Deep neural network architecture for automated soft surgical skills evaluation using objective structured assessment of technical skills criteria

Malik Benmansour, Abed Malti, Pierre Jannin

Deep Neural Network Architecture for Automated Soft Surgical Skills Evaluation Using Objective Structured Assessment of Technical Skills Criteria

Malik Benmansour[1*], Abed Malti[2] and Pierre Jannin[3]

[1*]G.E.E., Université de Tlemcen, LAT, 13000, Tlemcen, Algeria.
[2]G.B.M., Université de Tlemcen, Laboratoire de Génie Biomédical, Tlemcen, 13000, Algeria.
[3]Université de Rennes, INSERM, LTSI-UMR, Rennes, F-35000, France.

*Corresponding author(s). E-mail(s): malik.benmansour.tlemcen@gmail.com;
Contributing authors: abed.malti@gmail.com; pierre.jannin@univ-rennes1.fr;

Abstract

Purpose: Classic methods of surgery skills evaluation tend to classify the surgeon performance in multi-categorical discrete classes. If this classification scheme has proven to be effective, it does not provide in-between evaluation levels. If these intermediate scoring levels were available, they would provide more accurate evaluation of the surgeon trainee.

Methods: We propose a novel approach to assess surgery skills on a continuous scale ranging from 1 to 5. We show that the proposed approach is flexible enough to be used either for scores of global performance or several sub-scores based on a surgical criteria set called Objective Structured Assessment of Technical Skills (OSATS). We established a combined CNN+BiLSTM architecture to take advantage of both temporal and spatial features of kinematic data. Our experimental validation relies on real world data obtained from JIGSAWS database. The surgeons are evaluated on three tasks: Knot-Tying, Needle-Passing and Suturing. The proposed framework of neural networks takes as inputs a

sequence of 76 kinematic variables and produce an output float score ranging from 1 to 5, reflecting the quality of the performed surgical task.
Results: Our proposed model achieves high quality OSATS scores predictions with means of Spearman correlation coefficients between the predicted outputs and the ground-truth outputs of 0.82, 0.60 and 0.65 for Knot-Tying, Needle-Passing and Suturing, respectively. To our knowledge, we are the first to achieve this regression performance using the OSATS criteria and the JIGSAWS kinematic data.
Conclusion: An effective deep learning tool was created for the purpose of surgical skills assessment. It was shown that our method could be a promising surgical skills evaluation tool for surgical training programs.

Keywords: Surgical skills assessment, surgical robotics, deep learning, convolutional neural networks, recurrent neural networks, kinematic data

Statements and declarations

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

Ethical approval. For this type of study, formal consent is not required.

Informed consent. This article does not contain patient data.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

1 Introduction

The lack of safety during surgical operations is yielding continually to an increasing need of important improvements in the surgery training. Indeed, post-surgical complications may occur if a surgeon delivers poor technical skills during an operation, including death [1, 2]. Surgical technical errors are the most common reason for post-surgical complications including re-operation and re-admission [3, 4]. Consequently, inventing new approaches to improve surgeons skill can positively affect the safety of the patient. The problem of poor or average acquired surgical skills may come from the process followed by trainees so far to learn to operate: it consists on replicating operations done by expert surgeons on cadavers or ex-vivo organs. The trainee first watches an expert surgeon performing a surgical task, then she/he tries her/his best to re-do the operation under expert supervision [5]. At the end, the expert evaluates the trainee based on a set of surgical criteria. Apart from the fact that these surgical assessment methods are time and energy consuming for senior

