Automatic Smoke Analysis in Minimally Invasive Surgery by Image-based Machine Learning

R. Sharifian[a,b,c,1] (MS), H. Mendonça Abrão[d] (MD) S. Madad-Zadeh[a,e] (MD), C. Seve[d] (MD), P. Chauvet[a,d] (MD), N. Bourdel[a,b,d] (MD, HDR, PhD), M. Canis[a,d] (MD, HDR, PhD) and A. Bartoli[a,b,c](HDR, PhD)

a - EnCoV, Institut Pascal, UMR 6602 CNRS/UCA, Clermont-Ferrand, France

b - SURGAR, Surgical Augmented Reality, Clermont-Ferrand, France

c - Department of Clinical Research and Innovation, Clermont-Ferrand University Hospital, France

d - Department of Obstetrics and Gynecology, University Hospital Clermont-Ferrand, 15 63000 Clermont Ferrand, France

e - Surgical Oncology Department, Centre Jean Perrin, 63011 Clermont-Ferrand, France

Abstract

Background

Minimally Invasive Surgery (MIS) uses electrosurgical tools that generate smoke. Smoke reduces the visibility of the surgical site and spreads harmful substances. Automatic image analysis may provide assistance. However, existing studies are restricted to simple clear versus smoky image classification.

Materials and Methods

We propose an approach using surgical image analysis with machine learning. We address three tasks: 1) smoke quantification, which estimates the visual level of smoke, 2) smoke evacuation confidence, which estimates the level of confidence to evacuate smoke, and 3) smoke evacuation recommendation, which estimates the evacuation decision. We collected three datasets with expert annotations. We trained end-to-end neural networks for the three tasks. We also created indirect predictors using task 1) followed by linear regression to solve task 2) and using task 2) followed by binary classification to solve task 3).

Results

We observe a reasonable inter-expert variability for task 1) and a large one for tasks 2) and 3). For task 1), the expert error is 17.61 pp and the neural network error is 18.45 pp. For tasks 2) the best results are obtained from the indirect predictor based on task 1). For this task, the expert error is 27.35 pp and the predictor error is 23.60 pp. For task 3), the expert accuracy is 76.78% and the predictor accuracy is 81.30%.

Conclusions

Smoke quantification, evacuation confidence and evaluation recommendation can be achieved by automatic surgical image analysis with similar or better accuracy as the experts.

Keywords: smoke, mini-invasive surgery, image analysis, neural network

[1] Corresponding author. EnCoV, Clermont-Ferrand's Faculty of Medicine, 1erR3 (first floor, third wing), 28 place Henri Dunant, 63000, Clermont-Ferrand, France.
Email address: rasoul.sharifian.cs@gmail.com (R. Sharifian).

Introduction

Minimally Invasive Surgery (MIS) is a set of modern techniques where the surgical instruments and an endoscopic camera are inserted in the patient body through small incisions. The internal organs are then visible to the surgeons on screens. The abdomino-pelvic cavity is concerned by MIS through specific procedures involving a laparoscope or robot-assistance. Compared to open surgery, MIS has many advantages, reducing patient trauma and improving outcomes.

Regular actions in MIS are to cut tissues, resect pathologies and handle bleeding. This generally involves the use of electrocautery, with heated electrosurgical tools, such as monopolar, bipolar or argon diathermy. Electrocautery generates smoke composed of approximately 95% steam water and 5% organic and inorganic physical particles [1]. This smoke has two major negative consequences. First, it significantly reduces the visibility of the surgical site. Second, it contains potentially dangerous substances for the surgical staff. These issues are mitigated by evacuating the smoke, either continuously using a smoke evacuation system inserted in an MIS port or by simply opening a port's valve. Manual smoke evacuation occurs upon request from the surgeon. We hypothesise that smoke management tasks may increase the surgeon's mental load. This hypothesis is supported by consistent feedback from the surgeons participating in this study, who have all suggested that surgical smoke management can impact their overall performance. Furthermore, previous studies such as [2] highlight the adverse effects of surgical smoke on the surgical staff. A solution might be to perform systematic smoke evacuation as reported in [3], where smoke was evacuated automatically with a delay after electrosurgical device activation. The obvious drawback of systematic evacuation is that it activates even when not strictly required and may disrupt the intervention: it restores visibility but decreases the pressure of the pneumoperitoneum, the artificial insufflation required to create a workspace in the cavity. An alternative approach is to use continuous smoke evacuation by valveless trocars such as the Airseal insufflation system. Although this has been shown to have advantages [4], the method lacks efficiency since it performs evacuation constantly without considering the smoke situation, generating noise [5] which may have implications on sustainability of the operating room. It may also lead to potential postoperative complications [6].

We propose a novel approach to smoke management, which is to automatically evaluate the need for smoke evacuation from the surgical image contents. Indeed, a central criterion to perform smoke evacuation is the lack of visibility caused by smoke. However, visually estimating the amount of smoke from an MIS image is challenging because of the wealth of other visual artefacts, including chromatic noise, spatial light variations, lens dirt and blur caused by camera or organ motion. Artificial intelligence has recently made tremendous progress for image analysis, in particular with deep neural networks in machine learning. Interestingly, recent preliminary results demonstrate the capability of artificial intelligence for clear versus smoky image classification [7,8,9,10]. A natural question towards automatically evaluating the need for smoke evacuation thus regards the possibility of training a neural network to automatically quantify the amount of smoke from an image.

We define our main research question as the study of automatic smoke evacuation recommendation by automatic image analysis. In order to answer this question, we propose a methodology involving three main contributions. First, we propose a dataset of laparoscopy images with annotations collected from experts[2]. Specifically, we collected three types of advanced annotations for each image: 1) the smoke level, which is a percentage representing the visual quantity of smoke present in the image, 2) the smoke evacuation confidence, which is a percentage representing the confidence in performing or not performing smoke evacuation, and 3) the smoke evacuation recommendation, which is a binary indicator representing the decision to evacuate or not evacuate smoke. We collected these annotations for a subset of the public datasets Smoke_Cholec80 [7] and LapGyn4_v1.2 [11], and for a new gynaecology dataset collected from our hospital. For a part of the dataset, each image is annotated multiple times by multiple experts. Second, we propose a statistical analysis of the inter-expert variability and investigate the relationships between the three types of annotations, namely, the smoke level, the smoke evacuation confidence and the smoke evacuation recommendation. Specifically, we study the extent to which the smoke evacuation confidence can be found from the smoke level and the extent to which the smoke evacuation recommendation can be found from the smoke level or from the smoke evacuation confidence. Third, we leverage these relationships to propose image-based automatic prediction methods for three tasks, corresponding to the three types of annotations: 1) smoke quantification, 2) smoke evacuation confidence, and 3) smoke evacuation recommendation. We first train neural networks to achieve these tasks end-to-end. We then devise indirect methods, where the result obtained from the neural network for task 1) is used to solve task 2) using linear regression and where the result obtained from the neural network for task 2) is used to solve task 3) using classification. We thus obtain one method for task 1), two methods for task 2) and four methods for task 3).

Methods

We first describe the proposed data collection and annotation process. We then describe the proposed methods.

Data Collection and Annotation

We explain how we collected and annotated different laparoscopy image datasets for smoke quantification and smoke evacuation.

[2] For the sake of terminological consistency, we use the term expert to refer specifically to surgeons in the context of surgical image annotation.

Data Sources

We have used three data sources: Smoke_Cholec80[3] and LapGyn4_v1.2[4], which are publicly available, and data collected specifically for this work from our hospital. The images were anonymised in the hospital and the patients signed a non-opposition form. Specifically, we have used a subset of Smoke_Cholec80 and LapGyn4_v1.2 as a source of images and annotated them for the sought tasks. We next describe these two datasets, and the image selection and annotation processes for the three data sources. We have taken care to balance the selected images in terms of the annotations, which is difficult as these annotations are not a priori available.

