University Library

MINERVA ACCESS

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Badgery, H;Zhou, Y;Bailey, J;Brotchie, P;Chong, L;Croagh, D;Page, M;Davey, CE;Read, M

Title:
Using neural networks to autonomously assess adequacy in intraoperative cholangiograms

Date:
2024

Citation:
Badgery, H., Zhou, Y., Bailey, J., Brotchie, P., Chong, L., Croagh, D., Page, M., Davey, C. E. & Read, M. (2024). Using neural networks to autonomously assess adequacy in intraoperative cholangiograms. Surgical Endoscopy, 38 (5), pp.2734-2745. https://doi.org/10.1007/s00464-024-10768-0.

Persistent Link:
https://hdl.handle.net/11343/353601

License:
CC BY

Surgical Endoscopy (2024) 38:2734–2745
https://doi.org/10.1007/s00464-024-10768-0

Using neural networks to autonomously assess adequacy in intraoperative cholangiograms

Henry Badgery[2,3] · Yuning Zhou[1,7] · James Bailey[4] · Peter Brotchie[5] · Lynn Chong[2,3] · Daniel Croagh[2,6] · Mark Page[5] · Catherine E. Davey[1,7] · Matthew Read[2,3]

Received: 1 October 2023 / Accepted: 22 February 2024 / Published online: 1 April 2024
© The Author(s) 2024

Abstract

Background Intraoperative cholangiography (IOC) is a contrast-enhanced X-ray acquired during laparoscopic cholecystectomy. IOC images the biliary tree whereby filling defects, anatomical anomalies and duct injuries can be identified. In Australia, IOC are performed in over 81% of cholecystectomies compared with 20 to 30% internationally (Welfare AIoHa in Australian Atlas of Healthcare Variation, 2017). In this study, we aim to train artificial intelligence (AI) algorithms to interpret anatomy and recognise abnormalities in IOC images. This has potential utility in (a) intraoperative safety mechanisms to limit the risk of missed ductal injury or stone, (b) surgical training and coaching, and (c) auditing of cholangiogram quality.

Methodology Semantic segmentation masks were applied to a dataset of 1000 cholangiograms with 10 classes. Classes corresponded to anatomy, filling defects and the cholangiogram catheter instrument. Segmentation masks were applied by a surgical trainee and reviewed by a radiologist. Two convolutional neural networks (CNNs), DeeplabV3+ and U-Net, were trained and validated using 900 (90%) labelled frames. Testing was conducted on 100 (10%) hold-out frames. CNN generated segmentation class masks were compared with ground truth segmentation masks to evaluate performance according to a pixel-wise comparison.

Results The trained CNNs recognised all classes.. U-Net and DeeplabV3+ achieved a mean F1 of 0.64 and 0.70 respectively in class segmentation, excluding the background class. The presence of individual classes was correctly recognised in over 80% of cases. Given the limited local dataset, these results provide proof of concept in the development of an accurate and clinically useful tool to aid in the interpretation and quality control of intraoperative cholangiograms.

Conclusion Our results demonstrate that a CNN can be trained to identify anatomical structures in IOC images. Future performance can be improved with the use of larger, more diverse training datasets. Implementation of this technology may provide cholangiogram quality control and improve intraoperative detection of ductal injuries or ductal injuries.

✉ Henry Badgery
henry.badgery@svha.org.au

1 Department of Biomedical Engineering, The University of Melbourne, Parkville, Australia

2 Department of Upper Gastrointestinal Surgery, St Vincent's Hospital Melbourne, Melbourne, Australia

3 Department of Surgery, The University of Melbourne, St Vincent's Hospital, 41 Victoria Parade, Fitzroy, Melbourne, VIC 3065, Australia

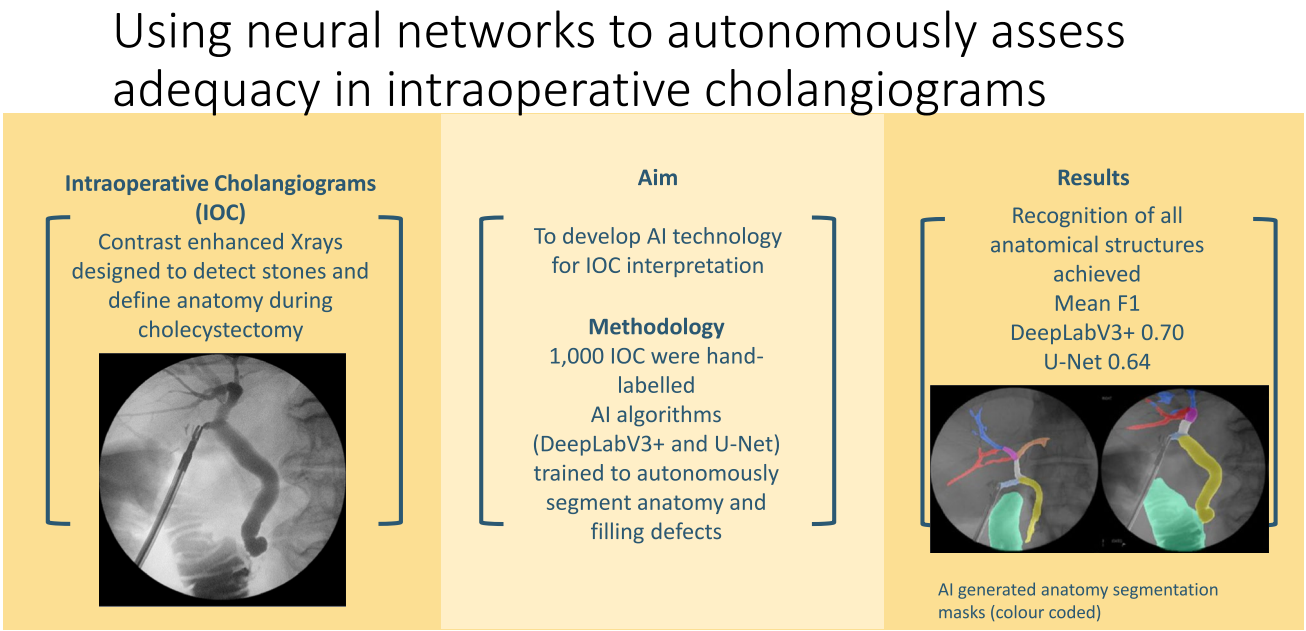
4 School of Computing and Information Systems, The University of Melbourne, Parkville, Australia

5 Department of Radiology, St Vincent's Hospital Melbourne, Melbourne, Australia

6 Department of Surgery, Monash Health, Melbourne, Australia

7 Graeme Clark Institute for Biomedical Engineering, The University of Melbourne, Melbourne, VIC, Australia

Graphical abstract



Keywords Convolutional neural network · Cholangiogram · Laparoscopic cholecystectomy · Artificial intelligence

Intraoperative cholangiography (IOC) is a contrast-enhanced X-ray study taken during laparoscopic cholecystectomy to display the biliary tree. IOC provides a dynamic method to image the biliary tree through radiographic visualisation of contrast flow through the biliary tree. IOC is used to detect the presence of stones in the common bile duct, define anatomy, and look for the presence of bile leak from a biliary tree injury. The use in Australia is comparatively high, being performed in over 80% of cholecystectomies, compared with rates internationally ranging from 20 to 30% [1–4]. Although IOC have not been demonstrated to reduce the rate of bile duct injury (BDI), they has clinical utility in being able to identify both filling defects and injuries to the biliary tree [5, 6].

