56%)		
No	189 (96.43%)		
Inflammation by indication and operator level of expertise		0.13	0.068
Yes	6 (3.06%)		
No	190 (96.94%)		
Stone in gallbladder neck and gallbladder wall thickened		0.084	0.2404
Yes	10 (5.10%)		
No	186 (94.90%)		
Stone in gallbladder neck and operator level of expertise		0.11	0.12
Yes	11 (5.61%)		
No	185 (94.39%)		
Gallbladder wall thickened and operator level of expertise		0.20	0.0059 [1]
Yes	19 (9.69%)		
No	177 (90.31%)		
Inflammation by indication and operation >6 weeks after diagnosis		0.041	0.56
Yes	2 (1.02%)		
No	194 (98.98%)		
Stone in gallbladder Neck and ERCP performed		0.080	0.26
Yes	17 (8.67%)		
No	179 (91.33%)		

[1] This constructed feature showed a statistically significant correlation with the dependent variable.

D. Feature selection plots of forward and backward feature selection

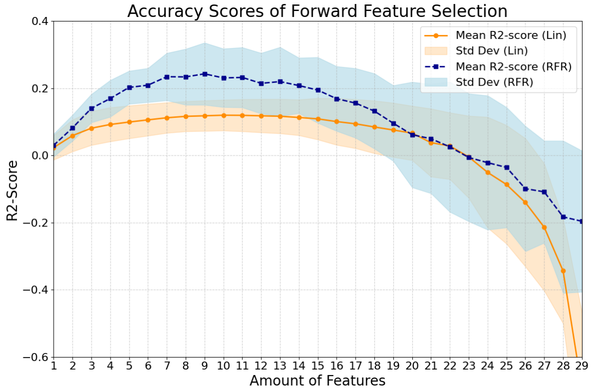


Figure D.1. Influence of amount of features on model performance after five fold cross-validation for forward feature selection

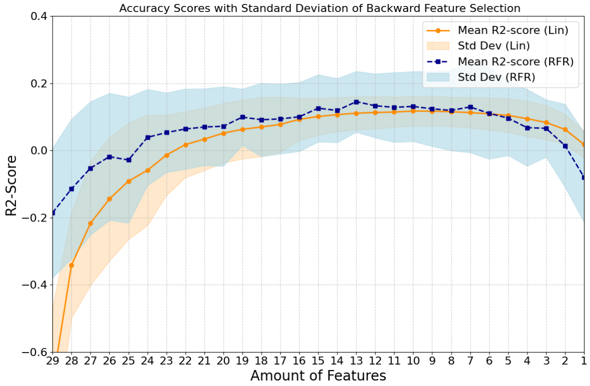


Figure D.2. Influence of amount of features on model performance after five fold cross-validation for backward feature selection

E. Bland-Altmann plot of the best prediction of the LR and RFR

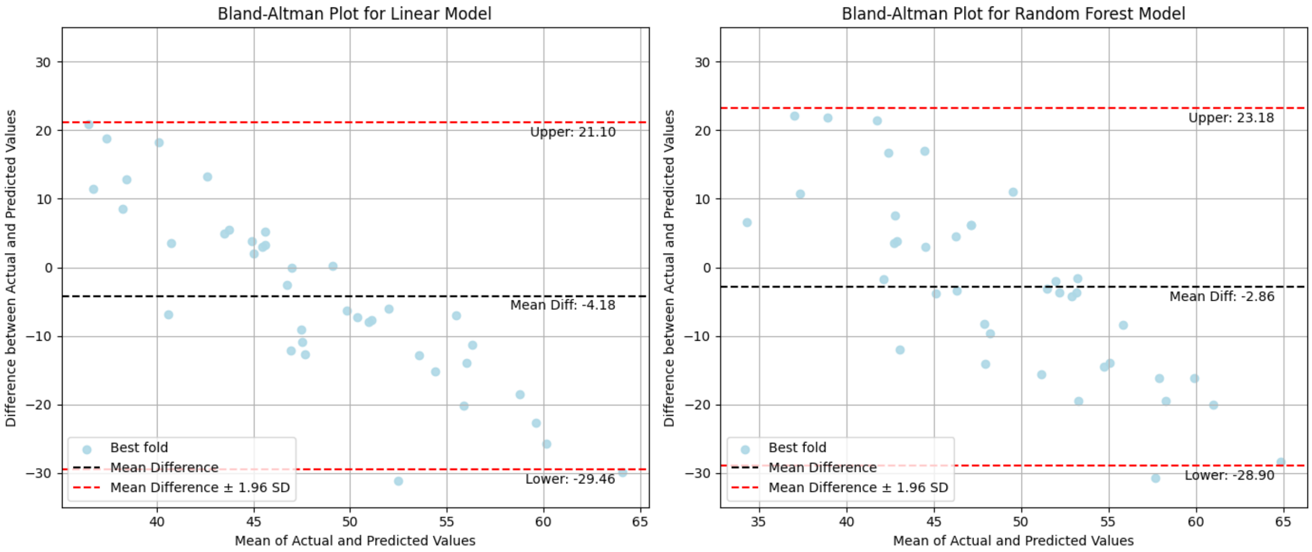


Figure E.1. The Bland-Altman plot for the best folds for the LR and RFR model. In these models, the features from the feature selection process are used.

5

Difficulty assessment based on laparoscopic cholecystectomy videos using a multiscale vision transformer

Difficulty assessment based on laparoscopic cholecystectomy videos using a multiscale vision transformer

V.P.S. Oosterhoff, BSc.[1] J.R. Abbing, MSc.[1,2] S.C. Baltus, MSc. [1,2] B.G.A. Gerats, MSc.[1,2] Dr. C.O. Tan.[1], Prof. Dr. I.A.M.J. Broeders[2]

Abstract

Background: Intraoperative difficulty assessment could be helpful in both operation time prediction and surgical benchmarking. However, objective difficulty assessment remains a challenge. Recent studies have shown the feasibility of deep-learning models using frame-wise classification. However, these methods fail to capture spatio-temporal context and therefore lack accuracy. In this study, we explore the possibility of using the Multi-scale Vision Transformer version 2 (MViTv2) to objectively determine the difficulty of laparoscopic cholecystectomy (LC) procedures.

Methods: To evaluate the effectiveness of the MViTv2 model, a surgical dataset consisting of 65 LC videos was utilized. A modified Nassar scale was created to classify the videos. The dataset was then divided into 10-second clips for model training. To assess the model's performance, accuracy, precision, recall, and F1 score were determined using a test set that was not used for training.

Results: The MViTv2 model successfully overfitted on a subset of our surgical dataset, indicating that usable features could be extracted. However, when trained on the complete dataset, the highest accuracy was 0.36%, indicating poor generalization to new data.

Conclusions: The MViTv2 model was not yet able to successfully determine the difficulty of an LC, as reflected by the best test accuracy of 36%. To improve the model, we recommend increasing the dataset size, employing various augmentation techniques, training on multiple GPUs to process more frames, and increasing the temporal step size of the training dataset.

Keywords
Difficulty — Cholecystectomy — Vision transformer — Nassar grade

[1] *Technical medicine, University Twente, Enschede, The Netherlands*
[2] *Department of Surgery, Meander Medical Center, Amersfoort, Netherlands*

1. Introduction

Laparoscopic surgery has not only transformed the field of cholecystectomy by improving post-operative outcomes for patients but also introduced a new dimension of data within the operating room [1, 2]. However, much of the valuable information embedded in this data remains inaccessible through traditional methods [3]. With recent advancements in AI, new models are emerging that can extract information from laparoscopic videos, offering enhanced decision-making tools [4, 5, 6]. One particularly valuable application is the objective assessment of surgical difficulty, where AI can quantify and predict challenges during surgery, enabling better preparation and response [7, 8].

Using difficulty assessment not only aids in real-time decision-making but also offers a significant advantage in surgical benchmarking [9, 10]. By objectively measuring intra-operative conditions during laparoscopic cholecystectomy (LC), surgeons gain a deeper understanding of the complexity of procedures. This objective assessment is crucial for predicting operation times and creating standardized benchmarks to evaluate surgical performance and outcomes [10]. A recent study demonstrated the feasibility of using computer vision models to determine surgical difficulty by assessing the Parkland difficulty grade [11]. While promising, this approach faced limitations. The first limitation was using the Parkland grading scale, which is not clinically validated [8]. A second limitation was that the model primarily focuses on gallbladder-specific features such as redness and vascularization and cannot identify adhesions correctly. Despite these challenges, integrating difficulty assessment into surgical benchmarking holds great potential for enhancing surgical training, improving patient outcomes, and standardizing procedural evaluation.

Building on this, previous research at the Meander Medical Center (MMC) explored a similar approach to difficulty assessment using a frame-wise method, where the model attempted to estimate gallbladder difficulty based on individual frames of the video [7]. In this study, the Nassar scale, a clinically validated measure of difficulty, was used [12]. They reported an accuracy of 88% in distinguishing between an easy and a difficult gallbladder. However, while the frame-based method could differentiate between inflamed and non-inflamed gallbladders, it struggled to capture the adhesion grade and failed to determine the total Nassar score reliably. We hypothesize that these shortcomings stem from the lack of spatiotemporal information, which is crucial for accurate difficulty assessment. In our institution, human annotators determine the Nassar scale by examining video clips and assessing the relationship between surrounding tissue and the gallbladder. This contextual information allows for better differentiation of tissue types, such as distinguishing adhesions, and provides a more comprehensive understanding of procedural difficulty.

Therefore, to improve the difficulty prediction further, a model must understand how different parts of the video relate to each other. A possible framework to achieve this is the multi-scale vision transformer version 2 (MViTv2) [13, 14]. This model, based on transformers, can understand contextual information within a video [15]. The model achieved a top-1 error of 86.1 % and a top-5 error of 97.0% on the Kinetics-400 dataset in terms of video classification, which outperforms temporal convolutional neural networks (CNNs) [13, 14, 16, 17]. Due to the high accuracy and the ability to understand video context, we expect that the MViTv2 model can better capture the relation between the gallbladder and surrounding tissue and achieve higher accuracy scores on the total grade than frame-based methods.