The first data source is Smoke_Cholec80, from which we created the SubCholec dataset, with an emphasis on class balancing and annotation on smoke level and smoke evacuation. Smoke_Cholec80 is the only public dataset related to smoke classification. It was created from Cholec80 [12] which contains 80 videos from cholecystectomy procedures. Smoke_Cholec80 contains a mapping file from specific segments of Cholec80 frames to smoky versus clear annotations. It includes a total of 100K images, half for the smoky class and half for the clear class. Although this dataset is balanced in terms of the smoke classification problem, it contains surgical smoke in various intensities, and mostly shows just a little amount of contamination with smoke. This dataset is therefore not balanced regarding the smoke level quantification and smoke evacuation classification problems. To handle this, we collected annotations both for the smoke levels and smoke evacuation and designed a procedure to extract a balanced dataset from it, which we describe in the Data Annotation section.

The second data source is LapGyn4_v1.2, from which we created the SubLapGyn dataset and annotated it with smoke level and smoke evacuation. LapGyn4_v1.2 is a public dataset collected from over 500 gynaecology surgeries designed for the task of automatic content analysis. This dataset contains four different categories concerning general surgical actions and anatomical structures. The surgical action collection itself involves eight general activities performed during surgery, including coagulation, cutting, injection and suturing. We selected our dataset from the coagulation folder since it is the most relevant category to smoke related tasks, containing images contaminated with different amounts of smoke. Specifically, it contains 3,480 images, from which we randomly selected 100 images to form SubLapGyn, with additional smoke level and smoke evacuation annotations, as described in the Data Annotation section.

The third data source is made of images which we collected from the gynaecology surgery department of our hospital and annotated to form a new dataset called SmokeGyn1. This dataset contains 50 images extracted from seven patients undergoing gynaecological surgeries. In these surgeries, both the monopolar and bipolar heating tools were used for coagulation. In order to find parts of videos that are contaminated with smoke, we used an initial neural network trained to quantify smoke, as described in the Smoke Quantification section, which provided an initial set of smoke contaminated candidate images with diversity. We then reviewed the images

[3] https://ftp.itec.aau.at/datasets/Smoke_cholec80
[4] https://ftp.itec.aau.at/datasets/LapGyn4

to finalise the selection by visually checking randomly selected images to verify that they were sufficiently different from each other. Furthermore, the dataset was constructed by purposefully selecting challenging images, including images taken with dirty lenses, under intense lighting and showing motion blur. Eventually, we forced the dataset to contain at least five frames from each patient, to ensure diversity, and that the smoky and clear classes were balanced.

Data Annotation

We require the smoke level and smoke evacuation annotations, which are not available in any public dataset, and in none of the three above-described datasets. We request the experts to enter two annotations per image, which take the form of two scores, both expressed as a percentage. The first score quantifies the amount of smoke: 0% is for a perfectly clear, smoke-free image, and 100% is for a completely foggy image. The second score quantifies the confidence level in recommending or not-recommending smoke evacuation. Indeed, during the annotation procedure, we realised that the experts were not always fully confident about their evacuation decision. Specifically, as the level and the location of smoke change, their confidence also changes; for some images, some experts could not reach a decision. Therefore, concerning evacuation recommendation, we requested the experts to provide a second score showing their confidence level, as follows: a score of -100% shows complete confidence to not evacuate and a score of 100% shows complete confidence to evacuate[5]. From these scores we obtain the class annotations for evacuation recommendation, namely E+ for positive scores and E- for negative scores. Concretely, we collected the annotations by a Graphical User Interface (GUI) which we created, as shown in figure 1, in which the experts can browse the images from the dataset and enter the two requested annotations by simply moving two slider widgets. This annotation procedure is used for all the three datasets we provide in this study. We repeated this approach with five experts for 233 images representing 26% of the SubCholec dataset to measure the inter-expert variability.

[5] Note that the evacuation confidence is signed, and the confidence is thus, strictly speaking, the absolute value of the evacuation confidence annotation. As described in the Experimental Results section, we use a 1D affine transformation to map the [-100,100] range to the normalised [0,1] range in order to comply with the standard sigmoid function used in neural networks.

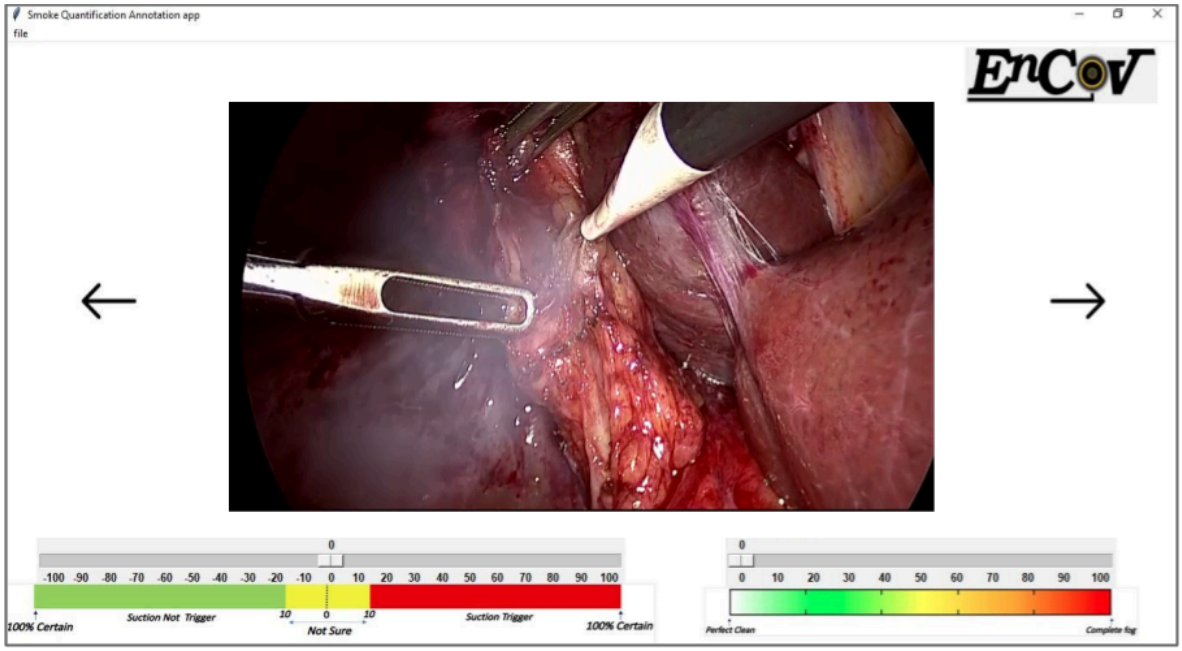

Figure 1. The GUI specifically developed for data annotation. The proposed GUI is based on a simple computer program displaying a single window. It is used in this study to collect annotations for smoke quantification and evacuation recommendation confidence. The image is shown on the top and the expert can use the left and right arrows to navigate through the dataset. The smoke evacuation confidence is chosen with the bottom left slider widget and the smoke quantification with the bottom right slider widget, simply by moving the cursor position on these. The program records the expert's annotations in a file for further processing.

As mentioned in the Data Sources section, most of the Smoke_Cholec80 images show a very low level of smoke. We thus carefully searched for the images with stronger smoke levels, in order to create a balanced SubCholec dataset. To this aim, we first performed a simple histogram analysis on the smoke class of the Smoke_Cholec80 dataset to create an initial estimate of how much smoke each image contains. We used SPA as in [7] with its default parameters. The figure illustrating the frequency of samples obtained with this method is shown in the Supplementary Materials section, figure 6. We then selected an initial subset of 3,715 images by splitting this histogram into five bins and selecting the same number of images per bin. This allowed us to create an initial balanced dataset. We then prevented the selected images from being too similar by introducing an image similarity measurement. For that, we extracted and compared Gabor features to remove overly similar images from the initial dataset, ending up with 1,005 images. We eventually requested an expert to annotate this dataset. We realised that our dataset was still unbalanced in terms of the E+ versus E- classes. We thus trained a classification neural network as described in the General Neural Network Training Strategy section on Smoke_Cholec80, and by investigating the estimated probability of classes, we selected another subset of images that makes our initial dataset balanced. Eventually, we ended up with 882 images. The properties of SubCholec are summarised in table 1.