An IOC is performed by injecting radio opaque contrast into the cystic duct while taking X-ray images using a portable X-ray machine (Fig. 1). Once the cystic duct is clearly identified and adequately dissected, a lateral incision is made, and the duct is cannulated. The cannula is used to infuse contrast into the duct under pressure to visualise the biliary tree. Interpretation of the cholangiogram involves recognition of five key features: (1) contrast flow into the duodenum; (2) distal filling of the common bile duct; (3) proximal filling of the three main hepatic ducts, i.e. the left hepatic duct (LHD), the right anterior hepatic duct (RAHD) and right posterior hepatic duct (RPHD); (4) the absence of filling defects in any ducts and (5) spiral valves visible within the cystic duct [7]. Recognition of filling defects intraoperatively provides an opportunity for early intervention or intervention at index operation. This can be achieved using a specialised camera and stone retrieval equipment via the cystic duct, or directly via choledochotomy (an incision into the bile duct). If surgical removal at index operation is not possible, early referral for endoscopic retrograde cholangiopancreatography (ERCP) is another option. Stents can also be placed at the time of operation to maintain duct patency. The immediate recognition of bile leak or BDI using IOC obviates the risk of delayed diagnosis, allowing for intervention at index operation or prompt transfer to a specialist centre for early intervention. Early recognition and management of bile leaks and BDI is key to improving outcomes [8]. Failure to achieve the five components of a cholangiogram should arouse suspicion for ductal injury or a retained stone or stricture. Furthermore, any technical issues should be rectified to ensure adequacy of the cholangiogram.

Artificial intelligence (AI) is the use of computers or machines to to perform tasks that typically require human intelligence. Investment of AI in medical and surgical applications has surged in recent years, in part owing to access to improved computing power and data collection [9]. AI-based visual tasks and applications have been developed across a wide array of medical fields including radiology, surgery,

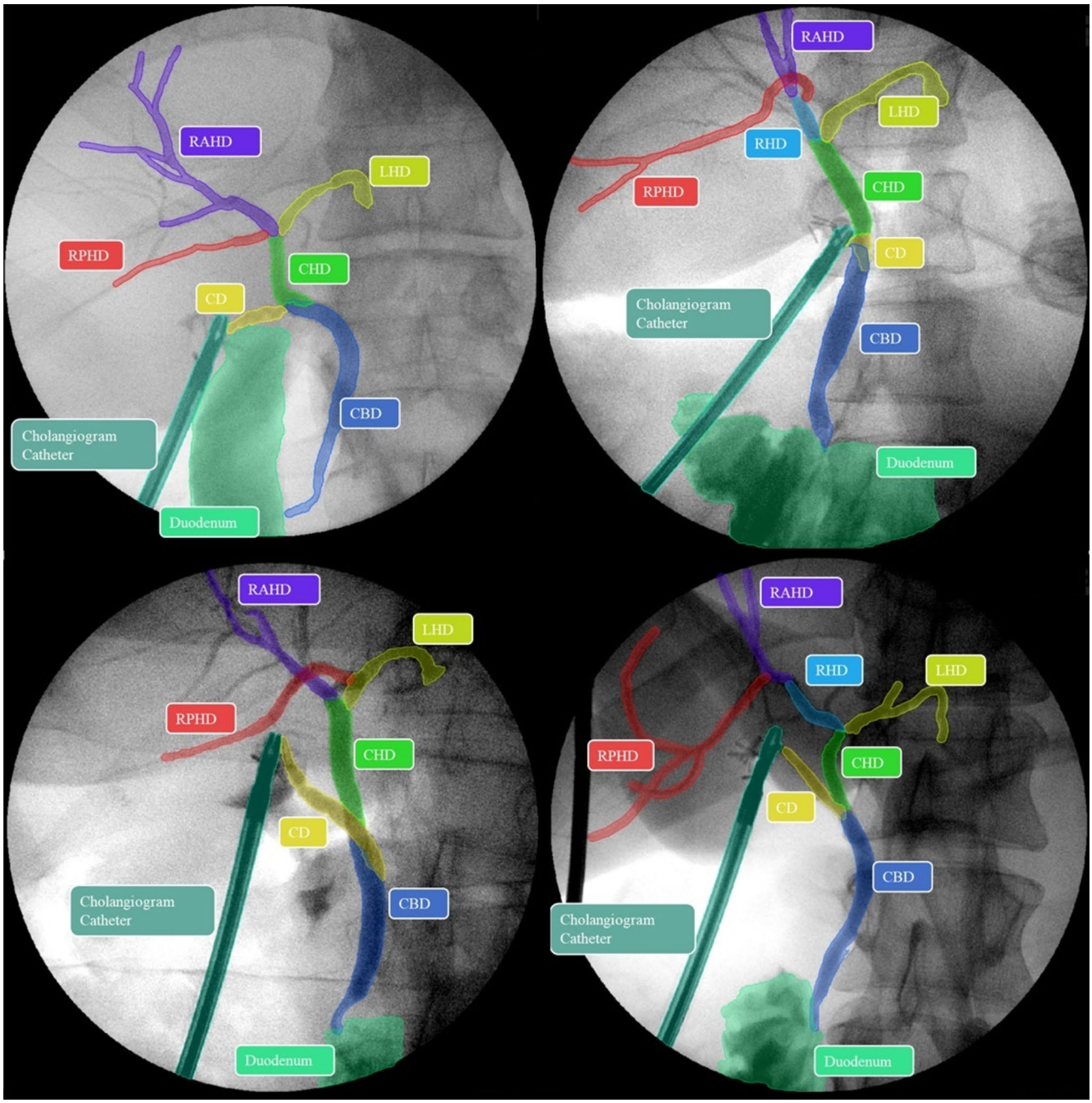


Fig. 1 Examples of human applied segmentation masks applied to intraoperative cholangiograms (IOC). Masks correspond to labels listed in Table 1: *CBD* common bile duct, *CHD* common hepatic duct, *CD* cystic duct, *LHD* left hepatic duct, *RAHD* right anterior hepatic duct, *RHD* right hepatic duct, *RPHD* right posterior hepatic duct

