MASTER THESIS TECHNICAL MEDICINE

AI-BASED PREDICTION MODEL OF SURGICAL DIFFICULTY IN LAPAROSCOPIC CHOLECYSTECTOMY

Ruby Mae Egging
S1567950

EXAMINATION COMMITTEE
prof. dr. I. A. M.J. Broeders
dr. ir. F. van der Heijden
dr. C.O. Tan
drs. P.A. van Katwijk
dr. A.T.M. Bellos-Grob

15-12-2021

meander MEDISCH CENTRUM

UNIVERSITY OF TWENTE.

Abstract

Laparoscopic cholecystectomy (LC) is the standard procedure to remove the gallbladder. Although this procedure has evolved to a relatively safe and tolerable daycare procedure, it can be difficult at times and complications can arise. Complicated gallstone disease, such as cholecystitis or gallstone pancreatitis, are risk factors for increased technical difficulty of a LC. Although it is possible to make a preoperative prediction of the surgical difficulty, perioperative findings can be surprising. Understanding the difficulty of the surgical scenario with AI-based models is important to allow benchmarking in surgical performance and improve planning on the OR. This study aimed to develop a Deep Learning (DL) to predict the difficulty of laparoscopic cholecystectomy on specific operative findings. A difficulty grading scale was used, based on the Nassar score. To train the DL network, frames were extracted from the recorded videos. All frames were labeled for 'gallbladder' difficulty grade 1-3 and 'adhesions' difficulty grade 1-3. Frames consisting of out-of-body images or in which the gallbladder was not visible were excluded. This resulted in a total of 26.483 frames. A ResNet was used as a backbone for the model. Hyperparameters were tuned to improve model results. Both multiclass and binary classification networks were trained. The network that was trained to classify gallbladder difficulty (3-grades) performed better (accuracy 74%) than the network trained to classify adhesions difficulty. It is possible to classify cholecystitis with an accuracy of 91% and classify easy cases with an accuracy of 87%. The results of this study could be used as a starting point for further research in classifying difficulty in LC. This is a first step to improve understanding of surgical scenery and allow benchmarking for surgeons in LC.

Preface

This thesis marks the end of my time as a Technical Medicine student. To this day I am still very glad about the choice I made seven years ago. This study program provided me with the knowledge and experience from the best of both worlds: technology and medicine. In the past year, during my graduation internship in Meander Medical Center, I have been given the chance to contribute to the exciting work of the Artificial Intelligence Lab. It goes for almost everyone that with a nice working environment, a good ambiance, an interesting subject, friendly colleagues and encouraging supervisors you are already halfway there. Luckily, I experienced all this during the last year at the AI lab and surgery department in Meander MC. I would like to thank my supervisor Ivo Broeders, who inspired me with his vision for the AI lab and who I got to learn from as a medical professional by the way he treats his patients. Thank you for letting me be a part of the team. I would like to express my appreciation to my other supervisors, Ferdi van der Heijden and Paul van Katwijk. Ferdi, thanks for your encouragement and advice. Paul, thank you for asking the right questions for me to reflect on the process and your kind words. Because of the daily support I received from Julian and all other researchers and students from the AI-lab I managed to finish this thesis. Finally, I would like to thank my friends and family for their loving support during the last years of my studies.

Table of Contents

1. Introduction

Laparoscopic cholecystectomy (LC) is widely accepted as the standard procedure for the surgical removal of the gallbladder. It is one of the most commonly performed operations in elective and emergency settings.[1] This procedure is indicated for the treatment of symptomatic cholelithiasis, cholecystitis, pancreatitis, or gallbladder polyps.[2] The minimally invasive approach has rapidly become the gold standard for routine gallbladder removal and has essentially replaced the open technique since the early 1990s.[2] The major benefits of laparoscopic versus open surgery include decreased morbidity, faster recovery, and shorter hospital stay.[3] Although this procedure has evolved into a relatively safe and tolerable daycare procedure, it can still be difficult at times. Serious operative complications can occur, such as bile duct injury, bile leaks, bleeding, and bowel injury.[4]

Laparoscopic surgery has challenges compared to open surgery because of the loss of a three-dimensional perception, indirect contact with tissue, and limited tactile feedback. This makes the operation difficult sometimes. A difficult procedure can make a conversion to open cholecystectomy necessary. The definition of a "difficult LC" is inconsistent, but in general, the term refers to multiple technical intraoperative difficulties that increase the risk of complications and significantly prolong operating time.[5] Numerous factors are important in the correct management of LC, because of the large heterogeneity in clinical manifestations.[1] Several factors are known to increase technical difficulties, such as the presence of inflammation and adhesions.[6] In addition, surgical difficulties not only depend on patient factors but also on the surgeon's experience and skills.

Prediction of surgical difficulty is required to identify high-risk patients to minimize the risk of complications and delays.[7] A difficult procedure requires an experienced surgical team, experienced at both laparoscopic and open cholecystectomy. Furthermore, estimating the level of complexity beforehand may improve the flexibility and accuracy of preoperative planning. Multiple LC grading scales of surgical difficulty have been developed based on several predictive preoperative factors. However, the discriminative value of these preoperative risk factors is low, and therefore surgical findings can still be surprising.[8,9] Therefore, it would be of additional value to incorporate the actual surgical findings in the first few minutes of surgery to adjust the prediction of surgical difficulty. There is a consensus on the fact that operative findings and difficulty hold the key to outcome.[1] LC has surprisingly variable approaches, findings, outcomes, and conversion rates. Standardizing the way surgical events are reported is important to allow qualitative studies and outcome comparisons. This is also important to gain more insight into the learning curve of surgeons in training. If it is clear which procedure was easy and which was difficult, it will enable more objective comparisons between surgeons in performance, procedure time, and/or complications. This is currently not documented regularly after surgery. Automatic, objective classification of surgical difficulty can improve outcome predictions, risk stratification, and surgical planning.

The laparoscopic video recordings show essential information regarding anatomy and lend themselves for analysis. Deep Learning (DL) is a subset of Artificial Intelligence (AI) that imitates the way humans learn. DL-based predictive models have already shown to be of great benefit in

improving healthcare quality and safety. For example by supporting healthcare personnel in clinical documentation and decision making.[10] This study aims to predict intraoperative difficulty by classifying and evaluating intraoperative images in the first phase of a LC. The concepts of DL and previous research on this topic will be further elaborated in the next chapter. The aim and research questions will be discussed at the end of the next chapter.

2. Clinical Background

This chapter discusses the clinical background of the LC procedure and the risk factors for a difficult LC. The definition of a difficult LC is elaborated, and different scoring systems are discussed. This is followed by an overview of previous research on DL in surgery, followed by the aim and research questions for this study.

2.1 Laparoscopic Cholecystectomy

LC is the standard procedure for the removal of the gallbladder. The indications for LC are cholelithiasis, pancreatitis, cholecystitis, gallbladder polyps, and choledocholithiasis.[11] The anatomical structure of the gallbladder is typically divided into three parts: the fundus, the body, and the neck *(Figure 1 and 2)*. The neck contains a mucosal fold, known as Hartmann's pouch. The LC procedure consists of the following steps: division of adhesions involving the gallbladder (gallbladder dissection), dissection of the hepatocystic triangle, clipping and transection of the cystic artery and duct, and the dissection and removal of the gallbladder from the liver bed.

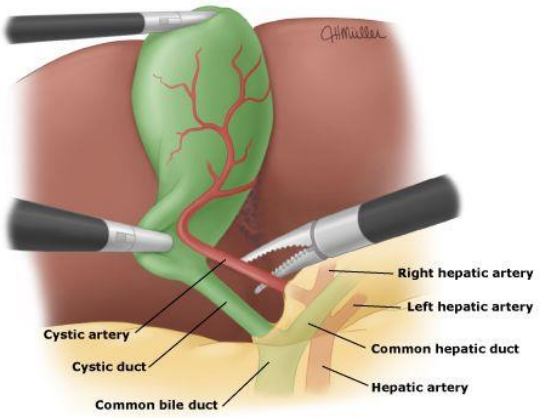


Figure 1 Overview of the anatomy

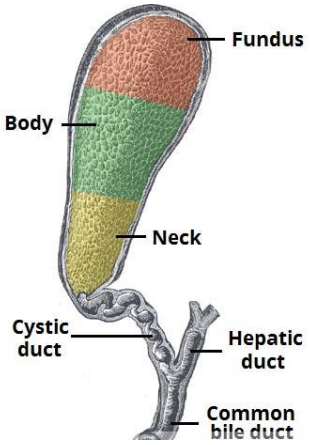


Figure 2 Anatomy of Gallbladder

In more detail, the procedure starts by establishing access to the abdomen and the creation of the pneumoperitoneum. When access has been established, the trocars and instruments are inserted. The gallbladder is elevated with an instrument by traction of the fundus *(Figure 3)*. If adhesions are present, they are taken down first. When a percutaneous cholecystectomy drain is in situ, it must be pulled out before the gallbladder can be removed. After percutaneous drain placement, several adhesions may be encountered, which should be separated first. (Figure 5) The infundibulum is subsequently retracted caudolaterally to straighten the cystic duct, maximizing visibility and elevating the gallbladder from the common bile duct. The next step is the dissection of the hepatocystic triangle. The dissection begins by incising the peritoneum along the edge of the gallbladder on both sides to open the hepatocystic triangle. This is performed mostly with blunt dissection.[4] Subsequently, the critical view of safety (CVS) is established. By achieving the CVS the cystic artery and duct are localized. Only when the lower part of the

gallbladder is separated from the cystic plate, the hepatocystic triangle is cleared of fat and fibrous tissue and there are only two structures attached to the gallbladder, the cystic artery and duct can be clipped and divided.[12] This is needed to prevent misidentification of the cystic artery and duct that could lead to biliary injury. Then the gallbladder is separated from the liver bed and removed from the abdominal cavity with a sterile retrieval bag.

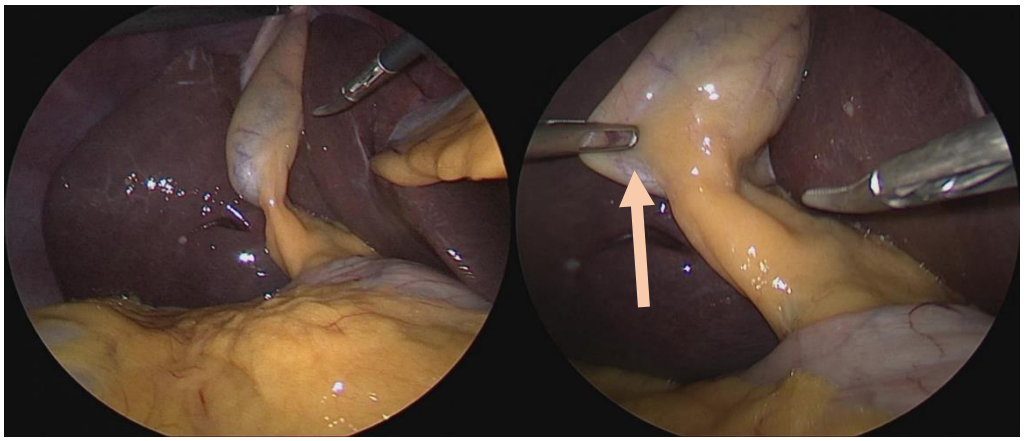


Figure 3 Start procedure: Evelation of the gallbladder

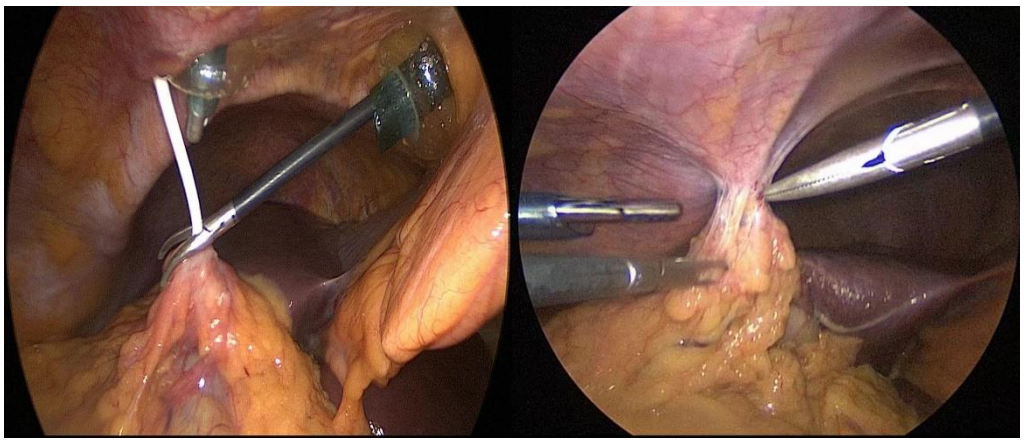


Figure 4 Hartmann's pouch

Figure 5 Laparoscopic images showing a percutaneous drain in situ

The average operating time for LC differs significantly among hospitals.[13] Here we define the operating time as the duration from the first skin incision to skin closure. In Meander MC the average time that is reserved for a LC is 45 minutes, although the actual time needed may vary depending on surgical difficulty. Thiels et al.[14] report a mean operating time of 89 ± 32 minutes. Sutcliffe et al.[15] report a median operating time of 60 minutes (45-88). In a study by Atta et al.[16] they found that the operating time of LCs performed by trained surgeons was significantly shorter

(median, 45 min; 30-70) compared with the operating time of surgeons still in training (60 min, 50-90).

Although LC is considered a safe procedure, morbidity can occur in approximately 6-8% of the patients. Complications include bleeding, abscess, wound complication, and bile duct injury.[4] Among postoperative complications, the most common ones are bleeding from the abdominal cavity and infection of the surgical wound. Biliary and vascular complications can be life-threatening, while minor complications cause patient discomfort and longer hospital stay.[17] The conversion rate has decreased over the last few years.[18] Mortality after LC is low. Sandblom et al report a 30-day overall mortality rate of 0.15%. The median hospital stay is 1 day.[15]

2.2 Risk factors for a difficult surgery

Surgical difficulty should also be estimated preoperatively. There are variations in studies when looking at predictive factors because they vary in research outcomes. Studies that focus on surgical difficulty either focus on the duration of the operation, complications during or after surgery, or conversion rate.[19] During surgery, most difficulties are experienced during the separation of all adhesions and ligation and division of the cystic artery and duct. Sahu et al.[5] state that 75% of all difficult surgeries (defined as procedures >90 min) are due to the separation of adhesions. Atta et al.[3] describe that patients with an impacted stone in the neck of the gallbladder, adhesions in Calot's triangle, and gallbladder rupture were more likely to result in a difficult LC. Inflammation can cause local changes in the tissue within and around the gallbladder. Scarring in the hepatocystic triangle can make it hard to identify the cystic artery and cystic duct. Predictive surgical findings for conversion are a completely obscured gallbladder, impacted stone, bile, or pus outside gallbladder and fistula.[20] Other factors that make the surgery difficult are dense adhesions, diffuse fibrosis, bleeding and necrotic tissue.[21] Difficulty can also be caused by the creation of the pneumoperitoneum, excision of the gallbladder from the bed, or extraction of the gallbladder.[5]

Many studies have reported different preoperative risk factors associated with difficult LC and conversion to open cholecystectomy.[222324] Male gender, high age, obesity, previous upper abdominal surgery, acute cholecystitis, and little surgical experience are predictors for a difficult LC in both acute and elective cases.[2526271820] Patients with a gallbladder wall thicker than 0.5 cm, a contracted gallbladder, high age, male gender, and acute cholecystitis have a higher risk for conversion.[8] The gallbladder wall will become thicker because of the inflammation and fibrosis caused by acute cholecystitis. In the first few days of acute cholecystitis, there is edema, hypervascularity and gallbladder distension. After a few days there will be adhesions and it becomes more difficult to dissect.[17]

Sutcliffe et al. proposed a preoperative scoring system to predict the need for conversion from laparoscopic to open surgery. This system gives a score to age, gender, indication, ASA score, gallbladder wall and common bile duct diameter. They correlated this score with intraoperative difficulty. Their scoring system could be used for preoperative prediction of surgical difficulty.[15]

2.3Scoring systems of intraoperative difficulty

Because of the variability in operative findings, LC is a very unpredictable surgery. Mainly because of the effect of cholecystitis and fibrosis surrounding the hepatocystic triangle. There are many preoperative grading scales that are used to predict a difficult surgery based on different anatomical and laboratory findings. It depends on how 'difficulty' is defined which parameters are of high importance for prediction. The definition is not clearly defined, because it also depends on the surgeon's experience and skills.[28] Lal et al [9] and colleagues suggest that the surgery is 'difficult' when operating time is more than 90 minutes or either taking down the adhesions surrounding the gallbladder or dissecting the hepatocystic triangle takes more than 20 minutes.[20] In the study of Kumar Sahu, difficult LC was defined in those procedures which exceeded 90 minutes in duration and/or converted to open procedure.[5]. In Atta et al [3] it is defined as the operative time of more than 60 min and/or cystic artery injury. Given the fact that LC is a common surgery that varies in operative difficulty, it is surprising that not many intra-operative difficulty grading systems have been published and none are widely used in clinical practice. With a simple scale to arrange intra-operative difficulty, it would be easier to improve intra-operative strategy and planning, to compare between studies, allow risk assessment for patient outcomes and provide insight into training progression for surgical trainees. It would also standardize the description of findings and reporting of outcomes. [29]

Numerous scoring systems have been developed in the past that aim to predict the level of difficulty for LC. The majority of these systems are based on preoperative clinical findings and do not incorporate intraoperative findings. It is clear that preoperative parameters have some predictive value, but it still lacks accuracy until the gallbladder is visualized, after which a true determination can be made of surgical difficulty. Not many intraoperative grading systems have been developed. Sugrue et al investigated if the proposed operative scoring system is useful to predict if conversion from laparoscopic to open surgery is necessary. This publication did not report clinical outcome data and has yet to be validated.[1] Their grading system is based on the severity of cholecystitis and difficulty grade with a score from 1 to 10. Cuschieri [23] published a 'scale of difficulty' for LC, but is probably outdated because skills have significantly improved over the last 20 years. The Tokyo guidelines to determine the severity of cholecystitis use three grades, but do not incorporate intraoperative findings.[30] The 'Parkland' scale though, published in Madni et al [31] is based on intraoperative images but they described outcome data for only 50 patients. The Nassar scale is a simple, clinically relevant, operative difficulty scale that could be used as a tool. Griffiths et al [29] proved the utilization of this grading system and clinical applicability in association with outcome data. In our opinion, the Nassar score is the most accurate grading scale to classify surgical difficulty within the first few minutes of the operation. *(Figure 6)* The scale is presented below. Automatically defining the status of the gallbladder intraoperatively will enable more standardized reporting and improve comparisons of outcomes.

2.4 The Nassar scale

This is a simple 4-grade LC difficulty scale that was published in 1995.[32] *(Table 1)* It provides a tool for reporting operative findings, disease severity, and technical difficulty.[29] Griffiths et al found that an increasing Nassar operative difficulty grade was consistently associated with significantly worse outcomes. Other publications used the scale to improve the management of complicated gallstone disease[3334] or investigate the suitability of patients for single-port LC[35].

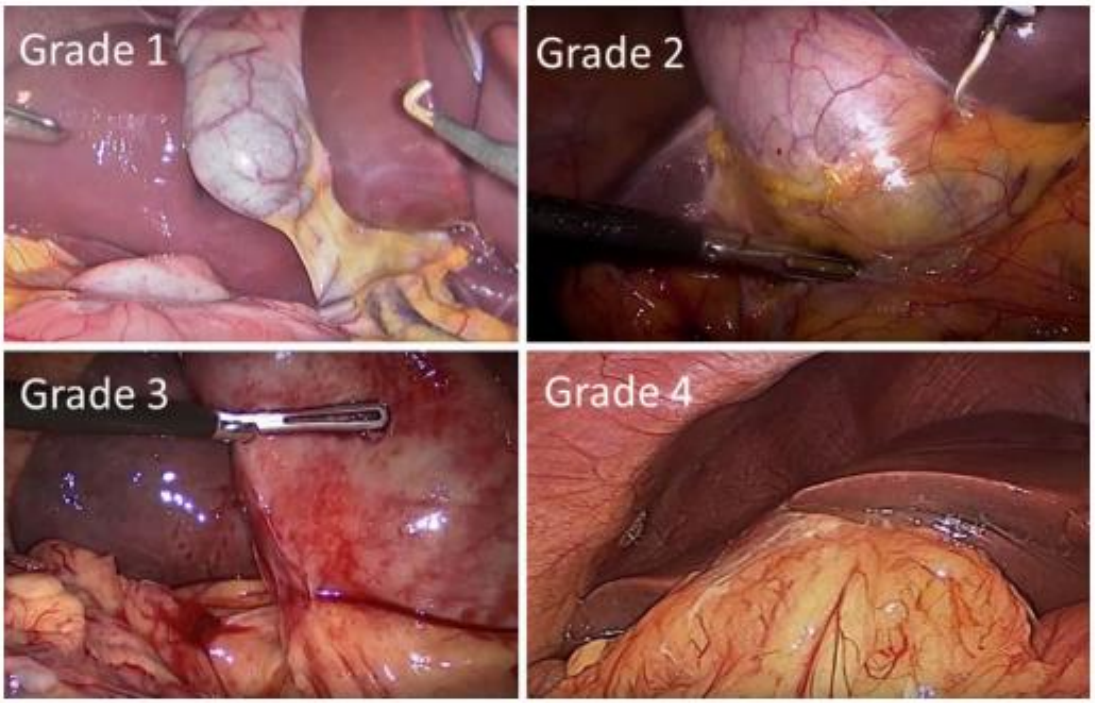


Figure 6 Laparoscopic images of each Nassar operative difficulty grade[29]

Table 1 Nassar scale definitions

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.	