Dataset Split and Preparation

We split the dataset SubCholec into training, validation and test sets in the following manner. Recall that the training and validation sets are both used for machine learning, in particular to train the neural networks, respectively to estimate their weights and hyperparameters, and that the test set is used to evaluate the result. We use a random split, selecting 600, 176 and 106 images for the training, validation, and test sets, which respectively represent typical rates of 68%, 20% and 12% of the dataset. In order to strengthen the evaluation, we then upgrade the dataset in two ways.

The first upgrade regards the SubCholec dataset and has an impact on all its training, validation and test splits. It is done by selecting 233 images from the SubCholec dataset and having them annotated five times by five annotators, with the goal to compute inter-expert variability. We name this dataset SubCholec233. Concretely, SubCholec233 is an excerpt of SubCholec and entirely includes its test set. We collected this dataset considering two facts. First, given the limitation in expert availability we had to limit the multiple annotations to a feasible number of images. SubCholec233 comprises 233 images, which allows one to obtain a meaningful measurement of inter-expert variability. Second, we selected these 233 images from the SubCholec dataset because this dataset was originally collected from eighty surgeries and is the largest and the most diverse one among our available datasets. We thus have five annotations per image which will be used to measure the inter-expert variability, as described in the Experimental Results section. There is still a challenge in defining the reference or consensus annotation for these images having multiple annotations each. Fortunately, standard approaches exist, such as the STAPLE method [13], which simultaneously estimates the reference annotation and each annotator's performance through an iterative process. This method is used in multiple studies, for instance in the context of prostate cancer staging analysis [14], to fuse the annotations from different pathologists and to create a reference segmentation annotation. We have used the same strategy to extract a consensus annotation per image by iteratively computing experts proficiency weights and adjusting their contribution to the reference annotation depending on their agreement with others. This leads to the Proficiency Weighted Smoke Levels (PWSL) and Proficiency Weighted Evacuation Confidences (PWEC). Note that as we have multiple annotations for SubCholec233, we extract the binary evacuation recommendation annotations from PWEC. Figure 2 summarises the procedure for preparing consensus annotations for SubCholec233.

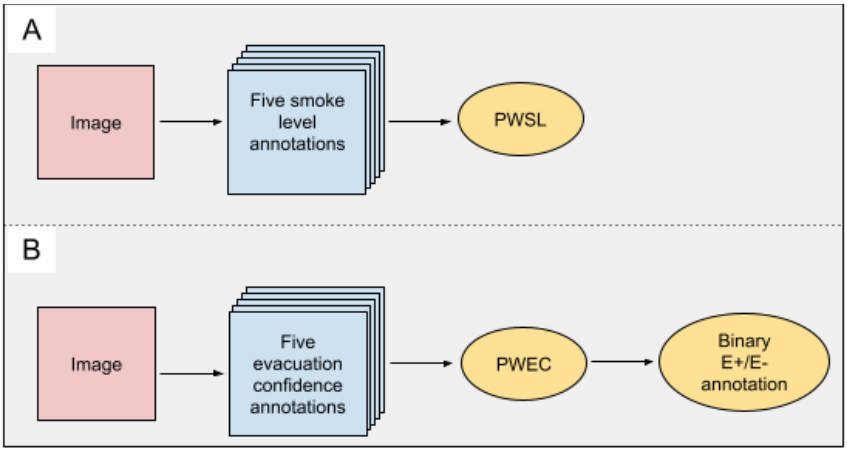

Figure 2. Steps taken in preparing the SubCholec233 dataset annotation consensus. A: Collecting smoke level annotations from five experts and estimating PWSL. B: Collecting smoke evacuation confidence annotations from five experts and estimating PWEC and E+/E-.

The second upgrade regards the test set is to add the datasets SubLapGyn and SmokeGyn1 entirely to it. Recall that SubLapGyn, which is an excerpt of the public dataset LapGyn4_v1.2, and SmokeGyn1 both exclusively contain gynecologic images, while our neural network is trained with the SubCholec dataset, exclusively consisting of cholecystectomy images. Adding these two datasets thus makes evaluation strong.

The properties of the collected datasets and their split in training, evaluation and test are summarised in table 1 and figure 3.

Dataset	**Surgery type**	**Number of images**	**Source**	**Usage**	**Number of expert annotations per image**
SubCholec	Cholecystectomy	882	Smoke-Cholec80	Train-Validation-Test	5 for the test set
SubLapGyn	Gynaecology	100	LapGyn4_v1.2	Test	1
SmokeGyn1	Gynaecology	50	Our Hospital	Test	1
SubCholec233	*Cholecystectomy*	*233*	*Smoke-Cholec80*	*Variability analysis*	*5*

Table 1. Summary of the collected datasets and their split in training, validation and test. SubCholec233 is an excerpt of SubCholec and thus not strictly speaking a dataset. Refer to figure 3 for a graphical visualisation of the datasets.

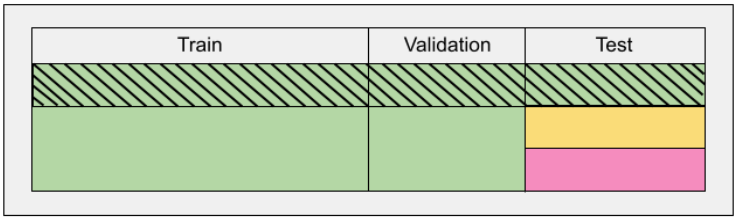

Figure 3. Schematic representation of the datasets and their train-validation-test split. The green, yellow and pink colours represent the SubCholec, SubLapGyn and SmokeGyn1 datasets, respectively. SubCholec is split over the train, validation and test sets while SubLapGyn and SmokeGyn1 are only used to form the test set. The green area with a hatch pattern represents SubCholec233, for which multiple annotations were collected. The consensus of these multiple annotations are used in the train, validation and test sets. Aside, the multiple annotations of SubCholec233 are independently used for inter-expert variability analysis.

Machine Learning

We describe the proposed machine learning methods used to address the three tasks at hand. Figure 4 shows the overall flowchart, depicting the tasks, the data and the ways to connect them. This clearly shows that the smoke evacuation confidence and smoke evacuation recommendation tasks can be approached by end-to-end neural networks or indirectly. We first provide general points on the end-to-end neural network training strategy. We then discuss the proposed machine learning approaches on a task-wise basis.