endoscopy and pathology. Convolutional neural networks (CNNs) are a specific type of machine learning algorithm modelled on the structure and function of the biological neural system [10]. CNNs are effective in computer vision tasks such as classification, detection, or segmentation of structures in medical images. This technology has broad potential; in the context of IOCs, it can be used to generate an AI-powered checklist to ensure that the key cholangiogram features have been adequately demonstrated and no abnormalities are overlooked, serving as a safety checkpoint to avoid a missed injury or retained stone. A segmented visual image can also be used as an improved form of documentation detailing a satisfactorily completed cholangiogram. Furthermore, this technology could serve as a training adjunct to surgical trainees and help from the basis for augmented reality environments were AI-defined anatomical maps can

help guide surgeons intraoperatively or provide opportunities for preoperative modelling and simulation.

In this project, we aim to train CNNs to accurately recognise and segment key anatomical structures in IOC. This is the first published work outlining the use of AI and computer vision in IOC.

Methods

Ethics approval for this study was obtained from the St Vincent's Hospital, Melbourne Human Research Ethics Committee (St Vincent's HREC reference HREC/67934/SVHM-2020-235987, protocol amendment V2 January 2022) with governance approval obtained for peripheral contributing sites.

Data collection

IOC was retrospectively obtained from three tertiary hospitals through the imaging archiving system with cases matched through the hospital coding systems. After retrieval of cholangiograms, images were manually selected based on the following criteria: (a) minimum of one and maximum of two frames per patient; (b) best quality frame(s) selected per patient; and (c) if the entire biliary tree is not visualised in a single frame, two included frames in combination should demonstrate the entire biliary tree. A maximum of two images per patient was implemented to limit imbalance within the dataset. The best quality frame(s) were selected for each patient based on visual assessment. Frames were assessed based on clear representation of all structures with minimal movement artefact. Frames with a presence of other obstructing structures such as instruments, cables, bony structures or leaked contrast were excluded. Metadata were stripped from all files and images were converted and stored securely as png files.

Dataset preparation

The dataset in its entirety was reviewed by a surgical trainee to ensure consistency. A purpose-written script was used to automatically deidentify images including removal of patient information visible on the Xray. A labelling protocol was written to encapsulate the key anatomical structures, filling defects and the cholangiogram catheter (Table 1). A total of 10 classes were included plus a background class for unlabelled pixels. Segmentation masks were applied by a general surgery trainee with previous labelling experience using Darwin V7 platform (V7 Labs, 2020) [11]. 1,000 frames taken from 586 patients were ultimately labelled and included in the dataset (Table 2). Codes were assigned to each cholangiogram, including patient-specific codes so that cholangiograms obtained from the same patient for the same procedure could be identified. Patients in the testing dataset were kept distinct from the training dataset to prevent data bleeding between the different sets.

The testing dataset of 100 frames (10%) was reviewed by an experienced abdominal radiologist to ensure that all anatomical labels were accurate. Minor adjustments were made where necessary. The test set masks were compared pre- and post- review adjustments, yielding a mean F1 score of 0.99 across the testing dataset. The remaining 900 frames (90%)

Table 2 Dataset summary

Dataset	Frames	Patients
Training	720	420
Validation	180	107
Testing	100	59
Total	1000	586

Table 1 Intraoperative cholangiogram (IOC) segmentation labels with corresponding colours as applied by Darwin V7

Cholangiogram catheter (forceps and catheter)	
Common Bile Duct (CBD)	
Common Hepatic Duct (CHD)	
Cystic Duct (CD)	
Duodenum (with contrast)	
Filling defect	
Left Hepatic Duct (LHD)	
Right Anterior Hepatic Duct (RAHD)	
Right Hepatic Duct (RHD)	
Right Posterior Hepatic Duct (RPHD)	

were then split into a training dataset of 720 frames (80%) and a validation dataset of 180 frames (20%).

Dataset augmentation and network selection training

Our dataset underwent augmentation to increase data volume. Augmentation techniques included horizontal flipping and random colour jittering within a specified range (brightness 0.25, contrast 0.25, saturation 0.25 and hue 0.0). In addition, random rotation was applied between -30 degrees and 30 degrees. Rotation was confined to this range as this reflects the real-world variability in cholangiogram X-ray orientation. For computational efficiency, frame resolution was downsized to 300 × 300 pixels. To mitigate the problem of dataset class and pixel imbalance, specific weightings were applied during training to each class. These weightings were inversely proportional to class prevalence. Two CNNs, namely DeeplabV3 + [12] with a ResNet101 [13] backbone (Supplementary Fig. 1) and U-Net [14] (Supplementary Fig. 2), were selected and trained separately using the same labelled and augmented dataset. Comparative analysis of the two networks was performed following training. DeeplabV3 + is a powerful CNN that performs well in computer vision tasks [12]. It was chosen after preliminary success in early experiments using a pilot dataset of 70 labelled cholangiogram frames. The second network trained was U-Net, a CNN that was specifically designed for biomedical image segmentation [14]. Deeplabv3 + was run with a ResNet101 backbone while U-Net was run without a backbone. DeeplabV3 + with ResNet101 backbone had almost 9 times the parameters of the U-Net architecture and conducted five times more multiply-accumulate computations for each feed-forward training iteration [13]. Training experiments were conducted on four NVIDIA A100 graphics processing units (NVIDIA, Santa Clara, California, USA) with PyTorch implementation on the Spartan high-powered computer housed at the University of Melbourne. Hyperparameters were optimised using the validation dataset. Both models were trained using AdamW as the optimizer for 100 epochs, with 10 warm up epochs [15, 16]. The batch size was set to 64, the initial learning rate 0.005, and weight decay 0.01. Given the relatively small testing dataset size, K-fold cross-validation was conducted (K = 5).

Evaluation metrics

After training, network accuracy was evaluated by comparing the network prediction with the ground truth of human annotations, as demonstrated using common evaluation metrics. These metrics included intersection over union (IoU), F1 coefficient, recall and precision as well as true positive (TP), false positive (FP), true negative (TN) and false negative [17] (FN) (Fig. 2). TP, FP, TN and FN refer to each pixel prediction and its concordance with the ground truth pixel-wise label. In addition to the described evaluation metrics, correct recognition of the presence of an object was also determined where the network generated segmentation mask overlapped with the ground truth segmentation mask. This was calculated by determining any degree of accurate overlap between ground truth and the prediction segmentation mask, without consideration of the pixel-wise segmentation accuracy.