surgeons, they are also characterized by subjectivity and a lack of accuracy [6]. To enhance the surgical training, several High-Tech platforms have emerged like surgical simulators based on virtual reality [7, 8] or surgical systems like the DaVinci robot [9, 8]. Also, many assessment methods based on artificial intelligence and especially on artificial neural networks were proposed to train surgeons and increase their performance without any human intervention [10]. These methods turned out to be very effective in classifying performance levels into discrete categories (for example: expert, medium and novice) [11]. However, they do not deliver a strong feedback to the trainee since they classify performance into a category once the result score reaches a certain threshold without taking into account intermediate scores. Consequently, the feedback suffers from a lack of accuracy and the trainee will not have enough information about her/his expertise level. In this work, we propose a deep learning architecture to provide an automatic and objective surgical skill assessment on a continuous scale ranging from 1 to 5 based on OSATS criteria [12, 13]. This continuous scale has been approximated from the OSATS scale which is originally a Likert scale and thus, composed of ordinal items. Such approximation is possible if intervals between each level in the Likert scale item can be presumed equal [14]. Moreover, a group of researchers maintains that Likert, or ordinal variables with five or more categories can often be used as continuous without any harm to the analysis the researcher plans to use them in [15, 16, 17]. Therefore, according to these researches and since every OSATS Likert item is composed of 5 assessment levels, an approximation with a continuous function is possible and it can be included in a regression process. The main purpose of proposing a continuous scale for the assessment of technical skill is to provide a more precise feedback to the trainee by allowing her/him to have a detailed information about her/his performance without restricting it into a category. Indeed, categorized performance does not show the difference between, for example, two novices when one is necessarily better than the other. However, a continuous scale allows differentiation of two or more surgeons with the same expertise level. We established several independent neural networks to evaluate surgeons on three tasks independently. This work is an extension to our previous contributions [18, 19, 20] where we presented approaches that allows a global evaluation based on a single global performance score using neural networks. We present briefly the contents of the JIGSAWS database which served as a validation ground for our experiments and also the general approach of these previous works in a separate document (Online Resource 1, Section 1 and Section 2, respectively). Briefly, it is a surgical database composed of real world kinematic and video data from the operating room obtained on the DaVinci robotic platform. After that, we present here the main contribution of this work that relies on a more rigorous assessment method: it consists on assessing automatically and objectively surgeons on the OSATS criteria. These criteria have been used previously to assess skills on many surgical operations like fetal blood sampling, manual removal of placenta and opening and closing the abdomen.A brief description of the

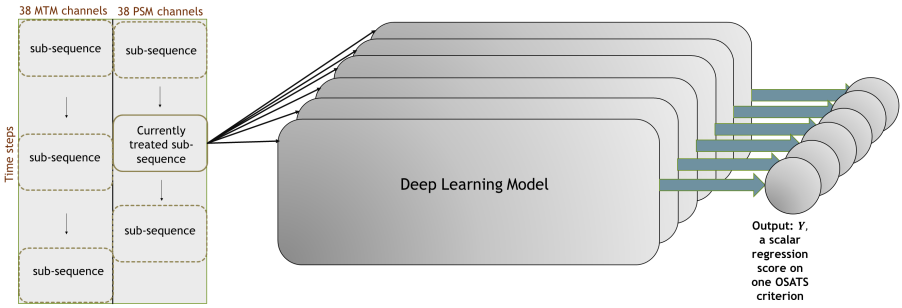

Fig. 1 A drawing of the general adopted approach for automatic surgical skill evaluation on OSATS criteria for one surgical task: each of the 6 deep learning models is responsible for assessing the surgeon trial in one criterion

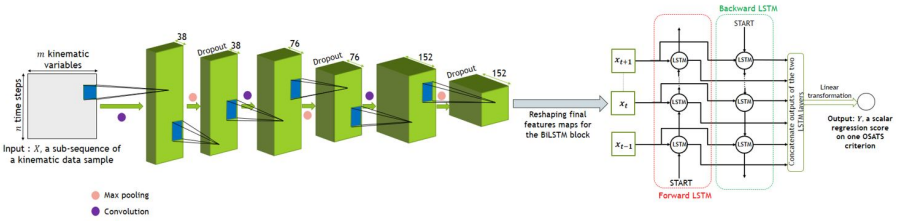

Fig. 2 Architecture of the proposed CNN+BiLSTM network: The network takes as input a crop of a kinematic data and yield a scalar output carrying the score on one OSATS criterion (Best view with colors).

OSATS criteria is provided in the additional document (Online Resource 1, Section 4) as well.

2 Related work based on regression

During the last decade, machine learning has been widely employed in the surgical skill assessment and surgical gesture recognition fields since the arrival of powerful computers and machine learning programming libraries. Multiple works have emerged for using various methods like Hidden Markov Models (HMM), Support Vector Machines (SVM) and Artificial Neural Networks (ANN). The studies and results of several works revealed that the ANN is the most efficient machine learning method till now. Especially Deep Neural Networks (DNN) like Deep Feed-forward Neural Networks, Recurrent Neural Networks (RNN), Long-Short Term Memory (LSTM) and Convolutional Neural Networks (CNN). We find in the literature some works based on skill classification into discrete categories and regression methods on soft continuous scales. We present, in the current section, some of the previous works done in the field of robot assisted surgery that adopted the regression

approach only. Related work on classification methods is provided in the additional document (Online Resource 1, Section 3).

