AE-CAI — CARE — OR 2.0 WORKSHOP PAPER

Towards an AI-Based Assessment Model of Surgical Difficulty During Early Phase Laparoscopic Cholecystectomy

Julian. R. Abbing[a, b] and Frank J. Voskens[a, b] and Beerend G. A. Gerats[a, b] and Ruby M. Egging[a, b] and Fausto Milletarì[c] and Ivo A.M.J. Broeders[a, b]

[a]Meander Medical Centre, Maatweg 3, Amersfoort 3813 TZ, The Netherlands; [b]University of Twente, Drienerlolaan 5 7522 NB Enschede, The Netherlands; [c]Johnson & Johnson Germany

ARTICLE HISTORY
Compiled September 12, 2022

ABSTRACT
Laparoscopic cholecystectomy (LC) is one of the most commonly performed minimally invasive (abdominal) procedures. It is the standard surgical treatment for patients with gallstone disease, ranging from symptomatic cholelithiasis to severe cholecystitis. Although complication rates are low, this procedure can be difficult due to severe inflammation or adhesions. As there is high variability in operative findings during LC, it is important to assess the surgical difficulty objectively. This is critical for further developing and using future artificial intelligence algorithms in LC surgery, as it allows more reliable benchmarking between surgeons and can help in adequate surgical OR planning. In this study, deep learning models were trained to assess the level of difficulty in the first phase of the procedure. First, 93 LC videos recorded at the Meander Medical Center were included and sampled at one fps. A modified Nassar scale was composed and used to annotate gallbladder difficulty (grade 1-3) and adhesion presence (grade 1-3). Various models, including Resnet-18, Resnet-50 and EfficientnetV2, were trained on different label combinations (binary and multi-label). On the multi-label test set, the best model reached accuracies of 66% and 40% for the classification of the gallbladder and adhesions, respectively. With binary labels, the best model classifies gallbladder difficulty grade 3 vs 1,2 (labels 1,2 were combined) with an accuracy of 88%, and gives an accuracy of 82% for grade 1 vs 2,3 (labels 2,3 were combined). This work shows the potential of difficulty understanding in surgical scenery based on early phase endoscopic video footage.

KEYWORDS
Surgical Difficulty Prediction; Laparoscopic Cholecystectomy; Resnet-50; EfficientnetV2; Deep Learning

1. Introduction

Laparoscopic cholecystectomy (LC) is currently one of the most commonly performed minimally invasive (abdominal) procedures in elective and emergency settings. (Sugrue et al. 2019) The procedure is considered the standard surgical treatment for patients with symptomatic cholelithiasis, cholecystitis, gallstone pancreatitis and gallbladder polyps. (Priego et al. 2009; Soper and Malladi 2022) The procedure has evolved into a relatively safe and tolerable daycare procedure. However, there is high variability in

CONTACT Julian R. Abbing. Email: jr.abbing@meandermc.nl

perioperative findings. (Griffiths et al. 2019; Ward et al. 2022) LC can be relatively easy in patients with a floppy, non-inflamed pink gallbladder. However, the procedure can be challenging in patients with dense adhesions and severe cholecystitis. These high complex procedures can result in prolonged OR time, more blood loss, and an increased risk of complications. (Atta et al. 2017; Sugrue et al. 2019; Madni et al. 2018)

One should consider these intra-operative conditions when including LC video recordings in an artificial intelligence computer vision (CV) model, as it affects the surgical course. There is a need for an objective assessment of surgical difficulty to allow a better understanding of LC procedures. Additionally, when comparing surgeons in a benchmark, it is important to consider both surgical performance and the level of difficulty of the considered procedures. The range of performed difficulty levels differs per surgeon, depending on their experience level, their degree of specialization and the institution where they are employed. For example, starting residents generally perform easier LC procedures and proceed to more difficult cases later in their education. Experienced surgeons could also perform the very difficult cases or the mean or easy cases, which makes comparing them on their performance only very challenging. Automatic difficulty prediction from laparoscopic video could enable more reliable benchmarking. Suppose the resident or surgeon's experience level is known. In that case, we can also schedule patient better on their difficulty score (surgeon-patient matching) or might help to estimate the procedure time (based on experience + difficulty of the patient).