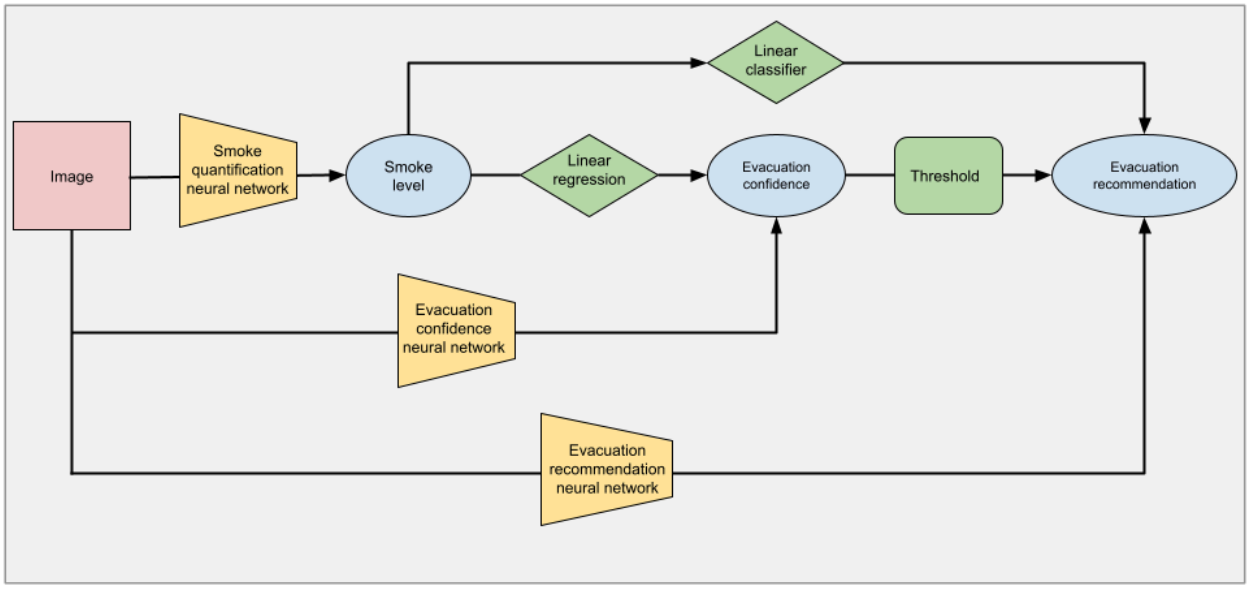

Figure 4. The surgical image shown left is the input data. The three tasks are shown as light blue ellipses. The black arrows indicate the flow between the data and the tasks. The end-to-end neural networks solving the tasks directly from the image are shown as yellow

squeezed boxes. The other machine learning steps connecting the tasks are shown as green shapes. We use them to construct an indirect method for the evacuation confidence task and three indirect methods for the evacuation recommendation task. We thus have a total of seven methods for the three tasks.

General Neural Network Training Strategy

We use CNNs, which are neural networks that proved to outperform classical methods in several machine learning tasks. We use ResNet50 [15] as backbone architecture. ResNet50 uses identity shortcut connections. These connections provide a direct path to the early layers and thus enable the gradient to flow through the deep layers. This results in efficient training of all layer parameters. This architecture contains over 23M parameters, and thus, as for many other medical tasks, there is insufficient data to train them from scratch. We use two strategies to overcome this issue. First, we use heavy random data augmentation with on-the-fly image generation. We use image rotation within 0 to 40 degrees, flipping left-right and up-down, shearing up to 20% along the horizontal axis, and zooming within 0.8 to 1.2 to increase the training dataset size. This ensures that the model is trained with new augmented images at each epoch, mitigating overfitting. Second, we use transfer learning, which re-purposes a pre-trained neural network to a new task. Concretely, we modified ResNet50, which was originally designed to solve the ImageNet challenge with 1,000 classes. We replaced its last classification layer by a layer adapted to the tasks at hand. The choice for this modified layer and for the loss function depend on the target task and are specifically explained below. We trained all networks with a batch size equal to 32 and using the ADAM optimiser with a low learning rate of 0.0001, well-adapted to the fine-tuning regime. We use Python with the Keras framework [16] to implement the proposed methods on a PC with an NVidia Geforce 2080 GPU card running Linux.

Smoke Quantification

Smoke quantification is a regression task, which we model with an end-to-end neural network, as shown in figure 7, which we name `Qu`. We modified the above-described ResNet50 classification architecture to solve this task, considering two challenges. First, if one simply uses the sigmoid activation function forming the last layer of ResNet50, one faces the issue of vanishing gradient. Second, if one substitutes the sigmoid with a linear activation function to prevent vanishing gradients, there is no guarantee that the final predictions fall within the desired range. To address these two challenges, we trained the neural network in two steps. In the first step, we replaced the sigmoid activation function with a linear one. This addresses the vanishing gradient issue and allows the neural network to learn the typical prediction range. In the second step, we continue training but replace the linear activation with a modified sigmoid activation function, ensuring that the final output of the neural network falls within the desired range; recall that the annotations are between 0 and 100, thus if we use the normalised values the desired range is between 0 and 1. We used the Mean Squared Error (MSE) loss function for

both training steps. The details regarding the modifications applied to the sigmoid function and the used loss function can be found in the Supplementary Materials section.

Smoke Evacuation Confidence

Smoke evacuation confidence is a regression task, which we model with an end-to-end neural network and an indirect method, as shown in the Supplementary Materials, figure 8.

End-to-end neural network method, `EC1`. We use the same end-to-end strategy as in smoke quantification, where we modified the ResNet50 architecture to solve the regression task. The only difference is that we use the smoke evacuation confidence annotations for training the neural network, instead of the smoke quantity annotations[6].
Indirect method via smoke quantification, `EC2`. We use the smoke level predictions obtained by the smoke quantification neural network `Qu` to estimate evacuation confidence, which we follow by a linear regression model as:

$$z = \theta y + \lambda \quad (2.4)$$

where z is the predicted smoke evacuation confidence and y is the smoke quantity, and θ and λ are the two model parameters to be estimated. More details for the estimation method are provided in the Supplementary Materials section.

Smoke Evacuation Recommendation

Smoke evacuation recommendation is a classification task, where one aims to classify images into two classes, E+ versus E-. We have proposed four approaches, as shown in the Supplementary Materials section, figure 3.

End-to-end neural network method, `ER1`. We used ResNet50 with an end-to-end strategy. Since ResNet50 was originally designed for the ImageNet multiclass classification challenge, we adapted the architecture to the binary classification task by using a sigmoid activation function at the last layer to obtain the class probability. We apply a simple threshold to this probability to obtain the classification output. Training is performed with a cross-entropy loss function.

Indirect methods via evacuation confidence, `ER2` and `ER3`. In the second and third methods we used the estimated evacuation confidence from methods `EC1` and `EC2` and applied a simple

[6] We used a simple 1D affine transformation to map the [-100,100] range to the normalised [0,1] range in order to comply with the standard sigmoid functions used in neural networks. Therefore, we also back-transformed the normalised value obtained by inference with the neural network to the original interval. This is straightforwardly achieved with the simple rule 200* (x-0.5), where x represents the normalised value.

threshold to estimate the evacuation recommendation. Specifically, the positive values are classified as E+ and the non-positive values are classified as E-.

Indirect method via quantification, ER4. In the fourth method we trained a linear classifier on smoke levels to estimate evacuation recommendations. Unlike the second and third methods which estimate the evacuation recommendation from the evacuation confidence, we now aim to directly classify the images into E+ or E- based on the smoke level. We have used evacuation recommendation annotations and trained a linear SVM binary classifier [17] to predict evacuation recommendation. The mathematical formulation of the classifier can be found in the Supplementary Materials section. The trained classifier applies directly to the smoke level in order to estimate the evacuation recommendation class.

Experimental Results

The following three sections are each related to one of the tasks at hand. For each task, we start with data analysis, in which we investigate the experts's annotations and measure the inter-expert variability. We then discuss the proposed methods' training. We finally evaluate the methods' performance.

Smoke Quantification

Data Analysis

As described in the Data Annotation section, we collected multiple annotations from five experts for the SubCholec233 dataset. In order to analyse the agreement between experts, we evaluated inter-expert variability for the *i*-th image by the standard deviation std_i, and formed the total inter-expert variability, denoted IEV, by averaging over all $N = 233$ images, giving:

$$IEV = \frac{1}{N}\sum_{i=1}^{N} std_i = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{1}{M}\sum_{j=1}^{M}(expert_{ji} - m_i)^2} \quad (3.1)$$

where m_i is the average of annotations for the *i*-th image, $M = 5$ is the number of experts, and $expert_{ji}$ is the annotation of the *j*-th expert for the *i*-th image. The results are shown in table 2. We observe that the experts have on average 13.98 pp (Percentage Points) variability in annotating smoke levels. This variability among the annotators was expected and is consistent with the high level of subjectivity inherent to the requested annotation. We continue the analysis by computing a second statistic named the average experts's error, denoted as AEE. Recall that there was no definitive unambiguous annotation for smoke quantification annotation. We thus defined a consensus annotation using proficiency weighting in the form of PWSL in the Dataset Split and Preparation section. Computing AEE facilitates a direct comparison between data variability and predictor performance. Specifically, we have:

$$AEE = \frac{1}{M}\sum_{j=1}^{M}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(expert_{ji} - pwsl_i)^2} \quad (3.2)$$

In other words, we first compute the consensus score to serve as reference annotation and then average each annotator's discrepancy with respect to this reference score. The AEE is shown in table 2, which is 17.61 pp. Importantly, in order to have a fair comparison with machine learning models, a similar statistic is used in the Prediction Performance section to compute the test error of our neural network predictor.