This research aims to determine MViTv2's strengths and limitations in surgical video analysis and improve difficulty assessments.

2. Methods

To evaluate the usability of the MViTv2 model in difficulty assessment, we first explain its working mechanism. Next, we define the datasets and labels used to analyze the model's behaviour further. Lastly, we elaborate on how the models are trained and discuss the evaluation metrics.

2.1 MViTv2 model structure

The building block of the MViTv2 model is the vision transformer. An overview of the model is visualized in Figure 1. First, a clip from the input video is randomly selected (a). The length of each clip is determined before training and is typically around 1 to 2 seconds. Additionally, frames are sampled within this clip. Using all frames in the selected clip yields better accuracy but has a bigger computational burden. Next, the frames are divided into several patches (b). Because the input for the transformer block must be a vector, these patches are fed into a ResNet50 [18] (c). Using ResNet50 to extract features, each patch is assigned a unique vector representing the original image's characteristics. Position embedding is added to ensure the model knows the original position of the patches [19]. After position embedding, the obtained vectors are fed into the transformer encoder (d). This transformer encoder is built up using multiple transformer blocks. The detailed structure of a transformer block is presented in Figure 2.

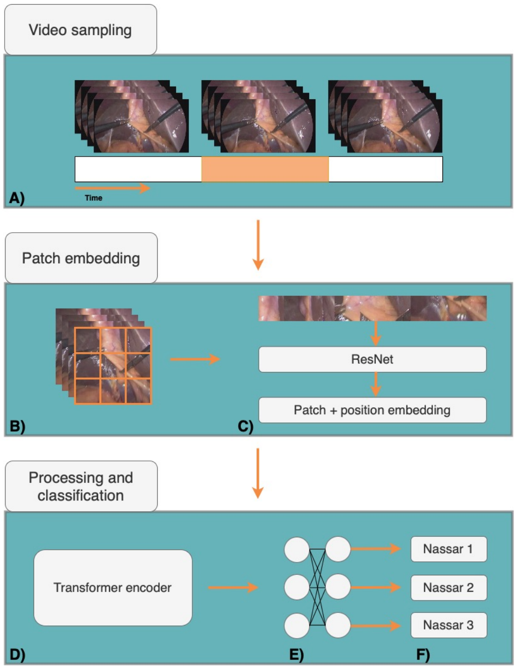


Figure 1. Overview of the MViTv2 work flow. A) a subset of frames is randomly sampled from the video. B) The frames are divided into several patches. C) The patches are fed into a ResNet to encode them as input for the transformer encoder. D) The embedded patches are fed into the transformer encoder structure of the MViT. E) A multi-head perceptron is added to classify the output of the Transformer encoder. F) The model outputs the probability of the label.

These encoder blocks are chained to each other in layers. Subsequently, the model is built in different layers. The first layers of the model primarily focus on gathering high spatial and temporal detail, and the later layers primarily focus on image features. In other words, the depth of the feature maps in the model expands over the layers while the processed number of frames reduces. In our model architecture, three scaling layers contain 3, 7, and 6 transformer blocks, respectively. The input for the next block is the output of the previous. To account for the change in dimension across layers, the MViTv2 uses residual pooling layers in the transformer blocks, which is visible in Figure 2. After all transformer blocks are passed, the output is fed into a multi-head perceptron. The last step is a fully connected softmax layer in which the found features are assigned to one of the pre-defined labels.

2.2 Acquiring labels

For acquiring the difficulty labels, the videos were loaded into label studio (v 0.9.0) [20]. We adopted the Nassar scale,

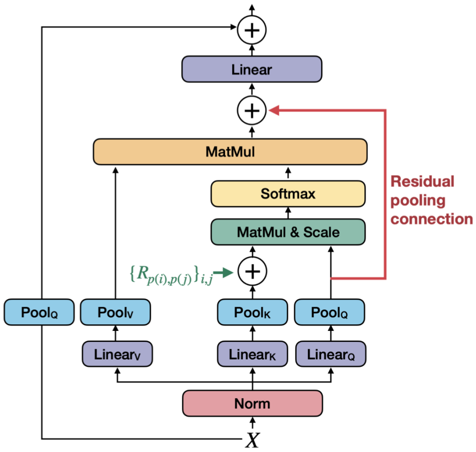


Figure 2. Visualisation of the MViTv2 model transformer block [14].

which has been validated for surgical outcomes [21, 12]. A label guide was constructed with a surgeon performing multiple LCs per week, shown in Appendix A. In this label guide, we attempted to make objective cut-off points suitable for a deep-learning model without impeding the clinical significance of the Nassar scale. After the construction of the label guide, the available videos were labelled by two annotators, one technical physician, and one technical medicine student. After labelling all videos separately, the cases with different grades were discussed until the annotators reached a consensus. Because we focus primarily on elective cases, the amount of Nassar grade 4 cases in the dataset was limited. Therefore, we decided to combine Nassar grades 3 and 4. In this study, we only used the final Nassar grade. This is determined by the category with the highest grade [12]. For example, if a video has gallbladder grade 2, adhesions grade 3, and cystic pedicle grade 2, the final Nassar grade is 3.

2.3 Dataset

The dataset consisted of laparoscopic videos from 70 patients undergoing LC surgery at Meander Medical Center between 2021 and 2024. The study was approved by the local Institutional Review Board (Protocol No: TWO 21–007). Three videos were excluded because the recording started after the start of the operation. Two videos were excluded because they were robot-assisted. Because the MViTv2 model randomly samples clips within the input video, training on the entire video was expected not to give the best results [14]. Therefore, clips were generated for each video. The clip started at the first moment the gallbladder came into view. If, in the video, the surgeon began to dissect the cystic duct, clipping was stopped. The minimum distance between the clips was 10 seconds. The length of the clips can be changed if necessary. For each video, a minimum of two and a maximum of seven clips were sampled. This is done to prevent under or over-sampling of videos.

Two separate datasets are made. **Dataset one** is primarily used to investigate model behaviour and determine how the model samples the input videos. This dataset contained six surgical videos, with each Nassar label represented two times. In this dataset, the train, validation, and test set contain the same videos. For this set, there are also clips available. However, for each of the videos, only one clip is made.

Dataset two contains clips of all available surgical videos. The video clips were made after data splitting to ensure there were no clips of the same video in the train, validation, or test set. In Figure 3, the final distribution of the labels is visualized. Due to the class imbalance in the dataset, class weighting was applied [22]. This was done using the Equation 1, where $n_{samples}$ denotes the total amount of labels present in the dataset, $n_{classes}$ the amount of different classes present in the dataset and n_{label_i} the amount samples of a single label [23].

$$\text{Class weights} = \frac{n_{samples}}{n_{clasess} \times n_{label_i}} \tag{1}$$

2.4 Trained models

Table 1 shows the different models we trained. First, we identified the ability of the model to overfit on surgical data. This was done using **dataset one**. One model was trained using the whole video as input. The other model was trained using only clips of 1 second as input. Both models' training and validation loss were logged as the primary output.

Next, we trained several models using **dataset two**. One model was trained with clips of 1 second, and one was trained with clips of 10 seconds. This was done to determine the influence of adding more data to the model. Both models were trained without data augmentation and using a model from scratch. To complete the analysis, two additional models were trained to identify the influence of a pre-trained model and data augmentation. To train these models, the clips of 10 seconds were used, because we hypothesized that the clips containing more data should yield more accurate results and are less prone to overfitting. One model was trained using the 10-second video clips using an MViTv2 model pre-trained on the kinetics400 dataset [16]. The other model was trained using 10-second video clips, a pre-trained model, and data augmentation. An overview of the different models and the corresponding metrics is shown in Table 2.

For models 3,4,5 and 6, the precision, accuracy, recall, and F1 score were calculated. The formulas for these metrics are shown in Equations (1)–(4). Here, the TP are the true positive classifications of the frames, TN the true negatives, FP false positives, and FN the false negatives. These were also used to generate the (normalized) confusion matrices [24]. The metrics were determined by performing inference on the first 60 seconds of the operation. This time frame was chosen, because it gives a good indication of model performance at the start of the operation. For inference, the models with the lowest validation loss during training were selected, which were manually extracted from the resulting validation plots.

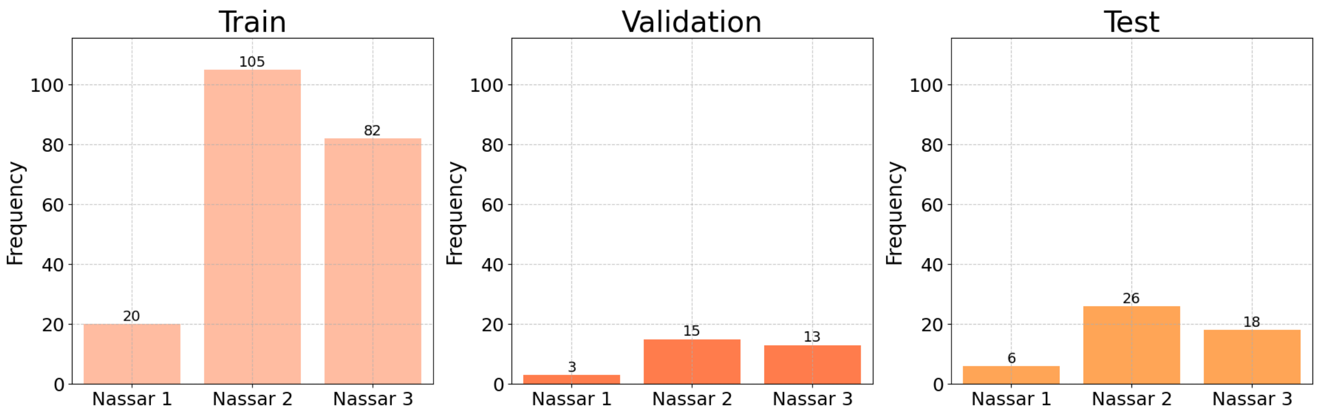


Figure 3. Distribution of the Nassar grade clips in dataset **dataset 2** as used for training.