In [21], a BiLSTM autoencoder has been proposed to assess surgical video clips by categorizing them based on the surgical step being performed and the level of the surgeon's competence . The authors worked on regression using OSATS scores [13]. They used a BiLSTM with an attention mechanism relying on video data. In this paper, we propose a method that relies only on kinematic data. In [22] is presented a contribution that compares several machine learning algorithms (decision forest, neural networks, boosted decision tree) for OSATS scores predictions on continuous scales to detect expert level in laparoscopic suturing and knot-tying. While most of the prior works focused on surgical skill assessment using simulated datasets, this work [23] proposed to use a real clinical dataset which consists of in-vivo laparoscopic surgeries. An objective and automated framework based on a Multi Layer Perceptron is proposed to predict surgical skills using a regression method achieving a Spearman's correlation of 0.55 with the ground truth of overall technical skill. In [24], three machine learning algorithms involving SVMs have been trained in order to evaluate surgical performance and predict clinical outcomes of patients who have had a robot-assisted radical prostatectomy. Finally, in [25] is proposed a 1D CNN to assess surgeons directly from the kinematic data of the JIGSAWS dataset. The framework is composed of a Fully Convolutional Network (FCN) and a Class Activation Map component that highlights the fractions of the surgical kinematic trial that contributed highly to classify the surgeon as novice, intermediate or expert. In addition, the authors proposed a regression model to predict OSATS scores [12, 13] on a continuous scale. The authors did not apply a data augmentation pipeline. Thus, their model is trained with raw kinematic data with variable length. In our proposed method, we apply a data augmentation strategy to overcome the limited number of training data in the JIGSAWS dataset and our model contains a 2D CNN that capture spatial features. We compare the results of this paper with ours in the results section.
Regarding our previous works based on regression, we used several types of neural networks and compared their results. In [18], we used a DNN architecture to assess all surgeons of the JIGSAWS dataset on the three tasks separately, using floating performance scores as a ground truth. Despite the good result obtained with this approach, we knew that it can be greatly improved by changing the type of the neural network. Since the kinematic data that we are dealing with represent dynamic sequences of kinematic variables that depend on time (time series), then we wanted to incorporate the notion of time in our approach. A suitable type of neural network that seemed to be effective for such type of data is the recurrent neural network (RNN). Therefore, in [20], we proposed to use a RNN architecture to provide a dynamic evaluation of the performed surgery tasks. RNN are well known for their capability to allow feedback connections between states at different time

steps. Indeed, the results obtained using the same amount of data samples for the training and the test steps as in [18] are more plausible and match more accurately our expectations. However, RNN are known to be vulnerable to two major problems possibly encountered during the training step: the vanishing gradient and the exploding gradient [26]. Long term dependencies in time series data may cause these problems. Besides, a kinematic data of a surgical task is composed of gestures as it can be checked in [27] and these gestures are largely spaced over time in a data sample. So, with the lack of memory, a RNN is not capable to learn correlations between largely time-spaced events. Consequently, we used in our third work [19] a LSTM Network architecture which represents an evolution of a traditional RNN and it is suited to overcome the problems encountered with the RNNs. The LSTM, indeed, have feedback connections and in addition, a notion of memory allowing it to learn long term dependencies. These backward feedback allow the LSTM to take into account length-customizable previous entries. This specificity gives LSTM advantages over the RNN and in the context of surgery tasks, this enable us to take into account the evaluation of current gesture together with former executed gesture. The results obtained using this architecture are a bit different from those obtained with classic RNN but are more logical since the whole surgical task is considered, as explained previously.

As an extension of our previous works, we propose in the present paper, a CNN+BiLSTM architecture considering a set of criteria and not just a unique global performance score. We believe that this approach delivers a more accurate and more complete assessment for the surgery trainees.