IoU, otherwise known as the Jaccard similarity coefficient, is one of the most commonly used metrics for computer vision evaluation [18]. It is the area of common overlap between the ground truth and the network prediction, ranging from 0 to 1 (0% to 100%) with 0 being no overlap and 100 being perfect concordance [19]. In object detection applications, IoU > 0.5 is considered a good score and represents adequate localisation, though the required precision varies depending upon network application [20, 21].

The F1 coefficient, otherwise known as the Dice similarity coefficient or the Sorensen-Dice index, is similar to IoU in that it measures overlap between ground truth and prediction. It differs in that it represents the harmonic mean between sensitivity and precision. IoU penalises over and under-segmentation more than the F1 coefficient [19, 21]. Similar to IoU, the F1 range is from 0 (no concordance) to 1 (perfect concordance) with a value greater than 0.5 considered good, depending upon the application [21].

Recall, otherwise known as the sensitivity or true positive rate, demonstrates the rate of correctly attributed pixels by calculating the ratio between the network prediction attribution of positive pixels and all pixels attributed to that class.

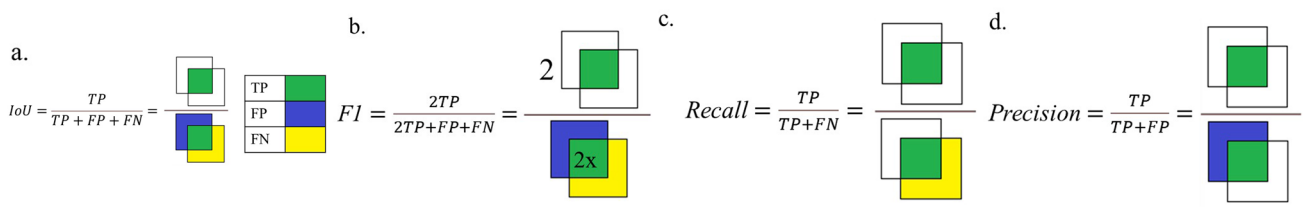


Fig. 2 Equation and diagram for evaluation metrics. **a** Intersection-over-Union (IoU); **b** F1/Dice Coefficient; **c** Recall; **d** Precision

Recall is particularly useful in medical diagnosis applications where false-negative rate is penalised and correct attribution of pixels is rewarded [17].

Class precision is calculated as the ratio between correct class pixel predictions and all pixels assigned to the relevant class. False-positive predictions are penalised in precision metrics [17].

Results

Both the DeeplabV3 + and U-Net networks performed well, achieving a mean F1 coefficient of 0.70 and 0.64, respectively, excluding the background class (Table 3). There was a degree of imbalance in classes in terms of incidence of structure representation as well as proportion of pixels

Table 3 Results of trained Deeplab V3+ convolutional neural network (CNN) (above) and U-Net (below) networks segmenting image frames using an unseen, hold-out test dataset

DeepLabV3+

Class	IoU	TP	FP	FN	Precision	Recall	F1
Mean	56.69	69.94	27.40	23.79	0.73	0.76	0.73
Mean (excl BG)	56.69	69.94	30.06	25.93	0.70	0.74	0.70
BG	96.89	99.21	0.79	2.36	0.99	0.98	0.98
Catheter	82.7	84.16	15.84	2.06	0.84	0.98	0.91
CBD	76.64	81.88	18.12	7.71	0.82	0.92	0.87
CHD	68.62	75.03	24.97	11.06	0.75	0.89	0.81
CD	58.26	69.23	30.77	21.39	0.69	0.79	0.74
Duodenum	80.65	85.31	14.69	6.33	0.85	0.94	0.89
Filling defect	20.59	76.77	23.23	78.04	0.77	0.22	0.34
LHD	62.23	71.37	28.63	17.06	0.71	0.83	0.77
RAHD	39.34	52.63	47.37	39.09	0.53	0.61	0.56
RHD	39.77	50.19	49.81	34.31	0.5	0.66	0.57
RPHD	38.08	52.79	47.21	42.25	0.53	0.58	0.55

Unet

Class	IoU	TP	FP	FN	Precision	Recall	F1
Mean	54.34	63.14	36.86	24.85	0.63	0.75	0.67
Mean (excl BG)	50.21	59.51	40.50	26.95	0.60	0.73	0.64
BG	95.62	99.44	0.56	3.86	0.99	0.96	0.98
Catheter	79.14	80.15	19.85	1.57	0.80	0.98	0.88
CBD	70.54	76.62	23.38	10.11	0.77	0.90	0.83
CHD	65.18	72.00	28.00	12.69	0.72	0.87	0.79
CD	50.44	62.21	37.79	27.28	0.62	0.73	0.67
Duodenum	73.15	76.11	23.89	5.04	0.76	0.95	0.84
Filling defect	13.80	46.95	53.05	83.66	0.47	0.16	0.24
LHD	48.53	54.33	45.67	18.02	0.54	0.82	0.65
RAHD	32.61	40.98	59.02	38.52	0.41	0.61	0.49
RHD	33.85	43.70	56.30	39.96	0.44	0.60	0.51
RPHD	34.90	42.00	58.00	32.63	0.42	0.67	0.52



IoU, TP	0	50	100
Prec, Recall, F1	0	0.5	1
FP, FN	100	50	0

IoU intersection over union, *TP* true positive, *FP* false positive, *FN* false negative, *BG* Background

attributable to each structure in the dataset. Larger and more distal structures (e.g. CBD, duodenum, CHD) had greater representation than the finer and more proximal higher order ductal structures (e.g., hepatic ducts) (Table 3). The cholangiogram catheter, CBD, CHD, CD, duodenum, LHD and RAHD were correctly identified without perfect segmentation in over 80% of the testing dataset. Filling defects were present in 21% of frames in the testing dataset (21/100). While the pixel-wise accuracy of filling defect segmentation was poor in both networks (DeeplabV3 + F1 0.34, UNet F1 0.24), the presence or absence of filling defects was correctly characterised in 89% of cases for the whole testing dataset (89/100) and in 66% of cases (14/25) in cholangiograms where a filling defect was present. Of the filling defect errors, 37% (4/11) represented a network detection where a filling defect was not present, and 63% represented failure to recognise a filling defect where one was present. This is reflected by a high false negative rate (78% in DeeplabV3 +). K-fold cross validation performed on DeeplabV3 + demonstrated a mean IoU of 0.61, (SD = 0.0045), while the UNet mean IoU was 0.52 (SD = 0.017), suggesting superior performance and greater stability of DeeplabV3 + when trained and tested across different dataset subsets.

