University of Twente
&
Meander Medical Centre

Thesis Technical Medicine

Deep learning for identification of gallbladder leakage during laparoscopic cholecystectomy

Maria Henrike Gerkema
s1350080

Medical Supervisor:
Prof. Dr. I.A.M.J. Broeders

Technical Supervisor UT:
Dr. Ir. F. van der Heijden

Process Supervisor UT:
Drs. A.G. Lovink

External Member UT:
M.E. Kamphuis, MSc

Tuesday 7th July, 2020

UNIVERSITY OF TWENTE.

Abstract

This study aimed to develop a deep learning algorithm which is able to detect bile leakage in laparoscopic cholecystectomy video frames. The occurrence of bile leakage during laparoscopic cholecystectomy varies between 1.3% and 40%. Although complication rates due to bile leakage and lost gallstones are low, they are avoidable. More research into complications could be done if bile leakage is reported automatically, since studies showed that 13.0% till 73.8% of the bile leakages is not reported correctly. The purpose of this study is to achieve bile leakage detection rate that has clinical added value by having a reporting rate which is above the current 87% reporting rate.

In total 172 patients are included which laparoscopic cholecystectomies are performed by 23 different surgeons. The videos are derived from the Cholec80 dataset and from surgeries performed in the Meander Medical Centre. Video data is transformed to video frames and hereby 62380 bile and no bile leakage images are included in this study. Two convolutional neural networks and different parameters settings were used for creating an optimal bile leakage detection algorithm.

Training of the deep learning algorithm and testing of the trained network, resulted in a trained model which showed 83% sensitivity, 80% specificity and an AUC score of 0.91 for the testing dataset. The colour based feature extraction dataset achieved better results when comparing the best performing model with its no feature extraction version. However, the results were more ambiguous when both models and multiple training sessions are compared. The most important outcome is that this trained model currently does not have clinical added value when compared to the standards of reporting bile leakage in surgery reports in the Netherlands.

Although results should be improved by extending the dataset and optimizing the hyperparameters, good results are achieved by this study and first insights are given into bile leakage detection by using a deep learning algorithm.

Preface

This thesis is written to complete my master Technical Medicine at the University of Twente. The thesis was part of my graduation internship at the Meander Medical Centre. During my master internships, data-analysis caught my interest. The upcoming field of artificial intelligence and the collaboration of the Meander Medical Centre with a company which aims to apply artificial intelligence into healthcare, were important reasons to start my internship at this hospital. During an exploratory conversation, it was concluded that I could contribute to one of their goals of creating an environment for benchmarking of surgeons. By using a high volume surgery, namely laparoscopic cholecystectomy, a large video dataset would be available. Besides, this surgery is often performed by surgical trainees and a learning curve could be observed. Therefore laparoscopic cholecystectomy is a perfect fit for testing the use of artificial intelligence for benchmarking of surgeons. Eventually, this led to the subject of my thesis: deep learning for identification of gallbladder leakage during laparoscopic cholecystectomy.

During this internship a lot of people were of great help. At first, I wish to thank my supervisors Annelies Lovink, Ivo Broeders and Ferdi van der Heijden for their patience and good advise. Annelies, your encouraging words and 'twentse nuchterheid' during the last three years, helped me to continue my internships and finish my study. Ivo, thank you for your honesty and for being understanding with my energy level. I would like to express my appreciation to Ferdi for helping me in exploring the field of artificial intelligence. It was reassuring that you kept pointing out that artificial intelligence is an enormous field and one will always discover new things during research. It helped me to feel less incompetent. To the TM-students at the Meander Medical Centre, thanks for the overwhelming amount of nice coffee and lunch breaks.

I would not have been able to finish my study without the help of my dear family and friends, who were patience and took care of me during difficult times in recent years. At last, Folkert, thank you for your loving support and your stupid jokes that kept me laughing during the frustrating process of writing a thesis.

Table of Contents

List Of Abbreviations

Adam Adaptive moment estimation.
AI Artificial Intelligence.
AUC Area Under the receiver operating characteristic Curve.

CBFE Colour Based Feature Extraction.
CNN Convolutional Neural Network.
CVS Critical View of Safety.

DL Deep Learning.
DLC Difficult Laparoscopic Cholecystectomy.

EHR Electronic Health Record.

FE Feature Extraction.
fps frames per second.

L Leakage.
LC Laparoscopic Cholecystectomy.
LT Limited Time.

M1 Meander dataset that was created as first. It includes 70 videos of the total of 507 videos of the Meander dataset.
M2 Meander dataset that was created secondly. It comprises 50 videos of the total of 507 videos of the Meander dataset.
MMC Meander Medical Centre.

NoL No Leakage.
NVL No Visible Leakage.

PQ Poor Quality.

ReLU Rectified Linear activation Unit.
ROC Receiver Operating Characteristics.
ROI Region Of Interest.

TEP Totally Extra-Peritoneal.

CHAPTER 1

Introduction

This chapter discusses the clinical background of gallbladder leakage and laparoscopic cholecystectomy surgery, the risks for gallbladder perforation and an overview of previous studies into the use of artificial intelligence for laparoscopic cholecystectomy surgery. This will lead to defining the clinical problem, research questions and the aim of this study.

1.1 Gallbladder leakage

In the Netherlands, around 25,000 gallbladders are surgically removed by cholecystectomy every year [1]. Most common indications for surgery are symptomatic gallstones and complications due to gallstones like cholecystitis, jaundice and pancreatitis [2]. More than 30 years after the introduction of laparoscopic cholecystectomy (LC) by Mouret, the majority of cholecystectomies are performed laparoscopically. Two advantages of LC are shortened recovery time after surgery and decreased discomfort for patients [3]. Shortly after introduction of LC, increased numbers of complications of the major bile ducts and gallbladder leakage were reported [4–6]. Although complication rates vary between 1.3% and 40%, studies have shown that the switch to laparoscopic surgery resulted in increased gallbladder leakage [4–8]. During the early years of LC, gallbladder leakage was not considered as a harmful complication. After several years more and more case reports have shown that bile leakage and lost stones resulted in formations of abscesses and fistulas in the peritoneal cavity [5–8]. Although complication numbers after gallbladder perforation are low, they are avoidable [4, 5, 8]. To prevent complications due to unretrieved gallstones, it is advisable to retrieve as many gallstones as possible and wash the abdominal cavity to remove bile [5, 6, 8]. Currently, an important issue is the non-reporting of gallbladder leakage, the numbers vary between 13.0% and 73.8%. It is negatively influencing research to the incidence of gallbladder leakage and its complications. Especially when considering the combination of the wide range of non-reporting numbers and incidence numbers and the limited amount of articles about gallbladder leakage [4, 7, 9]. Patient safety is at stake since incomplete reports could result in delayed diagnosis of LC related complications and underestimation of complications during research [4, 6]. Therefore, correct reporting of gallbladder leakage and informing patients about possible complications, is advised. Aforementioned is required to gain insight into gallbladder leakage and its consequences [5, 6, 8].

To improve reporting of gallbladder leakage, the introduction of Artificial Intelligence (AI)

into healthcare could open up new perspectives. Combining Deep Learning (DL) and the search for complications during laparoscopic cholecystectomy will improve patient safety and research outcomes. The amount of gallbladder perforation during LC and postoperative complications decreases when gallbladder perforations are automatically reported; surgeons can learn from previous mistakes, patients are correctly informed about possible complications and study outcomes will become more reliable.

1.2 Defining gallbladder leakage

It is important to note that there are two different situations which both could be described by gallbladder leakage or rupture. The first one is when the gallbladder ruptures without any surgical intervention, this is a rare complication and not part of this study. The second situation is during LC by perforating the gallbladder by a surgical tool, which is researched in this study. Multiple terms are used to describe this form of bile leakage, namely leakage, spillage and gallbladder perforation. Bile spillage is when a minimal amount of bile is leaking out of the gallbladder. When a hole is present in the gallbladder and the bile and stones are pouring out, it is defined as perforation. Both could be described as bile/gallbladder leakage, but only the occurrence of gallbladder perforation could cause loss of gallstones. For this research, bile spillage and gallbladder perforation are included. It is not in the scope of this study to distinguish between severity of gallbladder leakage.

1.3 Laparoscopic cholecystectomy

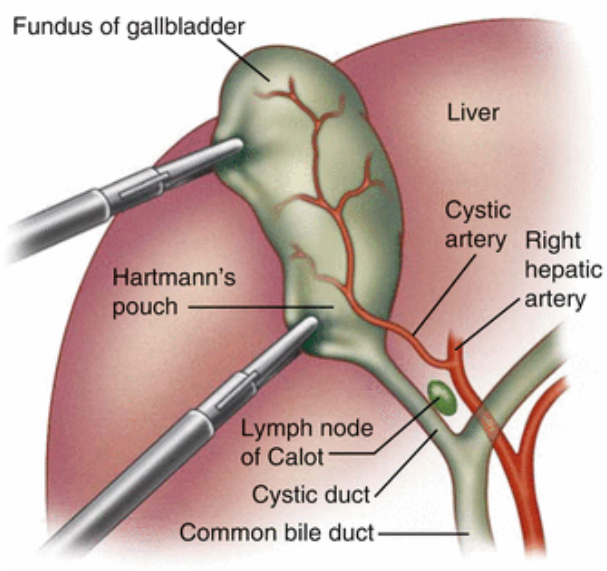
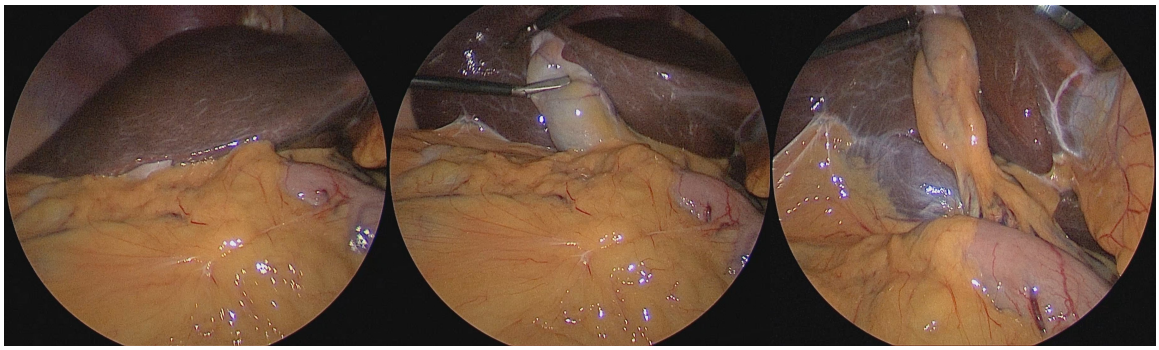


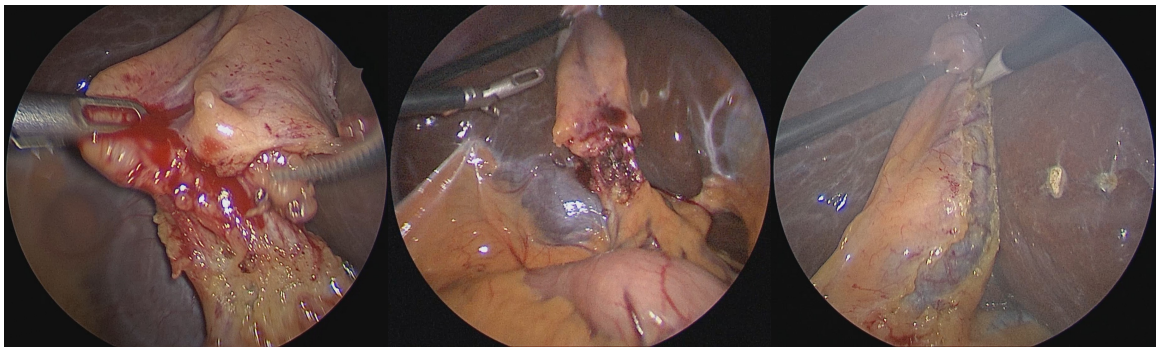
FIGURE 1.1: Anatomy of the gallbladder [10]

At the start of an LC procedure, the liver needs to be elevated to provide a sufficient overview of the gallbladder and other structures (Fig. 1.2A and 1.2B). It is done by using a fan retractor which lifts the right lobe of the liver [11]. It is important to lift the fundus of the gallbladder and give traction to the Hartmann's pouch to optimize visibility of the ducts and arteries (Fig. 1.1). These steps are also shown in figure 1.2C and 1.2D. Peritoneum, which is covering the cystic artery and cystic duct, is dissected to create a clear overview of these anatomical structures (Fig. 1.2E and 1.2F). It is essential to use a standardized method to identify the critical structures, also known as Critical View of Safety (CVS) [12]. It means that the cystic artery and cystic

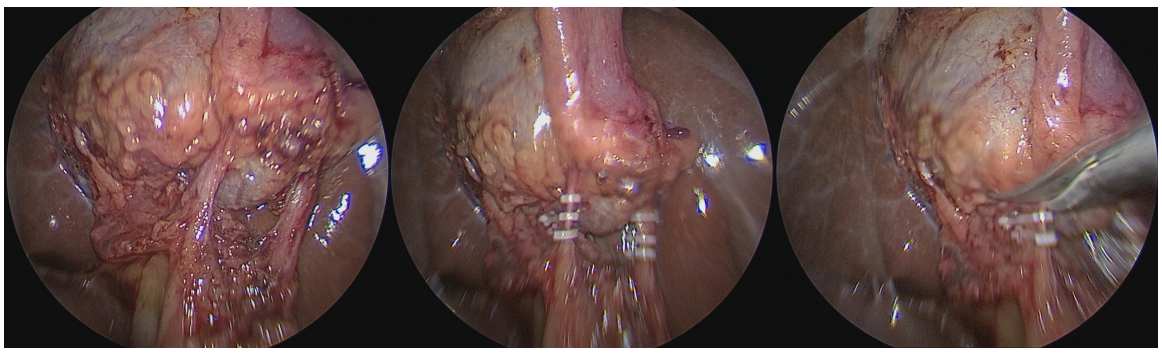
duct should only be dissected when both are clearly visible (Fig. 1.1). Identification of these structures and so the CVS is not always as straightforward as described. Both ducts and arteries show considerable variation in length and junction location. Therefore, this is a critical phase during surgery. If it is certain that the remaining structures, the cystic duct and cystic artery, are entering the gallbladder and the prescribed 360°view of both structures is possible, dissection of the cystic artery and cystic duct is safe (Fig. 1.1). The last step is to completely dissect the gallbladder from the liver bed which is already partly seen at figure 1.2F. Hereafter the gallbladder is removed out of the abdominal cavity by using a sterile plastic bag to prevent infections, bile leakage and lost stones [2, 11].



(A) Liver (B) Lifted liver (C) Stretching of fundus



(D) Hartmann's pouch (E) Surgery overview (F) Dissection of peritoneum



(G) Critical view of safety (H) Clipping of duct and artery (I) Cutting of cystic artery

FIGURE 1.2: Different phases during laparoscopic cholecystectomy

1.4 Risk factors for gallbladder perforation

1.4.1 High-risk surgery phases

Multiple studies have shown precarious phases during surgery with an increased risk of gallbladder rupture. Three phases were identified, namely when traction is given to the gallbladder with a grasper, which is occurring throughout the entire surgery. Additionally, dissection of the gallbladder from the liver bed is a procedure with an increased risk for rupture [6]. Impetuous dissection of the gallbladder from the liver fossa is mentioned as the most common cause of gallbladder perforation [5, 9]. Nooghabi et al. also mention retrieving the gallbladder out of the abdominal cavity as a high-risk procedure [6]. However, surgeons of the Meander Medical Centre (MMC) use a retrieval bag and prevent leakages of bile and stones when removing the gallbladder out of the abdominal cavity.

1.4.2 Difficult laparoscopic cholecystectomies

In addition to complications during difficult surgery phases, several articles describe predictive risk factors for gallbladder rupture. Patients who are at risk for gallbladder rupture are patients with gallbladder hydrops due to obstruction, chronic cholecystitis with thickened walls above 7mm and patients who previously received laparoscopic surgery [13]. Nooghabi et al. also mentioned male sex, higher weight, older patients and acute cholecystitis as risk factors. Since the study was retrospective, peroperative risk factors are determined: the presence of adhesions, challenging dissection of CVS, clip slippage and presence of infected bile and pigment stones [6]. Some of these factors are correlated: previous laparoscopic surgeries and the presence of adhesions, acute or chronic cholecystitis and infected bile. Besides the presence of (pigment) stones makes it more likely that there is obstruction. Some of these factors; male sex, older age, acute cholecystitis, spillage of pigment stones, number and size of stones and location of spilled stones, are also a predictive value for developing complications due to stone spillage [14]. All complications mentioned before are risk factors for gallbladder rupture. These partially correspond to risk factors for a difficult laparoscopic cholecystectomy (DLC). Risk factors for a DLC are impacted stone in a gallbladder neck, adhesions around the cystic artery and cystic duct and rupture of the gallbladder. Some identified risk factors, also define what a DLC is, namely injury of the cystic artery, blood loss above 50 mL and increased surgery time. When easy and difficult surgeries are compared, these risk factors are also significantly different [15].

1.4.3 Surgical experience

In the MMC, a high volume surgery like laparoscopic cholecystectomy, is often performed by surgical trainees and supervised by a surgeon. It is a suitable surgery to develop surgical experience. A potential risk factor is the correlation between surgical experience and number of complications. Two recent studies about gallbladder rupture and surgeons experience, estimated beforehand that complications could be correlated with surgery experience. Both studies did not find increased complication rates; only surgery time was increased [9, 15]. On the other hand, older studies found significant differences when gallbladder perforation was compared between experienced surgeons and surgical trainees [16, 17].

1.5 Artificial intelligence for LC

1.5.1 Previous Research

To improve reporting of gallbladder leakage, the introduction of Artificial Intelligence (AI) could improve quality of healthcare. Recently more and more papers are published about AI and laparoscopic cholecystectomy. One reason is that LC is a high volume surgery, resulting in a large

data set. Another important reason is the availability of two extensive datasets, Cholec80 and EndoVis, containing LC videos with annotation of surgery phase and used instruments [18, 19]. Thus far, these datasets are used for benchmarking, education, keyframe extraction and predicting the remaining surgery time. Other studies focused on combining these annotated datasets with external cameras or creating software for automatically annotation of data [18, 20–25]. Initially, studies focused on the improvement of results of previous studies about phase recognition and instrument usage [20]. These two recognition tasks are beneficial for the more difficult task of skill assessment. Benchmarking or skill assessment for surgeons has proven to increase their level of performance [20]. It is achieved by analyzing surgery steps and tasks, instrument usage and additional information about instrument path length, the number of hand motions, usage time of each instrument, applied force and how smoothly movements are [20, 21]. By evaluating these parameters, the learning process of (junior) surgeons is supported. More specifically, it enables personalized training, surgery evaluation and creation of skill-related feedback for (junior) surgeons [20]. Another promising subject is the study of Loukas et al. into keyframe extraction. They managed to extract 81% of the ground truth keyframes by using their trained network. This application is helpful for education, automatic generation of summaries for surgery reports and it could be used as a support tool for specific training for surgery phase and task recognition [22]. An innovative application of surgery phase information is the calculation of the remaining surgery time. When accurate estimation is possible, the preparations for the next surgery are more efficiently done by notifying staff automatically at the correct time. The use of surgery rooms and medical staff are optimally planned and more patients could be treated with the same healthcare budget and shortened waiting time [23, 25]. When the use of AI is extended to incorporation of the EHR and surgeon specific information, more accurate estimations could be made [25]. Padoy et al. describe the use of external cameras combined with surgery videos to extract more information about surgery phase and instrument usage. Although new information is added about the surgeons and medical staff's position and movement, it is still difficult to visualize all the members and movements and prove the added value of external cameras for patient outcome and surgery efficiently [23]. At last, a recently published article described the advantageous approach of automatic segmentation. Usually, this manual process is time-consuming because a medical expert manually annotates the videos. Bodenstedt et al. developed a method that only requires a limited amount of manual segmentations. Hereafter, similar regions in new data are detected by using a deep learning network and the probability of correct segmentation is calculated. Only segmentations with a very low probability for accurate annotation are verified and, if necessary, adjusted. Subsequently, all these segmentations are added to the training set and the next iteration starts. Hereby, a minimal amount of video frames needs manual segmentation and only the more complicated video frames will be annotated by an expert [24].

1.5.2 Research group Meander Medical Centre and Verb Surgical

In the MMC, different studies into AI and surgery are performed. The first project, the identification of five anatomical structures; ureter, tendon, artery, white line of Toldt and colon, was completed in August 2018. The next project aimed to remove video frames from surgery videos which contain personal information, most importantly, frames that contain medical staff. Verb Surgical, a collaboration between J&J and Google, is interested in this project, which is still ongoing. During multiple conversations, it was decided that a project about bile leakage during LC surgery would fit in their aim of creating a preoperative risk analysis for each patient, being able to estimate the remaining surgery time and offer benchmarking for surgeons. Another ongoing project is about identification of the Nervus Vagus. During anti-reflux surgery, the Nervus Vagus is injured in around 20 % of the patients. The goal of this study is to identify the nerve during surgery and support the surgeon in preventing collateral damage. Recently, a study about phase recognition during totally extra-peritoneal (TEP) repair started. Earlier

studies into phase recognition for LC surgery were used for this benchmark project. The goal is to give surgical trainees insight into their surgical skills. Since this operation includes many different steps, guidance during surgery and feedback per phase after a surgery, could be helpful. Eventually, the goal is to assist surgical trainees in learning to operate more systematically and focus training on specific phases which could be done faster.

1.5.3 Benchmarking

Although research is done into skill assessment, the development of a surgery robot by Verb surgical and their interest in AI opens up new perspectives. Besides improvement of skill assessment algorithms, there is a need for objective classification of the level of complexity of a surgery. As mentioned before, the definition of a DLC surgery is related to the health condition of the patient and the complications that occur during surgery. When it is possible to define what an easy, moderate and difficult LC surgery is, it is possible to determine if surgery times and number of complications are increased compared to other colleagues. Otherwise, increased mean surgery time and number of complications due to a lot of difficult patients, could incorrectly mark a surgeon as too slow or even incompetent. Combining the objective level of complexity of a surgery, surgery time, complications like gallbladder leakage and skill assessment, will result in fair benchmarking of surgeons and eventually improve healthcare.