Inter-Expert Variability	Average Experts's Error	Qu error on training dataset	Qu error on validation dataset	Qu error on test datasets
13.98 pp	17.61 pp	13.58 pp	15.27 pp	15.77 pp - 16.00 pp - 25.87 pp

Table 2. Inter-expert variability, neural network predictor Qu error and average experts's error for the smoke quantification task. The results in the Qu error on test datasets column are related to the SubCholec-Test, SubLapGyn and SmokeGyn1 datasets, respectively. Hence, the first number is directly comparable to the average experts's error.

Training

During the training phase, we monitor the convergence of the model using the Root Mean Square Error (RMSE), which represents the square root of the MSE from equation (2.3) on the SubCholec training and validation sets. We used a stopping criterion based on the improvement in performance. Specifically, if there is no more than a 0.5 pp improvement in the RMSE over the last 50 epochs, we considered the training process converged and terminated it. The final RMSE obtained for the SubCholec validation and training datasets are reported in table 2. We observe that the Qu error for both the training and validation datasets fall within the range of the experts's error available from the AEE analysis.

Prediction Performance

We evaluate the performance of Qu using the PWSL from the Dataset Split and Preparation section, similarly to the average experts's error statistic. The test error is thus defined by the Consensus Root Mean Square (CRMS) error over differences between the neural network output and the PWSL:

$$CRMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_i - pwsl_i)^2} \quad (3.3)$$

where p_i is the smoke level predicted for the i-th sample and $pwsl_i$ is its consensus annotation score. The `Qu` neural network error is reported in table 2 for all three collected datasets. The weighted average error over all datasets is 18.45 pp. Comparing this error with the average experts's error of 17.61 pp, suggests that the predictions are on par with the experts.

Evacuation Confidence

Data Analysis

We follow the same procedure as for smoke quantification. Using equation (3.1), the IEV is 29.35 pp for smoke evacuation confidence. This value is significantly higher than the IEV observed for smoke quantification, which was 13.98 pp. This substantial difference is explained by the fact that experts need more temporal information to annotate the evacuation confidently. We observed that some experts are more sensitive to the location of the smoke, meaning that with smoke occluding the target surgical site, they confidently activated evacuation. In contrast, other experts are more sensitive to the intensity of the smoke. We thus conclude that annotating evacuation confidence from still images leads to large IEV.

Training

Method `EC1`. For this end-to-end method we followed the same strategy and stopping criterion as for `Qu`. The final RMSE obtained for the SubCholec training and validation datasets are reported in table 3. Notably, the `EC1` errors for both the training and validation datasets are lower than the average errors made by human experts. This is explained by the selection of the expert used to annotate the training dataset, who was the expert with minimal divergence from consensus score. Nonetheless, this also shows that the method compares favourably with the experts.

Method `EC2`. For this indirect method, we train the linear model defined in equation (2.4) to fit our data. Using the training SubCholec samples and minimising the SSE defined in equation (2.5), the θ and λ parameters are obtained as 1.1 and -63.37% respectively. This trained linear predictor is shown in figure 5 with a red line alongside the training samples. This linear model leads to RMSE of 16.35 pp and 17.35 pp on the SubCholec training and validation datasets. It is noteworthy that these errors are obtained using the annotated smoke levels as input to the linear model, indicating that with an ideal `Qu` network and a simple linear model we should observe a similar error range. We observe in figure 5 that an increase in the smoke quantity does not consistently result in an increase in the level of evacuation confidence. There exist samples with high evacuation confidence levels associated with low smoke levels and vice versa. This is due to the fact that experts take into account additional information beyond the smoke quantity when making evacuation recommendations.

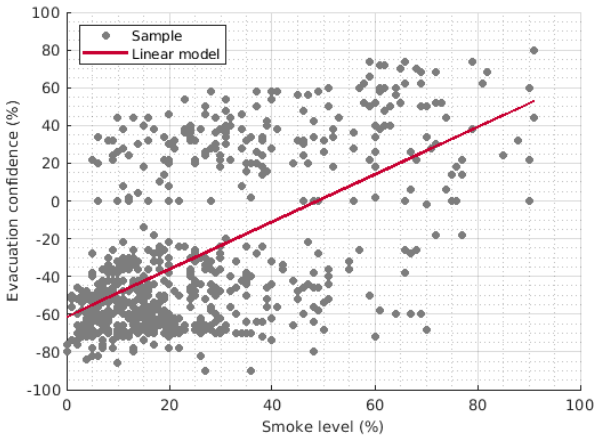

Figure 5. The linear regression model that estimates evacuation confidence from smoke levels.

Prediction Performance

We followed the same definitions as equations (3.3) and (3.4) to measure the performance of the proposed `EC1` and `EC2` methods. The only difference is that we used PWEC instead of PWSL. The error on the three test datasets is shown in table 3. The average weighted error for `EC1` and `EC2` are 25.90 pp and 23.60 pp, respectively. The results indicate that both of the proposed methods' errors fall within the range of AEE of 27.35 pp. Additionally, we observe that the performance of `EC2` is close to `EC1`. This can be attributed to the fact that the `Qu` model demonstrated a performance close to ideal.

Inter-Expert Variability	Average Experts's Error	EC1 error on training dataset	EC1 error on validation dataset	EC1 error on test datasets	EC2 error on test datasets
29.35 pp	27.35 pp	19.26 pp	20.35 pp	21.20 pp - 30.27 pp - 27.15 pp	19.3 pp - 24.21 pp - 31.5 pp

Table 3. Inter-expert variability, proposed methods' error and average experts's error for the evacuation confidence task. The results in the `EC1(EC2)` error on test datasets columns are related to the SubCholec-Test, SubLapGyn and SmokeGyn1 datasets, respectively. Recall that `EC1` is the end-to-end neural network method and `EC2` is the indirect method via smoke quantification.

Evacuation Recommendation

Data Analysis

We present the inter-expert variabilities concerning the evacuation recommendation annotations. Unlike the case of evacuation confidence, where we used IEV to analyse the confidence level annotations, the evacuation recommendations are binary in nature. Therefore, one cannot use the same metric in this context. We thus study the inter-expert agreement using Fleiss's kappa [18], a generalisation of the kappa agreement measure which works for any number of experts. This statistic expresses the extent to which the observed agreement among experts exceeds the expectation that all the experts make their decisions by chance. This statistic is presented in table 4. Based on this measurement we reject the null hypothesis, meaning that the observed inter-expert agreement is not accidental. According to the guidelines [19,20], the agreement level is moderate, presenting some class ambiguity for the experts.

Fleiss Kappa	Confidence Interval	Agreement
0.0402	0.39127 - 0.4124	Moderate

Table 4, Fleiss's kappa statistic for inter-expert variability of E+ versus E- classes.

Training

Method `ER1`. We monitored the accuracy during the training phase to devise a stopping criterion. The accuracy for a binary classification task is defined as:

$$Acc = \frac{T_p + T_n}{N} \quad (3.4)$$

where *Tp* is the number of samples truly classified as E+ and *Tn* is the number of samples truly classified as E-. We considered the model to be converged if we did not observe more than 0.1% progress in the model accuracy on the validation dataset in the last 50 epochs. The best accuracies and their corresponding sensitivities achieved by `ER1` on the SubCholec training and validation datasets are reported in table 5. We observe that a substantial difference exists between the performance of `ER1` on the training and the validation datasets. This discrepancy suggests that `ER1` overfits and that the features it learned do not generalise to other datasets.

