




Explaining a model predicting quality of surgical practice: a first presentation to and review by clinical experts

Arthur Derathé¹ · Fabian Reche^{1,2} · Pierre Jannin^{3,4} · Alexandre Moreau-Gaudry^{1,5} · Bernard Gibaud^{3,4} · Sandrine Voros^{1,4} 

Received: 30 November 2020 / Accepted: 27 May 2021 / Published online: 18 June 2021
© CARS 2021

Abstract

Purpose Surgical Data Science (SDS) is an emerging research domain offering data-driven answers to challenges encountered by clinicians during training and practice. We previously developed a framework to assess quality of practice based on two aspects: exposure of the surgical scene (*ESS*) and the surgeon's profile of practice (*SPP*). Here, we wished to investigate the clinical relevance of the parameters learned by this model by (1) interpreting these parameters and identifying associated representative video samples and (2) presenting this information to surgeons in the form of a video-enhanced questionnaire. To our knowledge, this is the first approach in the field of SDS for laparoscopy linking the choices made by a machine learning model predicting surgical quality to clinical expertise.

Method Spatial features and quality of practice scores extracted from labeled and segmented frames in 30 laparoscopic videos were used to predict the *ESS* and the *SPP*. The relationships between the inputs and outputs of the model were then analyzed and translated into meaningful sentences (statements, e.g., “To optimize the *ESS*, it is very important to correctly handle the spleen”). Representative video clips illustrating these statements were semi-automatically identified. Eleven statements and video clips were used in a survey presented to six experienced digestive surgeons to gather their opinions on the algorithmic analyses.

Results All but one of the surgeons agreed with the proposed questionnaire overall. On average, surgeons agreed with 7/11 statements.

Conclusion This proof-of-concept study provides preliminary validation of our model which has a high potential for use to analyze and understand surgical practices.

Keywords Surgical skills · Video-based assessment · Explainable artificial intelligence

Introduction

Although it presents many advantages for the patient compared to open surgery [15], minimally invasive surgery (MIS)

is very complex and challenging for surgeons. Guidelines, good practices, and checklists are key tools to preserve the patient's safety [10], while simultaneously reducing costs [16]. These tools are generally developed by a community of expert surgeons in a surgical specialty who reach a consensus following discussion and compromise [3,4]. In addition, global rating scales (GRS) have been developed to formalize and assess some aspects of surgical practice [2,9].

In recent years, with the widespread availability of artificial intelligence (AI) and machine learning, Surgical Data Science [8], which aims to extract knowledge from surgical data, has made tremendous progress. For instance, in relation to training and evaluation of laparoscopy, it allows the recognition of surgical tasks and can be used to assess technical skills [12]. In both cases, image content and infor-

✉ Sandrine Voros
sandrine.voros@univ-grenoble-alpes.fr

¹ Université Grenoble Alpes, CNRS, Grenoble INP, TIMC-IMAG, 38000 Grenoble, France

² Department of Digestive Surgery, Grenoble University Hospital, Grenoble, France

³ Université Rennes 1, LTSI, UMR_S 1099, 35000 Rennes, France

⁴ Inserm, Paris, France

⁵ Clinical Investigation Center, Innovative Technology, CHU de Grenoble, Grenoble, France

mation on the presence/motion of surgical instruments were of paramount importance in the recognition.

Due to numerous challenges summarized in [8], one of which is lack of explicability, SDS has not yet penetrated laparoscopic operating rooms. How can we trust an AI “black box” for decision-making if the clinical relevance of the information used by the algorithms to make their predictions cannot be explained to or discussed with the experts in the field? This question of explainable AI (XAI) is key in various domains. In SDS, even though some papers do address this XAI question [5], to our knowledge no work has been performed to present the selection criterion of an algorithm to the non-formalized expertise/knowledge of surgeons.

In a previous study [1], we developed a prediction model, which was used to investigate two specific indicators of surgical practice: how well the surgeon exposes the surgical target in the scene (ESS) and the profile of practice (SPP). Based on this work (briefly summarized in section “Previous work”), we now wished to address the explicability question by proposing a proof-of-concept framework (3.2) through which to interpret the model’s prediction results. Similarly to [14,17], we used classical methods of feature extraction for the Support Vector Machine (SVM) technique. Our interpretations were submitted to 6 digestive surgeons from two hospitals, to assess whether they agreed with the model’s choices (section “Results of the survey submitted to experienced surgeons”). We end our article with a discussion of our results and present some perspectives for future studies (section “Discussion”).

Previous work

In [1], we predicted two facets of surgical practice during laparoscopic sleeve gastrectomy. The first was the exposure of the surgical scene (ESS), defined as the accessibility and visibility of the surgical target during the intervention. This notion characterizes the management of the tight surgical space, as well as the restricted view available in MIS. For the surgeon, it is crucial to maintain a clear surgical space as well as good visibility of the surgical target. The second facet was the surgeon’s profile of practice (SPP), which reflected the specific approach used by a particular surgeon, his/her habits of practice. This parameter was important as, even though surgeons follow guidelines and good practices, each individual has a personal way of operating that could be characterized. Overall, we assumed that ESS and SPP would describe different aspects of surgical practice and that video-based information was sufficient to classify them.

For our analysis, we focused on the Dissection of fundus step as it is critical in this surgical technique. Within this surgical step, we particularly analyzed the “dissection activities,” as their aim is to improve the ESS quality for subsequent

surgical actions. The ESS and SPP ground truth were binary variables obtained based on annotation by an expert surgeon involved in the study (for ESS) and knowledge of the practicing surgeon (for SPP). To predict the ESS and the SPP, we created an annotated dataset from laparoscopic images of dissection gestures: each object visible in the images was segmented, and features were extracted from these segmentations. The dataset is described in greater detail in section “Dataset.”