1.6 Research questions and aims

1.6.1 Clinical problem

Although studies confirmed that gallbladder perforation could result in severe complications and they stated that it should be reported correctly, surgeons still do not consistently mention gallbladder leakage in surgical reports. Hereby, it is not possible to conduct a comprehensive study on the incidence of complications related to gallbladder rupture. Information about risk factors for gallbladder leakage is available. It is defined how surgeries could be classified as an easy, a moderate or a difficult LC. Besides, the possible effect of surgical experience is researched. Nevertheless, to confirm and combine these findings more reliable data is needed. To improve patient safety before, during and after an LC, more feedback and information should be collected.

1.6.2 Aim

The aim of this study is to detect bile leakage in videos of laparoscopic cholecystectomy surgeries. When the created deep learning network is outperforming the manual reporting of gallbladder leakage, the result is clinically relevant. Only then, the network is suitable for automatic reporting of gallbladder leakage in surgery reports and research into gallbladder complications will become more reliable. The ultimate goal for gallbladder surgery is that reliable preoperative risk assessment for each LC patient is done automatically before the surgical procedure by using previous mentioned high-risk factors. Besides, complications are detected during surgery and are reported automatically. Both surgeon and surgical trainees can learn from a gallbladder perforation, because data of perforation is annotated correctly and therefore available. Additionally, benchmarking, so comparing skills between surgeons, is possible and personalized training sessions will improve skill and speed during specific phases and procedures. Still most importantly, quality of care is improved when complications during laparoscopic cholecystectomies are reported correctly and patients are informed about possible postoperative complications.

The aim of this study, the identification of gallbladder leakage by using a deep learning network, will be a small contribution to this ultimate goal of improving quality of care for patients who receive a laparoscopic cholecystectomy.

1.6.3 Research questions

1. To what extent is it possible to detect bile leakage in laparoscopic cholecystectomy videos by using a deep learning algorithm?
2. What is the clinical added value of the deep learning network when comparing its bile leakage detection rate to the reporting rate of bile leakage in surgery reports?
3. How does the use of colour based feature extraction contribute to the gallbladder leakage detection rates in laparoscopic cholecystectomy video frames?

Primary objective: To detect gallbladder leakage post-operatively in laparoscopic cholecystectomy video frames by using a deep learning algorithm.

Secondary objective: To create an algorithm with a detection accuracy that has more clinical added value in comparison with current standards of bile leakage reporting in surgery reports, based on literature studies. Besides, a parameter study is performed to improve results and understanding of deep learning algorithms.

1.7 Outline of this study

During this study, five elements were carried out to create a working algorithm for laparoscopic cholecystectomy videos. At first, a parameter study is done to decide which network is suitable and which hyperparameters should be used and how they should be tuned. A second part of the study consists of creating an LC dataset with gallbladder leakage images out of the previous mentioned Cholec80 dataset. This enabled performing binary classification on a gallbladder leakage dataset. During this study phase, more information was obtained about tuning of hyperparameters and how to evaluate the model. The fourth element, colour based feature extraction, was performed on this dataset to decide if results of a deep learning network could be enhanced. At last, data was collected in the MMC to enable evaluation of previous performed network training and a larger dataset was created with Meander data and the Cholec80 dataset.

CHAPTER 2

Technical Background

In this chapter a brief introduction is given into deep learning and convolutional neural network architecture. Additionally, hyperparameters that were used during this study are explained. The third section of this chapter describes how network optimization could be performed. Hereafter, it is discussed how evaluation of deep learning networks could be performed. This chapter concludes with the introduction of feature extraction. This is used to reduce specific information in parts of laparoscopic cholecystectomy images and accentuate other elements in these images.

2.1 Convolutional neural network

A Convolutional Neural Network (CNN) is a specific type of deep learning network which is suitable for analyzing images. Three basic elements create such a network, which are: convolutional layers, pooling layers and fully-connected layers (Fig. 2.1).

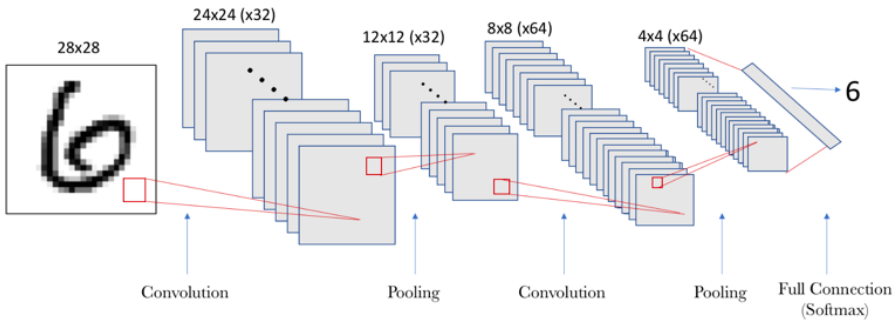


Figure 2.1: A convolutional neural network [26]

2.1.1 Convolutional layers

Convolutional layers are multiple neurons which operate as filters for the pixels of an image (input). The width of a network is determined by the amount of neurons or nodes in a layer and the depth of a network by the amount of layers. If a filter of size 5x5 moves with a step size

(stride) of one, the output (feature map) dimensions are downsized with four pixels. The input is processed by activation of neurons while moving the filter with specific weight over an image. The idea of a CNN is that there are a lot of these filters, 32 in first layer of Fig. 2.1, and each one of them is filtering another element because they have different weights. Hereby, different properties are detected for each image [27].

Activation

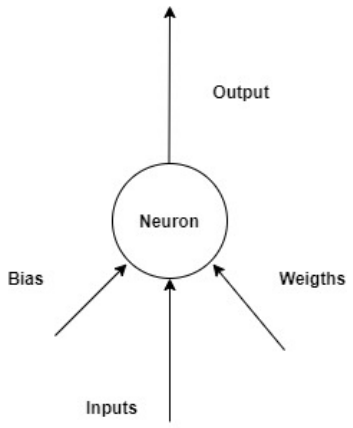


FIGURE 2.2: A network neuron

As described in previous paragraph, activation of a neuron is needed to process the input information. Inputs and bias are weighted and summed and an activation function will have a threshold which determines if the neuron is activated (Fig. 2.2).

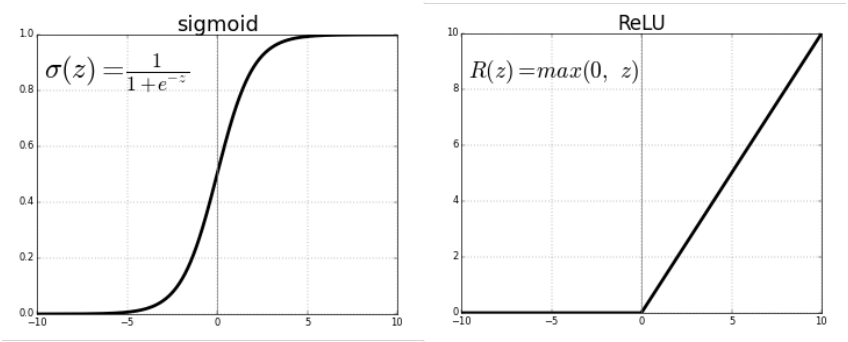


FIGURE 2.3: The sigmoid and ReLu function [28]

Nowadays, the two most used activation functions for binary problems are the rectified linear activation unit (ReLU) function and the sigmoid activation function at the end of a neural network. The ReLU activation function is based on the most simple activation function, namely a linear activation function. Since deep learning is often performed on complex data, the activation function not only needs to be adequate for this data, but also should be simple to enable less complex calculations. The ReLU is combining the linear activation function, but prevents that input below zero can activate the neuron and creates converging of the network towards zero (Fig. 2.3). This is important, because a neuron should not be activated if the weighed inputs will not contribute in the prediction of an outcome [27, 29]. Considering that the outcome value of a sigmoid function is between 0 and 1, probability predictions at the end of a network is often done by using this function. The last fully connected layer consists of one neuron with a sigmoid activation function. Hereby, outcomes for a binary classification problem

could be predicted with a cutoff value of 0.5. All values below 0 are assigned to one class and value of 0.5 and higher are assigned to the other class [27].

2.1.2 Pooling layers

The pooling layer is used to downsize the filters and thus lower the resolution and prevent overfitting. Otherwise, filters are created which are too specifically fitting the images. Besides, pooling layers provide feature maps which are more suitable for context recognition instead of detailed feature recognition.

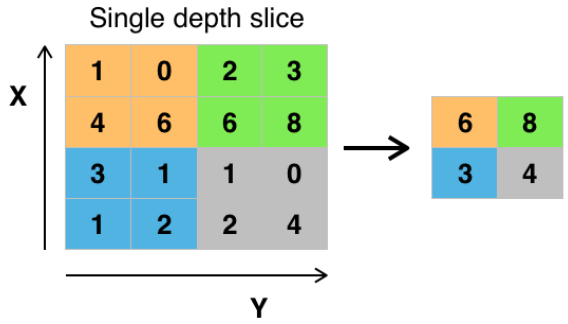


FIGURE 2.4: Max pooling

An often used pooling layer is maxpooling. It is a filter of size 2x2 which moves over each feature map with a stride of two. Only the highest pixel value of four pixels is kept. Hereby each feature map is reduced to one fourth (Fig. 2.4). It helps to prevent overfitting and less detailed feature maps are created for context recognition [27].

2.1.3 Flatten

The flatten layer is needed if data consist of multidimensional information. The use of a CNN enables working with colour images and this data is three dimensional since each pixel of these images has three colour channels (red, green, blue). A flatten layer transforms the three dimensional data into one dimensional data. For example, a None by 3 by 16 input is transformed to None by 48. A flatten layer allows the use of the fully connected layers as next layer in a network and this layer is needed to obtain predictions as output [30].

2.1.4 Fully connected layers

At the end of a network, fully connected layers are needed to combine information obtained in previous layers. These last layers of a network will predict what each image contains. In python language this is called a Dense layer. The activation functions that are used are the ReLU activation function and the sigmoid activation function in the last layer to have a final result between 0 and 1, which is useful for making predictions [27].

2.2 Network hyperparameters

When implementing a network, there are many options for the settings, also called hyperparameters. Network hyperparameters define the network structure, while optimizer hyperparameters will determine how training of a network will be done. Network hyperparameters are the number of layers and units in each layer, the use of dropout, the network weight initialization and the activation function which is explained before. Training hyperparameters are batch size, number of epochs, optimizers, loss functions, learning rates, momentum and learning rate decay [27, 31].

There are several goals that needs to be considered when creating a network: convergence, precision, robustness and the general performance of a network. An ideal network converges quickly to optimal hyperparameter settings. Large precision results in outcomes that are close to the reference outcome. Optimal robustness will create a trained network which generalizable to other LC datasets. It is difficult to create a perfectly performing network, so even for a well performing network, hyperparameters should be chosen carefully [32].

2.2.1 Epoch and batch size

Running through the entire training dataset once is called an epoch. The majority of datasets are too large to run at once, and running one by one makes it difficult to create a stable training of the network due to noise. Therefore, large datasets are divided into smaller parts called batches. To train a network, iterations of epochs are done tens and sometimes hundreds of times. [27]. It is important that the batch size is chosen carefully. A larger batch size means faster training, since one epoch is only a few batches and learning process is faster. But one needs to take into account that this means that when images are used, an image batch is loaded at once. There is a computational limitation for the GPU of a computer. On the other hand, when a batch size is too small it could induce overfitting of the model, since filters are trained too specifically when more feedback is given during training. Therefore a trade off needs to be found between a large batch size, but small enough to be loaded at once. Lastly, an important criteria for the batch size is that it needs to be a power of two to meet the memory requirements. In this way, calculations are done most efficiently [27, 33, 34].

2.2.2 Gradient descent optimization

To improve training results, the training output is compared to the reference outcome by using a gradient descent optimization algorithm. After training of a batch, the error between predicted output and reference output is calculated. The weights of each filter are updated based on the contribution of those filters to the error. This is called backpropagation. Updating is done by partial derivative computations to calculate the contribution of each layer to the error and hereafter use this outcome to calculate contribution to the error of the previous layer and so on. The purpose of this updating is to minimize the error by adjusting the weights of filters and find optimal parameters for a model. Updating of parameters could be done after running the entire dataset (batch gradient descent) or one-by-one (stochastic gradient descent). The earlier mentioned batch size, also called mini-batch, can help to train a large dataset faster. More importantly, a more precise model is created and results will improve. Efficient updating is achieved when using mini-batch gradient descent which updates the model after each mini-batch. Hereby, the advantage of batch gradient descent is used, namely stable updating with accurate prediction of the error. On the other hand, by using a batch-size closer to one, the advantage of stochastic gradient descent is used and efficient calculation is done with less computational power. [27, 34–36].

2.2.3 Loss function

The prediction of the error for updating by mini-batch gradient descent is done by an error function, also called loss function. For binary classification of gallbladder leakage, the binary cross entropy loss function is the most common choice. The loss is a maximum likelihood estimate expressed by the loss function. This function calculates the mean difference between predicted output and reference output, for which an optimal outcome is zero. So, an optimal situation is when the loss calculation becomes zero or close to zero. Thus, when a maximum likelihood estimate is performed, updating the weights by using the loss function is done to find model weights for which the predicted output is most resembling the reference class. This

method is called binary cross entropy, because the difference between predicted output and reference output is expressed in bits. [37, 38].

2.2.4 Weight initialization

Weight initialization is useful to prevent vanishing or exploding gradients. Backpropagation by the partial derivatives will be more unstable if each derivative of the layers is large, complexity of the weight update calculations increase and gradient is larger after each layer. Hereby the training is slowing down, since weight updating is taking more time. When derivatives are too small, the gradient is small and gets smaller after each layer and converge towards zero. This will slow down learning, since updating weights is only done by very small steps and it will take more time to find an optimum for the weights. [39]. Initializing all weights with the same value, creates filters with roughly the same property, which will limit optimal learning. By random initialization of the weights which are not too small or large, the learning process will be improved. For a ReLU activation function, an often used weight initialization method is the He Normal or He Uniform initialization [40].

2.2.5 Optimizers

An optimizer uses backpropagation, but other parameters are needed to improve optimization. For most optimizers, these other parameters are momentum, learning rate and learning rate decay. All available optimizers combine these parameters in different ways and will perform differently.

Momentum

Momentum is used to move the gradient vector in the correct direction and decreases oscillations. This is achieved by using the vector of the previous updates whereas most recent gradient updates are more important than older vectors. When updates proceeds in the same direction to a minimum or maximum, the use of momentum will accelerate this process. This is achieved, because the direction of the vector of the most recent updates are in the same direction and added to the current vector. Fluctuations of the gradient are reduced, since a more average gradient vector is used by combining current vector with previous vectors. Small changes in direction are prevented and a more smooth curve of the learning process is accomplished. Hereby, the optimization process is improved [40].

Learning rate

A learning rate is needed to determine how much the current weight of a filter changes by the loss calculations. When choosing a large value for learning rate, the weights can change rapidly which could create an unstable learning process or less suitable weights. Contrarily, smaller learning rates could result in more accurate adjusting of the weights, but a very slow learning process. [27, 41]

Learning rate decay

Learning rate decay is added to a network to combine positive fast converging with a large learning rate and the more precise tuning with a smaller learning rate. The network will learn fast at the beginning of training and when learning proceeds, only fine-tuning is allowed, so adjustments to weights are limited. This will speed up the process of finding suitable weights and creating a suitable model. This decay is done by using a learning rate schedule which changes the learning rate based on time, amount of epochs or the current performance during training [27, 41].

Adaptive Moment Estimation

The Adaptive Moment Estimation (Adam) is the most used optimizer for neural networks at this moment. It is a combination of RMSprop and momentum and provided by Kingma et al. [42]. RMSprop is an optimizer which divides the learning rate, η, by a squared decaying average of the previous gradients. This learning rate decay is accomplished by using variable β_2 which is getting smaller during training. Therefore, the learning rate will be larger at the beginning and smaller at the end of training, which will slow down training. Momentum is added by variable β_1 to accelerate the weight update in the right direction. The update equation of Adam is given in Eq. 2.1 [42]. θ is the weight update parameter, ϵ is a small value which prevents that η is divided by zero. m_t is defined in Eq. 2.2 and v_t in Eq. 2.3. These equations show how Adam is updated by parameters β_1 and β_2 [39, 40, 42].

$$\theta_t = \theta_{t-1} - \frac{\eta \cdot \hat{m}_t}{(\sqrt{\hat{v}_t} + \epsilon)} \tag{2.1}$$

$$m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t \tag{2.2}$$

$$v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2 \tag{2.3}$$

Bias correction is performed for the possibility that moment estimates move towards zero if β gets close to one (Eq. 2.4 and Eq. 2.5) [42]. For bias correction, the m_t and v_t are divided by $(1 - \beta^t)$. That is why the $\hat{m}_t$ and $\hat{v}_t$ are mentioned in Eq. 2.1 instead of the earlier defined m_t and v_t [42, 43].

$$\hat{m}_t = \frac{m_t}{(1 - \beta_1^t)} \tag{2.4}$$

$$\hat{v}_t = \frac{v_t}{(1 - \beta_2^t)} \tag{2.5}$$

2.3 Network optimization

2.3.1 Model complexity

Multiple layers and more neurons create a deeper and wider network. This enables solving of more complex data. When creating and testing a network, it is important to notice whether training results are converging to lower loss and higher accuracy. When overfitting occurs, perfect results on training data are achieved but too many neurons are used. As a result, every neuron learns only a small piece of the data and achieve high accuracy on the training set, but it is not flexible enough to interpret new data. On the other hand, when a network is too complex for a dataset, too many layers are used. Not enough information is present in the dataset to accurately train all neurons by the training examples.

2.3.2 Using validation and test set

When training a network, the data will be divided in three different groups, namely a training, validation and test set. The training set is used to train a network. After training a mini-batch, the validation set is used to check how the network is performing and how parameters should be updated. After training of a network, a test set is used, which is a new dataset, to check how the final network is performing on new data. After training, accuracy and loss are stored, multiple graphs are created of the training session and settings of the created algorithm are also stored. By using the validation set, it is tested how well the network is training. If only a training set is used, overfitting can occur since adjustments to the weights of the network will be done based on training data itself. When creating a validation set, a disadvantage is that a part of the data is not used for training of a network. Since annotated data is costly, you want to

reduce the validation set as much as possible, but still obtain optimal feedback during training. An often described distribution is 80% of the data for the training set, 10% validation set and 10% for testing. set [33].

2.3.3 Performance evaluation during training

To monitor the progress in network training, loss and accuracy of the training and validation set are useful parameters. The loss function is the sum of errors during one training iteration of a mini-batch. The accuracy shows the rate of correctly identified reference outcomes. When training a network, these parameters could be monitored to quit a training when accuracy and loss are not improving anymore. This is called early stopping. It is a useful addition to a network, since time is saved and unnecessary calculations are prevented [33, 44]. Besides, model checkpoint could be used to save the weights of the model. To avoid accumulation of files, only the best model is saved during training. Hereby, it is possible to reload the network weights and use this for testing of the final model by using test data. Besides, when an error occurs, loss of valuable training time is prevented [27].

2.3.4 Dropouts

Another clever tool in deep learning is the addition of dropouts which reduces overfitting of the model. This procedure leaves out one or multiple neurons during an iteration. Hereby, the weight updates will not be applied to these neurons and connected neurons in previous layers. During training, each layer and neuron tends to specialize in specific feature detection. By leaving neurons out for one training iteration, other neurons need to anticipate which results in less specialized neurons and hereby prevent overfitting [27, 35].

2.3.5 Batch normalization

Batch normalization is useful because of the internal covariate shift. Ioffe et al. formulate this as followed: “We define internal covariate shift as the change in the distribution of network activations due to the change in network parameters during training” [45]. This occurs during backpropagation after each batch, by which the weights of a neuron and the contribution of inputs to each layer changes. These changes are more difficult to predict when a neural network has more layers. After each layer, it becomes more difficult to predict the contribution of the following layer. Hereby, weights could become very large or small after multiple epochs. To simplify backpropagation, batch normalization is applied. Each input is standardized in order that the mean is zero and standard deviation is one. It will create smaller weight changes, while nonlinear relations between layers remain and the effect is that a more predictable network for backpropagation is created. One advantage is that a larger learning rate could be used and hereby network convergence is going faster. This will speed up training, since significant less epochs are needed. Another advantage of batch normalization is that less dropout is needed and weight initialization is less important, since batch normalization prevents exploding or vanishing of the gradient. Consequently, less dropout means that more data could be used during training [45–47].

2.3.6 Data augmentation

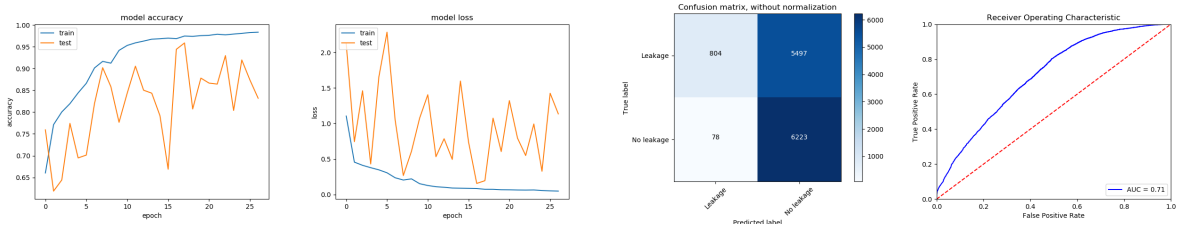
If images in one dataset show similarities or only a small dataset is available, data augmentation is a suitable solution. Images of the training dataset are adjusted to create a more diverse dataset, but these adjusted images are still representative for the initial purpose. A few examples of adjustments that could be made are: rotation, flipping, brightness adjustments, zooming and whitening [27]. It has to be considered that not all data augmentation techniques are useful for training of a specific dataset. In case of LC videos, 180 degree flips of video images during

surgery are rarely seen, since videos are made by keeping the horizon as stable as possible. This needs to be considered when implementing data augmentation [48].