2.5 Artificial Intelligence in surgery

The huge amount of laparoscopic video data holds a lot of information that could be used to learn from and make valuable predictions. Analyzing these videos can assist surgeons during surgery and postoperatively improve the evaluation of the procedure and the performance of the surgeon. AI techniques have gained significant popularity as a tool to analyze surgical videos. AI methods use algorithms to give computers the ability to perform tasks that usually require human intelligence. Algorithms such as deep neural networks can be trained without explicit programming, using large quantities of data to learn to predict an outcome on new data. AI techniques are not new as they have existed for decades, but due to developments in computing power and the increase of information, it has gained possibilities and popularity. In short, it describes a scientific field that develops algorithms to train computers at performing a specific task. It has already proven to be beneficial to healthcare quality and safety.[10] While interest and research on AI applications on surgery are increasing, much of the focus has been on other specialties such as radiology, pathology, or dermatology.

An overview of current DL applications in surgery is given in Figure 7. Different techniques include anatomy detection, instrument detection, action recognition, phase recognition, and remaining surgery time prediction.[36] Twinanda et al. developed a prediction model for the estimation of remaining surgery time. The average error in the estimation was 15.6 min.[37] Models that can detect tools are already showing good results.[38] AI algorithms are studied to enhance the safety of LC, enable benchmarking, and improve surgical training programs. It could be used for decision support and digital documentation of operative findings.[39] Madani et al.[40] suggest that DL can be used to identify safe and dangerous zones of dissection and other anatomical structures in the surgical field during LC with high accuracy. Mascagni et al [41] formalized a reproducible method for objective video reporting of CVS in LC. Tokuyase et al [42] developed a system that outlines landmarks in endoscopic images in real-time to prevent intraoperative bile duct injury. In addition to laparoscopic videos that are used for action recognition, the use of external cameras in the operating room (OR) is also being studied. These cameras capture more general activities in the OR process rather than anatomical structures.[43] All these developments mainly go towards context-aware systems that can give automated assistance and intelligent surgical training systems.[38]

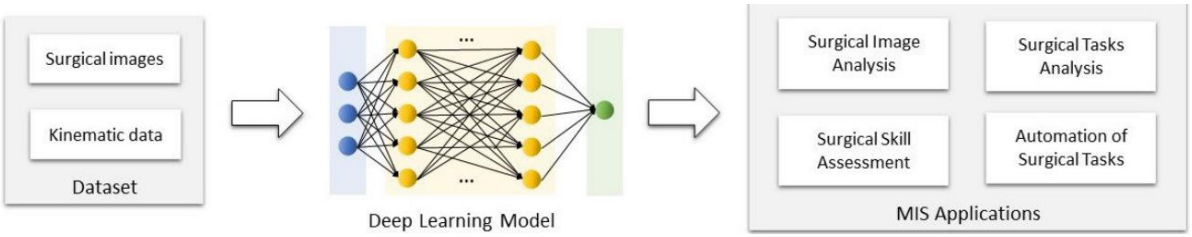


Figure 7 Main DL applications in minimally invasive surgery[38]

Because LC is a basic laparoscopic surgery that is widely used in many different hospitals, this procedure is often chosen to explore the possibilities of AI and evaluate the feasibility of the techniques.[444546] Nevertheless, LC is not fully standard because of the variance in difficulty and

experience of the surgeon.[44] We did not find other studies on difficulty prediction using laparoscopic video data.

Challenges in DL for medical purposes

Although the results of the abovementioned studies are promising, the application and value for real-time surgical guidance and decision support are complex and have yet to be demonstrated. Laparoscopic video data gives an extra challenge to image analysis. These videos have huge variability in terms of background noise, image quality, and camera angles, and blood and smoke in the images. If a model is trained on data from a certain institution, it is hard to use it for the same task in another institution, due to other camera settings or labeling protocols. Hashimoto et al.[47] suggest the creation of publicly available surgical datasets, allowing standardized protocols. Anteby et al.[36] state in their review that it is of great importance to give an accurate description of how the data is labeled and assess inter-annotator reliability. Another obstacle is the recruitment of experts (surgeons) to annotate data to provide clinically meaningful data. It would be beneficial if more surgeons are engaged in this type of research. Because of the poor quality of datasets, the implementation in the OR is still a challenge.

2.6 AI-lab in Meander Medical Center

Recently the 'Artificial Intelligence Lab' has been established in the Meander MC. The goal is to explore the possibilities and create a platform for AI projects within Meander and support these projects. Allowing surgeons to be at their best by giving insight into their performance is one of the main goals. In addition, AI-based models can be used to assess intraoperative decision-making and give real-time feedback during surgery. Previous work includes the identification of anatomical structures, surgical phase recognition, and detection of bile leakage. The focus is also on the use of external cameras in the operating room to recognize actions and improve planning. Patient history data from Meander will be used to create a preoperative prediction model for surgical difficulty in LC. Johnson & Johnson also has an interest in these projects and is available for technical support. The goal is to eventually combine the results of these projects using surgical phase recognition, preoperative and intraoperative prediction of surgical difficulty.

2.7 Aim and research questions

Below the long-term objectives of this study are presented: benchmarking and improving planning

Benchmarking

To improve surgical performance, it would be beneficial to set a benchmark for surgeons. When surgeons gain more insight into their performance and can compare it to the benchmark, they can apply a more targeted approach to improve their skills. If it is registered which actions are performed in a certain phase, the focus can be shifted in the learning process. Therefore, applications such as surgical phase recognition, instrument tracking, and action recognition are required. In addition, recognition of surgical events such as bile leakage or bleeding and the automatic assessment of the critical view of safety is needed. To allow benchmarking for surgeons

in LC it is also important to objectively classify surgical difficulty. The automatic reporting of surgical difficulty makes it possible to compare surgical outcomes.

Improve planning

The OR is a scarce and expensive environment and efficient planning is of great importance. Because of the variability in complexity, the operating time varies. Predicting remaining surgery duration is needed for optimal OR planning. Classifying surgical difficulty at the start of surgery could be used to predict the remaining surgery time more accurately. When this is done automatically, the next patient could also be called in automatically and therefore save time. The ultimate goal is to make a preoperative prediction of required surgery time that is automatically adjusted accurately during surgery.

Aim of this study

To contribute to this objective to realize benchmarking in LC and improve surgical planning, this study aims to objectively classify surgical difficulty. It would be clinically most relevant to do this at the start of surgery. Classifying surgical difficulty is a small step in improving patient care by giving surgeons more insight into their performance and improving planning.

Research question

- To what extent is it possible to predict surgical difficulty in the first phase of the operation with a Deep Learning model using intraoperative video data?

Subquestions

- Which aspect of the Nassar scale can be accurately classified using single frames?
- To what extent is the Nassar scale clinically relevant to solve this problem?
- To what extent can surgical planning be improved by predicting surgical difficulty in the first phase of the operation?

To answer these questions, the Nassar scale was used as a scoring system for surgical difficulty. A dataset was created using LC videos and these were labeled with the Nassar score as a starting point. A DL algorithm was developed to train the networks. Several experiments were executed to improve the results.

3. Technical Background

This chapter serves as support to provide the fundamental knowledge regarding the method used in this thesis: deep neural networks. First, an introduction to AI is given. Then, Artificial Neural Networks (ANNs) are described. Convolutional Neural Networks (CNNs), which can be viewed as a special type of ANNs, are later described in 3.2. In the next section, optimization techniques are presented to successfully train the network models. Ultimately, the neural network architecture is presented that will be used in this research.

3.1 Artificial Intelligence

It is hard to explain the overarching term 'AI' in one definition, but it can be described as the use of computers to simulate human intelligence. Such algorithms try to mimic human cognitive functions such as learning, decision-making, and problem-solving. They are applied and proven useful in a wide variety of fields, including medicine.[48] Machine learning is a subfield of AI which enables algorithms to learn patterns in large quantities of data without being explicitly programmed and to generate useful predictive outputs on new data.[49]

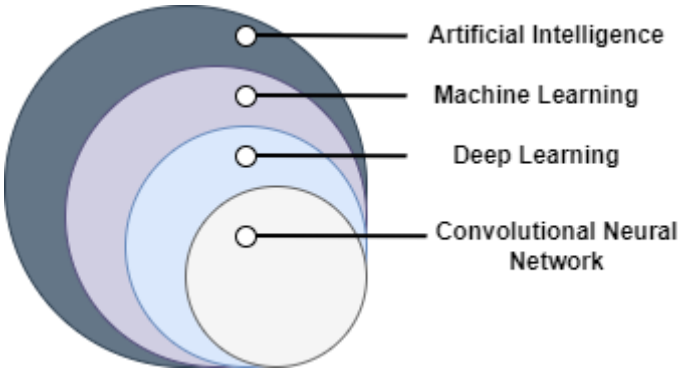


Figure 8 Overview Artificial Intelligence

In supervised learning, datasets that contain pre-labeled outcomes are used to train a model in such a way that it can make predictions on new, unseen data. The data is used to learn features that have a predictive value. Supervised learning can solve either a classification or a regression problem. Examples of classification techniques include support vector machines, neural networks, deep neural networks, decision trees, random forests, and naïve Bayesian classifiers approaches. Unsupervised learning uses an unlabeled dataset to train a model. The goal is to interpret and derive structure from data by extracting features and patterns in the data.[50] Unlike supervised and unsupervised learning, where a fixed dataset is used, reinforcement learning systems use a feedback loop between the system and its experiences. This type of learning uses the principles of behaviorism (reward and punishment) to train the algorithm.[51] It has been applied relatively little for medical purposes. Because fully labeled datasets are very hard to acquire, often semi-supervised learning methods are used.[52] In the present work, supervised learning is used.

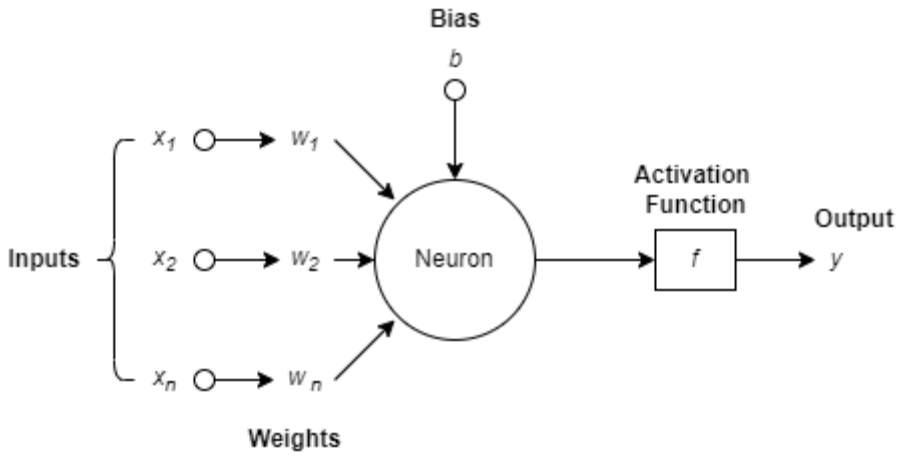


Figure 9 Neuron

ANNs are computational analytical tools inspired by the biological brain. They are made up of multiple neurons that are densely connected. These artificial neurons perform a computation on the input by multiplying it with a learnable weight and by adding a bias and applying a non-linear activation function to pass the output to the next neuron in the system *(Figure 9)*.[43]

An ANN consists of input neurons, several internal (hidden) layers, and output neurons. Each neuron in a layer is fully connected to all neurons in the previous and next layers. Figure 10 shows a network architecture with two hidden layers. The way all sequential layers are organized defines the architecture of the network. Increasing the number of layers increases the depth of the network. Expanded versions of ANNs containing more hidden layers are called Deep Neural Networks (DNN).[43]

During learning, the input passes through the network, and the network's predicted outcome label is compared with the annotated reference label using a pre-defined loss function. This is called the forward pass. The computed error is then backpropagated through the network to adjust all the weights and biases. Repeated iterations of this forward and backpropagation to optimize these parameters produce a network that can hopefully accurately make predictions. The network is a predictive function that is comprised of many non-linear neurons with learnable weights and biases.[53]

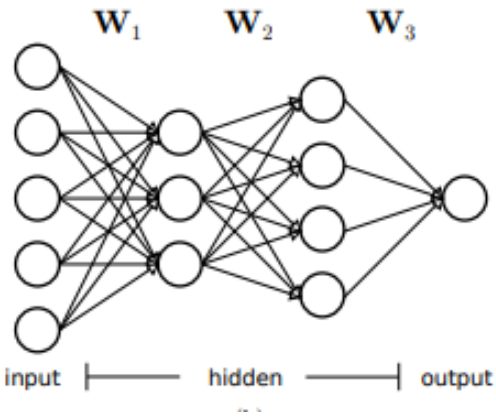


Figure 10 A 4-layered ANN architecture 43

The main benefit of DNNs is that they do not need manually defined features. They can learn features from large quantities of data with a high level of abstractness. They have proven the ability to solve numerous complex problems, including for example speech recognition.[54] Although DNNs can be very effective, they are not sufficient for input data containing spatial information, because the network takes the vectorized version of the input. That is why for image processing, convolutional neural networks (CNNs) are the solution because of the use of convolutional filters. It may sound counterintuitive, but the hardest problems for AI to solve are often the easiest for humans. Humans can easily solve these problems by intuition but can hardly describe how they

know what they see. CNN's are inspired by the structure of the visual cortex in the brain.[55] They make it possible to extract abstract features from images and make it possible to solve image classification problems. Image classification forms the basis for many other computer vision tasks, such as localization, detection, and segmentation.[55] The hierarchy of all these techniques is visualized in Figure 8.

3.2 Convolutional Neural Networks

CNN's have dramatically improved the results of image recognition tasks. Pixel values in medical images contain spatial information. A CNN aims to reduce the image into a simpler form without losing features because features are essential to achieve an accurate prediction. A CNN consists of three main layer types: convolutional layers, pooling layers, and fully connected layers. [43]

Convolutional layer

A CNN can learn the spatial dependencies in an image through the use of convolutional filters. These filters work as operators and extract features of the input images. In this context, these filters are usually called kernels. A kernel is a 2D convolution matrix which, depending on the parameter values, filters the input image in different ways. During the forward pass, many different filters convolve through the entire image by shifting with a certain *stride*. The result is a feature map that gives the responses of that filter at every position of the image. These filters have the same depth as the input image dimension. In color images, this means the network has to learn multi-channel filters.[56]

Activation layer

All learnable convolution layers are often followed by an activation function that is usually a ReLu function. This is a nonlinear function that outputs the input only if it's positive.[57] These nonlinear elements in the network are needed because class information is usually hidden in entangled distributions of the image data. The nonlinearity of the mapping of the network is needed to allow disentangling these distributions.

When the desired output is a probability, the sigmoid function is the right choice. The function exists between 0 and 1. For a binary classification problem, the output layer consists of one neuron with a sigmoid activation function. The threshold is set at 0.5. For multiclass classification, the softmax function is preferable. This function outputs the relative probabilities of each class, which all sum up to 1 *(Figure 11)*.[57]

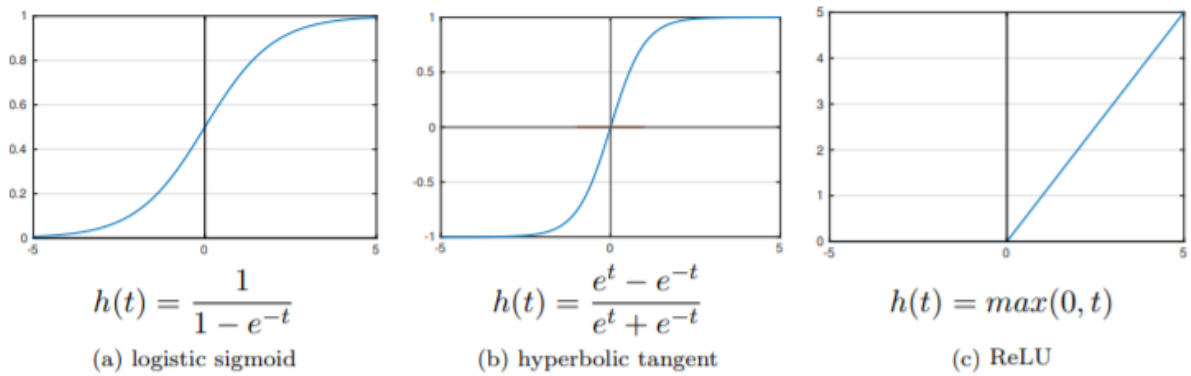


Figure 11 Illustration of non-linear activation functions[43]

Pooling layer

The function of the pooling layer is to reduce the spatial size of the input features to represent the information in the input data in a more compact way. It extracts the dominant features from the image by simplifying the output. Two common types are max pooling and average pooling. In Figure 12, max pooling is visualized. It uses the maximum value in the feature map.[58] Therefore, it lowers the resolution and reduces overfitting to the specific input image. Adding more convolutional and pooling layers may improve capturing low-level features but require more computational power.[56]

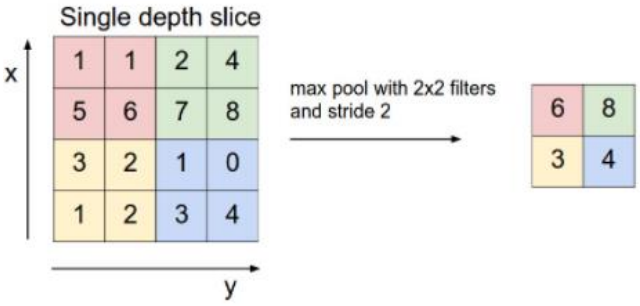


Figure 12 Max pooling [56]

Fully connected layer

The fully connected layer combines the outputs of the previous layers to estimate the class probability scores. If multidimensional data is used, as with color images, a flatten layer is used. The flatten layer transforms the three-dimensional image to the fully connected layer. An overview of all described layers is given in the figure below.

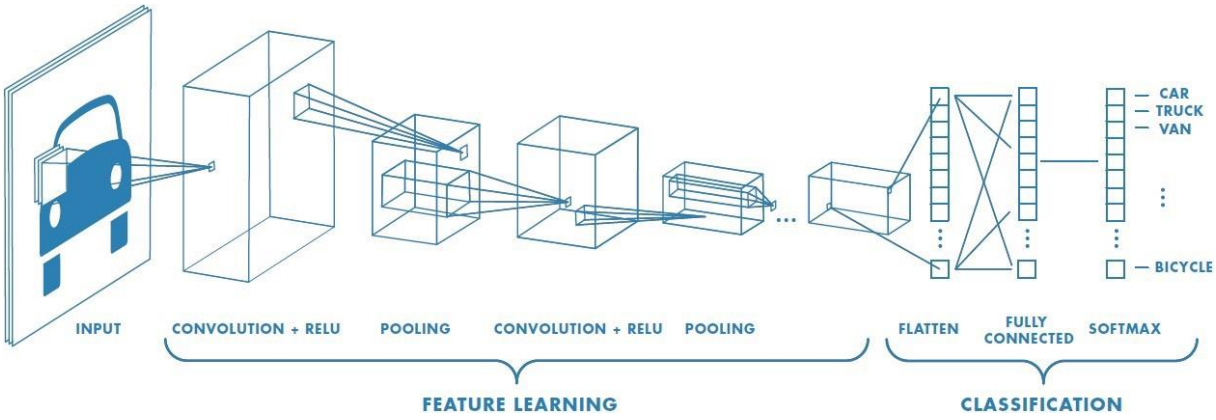


Figure 13 Overview CNN architecture [59]

3.3 Hyperparameters: Optimization

An overview of a training process is given in Figure 14. The weights and biases of a network are trainable parameters which means that they can be optimized using the training set. Apart from these, there are also so-called hyperparameters that cannot be optimized in such an optimization session. There are different categories of hyperparameters:

- The parameters of the optimization process. For instance, in a gradient descent approach for optimization, there is a parameter called 'learning rate'.
- The parameters determining the network configuration. Examples are the number of filters per layer, the filter sizes, the number of hidden layers, etc.
- Additional parameters for tweaking the training set and the network. For instance, parameters for data augmentation.

For the selection of these hyperparameters, a second dataset is needed, called the validation set.