Table 1. Overview of the trained model. Dataset 1 indicates the dataset with six videos and an equal train, validation, and test set. Dataset 2 contains all surgical videos and the corresponding clips

Model	Dataset	Clip length	Number of clips in train set	Pre-train	Augmentation
1	1	Whole video	6	No	No
2	1	1 second	6	No	No
3	2	1 second	207	No	No
4	2	10 seconds	207	No	No
5	2	10 seconds	207	Yes	No
6	2	10 seconds	207	Yes	Yes

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{4}$$

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

3. Results

In Figure 4, the results of the MViTv2 training on **dataset 1** and **dataset 2** are shown. It shows that the training and validation loss were lower when using only clips of 1 second for training. This result aligns with expectations, as the MViTv2 model randomly samples frames during training [25]. It is important to note that the validation loss for the training with clips is lower than the corresponding training loss. This happens because, during validation, the model is set to evaluation mode, where regularization techniques are not applied [26]. During training, regularisation prevents overfitting, which can affect the model's output. When the same dataset is used for both training and validation, the validation loss is expected to be lower than the training loss, as all model weights are used without regularization during validation.

Next, we trained the model on the complete dataset using both clips of 1 second and clips of 10 seconds. The class weights were computed as mentioned in the methods, resulting in the following weights: 3.45 for Nassar grade 1, 0.66 for Nassar grade 2, and 0.84 for Nassar grade 3. The metrics of the models are shown in Table 2. It is visible that the model using 1-second clips performs best on the test set. Moreover, none of the models gave Nassar grade 1 as an output. This is further emphasized by the confusion matrix of the 1-second and the 10-second clips without data augmentation in Figure 5.

4. Discussion

In this study, we developed a pipeline for labelling surgical videos and trained an MViTv2 model to assess the difficulty of LC procedures. We successfully trained the MViTv2 models on our surgical dataset. Our best model achieved an accuracy of 36%, significantly lower than the frame-wise difficulty predictions reported in previous research [7, 8]. In all trained models, we saw that the model tended to overfit. This resulted in a model primarily predicting Nassar grade 2.

Our study has several limitations. First, our model was trained on a relatively small dataset. The final set consisted of 256 surgical video clips, each lasting one second. When using pre-trained models, less data is needed to achieve appropriate conversion [27]. Despite the advantages of pre-trained models, our dataset would still be relatively small, especially for a ViT, which is known for its significant amount of tunable parameters [28]. To further enhance the model predictions, we suggest extending the dataset by selecting more clips within the videos where the gallbladder is visible.

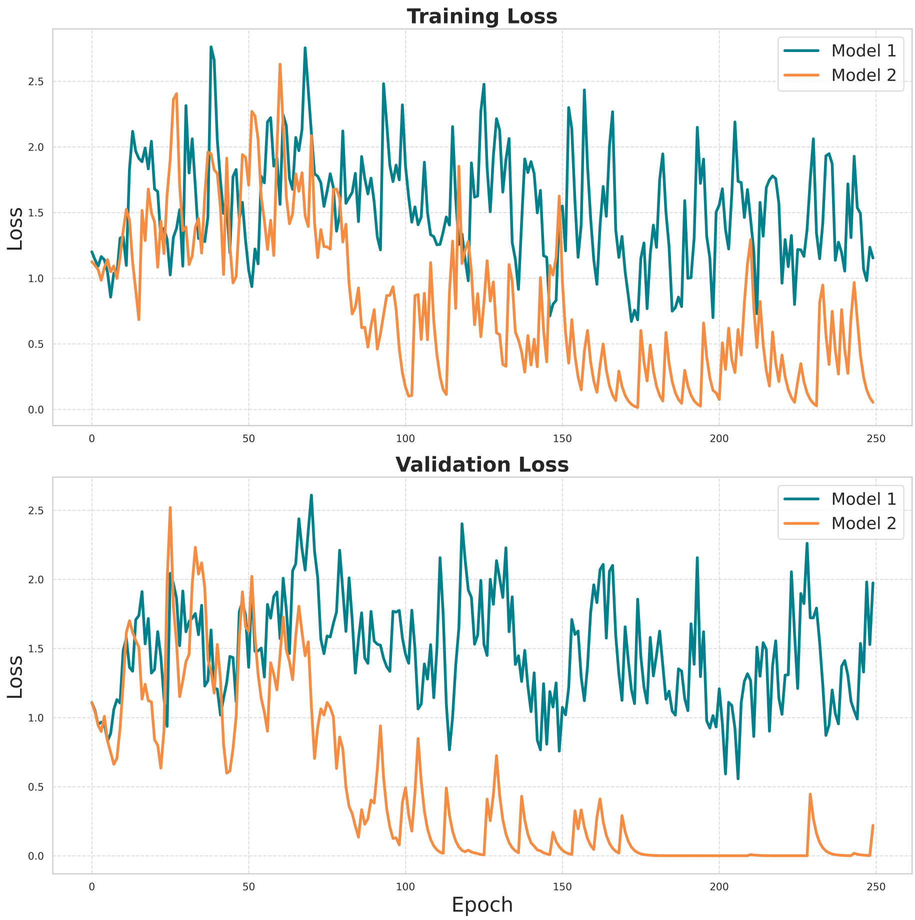
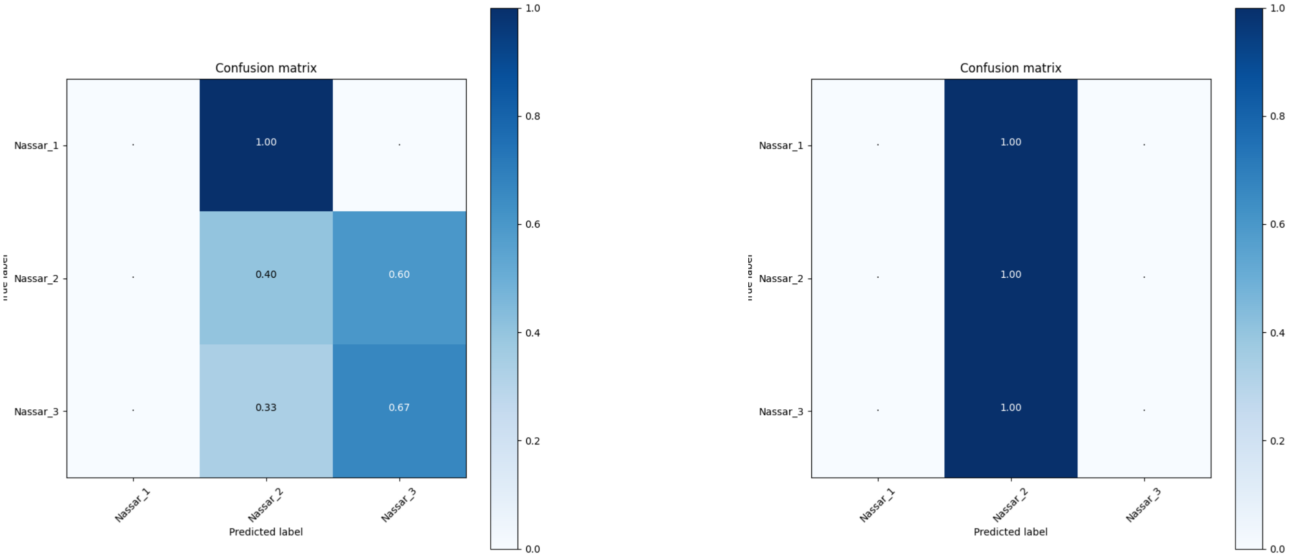


Figure 4. Training and validation loss training on video clips and the whole video using **dataset 1**.



(a) 1 second clips

(b) 10 seconds clips

Figure 5. Confusion matrices of the models using 1-second and 10-second clips for training. Both models are not pre-trained or use data augmentation. The inference is done on the first 60 seconds of the operation in the test set videos.

Table 2. Metrics of the models trained on **dataset 2**. On the left side of the table, the model characteristics are shown. The performance metrics are shown on the right side. The accuracy is determined for the entire model, and the other metrics are calculated for each label.

Model	Clip length	Pre-traiend	Data augmentation	Grade	Accuracy[1]	Precision	Recall	F1-score
3	1 second	No	No	1	0.36	0	0	0
				2		0.23	0.40	0.29
				3		0.53	0.67	0.59
4	10 seconds	No	No	1	0.33	0	0	0
				2		0.33	1	0.50
				3		0	0	0
5	10 seconds	Yes	No	1	0.24	0	0	0
				2		0.19	0.40	0.26
				3		0.35	0.33	0.34
6	10 seconds	Yes	Yes	1	0.33	0	0	0
				2		0.33	1	0.50
				3		0	0	0

[1] Accuracy is determined for the entire model.

Initially, we trained the model with minimal data augmentation to determine if it could extract features from the images. While this method can lead to overfitting, it allowed us to determine the model's behaviour. However, overfitting was still present after incorporating some basic augmentation and embedding a dropout layer. It is very important to fine-tune data augmentation further because it is known that vision transformers tend to overfit very fast on small datasets [29]. Moreover, recent studies report that incorporating augmentation and regularisation techniques can achieve the same results in improving the model as increasing the dataset [30, 31]. After incorporating more extensive data augmentation, we expect the model to be more generalizable over different datasets. Another method to reduce overfitting is the use of mixup. Mixup is a technique that combines synthetic images from the current batch of videos [32]. However, since mixup effectively increases the batch size by adding synthetic data, this leads to higher computational demands. As a result, training on multiple GPUs may be required to accommodate these increased costs.