2.4 Evaluation of the model

2.4.1 Plots

The evaluation of the model results into a Receiver Operating Characteristics (ROC) curve, an Area Under the receiver operating Characteristic (AUC) value, a confusion matrix, and accuracy and loss plots. The accuracy and loss plots will help to observe training progression and evaluate how training is performed. The ROC curve shows how classification is performed with different trade-offs between sensitivity and 1 - specificity [49]. The AUC value shows the probability that an image is classified correctly. If the AUC value is between 0.70 and 0.80, it is an acceptable outcome, between 0.80 and 0.90 is good and higher than 0.90 is an outstanding result. For clinical use, an AUC value above 95% is preferred [50]. The confusion matrix is a table with the true positives and negatives and false positives and negatives. Besides, previous mentioned plots the sensitivity, specificity and specificity of each model are calculated [49].



(A) Accuracy (B) Loss (C) Confusion matrix (D) ROC curve

FIGURE 2.5: Model evaluation plots

2.4.2 Accuracy and loss

Trade-off for optimalization

When comparing the accuracy and loss of models, it is important to realize why it is impossible to create a perfect model. When loss calculation is done by binary cross entropy, it is a balance between incorrect assumptions of the model, so an imperfect model, and by learning information of the dataset too well. This overfitting will occur when training continues for too long, because it is impossible to have a complete dataset which takes all anatomical variations of patients into account. The balance between both is a trade-off, so it is not possible to reach optimal values of zero for both. Less mistakes by the model, requires longer training. While less overfitting demands shorter training of the dataset [36]. When accuracy and loss values are obtained after training, the size of this trade-off between a well trained model and overfitting will determine how the model is performing.

Relative uncertainty of loss and accuracy calculations

Although low loss and high accuracy and no overfitting could indicate that the model is performing well, an important note is that machine and deep learning accuracy calculations will always have an uncertainty. Calculations can be done to estimate the minimal dataset size to achieve a specific accuracy when taking a relative uncertainty into account [33]. Van der Heijden et al. describe Eq. 2.9 that estimate the needed dataset size when this relative uncertainty level γ is included [33]. Since $\hat{E}$ is the the estimated error rate, the writers combine Eq. 2.6, 2.7 and 2.8 and assume that $\hat{E}$ is close to the true error rate E. When combining the uncertainty of E and $\sigma_{\hat{E}}$, γ is fixed as combination of the uncertainty of both [33].

$$\hat{E} = \frac{n_{error}}{N_{test}} \tag{2.6}$$

$$\sigma_{\hat{E}} = \sqrt{\frac{(1-E)E}{N_{test}}} \tag{2.7}$$

$$\gamma = \frac{\sigma_{\hat{E}}}{E} \tag{2.8}$$

$$N_{test} = \frac{1-E}{\gamma^2 E} \tag{2.9}$$

By choosing a fixed γ and using the error rate, it could be determined what dataset size is desirable. Another method how it could be used is to calculate the uncertainty of test outcomes when dataset size and a specific error rate is known [33]. To conclude: the earlier mentioned trade-off and relative uncertainty should both be taken into account when comparing prediction outcomes of a model. The outcome will always be based on statistics. One of the reason why implementation of deep learning into clinic is difficult, is because a deep learning model is a statistical model and outcomes will never be perfect or 100% certain.

2.5 Colour based feature extraction

Image data contain a lot of information, since an image of 100x100 already contains 10.000 pixel values. When humans look at a picture, often only a small part of the images is relevant. When training a network, the same efficiency could be achieved by leaving out irrelevant image data and decrease unnecessary calculations. One method to do so, is feature extraction (FE). When FE is performed for data which contain two classes, best results could be achieved when the difference between two classes is substantial and less noise is present. This difference between two classes is expressed by the inter/intra class distance. The inter class distance represent the distance between two classes in the feature space, while intra class represents distance within one class. In Fig. 2.6 four classes are shown, but these inter and intra class distance principles are the same. Since colour images contain three features, namely a blue, green and red channel, FE these could be used to find colour differences between two classes. When colour based feature extraction (CBFE) is used on colour images, the optimal linear combination of color channels is found.

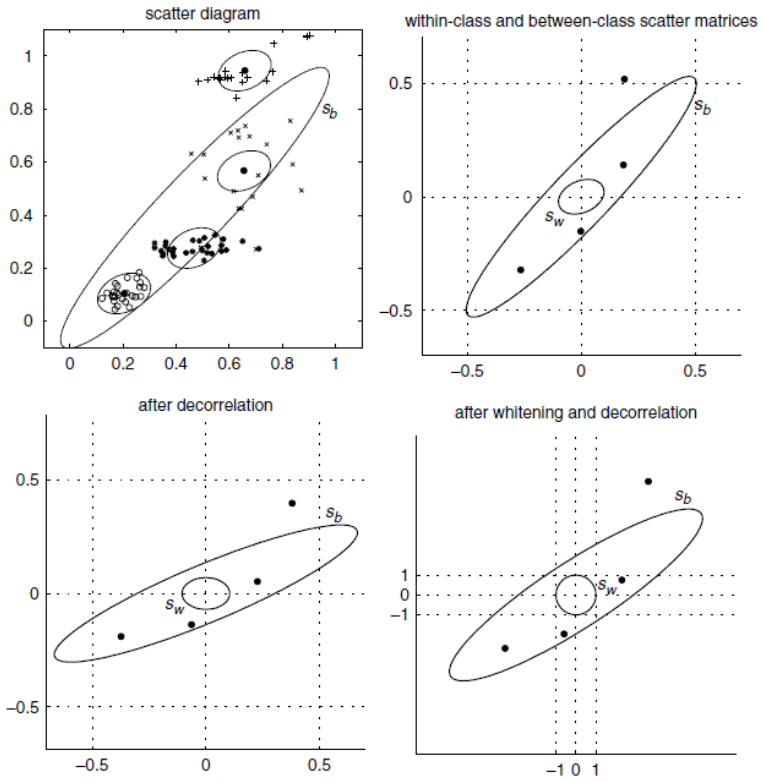


FIGURE 2.6: Inter/Intra class distance [33]

To find this optimal combination, which is distinguishing most accurately between two classes, intra-class whitening is applied (Fig. 2.6). Intra-class whitening normalizes the data within a class whereby the mean of the samples is centered in zero and standard deviation is one. Hereafter inter-class decorrelation is performed. It means that colour differences between classes are exaggerated and the optimal combination is searched for which leakage and no leakage images differ the most (Fig. 2.7).

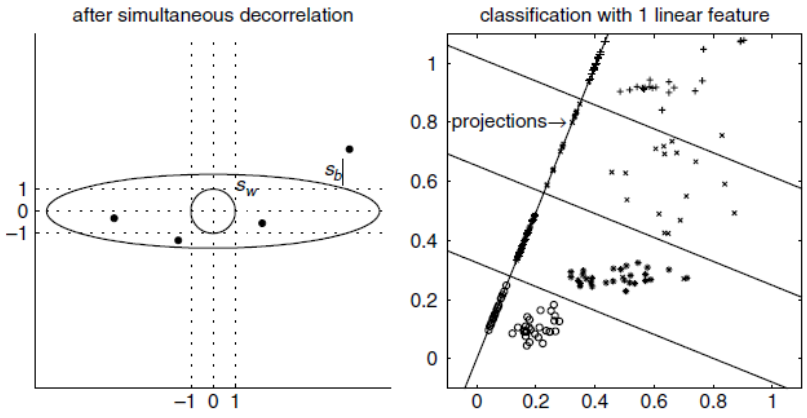


FIGURE 2.7: Decorrelation [33]

In the left images of Fig. 2.7 it is seen that after decorrelation and whitening, the x-axis is more valuable for classification of four classes compared to the y-axis. When three of these diagrams are created for colour images with two classes, one combination of two features will result in a highest inter class variability and the third feature should be nullified. [33]. Thus, inter/intra class decorrelations could be used to determine which two RGB values contain

valuable information for classification. By doing so, a transformation matrix could be created which transforms other images into a decorrelated and whitened version. This may help to increased classification rate.

CHAPTER 3

METHODS

3.1 Data Preparation

3.1.1 Explanations of different datasets

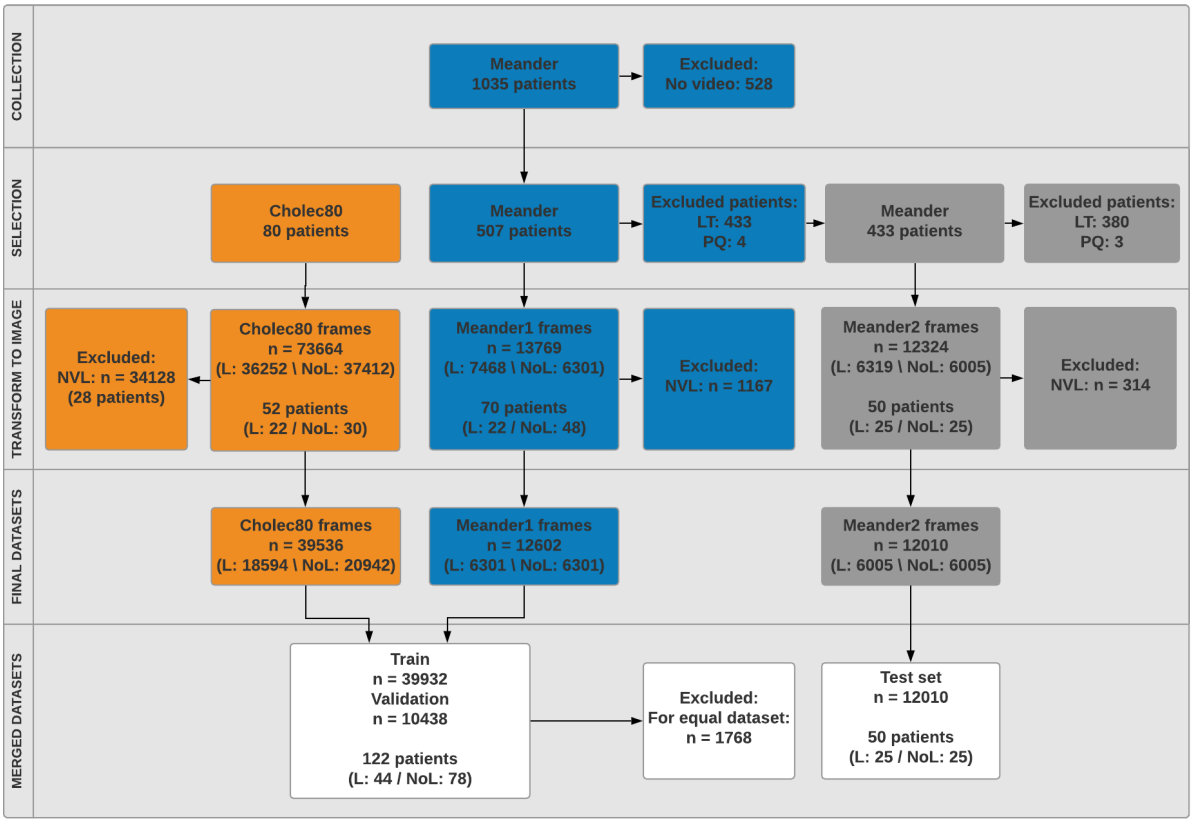


FIGURE 3.1: Flow chart of creating the datasets. Exclusion criteria abbreviations are No Visible Leakage (NVL), Limited Time (LT) and Poor Quality (PQ). The letter n is an abbreviation for number of frames

Five different steps were taken to create three datasets (Fig. 3.1). The first dataset was the collection dataset and this only included downloading of videos. During the selection step, videos were excluded. For the Meander dataset, limited time (LT) and poor quality (PQ) resulted in exclusion of videos. The third dataset (transform to image) was created after exclusion of unsuitable images. In the 'final datasets' step, three final datasets were created. In the last step, the Cholec80 and Meander1 dataset were merged to create one large training and validation set. The Meander2 dataset is used for testing after training the Merged dataset. In total, 62380 video frames of 172 patients are included in these three datasets.

3.1.2 Study population

The study population consists of two groups of patients. One dataset, the Cholec80 dataset, comprises 80 videos of laparoscopic cholecystectomy surgeries performed by 13 surgeons and this dataset is compiled by Twinanda et al. [18]. The second dataset, the Meander dataset, consists of 507 patients who underwent laparoscopic cholecystectomy surgery in the MMC between 01-01-2018 and 31-12-2019. These surgeries are performed by 15 different surgeons. By combining the Cholec80 and a part of the Meander dataset, a large dataset is created of LC surgeries. Due to lack of time, not all videos were included for this study, but data was stored to contribute to research in the future. The LC videos of 120 Meander patients and 52 Cholec80 patients are included, which are performed by 23 surgeons.

3.1.3 Collecting the video data

The Cholec80 dataset is provided by a research group of Nicolas Padoy, professor at the University of Strasbourg, and contains 80 LC videos and additional tool and surgery phase annotations. By filling in a request form, this data could be obtained. Only the 80 videos were used for this study [18].

Permission of the board of the MMC for collection of patient data was received after a research protocol was submitted and approved by the research committee (Appendix A). The Meander data was collected by using a form with dates of LC surgeries. By using the surgery planning, 1035 patients could be identified and their Electronic Health Record (EHR) was checked for existence of an LC video. These videos, which are made during surgery by using the laparoscopic camera, are downloaded from the EHR. For 507 patients, an LC video was available and therefore 507 videos were downloaded (Fig. 3.1).

3.1.4 Selection of videos for dataset

Initially, the LC dataset consisted of only Cholec80 videos. First training results of the Cholec80 dataset were not sufficient and therefore all videos were checked for visibility of gallbladder leakage. Hereafter, the Cholec80 dataset consisted of 52 videos (Fig. 3.1, transform to image lane). This review process of the dataset resulted in three exclusion criteria. Not all video frames with leakage should be included, a small amount of bile was too difficult and image quality should be sufficient. Besides, the python script is not able to split the video into a short video if there is less than 20 seconds in between start and end time. Therefore, visible leakage for less than 20 seconds was excluded. The inclusion criteria are based on the availability of videos and to define a time frame for inclusion of a suitable amount of surgeries.

Inclusion criteria:

- Videos of laparoscopic cholecystectomy
- Underwent surgery between 01-01-2018 and 31-12-2019 (Meander data)

Exclusion criteria:

- If gallbladder leakage occurs, but is difficult to identify
- If image quality is too low
- When video is too short (<20 seconds) or contains no valuable information

Selecting and annotating videos is time consuming. Therefore only 120 videos of the 507 videos of the Meander dataset were included in the Meander1 (M1) and Meander2 (M2) dataset. After the M1 dataset was created, the 433 remaining videos were used to created the M2 dataset. Seven videos were excluded during the selection process because of poor quality of the video (PQ), gallbladder leakage was not visible enough (NVL), the video was too short, the period of bile leakage was too short to create useful frames, only a small part of surgery was visible or the video did not contains surgery footage. 380 videos were excluded because of limited available time (LT). So the Meander database still comprises 380 videos which are not used for training, validation or testing (Fig. 3.1, selection lane).

3.1.5 Transform videos to image dataset

To create a training set, annotation of images is needed. By noting the timestamps of the first and last video frame with gallbladder leakage, suitable video frames can be selected. For the No Leakage (NoL) dataset, the timestamps are selected based on the surgery phase. Shortly before surgery starts is defined as the start time. When the gallbladder is dissected of the liver bed or the gallbladder is in the retrieval bag, this is assigned as the end time.

A script is used which creates short videos of the previous determined timestamps and subsequently splits these videos into images. For the Cholec80 Leakage (L) videos the parameter number of frames per second (fps) is 25. For the NoL videos, the frame rate was adapted in such a way that every video was transformed in approximately 690 video frames, which created a dataset of the same size as the leakage dataset. This is necessary to create a balanced dataset. Hereby, two different groups of images are created. One folder with bile leakage and one without bile leakage.

The M1 dataset and M2 dataset are created with an almost identical script as used for the Cholec80 dataset. The important difference is that Meander videos did not have the same fps. Therefore, the fps parameter is calculated per video and this is used to calculate the number of frames per short video and a suitable frame rate for extracting the images from the videos. Since these datasets are used as prediction datasets, a lot of resembling images due to a higher frame rate, would not give other results. For the M1 dataset and M2 dataset, five frames per second were included in the dataset. The maximum number of frames per video needs to be calculated, because the frame selection did not stop if the input 'end time' is not exactly synchronized with the time of the last frame of each video. For the NoL dataset an additional calculation was done to determine the optimal frame rate by taking the total time of the extracted shorter videos and create the frames with the same fps for each video. Hereby, a more comprehensive dataset is made, since short videos contribute less images to the dataset. So it is prevented that a lot of similar images of a short video are introduced into the Meander NoL datasets.

3.1.6 Selection of video frames

When video frames are created, all images should be checked. If the bile leakage was disguised by tools, tissue or surgical smoke, the frame was excluded. When defining gallbladder leakage, the definitions of spilling and perforation are used. At the start of spillage of bile, small amounts of bile are hardly visible. Besides, image quality could be low or lighting insufficient, which necessitate exclusion of these video frames.

Selection of the Cholec80 frames

Initially, the LC dataset consisted of the 80 Cholec80 videos which resulted in 73664 video frames (36252 L, 37412 NoL). The previous described checking of images for visibility of bile leakage, was carried out, but too difficult images were included. This means that little bile was visible, but the frames were still included into the dataset. As a consequence, first training results of the Cholec80 dataset were not acceptable and all videos were checked again for visibility of gallbladder leakage. Hereafter, 39536 frames (18594 L, 20942 NoL) remained in the Cholec80 dataset (Fig. 3.1). These frames are from 52 patients (22 L, 30 NoL).

Selection of the Meander1 dataset

Since the Cholec80 dataset was corrected after a first training, the Meander dataset is created based on the selection criteria that were used during the correction of the Cholec8o dataset. These are the same for video inclusion and exclusion. After splitting the videos into frames, the L dataset included 7468 images. After checking the images, 6301 frames remained. Hereafter, the NoL dataset was created by using the previous described calculations and therefore this dataset also contains 6301 video frames (Fig. 3.1). After checking the images, 70 patients were included (22 L, 48 NoL).

Selection of the Meander2 dataset

The M2 dataset is created with the same method as the M1 dataset. After splitting the videos into frames, the L dataset includes 6319 images. After checking the images, 6005 frames remained. Hereafter, the NoL dataset was created by using the previous described calculations and therefore this dataset also contains 6005 video frames (Fig. 3.1). 50 patients are included in the M1 dataset (25 L, 25 NoL).

3.1.7 Merged dataset

The Merged dataset was created by combining the Cholec80 dataset and the M1 dataset. The M2 dataset will be used as test set. 1768 images were excluded to create two balanced datasets for leakage and no leakage for the Merged training dataset (39932 frames) and validation dataset (10438 frames). 122 patients are included (44 L, 78 NoL).

3.2 Parameter study

3.2.1 Dataset

To start training a network, a standard dataset was chosen with images of cats and dogs to investigate how different parameters did influence the accuracy and loss during training of a model. This training dataset contained 8000 images (4000 cats, 4000 dogs) and the test set contained 2000 images (1000 cats, 1000 dogs).

3.2.2 Network architecture of Model 4 and used hardware

Table 3.1: Network architecture of Model 4

Type	**Filters**	**Size / Stride**	**Dropout**	**Output**
Convolutional layer	32	3x3		64 x 64 x 32
Batch normalization				64 x 64 x 32
Dropout			0.2	64 x 64 x 32
Convolutional layer	64	3x3		64 x 64 x 64
Max Pooling		2x2/2		32 x 32 x 64
Batch normalization				32 x 32 x 64
Convolutional layer	64	3x3		32 x 32 x 64
Batch normalization				32 x 32 x 64
Dropout			0.2	32 x 32 x 64
Convolutional layer	128	3x3		32 x 32 x 128
Max Pooling		2x2/2		16 x 16 x 128
Batch normalization				16 x 16 x 128
Flatten				None, 32768
Dropout			0.2	None, 32768
Type	**Units**		**Dropout**	**Output**
Dense	256			None, 256
Batch normalization				None, 256
Dropout			0.2	None, 256
Dense	128			None, 128
Batch normalization				None, 128
Dropout			0.2	None, 128
Dense	1			None, 1

Four different models were used to study the influence of different model architectures. Table 3.1 shows the model that was used for final parameter testing. Other models contained less convolutional layers, no batch normalization or no dropout. A windows 10 pc with an NVidea GPU was used for training and testing of the models. Python was used with a deep learning environment which contained all packages that are needed to run the deep learning scripts, like Keras and Tensorflow [51, 52].

3.2.3 Network parameters

The following parameters were tested during training: batch size, optimizers, early stopping with different patience values and data augmentation methods.

3.2.4 Evaluation of the study

A script is created which automatically stores the training and validation information that is obtained after training. The following parameters were stored: accuracy and loss of highest accuracy and lowest loss, for both training and validation set. Additionally, number of epochs for both highest accuracy and lowest loss, batch-size, patience which is used for early stopping, the optimizers and data augmentation options are stored. At last, the file location of the accuracy plots, the loss plots and the best weights of the model, were stored in the excel file.

The accuracy and loss of both training and validation were monitored during training. This information was plotted in two graphs which shows how training of the model is executed.

3.3 Laparoscopic cholecystectomy dataset

3.3.1 Network architecture

Two models are used during training with the LC dataset, namely Model 3 and Model 4. Model 4 was already used during training with the parameter study dataset. Its network architecture is showed in Table 3.1. The network architecture of Model 3 is showed in Table 3.2.