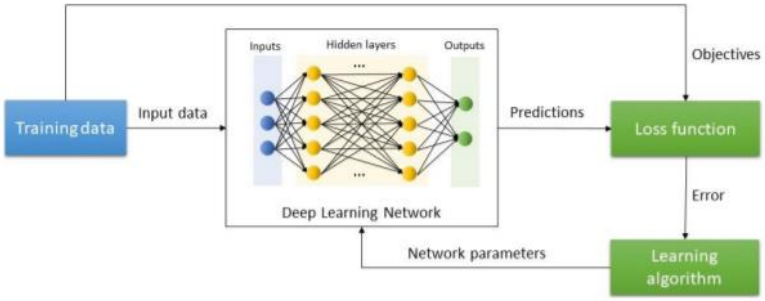


Figure 14 Training process in DL algorithms[38]

Loss function

The learning step is performed by minimizing a certain loss function. When input data passes through the network and the network's outcome is compared to the ground truth label, the error is quantified using a loss function that is computed over the full dataset or a subset of the dataset. As such, this loss function defines an error surface in a multi-dimensional landscape that is formed by the weights and biases. These weights and biases are adjusted in the opposite direction to the gradient of the loss function. This is the gradient descent approach. To optimize performance, the lowest point in the error surface has to be found.

Several loss functions can be used, but cross-entropy loss has shown to converge faster and better than others.[45] With this function, the difference between two probability distributions is measured. It tries to maximize the log-likelihood for both classes. If there are more than two classes, the loss is calculated separately for each class label and the results are summed.[60] When a class imbalance exists, extra weight must be given to the class that is less represented in the dataset. The last layer of a network consists of a softmax layer which holds as many neurons as that there are classes. Each class is associated with one of these neurons, and for each class, the associated neuron calculates a "probability" value between 0 and 1. In the ideally performing network, all

neurons would give an output of 0 except the one associated with the ground truth class. This one would give 1. In that case, the contribution of this data item to the loss function would be zero. A network, which is completely indecisive, that is, all neuron outputs are equal, would have a cross-entropy loss of $\log_2(N)$, where N is the number of classes. For instance, in an 8-class problem, an indecisive network would have a loss of 3 as $2^3=8$.

Optimization algorithm

To learn the weights and biases during training, an optimization algorithm must be chosen. In the gradient descent method, already mentioned above, the weights and biases are updated by computing the gradient vector of the loss function. Gradient descent is the most common method to optimize the network. Three variants are batch gradient descent, stochastic gradient descent (SGD), and mini-batch gradient descent.[61] SGD performs the computations on a small subset of the data instead of the whole dataset. This reduces the computational power required. Another popular algorithm that was proposed in 2015 is the Adaptive Moment Estimation (Adam). This algorithm is also based on gradient descent, but it computes individual adaptive learning rates for different parameters. It is supposed to be faster and more reliable reaching a global minimum.[62] Although Adam is now often the default optimization algorithm, an argument is that SGD generalizes better and results in better performance.[6364]

Learning rate and batch size

The magnitude of the adjustments in the parameters in each iteration is determined by the learning rate. Setting a small learning rate implies that minimizing the loss function takes a long time. Setting a learning rate that is too large results in too much change in response to the loss function. Therefore, it results in an unstable training process. To pass the entire dataset through the network once is called an epoch. The number of training samples that are run through the network for one training step is called the batch size. Setting a higher batch size means processing more training instances in each iteration, resulting in a faster training process. The batch size influences the stability of training and the generalization performance of the model. The learning rate, number of epochs, and batch size are hyperparameters that have to be optimized by trial and error, by testing the performance during the training process.[43]

3.4 Network optimization

The aim is to train a network that generalizes well from the training data to new, unseen examples. The dataset has to be split into a training set, a validation set, and a test set. During the training process, the validation set is used to give an independent estimate of the model's performance. By using a validation set during training, overfitting of the models can be detected. This makes it possible to tune the hyperparameters. The test set is another independent dataset, that is used to evaluate the model's performance after the training process. A commonly used split ratio is 70% for training, 10% for validation during training, and 20% for testing. To train a good model, a large dataset is needed that contains many examples of each class. It is also important to ensure the same class distribution in the training, validation, and test set, to correctly interpret the model's performance. When a model is trained too long on the training data, it learns the details and noise of this training data and can therefore not make accurate predictions on new data.

Overfitting means the network fits too well on the training data.[65] Overfitting on the training set is a common problem in DL. Several methods to reduce overfitting are elaborated below.

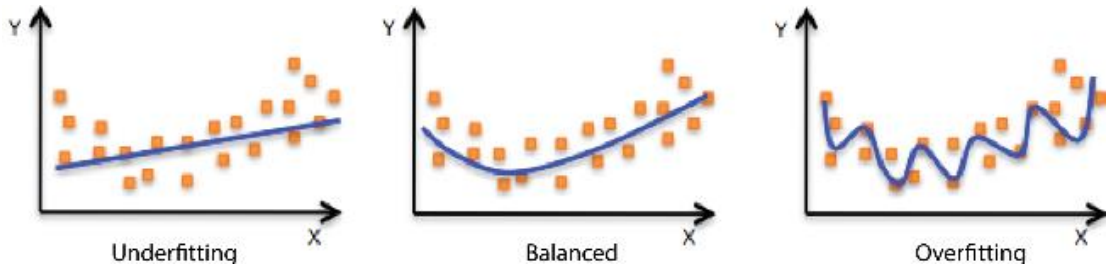


Figure 15 Visualization over- and underfitting of a model[65]

Train your model on more data

In medical imaging, the development of DL networks is often a challenge because highly complex features must be recognized in data with few variations. Variations are needed to ensure generalization. The most straightforward way to reduce overfitting is to add more data with more variability. Because the creation of correctly labeled, large medical datasets often require much time, work, and money, this is a big limitation. To address this issue, data augmentation can be used to artificially enlarge the dataset. Examples of augmentation strategies are horizontal flips, random crops, principal component analysis, the addition of Gaussian noise, scaling, random rotation, and shears. With these strategies, the variability is increased by looking at the same data but from different perspectives. Another rather simple way to improve the training process is to shuffle the training samples, ensuring the successive batches contain different classes. This improves the variation in each batch, making it less sensitive to overfitting. Another method to reduce overfitting is to repeat the whole training process multiple times, each time letting other parts of the dataset be the test set. This is computationally more expensive but ensures that all data is eventually used for training. This is called cross-validation.[66]

Change the complexity of the model: Make your model simpler

Adding more layers to a deep neural network can make the model too complex and more prone to overfit. A way to reduce overfitting is to remove hidden layers and neurons in the fully connected layer.[66]

Stop the learning process

After a certain number of epochs, the validation loss stops decreasing and reaches a plateau. It can be beneficial to reduce the learning rate with a certain factor after the learning reaches this plateau. Reducing the learning rate and stopping the learning process when the validation loss stops decreasing helps to improve the model's performance.[66]

Regularization

Regularization is a technique that prevents a model to become too complex. In other words, it forces the model to become simpler. Examples are L1 and L2 regularization and using dropouts. In regularization, we simply add a term to the loss function that penalizes for large weights. The most common one is 'L2 regularization. In L2 regularization the complexity of the model is

minimized by summing the squares of all weights. Small weights will therefore have little impact and large weights will be more important. Some weights will be close to zero and therefore reduce the impact of some layers. This makes the network less complex and reduces overfitting.[67]

In a publication of Hinton and Srivastava in 2014[68] a method called 'dropout' was proposed to address the problem of overfitting. The idea is to randomly select hidden neurons in each iteration and deactivate them. By deactivating them they do not contribute to the learning step. It slows down the converge of the learning process, but reduces overfitting, and ensures a more robust approach.

3.5 Performance evaluation

To be able to improve the performance of the network and determine the feasibility of the model in clinical practice, appropriate metrics have to be chosen. These are often statistical outputs. During training, it is useful to plot the training and validation loss as well as the accuracy. In DL, commonly used metrics are accuracy, precision, and recall.[43] The accuracy is the percentage of frames that are correctly recognized by the network and is the most obvious metric. This metric works well for a balanced dataset. For an unbalanced dataset, as is in many real-life problems, the precision or recall per class is a better metric. When the most important goal is to minimize the false negatives, the F1-score is a good metric because it gives a better measure of the incorrectly classified samples than accuracy does. To get a good view of the model's performance and what kind of errors it gives, a confusion matrix is created. A confusion matrix is a table that contains the counts of correct assignments per class, and the counts of incorrect assignments per class. The columns correspond to predicted classes. The rows correspond to true classes. A cell in the table at position (r,c) holds the count that class c was assigned whereas the true class was r. For a binary classification problem, the receiver operating characteristic curve (ROC curve) visualizes the performance at a certain trade-off between sensitivity and 1-specificity.[69] The Area Under the ROC curve (AUC) gives a good summary of the performance of the test. An AUC of 0 is the result of a complete inaccurate test and an AUC of 1 indicates a perfectly accurate test. A value of 0.5 means the network cannot discriminate between classes. A higher AUC value means better clinical applicability of the network.[69] All metrics are elaborated below.[70]

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

The percentage of correctly assigned cases with relation to the set of all cases.

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

The percentage of real positive assigned cases with relation to the set of all positive assigned cases.

$$Recall = \frac{TP}{TP+FN} =$$

The percentage of real positive assigned cases with relation to the set of all positive cases. This is also called the sensitivity

$$F1\ score = \frac{precision * recall * 2}{precision + recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} =$$

The harmonic mean of precision and recall

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

The percentage of real negative assigned cases with relation to the set of all negative cases

Transfer Learning

Training a DNN from scratch is very hard because the model contains many unknowns. Much labeled data and high computational power is needed to optimize such models. Transfer learning approaches give the possibility to get around this problem and have gained huge popularity in medical imaging.[56] A CNN architecture that was trained on a large natural image dataset such as ImageNet is reused, together with corresponding pre-trained weights.[71] ImageNet is an image database containing millions of annotated images and is widely used as a benchmark in image classification and object detection. It would be ideal to have a dataset of such magnitude for each domain, but that is difficult considering the time, effort, and availability of specific data. In transfer learning, we use the same model architecture that was used to train on a large dataset without the last layer. The pretrained weights of this model are transferred and used to extract features of the data for the particular task. *(Figure 16)* It is also an option to retrain some of the layers and freeze (fixate) the rest of the layers. Transfer learning speeds up the process of training. Some popular available models include VGG-16, Inception V3 and ResNet-50.[71]

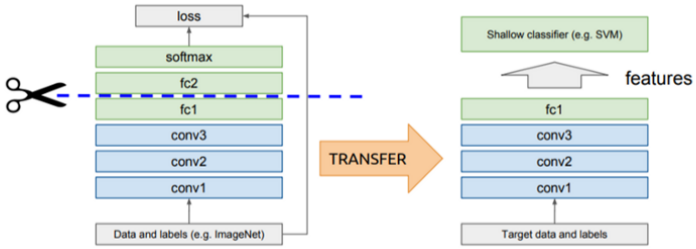


Figure 16 Transfer learning with pre-trained DL models as feature extractors [56]

One of the most groundbreaking developments in DL is the Residual Network (ResNet), proposed by Kaiming He et al in 2015.[72] It is by far the state-of-the-art image classification neural network. To avoid overfitting, it has become common practice to go deeper and deeper in CNN architecture. To update the weights, we need backpropagation. When the network is deeper, this makes it harder to train because of the vanishing gradient problem. While reaching the earlier layers in backpropagation this will result in extremely small values. A way to solve this problem is the residual learning framework. This framework makes it easier to optimize. It uses shortcut connections *(Figure 17)*. Instead of using only stacked convolutional layers, the original input is also added to the output of a convolutional block. Because some layers are skipped, the value will not be so small when reaching the earlier layers. This way these convolutional blocks can improve

the output of the previous block, instead of directly having to fit a desired mapping.[72] The ResNet comes in a different number of layers, for example, Resnet18 and Resnet50. Resnet18 is a version that uses 18 neural network layers.

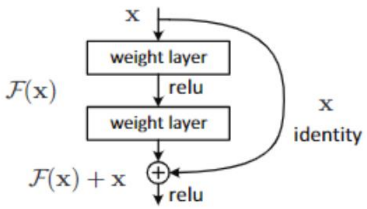


Figure 17 Residual learning: a building block[72]

4. Methods

This study aims to predict intraoperative difficulty by classifying and evaluating intraoperative images in the first phase of the surgery. In this section the methods are discussed. First, the selection of data is presented. Following, an explanation of the start of the surgery is provided. Thirdly, the label definitions based on the Nassar scale are explained and, lastly, an explanation about the experimental design and training is described.

4.1 Data selection

A total of 93 LC's recorded in the MMC in Amersfoort between 01-01-2018 to 10-10-2021 were selected to create the dataset. The patients underwent a classical LC, whose videos were recorded. The videos were already collected in a study protocol for previous research carried out by Maartje Gerkema[73], which was approved by the research committee in 2020. A signed consent form was not needed. This protocol permits the use of videos in an anonymous way for similar studies in the same center. All videos are strictly anonymous. The videos were already downloaded from the Electronic Health Record. A total of 257 videos was available at the start of this research. The first step was to exclude the videos that did not contain the first phase of the surgery. Some videos had to be excluded because of bad quality. Due to lack of time, not all suitable videos were used. The next step was to cut the videos into sub videos containing only the first phase of the surgery because the aim is to classify difficulty at this moment. An explanation of how this was done is elaborated below. The sub videos were converted to frames at one frame per second using a python script. This resulted in 93 folders, one for each video, containing the frames. This resulted in a total of 28.753 frames *(Figure 24)*. The frames were labeled and checked manually. The labels were provided by the researcher, supervised by a junior surgeon and an expert surgeon.

4.2 Start of the surgery

To predict surgical difficulty at the beginning of the surgery, only this part should be used to train the network. The prediction should be made when the condition of the gallbladder can be examined. This is the moment when the gallbladder is retracted. Before this moment, abdominal access has to be established by inserting the trocars and tools. Afterwards, the abdomen is first inspected and sometimes some abdominal adhesions have to be removed to clear the way for the tools to reach the gallbladder. Here we define the start of the surgery as the moment when the fundus of the gallbladder is grasped. Therefore, this part of the video was used to train the classification model. Timestamps were noted for each video to cut the videos into sub videos. *(Table 2)*The Nassar scale scores on the presentation of the Gallbladder, Adhesions, and Cystic Pedicle. In easy cases, the gallbladder is visible from the start, there are little to no adhesions and the surgeon can immediately proceed to the next phase: dissection of the hepatocystic triangle. For dissection, the surgeon zooms in on the cystic pedicle, where the ductus and artery are present. Because from this point the dissection starts, the second timestamp should be noted here. In many cases, adhesions first have to be removed to clear the way and visualize the cystic pedicle. In some cases, the gallbladder is partly or completely obscured and can therefore be scored at the first 'sight' on the gallbladder. This is also true for the cystic pedicle when there are many adhesions present.

Table 2 Timestamps definitions

Timestamp	Moment (description)
T1	Fundus of the gallbladder is grasped/Start removing adhesions
T2	All adhesions are removed/Start of dissection (often with a hook)

For easy cases, the timestamps are easily set because there is a clear transition from the first phase to the next phase. In difficult cases, for example with severe cholecystitis or dense adhesions, the transition is more gradual, and therefore it is more difficult to determine the exact moment. In severe cholecystitis cases, it is sometimes hard to see when the opening of the peritoneum starts when the gallbladder is completely obscured. This starting phase of the surgery in which preparations are made to be able to start dissection of the hepatocystic triangle, from now forth will be called 'phase 1'. Examples of the start and end of phase 1 are given in the figures below. *(Figure 18-21)*

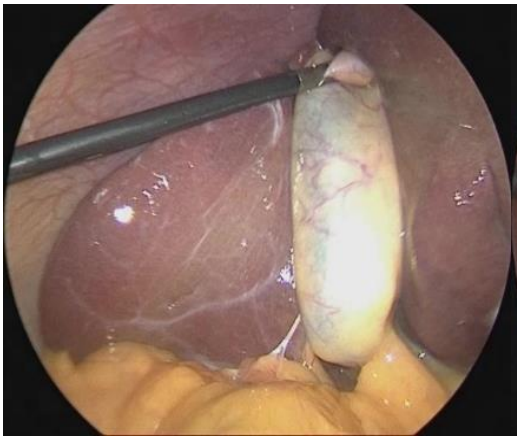


Figure 18, T1: Clear view on gallbladder

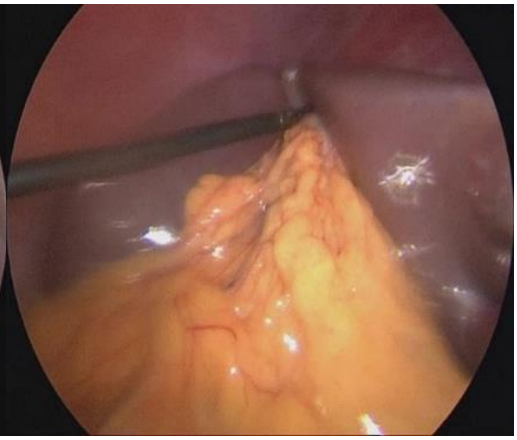


Figure 19, T1: Gallbladder obscured by adhesions

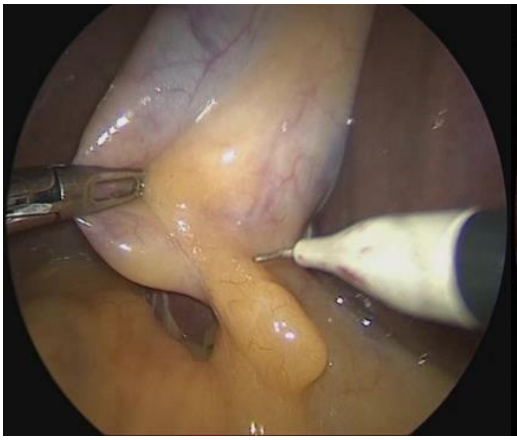


Figure 20, T2: Transition easy recognizable

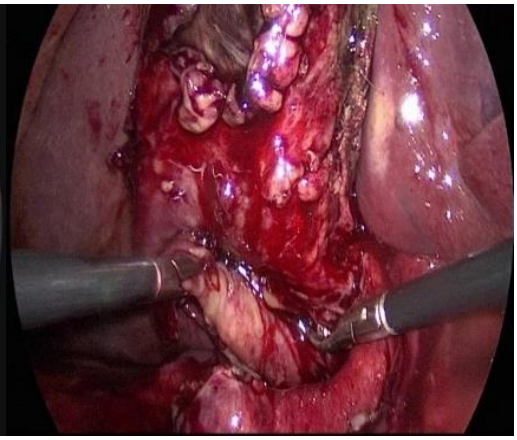


Figure 21, T2: Transition is unclear

All 93 were cut into sub videos using a python script. The duration of the sub videos varied from 24 seconds to 26 minutes and 40 seconds. To determine how the difficulty score classified in phase 1 correlates with the total surgery time, the total duration of each video was also noted. A part of the 93 videos included in the dataset was incomplete because the beginning was missing. On average, establishing access takes 141s (231s). This was calculated by taking the mean of all videos that do contain this part. The average time for establishing access was added to the total duration for the videos in which this part was missing. The mean and median of the durations are given in Table 3. The distribution is given in Figures 22 and 23. The time reserved for a LC is estimated at 45 minutes in Meander Medical Center.

Table 3 Duration of the videos

Phase	Duration (s)	median
Adhesions lysis	309 (401)	181
Total surgery	2328 (1205)	2214

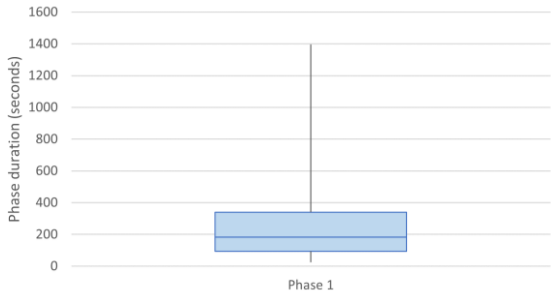


Figure 22 Distribution of duration of phase 1

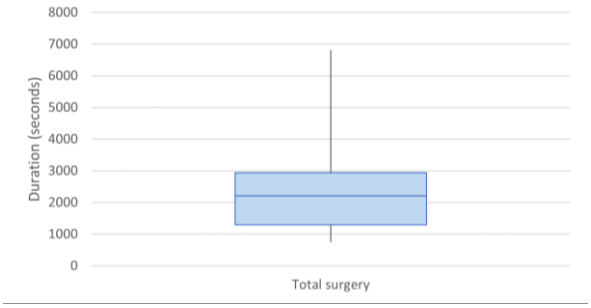


Figure 23 Distribution of duration of total surgery

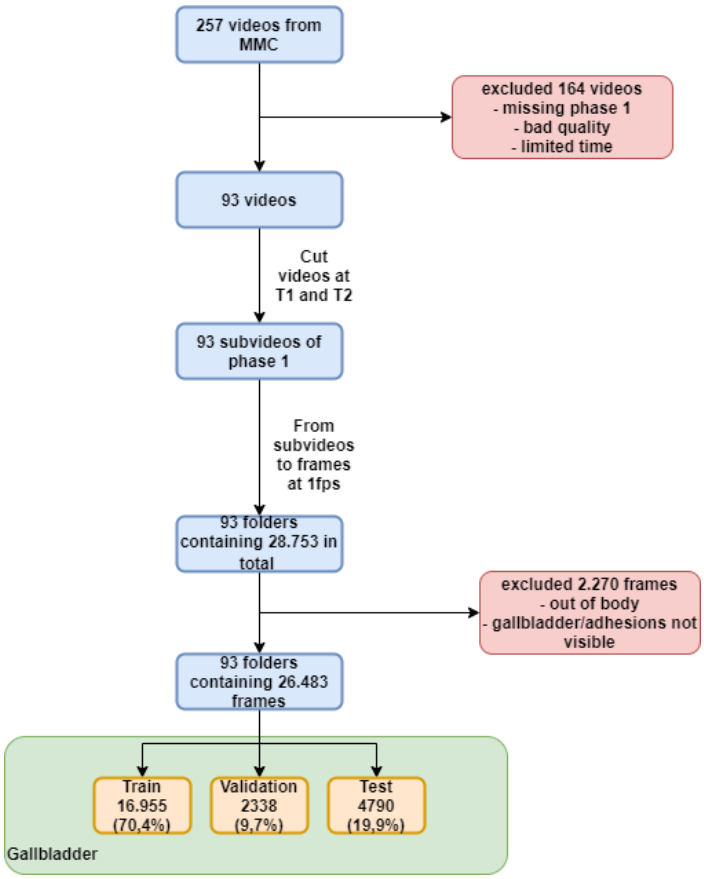


Figure 24 Dataset creation

4.3From Nassar to label definitions

The goal is to not only predict the overall grade for difficulty but also give the reason why it is difficult. The Nassar scale was used as a starting point to define the labels. This scale grades operative findings of the gallbladder, cystic pedicle, and associated adhesions from grades 1 to 4. Therefore, at the start of the labeling process, each laparoscopic image was given three separate labels for gallbladder, adhesions, and cystic Pedicle, ranging from grade 1 to 4. During the labeling process, some adjustments were made to the original definitions of the grades. This was done to make the results more clinically relevant.