Another issue we encountered was dataset imbalance, which we attempted to mitigate through class weighting. However, this approach did not fully compensate for the scarcity of Nassar grade 1 cases in the dataset [33]. We also tried binarizing the labels by combining grades 1 and 2, as suggested by Abbing et al., but this did not improve the model's performance. More Nassar grade 1 and 3 cases need to be included to balance the dataset, particularly by incorporating non-elective gallbladder cases. Since grade 1 gallbladders are more challenging to obtain, focusing on a subscore, such as the gallbladder or adhesion subscore, might provide a more balanced distribution.

Moreover, there are possibly some inconsistencies in the labelling process. Although we developed a labelling guide with input from a surgeon experienced in LC, the labelling was done without direct surgical oversight. Discrepancies between the two labellers were occasionally observed, particularly when distinguishing whether procedural difficulty stemmed from the intraoperative situation or the surgeon's skill. For instance, a straightforward procedure could appear more difficult if inadequate tension is applied, affecting visibility and leading to subjective labelling. Additionally, we modified the Nassar grading scale to better capture specific features in the operating room, which may affect its correlation with clinical outcomes.

Lastly, we used a set time window of 1 second for the model to select frames from. In this one second, 25 frames were extracted. If human reviewers label the gallbladder videos, we mostly jump into the video reviewing 2 or 3 seconds of video information. Therefore, training the model using longer video clips, including more frames, could be helpful. One drawback of this is that the computational costs increase when the number of frames loaded in the model increases.

A study by Kiyasseh et al. used another variant of a ViT to decode surgical videos for recognizing laparoscopic tool gestures, achieving accuracy rates of 0.85 in needle handling and 0.82 in needle driving [34]. Although their focus was on tool tracking rather than difficulty classification, their work underscores the utility of vision transformers in extracting meaningful features from surgical videos. Key differences between their study and ours include the volume of video data used and their application of a self-supervised, pre-trained ViT model. The self-supervised method used is DINO (knowledge distillation with no labels) and shows promising results with 530 - 912 clips per video. [35, 34]. They also trained the network on bigger sub-samples of the video, ranging between 10 - 30 seconds. Moreover, they made predictions of the entire surgical video, indicating the applicability of vision transformers for our tasks.

For future research, we recommend focusing on the model's performance in assessing adhesions. We hypothesize that the strength of vision transformers lies in comparing different video segments. While gallbladder inflammation, assessed by surface appearance, is well suited for ResNet models (as demonstrated by Abbing et al.), identifying adhesions is more challenging due to their visual similarity to healthy tissue. In such cases, contextual information becomes crucial. Focusing solely on adhesion grading may reveal whether the MViTv2 is better suited for this task. It is also advised to dive into the possibility of incorporating the DINO framework for self-supervised learning.

Moreover, we suggest verifying all labels with a clinical expert and performing laparoscopic cholecystectomies weekly. This should increase the label certainty and enhance the model's input data [36]. Additionally, the dataset should be extended to elective procedures and non-elective procedures, increasing the amount of Nassar grade 3 labels present in the dataset.

In conclusion, we developed a pipeline for manually labelling LC videos and training the MViTv2 model. Our research demonstrated the feasibility of training the MViTv2 model in-house and assessed its applicability to surgical data. While the model successfully extracted video features, it tended to overfit, and its generalizability to new data remains limited. To further explore the utility of MViTv2 for difficulty classification, future work should focus on expanding the dataset, applying appropriate data augmentation, consulting with clinical experts during labelling, and including non-elective gallbladder cases.

References

[1] Ahmad Guni, Piyush Varma, Joe Zhang, Matyas Fehervari, and Hutan Ashrafian. Artificial intelligence in surgery: the future is now. *European Surgical Research*, 65(1):22–39, 2024.

[2] Kangwei Guo, Haisu Tao, Yilin Zhu, Baihong Li, Chihua Fang, Yinling Qian, and Jian Yang. Current applications of artificial intelligence-based computer vision in laparoscopic surgery. *Laparoscopic, Endoscopic and Robotic Surgery*, 2023.

[3] David C Birkhoff, Anne Sophie HM van Dalen, and Marlies P Schijven. A review on the current applications of artificial intelligence in the operating room. *Surgical innovation*, 28(5):611–619, 2021.

[4] Constantinos Loukas, Ioannis Seimenis, Konstantina Prevezanou, and Dimitrios Schizas. Prediction of remaining surgery duration in laparoscopic videos based on visual saliency and the transformer network. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 20(2):e2632, 2024.

[5] Andru Putra Twinanda, Gaurav Yengera, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Rsdnet: Learning to predict remaining surgery duration from laparoscopic videos without manual annotations. *IEEE transactions on medical imaging*, 38(4):1069–1078, 2018.

[6] Ivan Aksamentov, Andru Putra Twinanda, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Deep neural networks predict remaining surgery duration from cholecystectomy videos. In *Medical Image Computing and Computer-Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part II 20*, pages 586–593. Springer, 2017.

[7] Julian R Abbing, Frank J Voskens, Beerend GA Gerats, Ruby M Egging, Fausto Milletari, and Ivo AMJ Broeders. Towards an ai-based assessment model of surgical difficulty during early phase laparoscopic cholecystectomy. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 11(4):1299–1306, 2023.

[8] Anup Shrestha, Abhishek Bhattarai, Kishor Kumar Tamrakar, Manoj Chand, Samjhana Yonjan Tamang, Sampada Adhikari, and Harish Chandra Neupane. Utility of the parkland grading scale to determine intraoperative challenges during laparoscopic cholecystectomy: a validation study on 206 patients at an academic medical center in nepal. *Patient Safety in Surgery*, 17(1):12, 2023.

[9] Ander Dorken-Gallastegi, Majed El Hechi, Maxime Amram, Leon Naar, Lydia R Maurer, Anthony Gebran, Jack Dunn, Ying Daisy Zhuo, Jordan Levine, Dimitris Bertsimas, et al. Use of artificial intelligence for nonlinear benchmarking of surgical care. *Surgery*, 174(6):1302–1308, 2023.

[10] RD Staiger, H Schwandt, Milo Alan Puhan, and PA Clavien. Improving surgical outcomes through benchmarking. *Journal of British Surgery*, 106(1):59–64, 2019.

[11] Thomas M Ward, Daniel A Hashimoto, Yutong Ban, Guy Rosman, and Ozanan R Meireles. Artificial intelligence prediction of cholecystectomy operative course from automated identification of gallbladder inflammation. *Surgical Endoscopy*, 36(9):6832–6840, 2022.

[12] Ewen A Griffiths, James Hodson, Ravi S Vohra, Paul Marriott, Tarek Katbeh, Samer Zino, Ahmad HM Nassar, and West Midlands Research Collaborative. Utilisation of an operative difficulty grading scale for laparoscopic cholecystectomy. *Surgical endoscopy*, 33:110–121, 2019.

[13] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6824–6835, 2021.

[14] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4804–4814, 2022.

[15] Satoshi Takahashi, Yusuke Sakaguchi, Nobuji Kouno, Ken Takasawa, Kenichi Ishizu, Yu Akagi, Rina Aoyama, Naoki Teraya, Amina Bolatkan, Norio Shinkai, et al. Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review. *Journal of Medical Systems*, 48(1):1–22, 2024.

[16] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The

kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[17] Subhash Nerella, Sabyasachi Bandyopadhyay, Jiaqing Zhang, Miguel Contreras, Scott Siegel, Aysegul Bumin, Brandon Silva, Jessica Sena, Benjamin Shickel, Azra Bihorac, et al. Transformers and large language models in healthcare: A review. *Artificial Intelligence in Medicine*, page 102900, 2024.

[18] Bum Jun Kim, Hyeyeon Choi, Hyeonah Jang, Dong Gu Lee, Wonseok Jeong, and Sang Woo Kim. Improved robustness of vision transformers via prelayernorm in patch embedding. *Pattern Recognition*, 141:109659, 2023.

[19] Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. *arXiv preprint arXiv:2403.13298*, 2024.

[20] Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. Label Studio: Data labeling software, 2020-2022. Open source software available from https://github.com/heartexlabs/label-studio.

[21] Masamichi Yokoe, Jiro Hata, Tadahiro Takada, Steven M Strasberg, Horacio J Asbun, Go Wakabayashi, Kazuto Kozaka, Itaru Endo, Daniel J Deziel, Fumihiko Miura, et al. Tokyo guidelines 2018: diagnostic criteria and severity grading of acute cholecystitis (with videos). *Journal of Hepato-biliary-pancreatic Sciences*, 25(1):41–54, 2018.

[22] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[23] Jay Gala and Pengtao Xie. Learning from mistakes based on class weighting with application to neural architecture search. *arXiv preprint arXiv:2112.00275*, 2021.

[24] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[25] Samuel Schmidgall, Ji Woong Kim, Jeffery Jopling, and Axel Krieger. General surgery vision transformer: A video pre-trained foundation model for general surgery. *arXiv preprint arXiv:2403.05949*, 2024.

[26] Jan Kukačka, Vladimir Golkov, and Daniel Cremers. Regularization for deep learning: A taxonomy. *arXiv preprint arXiv:1710.10686*, 2017.

[27] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021.

[28] Ibrahim Batuhan Akkaya, Senthilkumar S Kathiresan, Elahe Arani, and Bahram Zonooz. Enhancing performance of vision transformers on small datasets through local inductive bias incorporation. *Pattern Recognition*, 153:110510, 2024.

[29] Sachin Chhabra, Hemanth Venkateswara, and Baoxin Li. Patchswap: A regularization technique for vision transformers. In *BMVC*, page 996, 2022.