Table 3.2: Network architecture of Model 3

Type	**Filters**	**Size / Stride**	**Dropout**	**Output**
Convolutional layer	32	3x3		64 x 64 x 32
Batch normalization				64 x 64 x 32
Dropout			0.2	64 x 64 x 32
Convolutional layer	32	3x3		64 x 64 x 32
Max Pooling		2x2/2		32 x 32 x 32
Batch normalization				32 x 32 x 32
Convolutional layer	64	3x3		32 x 32 x 64
Batch normalization				32 x 32 x 64
Dropout			0.2	32 x 32 x 64
Convolutional layer	64	3x3		32 x 32 x 64
Max Pooling		2x2/2		16 x 16 x 64
Batch normalization				16 x 16 x 64
Convolutional layer	128	3x3		16 x 16 x 128
Batch normalization				16 x 16 x 128
Dropout			0.2	16 x 16 x 128
Convolutional layer	128	3x3		16 x 16 x 128
Max Pooling		2x2/2		8 x 8 x 128
Batch normalization				8 x 8 x 128
Flatten				None, 8192
Dropout			0.2	None, 8192
Type	**Units**		**Dropout**	**Output**
Dense	1024			None, 1024
Batch normalization				None, 1024
Dropout			0.2	None, 1024
Dense	512			None, 512
Batch normalization				None, 512
Dropout			0.2	None, 512
Dense	1			None, 1

3.3.2 Network parameters

The parameters that are tuned during this part of the study are: Adjustments to the Adam optimizer and the batch size. Other parameters were only incidentally trained at the beginning. The batch sizes are 256, 512 and 1024. For the Adam optimizer, the combinations are listed in Table 3.3 for which the last row is showing the default Adam settings.

Table 3.3: Tested Adam variations

α	β_1	β_2	ϵ
0.01	0.9	0.999	1E-08
0.01	0.9	0.999	1E-01
0.01	0.9	0.999	1E-03
0.01	0.9	0.999	1E-04
0.1	0.9	0.999	1E-08
0.1	0.95	0.99	1E-08
0.2	0.9	0.999	1E-08
0.001	0.9	0.999	1E-08

3.3.3 Evaluation of the model

The training of the LC dataset is monitored by the previous mentioned accuracy and loss graphs. After testing of the Meander dataset, a confusion matrix and ROC curve are created to evaluate the performance of the trained model. Besides, outcomes of training are automatically stored in an excel sheet. In section 2.4 the used evaluation graphs are shown in Fig. 2.5.

3.4 Colour based feature extraction

Colour based feature extraction is done by using four Matlab scripts. Three of which are provided by F. van der Heijden of the University of Twente. 26 gallbladder leakage images of 22 different patients of the Cholec80 dataset were used to determine the optimized inter and intra class distance.

3.4.1 Region of interest selection

The first script was created to draw a region of interest (ROI) in each of the 26 gallbladder leakage images. Hereby, pixel values are selected that contain bile, while the outside of the ROI could be used as NoL pixels. After drawing an ROI, a binary mask is created which is used to collect the pixel values of the bile leakage region. By creating the complement of the mask, the outer region of the ROI can be used to obtain the pixel values of the NoL pixels. Hereafter, the RGB pixel values of each image are stored in two matrices.

3.4.2 Colour based feature extraction scripts

The first CBFE script of F. van der Heijden, creates a balanced dataset, two arrays are created with the same number of randomly selected pixels. The second script calculates the mean of the pixel values and creates covariance matrices. Intra-class whitening and inter-class decorrelation are applied to these matrices. Hereafter, one feature is chosen. By transforming the images by the feature which has the optimal linear combinations, bile leakage in the image could be more noticeable. The last script uses the calculated transformation for the highest inter class distance to transform all the images of the dataset. A threshold of 0.5 is applied to the selected feature to partly filter pixels which are assigned as gallbladder leakage to create less false positive feature 1 pixels. Feature 1 is applied to the red channel of the image. Hereafter, two datasets could be used: the dataset with normal pixel values and a dataset with transformed pixel values.

3.4.3 Training and evaluation of the model

After CBFE, training of the network will be done with the same networks, a selection of the most suitable parameters of the LC dataset study and the same evaluation methods.

CHAPTER 4

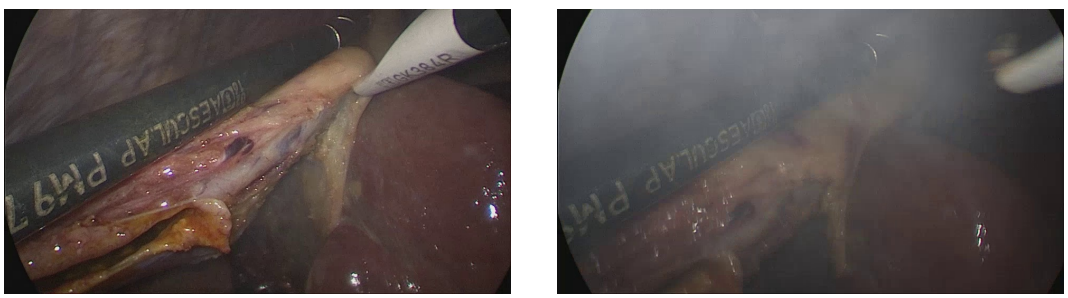
Results

This chapter shows the outcomes of this study. As first the outcomes of the dataset preparation are presented. Hereafter, outcomes of the parameter study and the binary classification study with the LC dataset are displayed. The outcomes of the study with the feature extraction images are presented hereafter. At last, training results per patient are displayed for several testing results and the M1 and M2 datasets are compared.

4.1 Dataset preparation

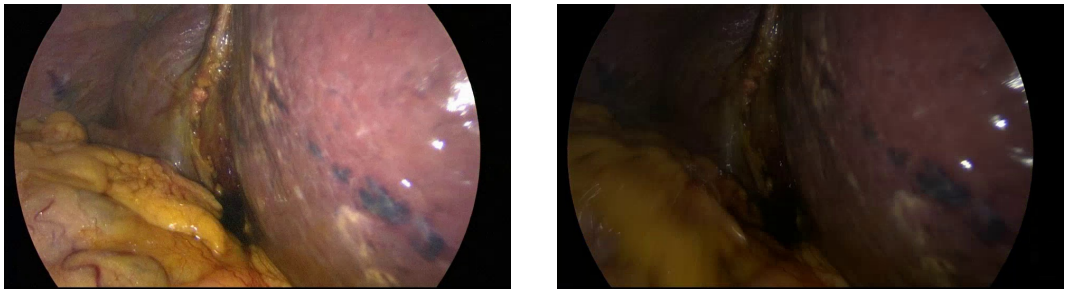
4.1.1 Selection based on image quality

The LC dataset consists of 122 patients who underwent a laparoscopic cholecystectomy. In this dataset 52138 images are included. Selection of the video frames was done by previous described selection criteria. Resulting consecutive images show how inclusion and exclusion is done based on these selection criteria. In Fig. 4.1 two consecutive frames show the appearance of surgical smoke which is created by surgical diathermy.



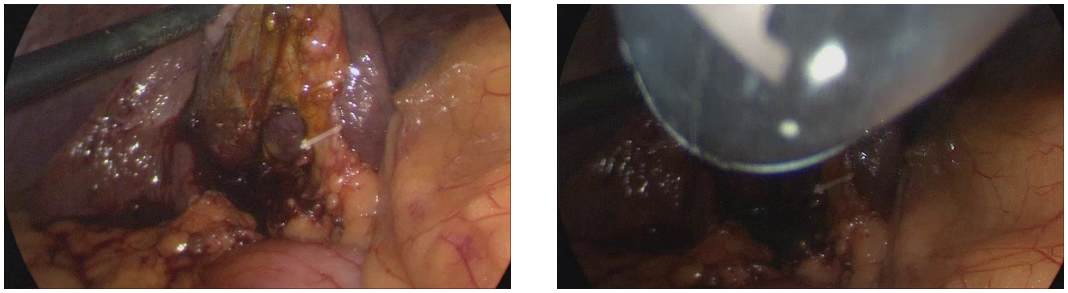
(A) Included (B) Excluded

Figure 4.1: Surgical smoke



(A) Included (B) Excluded

FIGURE 4.2: Changed lighting

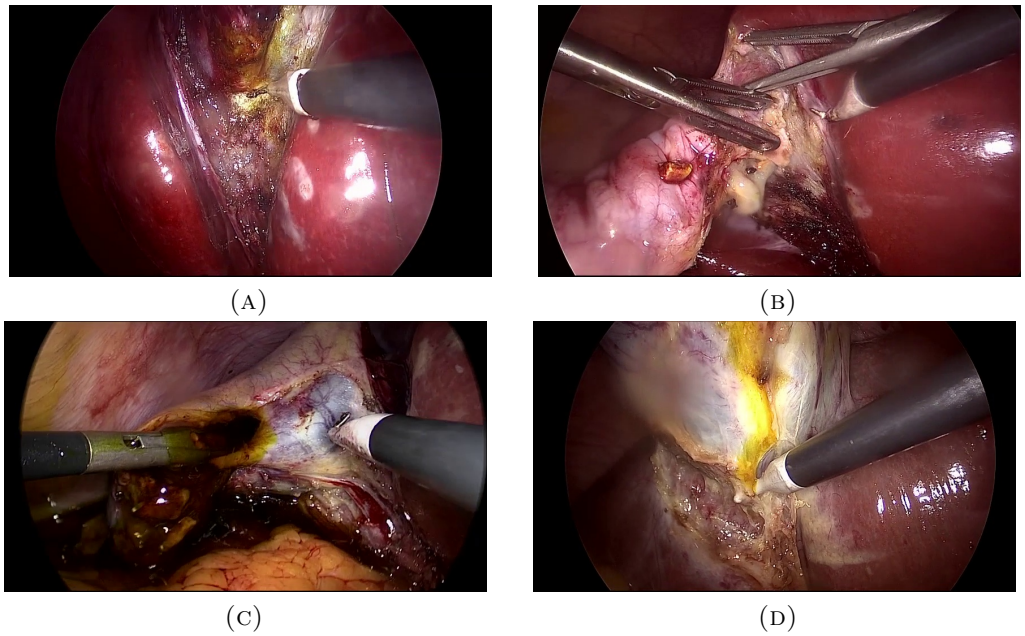


(A) Included (B) Excluded

FIGURE 4.3: Region disguised by tool

In Fig. 4.2 and 4.3 the left images (A) display a normal screen which is mostly seen during surgery. The right consecutive images (B) are darker and Fig. 4.3B mainly shows the trocar that is used to guide and stabilize surgical tools into the abdominal cavity.

4.1.2 Diversity of the dataset

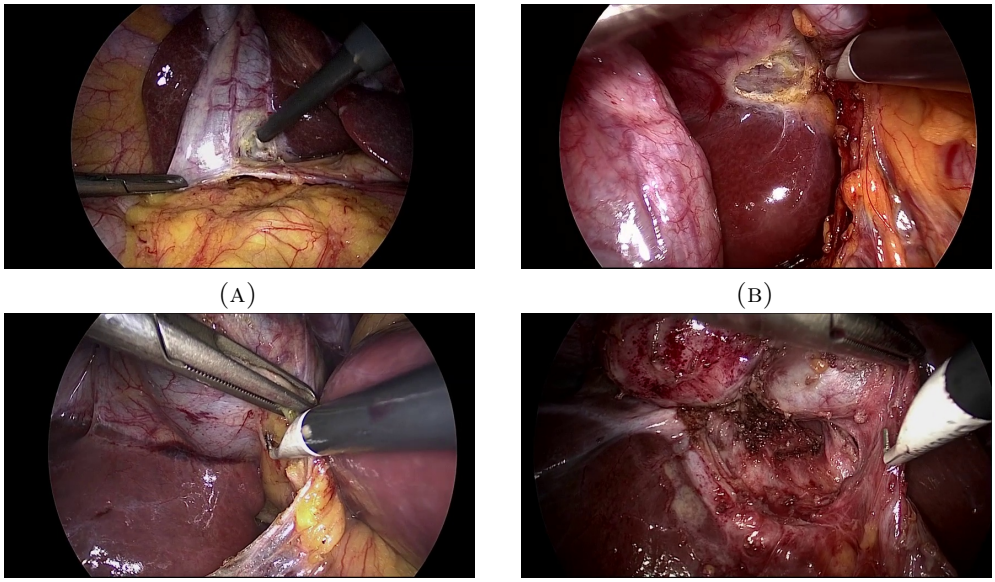


(A) (B)

(C) (D)

FIGURE 4.4: Bile leakage

When selecting bile leakage images, a diverse dataset is presented (Fig. 4.4). In image A, bile leakage is present combined with blood. Image B shows white bile and a gallstone. Dark green bile is seen in image C, while D is showing yellow-greenish bile.

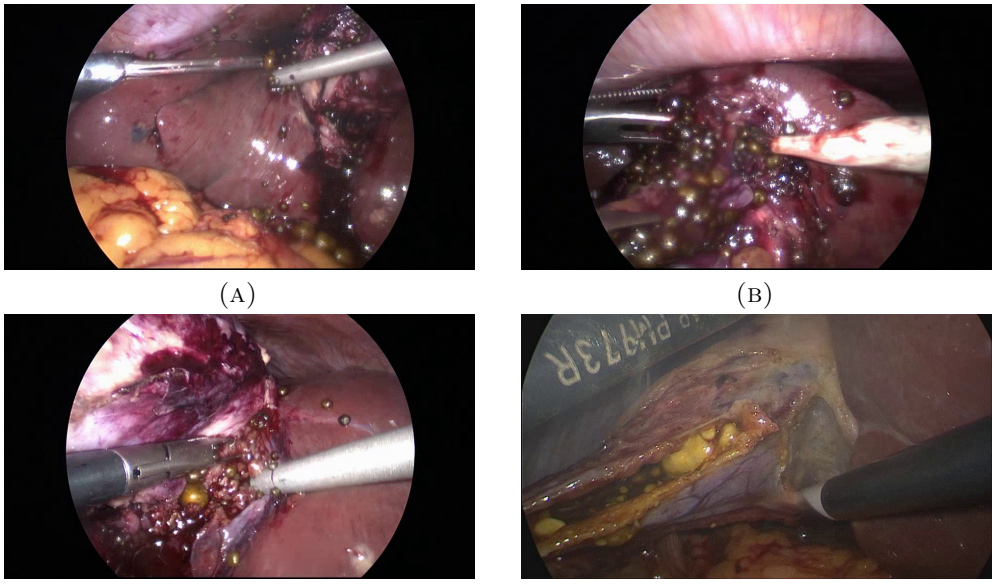


(A) (B)

(C) (D)

FIGURE 4.5: No bile leakage

A collection of NoL images is presented in Fig. 4.5. This figure shows straightforward LC images. Yellow fatty structures are seen in image A, B and C, while a more bloody sight is given in D. In B, the surgical diathermy tool is used to resect the gallbladder of the liver bed. The dried out tissue due to surgical diathermy, is more yellow compared to surrounding white, dark red or pink tissue.



(A) (B)

(C) (D)

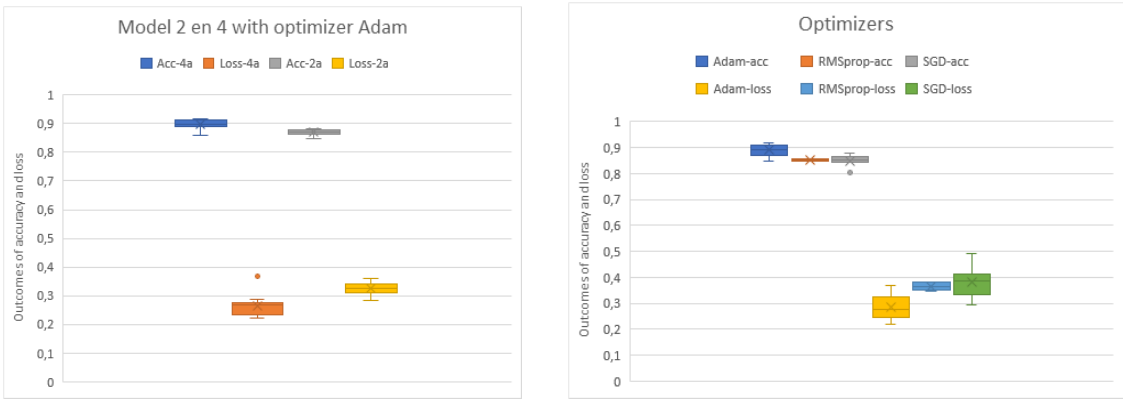
FIGURE 4.6: Gallstones

Gallstones could be present when gallbladder leakage occurs (Fig. 4.4B). Images 4.6A till C, show a case of gallbladder rupture which primarily consists of lost gallstones. In images A and

B more brown coloured stones are seen, while C and D show yellowish/goldish stones.

4.2 Effect of different parameters

4.2.1 Parameter study



(A) (B)

FIGURE 4.7: Model and optimizer comparisons

The models 2a and 4a are compared in Fig. 4.7A. Model 4a shows lower loss and higher accuracy. When optimizers are plotted in a box plot graph, optimizer Adam was showing the highest accuracy and lowest loss (Fig 4.7B). In two other graphs (Appendix B.1), the mean accuracy and loss are shown of four different models, for which model 2a and model 4a are showing the highest accuracy and loss (Table B.1). To display the effect of running for more epochs, the models 2a and 4a are plotted against number of epochs. The accuracy and loss tends to improve when the model is trained for more epochs.

4.2.2 Cholec80 dataset hyperparameters

(A)

FIGURE 4.8: Three different Adam optimizers.

The three most used optimizer settings for Adam during training of the LC models are displayed in figure 4.8. Additionally, the settings of the captions of this figure are showed in table 4.1. The settings Beta and LR show higher mean accuracy and lower mean loss compared to setting Adam E.

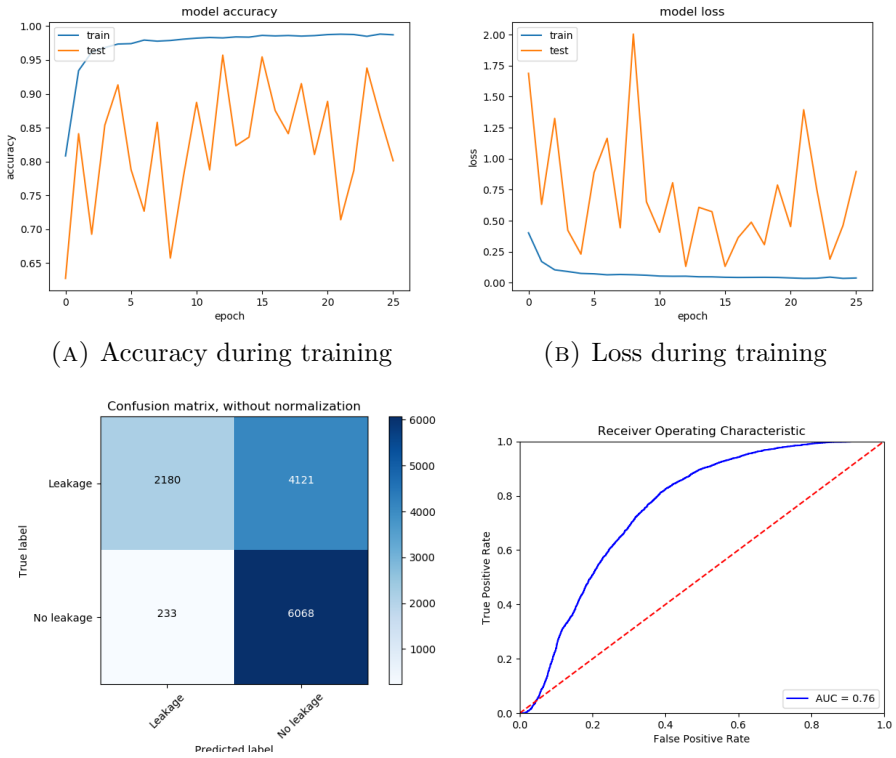
TABLE 4.1: Tested Adam settings

	α	β_1	β_2	ϵ
E	0.01	0.9	0.999	1E-04
Beta	0.1	0.95	0.99	1E-08
LR	0.01	0.9	0.999	1E-08

4.3 Binary classification of laparoscopic cholecystectomy images

4.3.1 Evaluation of trained models by using the M1 dataset

It is seen that the training of LC models is not very stable, a typical example is shown in the accuracy (A) and loss plot (B) of Fig. 4.9. The testing results of best performing model 3 and 4 during training on Cholec80 data are shown in the confusion matrix and ROC curve (Fig. B.3 and 4.9). These models had highest accuracy and/or lowest loss for the validation set during training. Measures of testing results of two training sessions per model are summarized in Table 4.2. Model 3b and 4b are the models which performed second best during training (Appendix B.2).



(A) Accuracy during training

(B) Loss during training

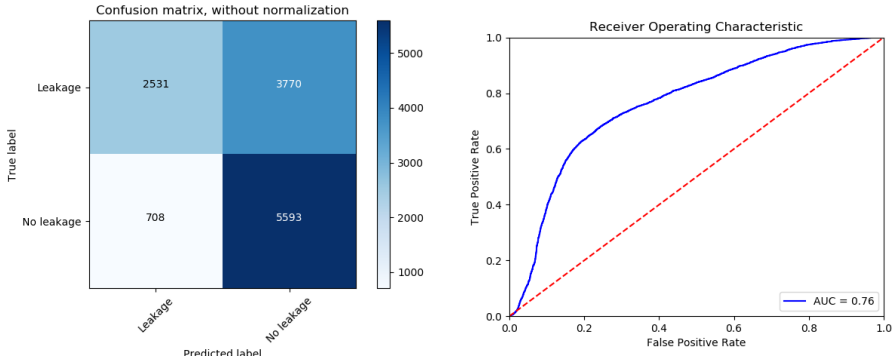
(C) Confusion matrix

(D) ROC curve

FIGURE 4.9: Best training model 4

TABLE 4.2: Testing results for the M1 dataset after training model 3, 3b, 4 and 4b.