From 4 to 3 grades

To make the learning task a little less complicated, it was decided to decrease the Nassar scale to three grades instead of four. After examing a random set of videos, it was noticed that grades three and four as described in the Nassar scale, often resemble each other. When the gallbladder is completely obscured, it takes little time to visualize some part of the gallbladder, therefore falling into grade three already. For this reason, grades 3 and 4 are taken together as grade 3.

Adjustments to Adhesions grade definitions

According to the Nassar scale, adhesions are described as ‘simple up to the neck’ (grade 1), ‘simple up to the body’ (grade 2), ‘dense up to the fundus’ (grade 3), and ‘dense, wrapping the

gallbladder' (grade 4). They are divided based on the severity and location. During the labeling process, it was noted that some adhesions were simple but up to the fundus, or really dense but only up to the body. In consultation with an experienced surgeon, it was decided to describe the grade definitions according to the difficulty to remove them. The location of the adhesions is less relevant when looking at difficulty. Therefore, a distinction is made between 'anatomical' (simple) adhesions and 'pathological' (dense) adhesions. In a few videos in the dataset, this resulted in a different grade, because there were pathological adhesions up to the body, or simple adhesions up to the fundus of the gallbladder.

Cystic Pedicle

To be able to determine the grade for 'cystic pedicle', the adhesions should be removed, and the camera should be zoomed in for a proper view. Because the endpoint of phase 1 is at the start of the dissection, in most sub videos this results in only seeing the cystic pedicle for a few seconds. The presentation of the cystic pedicle is more important in the next phase. This resulted in very few frames labeled for the cystic pedicle.

Furthermore, after labeling the first 40 videos, it was noted that almost all videos in which 'gallbladder' and 'adhesions' were labeled with grade 1, the 'cystic pedicle' had to be labeled with grade 2. Grade 2 is defined as 'fat laden', according to the Nassar scale. After 40 videos, it was noted that in almost all cases the cystic pedicle is mildly fat-laden. This would result in all these videos getting an overall difficulty grade of 2. When 'gallbladder' and/or 'adhesions' were labeled with grade 3, the label for 'cystic pedicle' was always also grade 3 or could not yet be determined because it was not visible. Therefore, the label for Cystic Pedicle does not add any information in these cases.

Mainly because the presentation of the cystic pedicle is more important in phase 2 and these images are not included in the dataset, it was decided to leave out this label in this dataset.

Grade definitions

Taking all the above-mentioned into account, each laparoscopic image was labeled for 'gallbladder', and 'adhesions', grade 1-3. In the sub videos, there were also moments when neither the gallbladder nor adhesions were clearly visible and could therefore not be given a label. In these frames, the gallbladder and/or adhesions were given the label '0'. Only images from within the body were given a label 0-4. All images from outside the body are not included in the dataset and were given the label 'excl', to be removed later on. When the gallbladder is completely obscured with adhesions, 'gallbladder' is labeled with '0'. If the gallbladder was disguised by tools, smoke, or other anatomical tissue, the frame was also labeled with '0'. Only when the gallbladder is visible, it is given a label (1-3). On average 91,57% of all frames contained a good view of the gallbladder and/or adhesions. The grade definitions are given in Table 4. Figure 25 shows an example of the proposed dataset with the assigned labels.

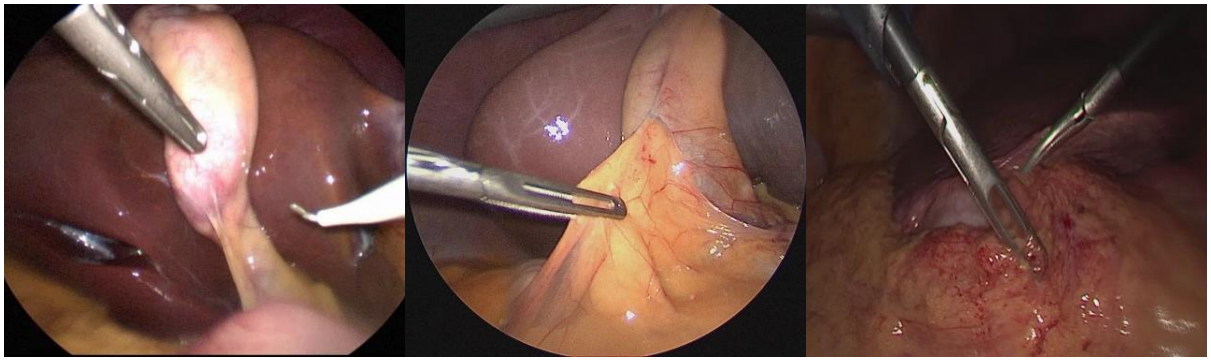
Table 4 Label definitions

Label	Gallbladder	Adhesions
1	Floppy, thin, gray/pink, fat-laden	No adhesions or simple up to the neck/Hartmanns pouch
2	Mucocele, hydropic or packed with big stones	Simple/anatomical adhesions
3	Cholecystitis, gangrene, empyema, mass	Pathological/dense adhesions/completely obscured

Gallbladder

Gallbladder: Grade 1 *Gallbladder: Grade 2* *Gallbladder: Grade 3*

Adhesions



Adhesions: Grade 1 *Adhesions: Grade 2* *Adhesions: Grade 3*

Figure 25 Examples of the dataset

4.4 Dataset

For the label 'gallbladder', each video only contained one grade (1, 2, or 3). For 'adhesions', some videos contained two different labels, because for example after the removal of some adhesions, the grade went from 2 to 1. The highest grade for either gallbladder or adhesions defines the overall grade for difficulty. Table 5 shows an overview of the number of videos that contain grade 1, 2, and 3.

Table 5 Number of videos for each grade

	Grade 1	Grade 2	Grade 3	Total
Gallbladder	58	15	20	93
Adhesions	56	22	15	93
Overall Grade	44	23	26	93

Looking at the distribution of videos and the grades, it would seem like there is more data available for grade 1 than grade 3. There are indeed more easy than difficult LC's in the dataset, and therefore there is more variability in the data in grade 1 than in grade 3. But videos labeled with Grade 3 are longer in duration *(Table 6)*, therefore resulting in more frames *(Table 7)*. The total number of frames is not the same for gallbladder and adhesions because in some frames either the gallbladder or adhesions label was '0'. It has to be mentioned that this dataset is not necessarily an accurate representation of the real-life distribution in easy or difficult surgeries. During the labeling process, more cholecystitis cases were added to the dataset on purpose.

Table 6 Average duration (s) per grade

	Grade 1	Grade 2	Grade 3
Adhesions lysis	113 (72)	332 (199)	622 (615)
Gallbladder	150 (108)	473 (331)	648 (666)
Adhesions	164 (171)	368 (248)	764 (734)
Total surgery	1857 (934)	2317 (1052)	3123 (1352)
Gallbladder	1864 (896)	2831 (961)	3300 (1431)
Adhesions	2092 (984)	2302 (1056)	3295 (1745)

Table 7 Number of frames in dataset per Grade

	Grade 1	Grade 2	Grade 3	Total
Gallbladder	7841	6256	9986	24.083
	(32,56%)	(25,98%)	(41,46%)	(100%)
Adhesions	6583	6652	10.149	23.384
	(28,15%)	(28,45%)	(43,40%)	(100%)
Overall Grade	4434	7076	14973	26.483
	(16,74%)	(26,72%)	(56,54%)	(100%)

To train the network, the dataset was divided into a train, validation, and test set. Because of the way the data was structured, it was not possible to use a 'random split' in the dataset. The dataset consisted of 93 folders containing frames, one for each video. Frames from one video should only be used in either the train, validation, or test set. A random split would result in unequal distributions in the different sets. Therefore, the data was split manually to make sets with the same distributions. The goal was to split the data in such a way that the train set contained 70%, the validation set 10% and the test set 20%. The results of the manual splits are given in Figure 26.

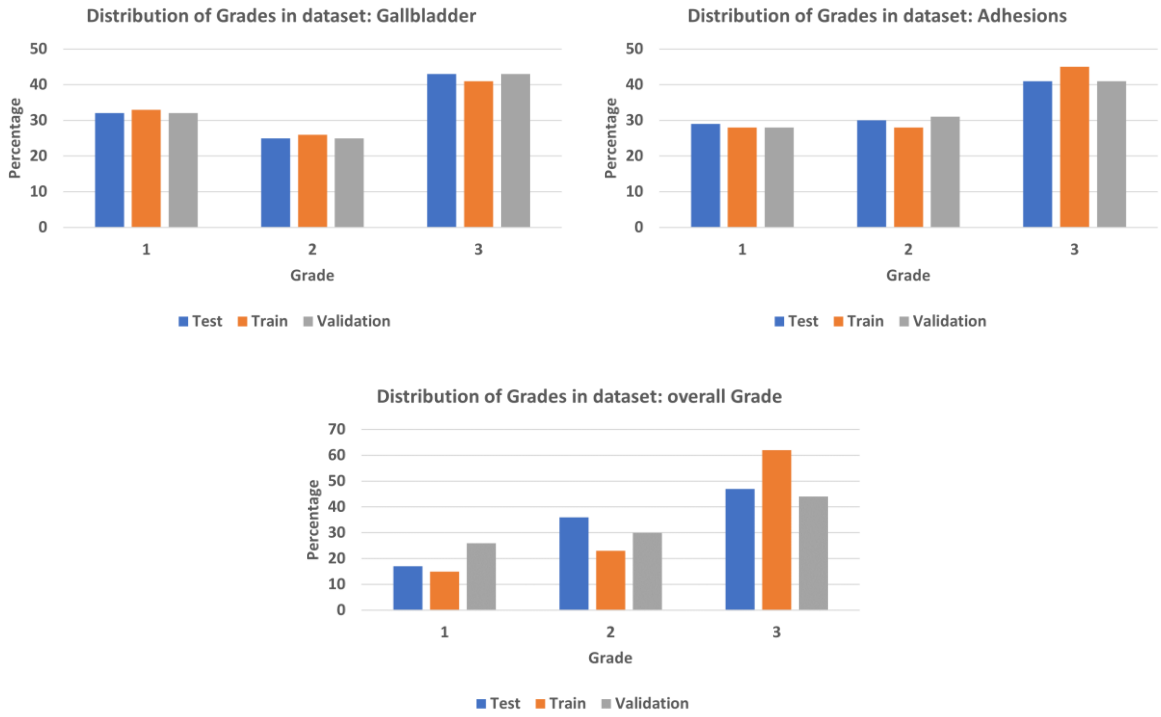


Figure 26 Distribution of grades in the train, validation, and test sets

4.5 Experimental setup

Several open-source frameworks and libraries are available that contain the complex mathematical functions and training algorithms required for developing DL models. For this research, the python library 'Pytorch' was used in Python 3.8 to train, test, and evaluate the networks.

The ResNet architecture was used as a backbone for the model. Both ResNet50 and ResNet18 were used for training. The network and the pre-trained weights are publicly available. The labeled frames serve as the input to the network. All frames were stored in PNG format in 93 separate folders. The images with the label 'excl' were excluded because these were out of body images. Also, all images with the label '0' were removed because the gallbladder or adhesions were not visible in these images. First, all images were resized to 256x256 and then center cropped to 224x224. This is the default input size that a ResNet takes. Then all images were normalized. Therefore, the mean and standard deviation was calculated for each color channel in the images in the dataset.

Depending on the type of classification task, the output layer of the model was adjusted. For binary classification, the output layer was set to 1, followed by a sigmoid activation function. The loss function that was used was the binary cross-entropy loss function. The sigmoid activation function squashes the output in the range (0,1). For multiclass classification, the output layer was set to 3, followed by a softmax activation function. Here, the cross-entropy loss was used. The softmax activation function squashes the output in the range (0,1) and all the resulting elements add up to 1. Figure 27 gives an overview of the networks.

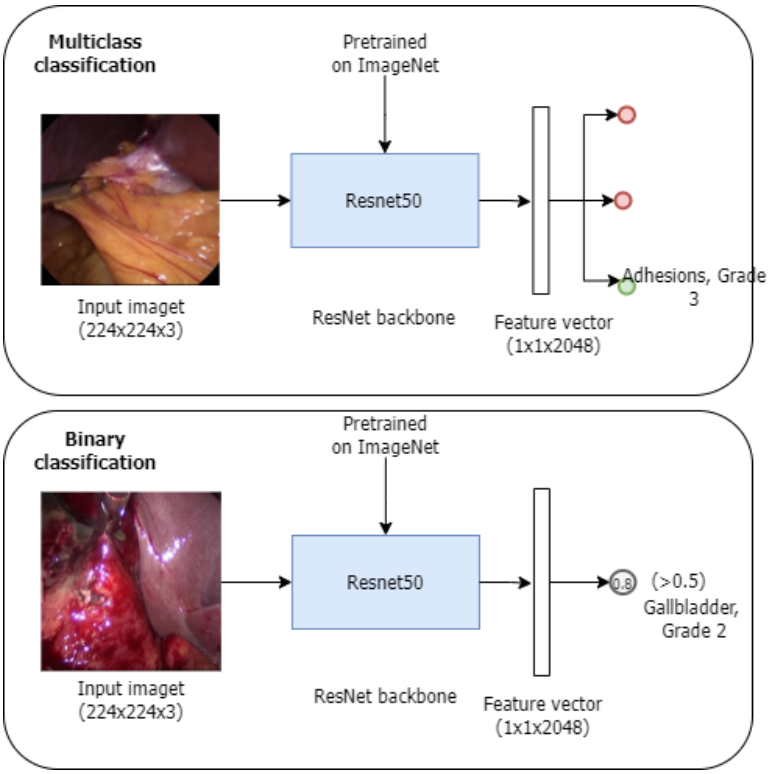


Figure 27 Network structure

Because the classes are not equally distributed, the network may tend to get biased towards the majority class. With class weighting, we can consider this uneven distribution of the classes. During training, we prioritize the minority class by giving a higher-class weight. The penalty will be higher if the minority class is misclassified. Using the class frequencies, a weight was given to each class in the calculation of the loss function. The calculation of the class weight is as follows:

$$class\ weight = 1 - \frac{number\ of\ samples\ of\ that\ class}{total\ number\ of\ all\ samples}$$

4.6 Training

Separate 3-class (or 3-grade) networks were trained for the classification of the gallbladder and the adhesions. In addition, to investigate if it is possible to recognize the easy LC's or for example cholecystitis cases, binary classification networks were trained. For these experiments, at first, the labels 2 and 3 were taken together and classified against label 1. In other words, 'easy' versus 'not-easy' cases. Three separate networks were trained. One with the 'Gallbladder' labels, one with the 'Adhesions' labels, and one with the 'Overall Grade' labels. The 'Overall Grade' labels

contain the maximum value of Gallbladder and Adhesions. For the next experiments, the labels 1 and 2 were taken together and classified against label 3. In other words, 'difficult' versus 'not-difficult' cases. Three separate networks were trained. One with the 'Gallbladder' labels, one with the 'Adhesions' labels, and one with the 'Overall Grade' labels. The 'Overall Grade' labels contain the maximum value of Gallbladder and Adhesions. Table 8 provides an overview of all classification tasks.

Table 8 Overview different classification tasks

Classification task	Gallbladder	Adhesions	Overall grade
	labels		
3-grades	1-2-3	1-2-3	
2-grades	1 – 2&3	1 – 2&3	1 – 2&3
2-grades	1&2 – 3	1&2 - 3	1&2 - 3

An Ubuntu PC with a GPU (NVidea) was used for training. The Docker platform was used to create a Docker image, providing a Python 3 environment and the selected packages. To train the network for 30 epochs with the entire train set took approximately 2.5 hours on the GPU.

During training, the loss and accuracy of both the training and validation set are monitored. The version of the model with the highest validation accuracy is saved during training. The proposed classification task is assessed by measuring the accuracy, precision, recall, and the F1-score. In addition, the confusion matrix and the ROC curve are calculated.

For the results to be clinically applicable, difficulty prediction has to be done after phase 1 of the surgery. Therefore, the percentage of correctly identified frames per test video was calculated for the best performing networks. If the majority of the frames were correctly identified, the video can be classified correctly.

Hyperparameter optimization

To be able to track the training process, evaluate the performance of the network, and visualize the results, the machine learning platform 'Weights & Biases' was used.[74] In the first experiments, the Adam optimizer was compared with the SGD optimizer. SGD showed evident better results, so we continued using SGD. 'Weights and Biases' allows for so-called 'sweeps', to optimize the hyperparameters. A sweep is a combination of a strategy for trying out hyperparameter values and code that evaluates them. That strategy could be as simple as trying every option or could be something complicated and optimal like combining bayesian and early stopping. Trying every single option can be computationally very costly. An easy but surprisingly effective method is 'random search'. In every sweep, a random value in a certain distribution is picked for each hyperparameter. The sweep that was configured for all experiments is shown in Table 9. The defined method was 'random search', meaning that every new combination is set at random according to the provided distribution.

Table 9 Sweep configuration

Method	Random
Metric	Goal: minimize loss
Parameters:	Value(s):
Architecture	Resnet18
	Resnet50
Batch size	16, 32
Classes	3
Dropout	0, 0.2, 0.4
Epochs	50
Learning rate	Distribution:uniform Max: 0.0001 Min: 1e-06
Loss function	BCEWithLogitsLoss
Optimizer	SGD
Weight decay	0, 0.0001

Prevent overfitting

After training all networks described in Table 9, we continued with the best performing network to improve the result. A big issue during training is overfitting on the train set. To reduce overfitting, different experiments were conducted. The network was trained again with adjustments such as data augmentation, early stopping, using dropouts, L2 regularization, and reducing the learning rate when it reaches a plateau.

5. Results

This chapter shows the results of the conducted experiments. First, the results for the multiclass classification networks are presented, followed by the results of the binary classification networks. In part 5.3 the number of frames that are correctly classified per video is visualized for the best performing networks. After these experiments, we proceeded with the best results from the binary gallbladder classification network.

5.1 Multiclass classification

Two separate networks were trained. One with the 'gallbladder' labels one with the 'adhesions' labels. The sweep configuration can be found in Table 9 in chapter 3. Figure 28 gives an example of how all hyperparameters configurations result in a certain validation accuracy. In this particular sweep, in which the network was trained for the gallbladder grade classification, the highest validation accuracy was reached by using a Resnet18 as a backbone network, with a batch size of 32, a learning rate of 0.00003346, a dropout of 0.4, and weight decay set at 0.0001. From all training sessions in the sweeps, the ones that resulted in the highest test accuracy are presented in the next sections. The networks that were trained with a batch size of 32 needed half the number of training steps than those with a batch size of 16. In the section below, the gallbladder 3- class classification and adhesions 3-class classification results are presented.