3 Surgical skill evaluation method using OSATS criteria

3.1 Proposed method and problem formulation

We consider the assessment of surgical skills as a supervised regression problem, where the input is multivariate time series (MTS) of motion kinematics measured from the DaVinci surgical robot end-effectors, X, and the output is the predicted OSATS score representing the quality of the gestures regarding one of the six OSATS criteria. Specifically, The predicted OSATS score is a scalar between 1 and 5 reflecting the performance of a subject S performing a surgical task T on one OSATS criterion C. The true OSATS outputs (ground-truth) are acquired from the metadata files of each task in the JIGSAWS dataset folder. These ground-truth scores have been attributed to trainees by a gynecologic surgeon with extensive robotic and laparoscopic surgical experience who watched each video of each trainee and each task. The objective cost function for training the network is defined as a Mean Squared Error loss function.

3.2 Architecture of the proposed model

The proposed model is composed of a CNN block followed by a BiLSTM block.

The CNN block: The CNN component architecture is highly inspired from the CNN in [28]. It contains three convolution stages with 38, 76 and 152 filters, respectively. All filters are size (2,2) and move across the input with a stride of 1. A 2D Max pooling operation is performed after each convolution operation with a pool size of (2,2) and a stride of 2. To enhance the model generalization over the input training data and to prevent overfitting, we added Gaussian noise at the input layer and after each convolution operation. A 20% Dropout regularization is also included after each max pooling stage. In addition, we employed a batch normalization process that standardizes the inputs to a layer for each mini-batch. This normalization helps stabilizing the learning process and reducing the time required to train the network efficiently. The ReLu activation function has been used in each convolution layer. At the end of the last convolution stage, the extracted features are reshaped in a way they can fit in the next temporal component, which is the BiLSTM block.

The BiLSTM block: The BiLSTM component contains two LSTM layers with 16 units each: the first one reads the input in a forward way regarding the time steps of the kinematic data sample, while the second one reads the inputs in a backward way. This structure allows the networks to capture both past and future information about the sequence at every time step. Next, the outputs of both LSTM layers are concatenated and transmitted to a single output node that carries the predicted OSATS score. The activation function of the bidirectional LSTM layer is a hyperbolic tangent and the activation function of the dense output node is a linear function. Batch normalization process is included in this block as well.

The main reason of choosing a mixed CNN+BiLSTM architecture is to take benefit from both spatial and temporal features extracted by both components from the kinematic data. The CNN component is able to capture spatial features while the BiLSTM can capture temporal information widely spaced in time. Moreover, choosing a BiLSTM over a LSTM relies on the ability of the BiLSTM in reading time series data in both forward and backward ways and thus doubling information provided to the network.

3.3 Data augmentation method

Each time-series has a typical length of between 1 and 5 minutes. With 8 surgeons and 5 trials per surgeon per procedure, the JIGSAWS dataset contains only 40 independent samples for training each surgical task's model. To overcome the training limitation, we follow a two-step data augmentation procedure used in this paper [28]. First, for each row sample in the kinematic data, we separate the MTM and PSM channels into two instances and cast them

as distinct samples. Instead of 76 sensors per row sample, for example, there will be 38 sensors for every two row samples. This "splitting and doubling" step is allowable because the MTM and PSM sensors are uncorrelated due to differences in position of the robot control arms. Each channel of raw sensor data is z-normalized to minimize the differences in scaling ranges of each sensor. Then, to further boost the training data, we withdraw a large volume of cropped sequences using a sliding window algorithm. It works by capturing observations of a fixed length (window size) from the sensor data and shifting that window by steps (step size) across the series to extract sub-sequences. In addition, we include some noise layers in the networks architectures to prevent overfitting.

4 Results and discussion

The experimental setup which provided the following results is described in the additional file (Online Resource 1, Section 5). It covers descriptions of the different validation schemes, details about the neural networks implementation and information about the model evaluation metrics (Spearman correlation coefficient, the median, the standard deviation and the mean).