Measures	Model 3	Model 3b	Model 4	Model 4b
Sensitivity	0.1968	0.4017	0.3460	0.1739
Specificity	0.9838	0.8876	0.9630	0.9598
Precision	0.9240	0.7814	0.9034	0.8125
AUC	0.73	0.76	0.76	0.65



(A) Confusion matrix (B) ROC curve

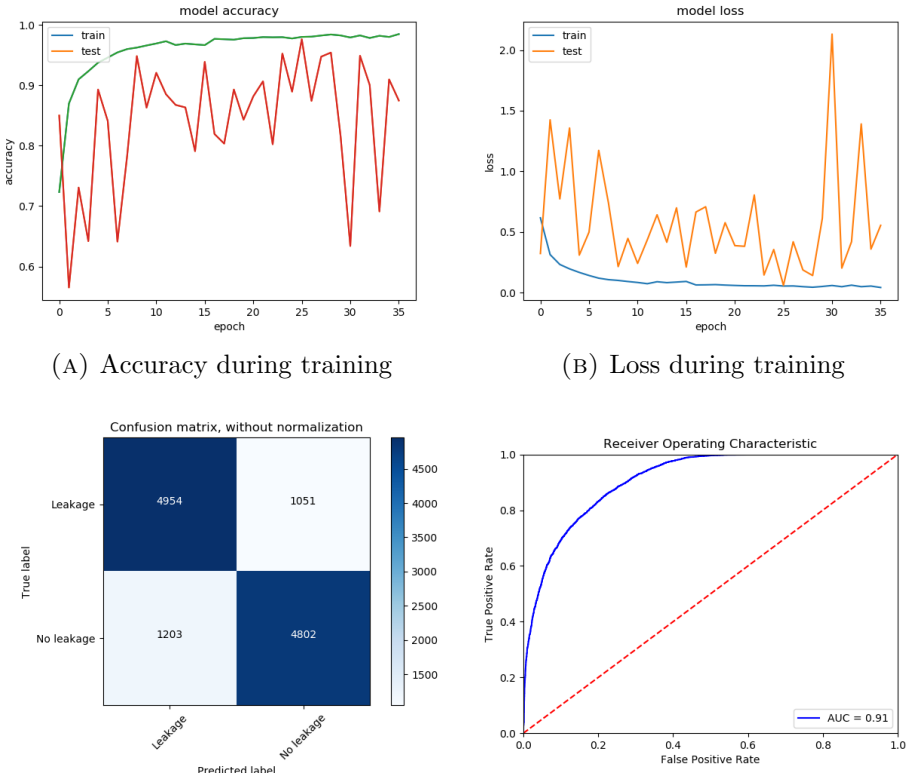
FIGURE 4.10: Second best model 3

4.3.2 Evaluation of training of the Merged dataset by using the M2 dataset

Model 3 and 4 are trained with the Merged dataset and the M2 dataset is used as test set. The confusion matrix and ROC curve are shown in Fig. 4.11 and B.4. Measures of testing model 3 and 4 are summarized in Table 4.3. The highest AUC value of 0.91 is obtained by using Model 3. The accuracy and loss plots are added for Model 3 and they show improvement of training until epoch 30 (Fig. 4.11).

TABLE 4.3: Testing with M2 dataset after training Model 3 and Model 4 with Merged dataset

Measures	Model 3	Model 4
Sensitivity	0.8250	0.8355
Specificity	0.7997	0.6893
Precision	0.8046	0.7289
AUC	0.91	0.87

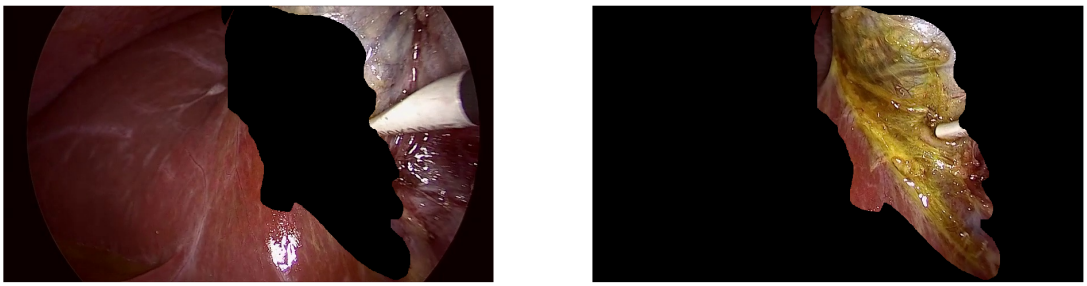


(A) Accuracy during training (B) Loss during training

(C) Confusion matrix (D) ROC curve

FIGURE 4.11: Testing with M2 dataset after training Model 3 with Merged dataset

4.4 Colour based feature extraction



(A) No bile leakage (B) Bile leakage

FIGURE 4.12: Masked leakage images for CBFE

For 26 bile leakage images of 17 patients, a CBFE masked images is created to calculate the highest inter/ intra class distance and create a transformation matrix. One example of a masked image with bile leakage and the surrounding pixels for no bile leakage is shown in Fig. 4.12.

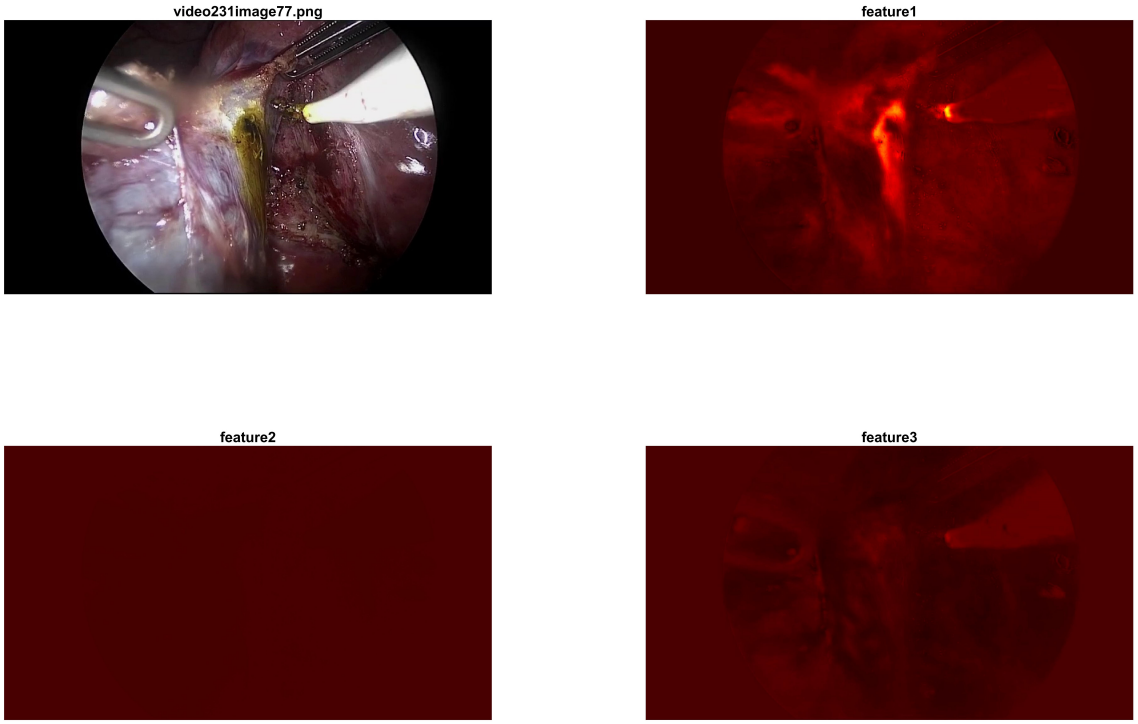


FIGURE 4.13: Colour based feature extraction

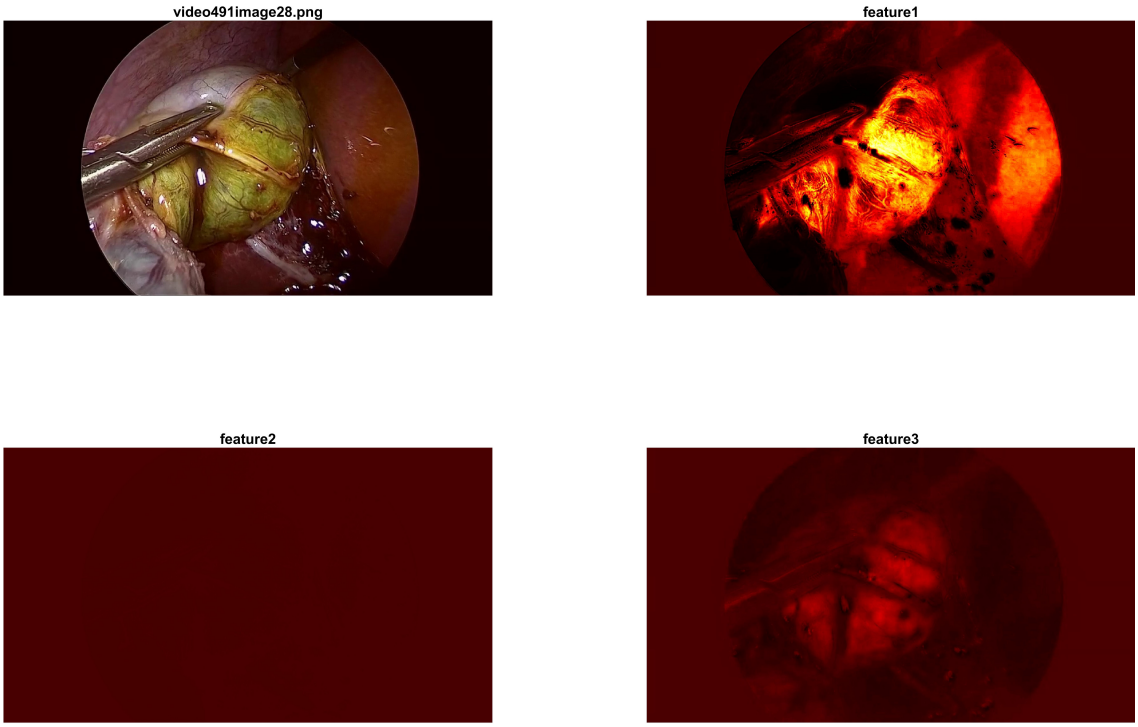
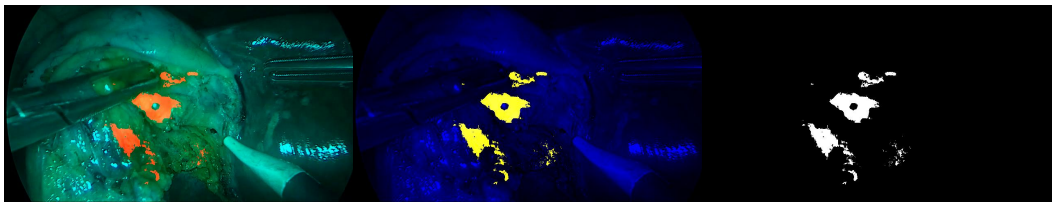


FIGURE 4.14: Colour based feature extraction

Two examples of these calculations show that feature 1 shows highlighted bile (Fig. 4.13 and 4.14). The bright yellow/orange areas are shown at the places of bile, while surrounding tissue without any bile is less or not highlighted. Feature 2 does not contain any information and feature 3 vaguely shows a surgery tool (Fig. 4.13) and the contours of the gallbladder (Fig. 4.14). For feature 1, the sensitivity and specificity per pixel for the 26 leakage/no leakage masks were 75% and 70%, respectively.

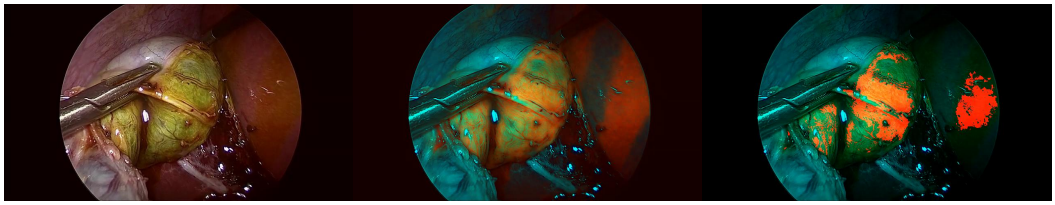
When feature 1 information is applied to the red channel of an image, surrounding tissue is still visible (Fig. 4.15A). When the blue channel or green channel is replaced by feature 1, surrounding tissue is not visible (Fig. 4.15A and 4.15B).



(A) Red channel (B) Green channel (C) Blue channel

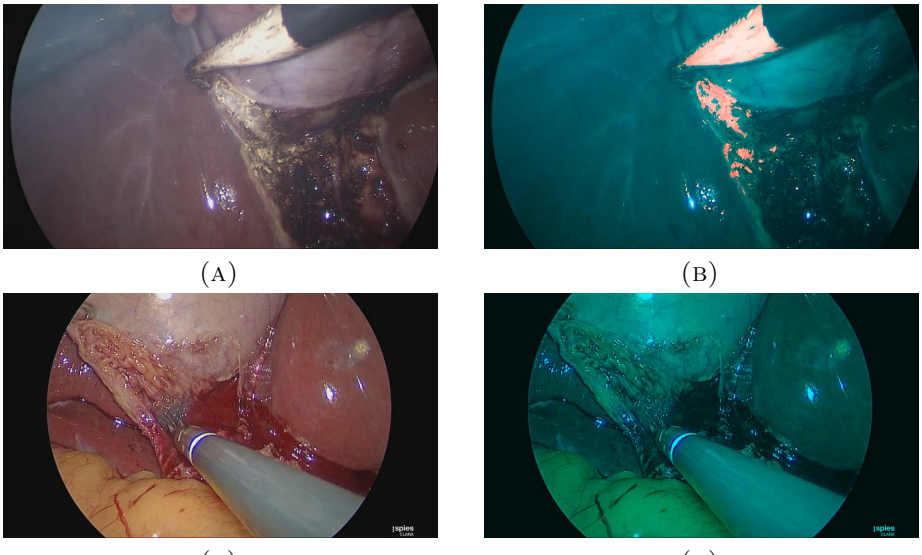
FIGURE 4.15: Feature 1 applied to red (A), green (B) and blue (C) channel of an image

Feature 1 transformation is applied to image 4.16A and replaces the red channel of the images (Fig 4.16B). Image 4.16C is created after applying a threshold of 0.5 and replacing the red channel of image 4.16A by a thresholded feature 1.



(A) normal (B) feature 1 (C) f1-segmentation

FIGURE 4.16: Normal bile leakage image (A), image after feature 1 transformation (B) and image after transformation with a threshold of 0.5 and higher for feature 1 (C)



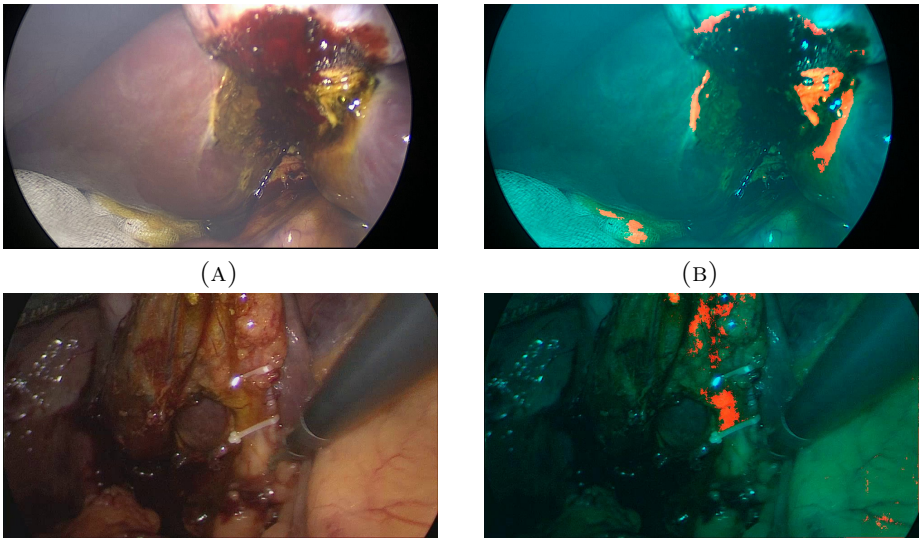
(A) (B)

(C) (D)

FIGURE 4.17: Transformation of the NoL dataset (A,C) to images transformed by thresholded feature 1 (B,D)

Previous images illustrate how transformation of the Cholec80 images is carried out. Most of the NoL dataset images did not contain highlighted areas caused by thresholded feature 1 (Fig. 4.17D). The images only show a blue version of the LC image. In Fig. 4.17B the tool and

some surrounding tissue is highlighted by the feature 1 transformation. This is seen as pink or orange pixels.



(A) (B) (C) (D)

FIGURE 4.18: Leakage images of the Cholec80 dataset before and after CBFE

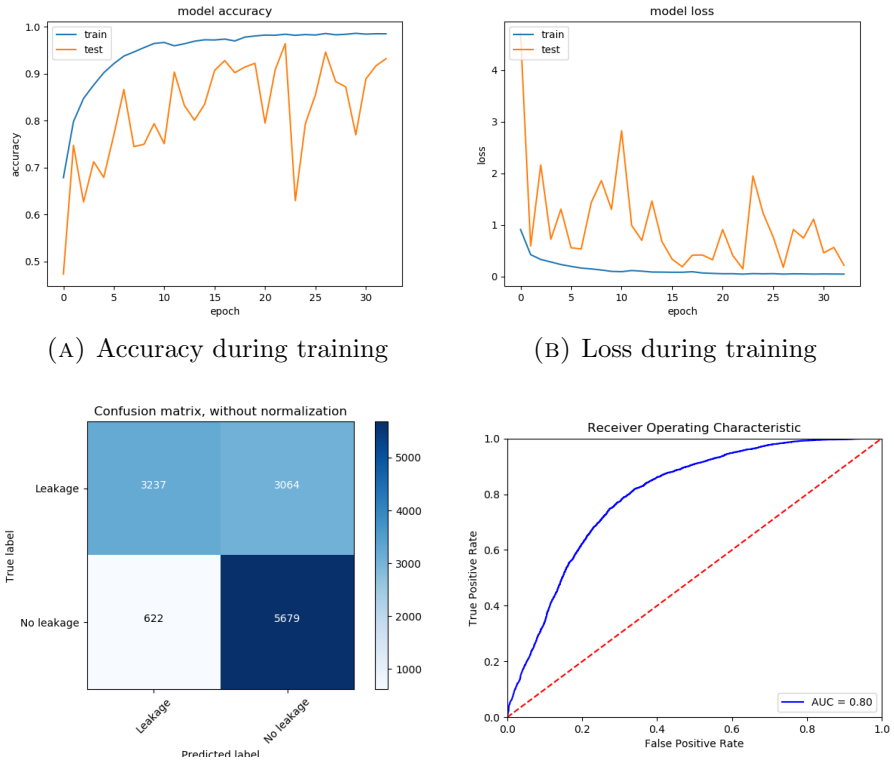
The L images of the Cholec80 dataset are shown before and after transformation by thresholded feature 1 (Fig. 4.18). Image 4.18A contains more distinguishable bile leakage and fresh red blood, compared to 4.18C which also shows some clotted blood. The thresholded feature 1 components in image 4.18B and 4.18D do not completely cover the bile leakage which is seen at the 4.18A and 4.18C

4.4.1 Evaluation of model 3 and 4 with CBFE M1 dataset

The transformed image M1 dataset has the same size as the previous used LC datasets. Training and validation of model 3 with CBFE Cholec80 images resulted in accuracy and loss plot (Fig. 4.19). Testing of the trained model by using the CBFE transformed M1 dataset resulted in three confusion matrices and ROC curves (Fig. 4.19, B.6 and B.5). Model 3 has the highest AUC value for these three training sessions, namely 0.80.

TABLE 4.4: Testing model training of CBFE M1 dataset for model 3, 3b and 4

Measures	**Model 3**	**Model 3b**	**Model 4**
Sensitivity	0.5137	0.4658	0.4480
Specificity	0.9013	0.8946	0.8334
Precision	0.8388	0.8155	0.7289
AUC	0.80	0.71	0.71



(A) Accuracy during training (B) Loss during training

(C) Confusion matrix (D) ROC curve

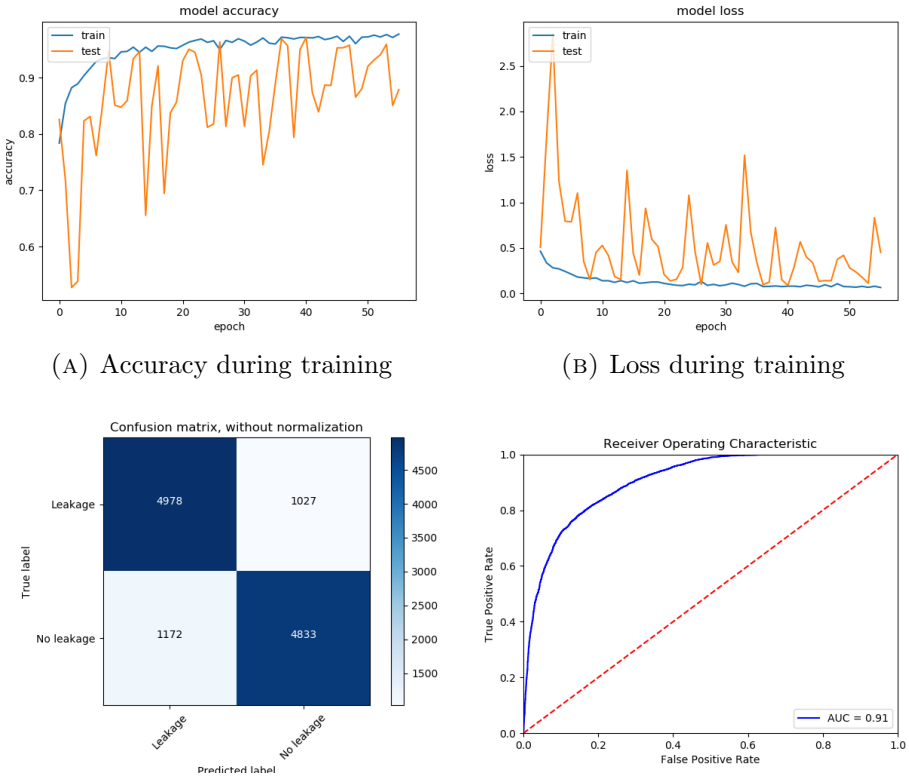
FIGURE 4.19: Best training and testing results for CBFE images with model 3

4.4.2 Evaluation of the training of the CBFE Merged dataset by the M2 dataset

Training of Model 3 and 4 with the CBFE Merged dataset is evaluated by using the M2 dataset and resulted in two confusion matrices and ROC curves and loss and accuracy plots (Fig. 4.20 and B.7). Statistical measures of testing the models are shown in 4.5. Almost 83% of the leakage images and 80% of the no leakage images are correctly identified by the trained model 4 which leaded to an AUC value of 0.91.