Although a lot of surgical scene understanding research is conducted mainly on; the phase detection (Twinanda et al. 2016), tool recognition (Ali et al. 2022), safety steps/adverse event (prevention) (Mascagni et al. 2022), surgical task/skill assessment (Lam et al. 2022). The majority of those techniques uses video data to generate the model outcomes.

In this work, we elaborate on developing and assessing deep learning for surgical difficulty prediction.

2. Methods

To perform difficulty classification experiments, a dataset was created of LC videos. Those frames were labelled on surgical difficulty based on existing grading features. We trained different models to classify the grading score for gallbladder and adhesion difficulty. Furthermore, grading labels were combined to allow binary difficulty classification: easy versus difficult. Finally, we analysed the models on performance metrics, k-fold cross-validation, degraded input response and used Class Activation Mapping (CAM) to perform a qualitative evaluation of the model's decision process.

2.1. Dataset and Labelling

2.1.1. Dataset

We collected 257 LC videos that were conducted in the Meander Medical Center, Amersfoort, The Netherlands. All procedures took place between 1-1-2018 and 10-10-2021 and were fully anonymised. The videos were filtered (missing/incomplete first phase of the procedure, bad video quality or limited recording); 164 were excluded, 93 videos were used. The study was approved by the local Institutional Review Board of the Meander Medical Center (Protocol No: TWO 21-081). LC videos were edited

to show only the first phase of the surgical procedure. This phase started at the first moment the fundus of the gallbladder or adhesions on the gallbladder was grasped. The endpoint was set prior to the start of the dissection of Calot. Finally, the videos were converted to images at one fps to enable labelling. Labelling was performed by a technical medicine student and a surgeon.

2.1.2. Annotation with the Nassar scale

For labelling, we modified an existing grading system, the *Nassar Scale*, which has already been validated in relation to surgical outcome data. (Griffiths et al. 2019; Yokoe et al. 2018) Various grading scales exist based on anatomical findings, such as the *Parkland scale* or cholecystitis severity scale by Tang *et al.* (Tang and Cuschieri 2006) However, these grading scales are less suitable for labelling. The advantage of the Nassar scale is its use of separate categories (e.g. gallbladder, adhesions, cystic pedicle). Furthermore, the Parkland scale is only validated on a small set of patients, and Tang *et al.* dates from procedures that were conducted between 1995 and 2005, which might be outdated considering that the tool(set) changed over time. (Tang and Cuschieri 2006; Madni et al. 2018)

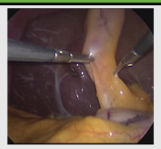
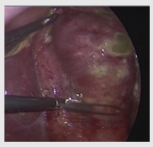
Modifications on the Nassar scale were made by the authors to support objective and consistent labelling, early detection of surgical difficulty, and clinical relevance. This resulted in merging similar gradings. Our label guide is provided in Table 2. Merging grades resulted in a scoring system change from 1–4 to 1–3 on the gallbladder and adhesions. The cystic pedicle grading scale was not used in this study. We included a label for the out-of-body frames (excl) and frames where no gallbladder was visible (0), for example, during a diagnostic laparoscopy. These frames (excl and 0) were excluded from training, validation and testing. A total of 2270 frames were excluded due to out-of-body footage or where gallbladder/adhesions were not scorable).

The videos were split into a train, validation and test set, where the class distribution was taken into account to ensure an equal distribution of labels. This was performed manually on a video level because of the intense class imbalance if selected randomly. The resulting label distribution of the frames is visualised in Table 1. The video combination used for the gallbladder labels was different from those for the adhesion labels. Also, labels 1 and 2 or labels 2 and 3 were combined for binary classification.

Table 1.: Label distribution of the different label sets. Note that in the binary models, labels 1 and 2 or labels 2 and 3 were combined.