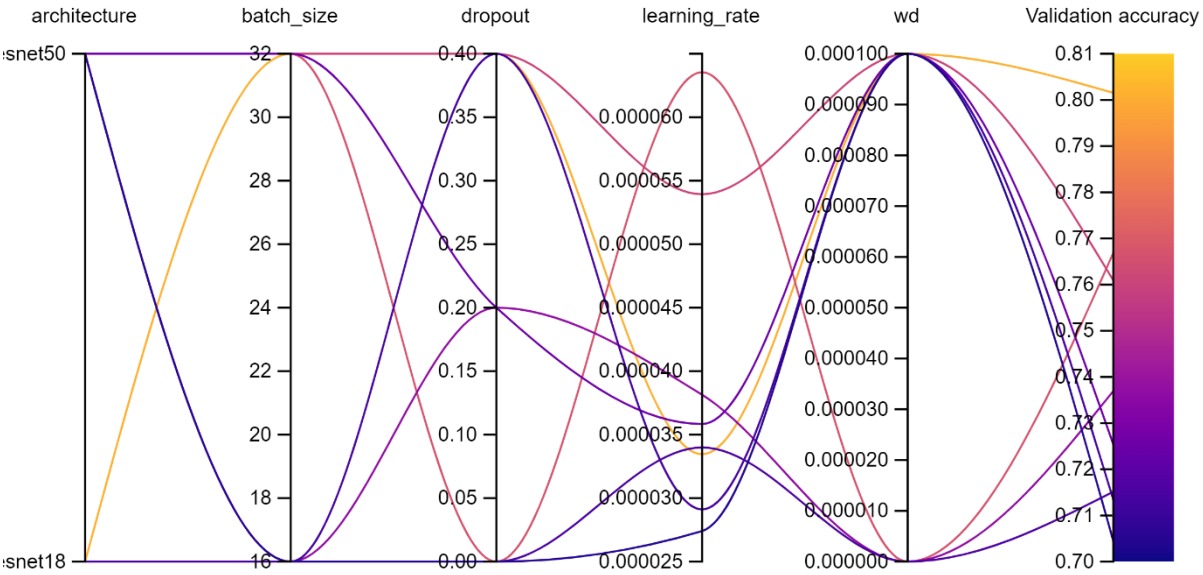


Figure 28 Hyperparameter sweep

Gallbladder 3-class classification

Figure 29 shows the accuracy and loss during training of the gallbladder 3-grade network. It is seen that the network overfits on the train set after a few epochs. The hyperparameter configuration used in this training session is given in Table 11. The confusion matrix of the test set is given in Figure 30. Table 10 shows the test results of the trained network.

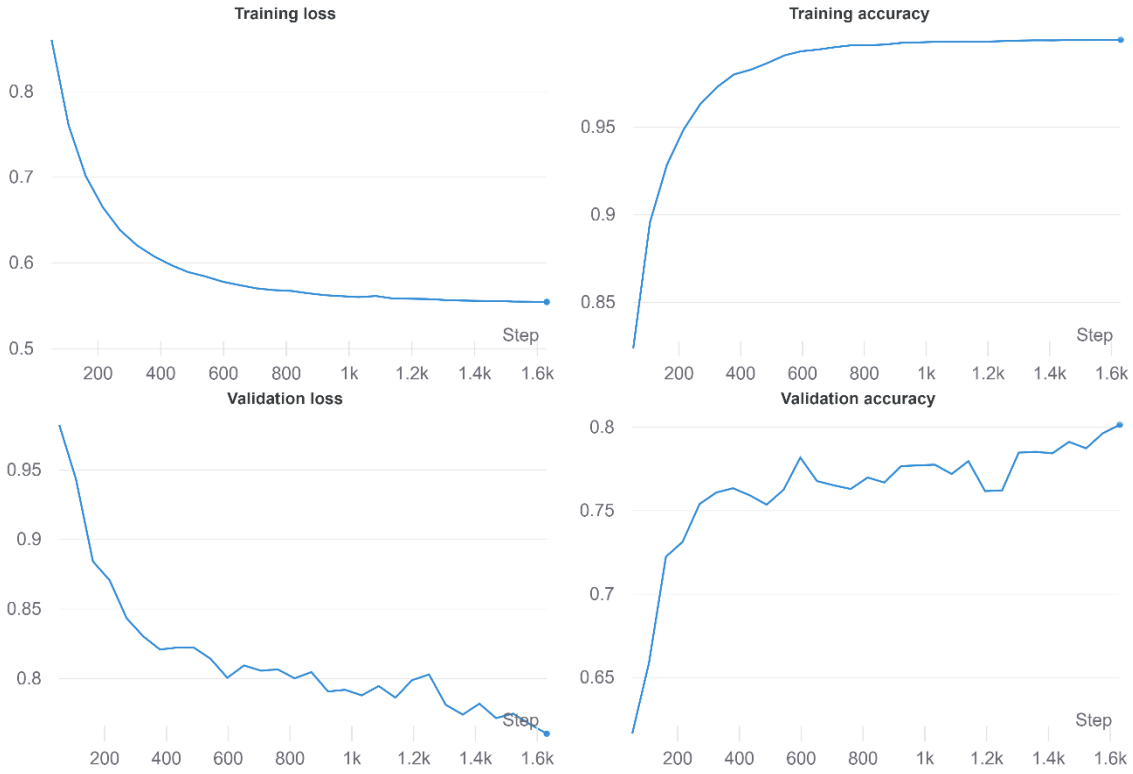


Figure 29 Gallbladder 3-grade, best result from sweep: accuracy and loss during training

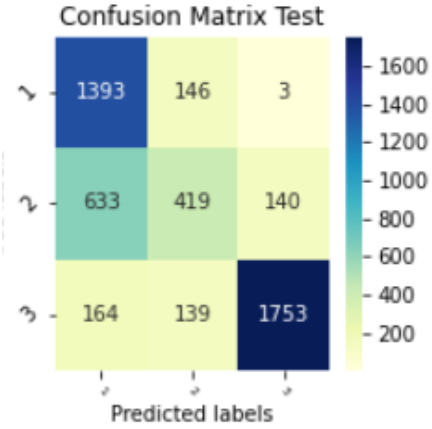


Figure 30 Confusion matrix

Adhesions 3-class classification

Figure 31 shows the accuracy and loss during training of the adhesions 3-grade network. It is seen that the network overfits on the train set after a few epochs. The validation accuracy does not increase effectively. The hyperparameter configuration used in this training session is given in Table 11. The confusion matrix of the test set is given in Figure 32. It can be seen that many test frames have predicted label 1 where it should be 2 or 3. Table 10 shows the test results of the trained network.

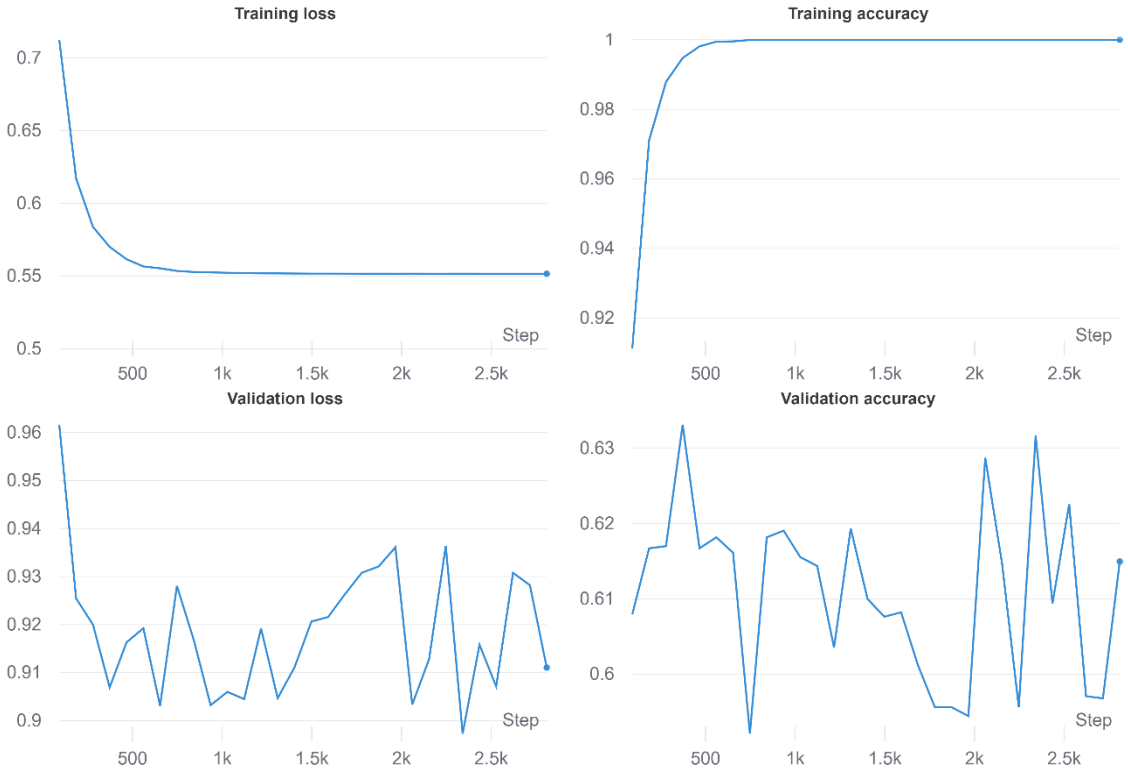


Figure 31 Adhesions 3-grade, best result from sweep: accuracy and loss during training

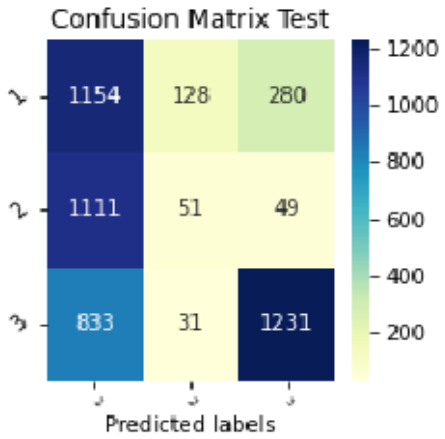


Figure 32 Confusion matrix

Table 10 Evaluation Multiclass Networks

	Grade	Accuracy	Precision	Recall	F1
Gallbladder	1	0.74	0.62	0.9	0.68
	2		0.59	0.33	
	3		0.93	0.84	
Adhesions	1	0.5	0.38	0.77	0.42
	2		0.26	0.04	
	3		0.81	0.58	
	Unweighted average				
	Gallbladder		0.71	0.69	
	Adhesions		0.50	0.56	

The test results are presented in Table 10. The network that was trained to classify gallbladder performs better (accuracy 74%) than the network trained to classify adhesions (50%). The gallbladder classification network has a high sensitivity (or recall) for grade 1 and grade 3, but not for grade 2 (33%).

Table 11 Hyperparameter configuration of the best performing networks

	Gallbladder	Adhesions
Architecture	Resnet18	Resnet50
Classes	3	3
Epochs	30	30
Batch size	32	16
Learning rate	0.00003346	0.00009184
Dropout	0.4	0.2
Weight decay	0.0001	0.0001

5.2 Binary classification

In this section the training results of six binary networks are presented. At first, the frames with labels 2 and 3 were taken together and distinguished from the frames with label 1 ('easy' versus 'not-easy). This was done for gallbladder, adhesions, and overall grade. For the last three networks, the frames with labels 1 and 2 were taken together and distinguished from the frames with label 3 ('difficult' versus 'not-difficult'). The best training results of the hyperparameter sweeps are visualized in accuracy and loss plots and are presented below. First, gallbladder 2-grade classification, followed by adhesions 2- grade classification and last overall 2-grade classification. The sweep configuration can be found in Table 9 in chapter 3. The networks that were trained with a batch size of 32 needed half the number of training steps as those with a batch size of 16.

Gallbladder 2-class classification 'easy' versus 'not-easy'

The training results of the binary gallbladder classification network are shown in Figure 34. The labels 1 and 2-3 were used to train the network to recognize gallbladder grade 1, or in other words, to recognize a normal gallbladder. As can be seen in Figure 33, the network overfits on the train set after a few epochs. The test results are shown in the confusion matrix and ROC curve in Figure 34 and 35. Table 12 shows the test results of the trained network. The combination of hyperparameters that was used for this training is given in table 13.

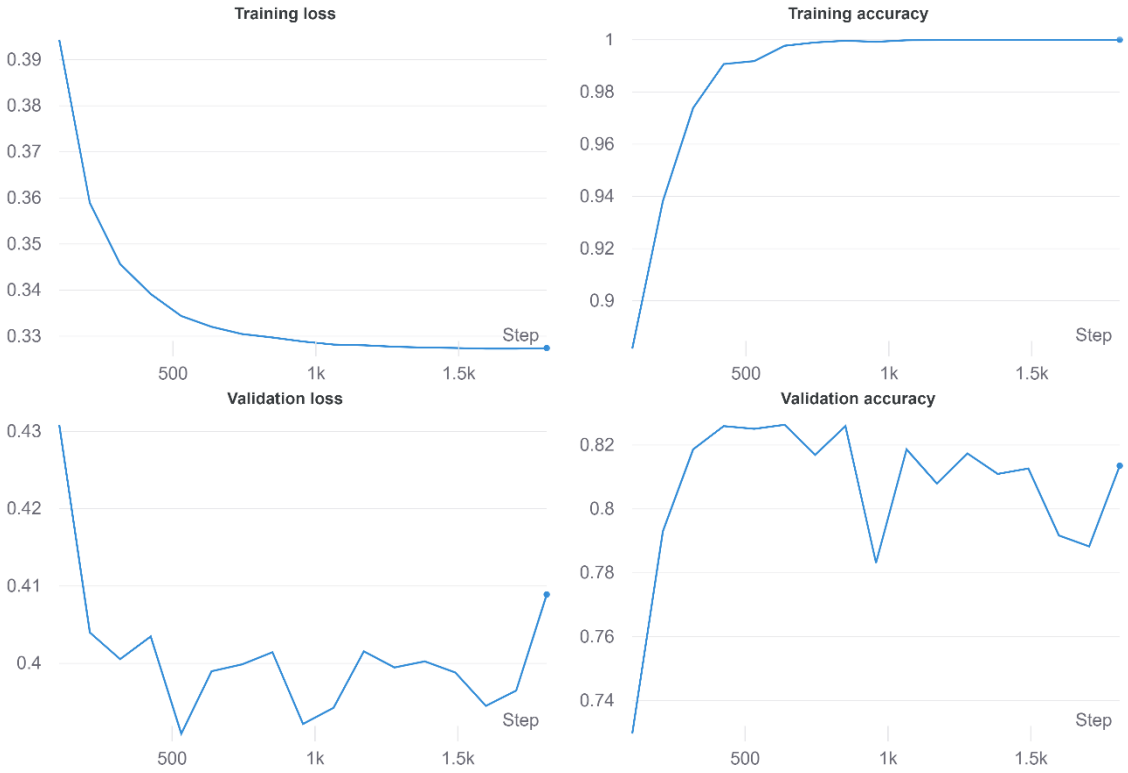


Figure 33 Gallbladder 2-grade (1 versus not-1), best result from sweep: accuracy and loss during training

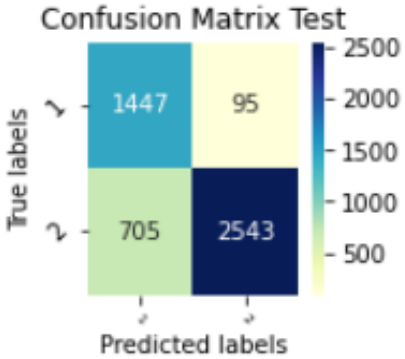


Figure 34 Confusion matrix

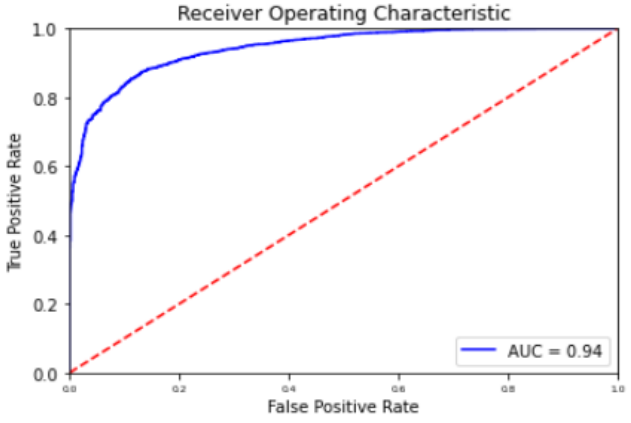


Figure 35 ROC curve

Adhesions 2-class classification 'easy' versus 'not-easy'

The training results of the binary gallbladder classification network are shown in Figure 36. The labels 1 and 2-3 were used to train the network to recognize gallbladder grade 1, or in other words, to recognize a normal gallbladder. As can be seen the network overfits on the train set after a few epochs. The test results are shown in the confusion matrix and ROC curve in Figure 37 and 38. Table 12 shows the test results of the trained network. The combination of hyperparameters that was used for this training is given in table 13.

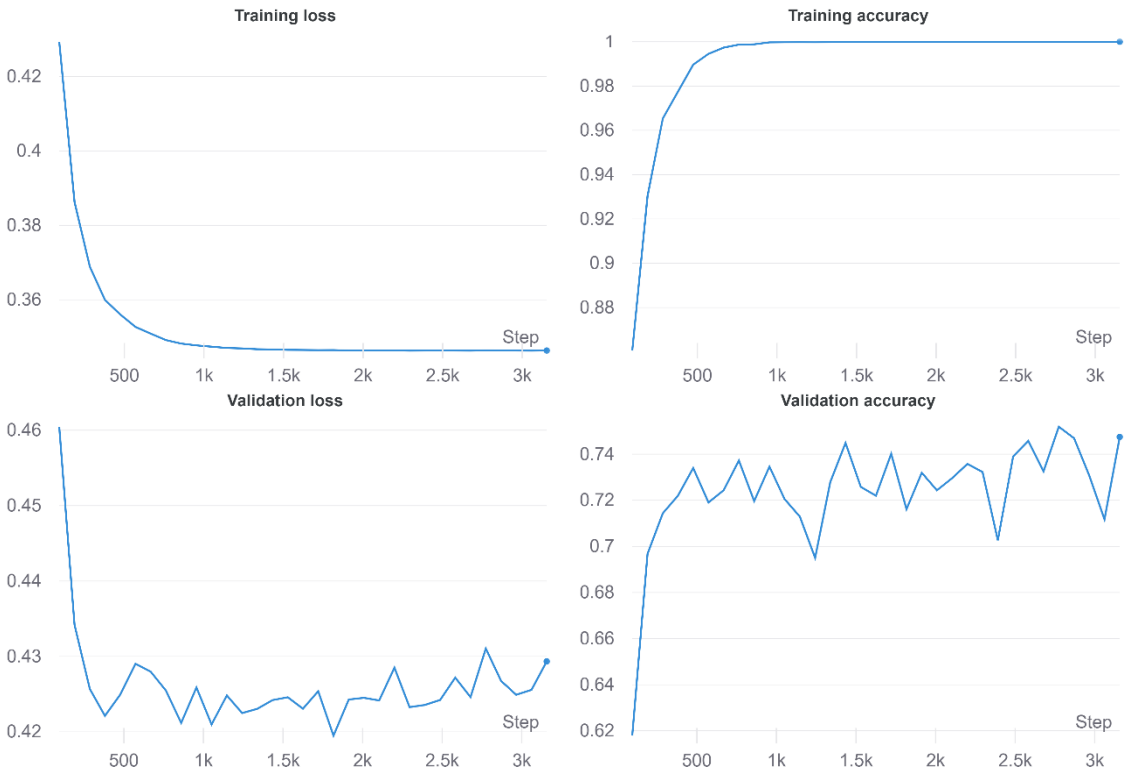


Figure 36 Adhesions 2-grade (1 versus not-1), best result from sweep: accuracy and loss during training

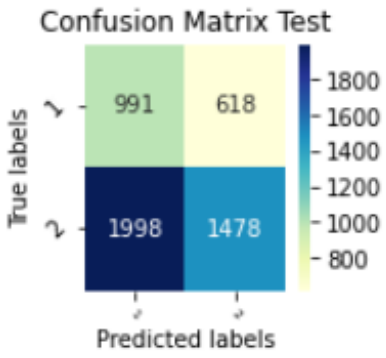


Figure 37 Confusion matrix

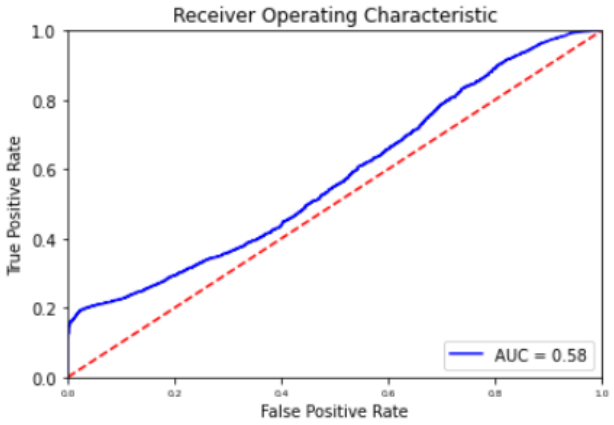


Figure 38 ROC curve

Overall Grade 2-class classification 'easy' versus 'not-easy'

The training results of the binary gallbladder classification network are shown in Figure 39. The labels 1 and 2-3 were used to train the network to recognize gallbladder grade 1, or in other words, to recognize a normal gallbladder. As can be seen the network overfits on the train set after a few epochs. The test results are shown in the confusion matrix and ROC curve in Figure 40 and 41. Table 12 shows the test results of the trained network. The combination of hyperparameters that was used for this training is given in table 13.