[30] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

[31] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[32] Hongyi Zhang. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[33] Batuhan Bakırarar and Atilla ELHAN. Class weighting technique to deal with imbalanced class problem in machine learning: Methodological research. *Turkiye Klinikleri Journal of Biostatistics*, 15:19–29, 01 2023.

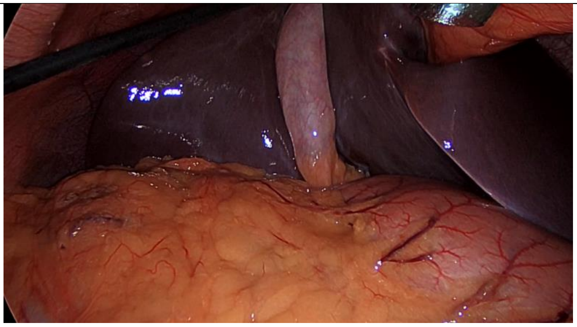
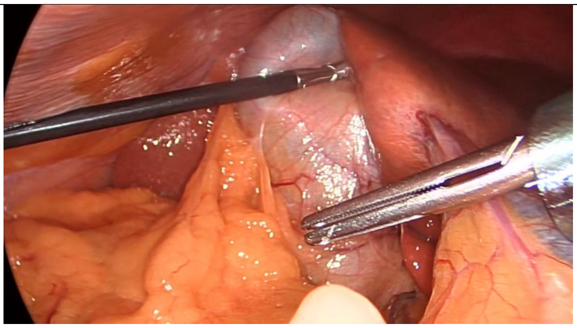
[34] D Kiyasseh, R Ma, TF Haque, BJ Miles, C Wagner, DA Donoho, A Anandkumar, and AJ Hung. A vision transformer for decoding surgeon activity from surgical videos. nat biomed eng, 2023.

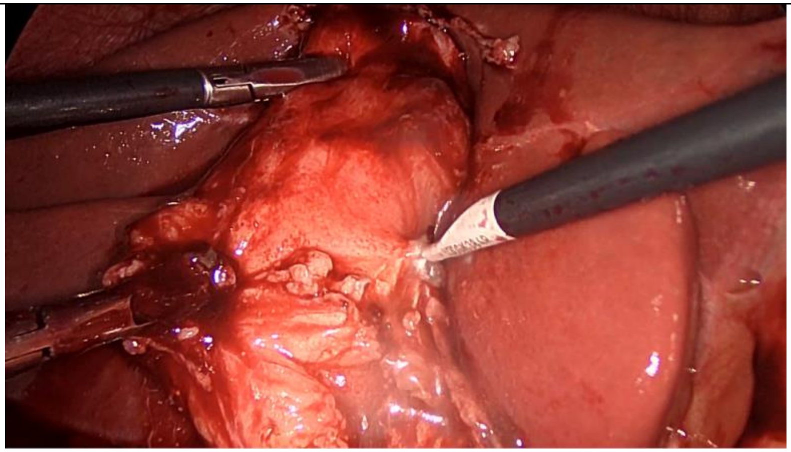
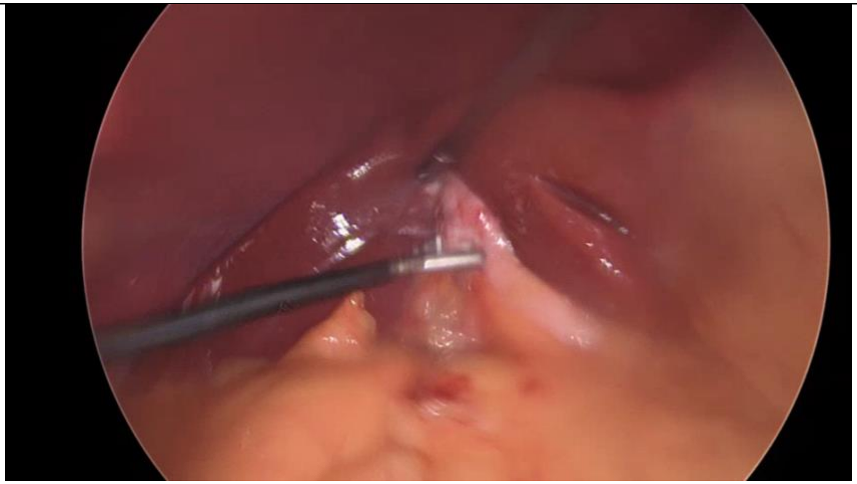
[35] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

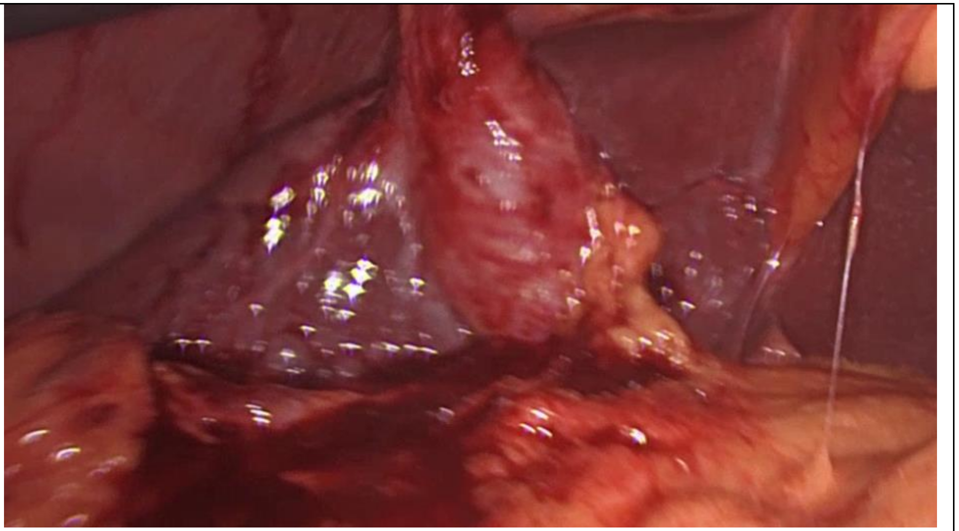
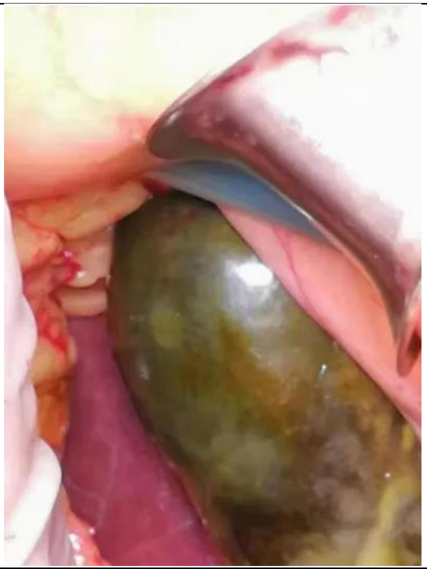
[36] Monique Kilkenny and Kerin Robinson. Data quality: “garbage in – garbage out”. *Health Information Management Journal*, 47:183335831877435, 05 2018.

Appendix

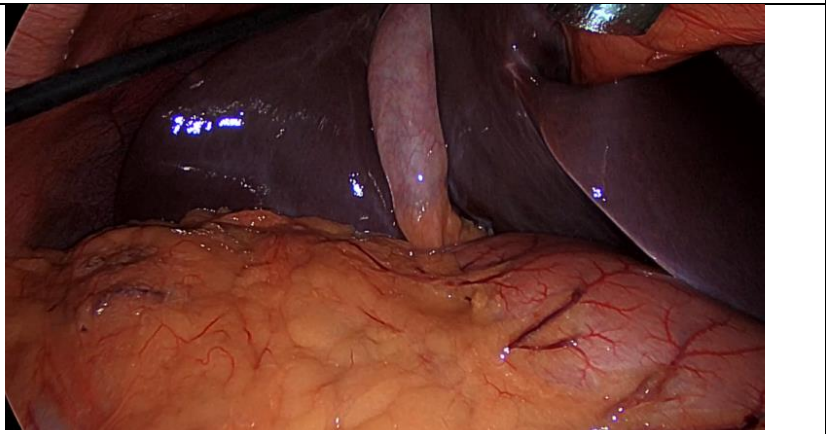
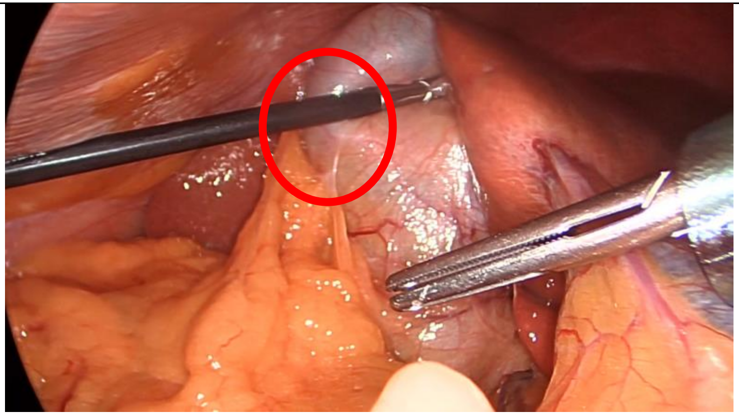
A. Label guide for labelling laparoscopic videos using an adjusted Nassar scale