Label	**Training Data**	**Validation Data**	**Testing Data**
Gallbladder	16739	2338	4790
I (easy)	5535	746	1542
II (moderate)	4246	594	1192
III (difficult)	6958	998	2056
Adhesion	14669	3410	4868
I (easy)	5183	1155	1562
II (moderate)	3082	1025	1211
III (difficult)	6404	1230	2095

Table 2.: Annotation guide for surgical difficulty in laparoscopic cholecystectomy. This can be applied to the first phase, which **starts** at the moment that the fundus of gallbladder or the adhesions on gallbladder are grasped, and **ends** the frame before the first dissection of Calot Triangle.

Grade	Gallbladder		Adhesions	
I (easy)	Floppy, thin, gray/pink or fat laden		No adhesions, or simple adhesions up to Hartmann's pouch	
II (moderate)	Mucocele, hydropic or packed with big stones		Simple anatomical adhesions	
III (difficult)	Cholecystitis. empyema, gangene, fibrosis		Pathological dense adhesions and /or completely obsucrred, or involvement duodenum	

2.2. Training and Evaluation

2.2.1. Models

In our experiments, we compare the performance of various basic convolutional neural networks, including *Resnet-18*, *Resnet-50* and *EfficientnetV2*. (He et al. 2016; Tan and Le 2021) The Resnet models are pre-trained on ImageNet (Pytorch 2022), while EfficientnetV2 is pre-trained on ImageNet-21k and fine-tuned on 1k. (Wightman 2022) Class weighting was applied due to class imbalance. Models were trained until the loss stabilized and stopped before the network overfitting. Per label combination (see Figure 2), the best model is selected based on validation accuracy (see Equation 1) and is used for further analysis. As visualised in Figure 1, with the multi-label models we endeavour to classify the actual grade of the *Nassar* feature. With the binary labels, we only allow the models to classify easy versus difficult cases.

2.2.2. Evaluation

2.2.2.1. Metrics. During training and evaluation logging the *weights and biases* package is used in combination with *Scikit-Learn* package functions to calculate the metrics. (Biewald 2020; Pedregosa et al. 2011) The metrics used are the Accuracy, Precision, Recall and F1 score (see Equation 1,2,3 and 4) calculated of the frame-wise predictions of the model. Here the TP are the true positive classifications of the frames, TN the true negatives, FP false positives and FN the false negatives. These are also used to generate the (normalised) confusion matrixes. (Pedregosa et al. 2011)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

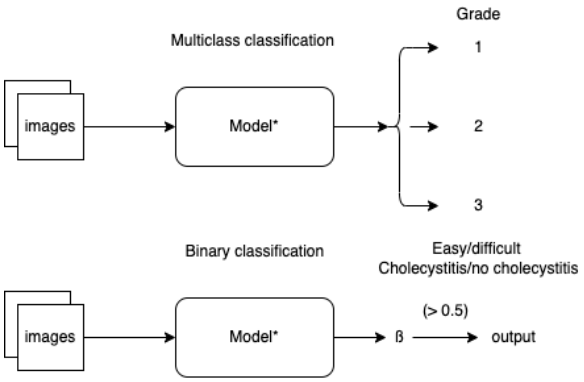


Figure 1.: Two setups are used; multiclass classification and binary classification. The models used are Resnet18(He et al. 2016), Resnet50 (He et al. 2016) and EfficientnetV2 (Tan and Le 2021).

Figure 2.: Label set combinations used for training the different models.

Classification task	**Gallbladder**	**Adhesions**
	Labels	
Multiclass	1, 2, 3	1, 2, 3
Binary	1 vs. 2/3	1 vs 2/3
	1/2 vs. 3	1/2 vs/ 3

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{4}$$

2.2.2.2. Model evaluation. **K-fold cross-validation** Every model was evaluated on its final hyperparameters with a 5 fold cross-validation. Because we use such a small dataset, this is performed to evaluate the model performance on the dataset without (manual) distribution of the labels/videos. (Pedregosa et al. 2011; sklearn 2022)
Input response. For the best model, we evaluate the model response based on different types of disturbance in image quality, such as; adding noise, blur or lowering the resolution. This is performed on two levels; medium degradation and hard degradation.