The model took as input the features extracted for all the images in the dataset and was trained once to predict either the ESS or the SPP. Prediction consisted in a normalization followed by a Linear Discriminant Analysis (LDA) for the dimension reduction step followed by application of a Support Vector Machine (SVM) for the prediction per se. We optimized our model in a Train-and-Test Cross-Validation (TnT-CV) environment to avoid an over-fitting bias. Prediction performances were on par with other studies in the field, as summarized in Table 2 (lines 1 and 2).

Interpreting and questioning the prediction of surgical practice

Having produced a model capable of predicting ESS and SPP based on video information, our goals with this study were:

- to identify what input information had been used by the model to make its predictions,
- to assess the relevance of this information, based on feedback from experienced digestive surgeons.

We start by describing the video data used as input of our model (section “Dataset”); then, we present the methodology used to identify the most important input features for the prediction and to build the survey administered to the experienced surgeons (3.2).

Dataset

An expert surgeon annotated the ESS and SPP as binary variables for each image of a dissection gesture (referred to as a “dissection sample”). The spatial state of each of these 10 objects visible in the video (see Fig. 1) was described by 13 features (see Table 1). The 130 resulting variables could then be analyzed by applying “feature-object” semantics.

Our dataset was composed of an input matrix X of size $m \times n$, and two potential output vectors y_{ESS} characterizing the ESS, and y_{SPP} characterizing the SPP, both of size $m \times 1$. We had $m = 735$ samples, and given the 10 objects and 13 features, $n = 130$ variables. These data were used as such in the prediction model presented in section “Previous work.” We constituted 6 clusters allowing us to more closely inves-

Fig. 1 Illustration of the 6 organs and 4 surgical instruments segmented in the surgical videos

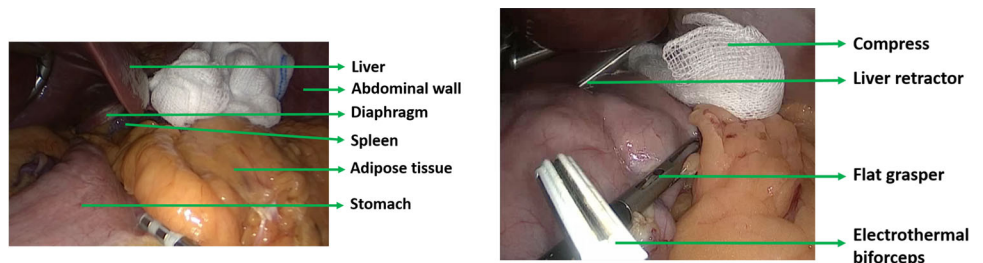


Table 1 Spatial features extracted for each object segmented in the image

Name	#	Description
Barycenter	2	x and y coordinates of the object’s central pixel
Color	3	Color of the object’s pixels in the CIE-Lab color space
Eigenvalue	1	Ratio of the 2 eigenvalues characterizing the object’s shape
Eigenvectors	4	x and y coordinates of the object’s 2 main eigenvectors (determined by PCA)
Perimeter	1	# of pixels on the object’s contour
Surface	1	# of pixel s on the object
Texture	1	Standard deviation of the histogram obtained by a local binary pattern algorithm

tigate the links between the input data and output vectors. Figure 2 illustrates the construction and inclusion relationships used to establish these clusters.

Three clusters aimed to predict the *ESS*:

- Cluster C_{ESS} associates the input data X and the output vector y_{ESS} . It answers the question “*What characterizes the quality of exposure for both surgeons overall?*”
- Cluster C_{ESS}^A associates X_{ESS}^A , the subset of input data X restricted to surgeries performed by surgeon 0, with y_{ESS}^A , the corresponding output subset of y_{ESS} . It answers the question “*What characterizes the quality of exposure as surgeon 0 operates?*”
- Similarly, cluster C_{ESS}^B associates X_{ESS}^B , the subset of X restricted to surgeries performed by surgeon 1, with y_{ESS}^B , the corresponding output subset of y_{ESS} . It answers the question “*What characterizes the quality of exposure as surgeon 1 operates?*”

Three clusters aimed to predict the *SPP*:

- C_{SPP} associates the input data X and the output vector y_{SPP} . It answers the question “*What characterizes the surgeon’s profile of practice independently of the quality of exposure?*”
- C_{SPP}^A associates X_{SPP}^A —the subset of input data X restricted to the 3 surgeries with the highest mean y_{ESS} score - with y_{SPP}^A , the corresponding output subset of y_{SPP} . It answers the question “*What characterizes the surgeon’s profile of practice when the quality of exposure is good?*”

- C_{SPP}^B associates X_{SPP}^B —the subset of input data X restricted to the 3 surgeries with lowest mean y_{ESS} score—with y_{SPP}^B , the corresponding output subset of y_{SPP} . It answers the question “*What characterizes the surgeon’s profile of practice when the quality of exposure is non-optimal?*”

Validation methodology

In order to construct the survey, we studied the input variables that had the most impact on the prediction results, and translated our findings in the form of statements. These statements were submitted to experienced surgeons alongside illustrative video segments. Our validation methodology consisted in 4 steps described in the 4 following sub-sections (see Fig. 3).

