Supplementary information

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, Algéria.
[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;

Title of the main manuscript: Deep Neural Network Architecture for Automated Soft Surgical Skills Evaluation Using Objective Structured Assessment of Technical Skills Criteria.

Journal: *International Journal of Computer Assisted Radiology and Surgery.*

Introduction

In order to not overwhelm the main manuscript, we present some supplementary information in this additional document. It is intended to provide information about:

- The JIGSAWS dataset [1] that has been used to train and test all our models.
- The general training approach for task-specific surgical skills using floating score that has been adopted is our previous works.
- The literature about surgical gestures and skills classification methods.
- The modified Objective Structured Assessment of Technical Skill (OSATS) criteria. A brief description is provided.

- The experimental setup which present descriptions of adopted validation schemes, the details of the neural networks implementation and the model evaluation metrics.
- An additional study about training the proposed CNN+BiLSTM with different batch sizes and dropout values.
- A regression versus classification study where we compare regression results on OSATS scores with classification results where we considered OSATS scores as categorical.

Each section is referenced in the main manuscript when discussed.

1 The JIGSAWS Dataset

The JIGSAWS dataset [1] has been collected from eight right-handed subjects (B, G, H, I, C, F, E, D) with three different expertise levels: Novice, Intermediate and Expert performing three basic surgical tasks : Suturing (ST), Needle-Passing (NP) and Knot-Tying (KT) using the DaVinci surgical system [2]. Each subject performed five trials for each task. For each trial, the kinematic and video data were recorded but in our works, we focused on kinematic data which are numeric variables of the four components of the DaVinci surgical system: two masters tools (right and left) controlled directly by the subject's hands and two patient side tools, also called slave tools (left and right) controlled indirectly by the subject via the master manipulators. So, the subject tele-operates the slaves tools by manipulating the master tools. Consequently, these kinematic data describe how each surgeon performed the operation. Each trial is represented by a matrix of 76 kinematic variables and each variable represents the changing value of a certain physical quantity of a certain DaVinci tool. For example, the Cartesian positions of the master tools/slave tools, their linear velocities and their rotational velocities. These physical quantities change over time throughout the surgical operation. Each row in a kinematic matrix represents the values of all the kinematic variable at a frame f. In addition, the JIGSAWS dataset contains grades that describe performance of surgeon on each OSATS criteria for each task. These grades have been attributed by senior expert surgeons while supervising subjects sur-geons. These scores will serve as ground truth output for the model of our present work. For more details about the JIGSAWS dataset, please refer to[1]. Here are some quick descriptions of the three tasks :

- **Knot-Tying**: a basic surgery task which consists on performing a single loop knot using a suture floss.
- **Needle-Passing**: this task consists on passing a needle through hoops.
- **Suturing**: a well-known surgery task that consists of entering a "tissue" from one side and exiting it from another side using a needle.

The following images depict some sample frames extracted from recorded videos while subjects were performing the operations.

Fig. 1 Sample frames from the three tasks in the JIGSAWS dataset, from left to right: Suturing frame, Knot-Tying frame, Needle-Passing frame.

2 General training approach for task-specific surgical skills using floating scores

We present here the general adopted approach in some of our previous works [3, 4, 5]. The suggested approach relies on feeding JIGSAWS kinematic data to three independent deep neural networks. Each of these networks is responsible for the assessment of all subjects but in only one of the three basic surgical tasks described previously. The way we have chosen to assess subjects on each task relies on giving float scores as desired outputs to the networks. These float scores reflect the global quality of the performed operation. Therefore, we give a complete score of 1/1 if a sample of expert data is given as input, an average score of 0.7/1 if a sample of intermediate data is given as input and a low score of 0.4/1 if a sample of beginner data is given as input. The intuition here is to inform the networks about the scores deserved by each subject according to his/her expertise level. The general training approach can be summarized in the following items:

- We design three independents neural networks to evaluate the Knot-Tying task, the Needle-Passing task and the Suturing task independently.
- Each network receives some kinematic data from the JIGSAWS database of surgeons with different expertise levels for the training procedure and some remaining data for the testing procedure. The desired outputs depend on whether the current subject is a beginner, intermediate or expert.
- After the training step, the network will be able to make predictions about unseen kinematic data by giving and adequate score to the data sample of the tested surgeon.