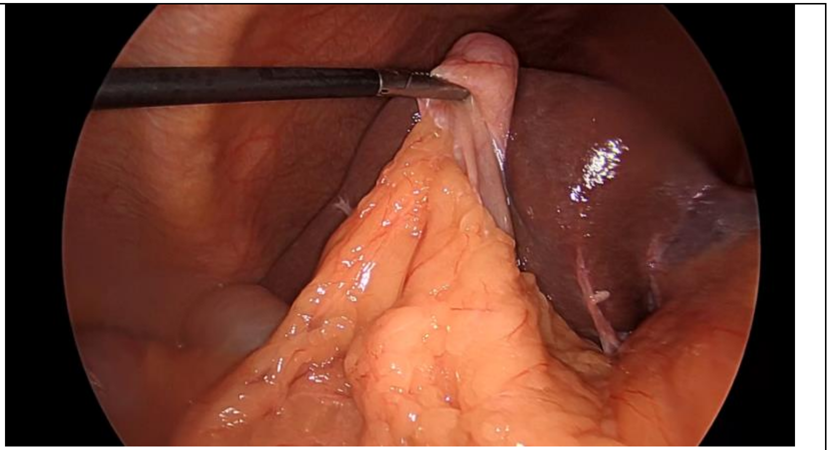
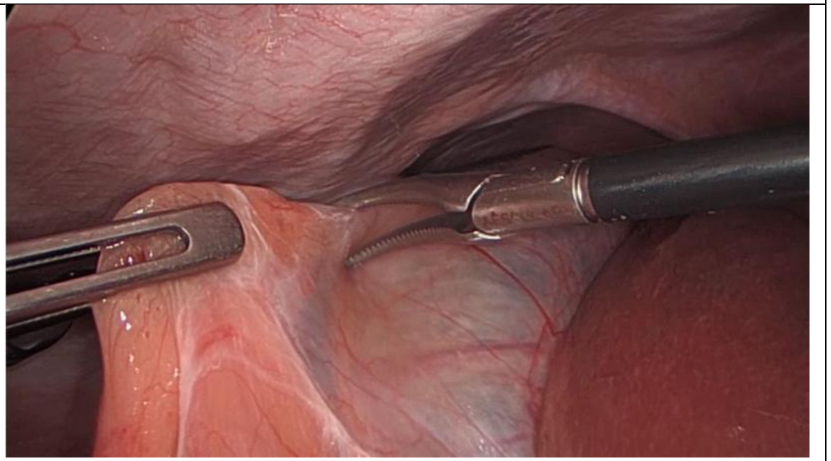
Galbladder grading				
Grade	**Criteria**	**Characteristics**	**Notes**	**Images**
1	Floppy, non-adherent	Gallbladder easy graspable and is easily moved out of the liver bed Slim galbladder, the pedicle is already clearly visible, and the first structures are recognisable (it is practicle impossible to have a very slim galbaldder with difficult cystic pedicle). Within first minute no doubt this is an easy gallbladder		 Frame: 4122 Video: 03913994a1.mp4
		Can be filled, but must be easy moveable		
2	Voluminous gallbladder, slightly withdrawn in liverbed	Slightly withdrawn in liver bed. Easy graspable without punctioning. Filled gallbladder, less pink than grade 1.		 Frame: 2912 Video: 004fdea34e.mp4

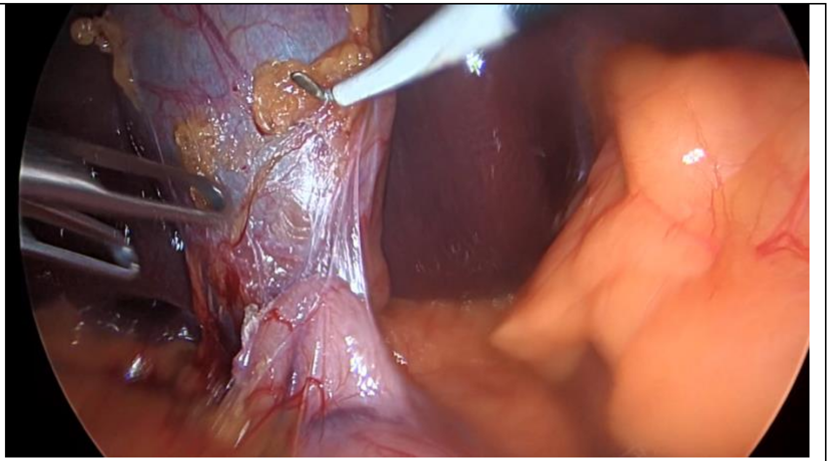
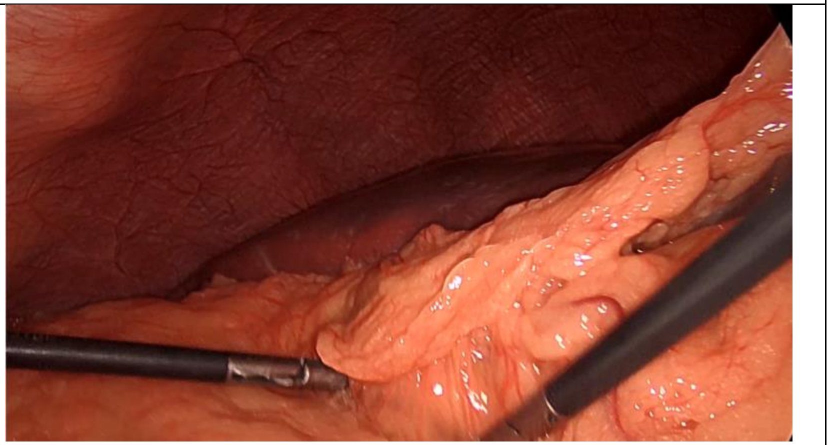
3	Deep fossa, cholecystitis, contracted, stone stuck in gallbladder neck	Fluid coming out of tissue when dissecting Red gallbladder with prominent vasculature Bleedy Needs to be punctured before grasping Totally withdrawn in galbladder bed		 Frame: 14120 Video: 1b52425fb8.mp4
		Empty gallbladder contracted in the liverbed. Difficult to grasp		 Video: 06064559d3.mp4 Frame 2767

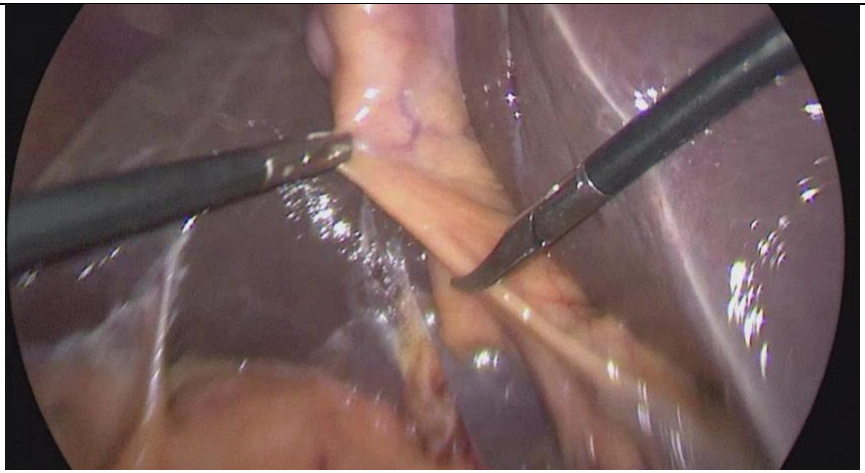
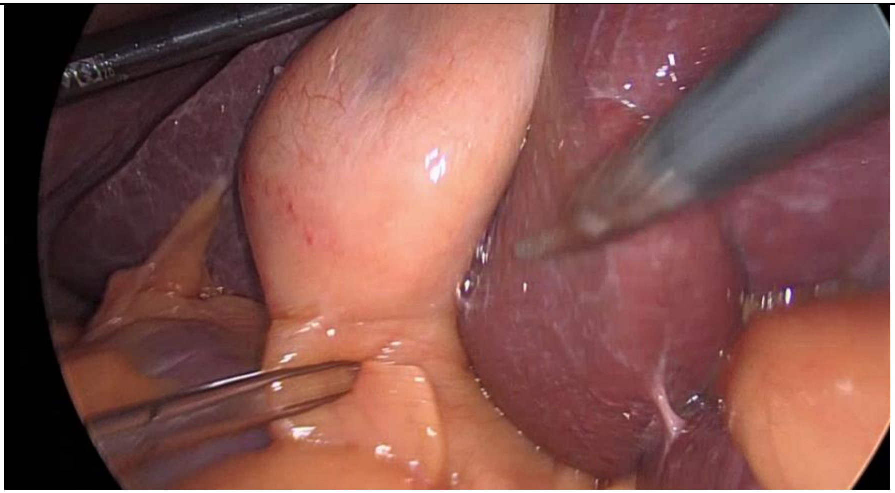
		A filled gallbladder which is also embedded in the liver bed. Difficult to grasp.		 Video: 06372a071.mp4 Frame: 12098
4	Completely obscured, gangrene, mass	Green/back gallbladder, easily damaged by impact.		