Class-specific performance was strong on most structures. Both networks achieved an IoU of over 0.5 on all anatomical structures, except for the filling defects and the right hepatic duct group (RHD, RAHD and RPHD) (Fig. 3). F1 generated by DeeplabV3 + for CBD, CHD, CD and LHD were > 0.7 whereas F1 scores for the right ducts were between 0.5 and 0.6. F1 scores were higher than IoU scores in all classes. The networks generally performed better on classes with a higher pixel representation (eg. Duodenum, CBD, catheter, CHD). Higher order structures or structures with a lower pixel representation, such as the hepatic duct branches and filling defects were less accurately segmented. DeepLabV3 + outperformed U-Net on nearly all classes and evaluation metrics with the exception being the recall for the RPHD.

Direct visual comparison between network prediction on the 100frame testing dataset reveal superior performance by DeeplabV3 + . This is demonstrated on the colour coded segmentation masks (Fig. 4) as well as composite images of segmentation masks superimpose upon original images. (Fig. 5). DeepLabV3 + appears to make fewer mistakes globally (Fig. 5). In some instances, peripheral minor hepatic duct branches that were left unlabelled in the ground truth dataset were accurately labelled by the predictive networks (Fig. 5).

Discussion

In this feasibility study, we have demonstrated a novel application of computer vision in laparoscopic cholecystectomy surgery. We have outlined our dataset of cholangiogram comprehensively segmented into ten classes. Using this locally acquired pilot dataset of labelled cholangiograms,

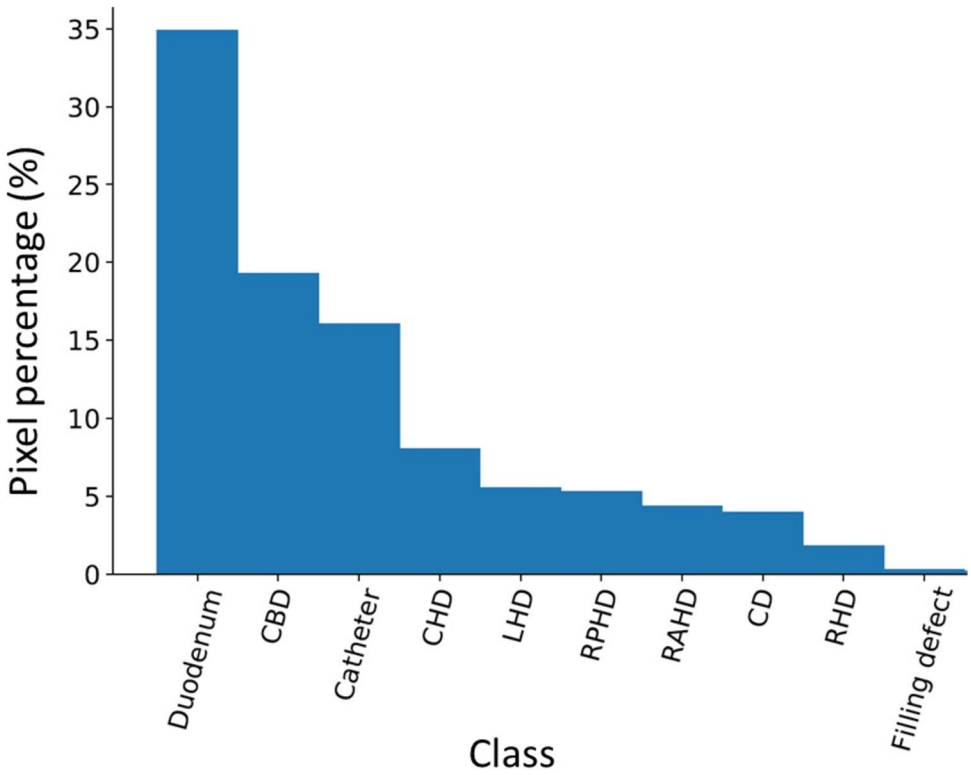


Fig. 3 Class frequency: The pixel class distribution expressed as a percentage of all pixels attributable to each class after exclusion of background