Estimating the importance of input variables

Similarly to [7], we were interested in observing the impact of the input variables on performance in the prediction task: for each variable we applied a value permutation and observed its impact on the resulting prediction. The value permutation consisted in shuffling values over each input variable iteratively, to observe the induced prediction shift [13].

We implemented the permutation process within the *TnT-CV* environment in our prediction model to avoid introducing an overfitting bias. For each *CV* split configuration, we first trained the model on a train-fold and evaluated the baseline prediction performance of the model $p_{ref,cv}$. Then, value permutation was performed for each input variable v , the pre-

Fig. 2 Construction and inclusion relationship of clusters—the C_{ESS} cluster with its two subclusters C_{ESS}^A and C_{ESS}^B —the C_{SPP} cluster and its three subclusters C_{SPP}^A , C_{SPP}^B and R (the remaining samples)

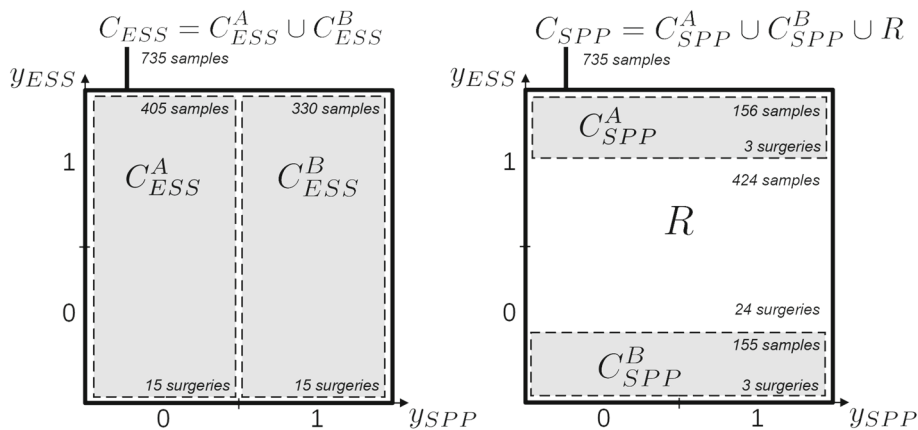
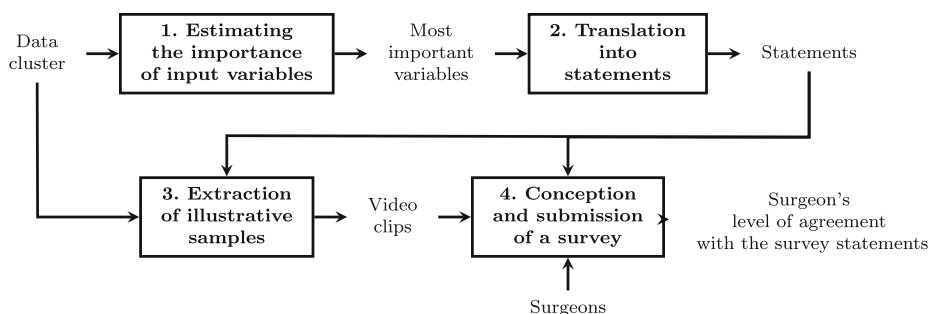


Fig. 3 Methodological steps in model interpretation and assessment



diction performance $p_{v,cv}$ was determined, and the distance to $p_{ref,cv}$ was computed. This distance corresponds to the amplitude of the shift in performance induced by the value permutation: variables that have a smaller distance from the reference are more important to the prediction than those further from the reference.

We defined the importance i_v of input variable v as this distance averaged over the n_{cv} split configurations of the $TnT-CV$ environment:

$$i_v = \frac{1}{n_{cv}} \sum_{cv=1}^{n_{cv}} p_{ref,cv} - p_{v,cv} \tag{1}$$

The vector of importance I_V is composed of the importance i_v of each variable (I_V is thus of size $1 \times n$). We also computed the mean importance $i_{v,obj}$ of each object over its 13 features.

Experimental protocol We performed two experiments on the 6 clusters (see section “Dataset”):

- *Experiment 1*: using the permutation process, we computed the mean importance of each object $i_{v,obj}$. For each of the 6 clusters, we extracted the object with the highest mean importance $i_{v,obj}$.
- *Experiment 2*: having identified the object with the highest mean importance for the 6 clusters, we restricted our input variables to the 13 features of this object and computed corresponding I_V using the permutation process.

Thus, for each of the 6 clusters, we extracted the feature with the highest importance i_v .

In this way, we extracted the object and its corresponding most important feature predicting each cluster’s output.

Translating the importance results into statements

Based on the analysis of the importance results, a scientist with extensive knowledge of the clinical context translated each importance result into a meaningful sentence, or *statement*. Each importance result referred to a cluster and a set of most important variables. Considering these constraints, the scientist translated each importance result by answering the question related to the cluster with the most important variables of the importance result using the “feature-object” semantic (see section “Dataset”).

Each statement satisfied the following constraints:

- Each importance result has a unique associated statement.
- A statement refers to one cluster.
- A statement refers to a set of input variables characterized by one or more features related to a single object.
- A statement refers either to the ESS output, or to the SPP output.

Illustrating statements with samples and associated video clips

We wished to illustrate each *statement* with two video clips, one for each output value. In these two situations, we automatically pre-selected m_{clip} samples and their related video clips. Then, among the m_{clip} clips, we manually selected the one best illustrating the situation, where the output was either 0 or 1.