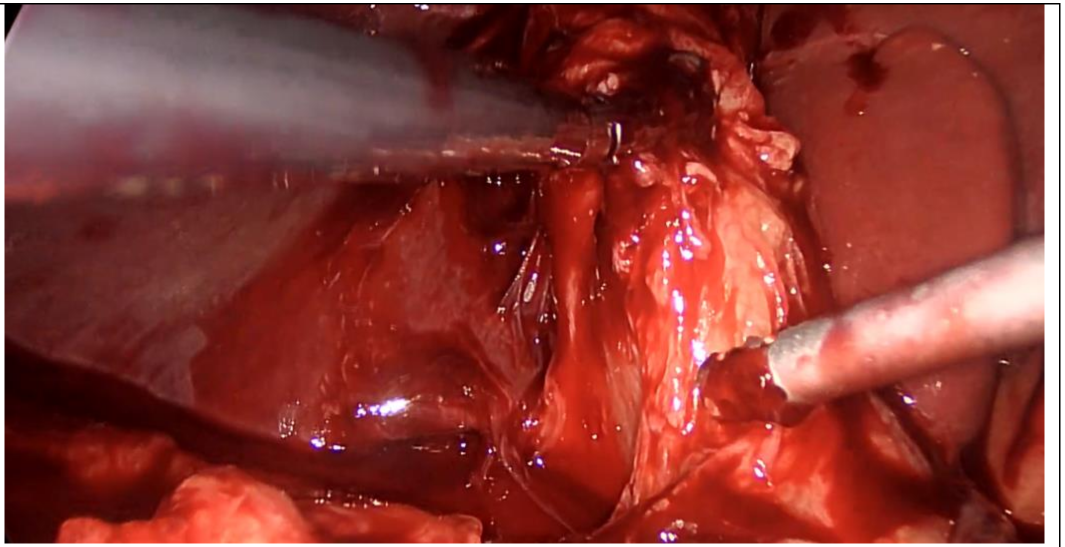
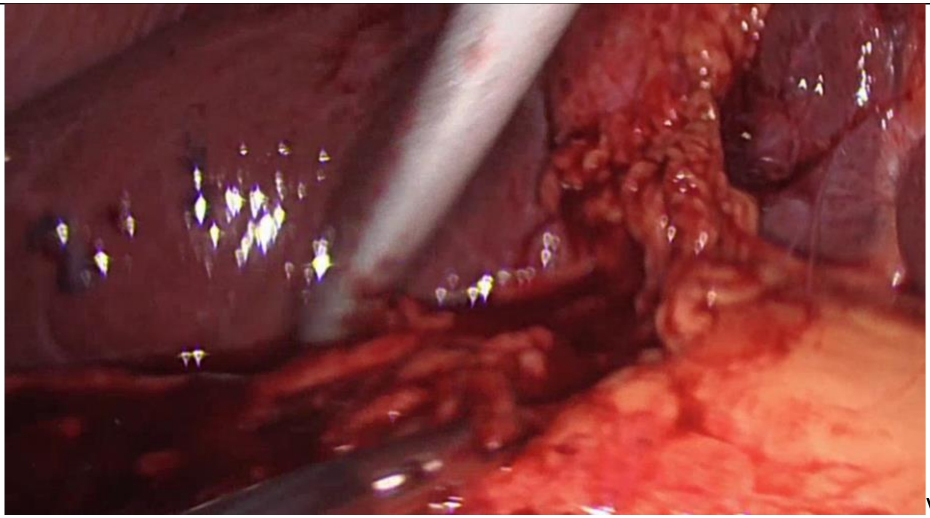
Adhesion grading

Grade	Criteria	Characteristics	Notes	Images
1	Simple up to the neck/Hartmann	Easy dissectible, not on fundus. Are almost immediately gone when grasping. Situated beneath Hartmann	Location is important	 Frame: 4122 Video: 03913994a1.mp4
2	Simple. Up to the body	Easy dissectible on gallbladder fundus. Are almost immediately gone when grasping. Location is of importance!	In some cases, an adhesion on the body can be small/insignificant (image red circle). However, based on location it is graded a 2.	 Frame: 2912 Video: 004fdea34e.mp4

			Adhesions up to the body, but fatty tissue lowers after dissecting, creating a clear overview.	 Video: 27917b6748.mp4 Frame: 3395
			Light adhesions over the body, easily removable.	 Video: 04b305ff0e.mp4 Frame: 5797

3	Dense up to the fundus, involving hepatic flexure or duodenum.	Location is important. Light adhesion involving duodenum are also graded a 3		 Video: 3ddfdeac07.mp4 Frame: 2927
4	Dense fibrosis, wrapping the gallbladder, involving duodenum and hepatic flexure	Wrapping entire gallbladder. Gallbladder is not/barely graspable without dissecting the adhesions		 Video: 1b52425fb8.mp4 Frame: 1140

Pedicle				
Grade	**Criteria**	**Characteristics**	**Notes**	**Images**
1	Thin and clear Structures are visible through pedicle. You know were you need to go. Choledochus is visible through the pedicle	Thin and easy graspable cystic pedicle		 Video: 0302f439fe.mp4 Frame: 202
2	Fat laden	There is easy dissectible fat present on the cystic pedicle which impedes the overview. However, it is clear where to look for the different structures.		 Video: 0e54abc631.mp4 Frame: 7760

3	Abnormal anatomy, cystic duct dilated or obscured.	No good resection plane visible Not clear were the lower border of the galbladder is. No clear overview of were the ductus choledechus is situated		 Video: 1b52425fb8 Frame: 13288
4	Impossible to clarify	Pedicle is impossible to grade the pedicle.		Video: 06372a071.mp4 Frame: 12098

6

General discussion and conclusion

General discussion and conclusion

Scientific and clinical relevance

In this master's thesis, we investigated the possibility to improve the operation time planning of an LC using a model that incorporates both preoperative and intraoperative information. We developed a preoperative pipeline to utilize a wide variety of patient characteristics to estimate the required operation time for an LC. Within this pipeline, parameters can be easily added to expand the model in the future. Different ML models can also be integrated into the pipeline to enhance the initial prediction further. Moreover, we created a pipeline that uses a state-of-the-art model to extract the Nassar grade from LC videos, which could adjust the initial estimation of the preoperative model. New data can be easily added within this model, and the focus can shift from training on the entire Nassar grade to training on a specific sub-score.

Unfortunately, the individual models have not yet yielded suitable results for clinical practice. Based on our evaluation, the preoperative patient model achieved an RMSE of 12.28 ±1.20, similar to the RMSE achieved by the planners at 12.88 ±0.05 on the same dataset. Additionally, our deep learning model was not yet able to accurately predict the difficulty of the gallbladder based on the LC videos, yielding an accuracy of only 36%. However, this research lays the foundation for enhancing operation time predictions while also highlighting the limitations of deep learning and machine learning within a surgical context. Expanding this work to improve accuracy and apply it to a broader range of surgeries could increase patient satisfaction, reduce operating room costs, and improve staff morale. Furthermore, it serves as a stepping stone towards a better intraoperative difficulty assessment, which can be used for surgical benchmarking and ultimately improve surgical outcomes [1, 2]. To our knowledge, no previous research combined preoperative patient characteristics and intraoperative video analysis for dynamic operation time estimation.

Study limitations

Within our study, there are several limitations present in both the preoperative model and the intraoperative model. Therefore, the limitations per model are highlighted. In the preoperative model, our current approach mainly focused on determining operation time based on patient-specific factors. However, operation time is also heavily influenced by factors related to the operating team [3, 4, 5, 6, 7]. We attempted to capture part of this variability by distinguishing between the expertise levels of the operating surgeon. However, this variable is highly simplified, as it only distinguishes between a resident and a fully trained surgeon. Moreover, in some cases, the surgeon listed in the EHR performs only part of the operation, while a resident does the other part. Therefore, a case can be classified as a simple case according the Nassar grade but takes more time due to the composition of the surgical team. It is expected that by extending the surgical factors to include the specific surgeon, the anesthesiologist, and the OR assistant, the accuracy of the operation time prediction can be significantly improved, and a better initial estimation can be made [8].

The second limitation of the preoperative model is the measurement of the actual operation time. The operation time may be prone to measurement errors since it is manually recorded by anesthesiology staff. Operational stress, distractions, and individual work habits can contribute to inconsistencies in operation time measurements, introducing variability that could affect model accuracy. The dependent variable must be reliably extracted to develop a model that can accurately predict operation time. This can be achieved by manually selecting a starting and an ending point in the laparoscopic video. In this way, a more objective operation time can be extracted.

Our intraoperative model also has some limitations. The first limitation is that the intraoperative difficulty assessment is conducted at the start of the operation, allowing only the initial time estimation to be adjusted. However, this adjustment does not account for intraoperative events. For example, if our model predicts an operating time corresponding to a Nassar grade 1 but bleeding occurs, the model does not update its prediction [9]. This results in an incorrect operation time estimation. The primary reason for using the Nassar grade for time predictions was that it enables both operation time estimation and surgical benchmarking simultaneously [10]. While objective difficulty assessment could be precious for further research and surgical performance evaluation, it may not be optimal for precise operation time estimations. To achieve the most accurate predictions, it would be worth considering implementing a deep-learning model that incorporates intraoperative events. A suitable framework for this is a long short-term memory (LSTM) model [11]. A model based on this framework, the residual surgery duration network (RSDnet), can accurately provide real-time residual operation time estimates based on intraoperative findings. In this way, the model accounts for the objective difficulty grade and intraoperative tool handling [12].

Another limitation in the assessment of intraoperative difficulty lies in our labelling methodology. As mentioned earlier, the labelling was performed by a technical physician and a medical student, following a labeling guide developed in collaboration with a trained surgeon. However, the labelling process itself occurred without direct supervision from the surgeon. As a result, the labels may diverge from the surgeon's expert opinion, potentially introducing inconsistencies that could negatively impact model performance. Furthermore, the skewed label distribution and the relatively small dataset may have also influenced the model's performance [13]. Additionally, the modification of the Nassar grade might have reduced the clinical relevance of the Nassar grade, possibly leading to correlations with operation time that are weaker or different from those reported in the literature [9, 14, 15]. Therefore, it is essential to establish a clear correlation between operation time and the adjusted Nassar grade. This is also an important step in combining our model's preoperative and intraoperative parts.

A final limitation of our intraoperative difficulty assessment is the amount of data. Most of our models are trained with random initialized weights. To successfully tune all the weights in these models, a lot of data is required [16]. Although no specific dataset size has been identified for video classification tasks, it is reasonable to assume that a large dataset is crucial. Another limitation of the current model for difficulty analysis is the computational load needed to train the model. If all different data augmentation and mixup methods are implied, the model can no longer be trained on a single GPU [17]. Therefore, in this research model, the complexity and the amount of sampled frames are limited. Regarding vision transformers, the model performance increases if there are more layers and the patch size is reduced, but they expand the number of parameters exponentially [18].