	ER1 (train)	**ER1 (validation)**	**ER1 (test)**	**ER2 (test)**	**ER3 (test)**	**ER4 (test)**
Accuracy	99.6	77.5	83.19 -82.2-75.5	84.91-77.72-64	85.85-75.59-63	76.42-88.19-68.19
Sensitivity	99.5	78.7	84.6-81.8-74.5	72.96-75.45-61.23	72.85-68.64-42.2	81.48-72.73-52.2

Table 5. Inter-expert variability, methods' accuracy and sensitivity for the evacuation recommendation task. The results for test datasets are related to SubCholec-Test, SubLapGyn and SmokeGyn1, respectively. Recall that `ER1` is the end-to-end neural network method, `ER2` and `ER3` are indirect methods via evacuation confidence, and `ER4` is the indirect method via smoke quantification.

Methods `ER2` and `ER3`. These two methods do not have new parameters to be learned since they use a zero threshold on the predicted evacuation confidence level, meaning that the predictions greater than zero are classified as E+ and the predictions lower than zero as E-.

Method `ER4`. The estimated parameters for the trained linear SVM binary classification are 0.069 and -1.7586% for *w* and *b* parameters defined in equation (2.6) respectively. Since our input feature to the SVM classifier is solely the smoke level, the |*b/w*| = 25.487% represents the trained threshold value that we can apply to the predicted smoke level in order to classify evacuation recommendations.

Prediction Performance

For all of the proposed methods, we calculate the accuracy as defined in equation (3.4) and sensitivity as:

$$Sen = \frac{T_p}{T_p + F_N} \quad (3.5)$$

where *Fn* is the number of samples wrongly classified as E-. In order to calculate these two metrics we used the binary annotations obtained from PWEC as consensus annotations described in the Dataset Split and Preparation section and shown in figure 2. The results are provided in table 5 for all of the four proposed methods and the three test datasets. The average weighted accuracies for the proposed methods are 81.30%, 78.02%, 77.37%, and 79.41%, respectively. From these results we have that the performance of `ER1` on test datasets falls in the range of its performance on validation dataset, but still considerably far from its performance on the training dataset. This trend is also observable for the other methods where there exists a substantial difference in performance on test datasets. This suggests that the model may overfit the training data, which can be explained by the moderate agreement between the experts in annotating evacuation recommendations. Furthermore, using the consensus binary annotations shown in figure 3, the average experts's accuracy is 76.78%. Notably, the proposed methods exhibit accuracies falling within the same range, validating the reliability of our proposed approaches. In addition, the performance of `ER3` and `ER4` on test datasets is on par with `ER1` and `ER2`. This can be explained by the fact that they have `Qu` in their back-end, whose performance is close to ideal.

Discussion

Smoke analysis involves several tasks, including clear versus smoky image classification and smoke quantification. While the classification task attempts to reach a binary decision regarding the amount of smoke, the quantification task goes a step further and attempts to estimate the amount of smoke in the image on a continuous scale. The literature has focused on the classification task only; the other possible tasks, in particular the quantification tasks, have not been studied. Smoke classification has been addressed with two main approaches. The first approach uses classical image processing and machine learning. The method in [21] analyses the visual motion in a laparoscopic video, under the assumption that this motion is primarily due to smoke. The method extracts a set of kinematic features from optical flow to describe motion and uses a one-class Support Vector Machine (SVM) to classify short video clips as smoky versus clear. The classifier is trained along with feature selection. This method has a limited performance when visual motion occurs owing to other phenomena than smoke, which is problematic in practice. The method in [7] uses the observation that smoke tends to create some greyish or colourless regions in the images. The images are classified by finding the local maxima related to these regions in the histogram of the saturation channel. The main problem with this method is that other colourless parts such as surgical instruments or specular light reflections have the same behaviour in the histogram and may confound the method. The method in [22] extracts several features based on colour, texture patterns and motion from a short video clip and uses a one-class SVM as classifier. The classifier is trained without feature selection. Although the authors did not report this in their investigation, the proposed model is expected to suffer from motion-related activities in laparoscopic videos and tissues with colours comparable to smoke. The method in [8] uses Saturation Peak Analysis (SPA), which converts an input image to the HSV (Hue, Saturation and Value) colour space and explores its saturation channel histogram. Smoky areas are segmented as the significant local bin maxima in the histogram. Smoky versus clear image classification is then decided by thresholding the smoky image area. The method is simple but highly sensitive to the presence of unicolour elements in the image, such as the surgical instruments and light reflections.

The second approach to smoke classification uses deep neural networks. The method in [6] uses a pre-trained GoogLeNet model, trained with a supervised approach from approximately 30K private laparoscopic images with transfer learning. The results are compared with a classical classification method based on SPA, which is described directly above. The main finding is that the deep learning approach achieves excellent results and that the classical method is a good indicator for the presence of smoke while being computationally much less expensive. The same team extended their work in [8] by creating a balanced smoky versus clear dataset called Smoke_Cholec80 containing about 100K images collected from the public dataset Cholec80 [12]. They used this dataset to attempt creating a real-time smoke classification neural network. They tried a shallow AlexNet architecture, trained with colour images and the GoogLeNet architecture, trained from the saturation channel. Eventually, both neural networks are fast but not real-time, whilst only slightly improving classification performance against the much faster SPA approach. The method in [10] uses the DarkNet architecture. The neural network is trained by transfer learning on the public smoke dataset [8]. The authors reported improvements in

performance by using other unannotated datasets and applying a self-training approach with semi-supervised noisy student models. The method in [9] uses the temporal context to improve classification performance. A DarkNet architecture is trained as a feature extractor on single-frame inputs and appended to the head of a recurrent neural network for final classification. The neural network is then fine-tuned using unannotated data from the same domain with semi-supervised training. An attempt to reduce the effect of the presence of electrosurgical instruments on the classifier using a balancing strategy is proposed to prevent the neural network to predict the instruments instead of the smoke.

The above-reviewed smoke analysis methods all address the problem of classifying laparoscopic images as smoky versus clear, with compelling results obtained for several methods. However, none of them addresses the problems of quantifying the smoke level and automatically recommending smoke evacuation. There currently do not exist annotated datasets for these tasks. Our contributions address these two problems.

We have evaluated the usability of machine learning to automatically predict the smoke level and the need for smoke evacuation from a surgical image. To this aim, we have developed a comprehensive framework, including three datasets with expert annotations, inter-expert variability analysis and have proposed machine learning methods to handle smoke-related tasks. Specifically, we have investigated three tasks: smoke quantification, smoke evacuation confidence, and smoke evacuation recommendation.

Regarding smoke quantification, the neural network reached an error of 18.45 pp. This number means that, on average, the prediction is 18.45 pp away from the annotation. Note that the average is computed across three test datasets, where the contribution of each dataset is weighted by the inverse number of images it contains. It might be difficult to interpret the errors' numerical values because, even if expressed in pp, these errors do not directly reflect a physically or statistically meaningful quantity. This is the reason why we included both the experts's error (AEE) and the machine learning model's errors. The former sets a reference to which the latter can be compared, allowing one to use a relative evaluation rather than an absolute one. We obtained an AEE of 17.61 pp, meaning that, on average, the annotators have a 17.61 pp error in their annotation compared to the consensus score. We observe that the machine learning model's error of 18.45 pp is on par with the experts's, indicating a satisfying performance of the machine learning model. Despite achieving better average error rates on the SubCholec-Test and SubLapGyn datasets compared to the reference error of AEE, one may argue that the test metrics for the SmokeGyn1 dataset consistently exhibit poorer performance compared to the SubCholec and SubLapGyn datasets. We attribute this discrepancy to two key factors. First, the machine learning model is trained using the SubCholec dataset, sourced from cholecystectomy procedures, while SmokeGyn1 was derived from gynaecological procedures. This introduces a shift in the training dataset domain relative to this test dataset. Second, the SmokeGyn1 dataset was intentionally collected from challenging images. The degradation of the metrics on this dataset were thus expected. Yes, this demonstrates the machine learning model's capability to handle such challenging cases without breaking down, albeit with lower accuracy.