As an extension to these works, we propose in the main manuscript a deep learning approach based on a combined architecture between CNN and BiLSTM for the purpose of surgical skill assessment. The assessment is based on the Objective Structured Assessments of Surgical Skills criteria. These evaluation approach is more detailed since it relies on 6 criteria. Therefore, for each surgical task, we build and train a CNN+BiLSTM model for each OSATS criterion.

3 Related work based on classification

Surgical gestures, phases and action recognition: We noticed that the previous works on the Computer Assisted Surgery (CAS) field mostly concern gesture recognition, action segmentation, phase detection or tool tracking. In [6] is proposed a method for surgical phase recognition that uses a CNN called EndoNet to automatically learn features from cholecystectomy videos. In a similar work [7], a segmentation and an surgical action recognition method using a Spatiotemporal CNN is proposed. Although the authors of these two contributions have shown effectiveness of their methods in phase recognition and fine-grained action segmentation, surgical skill evaluation is not involved, while we solely focus on the latter. Another work involving phase recognition has been made in [8]. The authors proposed a pre-trained CNN with a transfer learning method for surgical workflow detection and estimation but skills assessment is not taken into account. *Dandan Zhang, Ruoxi Wang and Benny Lo* developed A bidirectional multi-layer independently RNN for surgical gesture recognition combined with a Deep Convolutional Neural Network (DCNN) model based on the VGG architecture for spatial feature extraction from surgical video frames [9]. The authors used the JIGSAWS dataset to validate their method that covers the surgical gesture recogntion. Skills assessment is not involved. In [10], *Robert DiPietro and Gregory D. Hager* proposed a work on recognizing surgical activities from robot kinematic data using LSTM and BiLSTM. The same first authors developed an architecture based on unsupervised learning that combines an RNN encoder-decoder and mixture density networks (MDNs) to model the conditional distribution over future motion given past motion, showing the possibility to learn meaningful representations of surgical skill motion, without supervision, by learning to predict the future [10]. In another work [11], a comparison is made between four RNN architectures (simple RNNs, long short-term memory, gated recurrent units, and mixed history RNNs) for the purpose of recognizing surgical activities from kinematic data. In [12], it is claimed that RNNs and LSTM are not effective in capturing the relationship of features with different temporal scales, leading to sub-optimal recognition performance of surgical activities containing complex motions at multiple time scales. Thus, a Multiscale recurrent neural networks (MS-RNN) that combines the strength of both wavelet scattering operations and LSTM has been developed for the purpose of surgical activities recognition. In [13], a work that focuses on manual dexterity by considering it as one of the most important surgical skills is proposed. A system is designed to track surgeon's hand movements during simulated open surgery tasks. Then, their manual expertise is evaluated using and an artificial neural network (ANN) classifier. Although the ANN achieved very good results, we believe that the inclusion of recurrent neural network or a CNN would yield better results than a simple ANN.Another work on surgical gesture recognition applied a Time Delay Neural Network (TDNN) to JIGSAWS kinematic data to introduce temporal modeling in gesture recognition [14].[15] proposed a multi-class SVM for surgical gestures dictionary learning and classification which can be

used to effectively analyze complex surgical gestures recorded by the Da Vinci robotic surgical system. In [16], the authors developed and evaluated a machine learning platform for the construction of a holistic biomechanical model of the surgeon and of the instruments used for minimally invasive surgery that uses a Markov chain for gesture analysis. The machine learning method is able to recognize expertise level of surgeons but only after setting an expert surgeon performance as a reference. Finally, in [17], six techniques based on a temporal approach for segmentation and recognition of gestures in robotic surgery including Hidden Markov Models and Semi HHMs have been presented. All these works involves surgical gestures recognition or action segmentation but do not propose a surgical skill assessment method. In this paper, we focus only on the skill assessment regardless of the surgical gestures or the surgical phases the surgeon is going through.