Recommendations

As stated above, it is essential to address the limitations of the individual models. Therefore, for the preoperative model, it is essential to develop an objective method for registering the operation time. This could be done by analyzing the laparoscopic videos and manually labelling the operation's start and end based on prespecified points or phases. In the future, this labelling could be automated by a deep-learning model to handle significant amounts of new data. Moreover, it is recommended that more surgical parameters be incorporated. When distinguishing between the individual surgeons and residents, it is expected that a lot more of the variability present in the operation time can be captured. Regarding the amount of data, it is essential to include more patients. Because we want to determine whether or not certain correlations in our data are present, we recommend to focus on extending the prospective data for a single center. In case the correlations are stronger (above 0.5) for more cases, it could be favorable to determine the correlation over multiple centers in a retrospective way. However, it should be noted that with the current amount of patients stronger correlations were already expected [19]. More data might increase the correlations, but this is not guaranteed. Therefore, we advise first to focus on acquiring objective operation time data and adding surgical parameters before focusing on gathering more data.

For the intraoperative model, assessing whether the MViTv2 model is appropriate for our task is crucial. In the current setup, the model did not outperform existing frame-based techniques when predicting based on the overall Nassar grade. The strength of MViTv2 lies in its ability to capture correlations between different positions within the image, which is expected to be particularly valuable in identifying and classifying adhesions in videos.

Therefore, we recommend focusing on training and testing on adhesions specifically. In this training, it is essential to consider the limitations previously discussed. We recommend using a pre-trained model on the Kinetics-400 dataset or applying a semi-supervised learning method to fine-tune model parameters before training. One possible approach is the DINO (knowledge distillation with no labels) framework [20, 21]. Additionally, we recommend expanding the dataset. A study by Kiyaseh et al. successfully trained a ViT for video classification using 512-930 clips per class [22]. Consequently, we recommend extending the dataset to 300 LC videos and continue extracting clips from these videos. The resulting dataset would contain around 1,200 clips across three classes by selecting approximately four clips per video. Although this is still lower than the number of video clips in the study of Kiyasseh et al., we expect improved results from the model.

To improve the generalizability of the intraoperative model, we recommend leveraging MViTv2's built-in data augmentation framework, with careful parameter tuning to suit our dataset. If performance remains suboptimal compared to frame-based methods after implementing these adjustments, alternative approaches for assessing intraoperative difficulty may be necessary. One potential approach involves building further on frame-based methods. To make such a model feasible, developing a gallbladder detection algorithm to identify frames showing the gallbladder during surgery would be essential. Although this approach could be functional, it may still lack accuracy in grading difficulty due to the absence of adhesion grading. Integrating a frame-wise ViT instead of a temporal ViT could help address this limitation [23]. Combining outputs from both models might enable more accurate difficulty prediction. One advantage of the frame-wise approach is the significant increase in available data due to the large number of frames. However, a limitation of this method is that it can only detect the presence of adhesions, not their severity. Consequently, the model would produce only a binary output.

A next important recommendation is integrating the different aspects of our current work into a single model. This process involves incorporating the Nassar grade into the prediction model and examining its correlation with operation time. Furthermore, it will be essential to determine the appropriate interval rate for making predictions regarding intraoperative Nassar grades. Preferably, this should be done at a small interval. However, this is not yet possible due to the computational time required to make predictions. Therefore, we advise making predictions every 30 seconds of the surgical video after the start of the procedure. This should give the model enough information to make predictions and indicate operational difficulty in the early stages of the operation. However, an optimal value for this should be determined in future research.

Another consideration for future studies is the separation of operation time estimation from difficulty assessment. While difficulty assessment is a valuable tool for surgical benchmarking, alternative deep-learning models may provide more accurate predictions of remaining procedure time. Although both approaches are likely necessary for improving operating room efficiency and overall clinical care, in our specific context, it may be more effective to handle these two objectives separately.

Conclusion

Accurately predicting operation time remains a challenge. Despite the factors of thickened gallbladder wall, indication by indication, and surgical level of expertise significantly correlate to the operation time, we are not yet able to accurately predict operation time. Our current model, using both pre- and intraoperative parameters, lacks the precision required for clinical use. Optimizing the individual components could significantly improve the model's accuracy. However, alternative methods may be needed for intraoperative time estimation, as difficulty is a static measure that is hard to capture with current approaches. Despite this, the work presented in this thesis lays the groundwork for further advancements in surgical planning.

Bibliography

[1] RD Staiger, H Schwandt, Milo Alan Puhan, and PA Clavien. Improving surgical outcomes through benchmarking. *Journal of British Surgery*, 106(1):59–64, 2019.

[2] Yi Wu, Shizhen Li, Jingxiong Yuan, Hang Zhang, Min Wang, Zhenxiong Zhang, and Renyi Qin. Benchmarking: a novel measuring tool for outcome comparisons in surgery. *International Journal of Surgery*, 109(3):419–428, 2023.

[3] Ingwon Yeo, Christian Klemt, Christopher M Melnic, Meghan H Pattavina, Bruna M Castro De Oliveira, and Young-Min Kwon. Predicting surgical operative time in primary total knee arthroplasty utilizing machine learning models. *Archives of Orthopaedic and Trauma Surgery*, 143(6):3299–3307, 2023.

[4] York Jiao, Bing Xue, Chenyang Lu, Michael S Avidan, and Thomas Kannampallil. Continuous real-time prediction of surgical case duration using a modular artificial neural network. *British journal of anaesthesia*, 128(5):829–837, 2022.

[5] Alain Joe Azzi, Karan Shah, Andrew Seely, James Patrick Villeneuve, Sudhir R Sundaresan, Farid M Shamji, Donna E Maziak, and Sebastien Gilbert. Surgical team turnover and operative time: an evaluation of operating room efficiency during pulmonary resection. *The Journal of Thoracic and Cardiovascular Surgery*, 151(5):1391–1395, 2016.

[6] Cornelius A Thiels, Denny Yu, Amro M Abdelrahman, Elizabeth B Habermann, Susan Hallbeck, Kalyan S Pasupathy, and Juliane Bingener. The use of patient factors to improve the prediction of operative duration using laparoscopic cholecystectomy. *Surgical endoscopy*, 31:333–340, 2017.

[7] Bin Zheng, Ormond NM Panton, and Thamer A Al-Tayeb. Operative length independently affected by surgical team size: data from 2 canadian hospitals. *Canadian Journal of Surgery*, 55(6):371, 2012.

[8] Vahid Riahi, Hamed Hassanzadeh, Sankalp Khanna, Justin Boyle, Faraz Syed, Barbara Biki, Ellen Borkwood, and Lianne Sweeney. Improving preoperative prediction of surgery duration. *BMC Health Services Research*, 23(1):1343, 2023.

[9] Ewen A Griffiths, James Hodson, Ravi S Vohra, Paul Marriott, Tarek Katbeh, Samer Zino, Ahmad HM Nassar, and West Midlands Research Collaborative. Utilisation of an operative difficulty grading scale for laparoscopic cholecystectomy. *Surgical endoscopy*, 33:110–121, 2019.

[10] Julian R Abbing, Frank J Voskens, Beerend GA Gerats, Ruby M Egging, Fausto Milletari, and Ivo AMJ Broeders. Towards an ai-based assessment model of surgical difficulty during early phase laparoscopic cholecystectomy. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 11(4):1299–1306, 2023.

[11] Greg Van Houdt, Carlos Mosquera, and Gonzalo Nápoles. A review on the long short-term memory model. *Artificial Intelligence Review*, 53, 12 2020.

[12] Andru Putra Twinanda, Gaurav Yengera, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Rsdnet: Learning to predict remaining surgery duration from laparoscopic videos without manual annotations. *IEEE transactions on medical imaging*, 38(4):1069–1078, 2018.

[13] Seung Hoon Lee, Seunghyun Lee, and Byung Cheol Song. Vision transformer for small-size datasets. *arXiv preprint arXiv:2112.13492*, 2021.

[14] Shakeel Ahmad Mir, Rajandeep Singh Bali, Aijaz Ahmad, and Alam Manzoor Khan. Analysis of laparoscopic cholecystectomy after endoscopic retrograde cholangiopancreatography for choledocholithiasis-a prospective study. *International Surgery Journal*, 9(12):1977–1980, 2022.

[15] Sameh Hashish. Validation of nassar difficulty grading scale in predicting difficult laparoscopic cholecystectomy. *Ain Shams Journal of Surgery*, 16(2):110–129, 2023.

[16] Haoran Zhu, Boyuan Chen, and Carter Yang. Understanding why vit trains badly on small datasets: an intuitive perspective. *arXiv preprint arXiv:2302.03751*, 2023.

[17] Hongyi Zhang. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[18] Hugo Touvron, Matthieu Cord, Alaaeldin El-Nouby, Jakob Verbeek, and Hervé Jégou. Three things everyone should know about vision transformers. In *European Conference on Computer Vision*, pages 497–515. Springer, 2022.

[19] Reshma Bharamgoudar, Aniket Sonsale, James Hodson, and Ewen Griffiths. The development and validation of a scoring tool to predict the operative duration of elective laparoscopic cholecystectomy. *Surgical endoscopy*, 32:3149–3157, 2018.

[20] Jen Hong Tan. Pre-training of lightweight vision transformers on small datasets with minimally scaled images. *arXiv preprint arXiv:2402.03752*, 2024.

[21] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[22] D Kiyasseh, R Ma, TF Haque, BJ Miles, C Wagner, DA Donoho, A Anandkumar, and AJ Hung. A vision transformer for decoding surgeon activity from surgical videos. nat biomed eng, 2023.

[23] Altaf Hussain, Tanveer Hussain, Waseem Ullah, and Sung Wook Baik. Vision transformer and deep sequence learning for human activity recognition in surveillance videos. *Computational Intelligence and Neuroscience*, 2022(1):3454167, 2022.