Regarding evacuation confidence, the errors of the proposed methods are larger than smoke quantification predictions, reaching 23.60 pp. Nevertheless, we have observed a larger AEE of 27.35 pp too, suggesting that even the experts find it challenging to confidently predict evacuation from still images. This is mainly because factors such as the spatial distribution of the smoke, in particular the extent at which it covers the surgical site, and temporal considerations such as the surgical phase play a crucial role in their decision-making process.

Subsequently, we have assessed several classifiers to measure the separability of the E+ and E- classes involved in evacuation recommendation. Concretely, we have designed four classification experiments. They all seek to achieve the E+ versus E- classification but using different approaches; `ER1` and `ER2` by direct methods from input images and `ER3` and `ER4` by indirect methods via predicted smoke level. The results shown in table 5 lead to the observation that the classification performance is on par for both approaches. This was expected since the smoke quantification error obtained by `Qu` was shown to be on par with the experts. This suggests that quantifying the smoke level and then predicting the evacuation recommendation may be a more effective approach compared to directly predicting evacuation recommendation from the input image. This is a strong finding of our study, showing that one can leverage the simpler task of smoke quantification to predict the need for smoke evacuation. This finding has a strong impact on selecting the right approach to solving smoke evacuation tasks. This shifts the problem of automatic smoke evacuation which, when faced directly, requires one to deal with noisy and uncertain evacuation annotations, to the much simpler problems of quantifying the smoke and correlating the smoke quantity with the need for smoke evacuation. This finding will likely have a huge impact on future work, from data collection to machine learning model selection.

Conclusions

We have presented a new approach for smoke management in MIS based on image contents analysis. Through performing multiple inter-expert analyses, we have established a benchmark based on expert proficiency to determine the discrepancy between machine learning predictions and consensus scores. Our finding is that smoke can be reliably quantified by neural networks with results on par with the experts for still images, while evacuation recommendation using still images is more challenging even for the experts as we found larger inter-expert variability. However, we have shown that smoke quantification can be leveraged to predict the need for evacuation on par with the experts. In future work, we plan to evaluate the effectiveness of a recommendation system that displays the predicted smoke quantity as a simple visual warning signal on the MIS screen. Another possibility would be to incorporate the predictions as a signal to automatically trigger a smoke evacuation system. Additionally, we plan to explore the incorporation of temporal and contextual information into our neural network model to overcome the limitations of still image analysis. By contextual information, we mean that knowing the location of the smoke and the surgical site, even roughly, can significantly impact one's decision on evacuation activation. By temporal information, we mean that the smoke necessarily moves

continuously within the surgical scene. In other words, it does not appear or disappear suddenly. This temporal information can be exploited by applying the proposed methods to the individual images of a continuous video and using a simple temporal filtering of the results or by using recurrent neural networks. Overall, the proposed methods have the potential to be integrated into the OR to improve patient safety and surgical outcomes by reducing the risks associated with surgical smoke exposure.

Supplementary Materials

Modification of the Sigmoid Activation Function

The general modified sigmoid function is defined as:

$$f(x) = \frac{1}{1+e^{-\alpha(x-\beta)}} \quad (2.1)$$

The β parameter shifts the function and the α parameter controls its slope. We choose these parameters so that the sigmoid approximates the linear activation on the desired range as best possible. For that, we choose $\beta = 0.5$, so that the sigmoid is centred within the desired range. We then choose α, so that the sigmoid has a slope of 1 at $x = 0.5$. The slope is given by the sigmoid's derivative as:

$$\frac{d}{dx}f(x) = \frac{\alpha\, e^{-\alpha\,(x-0.5)}}{(1+e^{-\alpha(x-0.5)})^2} \quad (2.2)$$

Equating it to 1 and setting $x = 0.5$ leads to $\alpha = 4$.

Qu method Loss Function:

Mean Squared Error (MSE) loss function is used for both training steps, which measures the difference between the predicted and true values of a continuous target function as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2 \quad (2.3)$$

where N is the number of images in the training dataset, y_i is the annotated smoke quantity of the *i*-th image and $\widehat{y}_i$ is the predicted value.

EC2 Parameters Estimation

The θ and λ parameters are estimated using the least squares method with the SubCholec training samples. The fitting hence minimises the following Sum of Squared Errors (SSE):

$$SSE = \sum_{i=1}^{N} (z_i - \hat{z}_i)^2 \quad (2.5)$$

where $\hat{z}_i = \theta y_i + \lambda$ is the *i*-th image evacuation confidence prediction for a given smoke quantity y_i in the *i*-th image.

SVM Classifier Formulation

For a given smoke quantification annotation y_i, the SVM establishes a separation plane between the two classes defined by $w \cdot y_i + b$, satisfying the following conditions:

$$\begin{aligned} w \cdot y_i + b &\geq +1 \qquad y_i \in E+ \\ w \cdot y_i + b &\leq -1 \qquad y_i \in E- \end{aligned} \quad (2.6)$$

where w is the normal vector to the separation plane and b is the bias. The SVM determines these parameters by maximising the margin between the two classes.

Supplementary Figures

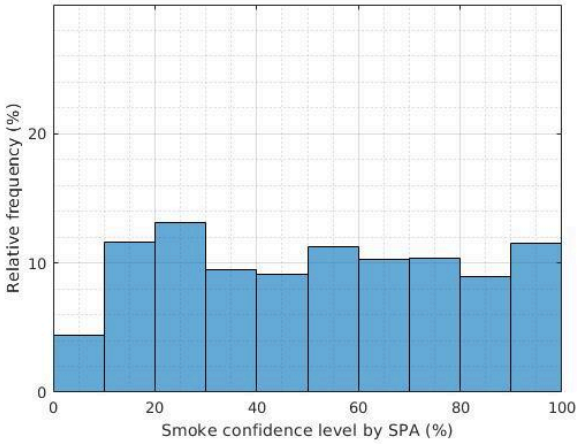

Figure 6. Relative histogram of smoke confidence levels for the SmokeCholec80 dataset obtained from the SPA method [7], which we used to create the balanced SubCholec dataset.

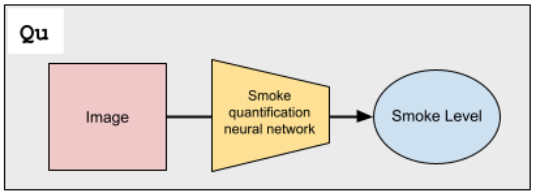

Figure 7. We propose an end-to-end neural network to address smoke quantification. This method is abbreviated as `Qu`.

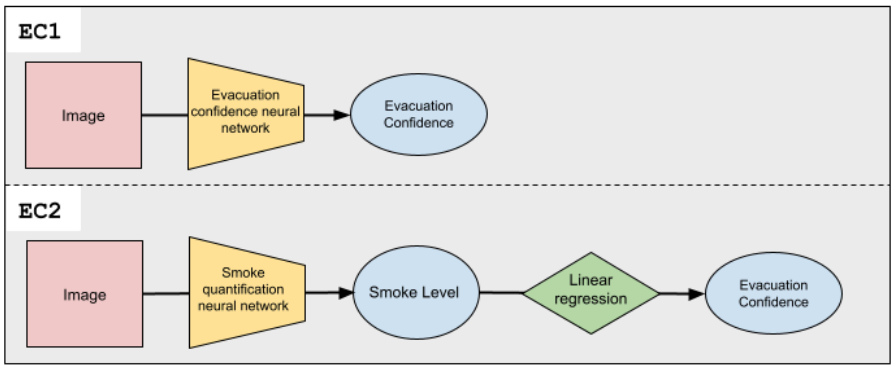

Figure 8. We propose two methods to address smoke evacuation confidence; `EC1`: an end-to-end neural network and `EC2`: an indirect method via smoke quantification.