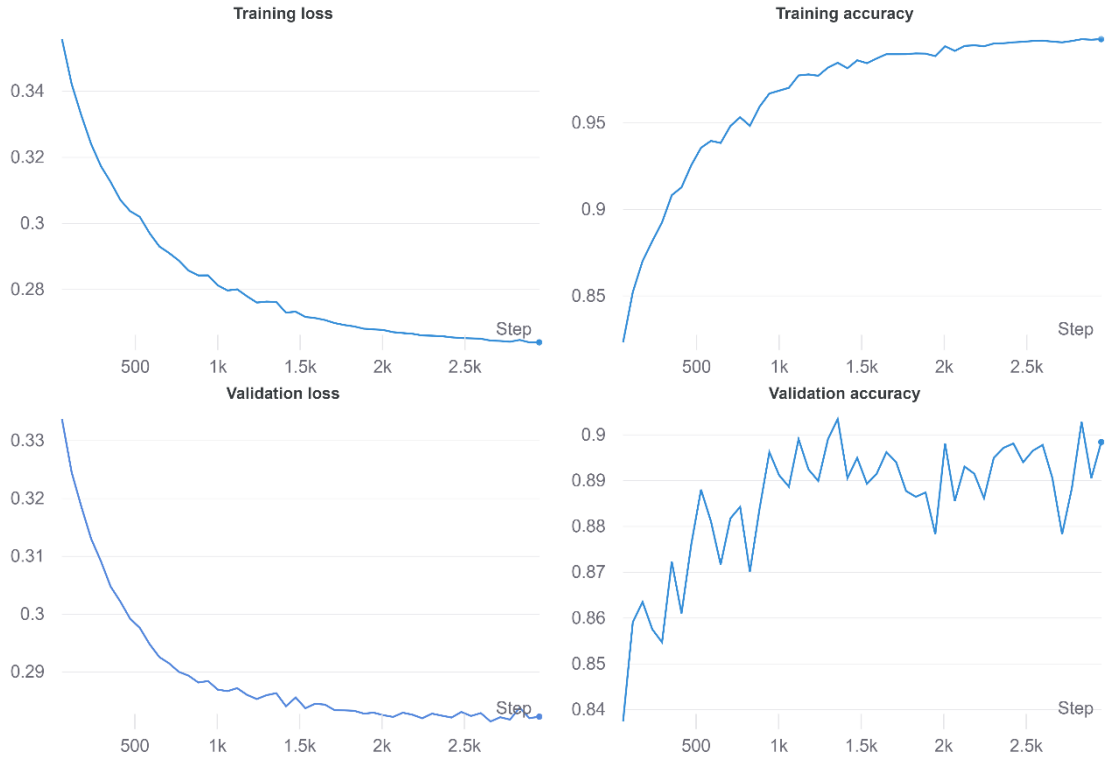


Figure 39 Overall Grade 2-grade (1 versus not-1), best result from sweep: accuracy and loss during training

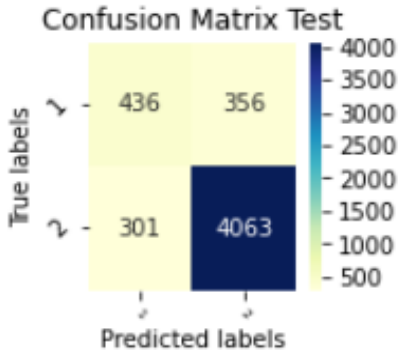


Figure 40 Confusion matrix

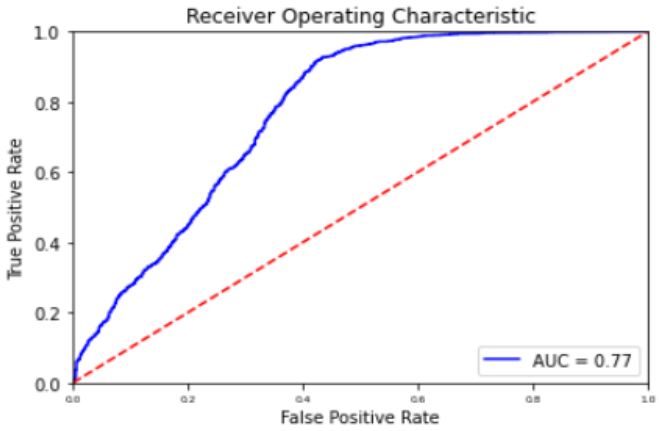


Figure 41 ROC curve

Table 12 Evaluation Binary classification Networks 1 vs 2&3

	Accuracy	Precision	Recall	F1
G-binary	0.83	0.96	0.78	0.86
A-binary	0.49	0.71	0.43	0.53
GA-binary	0.87	0.92	0.93	0.93

As can be seen, the network that was trained on the overall grade gives the highest accuracy on the test set. This means that the network is capable of recognizing the easy cases (gallbladder label 1 and adhesions label 1) with an accuracy of 87%. The test results of the network trained for the classification of the adhesions show a low accuracy and recall.

Table 13 Hyperparameter configuration of the best performing networks

	Gallbladder	Adhesions	Overall Grade
Architecture	Resnet50	Resnet50	Resnet18
Classes	2	2	2
Epochs	50	50	50
Batch size	16	16	32
Learning rate	0.0000492	0.00007911	0.00003635
Dropout	0.4	0	0.2
Weight decay	0.0001	0.0001	0

Gallbladder 2-class classification 'difficult' versus 'not-difficult'

In the figure below, the accuracy and loss plots of the trained network with the highest test accuracy are shown *(Figure 42)*. This network was trained to recognize gallbladder grade 3, or in other words, to recognize cholecystitis. As can be seen the network overfits on the train set after a few epochs. The test results are shown in the confusion matrix and ROC curve in Figure 43 and 44. Table 14 shows the test results of the trained network. The combination of hyperparameters that was used for this training is given in table 15.

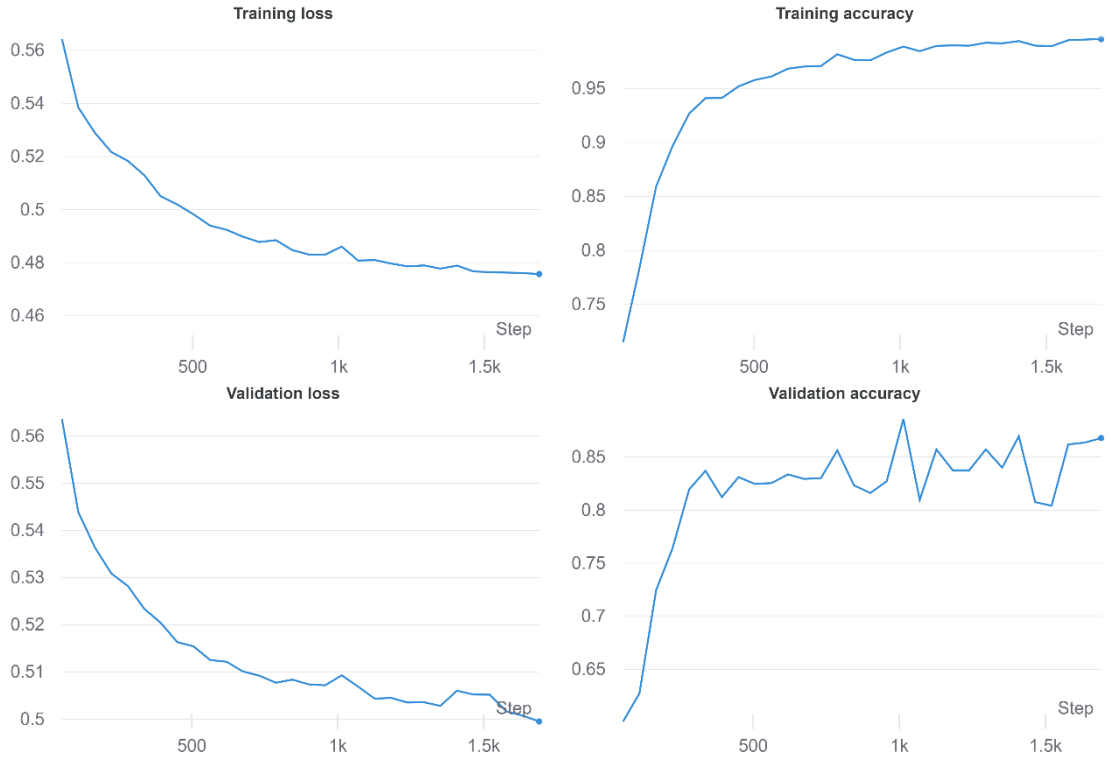


Figure 42 Gallbladder 2-grade (3 versus not-3), best result from sweep: accuracy and loss during training

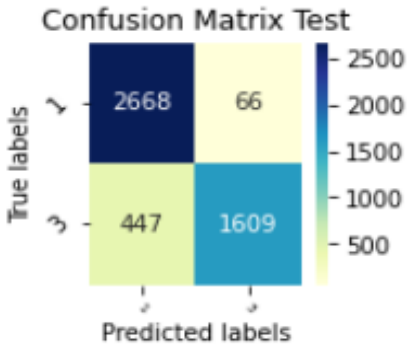


Figure 43 Confusion matrix

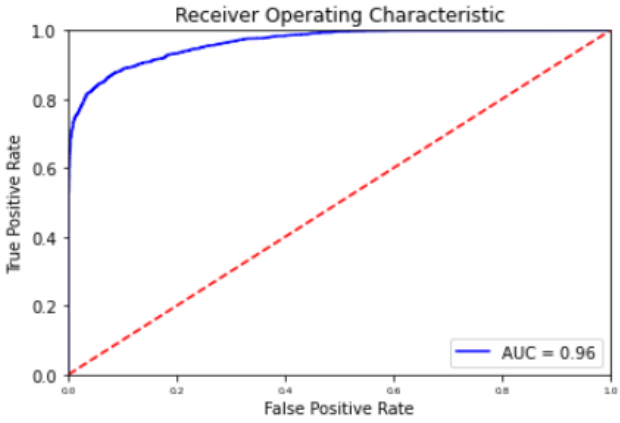


Figure 44 ROC curve

Adhesions 2-class classification 'difficult' versus 'not-difficult'

In the figure below, the accuracy and loss plots of the trained network with the highest test accuracy are shown *(Figure 45)*. This network was trained to recognize adhesions grade 3, or in other words, pathological adhesions. As can be seen the network overfits on the train set after a few epochs. The test results are shown in the confusion matrix and ROC curve Figure 46 and 47. Table 14 shows the test results of the trained network. The combination of hyperparameters that was used for this training is given in table 15.

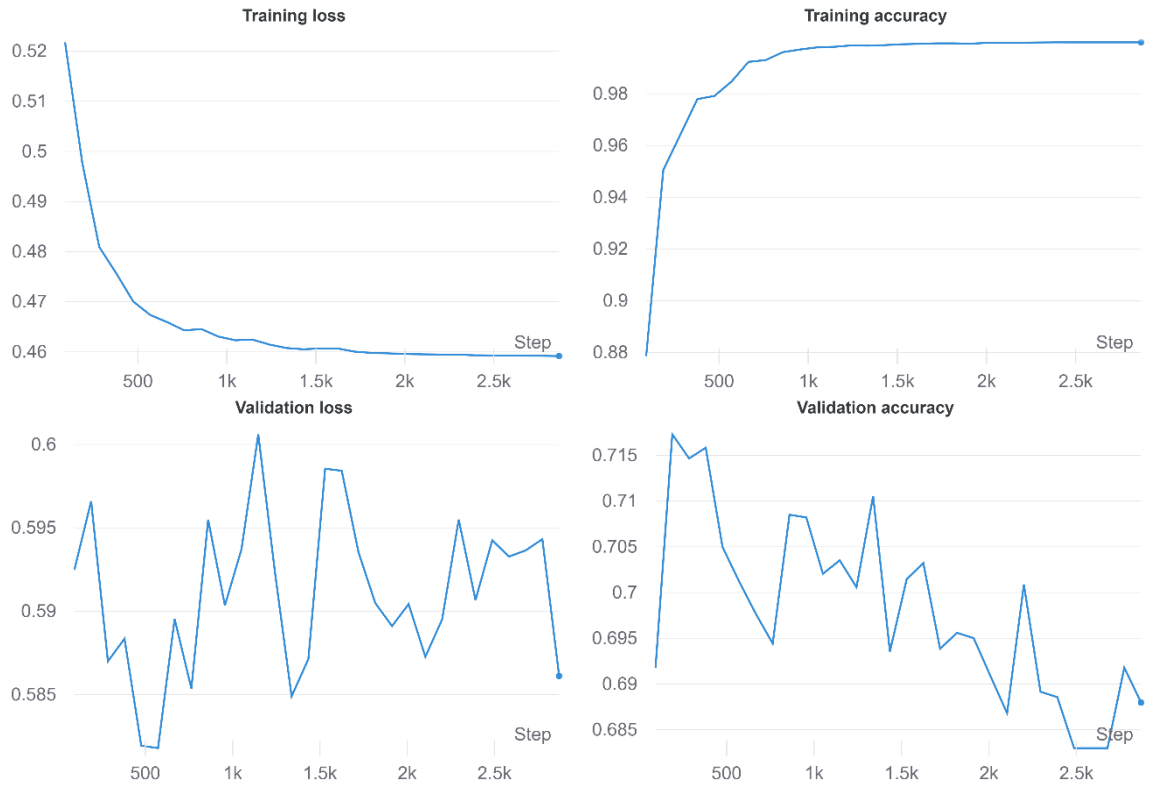


Figure 45 Adhesions 2-grade (3 versus not-3), best result from sweep: accuracy and loss during training

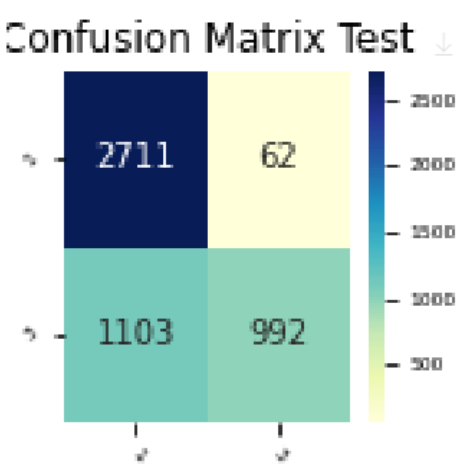


Figure 46 Confusion matrix

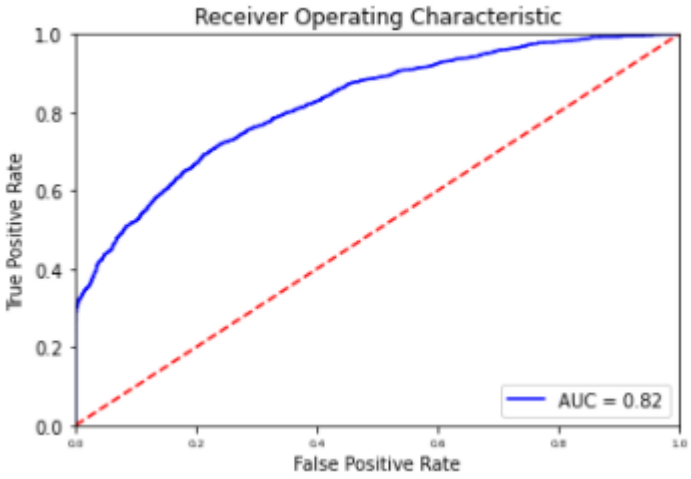


Figure 47 ROC curve

Overall Grade 2-class classification 'difficult' versus 'not-difficult'

In the figure below, the accuracy and loss plots of the trained network with the highest test accuracy are shown *(Figure 48).* This network was trained to recognize overall grade 3, or in other words, to recognize cholecystitis with pathological adhesions. As can be seen the network overfits on the train set after a few epochs. The test results are shown in the confusion matrix and ROC curve in Figure 49 and 50. Table 14 shows the test results of the trained network. The combination of hyperparameters that was used for this training is given in table 15.

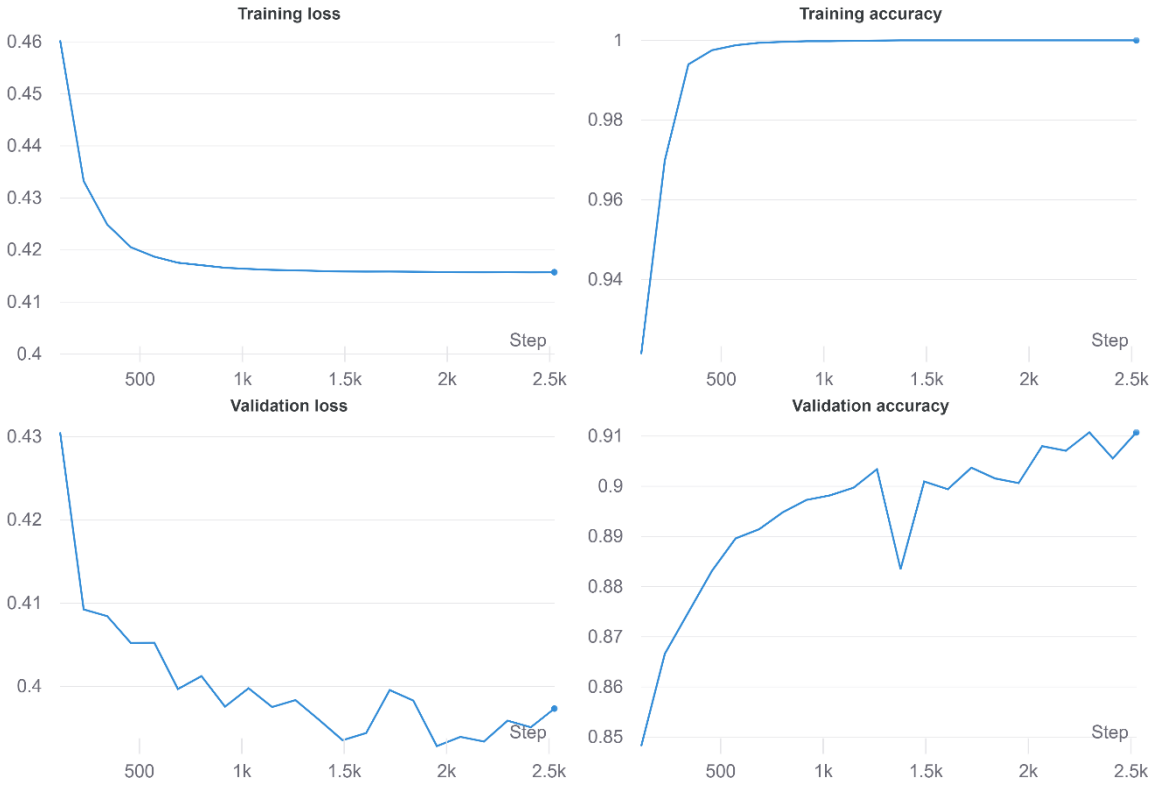


Figure 48 Overall Grade 2-grade (3 versus not-3), best result from sweep: accuracy and loss during training

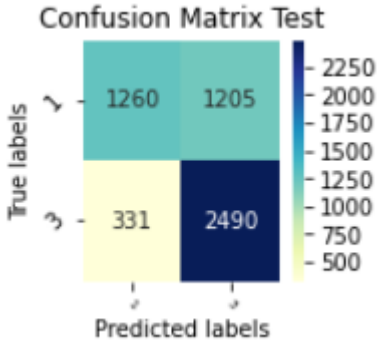


Figure 49 Confusion matrix

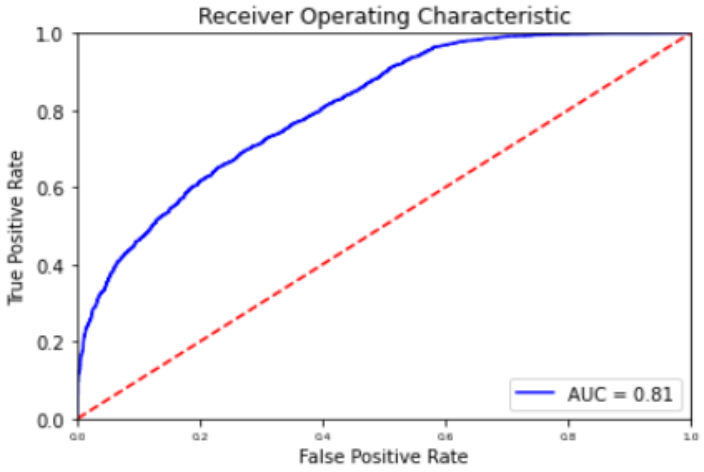


Figure 50 ROC curve

Table 14 Evaluation Binary classification Networks 1&2 - 3

	Accuracy	Precision	Recall	F1
G-binary	0.90	0.97	0.80	0.88
A-binary	0.72	0.94	0.34	0.50
GA-binary	0.71	0.67	0.88	0.76

As can be seen, the network that was trained to recognize gallbladder grade 3 gives the highest accuracy on the test set. This means that the network is capable of recognizing cholecystitis with an accuracy of 90%.

Table 15 Hyperparameter configuration of the best performing networks

	Gallbladder	Adhesions	Overall Grade
Architecture	Resnet18	Resnet18	Resnet50
Classes	2	16	2
Epochs	30	30	30
Batch size	16	16	16
Learning rate	0.00009142	0.00007376	0.00008534
Dropout	0.6	0.2	0
Weight decay	0	0.0001	0.0001

The best performing network of the previous section is the binary gallbladder classification network that is trained to recognize gallbladder grade 3. The evaluation of the training experiments with adjustments to the hyperparameters to reduce overfitting is presented in Table 16. Reducing the learning rate when the accuracy reaches a plateau, in combination with training for 40 epochs and setting dropout to 0.2 resulted in a test accuracy increase by 1%.