2.2.2.3. Class Activation Mapping (CAM). Because we label the images with visual-based grading features, also class activation maps (CAM) are generated. Class activation maps allow visualising what regions on the model's input contribute the most to the output. (Gagana et al. 2019) We generate the CAM for the winning class that is predicted and only for the multi-class Resnet models.

3. Results

3.0.1. Training results

An overview of the training results of the best-trained models is presented in Table 3 for multi-class labels and in Table 4 for binary labels. The performance of the model

trained on the multi-class labels remains low, with an accuracy of 66.6% for gallbladder and 40.3% for adhesions. Interestingly, the highest performance, in terms of F1 score, is seen for grade 3 of the difficulty scale for both gallbladder and adhesions.

The high classification scores for the gallbladder for the multi-class labels can also be seen for the binary model. More specifically, when grade 3 is set out against grades 1 and 2, the model performs slightly better than when grades 2 and 3 are combined and set out against grade 1. With the former label set, which can be seen as classes "non-cholecystitis" (1/2) and "cholecystitis" (3), the model achieves the highest accuracy: 88.2%. As can be expected from the precision and recall figures in Table 4, the normalised confusion matrices show the highest values on the diagonal. The performance of the model on adhesions remains low for binary labels.

Table 3.: Results of the multi-class classification. The metrics given are on the test data unless indicated otherwise. (**) is the test set accuracy.

	Grade	**Validation accuracy**	**Accuracy** **	**Precision**	**Recall**	**F1**	**best model**
Gallbladder	**I**	81.8%	66.6%	52.3%	91.0%	66.4%	EfficientnetV2
	II			41.0%	64.3%	50.1%	
	III			91.5%	84.7%	**87.9%**	
Adhesions	**I**	61.1%	40.3%	30.6%	77.2%	43.8%	Resent-18
	II			18.7%	60.0%	28.5%	
	III			75.9%	46.6%	57.7%	

3.0.2. Model evaluation; Results K-fold cross-validation

Model evaluation; degraded input response

Table 4.: Results of the binary models on the different label sets. The metrics given are on the test data unless indicated otherwise. (*) label set can be seen as "non-cholecystitis" vs "cholecystitis". (**) is the test set accuracy.

	Label set	**Validation accuracy**	**Accuracy****	**Precision**	**Recall**	**F1**	**Best model**
Gall-bladder	I, II vs. III*	94.4%	88.2%	92.6%	78.8%	85.1%	Resnet-50
Gall-bladder	I vs. II, III	86.8%	82.9%	98.3%	76.0%	85.7%	Resnet-50
Adhesions	I, II vs. III	74.1%	81.8%	90.0%	64.8%	75.3%	Resnet-50
Adhesions	I vs II, III	77.2%	54.3%	82.6%	41.4%	55.2%	EfficientnetV2

The results of the 5-fold cross-validation are visualised in Table 5. It can be observed that the highest accuracy is achieved on the Gallbladder models when comparing them either the binary models, multi-label or all together. The results of the mean accuracy are lower than in the results of the manually balanced dataset in Table 3 and Table 4 There can be a low accuracy observed with a smaller standard deviation in the adhesion

Table 5.: 5-fold cross-validation results.

Replication	Fold					Mean	Std. dev
	1	2	3	4	5		
Gallbaldder multi	0.65	0.79	0.86	0.58	0.75	0.78	0.111
Adhesion multi	0.58	0.58	0.70	0.62	0.65	0.63	0.048
Gallbladder I, II vs. III	0.84	0.86	0.71	0.77	0.96	0.83	0.094
Gallbladder I vs. II, III	0.72	0.92	0.89	0.96	0.85	0.87	0.092
Adhesion I, II vs. III	0.65	0.71	0.74	0.90	0.77	0.75	0.091
Adhesion I vs. II, III	0.92	0.86	0.60	0.69	0.83	0.78	0.128