A translated *statement* S refers to an importance result, its cluster and its most important variables. These constraints define some $(X_S; y_S)$ data for *statement* S , with a reduced number of input variables: 13 variables for Experiment 1 where the importance result related to the most important object and all its features, and 1 variable for Experiment 2 where we focused on the most important feature of the most important object. The y_S vector was either a subset of y_{ESS} or y_{SPP} (restricted to the samples of X_S) depending on the cluster studied.

In the pre-selection, we determined which samples were the most decisive for our model regarding a *statement* S by training our predictive model on the (X_S, y_S) data. After this training, the decision function in the *SVM* step of our model returned a decision value for each trained sample (i.e., a prediction probability). The n_{clip} highest decision values were automatically selected for the “1” and the “0” labels of y_S , allowing us to identify the most discriminant samples for *statement* S .

Once we had identified the m_{clip} most decisive samples, it was straightforward to obtain their associated timestamp from the video and extract representative video clips. A clip lasts a given duration t_{clip} and its beginning is shifted from the sample’s timestamp by t_{shift} . Finally, an operator considered each *statement* and its video clips. For each label of the output variable y_S , he selected one sample among the m_{clip} which best illustrated the clinical idea expressed by the *statement* S .

Experimental protocol We characterized the selection process by setting the parameters $m_{clip} = 5$, $t_{clip} = 10$ s and $t_{shift} = -7$ s. Based on these parameters, for each *statement*, 5 sample video clips were pre-selected for each output label. Each sample was illustrated with a 10 s video clip that started 7 s before the sample timestamp. The operator who selected one sample out of the 5 proposed was a non-clinician specialist. In total, two video clips were presented with each *statement*.

Design of the survey

A survey was designed with questions consisting in a statement associated with two video clips. Surgeons responded using a 5-level Likert scale of agreement: 0 = strongly dis-

agree, 1 = rather disagree, 2 = no opinion, 3 = rather agree, and 4 = strongly agree.

Experimental protocol The survey was produced with questions in a randomized order and submitted to 6 experienced digestive surgeons (1 PGY5 resident, and 5 experts).

Results

Extraction of the most important variables

The prediction performance of our model informs us on the level of confidence we can accord to the estimation of the importance vector I_V . Table 2 presents the prediction performance for the complete clusters C_{ESS} and C_{SPP} (lines 1 and 2) taken from [1], and for each sub-cluster defined in section “Dataset.” The number of input samples in each cluster is also provided.

Table 3 provides a summary of the results of Experiment 1, highlighting the most important objects identified for each cluster studied. For instance, the most important object when predicting the *ESS* in cluster C_{ESS} was identified as the spleen (id #1).

Table 4 provides a summary of the results from Experiment 2, highlighting the most important feature of the objects identified as most important in Experiment 1. For instance, id #3 indicates that the most important feature of the spleen to predict the *ESS* was the texture. Result id #12 was not translated into a statement as this feature had exactly the same importance as the 11 other features. More detailed results are presented in suppl. material #1.

Results of the survey submitted to experienced surgeons

Statements (see Table 5) and their related video clips were extracted for each of the most important objects and features to build a survey comprising 11 questions. Figure 4 illustrates two questions from the questionnaire, and the full questionnaire is provided as a Power Point file in the Supplementary Material #2 with the embedded videos.

We gathered responses from 6 surgeons expressed as a level of agreement on a Likert scale and present them here in summary form. The complete responses are provided in suppl. material #3.

Figure 5 presents the overall level of agreement of each surgeon with the statements. We observe that, apart from surgeon #1 who expressed a low overall level of agreement, the five other surgeons tended to agree with the statements in our survey.

Figure 6 shows a more detailed breakdown of the responses to each *statement* with a Likert chart. 7/11 answers show a positive agreement trend, while 4/11 answers show

Table 2 Prediction performances of the model for the 6 clusters—in each cluster, the performance of the binary classification was evaluated for each image sample

Data cluster	Sample count	Accuracy	Sensitivity	Specificity
C_{ESS}	735	0.64 ± 0.05	0.76 ± 0.15	0.56 ± 0.14
C_{SPP}	735	0.60 ± 0.08	0.58 ± 0.16	0.62 ± 0.13
C_{ESS}^A	405	0.66 ± 0.15	0.61 ± 0.30	0.70 ± 0.30
C_{ESS}^B	330	0.73 ± 0.05	0.58 ± 0.10	0.88 ± 0.07
C_{SPP}^A	156	0.99 ± 0.02	0.98 ± 0.04	1 ± 0
C_{SPP}^B	155	0.97 ± 0.03	0.98 ± 0.03	0.95 ± 0.04

Accuracy, sensitivity and specificity values indicated were averaged over all the samples in each cluster. Results are presented as mean \pm standard deviation

Table 3 Results of Experiment 1 with extraction of the most important objects for the prediction tasks performed with each of the 6 clusters (computed as $\overline{i_{v,obj}}$, the highest mean importance over features)

Id	Cluster	Most important object	$\log_{10}(\overline{i_{v,obj}})$
#1	C_{ESS}	Spleen	-1.84
#2	C_{SPP}	Stomach	-1.71
#5	C_{SPP}	Spleen	-1.71
#6	C_{ESS}^B	Flat grasper	-1.43
#7	C_{SPP}^A	Liver	-1.90
#8	C_{SPP}^B	Abdominal wall	-2.73
#9	C_{ESS}^A	Surface	-1.91

Table 4 Results of Experiment 2 with extraction of the most important feature for the most important object for the prediction task applied to each of the 6 clusters (computed as i_v , the highest importance)