4.1 Random Split validation scheme

For this first validation scheme, we tested and compared several neural network architectures (see Table 1). All the models of each architecture have the same purpose: predicting each of the six OSATS score for a surgeon trial, independently. Hence, we created 6 independent models based on each architecture and for each surgical task. Table 1 shows the means of the computed Spearman coefficients (ρ) over the six OSATS scores for the three surgical tasks
while adopting the Random Data Splitting validation scheme. The high correlations between the ground-truth labels and the predicted labels prove the similarity between the two distributions leading to the conclusion that all the networks are capable of giving an adequate grade deserved by tested surgeon in a 2-seconds window (We have chosen a sliding window of size 60 and a step size of 30 for this validation scheme). Specifically, CNN, LSTM, BiLSTM and the combined architectures CNN+LSTM and CNN+BiLSTM delivered the highest rank correlations. The DNN delivered lower ρ correlations because of its simple architecture and its weakness in capturing spatial and tempo-ral features of the kinematic data. On the other hand, CNN has the ability to capture spatial features while the LSTM can capture temporal information widely spaced in time without any information loss while processing the kine-matic data sample. BiLSTM benefits from its ability to read time series data in a forward direction, and in a backwards direction as well. BiLSTMs effec-tively increase the amount of information available to the network, improving the context available to the algorithm. Finally, the combined architectures take benefits from the LSTM, BiLSTM and CNN networks by capturing both spatial and temporal information.

Table 1 Means of Spearman's correlation coefficients calculated over the six OSATS scores between predicted and real values obtained with each network architecture (Random split validation scheme).

	Knot-Tying	Needle-Passing	Suturing
DNN	0.82	0.76	0.76
CNN	0.98	0.97	0.97
LSTM	0.95	0.93	0.94
BiLSTM	0.95	0.93	0.95
CNN+LSTM	0.95	0.93	0.94
CNN+BiLSTM	0.95	0.93	0.94

4.2 LOSO validation scheme

Due to GPU resource limitation, we tested only the CNN+BiLSTM architecture intended for this contribution. Table 2 shows the Spearman correlation coefficient obtained for each of the 6 models on each OSATS criterion and for each surgical task. Table 4 shows comparison between the ρ coefficients obtained by our model and the models of two others contributions. By adopting the same validation methodology (*LOSO*) proposed by [29] and [25], we are able to compare our proposed CNN+BiLSTM regression model to their best performing regression methods. Our regression model outperforms the FCN regression model [25] and the ApEn [29] regression method in the three surgical tasks. In other words, the prediction and the ground truth OSATS scores are more correlated when using CNN+BiLSTM than the ApEn-based and FCN-based solutions on the three surgical tasks. We point out the differences of our proposed method with [25]: The authors of this contribution did not apply any data augmentation strategy and they have trained their Fully Convolutional Network with raw kinematic data, while we have trained our models with crops of kinematic data obtained after applying the data augmentation method described in Sub-section 3.4. Also, they used one-dimensional convolution layers while we used two-dimensional convolution layers in our CNN component. the latter difference seems to explain why we have got better results: applying 2D filters allows the CNN to capture temporal information about the input sequence along its temporal axis but also to capture spatial information along the kinematic variables axis, which enables the CNN to choose the most significant features and capture correlation features between physical quantities (positions, velocities and accelerations). In addition, the BiLSTM block of our model allows the model to have a temporal understanding of the extracted features and to process them in two opposite directions. Finally, the authors designed a single FCN model to predict a vector of the six OSATS scores while we desgined six independent CNN+BiLSTM models for each OSATS score. Concerning the work [29], the authors adopted the approximate entropy (ApEn) algorithm to extract features from each surgical trial which are later fed to a nearest neighbor classifier. According to the authors, the features they use try to differentiate between different skill levels

using data repeatability. To justify the low performance on the Needle Passing task, they have claimed that the Needle Passing task is a less repetitive task as compared to the Knot-Tying and the Suturing tasks. According to us, their method that consists in evaluating holistic features for predicting skill level is outdated since the arrival of powerful deep learning libraries like Keras [30], that allows data scientists to easily build, train and test neural networks, knowing that neural networks are the most performing methods in the machine learning and deep learning fields.

In addition, we computed the mean squared error (MSE) between true and predicted OSATS scores for each criterion and reported the results in Table 3. This metric measures the average of the squares of the errors. What this means, is that it returns the average of the sums of the square of each difference between the estimated value and the true value. The mean squared error is always 0 or positive. When a MSE is small, this is an indication that the model accurately predicts the outputs. An important piece to note is that the MSE is sensitive to outliers. This is because it calculates the average of every data point's error. Because of this, a larger error on outliers will amplify the MSE.

Table 2 Individual Spearman's correlation coefficients computed for each of the six OSATS scores between predicted and real values obtained with the CNN+BiLSTM architecture (LOSO validation scheme).