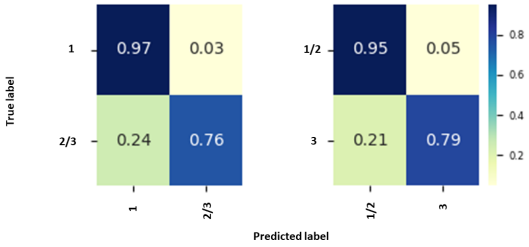


Figure 3.: Normalised confusion matrix on the test set of the Gallbladder binary model; label I, II vs III (No cholecystitis vs. cholecystitis). Normalised confusion matrix on the test set of the Gallbladder binary model; label I vs II, III (easy, vs. moderate/difficult)

model (multi-label and I,II vs III). It should be mentioned that this can be caused when model performance is poor, due to (mis)classification focus on one class.

Table 6.: Model response on different degraded image inputs

Test accuracy							
Image degradation type and level (appl. on input image 224x224)	No noise	Noise		Blur		Resolution	
		Mean / std		kernel/sigma		fraction of original	
Model		0./0.1	0./0.3	9/0.5	9/5	1/4	1/8
Gallbladder model (I, II vs. III)	88.2	83.2	42.9	89.5	70.8	86.8	49.7

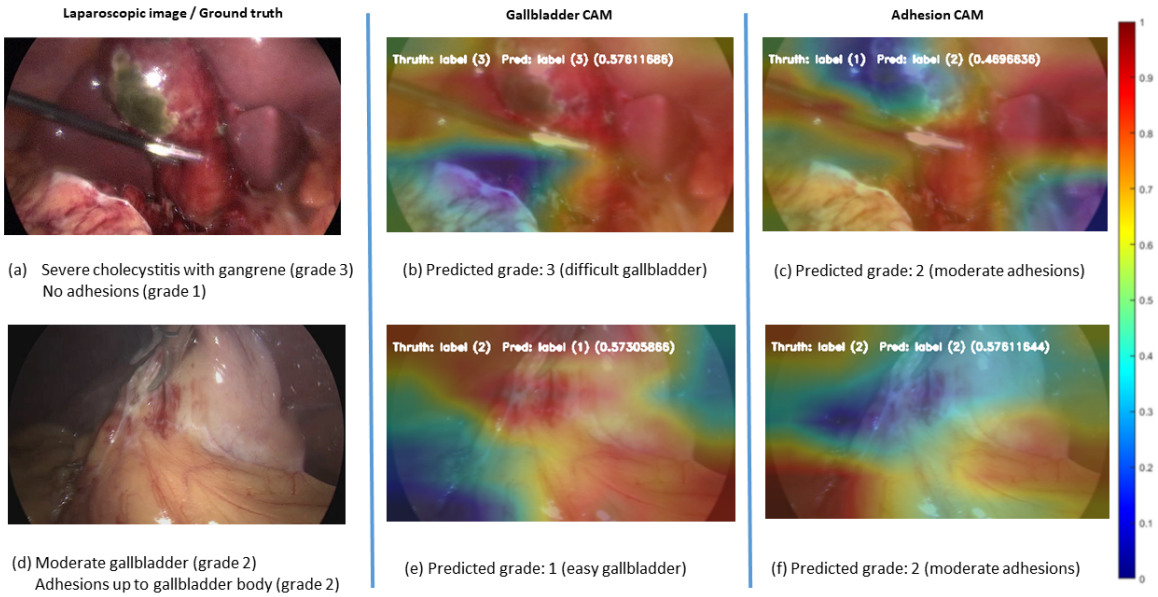
The model responses for degraded inputs are given in Table 6. Model responses on the medium disturbance are noticeable, but little and within the 5% range of the non-disturbed noise accuracy. When heavy noise or the resolution is brought down to an 8th the accuracy drops to the level below 50%, which can be considered useless or as worse as a coin flip. Heavy blurring allows the model to be able to classify (with an accuracy of 70%).