TABLE 4.5: Test outcomes of the CBFE Merged dataset

Measures	**Model 3**	**Model 4**
Sensitivity	0.5047	0.8290
Specificity	0.9504	0.8048
Precision	0.9119	0.8094
AUC	0.84	0.91



(A) Accuracy during training (B) Loss during training

(C) Confusion matrix (D) ROC curve

FIGURE 4.20: Testing of trained model 4 with CBFE merged dataset

4.5 Comparison between M1 and M2 dataset

4.5.1 Comparison of test results of first and second Meander dataset

The trained models which are previously used to create Table 4.2 are used for model evaluation by the M2 dataset which resulted in Table 4.6. The results of testing with the M2 data show increased sensitivity, specificity, precision and AUC values when compared to testing with the M2 dataset.

TABLE 4.6: Testing of best and second best training of Model 3 and 4 by using CBFE M2 dataset

Measures	Model 3	Model 3b	Model 4	Model 4b
Sensitivity	0.4654	0.6306	0.6208	0.4380
Specificity	0.9947	0.9157	0.9386	0.9659
Precision	0.9887	0.8821	0.9099	0.9277
AUC	0.81	0.81	0.86	0.68

The results of training model 3 and 4 with CBFE images and testing with the earlier mentioned results of the M1 dataset, are compared to outcomes of testing with the M2 dataset (Table 4.7). For Model 3 all measures increased, except specificity, while all measures improved for Model 4 when the M2 dataset is used.

TABLE 4.7: Testing of CBFE trained models with M1 and M2 datasets

Measures	Model 3 M1	Model 3 M2	Model 4 M1	Model 4 M2
Sensitivity	0,5137	0,6724	0,4480	0,6531
Specificity	0,9013	0,8873	0,8334	0,8966
Precision	0,8388	0,8564	0,7289	0,8633
AUC	0.80	0.81	0.71	0.81

4.5.2 Correctly identified frames per patient

The percentage of correctly identified frames of the M2 dataset by using four trained models, are plotted for each surgery video (Fig 4.21 and 4.22). These figures show that this percentage of identification is fluctuating per video and per model. The blue line is showing the mean percentage of identification by these four models. For five leakage videos, the mean identification percentage for the video frames is below 50%. For the no leakage videos this was only the case for two videos. In two of the leakage videos the mean percentage of identification were 10% and 12%. The other five badly identified leakage and no leakage videos, show mean percentages above 42%. On average, 72% of the leakage M2 dataset was correctly identified by the four models and 80% of the no leakage frames.

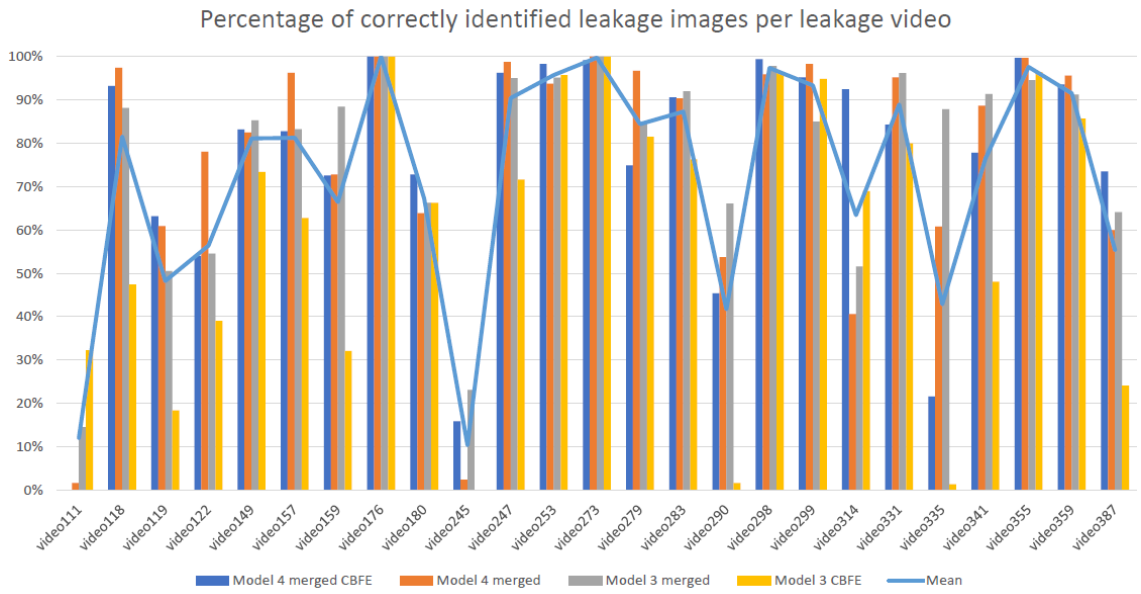


FIGURE 4.21: Identification rate per leakage video

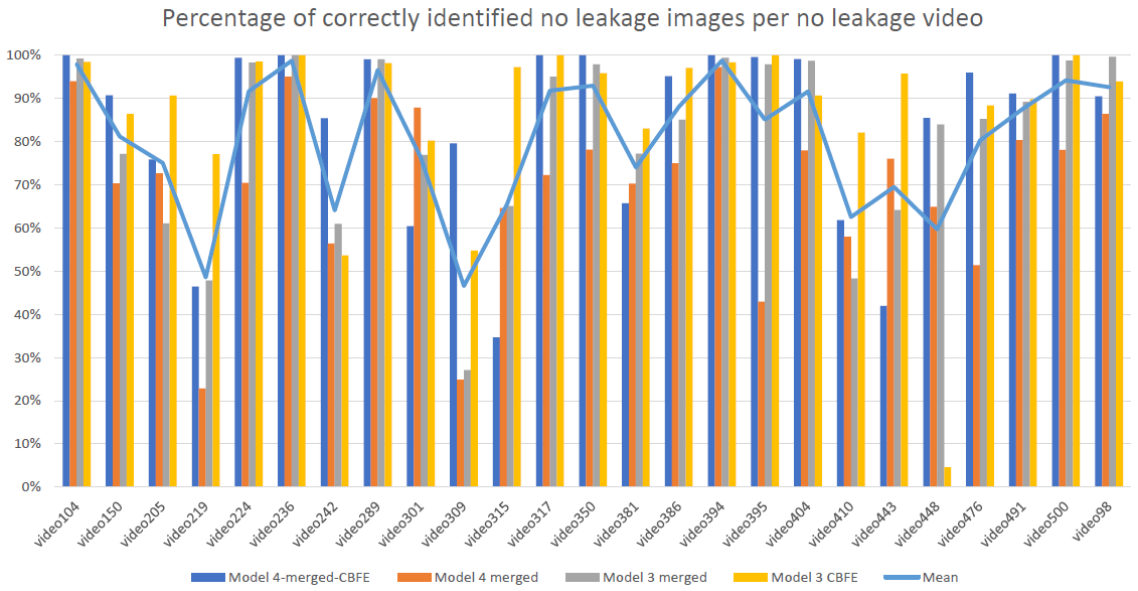


FIGURE 4.22: Identification rate per no leakage video

CHAPTER 5

Discussion

This chapter discusses the outcomes of the result section, namely the data preparation, parameter study, binary classification with the LC dataset, the colour based feature extraction results and comparison between the M1 and M2 datasets. At first, a brief summary of results will be given to shortly answer the research questions. Hereafter, the main results will be explained. Subsequently, the study procedure will be discussed and areas for improvement will be given in the limitation section. Lastly, recommendations will be done for future research and future perspective on AI and LC will be given.

5.1 Summary of results

5.1.1 Research questions and aim

The first research question was composed to determine to what extent it is possible to detect bile leakage in laparoscopic cholecystectomy videos by using a deep learning algorithm. If detection was possible it was important to determine if this bile leakage detection algorithm has clinical added value when compared to the current reporting rate of bile leakage in surgical reports. At last it was investigated if the use of feature extraction for a laparoscopic cholecystectomy dataset could contribute to the gallbladder leakage detection rates.

5.1.2 Summary of results

Data preparation was executed successfully. Although manual selection was needed to create correct leakage datasets, an extensive LC dataset was created which consists of 62380 images of 172 patients. These laparoscopic cholecystectomies are performed by 23 different surgeons. The parameter study led to the selection of a best performing model and more insight was obtained in tuning parameters. The most successful binary classification was performed by using the Merged dataset for training and validation and the M2 dataset for testing of the trained model. Best results are obtained by using the trained model 4 with CBFE images. For the M2 dataset a sensitivity of 83% and specificity of 80% are achieved with an AUC score of 0.91. Although the CBFE method created a dataset which achieved better results compared to the non-CBFE images, it was ambiguous when both Models and multiple training sessions are compared. The trained Model 3 and 4 show different results when the CBFE images are used.

This study provides first insights in automatic bile leakage detection. The best results of correct classification of bile leakage are close to the 87% reporting rate of bile leakage in surgery reports of a Dutch study [7]. Since the Merged dataset showed better results compared to the smaller Cholec80 dataset, it could be expected that a clinical applicable algorithm is within reach if the current dataset is extended.

5.2 Explanation of results

5.2.1 Data preparation

As expected, compiling of the dataset was partly done by manual selection of the images. The need of manual selection is substantiated by the fact that two consecutive frames could contain one good quality image and one image which needs to be excluded (Fig. 4.1, 4.2 and 4.3).

The diversity of the dataset was pointed out by four figures of bile leakage, no bile leakage and gallstones (Fig. 4.4, 4.5 and 4.6). These twelve images show that this dataset contains different colours of bile (white to dark green), gallstones of different colours are present, colour of surrounding (fatty) tissue and amount of blood. Although the created LC dataset already contained images of 172 patients, the diversity of the data demands extension of the current dataset with more patients. Especially, patients with less common colours of gallstones (dark green) and bile (white).

5.2.2 Effect of different parameters

Based on the results of the parameter study, Model 4a, later used as Model 4 during the binary study, and the Adam optimizer are chosen as network architecture and optimizer (Figure 4.7). At first it was thought that the longer trained models showed improved training results, but the addition of early stopping helped to stop training if results did not improve for several epochs (Table B.1). Therefore, if a model is trained for more epochs it means that the model is more trainable compared to models that are trained for less epochs.

During training with the Cholec80 dataset it was considered to use a slightly deeper and wider network architecture since the data was more complex. Therefore, the previous used Model 4 and the more complex Model 3 are used for the binary classification of LC data (Table 3.2 and 3.1). For both models, the LR Adam settings with a learning rate of 0.01, and the Beta settings with a learning rate of 0.1, $\beta 1$ of 0.95, $\beta 2$ of 0.99 and ϵ of 1E-04 showed best performances.

Grid search and K-fold cross validation

Initially, it was planned to search for an optimized batch size, Adam setting, data augmentation option and dropout setting. A start was made for training and testing with for example the earlier described different Adam settings (Table 3.3). The use of multiple parameter tests at once is highly preferred for this type of parameter testing, since training takes a long time if each parameter and combination with other parameters is trained individually.

In python the Scikit-learn software enables training of data for several parameters at once [53]. The term grid search is used, since a grid of different parameters and their settings could be made. Hereby, all combinations of parameters and settings are tested. However, K-fold cross validation is necessary when using the grid search. This method splits up the dataset in smaller sets, for example ten sets, and combines nine of them for a training set and one for a validation set. The advantage is that testing of parameters could be done ten times with the ten different sets as validation dataset. Hereby, optimal use of valuable patient data is achieved [33].

When medical data is used, these splits should be made per patient. Especially for the LC datasets, variation between images of one patient are low. When using images of one patient for

training and validation, biased results are obtained. The Scikit-software enables designation per patient to a group, called Group K-fold cross validation [53, 54].

When searching for examples for large image datasets per patient, only manually created groups were found. Since the dataset of this research consists of a large group of patients with a lot of images, no suitable method was found to create groups per individual patient and the use of Grid search and K-fold cross validation was not possible. Therefore, not all parameters were tested extensively. If results were insufficient for a few training sessions, another setting was tested. For the Adam settings, five of eight options were used four times or less and therefore not suitable for setting analysis. Thus, not using Grid search and K-fold cross validation caused lack of supporting information to choose the most optimal network and parameter settings. As a result, the study became more an exploratory type of study than initially planned.

5.2.3 Testing of model training on M1 dataset

Testing of the models on LC data, resulted in more unstable training then expected. More data, smaller batch size and lower learning rate for Adam, could improve stability of training but no results of these adjustments were seen. As mentioned before, sometimes small number of epochs are used for training when early stopping is applied. If a model runs for more epochs, accuracy and loss plots look more stable, because these fluctuations seem smaller when more epochs are presented in a plot. Besides, the large variations between patients could also cause large fluctuation in the accuracy and loss during training.

Although the ROC curves that are obtained during testing with the M1 dataset, all show AUC values of 0.65 and higher, sensitivity is an important measure for suitability of the trained models for bile leakage detection. Both Model 3b and 4 show AUC values of 0.76. When choosing one of these models, Model 3b is preferred for leakage detection, since this model shows higher sensitivity compared to Model 4. However, the acceptable AUC value is mainly achieved by high identification scores for the no leakage data, as seen by high specificity values. Therefore, these models are not useful for clinical practice. The most logical explanation for the high specificity for all testing results is that the patient group for the NoL dataset is more diverse, since more patients are included. Besides most videos were included almost entirely by using only one frame every few seconds, which creates more diverse images of different surgery phases.

5.2.4 Testing of the Merged dataset on M2 dataset

The extension of the Cholec80 dataset with the M1 dataset resulted in improved testing results for Model 3 and 4. When comparing testing outcomes with the smaller Cholec80 dataset, it is seen that specificity decreased for both models when the large dataset was used, while sensitivity significantly increased. Since, two different training and testing datasets are used, the difference in sensitivity and specificity could be explained by how difficult the testing sets are, which is explained in section 5.2.8.

When comparing the measures with clinical practice, the reporting rate of bile leakage for a Dutch study was 87%. When looking at the results of Model 3, the 82.5% sensitivity is close to the Dutch reporting rate. When looking at international numbers which are all lower than the Dutch 87%, these sensitivity results are definitely promising [7].

5.2.5 Colour based feature extraction transformation of datasets

The applied method for creating masked images, was an easy method to obtain specific leakage and no leakage pixel values. The resulting images of feature 1, 2 and 3 showed that the identification of gallbladder leakage could be done by using feature 1 (Fig. 4.13 and 4.14). When applying feature 1 transformation to different colour channels, it was easy to decide that the feature 1 would replace the red channel (Fig. 4.15). This is a logical outcome when looking at

both bile and no bile images, since both will contain a lot of red colour and more contribution of green and blue is expected for bile leakage images.

Although the calculated specificity and sensitivity of feature 1 for all pixels of the 26 images was quite high without any training, this was only achieved for the 26 images that were used. When looking at the eight images of Fig. 4.13 and 4.14 , it is already visible that darker green bile was less pronounced compared to more yellow bile. It was expected that adding a threshold to feature 1, could reduce false positives for gallbladder leakage. When looking at one of the 26 feature extraction images (Fig. 4.16), it looks like a successful method for identification of gallbladder leakage. Since this feature 1 transformation was based on only leakage images, no tests were done for how it would look like when applied to NoL images. Unfortunately, figure 4.17 and 4.18 show that a surgical tool could also be assigned as gallbladder leakage. In addition, while looking at the leakage image, more identified pixel values were expected, but only some bile leakage is highlighted. The identification of a feature for CBFE probably could have been more useful when a more diverse leakage dataset and NoL images are used, but this was not available and considered at the time when feature extraction was done.

5.2.6 Testing of CBFE trained model 3 and 4 by using the M1 dataset

When evaluating the training of Model 3 and 4 with CBFE Cholec80 images, some measurements did improve compared to the normal video frames. Figure 4.19 showed the best results. The most salient element of the results is that identification of leakage improved for all three models when compared to the testing of no-CBFE M1 images. Since specificity decreased, it substantiate the idea that CBFE is mainly an advantage for leakage images.

5.2.7 Testing of CBFE Merged dataset trained models by using the M2 dataset

When comparing the Merged dataset and CBFE Merged dataset results, Model 4 mainly showed increased results, while the statistical measures for Model 3 primarily decreased. For all trained and tested models, training with the CBFE Merged dataset resulted in the best trained model for laparoscopic cholecystectomy data with 82.9% sensitivity, 80.5% specificity, 80.9% precision and an AUC value of 0.91. However, when looking at the specificity value of Model 3 for the CBFE Merged data and other trained models mentioned previously, it could be seen that specificity values of Model 4 are relatively low. One explanation is that some models will specialize for leakage data during training while others are more specialized for no leakage data. For this study, a high sensitivity is preferred, since the aim is to identify bile leakage.

5.2.8 Comparison of M1 and M2 dataset

Comparison between testing outcomes

For the interpretations of previous results, it is important to notice that two testing datasets are used, namely M1 and M2. Since, M1 was added to the Cholec80 dataset to create the Merged dataset, a new testing set was made. However, results of testing are inevitably influenced when using two different datasets. Therefore, Cholec80 results and Merged results should not be compared.

To substantiate this, trained models which are used in chapter 4.3.1 and 4.4 are also validated with the M2 dataset. The results for the Cholec80 dataset show that all sensitivity and three of four AUC values increased significantly, while specificity and two of four precision values decreased (Table 4.6). The differences between no leakage data for the M1 and M2 could have caused the decrease in specificity which was seen for the Merged dataset testing. When comparing the outcomes of the CBFE Cholec80 dataset, seven of eight statistical measures increased when the M2 dataset was used.

Although extensive evaluation of the differences between the two datasets was not possible because of limited time, these findings already show that it is important to use a large and diverse testing dataset if an algorithm is used for clinical practice. Previous results still show that the best result of Model 4 with CBFE Merged dataset is an excellent achievement, but also point out that different results are obtained when two different datasets are used.

Testing outcomes per patient

Additional to conclusions about using two different datasets, some information per patient was also plotted in two figures (Fig. 4.21 and 4.22). When creating a generalizable dataset, these graphs show that some patient data is more difficult to identify compared to other patients. Especially two patients showed low mean percentages of identification. When extending the dataset, this is important information since similar difficult patient data should be added to the training and validation dataset to create a more generalizable trained model.

5.3 Limitations of the study

5.3.1 Parameter choices

As mentioned previously, the choice for specific parameters was not statistically substantiated since grid search and K-fold cross validation are not used. Additionally, an image size of 64x64 pixels was chosen at the start of this study, based on the settings of the script for the parameter study. Initially, a larger image size was used when starting training with the Cholec80 dataset. Since, this dataset contains more and larger images compared to the parameter study, the GPU of the computer reached its limit and one simple solution was to downsize the images. Hereby, information in the images was reduced, the original size was 854x480, and it would have been useful to investigate if larger images would have resulted in higher bile leakage detection rates. Lastly, an often used strategy is to use a (pre-trained) well known network architecture like a VGG-16, Inception or ResNet network [55]. It would have saved time since network architecture choices were not necessary and training results could have been compared to other articles.

5.3.2 Dataset preparations

The component of this research which consisted of the most difficult choices, was the creation of the dataset. At first, manual selection of the images took a lot of time and is sensitive to human errors. Although, this was expected, the python script that was used for creating the short videos and transforming these to images based on start and end time, was not precise enough. This was caused by fluctuating fps and the fact that the rounded start time did not always match with the corresponding frame times in the video. Hereby, more or less frames were excluded than planned and more manual checking was needed.

The time consuming process of creating a database and the limited amount of patients, was one of the reasons why initially no test set was created for testing at the start of the binary classification. Besides, data should be presented differently for testing and this took more time to figure it out. Some parameter choices would have been more easy if more information was present about the performance of the network on new data.

The most difficult consideration when making the dataset, was the balance between number of patients and number of frames per patient. When the Cholec80 dataset was created all 25 fps were used for the leakage dataset, since the goal was to create a large dataset. For the NoL dataset around 690 frames were created per patient and for this dataset more patients were included compared to the leakage dataset. When the M1 dataset was made, still more NoL patients were included, but the total time of NoL videos was used to calculate a frame rate. Hereby, a short NoL video contributed less frames compared to longer videos, which created a

more diverse dataset. For the M2 dataset even number of patients and the more balanced frame selection method for the NoL dataset was used. A more diverse and balanced dataset would have been made, if more consideration had gone into creating the dataset.

Lastly, the earlier mentioned differences between the M1 and M2 datasets show that the testing datasets are not generalizable and that more and more diverse data is needed to create a useful test set.

5.4 Recommendations for future research

5.4.1 K-fold cross validation and grid search

Future research into the application of deep learning for laparoscopic cholecystectomy could be done to improve and extend this study. At first, the implementation of group K-fold cross validation and grid search, will give more insights into optimal parameters settings. Besides all videos are used optimally, since validation and training is done with all surgery videos, compared to a fixed training and validation dataset. Eventually, this will result into a more comprehensive trained model.

5.4.2 Extension of the LC videos

More laparoscopic cholecystectomy videos are available in the Meander Medical Centre. 380 LC videos are not used, but are downloaded and stored at the deep learning computer. Therefore, extending of the dataset is relatively easy, but will improve reliability of the trained model.

5.4.3 More parameters

During this research, not only bile leakage, but gallstones were included as well. As mentioned earlier in the introduction, the presence of gallstones during a LC is a predictive value for extended surgery time and classifying the surgery as a difficult LC [15]. Besides, it is a risk factor for developing postoperative complications [5–8]. Therefore detecting gallstones could be a useful additional feature, besides detecting of bile leakage. The presence of gallstones during surgery could be reported in the surgery report. As a result, complications due to gallstones could be noticed more easily.

In the introduction, the high-risk surgery phases and the risk factors for a difficult LC are mentioned. All this information could one-by-one be included into a general report for preoperative risk assessment, classification of difficulty of the surgery and postoperative assessment of its complications.