	Knot-Tying	Needle-Passing	Suturing
Respect for tissue (RFT)	0.83	0.49	0.46
Suture/Needle handling (SNH)	0.82	0.79	0.75
Time and motion (TM)	0.87	0.85	0.68
Flow of operation (FO)	0.76	0.58	0.62
Overall performance (OP)	0.89	0.58	0.71
Quality of final product (QFP)	0.75	0.31	0.67
Mean over the six criteria	0.82	0.60	0.65

Table 3 Mean Squared Error between predicted and actual OSATS scores obtained using the CNN+BiLSTM architecture (LOSO validation scheme)

	Knot-Tying	Needle-Passing	Suturing
Respect for tissue (RFT)	0.22	0.72	0.42
Suture/Needle handling (SNH)	0.23	0.23	0.52
Time and motion (TM)	0.11	0.21	0.52
Flow of operation (FO)	0.31	0.23	0.42
Overall performance (OP)	0.23	0.47	0.43
Quality of final product (QFP)	0.43	0.93	0.35
Mean over the six criteria	0.25	0.47	0.44

Table 4 Comparative results obtained by our model and the models of two other contributions (Means of the Spearman coefficient over the six OSATS scores)

Compared methods	Knot-Tying	Needle-Passing	Suturing
ApEn [29] (*LOSO*)	0.66	0.45	0.59
FCN [25] (*LOSO*)	0.65	0.57	0.60
CNN+BiLSTM (proposed) (*LOSO*)	**0.82**	**0.60**	**0.65**

4.3 LOGO validation scheme

The challenge intended by this validation methodology is to check the performance of our CNN+BiLSTM model on predicting the regression outputs of unseen input kinematic data whose true outputs have been unseen,too, during the training step. This can help to significantly increase the level of confidence we will attribute to our network. table 5, table 6 and table 7 refer to the results of 3 statistical parameters computed for the predicted outputs by our proposed CNN+BiLSTM model for each OSATS criterion, for the Knot-Tying task, the Needle Passing task and the Suturing task, respectively. We can not compute the Spearman correlation coefficient in this case because the real outputs vector has constant values (knowing that we leave a unique grade for testing). Hence, there is no variation in the real outputs vector so its standard deviation is equal to 0 which will result in zero division in the Spearman function, thereby being undefined. Instead, we computed the median, the standard deviation and the mean of the predicted outputs for each OSATS criterion. We compared the mean of predicted outputs to the real outputs for each OSATS criterion. We consider that the result is quite good when the difference between ground truth scores and predicted scores is not greater than 1. We decided to take this cutoff to evaluate the network prediction performance according to the JIGSAWS surgery skills annotation levels. In the metadata files of the dataset, surgical performance is annotated in 1 grade steps, i.e. surgical performance reaches a new level every 1 step. We can notice that all the trained models for each criterion have difficulties in predicting a score that was unseen during the training phase. One can notice that the results are good on the Time and Motion (TM) criterion in the Knot-Tying and in the Suturing tasks (table 7, table 6. The reason behind this might be the fact that this criterion is more correlated with the kinematic data since the latter is a description of the surgical robot tools motions through time steps. However, a criterion like Respect for Tissue (RFT) seems to be more related to forces and dynamics, which are not described in the kinematic data.

Failure case: We consider the results obtained on the Needle Passing task (table 6) as a failure case because the network yielded constant prediction vectors, meaning that the network did not properly learn features in this case.

Table 5 Regression results with the CNN+BiLSTM architecture for the Knot-Tying task while adopting the LOGO validation scheme.RFT: Respect for tissue. SNH: Suture/Needle handling. TM: Time and motion. FO: Flow of operation. OP: Overall performance. QFP: Quality of final product.

Statistical parameters	RFT	SNH	TM	FO	OP	QFP
Median	3.07	2.89	2.08	3.07	2.91	1.88
Standard deviation	0.10	0.36	0.50	0.12	0.15	0.20
Mean of the predicted outputs	3.05	2.73	**1.98**	**3.06**	2.91	**1.90**
Real outputs	4.00	4.00	**2.00**	**4.00**	4.00	**1.00**

Table 6 Regression results with the CNN+BiLSTM architecture for the Needle Passing task while adopting the LOGO validation scheme.