3.0.3. Class activation Mapping

The class activation maps in Figure 7 are presented for the gallbladder trained models and focus (red) on the gallbladder and low values (blue) on the lower areas around the gallbladder. In the second image (d), the gallbladder is limited visualised, also the

classification of the gallbladder difficulty label is incorrect (image e). For the adhesion results, the highest values of the class activation maps of different images are observed at all different locations (images c and f). Our inspection could not find any logic in those outputs. The CAM of the Adhesions trained model appears/seems not the focus on one thing (anatomical structure or feature) in particular.

Table 7.: An set of examples of with class activation mapping on the best Resnet-50 of the multiclass classification.



4. Discussion

Video recordings of LC are currently of high interest for computer vision research. Important research areas on cholecystectomy datasets have been autonomous detection of surgical phases, anatomy recognition, surgery time prediction and instrument and action recognition. (Anteby et al. 2021) However, when studying the metrics mentioned above, one should ideally take into account the intra-operative conditions since it affects the surgical course. This work shows the potential of difficulty understanding in LC based on laparoscopic images. Current results are still not optimal; however, these are the first steps to enable more reliable comparisons between surgeons' performance, procedure time and/or complications.

The binary model achieved the highest accuracy between grade 1/2 (non-cholecystitis) and grade 3 (cholecystitis). However, it should be noted that it is unclear if the actual feature that is trained (infected gallbladder tissue (cholecystitis)) or some other closely related feature such as the redness of the gallbladder tissue. This is visible in the first image (b) of the gallbladder CAM, where the model focuses on the gallbladder and a major part of the liver. The higher performance/better trainability

of the binary model I, II vs III could have been expected if we look at the multi-label gallbladder models' higher Precision, Recall and F1 score on the 3rd label.

In our opinion, the current database is of use for further experimentation with "difficulty" related features. However, the usability and trainability of the adhesion labels or adhesions, in general, remains questionable. This study shows that adhesions are difficult anatomical structures to identify. Reasons could be the diversity of adhesions and the great amount of intestinal organs covered by the same fatty tissue or omentum. It could be suggested to first investigate the labels of multiple observers in a dataset analysis. Interrater reliability could provide us with interesting information on data and annotation quality. In this study, we used visually present labels that were based on strict definitions.

Additionally, other visual aspects that are highly related to "difficulty" should be examined. A suggested option is to research other existing difficulty manual grading systems for features that do relate some difficulty, longer procedure time or negative post-surgical outcomes. Those other features might support a final set of difficulty labels and grading system.

Furthermore, there are publically available cholecystectomy video datasets which we could have used for labelling. But those did contain fewer difficult cases when using the Nassar grading system.

There are limitations to be mentioned. As mentioned, the dataset constructed was not analysed on its labels and differences between multiple observers. Secondly, overfitting is more likely in imbalanced small datasets. More specifically, this could be the case without being aware of the overfitting due to general accuracy and loss we logged during multi-class training. This could be solved/visualised in the multilabel setting by calculating the model error per label. Although the publically available and commonly used Cholec80 (Twinanda et al. 2016) dataset only consists of 80 videos and our dataset is 96 videos, both can be considered small datasets.

This is also visible in the differences between the manually distributed video models and the k-fold cross-validation accuracies. However, the order of the models in their performance is the same. Generally speaking, the gallbladder-trained models perform better than the adhesion-based models, but the accuracies are noticeably lower on the accuracy (apart from 1 model; Gallbladder I vs II, III). The difference in accuracy is possible due to an imbalanced label distribution that is induced due to manually splitting the different sets versus the k-fold splits that were selected randomly. We should mention that the standard deviation of the models performing at 70%+ accuracy can be considered high. This points out that the feature could be trainable, but the dataset is too small and/or the labels are inequally distributed to generate a stable performance in each fold. We endorse that the dataset for application purposes is too small, though it shows the potential for surgical difficulty prediction.