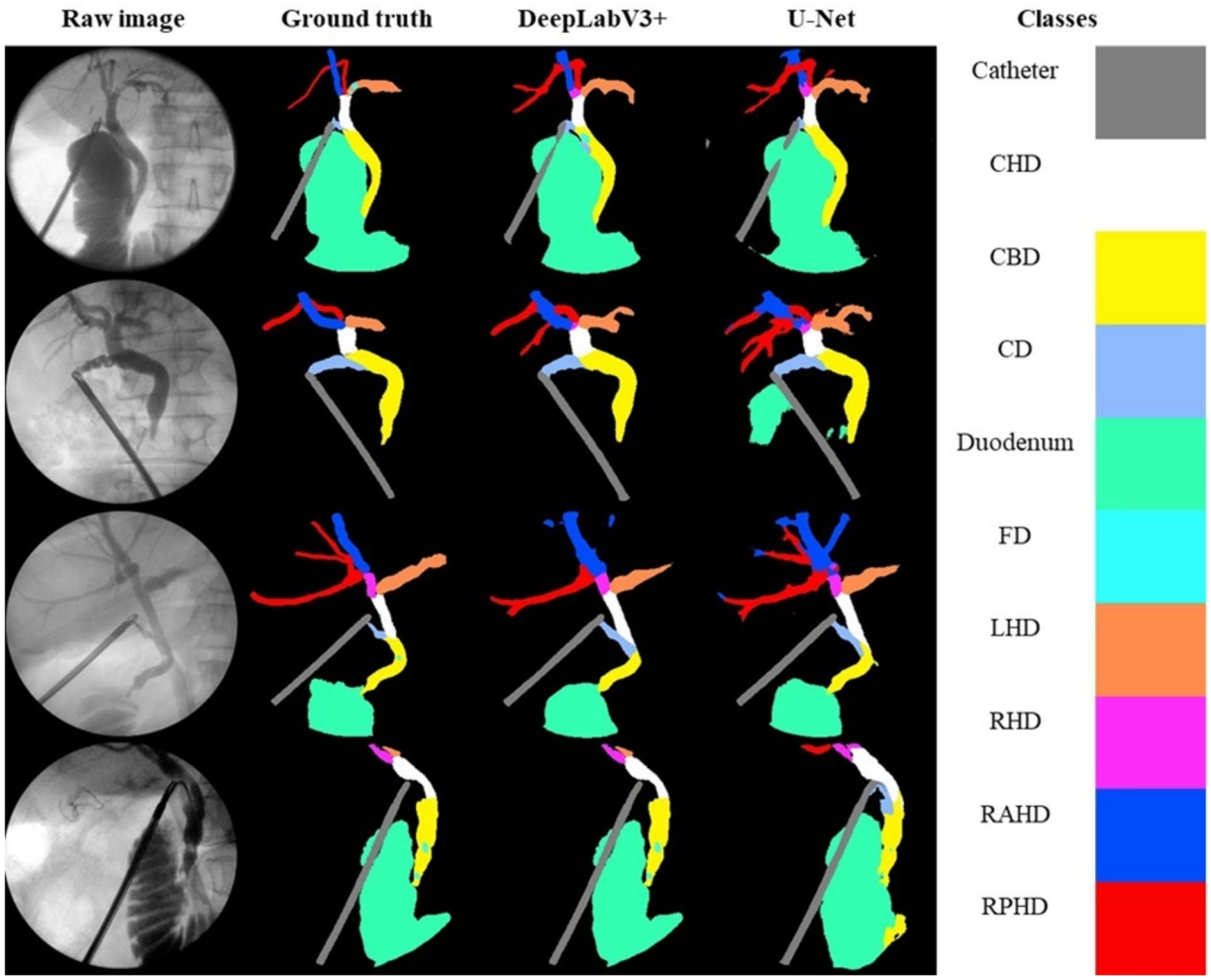


Fig. 4 Examples of convolutional neural network (CNN) generated masks compared with ground truth labels with accompanying colour legend from hold-out test dataset. Figure depicts original cholangiogram image and human labelled ground truth segmentation masks with DeepLabV3 + and U-Net prediction segmentation masks for side-by-side comparison. Colour legend altered from Fig. 1 and Table 1 for visual clarity. *CHD* common hepatic duct, *CBD* common bile duct, *CD* cystic duct, *FD* filling defect, *LHD* left hepatic duct, *RHD* right hepatic duct, *RAHD* right anterior hepatic duct, *RPHD* right posterior hepatic duct

we have achieved a high degree of accuracy in anatomical segmentation using two CNN models with anatomical structures correctly recognised in over 80% of cases and filling defects correctly characterised in 73% of cases. These results provide proof of concept in the development of an accurate and clinically useful tool to aid in the interpretation and quality control of intraoperative cholangiograms.

To our knowledge, this is the only existing dataset of semantic segmentation labelled cholangiograms. The use of CNNs to autonomously segment intraoperative cholangiograms is novel and has many potential applications. Despite the modest dataset size, we have trained two networks capable of identifying key structures. In its current state, the network can be implemented in an autonomous surgical checklist that identifies key cholangiogram features, including contrast flow to the duodenum, the presence of all three hepatic ducts and the visualisation of cystic duct draining to CBD with proximal filling.

One major strength of our trained networks is the capacity to recognise anatomy despite inconsistencies in cholangiogram acquisition and projection. Unlike other imaging modalities, such as frontal X-ray or axial CT that utilise consistent and protocolised projections, cholangiograms are taken with random oblique projections, depending upon patient and operating table position. This adds an additional layer of complexity in autonomous structure recognition using neural networks. Despite the inconsistency in projection, our trained networks were still able to achieve a good result. In addition to projection variability, different X-ray machines were used across multiple health services leading to heterogeneity in cholangiogram appearance. Despite the cholangiogram variability stemming from the use of

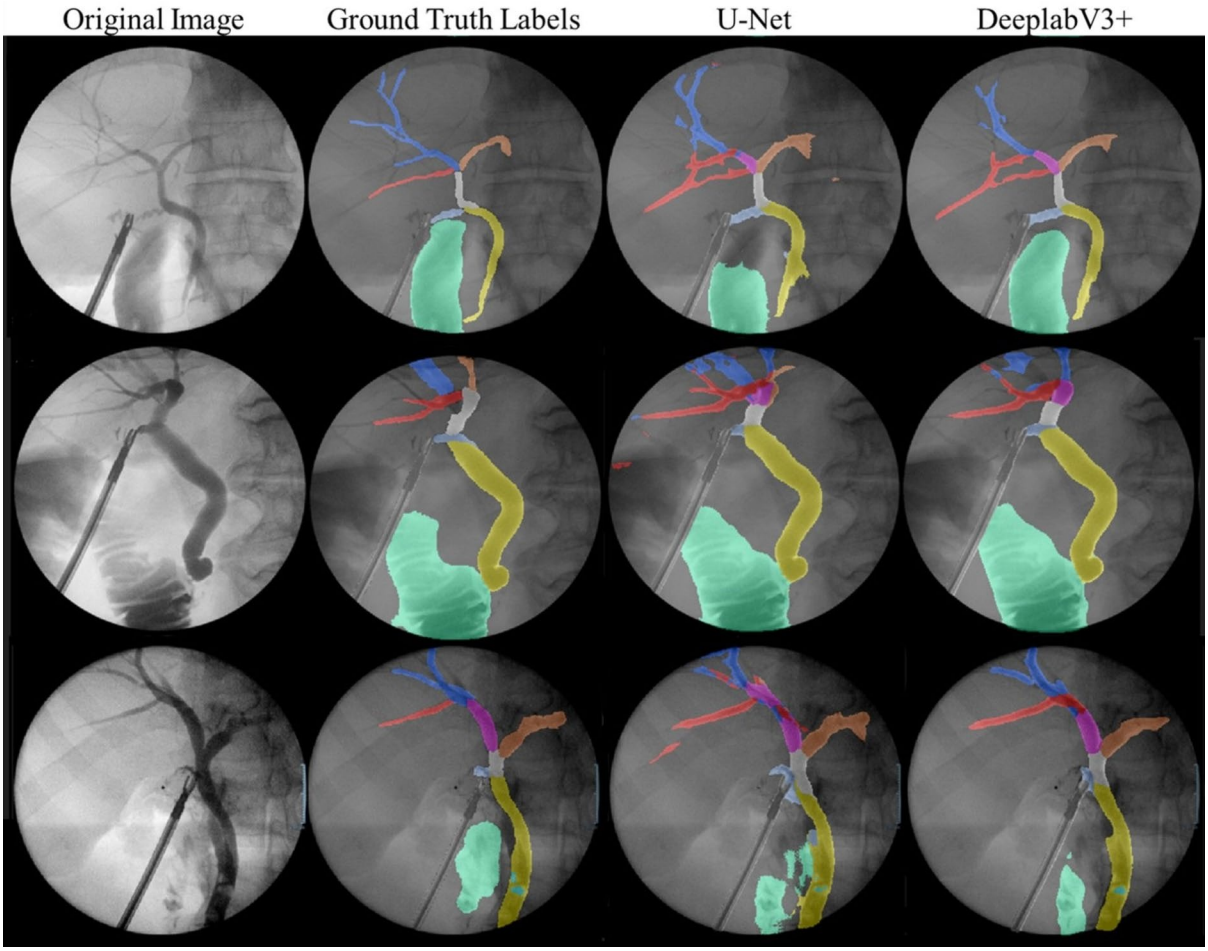


Fig. 5 Comparison of ground truth composite mask overlay with DeeplabV3 + and U-Net composite mask overlay

different machines, our trained networks performed well, highlighting the robustness of the algorithms.