Statistical parameters	RFT	SNH	TM	FO	OP	QFP
Median	2.47	2.99	2.30	2.81	2.10	2.70
Standard deviation	0.32	0.29	0.76	0.16	0.19	0.61
Mean of the predicted outputs	2.45	2.84	2.66	2.80	2.16	2.78
Real outputs	4.00	4.00	3.00	2.00	3.00	1.00

Table 7 Regression results with the CNN+BiLSTM architecture for the Suturing task while adopting the LOGO validation scheme.

Statistical parameters	RFT	SNH	TM	FO	OP	QFP
Median	2.15	2.04	4.127	4.02	4.15	4.08
Standard deviation	0.25	0.06	0.08	0.13	0.07	0.06
Mean of the predicted outputs	2.24	2.05	**4.12**	3.99	**4.14**	**4.07**
Real outputs	1.00	1.00	**5.00**	5.00	**5.00**	**5.00**

Limitations

Regardless of the results we obtained through our present contribution in the field of automated surgical skill assessment, deep learning models still turn out to be limited with regards to online skill assessment. First, The lack of data and variety in the JIGSAWS dataset constitutes a problem since training deep learning models mainly rely on huge amounts of data. In addition, we think that the OSATS labels in the JIGSAWS dataset need to be more accurate and maybe more correlated to the surgeon expertise level. One can see that some beginner surgeons obtained a much better global OSATS score than a expert surgeon. Finally, we need to improve our online assessment by searching for a better weights optimization of our proposed model. If our model excelled in predicting scores while adopting the Random Split Validation Scheme (see Table 1), we still notice significant drops in scores prediction performance while adopting the LOSO and the LOGO schemes (see table 2, table 5, table

6 and table 7). We can not rely only on the Random Split Validation scheme to validate our model since the two others schemes are way more robust.

5 Conclusion and future work

In this work, we presented a novel deep learning method for automatic and objective surgical skill assessment from kinematic data only. The JIGSAWS kinematic data was acquired from the DaVinci surgical robotic system in a training context. Kinematic data may be also available from other ways: 3D localization systems, image guided surgery tools, analysis of surgical videos, and virtual reality based surgical simulation systems for training. In these setups, kinematic data can be made available either using active sensors like external 3D trackers and IMUs, or passive sensors like a laparoscopic camera using surgical instrument recognition, segmentation and tracking. Our proposed model based on a combination between a CNN and a BiLSTM provided new state-of-the-art predictions results on the OSATS criteria. For future work, we aim to extent our model to a real-time surgical skill assessor since it is able to provide feedback by taking small crops of kinematic data. In addition, we aim to include it in a surgical simulation system to provide an online skill assessment feedback. Finally, we have to find a solution to overcome the black-box effect of the deep learning models. It would greatly help to justify decisions taken by a deep learning regression model.

Supplementary information. An additional document named Online Resource 1 is provided with the main manuscript containing supplementary information about the JIGSAWS dataset, a description of modified OSATS criteria, networks implementation details and validation schemes descriptions, the general training approach using soft scores and two studies involving training parameters and a comparison between regression and classification problems using OSATS criteria.

References

[1] Tevis SE, Kennedy GD (2013) Postoperative complications and implications on patient-centered outcomes. journal of surgical research 181(1):106–113

[2] Semel ME, Lipsitz SR, Funk LM, (2012) Rates and patterns of death after surgery in the united states, 1996 and 2006. Surgery 151(2):171–182

[3] Sweeney JF (2013) Postoperative complications and hospital readmissions in surgical patients: an important association. Annals of surgery 258(1):19

[4] Lawson EH, Hall BL, Louie R, (2013) Association between occurrence of a postoperative complication and readmission: implications for quality improvement and cost savings. Annals of surgery 258(1):10–18

[5] Polavarapu HV, Kulaylat AN, Sun S, (2013) 100 years of surgical education: the past, present, and future. Bull Am Coll Surg 98(7):22–27

[6] Kassahun Y, Yu B, Tibebu AT, (2016) Surgical robotics beyond enhanced dexterity instrumentation: a survey of machine learning techniques and their role in intelligent and autonomous surgical actions. International journal of computer assisted radiology and surgery 11(4):553–568

[7] Satava RM (1993) Virtual reality surgical simulator. Surgical endoscopy 7(3):203–205