Id	Cluster	Most important feature	Related object	$\log_{10}(i_v)$
#3	C_{ESS}	Texture	Spleen	-2.32
#4	C_{ESS}^A	Color_b	Stomach	-2.07
#10	C_{ESS}^B	Color_a	Flat grasper	-2.38
#11	C_{SPP}^A	Eigenvector_1_y	Liver	- Inf
-	C_{SPP}^B	Barycenter_x	Abdominal wall	-3.33

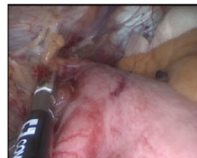
Table 5 Clinical statements based on the extraction results that were submitted to surgeons in the survey

Cluster	Id	Statement text
C_{ESS}	#1	To optimize the <i>ESS</i> , it is very important to correctly handle the spleen
C_{SPP}	#2	The difference in practice between the two surgeons involved in the study is mostly visible in their handling of the stomach
C_{ESS}	#3	Observation of the texture of the spleen is very helpful to assess the <i>ESS</i>
C_{SPP}	#4	Observation of the color of the stomach in the image is very helpful to recognize the <i>SPP</i>
C_{ESS}^A	#5	Here, to manage the <i>ESS</i> , the surgeon is very mindful of handling of the spleen
C_{ESS}^B	#6	Here, to manage the <i>ESS</i> , the surgeon is very mindful of manipulation of the flat grasper
C_{SPP}^A	#7	As the <i>ESS</i> is good, the <i>SPP</i> of the two surgeons are distinguishable mostly in their handling of the liver
C_{SPP}^B	#8	As the <i>ESS</i> is non-optimal, the <i>SPP</i> of the two surgeons are distinguishable mostly in their handling of the abdominal wall
C_{ESS}^A	#9	Here, to manage the <i>ESS</i> , the surgeon is very mindful of the visible surface of the spleen
C_{ESS}^B	#10	Here, to manage the <i>ESS</i> , the surgeon is very mindful of the color of the flat grasper present in the image
C_{SPP}^A	#11	As the <i>ESS</i> is good, the <i>SPP</i> of the two surgeons are distinguishable mostly in the shape and spreading of the liver in the image

Fig. 4 Two examples of statements with their two associated video clips, as submitted in the survey

Statement #2 / 11

The difference in practice between the two surgeons involved in the study is mostly visible in their handling of the stomach.

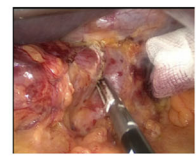
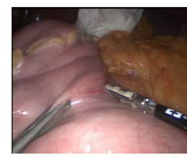


To complete:
You select an answer

Next slide

Statement #11 / 11

As the ESS is good, the SPP of the two surgeons are distinguishable mostly in the shape and spreading of the liver in the image.



To complete:
You select an answer

Next slide

Fig. 5 Overall agreement with the statements for each surgeon

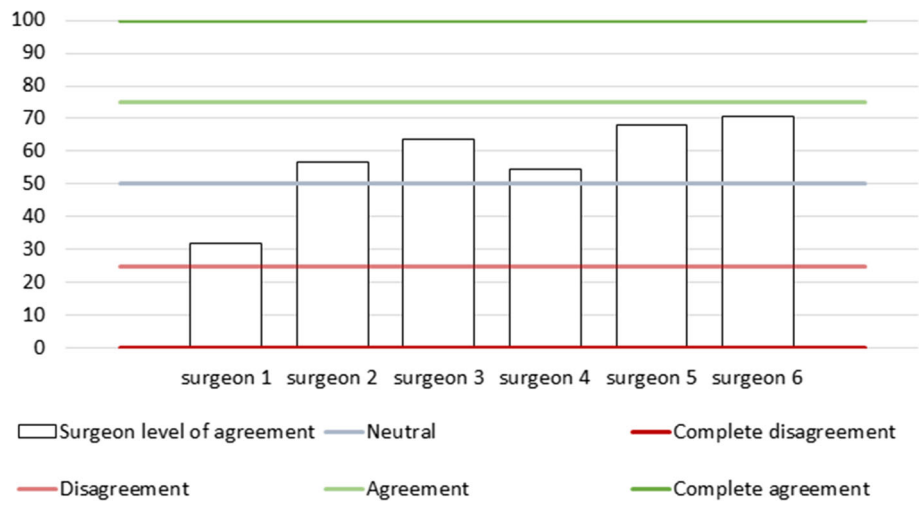
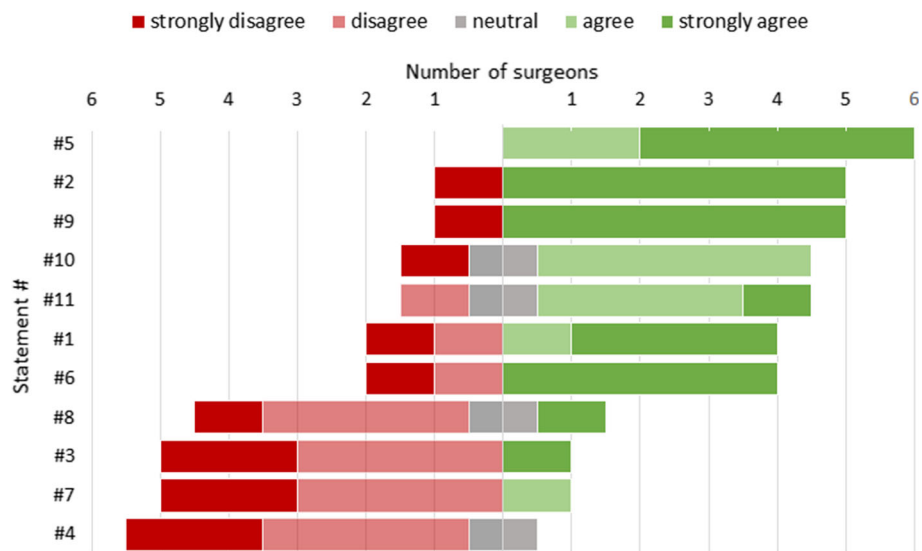


Fig. 6 Likert chart representing the level of agreement of surgeons, centered on the neutral answer, and ordered by agreement level



a trend for disagreement. For each statement, at most one neutral opinion was recorded.