5.4.4 Using videos instead of images

During this study only video frames are used. When gallbladder leakage occurs, several consecutive frames will contain bile and/or gallstones. By only using frames, this information is lost. A Long Short-Term Memory (LSTM) network was developed by Hochreiter et al. in the late nineties. The main idea of the network is that it contains cell blocks which have multiple memory cells. These cells can decide if the input should be stored or left out. By using this network, important long term information is stored and temporary changes of input value do not immediately disturb weights [27, 56]. When applied to the LC videos, it could make it easier to detect leakage in a video and splitting of videos into frames is not necessary which saves time. Besides, using consecutive frames resembles the real world and could eventually enable the use of a classification tool during surgery.

5.5 Clinical applicability and future perspective

When putting the results of this study into a clinical perspective, three main thoughts should be taken into account. At first, at this point there is no clinical added value for the trained network as long as the sensitivity measure is below the 87% reporting rate in Dutch surgery reports [7]. Especially, since the earlier mentioned relative uncertainty also needs to be taken into account [33].

However, when looking at the results from a scientific perspective, the first results of bile leakage detection could be considered as a promising and interesting first step. Extending the dataset and optimizing hyperparameter settings will eventually result in higher detection rates, which will hopefully create an algorithm with clinical added value.

At last, the developing of AI and application of AI for healthcare purposes is promising and will continue to develop at fast pace. It is a matter of time until EHR information is checked by trained algorithms which enables that preoperative risk assessment is done and surgery complications are reported automatically. Nevertheless, at this moment surgery is an art practiced by surgeons. Only when a large amount of surgery videos are watched, one could see the frustration of a surgeon or clumsiness of a surgical trainee shortly before bile leakage occurs and these human errors are difficult to describe in a few parameters.

CHAPTER 6

Conclusion

This study aimed to develop a deep learning algorithm which is able to detect bile leakage in laparoscopic cholecystectomy video frames. It can be concluded that it is possible to detect bile leakage by using a deep learning algorithm. The use of colour based feature extraction partly resulted in a better performing classification algorithm. However, more research is needed to substantiate the use of CBFE for bile leakage detection in LC videos. Although no clinical added value can be obtained based on the results of this study, the algorithm, datasets and parameter settings can be improved and may provide clinically relevant results in the future. At first it is important to extent the dataset and choose the best strategy for creating a dataset beforehand, since several options could be chosen for splitting the surgery videos. Additionally, it is likely that the use of grid search and K-fold cross validation will help to improve training of a model and will show increased bile leakage detection rates. When bile leakage detection will eventually be used in clinical practice, this study will be a small contribution to the ultimate goal of improving quality of care for patients who receive a laparoscopic cholecystectomy.

APPENDIX A

Research proposal

A research proposal was written to obtain permission for collection of patient data, namely LC videos. The original research proposal is added as a pdf file at the next page of this document.

Deep learning for identification of gallbladder leakage during laparoscopic cholecystectomy

Researcher:
M.H. Gerkema, Bsc
mh.gerkema@meandermc.nl

Supervisor:
Prof. dr. I.A.M.J. Broeders
iamj.broeders@meandermc.nl

1 Introduction

1.1 Clinical Background

In the Netherlands, around 25,000 gallbladders are surgically removed by cholecystectomy every year [1]. Most common indications for surgery are symptomatic gallstones and complications due to gallstones like cholecystitis, jaundice and pancreatitis [2]. Shortly after the introduction of laparoscopic cholecystectomy (LC) by Mouret, it appeared that there was an increased number of complications of the major bile ducts and gallbladder leakage [3–5]. Although complication rates and classification vary between 1.3 and 40 %, studies showed that the switch to laparoscopic surgery resulted in increased gallbladder leakage [3–7]. During the early years of LC, gallbladder leakage was not considered as a harmful complication. After several years more and more case reports showed that bile leakage and lost stones resulted in formations of abscesses and fistulas in the peritoneal cavity [4–7]. Although complication numbers after gallbladder perforation are low, it is avoidable [3, 4, 7]. At the moment, an important issue is poorly reporting of gallbladder leakage, which negatively influences research to the incidence of gallbladder leakage and its complications [3, 6, 8]. Patient safety is at stake since incomplete reports could result in delayed diagnosis of LC related complications and underestimation of complications during research [3,5]. Therefore, correct reporting of gallbladder leakage and, as necessary, informing patients about possible complications, is advised. Aforementioned is required to gain insight into gallbladder leakage and its consequences [4, 5, 7].

1.2 Previous research on gallbladder leakage

Multiple studies researched precarious phases during surgery with an increased risk of gallbladder rupture. Three phases were identified, namely when traction is given to the gallbladder with a grasper, which is occurring throughout the entire surgery. Additionally, dissection of the gallbladder from the liver bed is a specific procedure with an increased risk for rupture [5]. Impetuous dissection

of the gallbladder from the liver fossa is mentioned as the most common cause of gallbladder perforation [4, 8].

In addition to complications during difficult surgery phases, several articles describe predictive risk factors for gallbladder rupture. Patients who are at risk of gallbladder rupture are patients with gallbladder hydrops due to obstruction, chronic cholecystitis with thickened walls above 7mm and patients who received laparoscopic surgery previously [9]. Nooghabi et al. also mentioned male sex, higher weight, older patients and acute cholecystitis. Since the study is retrospective, peroperative risk factors are determined: the presence of adhesions, challenging dissection of CVS, clip slippage and presence of infected bile and pigment stones [5]. Some of these factors are correlated: previous laparoscopic surgeries and the presence of adhesion, acute or chronic cholecystitis and infected bile. Besides the presence of (pigment) stones makes it more likely that there is obstruction. Some of these factors; male sex, older age, acute cholecystitis, spillage of pigment stones, number and size of stones and location of spilled stones, are also a predictive value for developing complications due to stone spillage [10].

All complications mentioned before are risk factors for gallbladder rupture. These partially correspond to risk factors for a difficult laparoscopic cholecystectomy (DLC). Risk factors for a DLC are impacted stones in gallbladder neck, adhesions around the cystic artery and cystic duct and rupture of the gallbladder. Some identified risk factors, also define what a DLC is, namely injury of the cystic artery, blood loss above 50 mL and surgery time. These are also significantly different if easy and difficult surgeries are compared [11].

Another potential risk factor which could be interesting is the correlation between a surgeon's experience and complication rate. Two recent studies about gallbladder rupture and surgeons experience estimated beforehand that complications could be correlated with surgery experience. Both studies did not find any increased complication rate; only surgery time was increased [8, 11]. On the other hand, older studies found significant differences when gallbladder perforation was compared between experienced surgeons and surgical trainees [12, 13].

2 Clinical problem

Although studies confirmed that gallbladder perforation could result in severe complications and they stated that it should be reported correctly, surgeons still do not always mention gallbladder leakage in surgical reports. Hereby, it is not possible to conduct a comprehensive study on the incidence of complications related to gallbladder rupture. Information about risk factors for gallbladder leakage, distinguishing between easy and difficult LC and the possible effect of surgical experience is available, but more reliable data to confirm these findings is not present. To improve patient safety before, during and after an LC, more feedback and information is needed.

3 Artificial intelligence for LC

To improve reporting of gallbladder leakage, the introduction of Artificial Intelligence (AI) into health care could open new perspectives. Recently more and more papers are published about AI and laparoscopic cholecystectomy. One reason is the large number of surgeries that are performed every year, resulting in a large data set. Another important reason is the availability of two extensive datasets, Cholec80 and EndoVis, containing LC videos with annotation of surgery phase and instrument usage [14, 15]. Thus far, these datasets are used for benchmarking, education, keyframe extraction and predicting the remaining surgery time. Other studies focused on combining these

annotated datasets with external cameras or creating software for more automatically annotation of data [14, 16–21].

Initially, studies focused on the improvement of results of previous phase recognition and instrument usage studies [16]. These two recognition tasks are beneficial for the more difficult task of skill assessment. Benchmarking or skill assessment for surgeons has proven to increase their level of performance [16]. It is achieved by analyzing surgery steps and tasks, instrument usage and additional information about instrument path length, the number of hand motions, usage time of each instrument, applied force and how smoothly movements are [16, 17]. By evaluating these parameters, the learning process of (junior) surgeons is supported. More specifically, it enables personalized training, surgery evaluation and creation of skill-related feedback for (junior) surgeons [16].

Another promising subject is the study of Loukas et al. into keyframe extraction. They managed to extract 81% of the ground truth keyframes by using their trained network. This application is helpful for education, for automatic generation of summaries for surgery reports and it could be used as support for specific training for surgery phase and task recognition [18].

An innovative application of surgery phase information is the calculation of the remaining surgery time. When accurate estimation is possible, the preparations for the next surgery are done more efficiently by notifying staff automatically at the correct time. The use of surgery rooms and medical staff are optimally planned and more patients could be treated with the same health care budget and shortened waiting time [19, 21]. When the use of AI is extended to incorporation of medical record and surgeon specific information, even more accurate estimations could be made [21].

Padoy et al. describe the use of external cameras combined with surgery videos to extract more information about surgery phase en instrument usage. Although new information is added about the surgeon and medical staff their position and movement, it is still difficult to visualize all the members and movements and prove the added value of external cameras for patient outcome and surgery efficiency [19].

3.1 Benchmarking

Although research is done into skill assessment, the development of the surgery robot by Verb surgical and their interest in AI, opens new perspectives. Besides improvement of skill assessment algorithms, there is a need for objective classification of the level of complexity of a surgery. As mentioned before, the definition of a difficult LC surgery is related to the health condition of the patient and the complications that occur during surgery. When it is possible to define what an easy, moderate and difficult LC surgery is, it is possible to determine if surgery times and number of complications are increased compared to other colleagues. Otherwise, increased mean surgery time and number of complications due to a lot of difficult surgeries, could incorrectly mark a surgeon as too slow or even incompetent. Combining the objective level of complexity of a surgery, surgery time, complications like gallbladder leakage and skill assessment, will result in fair benchmarking of surgeons and eventually improve health care.

3.2 Research group Meander Medical Centre and Verb surgical

In the Meander Medical Centre, different projects about AI and surgery are done. The first project, the identification of five anatomical structures; ureter, tendon, artery, white line of Toldt and colon, was completed in August 2018. The next project aimed to remove video frames from surgery videos which contain personal information, most importantly, frames that contain medical staff. Verb surgical, a collaboration between J&J and Google, is interested in this project, which is still ongoing. During multiple conversations, it was decided that this study about bile leakage during

LC surgery would fit in their aim of creating a preoperative risk analysis for each patient, being able to estimate the remaining surgery time and offer benchmarking for surgeons. Another ongoing project is about identification of the Nervus Vagus. During anti-reflux surgery, the Nervus Vagus is injured in around 20 % of the patients. The goal of this study is to identify the nerve during surgery and support the surgeon in preventing collateral damage.

4 Aim

The aim of this study is to detect bile leakage in videos of laparoscopic cholecystectomy surgeries. When the created deep learning network is outperforming the manually reporting of gallbladder leakage, the result is clinically relevant. Only then, the network is suitable for automatic reporting of gallbladder leakage in surgery reports and research into gallbladder complications will become more reliable. The ultimate goal for gallbladder surgery is that reliable preoperative risk assessment for each LC patient is done automatically before the surgical procedure by using previous mentioned high-risk factors. Besides, complications are detected during surgery and are reported automatically. Both surgeon and surgical trainees can learn from a gallbladder perforation, because data of perforation is annotated correctly and therefore available. Additionally, benchmarking, so comparing skills between surgeons, is possible and personalized training sessions will improve skill and speed during specific phases and procedures. However most importantly, quality of care is improved when complications during laparoscopic cholecystectomies are reported correctly and patients are informed about possible postoperative complications. The aim of this study, the identification of gallbladder leakage by using a deep learning network, will be a small contribution to this ultimate goal of improving quality of care for patients who receive a laparoscopic cholecystectomy.

5 Research questions

1. To what extent is it possible to detect bile leakage in laparoscopic cholecystectomy videos?
2. What is the clinical added value of the created network when comparing the performance in bile leakage detection with the reported bile leakage in literature?

Primary objective: To detect gallbladder leakage post-operative in laparoscopic cholecystectomy videos by using deep learning algorithms

Secondary objective: To create an algorithm with a detection accuracy that has more clinical added value in comparison with current standards, based on literature studies. Besides, an extensive parameter study is performed to improve results and understanding of deep learning algorithms.

6 Study population

The study population is a group of patients who underwent laparoscopic cholecystectomy surgery between 01-01-2018 and 31-12-2019. Only the videos which are made during surgery by using the laparoscopic camera, are obtained from the EPD. Because the current set of 70 videos is not sufficient for creating an accurate self-learning network, an additional 120 videos are needed. Hereby a dataset is created with surgeries performed by a diverse group of surgeons and enough anatomical variations between patients. Selecting and annotating videos is time consuming, therefore only 120 video's are included, since annotating will take a month. If collecting data is more time consuming

than initially thought, only 80 videos will be included. To create a balanced dataset, 60 videos contain gallbladder leakage and 60 videos do not contain gallbladder leakage.

Inclusion criteria :

- Only videos of laparoscopic cholecystectomy
- Underwent surgery between 01-01-2018 and 31-12-2019

Exclusion criteria:

- If gallbladder leakage occurs but is difficult to identify
- If less then 60 gallbladder leakage videos are collected, only the same amount of normal laparoscopic cholecystectomy videos are included.

7 Methods

7.1 Study parameters

7.1.1 Primary endpoint

The primary endpoint is the accuracy and loss of the binary classification algorithm for gallbladder leakage detection.

7.1.2 Secondary endpoints

Secondary endpoints are results of a parameter study which is needed to optimize the outcomes of the primary endpoint. Besides, outcomes of literature study are compared to the primary endpoint to decide if there is any clinical value for the created algorithm.

7.1.3 Other study parameters

Another parameter that is included for this research is the presence of gallstones in LC videos. As mentioned earlier, the presence of gallstones in LC videos is a predictive value for extended surgery time and difficult LC and for developing post-operative complications. Therefore detecting gallstones is a useful additional feature, besides detecting of bile leakage. Another study parameter is the dataset size. The primary endpoint will be used to determine if the dataset size is sufficient after adding 120 additional videos.

7.2 Study procedures

7.2.1 Collecting the video data

There is no interference or change to interventions. Data is collected retrospective. The recorded LC videos are obtained by retrieving these videos out of the electronic health record if it meets the inclusion criteria. It is done by the treating physician, namely surgeon, assistant or a person who, commissioned by a surgeon, is assigned as part of the treatment team. They are transferred to a deep learning desktop which is situated in the meeting room of the surgeons of the Meander Medical Centre. This desktop is secured with a password and the room could only be accessed when a person owns a Meander access card. Although videos are automatically anonymized if downloaded, an algorithm which is designed for excluding frames which contain members of the

treatment team or the patient, is used to ensure the whole video is anonymous. These new videos which do not contain personal information, is saved at the computer. Other frames are deleted.

7.3 Creating a deep learning dataset

To create a training set, annotation of images is needed. This is done by annotating the first and last video frame with gallbladder leakage. Hereafter, a script will be used which split the images and save them in different folders. Hereby two different groups are created, one folder with bile leakage and one without bile leakage. It is important to note that there are multiple terms used to describe bile leakage, namely leakage, spillage and gallbladder perforation. Bile spillage is when a minimal amount of bile is leaking out of the gallbladder. When a hole is present in the gallbladder and the bile is pouring out, it is defined as perforation. Both could be described as bile leakage. One major difference is that the occurrence of gallbladder perforation could cause loss of gallstones. For this research, bile spillage and gallbladder perforation are included. Lost stones could be added as a fourth class in this project.

7.4 Training a network

A Convolutional Neural Network (CNN) is a specific type of deep learning network which is suitable for analyzing images. The frames will be divided in three groups, namely a training, validation and test set. These three groups are needed to train the network, validate if it is learning correctly and hereafter use the test set to see how the network is reacting to new images. After training, accuracy and loss are stored, multiple graphs are created of the training session and settings of the created algorithm are stored. If fewer LC videos with gallbladder leakage are available then needed, image augmentation is a suitable solution to increase the dataset size. In advance, it is difficult to determine the required size of the dataset. Therefore, the accuracy and loss graphs are used to determine if the learning curve is flattening with the available dataset size. If not, more data is required. When a larger dataset is available, it is possible to compare the results of different dataset sizes. This could also indicate if the current dataset size is sufficient.

8 Privacy and WMO

For the usage of data that is stored in the hospital patient record systems apply some strict regulations and laws. Two of them consider the use of medical data and privacy; Wet medisch-wetenschappelijk onderzoek met mensen (WMO) and the Algemene verordening gegevensbescherming (AVG or the English version: GDPR) The WMO states if ethical approval is needed by the METC (medical ethical review committee). This is the case if both rules apply to the study:[19] 1) Er sprake is van medisch-wetenschappelijk onderzoek. English: It concerns medical, scientific research 2) Personen worden onderworpen aan handelingen of aan hen wordt een bepaalde gedragswijze opgelegd. English: The patients/participants are subject to procedures or are required to follow rules of behaviour. Only the first rule applies to our study, based on that our study is not WMO plichtig. 1) Informed concent/Privacy: By dutch law (AVG) a persoonsgegevens/personal data is: All personal data from an identified or identifiable natural person. It is considered as information directly about someone or can be traced back to someone. [20] The data is fully anonymised (not pseude anonymised) and thereby it is not a persoonsgegeven. Because no persoonsgegevens/personal data is used, the data can/is legally allowed to be collected without informed consent of the patients. In our conclusion for this study, no approval is needed from the METC, and anonymous video data can be used if the created protocol is followed and no informed consent is needed. Besides the LC

videos, no additional information is stored which could link this videos to the patient. When M.H. Gerkema is graduated and leaving the hospital, the password of the deep learning desktop is known to prof. dr. I.A.M.J. Broeders.

References

[1] Centraal Bureau voor de Statistiek. Operaties in het ziekenhuis; soort opname, leeftijd en geslacht, 1995-2010, 2014.

[2] G.M. Fried and A. Neville. Laparoscopic Cholecystectomy. In L.L. Swanstrom and N.J. Soper, editors, *Mastery of Endoscopic and Laparoscopic Surgery*, chapter 33, pages 339–347. Wolters Kluwer Health, 2 edition, 2013.

[3] U. Kaplan, G. Shpoliansky, O. Abu Hatoum, B. Kimmel, and D. Kopelman. The lost stone — Laparoscopic exploration of abscess cavity and retrieval of lost gallstone post cholecystectomy: A case series and review of the literature. *International Journal of Surgery Case Reports*, 53:43–45, 1 2018.

[4] S. Virupaksha. Consequences of spilt gallstones during laparoscopic cholecystectomy. *The Indian journal of surgery*, 76(2):95–99, 2014.

[5] A.J. Nooghabi, M. Hassanpour, and A. Jangjoo. Consequences of Lost Gallstones during Laparoscopic Cholecystectomy: A Review Article. *Surgical Laparoscopy, Endoscopy and Percutaneous Techniques*, 26(3):183–192, 2016.

[6] A.H. van Dijk, M. van der Hoek, M. Rutgers, P. van Duijvendijk, S.C. Donkervoort, P.R. de Reuver, and M.A. Boermeester. Efficacy of Antibiotic Agents after Spill of Bile and Gallstones during Laparoscopic Cholecystectomy. *Surgical Infections*, 20(4):298–304, 2019.

[7] J. Zehetner, A. Shamiyeh, and W. Wayand. Lost gallstones in laparoscopic cholecystectomy: all possible complications. *American Journal of Surgery*, 193(1):73–78, 1 2007.

[8] Y.E. Altuntas, M. Oncel, M. Haksal, M. Kement, E. Gundogdu, N. Aksakal, and F.C. Gezen. Gallbladder perforation during elective laparoscopic cholecystectomy: Incidence, risk factors, and outcomes. *Northern clinics of Istanbul*, 5(1):47, 2018.

[9] P. De Simone, R. Donadio, and D. Urbano. The risk of gallbladder perforation at laparoscopic cholecystectomy. *Surgical Endoscopy*, 13(11):1099–1102, 1999.

[10] R.J.L.F Loffeld. The consequences of lost gallstones during laparoscopic cholecystectomy. *Netherlands Journal of Medicine*, 64(10):364–366, 2006.

[11] H.M. Atta, A.A. Mohamed, A.M. Sewefy, A.F.S. Abdel-Fatah, M.M. Mohammed, and A.M. Atiya. Difficult Laparoscopic Cholecystectomy and Trainees: Predictors and Results in an Academic Teaching Hospital. *Gastroenterology Research and Practice*, 2017, 2017.

[12] L Sarli, N Pietra, R Costi, and M Grattarola. Gallbladder perforation during laparoscopic cholecystectomy. *World journal of surgery*, 23(11):1186–90, 1999.

[13] C. Barrat, A. Champault, L. Matthyssens, and G. Champault. Iatrogenic perforation of the gallbladder during laparoscopic cholecystectomy does not influence the prognosis. Prospective study. *Annales de chirurgie*, 129(1):25–29, 2004.

[14] A.P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, and N. Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97, 2016.

[15] R. Stauder, D. Ostler, M. Kranzfelder, S. Koller, H. Feussner, and N. Navab. The TUM LapChole dataset for the M2CAI 2016 workflow challenge, 10 2016.

[16] C. Loukas. Video content analysis of surgical procedures. *Surgical Endoscopy*, 32(2):553–568, 2018.

[17] Z. Wang and A. Majewicz Fey. Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. *International Journal of Computer Assisted Radiology and Surgery*, 13(12):1959–1970, 2018.

[18] C. Loukas, C. Varytimidis, K. Rapantzikos, and M.A. Kanakis. Keyframe extraction from laparoscopic videos based on visual saliency detection. *Computer Methods and Programs in Biomedicine*, 165:13–23, 2018.

[19] N. Padoy. Machine and deep learning for workflow recognition during surgery. *Minimally Invasive Therapy and Allied Technologies*, 28(2):82–90, 2019.

[20] S. Bodenstedt, D. Rivoir, A. Jenke, M. Wagner, M. Breucha, B. Müller-Stich, S.T. Mees, J. Weitz, and S. Speidel. Active learning using deep Bayesian networks for surgical workflow analysis. *International Journal of Computer Assisted Radiology and Surgery*, 14(6):1079–1087, 2019.