[8] Sun LW, Van Meer F, Bailly Y, (2007) Design and development of a da vinci surgical system simulator. In: 2007 International Conference on Mechatronics and Automation, IEEE, pp 1050–1055

[9] DiMaio S, Hanuschik M, Kreaden U (2011) The da Vinci Surgical System, Springer, pp 199–217. https://doi.org/10.1007/978-1-4419-1126-1˜9

[10] Levin M, McKechnie T, Khalid S, (2019) Automated methods of technical skill assessment in surgery: a systematic review. Journal of surgical education 76(6):1629–1639

[11] Ahmidi N, Hager GD, Ishii L, (2010) Surgical task and skill classification from eye tracking and tool motion in minimally invasive surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, pp 295–302

[12] Asif H, McInnis C, Dang F, (2021) Objective structured assessment of technical skill (osats) in the surgical skills and technology elective program (sstep): Comparison of peer and expert raters. The American Journal of Surgery https://doi.org/https://doi.org/10.1016/j.amjsurg.2021.03.064, URL https://www.sciencedirect.com/science/article/pii/S0002961021002257

[13] Martin J, Regehr G, Reznick R, (1997) Objective structured assessment of technical skill (osats) for surgical residents. Journal of British Surgery 84(2):273–278

[14] Lubke GH, Muthén BO (2004) Applying multigroup confirmatory factor models for continuous outcomes to likert scale data complicates meaningful group comparisons. Structural equation modeling 11(4):514–534

[15] Sullivan GM, Artino Jr AR (2013) Analyzing and interpreting data from likert-type scales. Journal of graduate medical education 5(4):541–542

[16] Jamieson S (2004) Likert scales: How to (ab) use them? Medical education 38(12):1217–1218

[17] Carifio J, Perla R (2008) Resolving the 50-year debate around using and misusing likert scales

[18] Benmansour M, Handouzi W, Malti A (2018) Task-specific surgical skill assessment with neural networks. In: International Conference on Advanced Intelligent Systems for Sustainable Development, Springer, pp 159–167

[19] Benmansour M, Malti A (2019) Skills evaluation of specific surgical tasks using long short term memory networks. In: International Conference on Advanced Intelligent Systems for Sustainable Development, Springer, pp 331–339

[20] Benmansour M, Malti A (2018) Simple and efficient recurrent neural network to evaluate classified surgery tasks. In: 5th International Conference on Automation, Control Engineering and Computer Science-ACECS

[21] Khalid S, Goldenberg M, Grantcharov T, (2020) Evaluation of deep learning models for identifying surgical actions and measuring performance. JAMA network open 3(3):e201,664–e201,664

[22] Kowalewski KF, Garrow C, Schmidt M, (2019) Sensor-based machine learning for workflow detection and as key to detect expert level in laparoscopic suturing and knot-tying. Surgical Endoscopy 33. https://doi.org/10.1007/s00464-019-06667-4

[23] Liu D, Jiang T, Wang Y, (2019) Surgical skill assessment on in-vivo clinical data via the clearness of operating field. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, pp 476–484

[24] Hung A, Chen J, Che Z, (2018) Utilizing machine learning and automated performance metrics to evaluate robot-assisted radical prostatectomy performance and predict outcomes. Journal of Endourology 32. https://doi.org/10.1089/end.2018.0035

[25] Ismail Fawaz H, Forestier G, Weber J, (2019) Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks. International journal of computer assisted radiology and surgery 14(9):1611–1617

[26] Pascanu R, Mikolov T, Bengio Y (2012) Understanding the exploding gradient problem. CoRR abs/1211.5063. URL http://arxiv.org/abs/1211.5063, https://arxiv.org/abs/arXiv:1211.5063

[27] Gao Y, Vedula SS, Reiley CE, (2014) Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion

modeling. In: MICCAI workshop: M2cai

[28] Wang Z, Majewicz Fey A (2018) Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. International journal of computer assisted radiology and surgery 13(12):1959–1970

[29] Zia A, Essa I (2017) Automated surgical skill assessment in rmis training. International Journal of Computer Assisted Radiology and Surgery 13. https://doi.org/10.1007/s11548-018-1735-5

[30] Chollet F (2015) Keras. https://keras.io, URL https://github.com/fchollet/keras