As surgeons answered the survey and immediately after administering it, we collected their views on the clinical implications and the design of the survey. Here are some of their open comments:

1. Three of the 6 surgeons observed at least once during the survey that the video clips did not correctly illustrate their related *statement*. These observations were made for 5 different *statements*, not for a single one.
2. Three surgeons discussed the duration of the video clips, two out of three considered that they were too short.
3. In response to *statement* #3, four surgeons considered the texture of the spleen to be a tactile notion, whereas it represented a visual notion for one surgeon.

Discussion

In this study, we wished to investigate the explicability of a predictive model based on laparoscopic video data and to compare its interpretation with surgeons' expertise. AI applications are generally evaluated based on a set of metrics quantifying their performance when completing a specific task, but this type of assessment might miss significant clinical concerns. By presenting our model to clinicians, we were able to strengthen its clinical relevance.

A similar XAI method for the prediction of hypoxemia during surgeries based on temporal hemodynamic data, ventilation data and patient history data was proposed in [7]. In their study, four anesthetists were asked to rank the same variables as their model, and the two rankings were compared. Thanks to an established literature in the field of anesthesia, a consensus already exists on the major risk factors to monitor, and the results of the comparison mainly helped to validate the predictive model.

A majority of studies relating to surgical practice work with practice indicators. Only a few recent studies have exploited laparoscopic videos with the aim of better understanding clinical practice: [6] proposed to automatically detect intra-operative adverse events based on the annotation of surgical activities and hidden semi-Markov models, but the image content was not exploited. [11] used convolutional neural networks (CNNs) to automatically assess the critical view of safety in laparoscopic cholecystectomy, but the use of a CNN makes it challenging to interpret the network's predictions.

Here, we studied two more specific indicators of surgical practice: the quality of handling of exposure of the surgical target by the surgeon (*ESS*) and the profile of surgical practice (*SPP*), and proposed a method to link the choices made by a machine learning model to clinical expertise. Due to the lack

of literature on these two notions, the comparative results provided by our survey raise questions on the interpretations of our model, as well as the implicit, and non-formalized knowledge of the surgeon relative to these phenomena. In the future, we hope that XAI will become a tool to validate AI models but also to formalize complex and often implicit knowledge, with potential applications in surgical training and surgical quality control.

Clusters and prediction

Our dataset (X, y_{ESS}, y_{SPP}) led us to build 6 clusters, each of which was described by a meaningful question. The "feature-object" semantics describing the input variables were used to define Experiments 1 and 2. Overall, this rich conceptual context allowed us to propose a substantial and meaningful analysis of this dataset with 12 distinct scenarios (6 clusters \times 2 Experiments).

The prediction results for clusters C_{ESS} ($Acc = 0.64$) and C_{SPP} ($Acc = 0.6$) described in Table 2 are on par with the literature. Results for clusters C_{ESS}^A and C_{SPP}^B have similar mean accuracies but higher standard deviations, which might be imputed to the smaller numbers of samples.

The prediction performance of our model on clusters C_{SPP}^A and C_{SPP}^B may be questionable. These two clusters are very small (150 samples as against the 735 samples in the complete dataset), and almost perfect prediction performances were obtained. This type of result is typical of overfitting, and the interpretations presented for these two clusters should be considered with care.

We assumed that our model's predictive performances were acceptable and sufficient to build our method (importance vector computation, automatic selection of video clips). However, these performances are not optimal and may affect the resulting variable importance vectors and pre-selected video clips.

The variable of importance I_V

To select the most important objects and variables, we chose to look at the mean prediction differences with value permutation (Eq. 1). Using this approach, variable importance is easily interpretable as it represents how much impact a variable has on the predictive task along the CV splits.

However, examination of the detailed importance results (suppl. material #1) reveals relatively similar i_v values, and selecting a set of most important objects/features rather than a single one may give more realistic results. To complement our method, we could also evaluate the stability of the variables importance by including the standard deviation of the prediction shift.

Translating the importance results into clinical statements

Interpreting the variable importance i_v and translating it into a clinically meaningful *statement* is the key step in our methodology. We translated video-based variables to texts referring to the laparoscopic practice by placing ourselves at the interface between computer science and clinical practice. To our knowledge, this is a first attempt in the domain.

We chose to perform this translation manually and we avoided the human bias of the operator by defining sufficient constraints on the translation process:

1. The definition of the clusters, their related questions and the “feature-object” semantics provided a framework in which the translation gained clarity.
2. The operator had to answer the cluster’s question with an importance result: the most important object (Experiment 2) or its most important feature (Experiment 2). As discussed in section “The variable of importance I_V ,” other object/feature selection approaches could have been considered, but might have made the interpretation/translation process more complex.
3. Constraints were defined directly for the translations in section “Translating the importance results into statements.”