[21] A.P. Twinanda, G. Yengera, D. Mutter, J. Marescaux, and N. Padoy. RSDNet: Learning to Predict Remaining Surgery Duration from Laparoscopic Videos Without Manual Annotations. *IEEE Transactions on Medical Imaging*, 38(4):1069–1078, 2018.

APPENDIX B

Result section

B.1 Parameter study

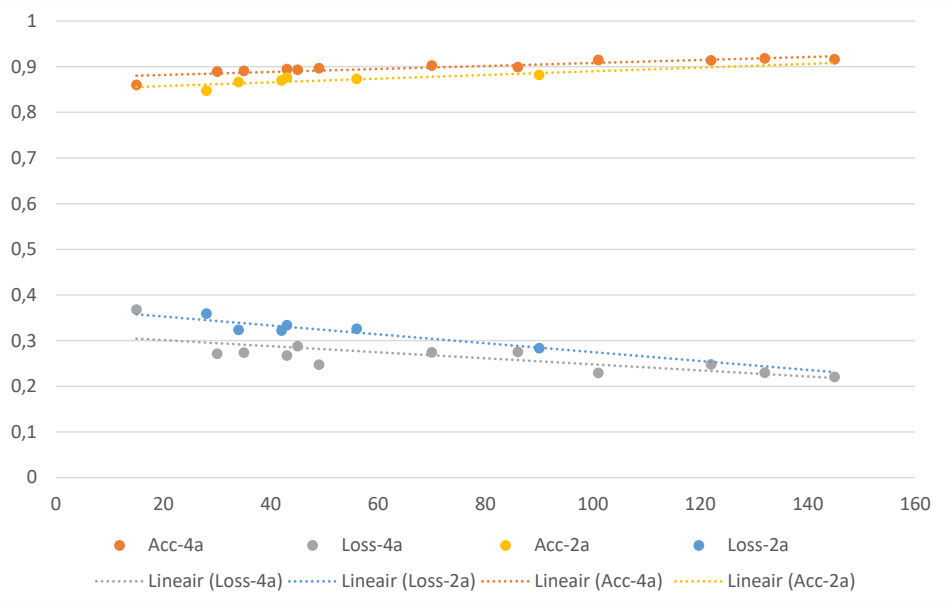


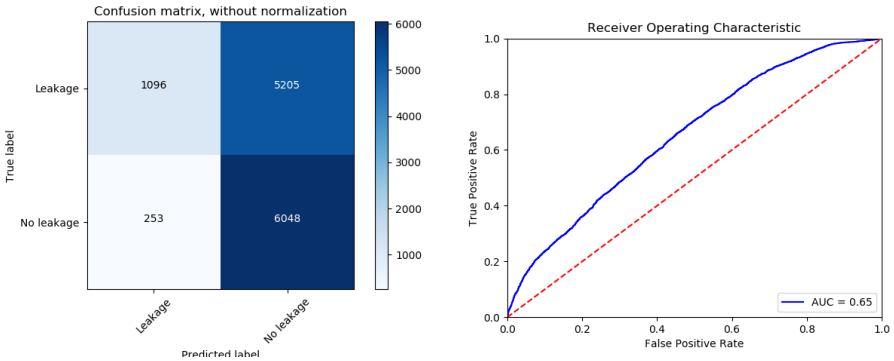
FIGURE B.1: Epochs vs. accuracy and loss

TABLE B.1: Four models for parameter study

Binary model	mean-loss	mean-acc	epochs
1a	0,5127	0,8029	23,25
2a	0,3459	0,8599	58,5
3a	0,4378	0,8339	53
4a	0,2575	0,8960	61

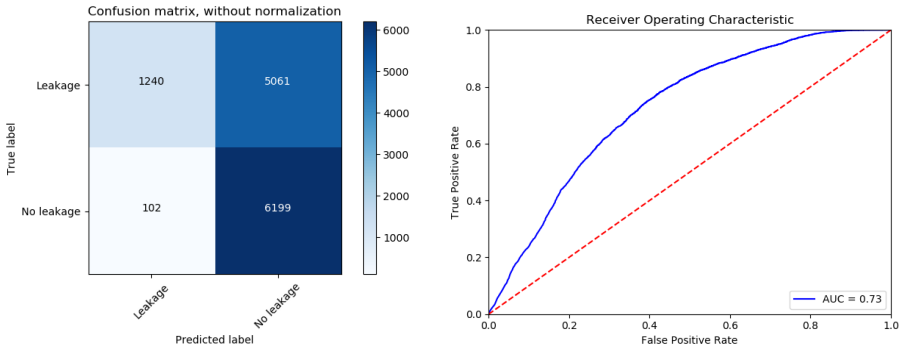
B.2 Binary classification

B.2.1 Evaluation of trained models by using the M1 dataset



(A) Confusion matrix (B) ROC curve

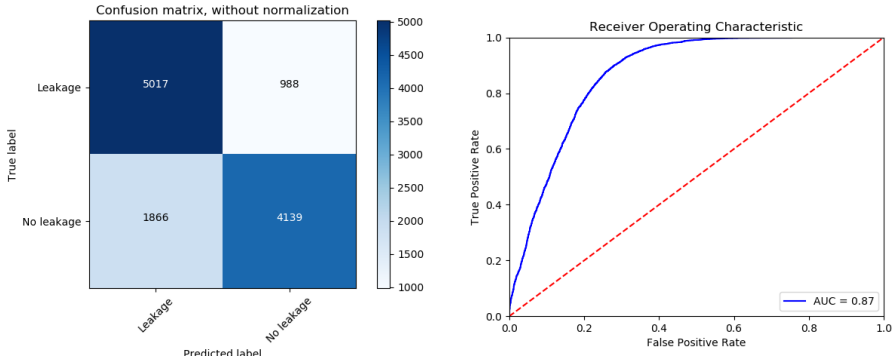
FIGURE B.2: Second best model 4



(A) Confusion matrix (B) ROC curve

FIGURE B.3: Best training model 3

B.2.2 Merged dataset

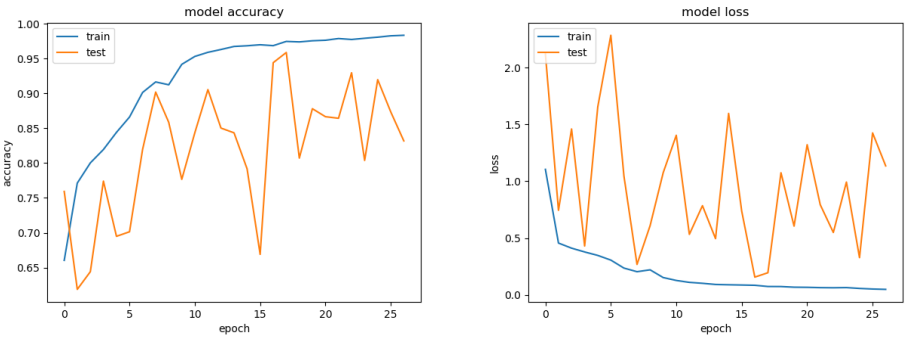


(A) Confusion matrix (B) ROC curve

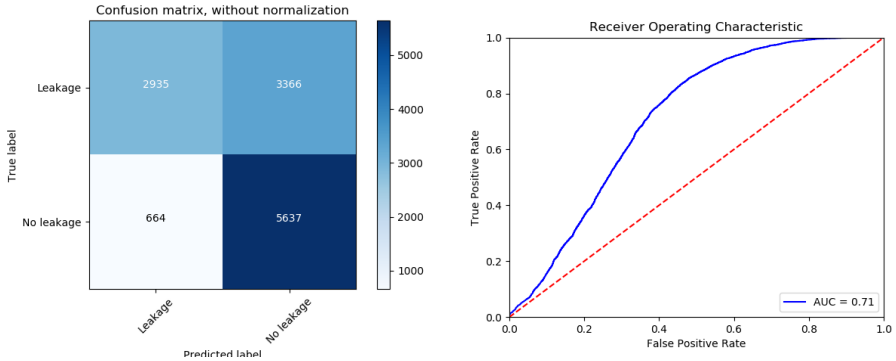
FIGURE B.4: Testing with M2 dataset after training model 4 with Merged dataset

B.3 Colour based feature extraction

B.3.1 Evaluation of training model 3 and 4 by testing with the CBFE M1 dataset

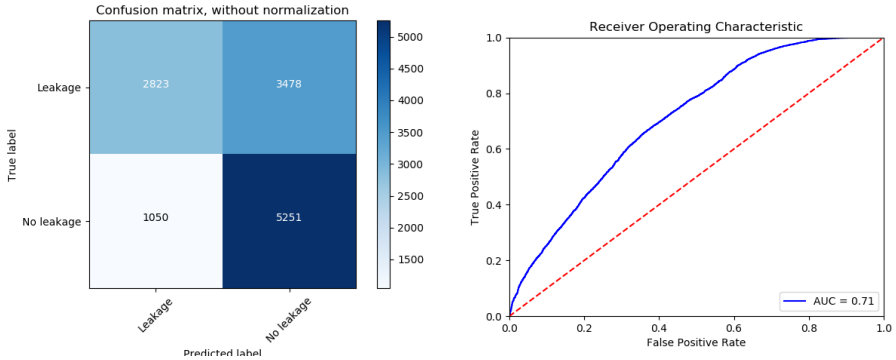


(A) Accuracy during training (B) Loss during training



(C) Confusion matrix (D) ROC curve

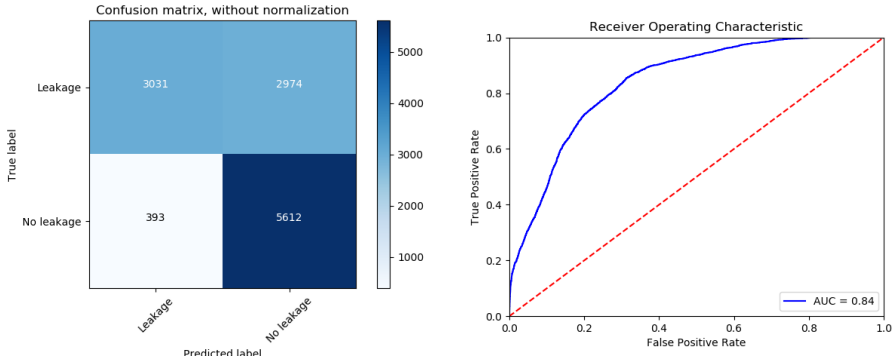
FIGURE B.5: Second best training of model 3 with CBFE images and testing with transformed M1 dataset



(A) Confusion matrix (B) ROC curve

FIGURE B.6: Best training and testing results for CBFE images with model 4

B.3.2 Evaluation of training of the CBFE Merged dataset by using the M2 dataset



(A) Confusion matrix (B) ROC curve

FIGURE B.7: Testing results of trained model 3 with CBFE merged dataset

[1] Centraal Bureau voor de Statistiek, "Operaties in het ziekenhuis; soort opname, leeftijd en geslacht, 1995-2010," 2014. [Online]. Available: https://opendata.cbs.nl/statline/#/CBS/nl/dataset/80386ned/table?dl=1615B

[2] G. Fried and A. Neville, "Laparoscopic Cholecystectomy," in *Mastery of Endoscopic and Laparoscopic Surgery*, 2nd ed., L. Swanstrom and N. Soper, Eds. Wolters Kluwer Health, 2013, ch. 33, pp. 339–347. [Online]. Available: https://books.google.nl/books?id=N1AtBAAAQBAJ

[3] A. Polychronidis, P. Laftsidis, A. Bounovas, and C. Simopoulos, "Twenty years of laparoscopic cholecystectomy: Philippe Mouret–March 17, 1987," *JSLS : Journal of the Society of Laparoendoscopic Surgeons*, vol. 12, no. 1, pp. 109–111, 2008. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/18402752

[4] U. Kaplan, G. Shpoliansky, O. Abu Hatoum, B. Kimmel, and D. Kopelman, "The lost stone Laparoscopic exploration of abscess cavity and retrieval of lost gallstone post cholecystectomy: A case series and review of the literature," *International Journal of Surgery Case Reports*, vol. 53, pp. 43–45, 1 2018.

[5] S. Virupaksha, "Consequences of spilt gallstones during laparoscopic cholecystectomy," *The Indian journal of surgery*, vol. 76, no. 2, pp. 95–99, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24891771https://www.ncbi.nlm.nih.gov/pubmed/24891771

[6] A. Nooghabi, M. Hassanpour, and A. Jangjoo, "Consequences of Lost Gallstones during Laparoscopic Cholecystectomy: A Review Article," *Surgical Laparoscopy, Endoscopy and Percutaneous Techniques*, vol. 26, no. 3, pp. 183–192, 2016. [Online]. Available: www.surgical-laparoscopy.comwww.surgical-laparoscopy.com

[7] A. van Dijk, M. van der Hoek, M. Rutgers, P. van Duijvendijk, S. Donkervoort, P. de Reuver, and M. Boermeester, "Efficacy of Antibiotic Agents after Spill of Bile and Gallstones during Laparoscopic Cholecystectomy," *Surgical Infections*, vol. 20, no. 4, pp. 298–304, 2019.

[8] J. Zehetner, A. Shamiyeh, and W. Wayand, "Lost gallstones in laparoscopic cholecystectomy: all possible complications," *American Journal of Surgery*, vol. 193, no. 1, pp. 73–78, 1 2007.

[9] Y. Altuntas, M. Oncel, M. Haksal, M. Kement, E. Gundogdu, N. Aksakal, and F. Gezen, "Gallbladder perforation during elective laparoscopic cholecystectomy: Incidence, risk factors, and outcomes," *Northern clinics of Istanbul*, vol. 5, no. 1, p. 47, 2018.

[10] U. Sarpel, "Cholecystectomy," in *Surgery: An Introductory Guide*, U. Sarpel, Ed. New York, NY: Springer New York, 2014, pp. 65–76. [Online]. Available: https://doi.org/10.1007/978-1-4939-0903-2_7

[11] G. Wind and M. Dudai, "The Biliary System," in *Applied Laparoscopic Anatomy: Abdomen and Pelvis*. Lippincott Williams & Wilkins, 1997, pp. 36–84.

[12] J. Lange and G. Kleinrensink, "The gallbladder and bile ducts," in *Surgical Anatomy of the Abdomen*, 1st ed. Elsevier, 2002, ch. 10, p. 274.

[13] P. De Simone, R. Donadio, and D. Urbano, "The risk of gallbladder perforation at laparoscopic cholecystectomy," *Surgical Endoscopy*, vol. 13, no. 11, pp. 1099–1102, 1999.

[14] R. Loffeld, "The consequences of lost gallstones during laparoscopic cholecystectomy," *Netherlands Journal of Medicine*, vol. 64, no. 10, pp. 364–366, 2006.

[15] H. Atta, A. Mohamed, A. Sewefy, A. Abdel-Fatah, M. Mohammed, and A. Atiya, "Difficult Laparoscopic Cholecystectomy and Trainees: Predictors and Results in an Academic Teaching Hospital," *Gastroenterology Research and Practice*, vol. 2017, 2017.

[16] L. Sarli, N. Pietra, R. Costi, and M. Grattarola, "Gallbladder perforation during laparoscopic cholecystectomy," *World journal of surgery*, vol. 23, no. 11, pp. 1186–90, 1999.

[17] C. Barrat, A. Champault, L. Matthyssens, and G. Champault, "Iatrogenic perforation of the gallbladder during laparoscopic cholecystectomy does not influence the prognosis. Prospective study," *Annales de chirurgie*, vol. 129, no. 1, pp. 25–29, 2004.

[18] A. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, and N. Padoy, "Endonet: a deep architecture for recognition tasks on laparoscopic videos," *IEEE transactions on medical imaging*, vol. 36, no. 1, pp. 86–97, 2016.

[19] R. Stauder, D. Ostler, M. Kranzfelder, S. Koller, H. Feussner, and N. Navab, "The TUM LapChole dataset for the M2CAI 2016 workflow challenge," *ArXiv*, vol. abs/1610.09278, 10 2016.

[20] C. Loukas, "Video content analysis of surgical procedures," *Surgical Endoscopy*, vol. 32, no. 2, pp. 553–568, 2018.

[21] Z. Wang and A. Majewicz Fey, "Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, no. 12, pp. 1959–1970, 2018.

[22] C. Loukas, C. Varytimidis, K. Rapantzikos, and M. Kanakis, "Keyframe extraction from laparoscopic videos based on visual saliency detection," *Computer Methods and Programs in Biomedicine*, vol. 165, pp. 13–23, 2018.

[23] N. Padoy, "Machine and deep learning for workflow recognition during surgery," *Minimally Invasive Therapy and Allied Technologies*, vol. 28, no. 2, pp. 82–90, 2019.

[24] S. Bodenstedt, D. Rivoir, A. Jenke, M. Wagner, M. Breucha, B. Müller-Stich, S. Mees, J. Weitz, and S. Speidel, "Active learning using deep Bayesian networks for surgical workflow analysis," *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 6, pp. 1079–1087, 2019.

[25] A. Twinanda, G. Yengera, D. Mutter, J. Marescaux, and N. Padoy, "RSDNet: Learning to Predict Remaining Surgery Duration from Laparoscopic Videos Without Manual Annotations," *IEEE Transactions on Medical Imaging*, vol. 38, no. 4, pp. 1069–1078, 2018.

[26] J. Torres, *First Contact with Deep Learning: Practical Introduction with Keras*, ser. Barcelona Series. Independently Published, 2018. [Online]. Available: https://books.google.nl/books?id=oHK2uQEACAAJ

[27] J. Brownlee and M. L. Mastery, *Deep Learning with Python: Develop Deep Learning Models on Theano and TensorFlow Using Keras*. Machine Learning Mastery, 2017. [Online]. Available: https://books.google.nl/books?id=eJw2nQAACAAJ

[28] S. Bhattarai, "What is activation functions in neural networks (NN)?" 2018. [Online]. Available: https://saugatbhattarai.com.np/what-is-activation-functions-in-neural-network-nn/

[29] D. Liu, "A Practical Guide to ReLU," 2017. [Online]. Available: https://medium.com/@danqing/a-practical-guide-to-relu-b83ca804f1f7

[30] Keras, "Core Layers." [Online]. Available: https://keras.io/layers/core/

[31] P. Radhakrishnan, "What are Hyperparameters ? and How to tune the Hyperparameters in a Deep Neural Network?" 2017.

[32] M. Stewart, "Simple Guide to Hyperparameter Tuning in Neural Networks," 2019. [Online]. Available: https://towardsdatascience.com/simple-guide-to-hyperparameter-tuning-in-neural-networks-3fe03dad8594

[33] F. Van der Heijden, R. P. W. Duin, D. de Ridder, and D. M. J. Tax, *Classification, Parameter Estimation and State Estimation: An Engineering Approach using MATLAB*. Chichester: Wiley, 9 2004.

[34] J. Brownlee, "A Gentle Introduction to Mini-Batch Gradient Descent and How to Configure Batch Size," 2017. [Online]. Available: https://machinelearningmastery.com/gentle-introduction-mini-batch-gradient-descent-configure-batch-size/

[35] S. K. Zhou, H. Greenspan, and D. Shen, *Deep Learning for Medical Image Analysis*. Elsevier Science, 2017. [Online]. Available: https://books.google.nl/books?id=WVqfDAAAQBAJ

[36] K. Bokka, "Guide to choosing Hyperparameters for your Neural Networks," 2019. [Online]. Available: https://towardsdatascience.com/guide-to-choosing-hyperparameters-for-your-neural-networks-38244e87dafe

[37] J. Brownlee, "How to Choose Loss Functions When Training Deep Learning Neural Networks," 2019. [Online]. Available: https://machinelearningmastery.com/how-to-choose-loss-functions-when-training-deep-learning-neural-networks/

[38] ——, "Loss and Loss Functions for Training Deep Learning Neural Networks," 2019. [Online]. Available: https://machinelearningmastery.com/loss-and-loss-functions-for-training-deep-learning-neural-networks/

[39] P. Sharma, "Improving Neural Networks Hyperparameter Tuning, Regularization, and More," 2018. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/11/neural-networks-hyperparameter-tuning-regularization-deeplearning/

[40] M. Stewart, "Neural Network Optimization," 2019. [Online]. Available: https://towardsdatascience.com/neural-network-optimization-7ca72d4db3e0

[41] J. Brownlee, "Understand the Impact of Learning Rate on Neural Network Performance," 2019. [Online]. Available: https://machinelearningmastery.com/understand-the-dynamics-of-learning-rate-on-deep-learning-neural-networks/

[42] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[43] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[44] J. Brownlee, "A Gentle Introduction to Early Stopping to Avoid Overtraining Neural Networks," 2018. [Online]. Available: https://machinelearningmastery.com/early-stopping-to-avoid-overtraining-neural-network-models/

[45] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *32nd International Conference on Machine Learning, ICML 2015*, 2015.

[46] J. Brownlee, "A Gentle Introduction to Batch Normalization for Deep Neural Networks," 2019. [Online]. Available: https://machinelearningmastery.com/batch-normalization-for-training-of-deep-neural-networks/

[47] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization?" in *Advances in Neural Information Processing Systems*, 2018.

[48] J. Brownlee, "How to Configure Image Data Augmentation in Keras," 2019. [Online]. Available: https://machinelearningmastery.com/how-to-configure-image-data-augmentation-when-training-deep-learning-neural-networks/

[49] P. Dangeti, C. Chung, and A. Yu, *Numerical Computing with Python*. Packt Publishing, 2018. [Online]. Available: https://books.google.nl/books?id=tVA8vwEACAAJ

[50] J. N. Mandrekar, "Receiver Operating Characteristic Curve in Diagnostic Test Assessment," *Journal of Thoracic Oncology*, vol. 5, no. 9, pp. 1315–1316, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1556086415306043

[51] G. Van Rossum and F. L. Drake Jr, *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[52] F. Chollet *et al.*, "Keras," 2015. [Online]. Available: https://keras.io

[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[54] A. C. Müller, M. A. C, and S. Guido, *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media, 2016. [Online]. Available: https://books.google.nl/books?id=1-4lDQAAQBAJ

[55] R. Karim, "Illustrated: 10 CNN Architectures A compiled visualisation of the common convolutional neural networks," 2019. [Online]. Available: https://towardsdatascience.com/illustrated-10-cnn-architectures-95d78ace614d

[56] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.