Table 16 Evaluation binary gallbladder classification networks (grade 1&2 – 3) of experiments to improve performance

	Accuracy	Precision	Recall
Reduce learning rate on plateau, 40 epochs, dropout 0.2	0.91	0.98	0.81
70 epochs, early stopping	0.87	0.99	0.71
Dropout 0.6	0.87	0.99	0.71
Dropout 0.8	0.87	0.99	0.71
Wd 0.01 Dropout 0.2	0.87	0.99	0.71
Data augmentation	0.89	0.99	0.75

5.3 Correctly identified frames per video

In this section we will further elaborate on the best performing networks, which are:

- Gallbladder 3 Grades: The multiclass network trained to classify the gallbladder in either grade 1, 2 or 3
- Gallbladder 2 Grades: The binary network trained to recognize gallbladder grade 3 (cholecystitis)
- Overall grade 2 Grades: The binary network trained to recognize overall grade 1 (normal gallbladder without adhesions)

Gallbladder 3 Grades

The percentage of correctly classified frames per video in the test set is presented in Figure 51. When more than 50% of the frames are correctly classified the video is considered correctly classified. The dotted yellow line marks this threshold. This figure shows that videos with gallbladder grade 1 or 3 can be correctly classified. The videos with gallbladder grade 2 do not beat the threshold of 50%. These videos are therefore not classified correctly.

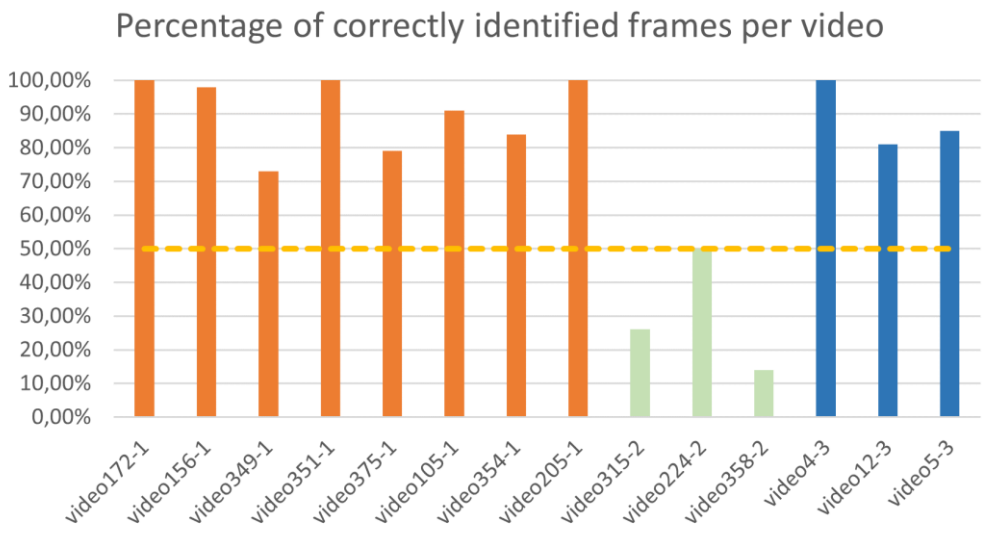


Figure 51 Gallbladder 3 grades classification rate for all test videos

Identifying cholecystitis

The results for this network on the test videos are given in Figure 52. In all videos, more than 50% of the frames are correctly classified and therefore all test videos are correctly classified as either 'cholecystitis' or 'no cholecystitis'.

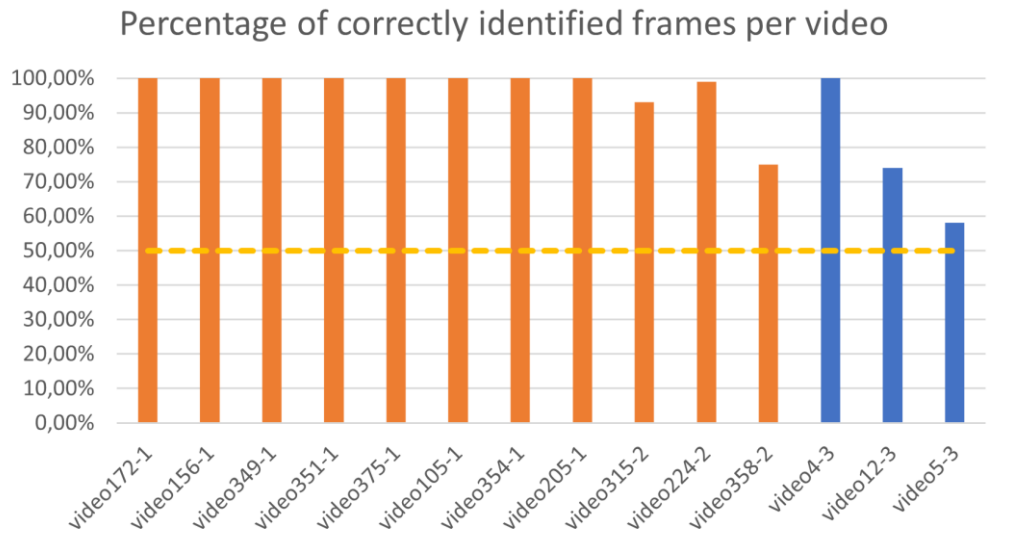


Figure 52 Gallbladder binary classification rate for all test videos

Identifying easy cases

The results for this network on the test videos are given in Figure 53. In all the videos with an overall grade of 2 or 3 (blue), more than 50% of the frames are correctly classified. In the videos with an overall grade of 1 (yellow), 4 out of 6 are correctly classified.

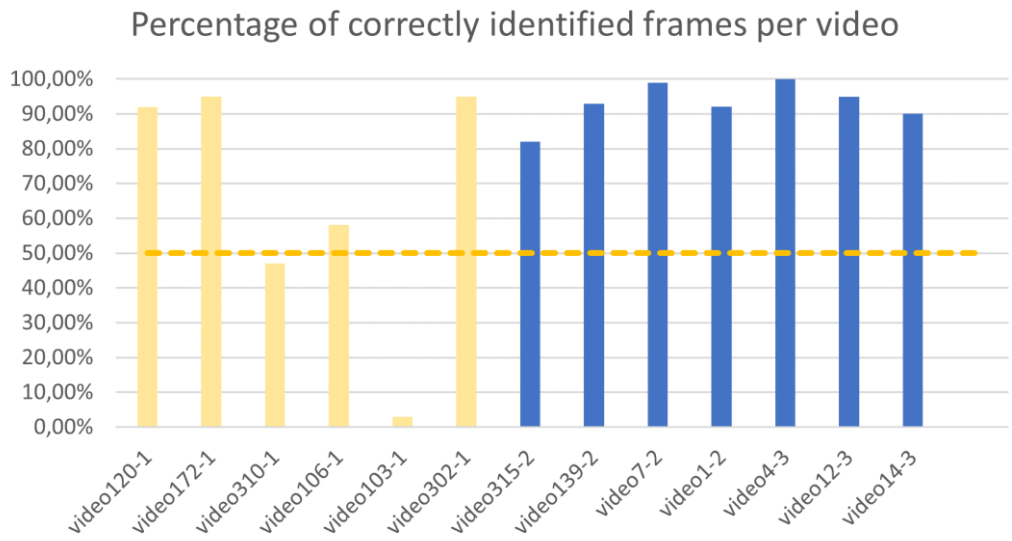


Figure 53 Overall grade binary classification rate for all test video

6. Discussion

This chapter discusses the outcomes of the result section. The summary of the results of the trained networks is given in 6.1. Hereafter, these results will be explained. Limitations of this study are given in 5.3 and this chapter ends with the recommendations for future research and the clinical applicability of the results.

6.1 Summary of results

This study aimed to develop a DL network that can predict the difficulty of a LC. The research question was to what extent this is possible in the first phase of surgery using laparoscopic videos. Therefore, a model is needed that recognizes the 3 defined difficulty grades based on the Nassar scale that describe the presentation of the gallbladder and the adhesions. The results are further interpreted below, but the short answer to this question is that it not possible to accurately predict difficulty in 3 grades based on both gallbladder and adhesions. When considering only 2 difficulty grades and therefore making it a binary classification problem, the results are better. It is possible to accurately recognize gallbladder grade 3. It is also possible to accurately recognize overall grade 1, or in other words, easy cases with a normal gallbladder without adhesions. The gallbladder difficulty is better recognizable than the adhesions difficulty using single frames. At this moment, the surgical planning cannot yet be improved accurately.

To achieve these results, frames were extracted from LC videos to use for training. The Nassar scale for difficulty was used as a scoring system. The Nassar scale scores on the appearance of the gallbladder, the adhesions, and the cystic pedicle. All frames were labeled for 'gallbladder' and 'adhesions' with grade 1, 2, or 3. The cystic pedicle was left out for the scope of this research. This was done for the frames of 93 LC videos. After excluding the 'out-of-body' frames and frames in which the gallbladder was not visible, this resulted in a total of 26.483 labeled frames.

Multiclass CNNs were trained to classify gallbladder and adhesions. Hyperparameter sweeps were used to find the most optimal hyperparameter configuration. The best results for the multiclass CNNs were obtained with the gallbladder classification network with an accuracy on the test set of 74%. The Adhesions classification network reached an accuracy of 50%. Out of 14 videos in the test set, 11 were classified correctly. Videos with grade 1 and grade 3 were all correctly classified, and videos with grade 2 were not.

To see if it was possible to recognize the easy cases or the difficult cases, also binary CNNs were trained for gallbladder, adhesions, and the overall grade. First, binary CNNs were trained to classify 'easy' (label: 1) versus 'not-easy' (label: 2&3). The best result was obtained with the network trained for the classification of the overall grade, with an accuracy of 87%. All 14 videos in the test set were correctly classified. In addition, CNN's were trained to classify 'difficult' (label: 3) versus 'not-difficult' (label: 1&2). The best result was obtained with the network trained for the classification of the gallbladder with an accuracy of 90%. A total of 11 out of 13 test videos were correctly classified. To improve the performance of the last-mentioned network we tried to find ways to prevent overfitting. Setting the dropout to 0.2 in combination with reducing the learning rate on plateau and training for 40 epochs increased the test accuracy to 91%.

6.2 Explanation of results

The most eminent lesson learned during the labeling process was the importance of assigning clear definitions to difficulty grades. These definitions must contain clinically valuable information, but they must also be visible in the image. The classification is only clinically applicable if the grades describe a scale of difficulty. We concluded that the composition of adhesions is more important when looking at difficulty than the location of the adhesions. For example, dense adhesions at the fundus of the gallbladder should result in assigning grade 3 as a label for adhesions. This may make the classification task even more difficult for a network to learn. Only when the adhesions are retracted with surgical instruments, the density or nature of the adhesions can be determined. This makes it very hard to determine the grade of the adhesions in only one frame.

With the way the dataset was labeled, using the Nassar scale as a starting point, it seems that the difficulty grades hold predictive value when looking at total surgery duration. Videos that were labeled with an overall grade of 3 result in a longer operating time (21 minutes and 5 seconds on average) than videos with grade 1. Unfortunately, this does not yet give an accurate estimation of total surgery time because of the high standard deviation. Furthermore, this dataset does not necessarily give an accurate representation of the actual distribution in easy or difficult LCs. To give an accurate estimation of total surgery time and the actual prediction error after classifying difficulty, the dataset should be expanded with many more patients.

The intention was to use K-fold cross-validation, making it possible to perform multiple training and testing sessions with different train, validation and test split. Because the frames of one patient should only be used in either the train, validation, or test set, splitting the data should be done per video. If this is done randomly this would not result in an equal distribution of the classes in all sets. Because one video often contains only one label, this could even result in sets where a class is completely absent. No method was found to split the data in a statistical, random, or automatic way because of the way the data was structured. Therefore, cross-validation was not possible. Unfortunately, this means that not all data was used for training. Because there is little variation in the frames in one patient, valuable information to improve the network is not used with the current method. The videos were manually split into train, validation, and test sets to achieve an equal distribution of the classes in all sets. This resulted in a different distribution of the videos in all sets when training for gallbladder, adhesions, or overall grade. To combine and use all trained networks, an additional test set is necessary, that consists of unseen data for all the trained networks.

Training

After the first test sweeps it was already noted that SGD optimizer performed better than Adam optimizer and therefore it was decided to continue using only SGD. SGD seemed better in generalization than Adam. Using hyperparameter sweeps was a good way to search for an optimized configuration of hyperparameters. When starting with the training sessions, it was quickly noted that starting with a small learning rate in an order of magnitude 0.00001 or lower gave better results. Also letting the learning rate decay when a plateau was reached was beneficial,

and adding early stopping as well. It varied per trained network which combination of the batch size and learning rate resulted in the best result. All networks were able to overfit on the training data after a few epochs. It should be noted that it matters which videos are present in the test set. Because there is not a lot of variability in the test data in only 14 videos. That is why sometimes there was a significant difference in the validation accuracy and the test accuracy.

Multiclass classification

Of the 3-class classification networks, the gallbladder classification gives the best results. Intuitively, this is not very surprising. Cholecystitis has many visible aspects, such as bloodiness, redness, gangrene, and pus. The sensitivity of grade 1 and 3 is high (90% and 84%), whereas the sensitivity of grade 2 is very low (33%). Grade 2 is very hard to recognize. During the labeling process, the biggest challenges were to set clear boundaries between the grades. It seems that the hydropic, swollen gallbladder is not recognizable in one frame. Most of the frames with grade 2 were classified as grade 1. After puncturing the swollen gallbladder, it looks like gallbladder grade 1 again. More videos with gallbladder grade 2 should be added to the dataset to improve the results and make it possible to recognize grade 2. During the labeling process, it was decided that one video could only have one label for gallbladder. For the adhesions, it could be the case that the video starts with label 2 and gradually goes to grade 1 when the adhesions are taken down.

For the classification of the adhesions, it can be concluded that this was not yet successful. The explanation for this is probably that action is needed to see whether it is adhesions grade 2 or grade 3. When the adhesion is retracted, it is visible if it is an anatomical or easy adhesion, or it is a pathological or difficult adhesion. The frame-wise classification does therefore not suffice, because the context of other frames is needed. In a single frame, the composition of the adhesion and its relation to the surrounding tissue is not visible. This makes it unsuitable for a CNN to learn features from. Because it was decided that the density of the adhesion was more important than its location in order of difficulty, the task may be even harder. A single dense adhesion makes it grade 3, but it may resemble grade 1 or 2 too much to be accurately distinguished. Much time was spent on deciding how to label the data for adhesions, and different ways were tried, but it remains hard to decide which adhesions are grade 2 or grade 3. Because it is already difficult for a human to distinguish the grades, it is even more difficult for a network to learn this task. It may be better to train a network that is able to recognize the action ‘dissecting adhesions’, although then there would still be little difference in grade 2 and grade 3.

Binary classification

It was expected that the binary classification would result in a higher accuracy because this is simply an easier task. With an accuracy of 84% on the test set, the binary classification network trained to recognize a gallbladder grade 1 already shows a promising result. The network trained to recognize overall grade 1 performed even better (87%). In these cases, the whole shape of the gallbladder is often completely visible and there are little to no adhesions present. This network has a sensitivity of 93%, which is an important measure for suitability to detect easy cases. This is probably because there is a lot of variability in the data with grade 1. Though, when classifying

whole videos, only 4 out of 6 were identified correctly as easy. The frames of one of these videos were almost completely classified as grade 2. A total of 13 test videos is still a very low amount. More videos should be added to the test set to conclude if this could be clinically applicable. When looking at the sensitivity, this network could be useful to quickly identify the easy cases.

From the binary classification networks that were trained with labels 1 and 2 taken together and labels 3, the gallbladder network showed the best results. With an accuracy of 90% on the test set, the network can recognize difficult cases. In other words, it can recognize cholecystitis. All three test videos with gallbladder grade 3 were correctly classified. This could be clinically useful to automatically recognize difficult gallbladder cases. The binary adhesions grade 3 classification reached an accuracy of 76%, but with a sensitivity of 47%. This unfortunately means that it is not yet useful for clinical practice.

Reduce overfitting

To increase the clinical applicability of the binary gallbladder classification network (recognizing cholecystitis), several experiments were conducted to improve the performance of the network. The only adjustments that increased the accuracy on the test set to 91% were the combination of setting dropout to 0.2, reducing the learning rate on plateau, and training for 40 epochs. It was expected that the other attempts would also improve the result, but these were ineffective. It is possible that the specific data augmentation techniques that were chosen (random rotation and random perspective) are not suitable for this particular task. Many other data augmentation techniques exist and have yet to be tested. It has to be mentioned, that the little variability in the test set may have a significant influence on the evaluation of all networks. In some evaluation results, the validation accuracy was even 5% lower than the test accuracy.

6.3 Limitations

During the labeling process, there was the possibility to discuss the choices that had to be made with an expert surgeon and a junior surgeon. Because eventually the frame-by-frame labeling was done by me, no inter observer variability test could be performed. It is possible that there are some inconsistencies in the labels of the frames and that another expert surgeon would label some of the adhesions grade 3 frames as grade 3 but as grade 2, for example. When there are inconsistencies in the labels, it causes problems for the network to learn the right features from the data for a specific class. It would be better if two expert surgeons would both label the data, and the variability could be studied.

The size of the dataset is also a limiting factor. The most time-consuming process of this study was to decide the label definitions and to create the dataset. Frame-by-frame labeling is a time-consuming process, but when the method is decided and one has spent hours and hours watching LC videos, it becomes easier to label the data even for a non-expert human. The expectation is that the results would improve when more variability is added to the dataset. Especially for the gallbladder 3-class classification task, it would be better if more data was available. In this study, one frame per second was used to create the dataset. Using only one frame per second, resulted in the possibility that two consecutive frames are significantly different. Valuable information may be lost. Using a higher frame rate, for example, 5 frames per second, could improve the

performance of the network in a relatively simple way. Although this would be a simple way to increase the size of the dataset, it may be better to use an equal number of frames per patient. This would result in a more equal distribution of the classes in the dataset.

When we started with the creation of the dataset, we only checked if the first phase of the surgery was fully recorded and cut this part into sub videos to use for training. Some of the videos did not contain the whole surgery and were either missing the beginning or the end. When calculating the average total operating time for each grade, we corrected for the parts that were missing as correctly as possible. Nevertheless, there exists an inaccuracy in the total surgery time. Adding the fact that there are only 93 patients included in this study, improving surgical planning by estimating total surgery time is not yet achievable using this data.

6.4 Recommendations for future research

To increase the variability in the training data, the dataset should be expanded. In Meander MC there are now hundreds of LC videos available for research. The labeling should be done by at least two surgeons, minimizing the possibility of label inconsistencies. For the adhesion's recognition task, possibilities for surgical action recognition should be explored. It is expected, that even with a larger dataset, recognizing adhesions would still give poor results using only single frames. Because the instrument-tissue interaction is important to grade adhesions, temporal networks should be considered. A Long Short-Term Memory (LSTM) network is a special type of Recurrent Neural Network that can learn long-term dependencies. It uses 'memory cells' that can choose whether the information coming in is to be remembered or is irrelevant and can be forgotten. If the consecutive frames are used, the contextual information can be incorporated. In addition, it is recommended to improve the results by using a multitask network to learn to recognize the gallbladder grade and adhesions grade simultaneously.

Another method that should be further explored is multitask learning. During this research, we already worked on the creation of a multitask network, but unfortunately due to lack of time these results are not included in the scope of this study. In a multitask learning network, the goal is to learn multiple tasks at the same time from the same input data. The network would have to learn the gallbladder grade and the adhesions grade simultaneously. In multitask learning, a loss function should be defined for each task and the total loss would be the sum of those loss functions. The back-propagation step during training is executed similarly to single-task learning. Because the two tasks are correlated to each other it could result in improved parameters in the learned layers.[43] By using the similarities between two related tasks, we allow the model to generalize better on the original task. When a specific task is related to the other, a multitask model can improve data efficiency, reduce overfitting through shared representations and decrease training time.[75]

6.5 Clinical applicability and future perspective

The goal is to predict surgical difficulty at the start of surgery. This must be done automatically at the right moment and the surgeon should not have to perform an extra action. To implement the results of this study in clinical practice, it is important to classify difficulty at the right moment. It is therefore necessary to incorporate phase recognition for this task. The difficulty prediction

should be made right after the first phase is over. If this phase is accurately recognized, using the results of this study it is possible to classify cholecystitis with an accuracy of 91% or classify easy cases with an accuracy of 87%. This makes it possible to automatically report cholecystitis cases or easy cases with reasonable certainty. The results of this study show that it is still not possible to accurately recognize the adhesions grade in LC. To accurately predict surgical difficulty in 3 grades, further research is necessary to improve the results. To improve surgical planning, it is also important to combine the difficulty prediction with a preoperative prediction of surgical difficulty. Combining preoperative prediction of surgical difficulty, phase recognition, and intraoperative difficulty prediction will improve surgical planning and benchmarking for surgeons. Developments in AI techniques will enable context-aware assistance during surgery in the future. This will make it possible to improve surgical planning and give surgeons more insight into their performance. The future of healthcare will increasingly be influenced by these developments through risk prediction and help in decision making.