The best model was tested on performance when image quality was degraded. Poor image quality is something quite common in surgery, although in real surgery, the surgeon will always try to preserve the best possible visibility. As mentioned above, the models were trained on a small dataset, so there is a noticeable drop in accuracy when the image quality was degraded (either blur, noise or resolution drop). It could be argued that the response on poor image quality episodes is less in temporal/clip-based models.

Ideally, surgical difficulty prediction might be more suitable as a video analysis task than individual frames. It might be a suggestion to define the task of surgical difficulty prediction in terms of video classification instead of classifying individual frames. However, in our approach, this new problem is tested with simple frame-by-

frame models first to see what could already be achieved. More specifically, it shows that some features related to the Nassar scale's difficulty are trainable. It shows that a simple and widely used neural network can detect cholecystitis in single frames from the first phase with 88% accuracy.

The training of the different models was only performed during the sweeps of the hyperparameter optimisation. Manual fine-tuning might optimise the model performance further. Also, the only optimisation technique used regarding the class imbalance was applying class weighting.

Lastly, other approaches could potentially reach higher performance such as regression-based solution.

5. Conclusion

The work in this paper shows surgical scene understanding for difficulty evaluation of the patient with minimal surgeon intervention. The models can classify with reasonable accuracy *'difficulty'* on the gallbladder features, with the best result when the third label is set out against the label 1, 2. These are the cholecystitis vs no cholecystitis (88% accuracy). Although, the model is already able to also reach an accuracy of 82% on the other label combination (1, vs 2,3). Based on the "Gallbladder" we are able to distinguish cholecystitis the more complicated cases. We encourage to use an larger dataset in similar setups. We cannot conclude that the model can classify the adhesions grade nor the adhesion features in this setup.

Acknowledgement(s)

We thank Ruby M. Egging for her support and valuable discussions. We thank Fausto Milletarì for assistance and help regarding model development.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Ali M, Ochoa-Ruiz G, Ali S. 2022. A semi-supervised teacher-student framework for surgical tool detection and localization. Available from: https://arxiv.org/abs/2208.09926.

Anteby R, Horesh N, Soffer S, Zager Y, Barash Y, Amiel I, Rosin D, Gutman M, Klang E. 2021. Deep learning visual analysis in laparoscopic surgery: a systematic review and diagnostic test accuracy meta-analysis. Surgical Endoscopy. 35(4):1521–1533.

Atta HM, Mohamed AA, Sewefy AM, Abdel-Fatah AFS, Mohammed MM, Atiya AM. 2017. Difficult laparoscopic cholecystectomy and trainees: predictors and results in an academic teaching hospital. Gastroenterology research and practice. 2017.

Biewald L. 2020. Experiment tracking with weights and biases. Software available from wandb.com; Available from: https://www.wandb.com/.

Gagana B, Shukla N, Pathak C, Garda K, Holdroyd T, Broz DJ. 2019. Class activation maps; Oct. Available from: https://medium.com/@GaganaB/class-activation-maps-551477720679.

Griffiths EA, Hodson J, Vohra RS, Marriott P, Katbeh T, Zino S, Nassar AH. 2019. Utilisation of an operative difficulty grading scale for laparoscopic cholecystectomy. Surgical endoscopy. 33(1):110–121.

He K, Zhang X, Ren S, Sun J. 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 770–778.

Lam K, Chen J, Wang Z, Iqbal FM, Darzi A, Lo B, Purkayastha S, Kinross JM. 2022. Machine learning for technical skill assessment in surgery: a systematic review. NPJ digital medicine. 5(1):1–16.

Madni TD, Leshikar DE, Minshall CT, Nakonezny PA, Cornelius CC, Imran JB, Clark AT, Williams BH, Eastman AL, Minei JP, et al. 2018. The parkland grading scale for cholecystitis. The American Journal of Surgery. 215(4):625–630.

Mascagni P, Vardazaryan A, Alapatt D, Urade T, Emre T, Fiorillo C, Pessaux P, Mutter D, Marescaux J, Costamagna G, et al. 2022. Artificial intelligence for surgical safety: automatic assessment of the critical view of safety in laparoscopic cholecystectomy using deep learning. Annals of surgery. 275(5):955–961.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 12:2825–2830.