The detection of fillings defects and stones is an important function of cholangiogram. Accurate detection by a network is therefore an important capability to justify clinical implementation. While we achieved correct characterisation of filling defects in most cases, the network failed in several instances. The failure to consistently recognise the presence of filling defects likely stems from several factors. Stones are detected in a small minority of cholangiograms with reported rates ranging from 5 to 20% [22–25]. In our dataset, there were 227 frames containing fillings defects and a total of 363 instances representing just over a quarter of total frames. This rate of filling defects is higher than the general population cholangiogram rate for stones. It is also important to note that cholangiogram filling defects are small and make up a disproportionately low fraction of the segmented area. Furthermore, the appearance of stones and fillings defects on cholangiogram is not consistent. Stones can appear as an absence of downstream contrast flow or a single or multiple rounded filling defects. Another issue with filling defect detection is their similarity in appearance to the background given that they represent a radiolucent absence of contrast. It can be difficult to distinguish filling defects from other cholangiogram features on single static images. The operating surgeon obtains important information from the dynamic images as the cholangiogram is being taken. The subtle signs suggestive of a stone are better appreciated on these dynamic images. These include the pattern of movement of a filling defect distinguishing it from an air bubble and the dynamic response of a filling defect to contrast flow. Improved recognition of stones could be achieved in several ways. The addition of more cholangiogram images positive for stones may improve results. Another method might be to explicitly classify features that are predictive of stone presence. These features might include an absence of downstream flow and dilatation of the cystic or common bile duct. Incorporation of these features in the training pipeline might improve downstream detection of stones.

Within individual cholangiograms there is a degree of ambiguity. Cholangiograms are 2-dimensional representations of 3-dimensional structures. The true anatomy,

particularly in the hepatic ducts, can be misinterpreted due to X-ray projection and superimposition of ductal structure, bony structures, or instruments. Our network utilised exclusive classes, whereby each pixel could only be attributed to a single class. The use of non-exclusive classes, whereby a pixel could be attributed to multiple classes may further improve the accuracy and representation of the biliary tree by allowing for and identifying overlapping structures.

The assessment and evaluation of network performance warrants interrogation. All evaluation metrics employed in this study are a comparison of the network prediction with the ground truth, as determined by surgical trainee labellers. There are limitations in this approach. The ground truth labels contain a degree of subjectivity. The objectivity and truth of the training and testing datasets, therefore, are imperfect. The specific evaluation metric chosen to reflect performance must take into consideration the intended function or application of the network. In our trained networks, smaller higher order ducts tend to be less well segmented than larger calibre distal ducts. The segmentation prediction may not comprehensively detect all higher order hepatic duct branches however does succeed in recognising that the three main hepatic duct branches (LHD, RPHD, RAHD) have been adequately demonstrated. Failure to accurately segment all higher order branches will be penalised by the mathematical evaluation metrics but does not limit the clinical utility. This important point is demonstrated in Fig. 5. In the ground truth labels, higher order branches of the RPSD were not labelled however these were accurately segmented by U-Net and DeeplabV3 + . The mathematical evaluation metrics will categorise the predicted segmentation of these higher order branches as false positives given the discordance with the ground truth. While the accurate recognition of these higher order branches does not alter the clinical utility of the network, the performance is penalised. Similarly, accurate segmentation of all duodenal contrast by a network is not necessary where the function is the binary detection of the presence or absence of contrast flow into the duodenum. In such applications, a lower IoU can be tolerated provided the false positive rate is also low. Conversely, applications demanding a higher degree of accuracy such as measurement of the CBD diameter or cystic duct length demand more accurate segmentation ability of these structures. The intended output or clinical application of a network therefore must be considered when choosing appropriate evaluation metrics and interpretating prediction results.

Visual inspection of the prediction masks demonstrates superior performance by DeeplabV3 + as also reflected in the calculated evaluation metrics. Important structures, on visual assessment, are more accurately and consistently segmented. There is also comparatively less misrecognition. The modestly superior global performance of the DeepLabV3 + may be attributed to both the model size and its use of atrous separable convolutions that improve computational efficiency and reduce complexity [12]. Using a large backbone allows preservation of the understanding of the relative relationship and location of the structures and their adjacent objects. In Addition, atrous convolution allows the model decoder to receive larger contextual information from the previous feature maps while retaining spatial resolution. This can help the model to attain precise localization ability for the structures' location. However, we argue that a larger model is not necessary to achieve better performance. We observe that although the model architecture and size are very different, the performance between the two networks is competitive. In some individual classes DeepLabV3 + performs worse than U-Net. U-Net identifies the subtle structure boundaries more precisely and has fewer FN, which suggest that the model is less likely to misattribute one structure to another. It is also superior at identifying small and underrepresented structures like filling defects. Using a heavy backbone like ResNet101 in DeeplabV3 + may also require more data to converge the larger model. In the biomedical and surgical context, adopting a large model on small datasets may lead to poorer performance, especially in identifying the under-represented classes with overfitting more likely to become a problem [26].

There were several limitations in our methodology. All cholangiograms in the dataset were labelled or finalised by one surgical trainee. The gold standard for semantic segmentation would be to have multiple trained experienced labellers segment each image and then use a concordance map or heatmap to determine the final semantic segmentation masks. The testing dataset was reviewed by a consultant radiologist and then adjusted accordingly by the surgical registrar to ensure ground truth was as accurate as possible. A more robust labelling and validation pathway would be justified in future projects. Another limitation was the local source of the cholangiograms. Our dataset was collected from three hospitals with cholangiograms conducted by a small group of surgeons. The CNN trained from a local dataset may not be as effective or applicable to cholangiograms performed internationally where local protocol and equipment may differ. Our network was trained on still images. However, intraoperative cholangiograms are dynamic investigations where information can be gained through tactile feedback from the pressure in the contrast syringe, observation of the rate of contrast flow through ducts and the movement of fillings defects to help distinguish stones from air bubbles. In some cases, static images are obtained to the satisfaction of the surgeon, but the corresponding stills are not captured for storage which may impact the apparent accuracy of the networks when considering the retrospective training datasets. Training from still images is therefore limiting. Future work should include dynamic or prospective real-time cholangiogram videos that can provide more

information. An additional limitation relates to the chosen class list (Table 1). The labels defined for the hepatic ducts only took into account those demonstrated in the most common anatomical configurations. Accessory hepatic ducts that are seen in rare anatomically aberrant cases were not explicitly labelled. Furthermore, other structures commonly seen such as the pancreatic duct, or the gallbladder in the case of retrograde contrast flow, were not explicitly labelled. Therefore, in cholangiograms where these structures are demonstrated, our network will not be able to segment and label them accurately. Given these structures are uncommon, a substantially larger training dataset would be required to accurately incorporate these labels into a network.