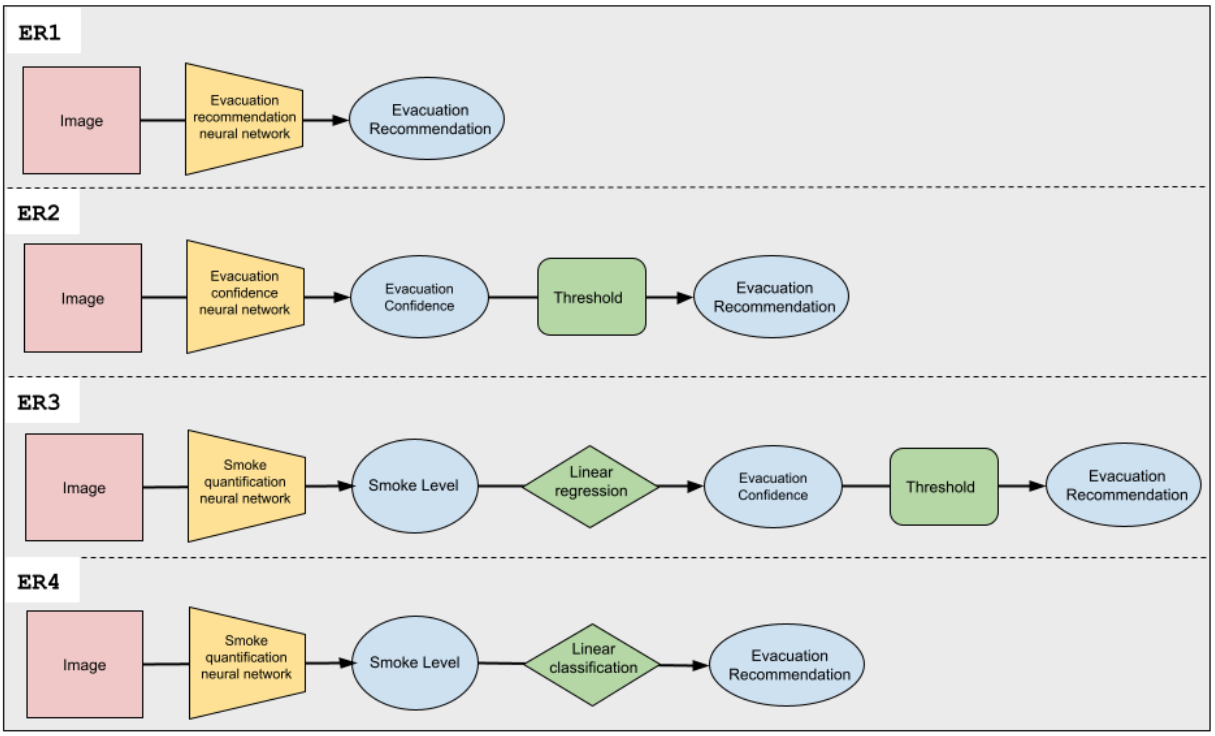

Figure 9. We propose four methods to address smoke evacuation recommendation; `ER1`: an end-to-end neural network, `ER2` and `ER3`: indirect methods via evacuation confidence, and `ER4`: an indirect method via smoke quantification.

Author Contributions

RS and AB developed the scientific method and wrote the paper. RS coded the computer programs. HMA, SMZ, CS, PC, NB and MC annotated the data as surgical experts and gave general advice on surgical smoke management problems.

Disclosure

Nothing to disclose.

Funding

This research has been partially funded by Cancéropôle CLARA within the Proof-of-Concept project AIALO (2020-2023).

Disclaimer

References

[1] Ulmer BC. The hazards of surgical smoke. AORN journal. 2008 Apr 1;87(4):721-38.

[2] Merajikhah, A., Imani, B., Khazaei, S., & Bouraghi, H. (2022). Impact of surgical smoke on the surgical team and operating room nurses and its reduction strategies: A systematic review. Iranian Journal of Public Health, 51(1), 27.

[3] Takahashi H, Yamasaki M, Hirota M, Miyazaki Y, Moon JH, Souma Y, Mori M, Doki Y, Nakajima K. Automatic smoke evacuation in laparoscopic surgery: a simplified method for objective evaluation. Surgical endoscopy. 2013 Aug;27:2980-7.

[4] Balayssac D, Selvy M, Martelin A, Giroudon C, Cabelguenne D, Armoiry X. Clinical and organizational impact of the AIRSEAL® insufflation system during laparoscopic surgery: a systematic review. World Journal of Surgery. 2021 Mar;45:705-18.

[5] Uppal S, Frumovitz M, Escobar P, Ramirez PT. Laparoendoscopic single-site surgery in gynecology: review of literature and available technology. Journal of Minimally Invasive Gynecology. 2011 Jan 1;18(1):12-23.

[6] Weenink RP, Kloosterman M, Hompes R, Zondervan PJ, Beerlage HP, Tanis PJ, van Hulst RA. The AirSeal® insufflation device can entrain room air during routine operation. Techniques in coloproctology. 2020 Oct;24(10):1077-82.

[7] Leibetseder A, Primus MJ, Petscharnig S, Schoeffmann K. Image-based smoke detection in laparoscopic videos. InComputer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures: 4th International Workshop, CARE 2017, and 6th International Workshop, CLIP 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, 2017, Proceedings 4 2017 (pp. 70-87). Springer International Publishing.

[8] Leibetseder A, Primus MJ, Petscharnig S, Schoeffmann K. Real-time image-based smoke detection in endoscopic videos. InProceedings of the on thematic workshops of ACM multimedia 2017 2017 Oct 23 (pp. 296-304).

[9] Reiter W. Co-occurrence balanced time series classification for the semi-supervised recognition of surgical smoke. International Journal of Computer Assisted Radiology and Surgery. 2021 Nov;16(11):2021-7.

[10] Reiter W. Improving endoscopic smoke detection with semi-supervised noisy student models. InCurrent Directions in Biomedical Engineering 2020 Sep 17 (Vol. 6, No. 1, p. 20200026). De Gruyter.

[11] Leibetseder A, Petscharnig S, Primus MJ, Kletz S, Münzer B, Schoeffmann K, Keckstein J. Lapgyn4: a dataset for 4 automatic content analysis problems in the domain of laparoscopic gynecology. InProceedings of the 9th ACM multimedia systems conference 2018 Jun 12 (pp. 357-362).

[12] Twinanda AP, Shehata S, Mutter D, Marescaux J, De Mathelin M, Padoy N. Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE transactions on medical imaging. 2016 Jul 22;36(1):86-97.

[13] Warfield, S. K., Zou, K. H., & Wells, W. M. (2004). Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE transactions on medical imaging, 23(7), 903-921.

[14] Qiu, Yali, et al. "Automatic Prostate Gleason Grading Using Pyramid Semantic Parsing Network in Digital Histopathology." Frontiers in Oncology 12 (2022)

[15] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. InProceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 770-778).

[16] Chollet F. Keras: The python deep learning library. Astrophysics source code library. 2018 Jun:ascl-1806.

[17] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press; 2000 Mar 23.

[18] Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and psychological measurement. 1973 Oct;33(3):613-9.

[19] Cohen J. A coefficient of agreement for nominal scales. Educational and psychological measurement. 1960 Apr;20(1):37-46.

[20] McHugh ML. Interrater reliability: the kappa statistic. Biochemia medica. 2012 Oct 15;22(3):276-82.

[21] Loukas C, Georgiou E. Smoke detection in endoscopic surgery videos: a first step towards retrieval of semantic events. The International Journal of Medical Robotics and Computer Assisted Surgery. 2015 Mar;11(1):80-94.

[22] Alshirbaji TA, Jalal NA, Mündermann L, Möller K. Classifying smoke in laparoscopic videos using SVM. Current Directions in Biomedical Engineering. 2017 Sep 7;3(2):191-4.