7. Conclusion

This is the first step in predicting surgical difficulty at the start of a LC. This study aimed to develop a DL algorithm that can predict surgical difficulty using laparoscopic videos. It is not yet possible to classify the overall difficulty grade based on the Nassar scale by classifying both 'gallbladder' and 'adhesions' with reasonable accuracy. It is possible to recognize easy cases and cholecystitis cases with the binary classification networks with a high accuracy. The gallbladder grade is better recognizable using single frames than the adhesions grade. It is recommended to explore the use of temporal networks to recognize the density of adhesions. To make the results clinically applicable, it is recommended to use a surgical phase recognition model. The results of this study could be used as a starting point for further research in classifying difficulty in LC. This would be the first step to improve understanding of surgical scenery and allow benchmarking for surgeons in LC.

References

1. Sugrue M, Coccolini F, Bucholc M, Johnston A, Wses C. Intra-operative gallbladder scoring predicts conversion of laparoscopic to open cholecystectomy : a WSES prospective collaborative study. 2019;9:10-17.

2. Hassler KR, Jones MW. *Gallbladder, Cholecystectomy, Laparoscopic*. StatPearls Publishing; 2017. Accessed January 21, 2021. https://www.ncbi.nlm.nih.gov/books/NBK448145/

3. Atta HM, Mohamed AA, Sewefy AM, Abdel-Fatah A-FS, Mohammed MM, Atiya AM. Difficult Laparoscopic Cholecystectomy and Trainees: Predictors and Results in an Academic Teaching Hospital. Published online 2017. doi:10.1155/2017/6467814

4. Majumder A, Altieri MS, Brunt LM. How do I do it : laparoscopic cholecystectomy. Published online 2020:4-7. doi:10.21037/ales.2020.02.06

5. Kumar Sahu S, Sahu SK, Agrawal A, Sachan PK. INTRAOPERATIVE DIFFICULTIES IN LAPAROSCOPIC CHOLECYSTECTOMY. *Jurnalul Chir*. 2013;9(2). doi:10.7438/1584-9341-9-2-5

6. Bat O. The analysis of 146 patients with difficult laparoscopic cholecystectomy. 2015;8(9):16127-16131.

7. Bouarfa L, Schneider A, Feussner H, et al. Prediction of intraoperative complexity from preoperative patient data for laparoscopic cholecystectomy. *Artif Intell Med*. 2011;52(3):169-176. doi:10.1016/j.artmed.2011.04.012

8. Philip J, Burcharth J, Pommergaard H. Preoperative Risk Factors for Conversion of Laparoscopic Cholecystectomy to Open Surgery – A Systematic Review and Meta-Analysis of Observational Studies. Published online 2016:414-423. doi:10.1159/000445505

9. Lal P, Agarwal PN, Malik VK, Chakravarti AL. A difficult laparoscopic cholecystectomy that requires conversion to open procedure can be predicted by preoperative ultrasonography. *JSLS*. 2002;6(1):59-63. Accessed January 5, 2021. /pmc/articles/PMC3043388/?report=abstract

10. Buchlak QD, Esmaili N, Leveque JC, et al. Machine learning applications to clinical decision support in neurosurgery: an artificial intelligence augmented systematic review. *Neurosurg Rev*. 2020;43(5):1235-1253. doi:10.1007/s10143-019-01163-8

11. Priego P, Ramiro C, Molina JM, et al. Results of laparoscopic cholecystectomy in a third-level university hospital after 17 years of experience. *Rev Española Enfermedades Dig*. 2009;101(1). doi:10.4321/S1130-01082009000100003

12. Strasberg SM. A perspective on the critical view of safety in laparoscopic cholecystectomy. *Ann Laparosc Endosc Surg*. 2017;2(5):91-91. doi:10.21037/ales.2017.04.08

13. D F, D M, H CE, M R, M J. Mean operating room times differ by 50 % among hospitals in different countries for laparoscopic cholecystectomy and lung lobectomy. Published

online 2006:319-322. doi:10.1007/s00540-006-0419-4

14. Thiels CA, Yu D, Abdelrahman AM, et al. The use of patient factors to improve the prediction of operative duration using laparoscopic cholecystectomy. *Surg Endosc*. 2017;31(1):333-340. doi:10.1007/s00464-016-4976-9

15. Sutcliffe RP, Hollyman M, Hodson J, Bonney G, Vohra RS. Preoperative risk factors for conversion from laparoscopic to open cholecystectomy : a validated risk score derived from a prospective U . K . database of 8820 patients. Published online 2016:922-928. doi:10.1016/j.hpb.2016.07.015

16. Atta HM, Mohamed AA, Sewefy AM, Abdel-Fatah AFS, Mohammed MM, Atiya AM. Difficult Laparoscopic Cholecystectomy and Trainees: Predictors and Results in an Academic Teaching Hospital. *Gastroenterol Res Pract*. 2017;2017. doi:10.1155/2017/6467814

17. Radunovic M, Lazovic R, Popovic N, et al. Complications of laparoscopic cholecystectomy: Our experience from a retrospective analysis. *Maced J Med Sci*. 2016;4(4):641-646. doi:10.3889/oamjms.2016.128

18. Hussain A. Difficult laparoscopic cholecystectomy: Current evidence and strategies of management. *Surg Laparosc Endosc Percutaneous Tech*. 2011;21(4):211-217. doi:10.1097/SLE.0b013e318220f1b1

19. Iwashita Y, Hibi T, Ohyama T, et al. An opportunity in difficulty: Japan–Korea–Taiwan expert Delphi consensus on surgical difficulty during laparoscopic cholecystectomy. *J Hepatobiliary Pancreat Sci*. 2017;24(4):191-198. doi:10.1002/jhbp.440

20. Sugrue M, Sahebally SM, Ansaloni L, Zielinski MD. Grading operative findings at laparoscopic cholecystectomy- A new scoring system. *World J Emerg Surg*. 2015;10(1):14. doi:10.1186/s13017-015-0005-x

21. Hardman RL, Jazaeri O, Yi J, Smith M, Gupta R. Overview of classification systems in peripheral artery disease. *Semin Intervent Radiol*. 2014;31(4):378-388. doi:10.1055/s-0034-1393976

22. Low SW, Iyer SG, Chang SKY, Mak KSW, Lee VTW, Madhavan K. Laparoscopic cholecystectomy for acute cholecystitis: safe implementation of successful strategies to reduce conversion rates. *Surg Endosc*. 2009;23(11):2424-2429. doi:10.1007/s00464-009-0374-x

23. Tang B, Cuschieri A. Conversions During Laparoscopic Cholecystectomy: Risk Factors and Effects on Patient Outcome. *J Gastrointest Surg*. 2006;10(7):1081-1091. doi:10.1016/j.gassur.2005.12.001

24. Rosen M, Brody F, Ponsky J. *Predictive Factors for Conversion of Laparoscopic Cholecystectomy.*; 2002.

25. Gupta V, Jain G. Wj g s. 2019;11(2):62-85. doi:10.4240/wjgs.v11.i2.62

26. Sato N, Yabuki K, Shibao K, et al. Risk factors for a prolonged operative time in a single-incision laparoscopic cholecystectomy. *Hpb*. 2014;16(2):177-182. doi:10.1111/hpb.12100

27. Subhas G, Gupta A, Bhullar J, et al. Prolonged (longer than 3 hours) laparoscopic cholecystectomy: Reasons and results. *Am Surg*. 2011;77(8):981-984. doi:10.1177/000313481107700814

28. Jameel SM, Bahaddin MM, Mohammed AA. Grading operative findings at laparoscopic cholecystectomy following the new scoring system in Duhok governorate: Cross sectional study. *Ann Med Surg*. 2020;60:266-270. doi:10.1016/j.amsu.2020.10.035

29. Griffiths EA, Hodson J, Vohra RS, et al. Utilisation of an operative difficulty grading scale for laparoscopic cholecystectomy. *Surg Endosc*. 2019;33(1):1-12. doi:10.1007/s00464-018-6281-2

30. Yokoe M, Hata J, Takada T, et al. Tokyo Guidelines 2018: diagnostic criteria and severity grading of acute cholecystitis (with videos). *J Hepatobiliary Pancreat Sci*. 2018;25(1):41-54. doi:10.1002/jhbp.515

31. Madni TD, Leshikar DE, Minshall CT, et al. The Parkland grading scale for cholecystitis. Published online 2017. doi:10.1016/j.amjsurg.2017.05.017

32. Nassar AHM, Ashkar KA, Mohamed AY, Hafiz AA. Is laparoscopic cholecystectomy possible without video technology? *Minim Invasive Ther Allied Technol*. 1995;4(2):63-65. doi:10.3109/13645709509152757

33. Amboldi M, Amboldi A, Gherardi G, Bonandrini L. Complications of Videolaparoscopic Cholecystectomy: A Retrospective Analysis of 1037 Consecutive Cases. *Int Surg*. 2011;96:35-44. Accessed December 7, 2021. http://meridian.allenpress.com/international-surgery/article-pdf/96/1/35/2213921/1385_1.pdf

34. Lirici MM, Califano A. Management of complicated gallstones: Results of an alternative approach to difficult cholecystectomies. *Minim Invasive Ther Allied Technol*. 2010;19(5):304-315. doi:10.3109/13645706.2010.507339

35. Ye G, Qin Y, Xu S, et al. Comparison of transumbilical single-port laparoscopic cholecystectomy and fourth-port laparoscopic cholecystectomy. *Int J Clin Exp Med*. 2015;8(5):7746-7753.

36. Anteby R, Horesh N, Soffer S, et al. Deep learning visual analysis in laparoscopic surgery: a systematic review and diagnostic test accuracy meta-analysis. *Surg Endosc*. 2021;35(4):1521-1533. doi:10.1007/s00464-020-08168-1

37. Twinanda AP, Yengera G, Mutter D, Marescaux J, Padoy N. RSDNet: Learning to Predict Remaining Surgery Duration from Laparoscopic Videos Without Manual Annotations. *IEEE Trans Med Imaging*. 2019;38(4):1069-1078. doi:10.1109/TMI.2018.2878055

38. Rivas-blanco I, Pérez-del-pulgar CJ, García-morales I, Muñoz VF. A Review on Deep Learning in Minimally Invasive Surgery. 2021;9.

39. Korndorffer JR, Hawn MT, Spain DA, et al. Situating Artificial Intelligence in Surgery. *Ann Surg*. 2020;272(3):523-528. doi:10.1097/sla.0000000000004207

40. Madani A, Namazi B, Altieri MS, et al. Artificial Intelligence for Intraoperative Guidance.

Ann Surg. 2020;Publish Ah. doi:10.1097/sla.0000000000004594

41. Mascagni P, Fiorillo C, Urade T, et al. Formalizing video documentation of the Critical View of Safety in laparoscopic cholecystectomy: a step towards artificial intelligence assistance to improve surgical safety. *Surg Endosc*. 2020;34(6):2709-2714. doi:10.1007/s00464-019-07149-3

42. Tokuyasu T, Iwashita Y, Matsunobu Y, et al. Development of an artificial intelligence system using deep learning to indicate anatomical landmarks during laparoscopic cholecystectomy. *Surg Endosc*. 2020;(0123456789). doi:10.1007/s00464-020-07548-x

43. Troccaz MJ, Hager MG, Hopkins J, Gwenolé Quellec UM. Vision-based Approaches for Surgical Activity Recognition Using Laparoscopic and RBGD Videos Chair of the Committee: Examiners. Published online 2017.

44. Cheng K, You J, Wu S, et al. Artificial intelligence-based automated laparoscopic cholecystectomy surgical phase recognition and analysis. *Surg Endosc*. 1:3. doi:10.1007/s00464-021-08619-3

45. Twinanda AP, Shehata S, Mutter D, Marescaux J, Mathelin M De, Padoy N. EndoNet : A Deep Architecture for Recognition Tasks on Laparoscopic Videos. 2017;36(1):86-97.

46. Dergachyova O, Bouget D, Huaulmé A, Morandi X, Jannin P. Automatic data-driven real-time segmentation and recognition of surgical workflow. Published online 2016:1081-1089. doi:10.1007/s11548-016-1371-x

47. Ward TM, Fer DM, Ban Y, Rosman G, Meireles OR, Hashimoto DA. Challenges in surgical video annotation. *Comput Assist Surg*. 2021;26(1):58-68. doi:10.1080/24699322.2021.1937320

48. Toh C, Brody JP. Applications of Machine Learning in Healthcare. *Smart Manuf - When Artif Intell Meets Internet Things*. Published online January 14, 2021. doi:10.5772/INTECHOPEN.92297

49. What is the Definition of Machine Learning? | Expert.ai | Expert.ai. Accessed December 14, 2021. https://www.expert.ai/blog/machine-learning-definition/

50. Machine Learning For Beginners. Machine learning was defined in 90's by... | by Divyansh Dwivedi | Towards Data Science. Accessed December 14, 2021. https://towardsdatascience.com/machine-learning-for-beginners-d247a9420dab

51. Reinforcement Learning 101. Learn the essentials of Reinforcement... | by Shweta Bhatt | Towards Data Science. Accessed December 14, 2021. https://towardsdatascience.com/reinforcement-learning-101-e24b50e1d292

52. Jackson AH. Machine learning. *Expert Syst*. 1988;5(2):132-150. doi:10.1111/j.1468-0394.1988.tb00341.x

53. Kwon SJ. Artificial neural networks. *Artif Neural Networks*. Published online 2011:1-426. doi:10.15864/jmscm.1104

54. Graves A, Mohamed A-R, Hinton G. SPEECH RECOGNITION WITH DEEP RECURRENT NEURAL NETWORKS.

55. Rawat W, Wang Z. Deep Convolutional Neural Networks for Image Classification : A Comprehensive Review Deep Convolutional Neural Networks for Image Classification : A Comprehensive Review. 2017;(October). doi:10.1162/NECO

56. A Comprehensive Hands-on Guide to Transfer Learning with Real-World Applications in Deep Learning | by Dipanjan (DJ) Sarkar | Towards Data Science. Accessed December 7, 2021. https://towardsdatascience.com/a-comprehensive-hands-on-guide-to-transfer-learning-with-real-world-applications-in-deep-learning-212bf3b2f27a

57. Activation Functions | Fundamentals Of Deep Learning. Accessed December 15, 2021. https://www.analyticsvidhya.com/blog/2020/01/fundamentals-deep-learning-activation-functions-when-to-use-them/

58. CS231n Convolutional Neural Networks for Visual Recognition. Accessed December 7, 2021. https://cs231n.github.io/convolutional-networks/#conv

59. A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way | by Sumit Saha | Towards Data Science. Accessed December 7, 2021. https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

60. Loss Functions — ML Glossary documentation. Accessed December 15, 2021. https://ml-cheatsheet.readthedocs.io/en/latest/loss_functions.html

61. Zaheer R. A Study of the Optimization Algorithms in Deep Learning. *2019 Third Int Conf Inven Syst Control*. 2020;(August):536-539. doi:10.1109/ICISC44355.2019.9036442

62. Kingma DP, Ba JL. A : a m s o. Published online 2015:1-15.

63. Wilson AC, Roelofs R, Stern M, Srebro N, Recht B. The marginal value of adaptive gradient methods in machine learning. *Adv Neural Inf Process Syst*. 2017;2017-Decem(Nips):4149-4159.

64. Hardt M, Recht B, Singer Y. Train faster, generalize better: Stability of stochastic gradient descent. Published online 2016.

65. Model Fit: Underfitting vs. Overfitting - Amazon Machine Learning. Accessed December 2, 2021. https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html

66. 8 Simple Techniques to Prevent Overfitting | by David Chuan-En Lin | Towards Data Science. Accessed November 23, 2021. https://towardsdatascience.com/8-simple-techniques-to-prevent-overfitting-4d443da2ef7d

67. L1 and L2 Regularization Methods. Machine Learning | by Anuja Nagpal | Towards Data Science. Accessed December 15, 2021. https://towardsdatascience.com/l1-and-l2-regularization-methods-ce25e7fc831c

68. Hinton G. Dropout : A Simple Way to Prevent Neural Networks from Overfitting. 2014;15:1929-1958.

69. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol*. 2010;5(9):1315-1316. doi:10.1097/JTO.0b013e3181ec173d

70. Grandini M, Bagli E, Visani G. Metrics for Multi-Class Classification: an Overview. Published online 2020:1-17. http://arxiv.org/abs/2008.05756

71. Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: Understanding transfer learning for medical imaging. *arXiv*. 2019;(NeurIPS).

72. He K, Sun J. Deep Residual Learning for Image Recognition. :1-9.

73. Gerkema MH. Deep learning for identification of gallbladder leakage during laparoscopic cholecystectomy. Published online 2020.

74. Weights & Biases - Documentation. Accessed December 10, 2021. https://docs.wandb.ai/

75. Ruder S. An Overview of Multi-Task Learning in Deep Neural Networks. 2017;(May). http://arxiv.org/abs/1706.05098

Appendix

Overall Grade 3-class classification

Figure 55 shows the accuracy and loss during training. It is seen that the network overfits on the train set after a few epochs. The validation accuracy does not increase effectively. The hyperparameter configuration used in this training session is given in table 15. The confusion matrix of the test set is given in igure 56. Table 14 shows the test results of the trained network.

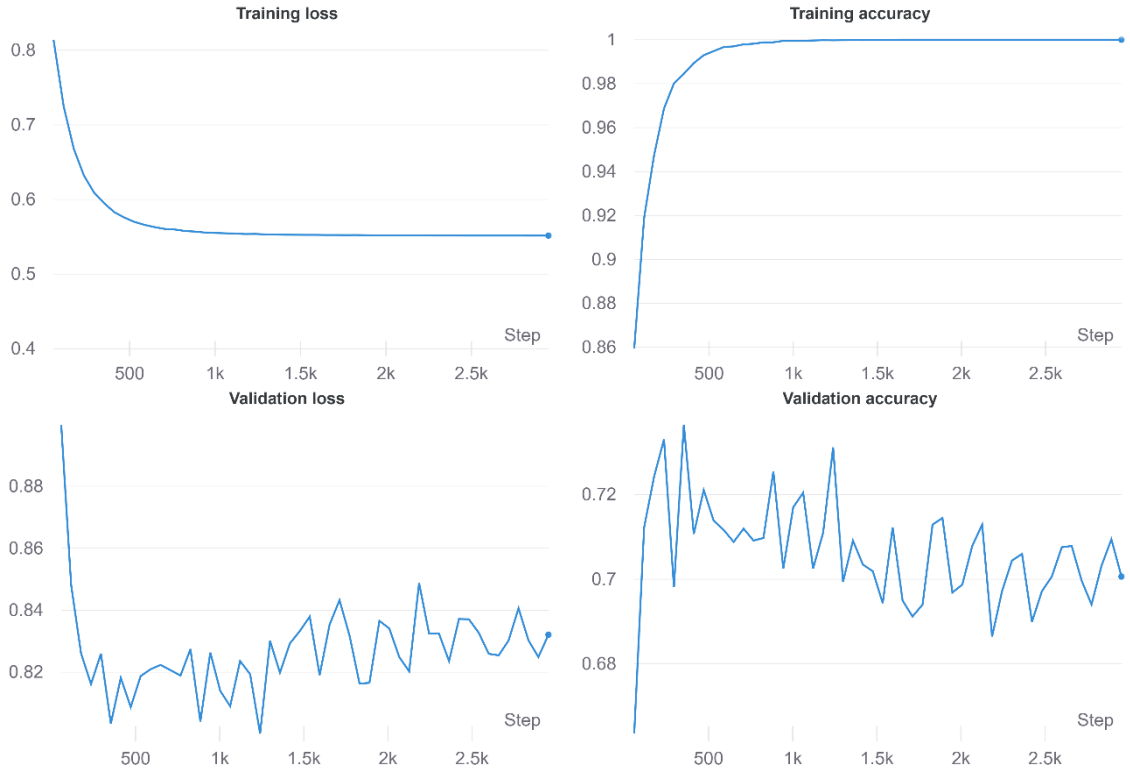


Figure 55 Accuracy and loss during training

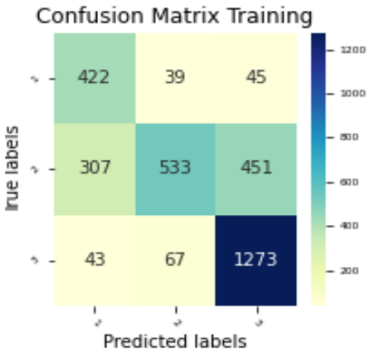


Figure 56 Confusion matrix

Gallbladder 3-class classification

Tabel 14 Evaluation 3 grade classification network (overall grade)

	Grade	Accuracy	Precision	Recall	F1
Overall Grade	1	0.58	0.39	0.64	0.49
	2		0.47	0.16	
	3		0.64	0.89	

Tabel 15 Hyperparameter configuration of the best performing network

	Overall Grade
Architecture	Resnet18
Classes	3
Epochs	30
Batch size	32
Learning rate	0.00008133
Dropout	0.4
Weight decay	0