There is great potential for the future direction of this work. With a local pilot dataset, we have achieved strong results while identifying clear strategies where these results can be improved. As discussed, larger more balanced datasets are needed to improve the performance of this CNN. Prospective evaluation of the network will demonstrate and better elucidate the clinical potential. The development of software aimed at autonomously checking that all five cholangiogram features have been satisfied is an important future aim that has real translational potential. Furthermore, international, or multi-site dataset collaborations will improve the performance and generalisability of this network.

Conclusion

In this feasibility study, we have demonstrated the use of CNN based methods to detect and segment key anatomical structures on intraoperative cholangiogram. This work provides a platform for the development of ML based software for use in intraoperative cholangiograms. This software can serve as an autonomous safety checklist as well as a training and education tool. This may have utility in (a) surgical training and coaching, (b) auditing of cholangiogram quality, and (c) intraoperative safety mechanisms minimising the risk of missed ductal injury or stone. While the ability to detect and segment filling defects was average, we have discussed and identified methods of improving this performance that can be implemented in future work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00464-024-10768-0.

Acknowledgements This research was undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This Facility was established with the assistance of LIEF Grant LE170100200. We acknowledge the support of V7 labs for the use of the Darwin data labelling platform. This work is made possible with thanks to the generosity of donors of the Epworth Medical Foundation. Additional thanks to Cassius Fernando for data labelling assistance.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. Epworth Foundation 2022 Capacity Building Grant.

Declarations

Disclosures Dr Peter Brotchie is a clinical consultant with Annalise.ai. Drs. Henry Badgery, Yuning Zhou, James Bailey, Lynn Chong, Daniel Croagh, Mark Page, Catherine Davey and Matthew Read have no conflicts of interest or financial ties to disclose.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Australian Institute of Health and Welfare (2017) Australian atlas of healthcare variation 2017
2. Mirizzi PL (1932) La cholangiografia durante las operaciones de las vias biliares. Bol Soc Cir Buenos Aires 16(1133)
3. Mui J, Mayne DJ, Davis KJ, Cuenca J, Craig SJ (2021) Increasing use of intraoperative cholangiogram in Australia: is it evidence-based? ANZ J Surg 91(7–8):1534–1541
4. Donnellan E, Coulter J, Mathew C, Choynowski M, Flanagan L, Bucholc M et al (2021) A meta-analysis of the use of intraoperative cholangiography; time to revisit our approach to cholecystectomy? Surg Open Sci 3:8–15
5. de'Angelis N, Catena F, Memeo R, Coccolini F, Martínez-Pérez A, Romeo OM et al (2021) 2020 WSES guidelines for the detection and management of bile duct injury during cholecystectomy. World J Emerg Surg. 16(1):30
6. Ford JA, Soop M, Du J, Loveday BPT, Rodgers M (2012) Systematic review of intraoperative cholangiography in cholecystectomy. Br J Surg 99(2):160–167
7. Connor SJ, Perry W, Nathanson L, Hugh TB, Hugh TJ (2014) Using a standardized method for laparoscopic cholecystectomy to create a concept operation-specific checklist. HPB (Oxford) 16(5):422–429
8. Jabłońska B, Lampe P (2009) Iatrogenic bile duct injuries: etiology, diagnosis and management. World J Gastroenterol 15(33):4097–4104
9. Badgery H, Zhou Y, Siderellis A, Read M, Davey C (2022) Machine learning in laparoscopic surgery. Artificial intelligence in medicine: applications, limitations and future directions. Springer, Cham, pp 175–90
10. Soffer S, Ben-Cohen A, Shimon O, Amitai MM, Greenspan H, Klang E (2019) Convolutional neural networks for radiologic images: a radiologist's guide. Radiology 290(3):590–606
11. V7 Labs (n.d.). Darwin V7 2022 https://www.v7labs.com/
12. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. arXiv:1802.02611

13. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. 2016 IEEE conference on computer vision and pattern recognition (CVPR) 27–30 June 2016, pp. 770–788
14. Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. arXiv:1505.04597
15. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv:1412.6980
16. Loshchilov I, Hutter F (2017) Decoupled weight decay regularization. arXiv:1711.05101
17. Hicks SA, Strümke I, Thambawita V, Hammou M, Riegler MA, Halvorsen P et al (2021) On evaluation metrics for medical applications of artificial intelligence. medRxiv. https://doi.org/10.1101/2021.04.07.21254975
18. Maier-Hein L, Reinke A, Godau P, Tizabi MD, Buettner F, Christodoulou E, Glocker B, Isensee F, et al (2022) Metrics reloaded: recommendations for image analysis validation. arXiv:2206.01653
19. Müller D, Soto-Rey I, Kramer F (2022) Towards a guideline for evaluation metrics in medical image segmentation. arXiv:2202.05273
20. Dai J, He K, Sun J (2015) Instance-aware semantic segmentation via multi-task network cascades. arXiv:1512.04412
21. Reinke A, Tizabi MD, Sudre CH, Eisenmann M, Rädsch T, Baumgartner M, et al (2021) Common limitations of image processing metrics: a picture story. arXiv:2104.05642
22. Lai H-Y, Tsai K-Y, Chen H-A (2022) Routine intraoperative cholangiography during laparoscopic cholecystectomy: application of the 2016 WSES guidelines for predicting choledocholithiasis. Surg Endosc 36(1):461–467
23. Collins C, Maguire D, Ireland A, Fitzgerald E, O'Sullivan GC (2004) A prospective study of common bile duct calculi in patients undergoing laparoscopic cholecystectomy: natural history of choledocholithiasis revisited. Ann Surg 239(1):28–33
24. Varadarajulu S, Eloubeidi MA, Wilcox CM, Hawes RH, Cotton PB (2006) Do all patients with abnormal intraoperative cholangiogram merit endoscopic retrograde cholangiopancreatography? Surg Endosc Other Interv Tech 20(5):801–805
25. Ng J, Teng R, Izwan S, Chan E, Kumar M, Damodaran Prabha R et al (2023) Incidence and management of choledocholithiasis on routine intraoperative cholangiogram: a 5-year tertiary centre experience. ANZ J Surg 93(1–2):139–144
26. Yamashita R, Nishio M, Do RKG, Togashi K (2018) Convolutional neural networks: an overview and application in radiology. Insights Imaging 9(4):611–629

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.