Priego P, Ramiro C, Molina J, Rodríguez Velasco G, Lobo E, Galindo J, Fresneda V. 2009. Results of laparoscopic cholecystectomy in a third-level university hospital after 17 years of experience. Revista Espanola de Enfermedades Digestivas. 101(1):20.

Pytorch. 2022. Torchvision.models; July. Available from: https://pytorch.org/vision/stable/models.html.

sklearn. 2022. sklearn model_selection kfold; Feb. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html.

Soper NJ, Malladi P. 2022. Laparoscopic cholecystectomy; July. Available from: https://www.uptodate.com/contents/laparoscopic-cholecystectomy.

Sugrue M, Coccolini F, Bucholc M, Johnston A. 2019. Intra-operative gallbladder scoring predicts conversion of laparoscopic to open cholecystectomy: a wses prospective collaborative study. World Journal of Emergency Surgery. 14(1):1–8.

Tan M, Le Q. 2021. Efficientnetv2: Smaller models and faster training. In: International Conference on Machine Learning. PMLR. p. 10096–10106.

Tang B, Cuschieri A. 2006. Conversions during laparoscopic cholecystectomy: risk factors and effects on patient outcome. Journal of gastrointestinal surgery. 10(7):1081–1091.

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

Ward TM, Hashimoto DA, Ban Y, Rosman G, Meireles OR. 2022. Artificial intelligence prediction of cholecystectomy operative course from automated identification of gallbladder inflammation. Surgical Endoscopy:1–9.

Wightman R. 2022. Rwightman/pytorch-image-models: Pytorch image models, scripts, pretrained weights – resnet, resnext, efficientnet, efficientnetv2, nfnet, vision transformer, mixnet, mobilenet-v3/v2, regnet, dpn, cspnet, and more; Feb. Available from: https://github.com/rwightman/pytorch-image-models.

Yokoe M, Hata J, Takada T, Strasberg SM, Asbun HJ, Wakabayashi G, Kozaka K, Endo I, Deziel DJ, Miura F, et al. 2018. Tokyo guidelines 2018: diagnostic criteria and severity grading of acute cholecystitis (with videos). Journal of Hepato-Biliary-Pancreatic Sciences. 25(1):41–54.

6. Appendices

Appendix A. Train and model parameters

For hyperparameter optimisation of the different models, sweeps were conducted with weights and biases with the following configurations:

- Loss:
  - Multiclass: Crossentropy (final layer: softmax)
  - Binary: Binary Crossentropy with logits (final layer: sigmoid)
- optimizer: adam
- Classweighting
- Models:
  - Pre-trained Resnet-18 and Resnet-50
  - Pre-trained *EfficientnetV2*
- Learning rate ranging between 0.0001 and 1e-06
- dropout ranging between 0.8 and 0.2
- Batch size ranging between 16 and 64 (only when possible to load)
- image model input size 224 (original resized to 256 and centre-cropped to 224)
- Bayes optimisation with Maximize validation accuracy
- All models trained on a NVIDIA GeForce GTX 1080 Ti and NVidia RTX A4000 (for the K-fold cross-validation and degraded input-model response (Ubuntu 20.04 LTS))

config param \ model	Gallbladder 3 classes	Adhesions 3 classes	Gallbladder 1,2 vs 3	Gallbladder 1 vs 2,3	Adhesions 1 vs 2,3	Adhesions 1,2 vs 3
Batchsize	32	62	32	25	18	58
Loss	CrossEntropy Loss	CrossEntropy Loss	BCE With-LogitsLoss	BCE With-LogitsLoss	BCE With-LogitsLoss	BCE With-LogitsLoss
Model	EfficientNetV2	Resnet50	Resnet50	Resnet50	EfficientNetV2	Resnet50
Dropout	0.4	0.71147110161	0.7958115777	0.2183149701	0.2293131140	0.42504033628
Learning rate	0.00003346	0.00000868960	0.0000750107	0.0000382985	0.0000244118	0.00001182659