Despite these constraints, we faced challenges related to the “feature-object” semantics which required different translation approaches: `barycenter_x` and `barycenter_y` have intricate meanings and would be difficult to understand separately when characterizing the position of the object. Surgeons gave us contradictory feedback on the texture feature: two surgeons observed that it is a tactile sensation, whereas another considered it to be visually interpretable. Similarly, we processed instruments and organs as equivalents despite their clear clinical distinctions. Thus, depending on the context (computer science or medical), some concepts take on a very different meaning and should be considered with extreme caution.

To reduce the human bias, several experts should collaborate to translate *statements*. To avoid the human bias, completely automated translation could be used, but this is an open, complex question. Automated translation implies a Natural Language Processing approach with input data composed of labels (variable name) and values (importance value) to generate text content at the interface between computing and laparoscopy expertise.

Illustrating statements with samples and associated video clips

To complement the *statements*, we selected video clips from the initial video dataset. Surgeons appreciated this mixture of text and image as it encouraged them to investigate very specific aspects of their practice.

We performed an automated pre-selection of the most illustrative samples to remain as objective as possible. For each *statement*, we reduced the variable count (13 variables for Experiment 1 and 1 variable for Experiment 2). Despite the high importance of these variables, our model was designed and optimized for a matrix with $n = 130$ variables, and not with 13 or less. Thus, the automated selection of images requires external validation. Future studies will include a focus on the impact of the variable count on the model’s performance.

To validate the automated pre-selection, we added a manual selection step which introduced a human bias. Choosing the best clip among those automatically pre-selected was non-trivial, and when the survey was presented to surgeons, three of them observed at least once that the video clips did not correctly illustrate their associated *statement*. *Statement* #10 focused on the stomach, but one of its clips failed to show the stomach correctly. The two clips for *statement* #6 showed different, incomparable, surgical steps. This gap between the *statement* and its illustration shows the complexity of selecting a satisfactory clip.

We defined $t_{clip} = 10$ s as the clip duration with regard to the mean period between two successive annotations of *ESS* ($19.1 \text{ s} \pm 21.9 \text{ s}$). One surgeon confirmed that this duration suited him/her, but two others found that the clips were too short to correctly analyze the surgical practice. We will need to investigate the clip duration in future studies.

The survey submitted to experienced surgeons

We proposed a very novel clinically-oriented validation methodology with the presentation of a survey to 6 surgeons: 4 expert surgeons and 1 PGY5 resident came from one hospital, the last expert worked at another hospital. This proof-of-concept study was limited to assessing the relevance of our XAI approach and the distribution of the levels of agreement of the surgeons surveyed reveals their interest in our approach. We aim to include more surgeons from different hospitals in a larger-scale study in the future.

Out of the 11 *statements* presented, surgeons indicated positive agreement for 7. This positive agreement trend throughout our survey indicates we have validated our proof-of-concept study. Through this survey, 5/6 surgeons agreed with the interpretations of our predictive model and con-

firmed the overall clinical relevance of 7/11 *statements*. Only the combined responses for 1 of the 6 surgeons revealed an overall disagreement with our survey (7/11 disagreements). In section “Clusters and prediction” we already mentioned the potentially overfitted prediction result due to the small number of samples in clusters C_{SPP}^A and C_{SPP}^B . Among the 4 *statements* with which surgeons disagreed, 3 were related to these two clusters. This result may confirm the weakness of the model’s prediction for these two clusters.

In section “Illustrating statements with samples and associated video clips,” we discussed the lack of complete concordance between some *statements* and their video clips. When faced with this situation, surgeons were tempted to assess the ability of the video clips to correctly illustrate their *statement* rather than assessing the relevance of the *statement* itself. We asked them to keep the initial goal of the survey in mind, but we should have anticipated this ambiguity. In the future, we will work with sociologists and/or linguists to build a systematic survey methodology correctly answering questions related to the presence/absence or the appropriate duration of the clips.

Conclusion

In this paper, we propose a proof-of-concept method to validate the clinical relevance of a model predicting two indicators of the quality of surgical practice. This validation was based on a survey administered to 6 surgeons that took the form of statements illustrated by short videos. To our knowledge, no previous study proposed similar clinical interpretations, or validated such an approach with a survey presented to surgeons. Overall, we developed a novel approach with a high potential for the analysis and understanding of surgical practice.

If we could entirely automate our approach, it would allow us to automatically provide the surgeon with a video associated with relevant subtitles. Such an algorithmic approach would make it possible to produce some clinical applications for surgical training, post-training review, or postoperative feedback.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11548-021-02422-0>.

Acknowledgements This work was supported by funding from the French government managed by the ANR as part of the Investissements d’Avenir Programme (Labex CAMI) under reference ANR-11-LABX-0004. This work was partially supported by MIAI @ Grenoble Alpes, (ANR-19-P3IA-0003). The authors thank the IRT b <> com for providing the “Surgery Workflow Toolbox [annotate],” software used in this study.

Declarations

Conflict of interest The authors declare that they have no conflict of interest to declare.

Ethical approval All procedures performed in the study presented involving human participants were carried out in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Declaration of Helsinki and its later amendments, or comparable ethical standards.

Informed consent The study protocol was approved by the institutional ethics committee on May 24, 2018 (CECIC Rhône-Alpes-Auvergne, Clermont-Ferrand, IRB 5891).

References

1. Derathé A, Reche F, Moreau-Gaudry A, Jannin P, Gibaud B, Voros S (2019) Predicting the quality of surgical exposure using spatial and procedural features from laparoscopic videos. *Int J Comput Assist Radiol Surg*. <https://doi.org/10.1007/s11548-019-02072-3>
2. El Ahmadieh TY, Harrop J, Batjer HH, Resnick DK, Bendok BR (2014) Blinded peer assessment of surgical skill is feasible and can predict complication rates: a step toward measuring surgical quality. *Neurosurgery* 74(6):N12–14. <https://doi.org/10.1227/01.neu.0000450232.06740.ef>
3. Foster JD, Miskovic D, Allison AS, Conti JA, Ockrim J, Cooper EJ, Hanna GB, Francis NK (2016) Application of objective clinical human reliability analysis (OCHRA) in assessment of technical performance in laparoscopic rectal cancer surgery. *Tech Coloproctol* 20(6):361–367. <https://doi.org/10.1007/s10151-016-1444-4>
4. Gagner M, Hutchinson C, Rosenthal R (2016) Fifth international consensus conference: current status of sleeve gastrectomy. *Surg Obes Relat Dis* 12(4):750–756. <https://doi.org/10.1016/j.soard.2016.01.022>
5. Gordon L, Grantcharov T, Rudzicz F (2019) Explainable artificial intelligence for safe intraoperative decision support. *JAMA Surg* 154(11):1064–1065. <https://doi.org/10.1001/jamasurg.2019.2821>
6. Huaulmé A, Jannin P, Reche F, Faucheron JL, Moreau-Gaudry A, Voros S (2020) Offline identification of surgical deviations in laparoscopic rectopexy. *Artif Intell Med* 104:101837. <https://doi.org/10.1016/j.artmed.2020.101837>
7. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, Liston DE, Low DKW, Newman SF, Kim J, Lee SI (2018) Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng* 2(10):749–760. <https://doi.org/10.1038/s41551-018-0304-0>
8. Maier-Hein L, Vedula SS, Speidel S, Navab N, Kikinis R, Park A, Eisenmann M, Feussner H, Forestier G, Giannarou S, Hashizume M, Katic D, Kennigott H, Kranzfelder M, Malpani A, März K, Neumuth T, Padoy N, Pugh C, Schoch N, Stoyanov D, Taylor R, Wagner M, Hager GD, Jannin P (2017) Surgical data science for next-generation interventions. *Nat Biomed Eng* 1(9):691–696. <https://doi.org/10.1038/s41551-017-0132-7>
9. Malpani A, Vedula SS, Chen CCG, Hager GD (2015) A study of crowdsourced segment-level surgical skill assessment using pairwise rankings. *Int J Comput Assist Radiol Surg* 10(9):1435–1447. <https://doi.org/10.1007/s11548-015-1238-6>
10. Manganò A, Kim HY, Wu CW, Rausei S, Hui S, Xiaoli L, Chiang FY, Roukos DH, Lianos GD, Volpi E, Dionigi G (2016) Continuous intraoperative neuromonitoring in thyroid surgery: safety analysis of 400 consecutive electrode probe placements with standardized

- procedures. *Head Neck* 38(S1):E1568–E1574. <https://doi.org/10.1002/hed.24280>
11. Mascagni P, Vardazaryan A, Alapatt D, Urade T, Emre T, Fiorillo C, Pessaux P, Mutter D, Marescaux J, Costamagna G, Dallemagne B, Padoy N (2020) Artificial intelligence for surgical safety: automatic assessment of the critical view of safety in laparoscopic cholecystectomy using deep learning. *Ann Surg*. <https://doi.org/10.1097/SLA.0000000000004351>
 12. Pernek I, Ferscha A (2017) A survey of context recognition in surgery. *Med Biol Eng Comput* 55(10):1719–1734. <https://doi.org/10.1007/s11517-017-1670-6>
 13. Radivojac P, Obradovic Z, Dunker AK, Vucetic S (2004) Feature selection filters based on the permutation test. In: Boulicaut JF, Esposito F, Giannotti F, Pedreschi D (eds) *Machine learning: ECML 2004. Lecture notes in computer science*. Springer, Berlin, pp 334–346. <https://doi.org/10.1007/978-3-540-30115-8-32>
 14. Rätsch G, Sonnenburg S, Schäfer C (2006) Learning interpretable SVMs for biological sequence classification. *BMC Bioinform* 7(1):S9. <https://doi.org/10.1186/1471-2105-7-S1-S9>
 15. The Colon cancer Laparoscopic or Open Resection Study Group (2005) Laparoscopic surgery versus open surgery for colon cancer: short-term outcomes of a randomised trial. *Lancet Oncol* 6(7):477–484. [https://doi.org/10.1016/S1470-2045\(05\)70221-7](https://doi.org/10.1016/S1470-2045(05)70221-7)
 16. Thiele RH, Rea KM, Turrentine FE, Friel CM, Hassinger TE, Goudreau BJ, Umaphathi BA, Kron IL, Sawyer RG, Hedrick TL, McMurry TL (2015) Standardization of care: impact of an enhanced recovery protocol on length of stay, complications, and direct costs after colorectal surgery. *J Am Coll Surg* 220(4):430–443. <https://doi.org/10.1016/j.jamcollsurg.2014.12.042>
 17. Zhang W, Wang L, Wang Z (2019) Interpretable machine learning study of the many-body localization transition in disordered quantum Ising spin chains. *Phys Rev B* 99(5):054208. <https://doi.org/10.1103/PhysRevB.99.054208>

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