Surgical skill classification: On the other hand, *Hei Law, Khurshid Ghani and Jia Deng* proposed an automated method [18] for surgical skill evaluation by tracking surgical instruments using a Hourglass Network. The evaluation relies on the five GEARS criterion [19]. However, surgeons are classified by adopting scores thresholds. Thus, performance of surgeons is not faithfully assessed. Our proposed model automatically and objectively evaluates surgeons on a continuous scale without using any score thresholds. In [20] an automatic surgical skill assessment approach based on tool tracking and analyzing tool movements in surgical videos is presented, using region-based convolutional neural networks. This method was the first to not only detect presence but also spatially localize surgical tools in real-world laparoscopic surgical videos. In [21] is proposed a 3D convolutional neural networks have to classify snippets (a batch of few consecutive frames extracted from surgical video) from the JIGSAWS dataset into three expertise classes and without involving the OSATS scores. Although the proposed method achieves high classification accuracy, it does not demonstrate its capability of ranking subjects on a soft continuous scale. [22] used a pairwise deep ranking model for skill comparison in video. The proposed model, which is a two stream CNN, characterizes the relative differences in performance between a pair of videos containing a high ranked user and a low ranked one. JIGSAWS surgical videos have been used for training and testing the model. An interesting work done by *Wang and Fey* involves an in-real-time surgical skill assessment method [23] by building a deep CNN that can reliably interpret skills within 1-3 seconds window. The same authors presented a combined architecture of a CNN and a Gated Recurrent Unit (GRU) for online trainee skill analysis and task recognition [24]. The CNN component learns spatial abstract representations within the interval of input frame while the GRU component learns the temporal dynamics of multiple channels at each time step in raw motion data. Consequently, the model can characterize the nature of surgery motion relative to both the surgeon level of expertise and operation task. Although these works contribute to both surgical skill classification and task recognition

domains, they still do not propose a robust method to truly grade a surgical performance by giving an exact score, instead of categorizing it. In [23], *Wang and Fey* actually trained their CNN to assess surgeons using the Global Rating Score (GRS) of the JIGSAWS dataset, but they did not set the original GRS outputs for a regression approach and they opted instead for a GRS classfication approach by establishing scores thresholds. On the other hand, we propose not only to assess surgeons by using the raw outputs, thus following a regression approach, but also, we take into account the six OSATS criteria. In [25], the authors took advantage of bidirectional LSTM (BiLSTM) that is able to read input time series data in a forward direction and in a backward direction as well, in order to classify surgeon expertise level from the Basic Laparoscopic Urologic Skills dataset [26]. This work has proven the effectiveness of the BiLSTM in surgeon level categorization. Hence, we decided to include a BiLSTM block in our model architecture but for the purpose of predicting the exact surgical performance score. In [27], a descriptive structure for nasal septoplasty is provided by automatically segmenting it into higher-level meaningful activities called strokes . The sequence of strokes has been used to train a SVM classifier to distinguish between novice and expert surgeons. However, the best classification accuracy they obtained is 73%, which remains pretty low for a classification task and there is no floating score for an exact performance evaluation. Finally, another work [28] using SVMs combined to a logistic regression method for robotic minimally invasive surgery skill assessment provided an evaluation method on six important movement features, leading to a strong surgical performance evaluation scheme.

4 The modified Objective Structured Assessment of Technical Skill by JIGSAWS

The original OSATS method [29] was established in 1997 by the Surgical Education Research Group which belongs to the University of Toronto, Canada. The authors claimed that the surgeons were not well assessed regarding to their technical skill. Therefore, they came up with an approach that can deliver an accurate evaluation of residents who performed a variety of structured operative tasks under the supervision of mentor surgeons certified by the Royal College of Physicians and Surgeons of Canada . Both live animals and bench models were used for this. They developed two types of scoring system. A first one called operation-specific checklist which consists of attributing a binary value (1 for "Done Correctly" and 0 for "Not Done or Done Incorrectly") for the performed steps composing a certain surgical task. The second scoring system consists of a detailed global rating scale. This second scoring system aims to give a score on a scale of 1 to 5 on seven criteria: (1) Respect for Tissue, (2) Time and motion, (3) Instrument handling, (4) Knowledge of instruments, (5) Use of assistants, (6) Flow of operation and forward planning, (7) Knowledge of specific procedure. Each score (from 1 to 5) describes the quality of the skill on

the concerned criterion. The goal of the whole work is to validate the two scoring systems and compare them. At the end, The authors have strongly proven by the obtained results that the Objective Structured Assessment of Technical Skill can reliably and validly assess surgical skills. They demonstrated and concluded that global ratings are a better method of assessment than task-specific checklists. This is why the JIGSAWS authors have chosen to assess their own trainees by using the global ratings method. Nevertheless, they did not perform the evaluation on live animals but only on workbench models and they also barely modified the assessment criteria by deleting or changing some items according to their own working conditions. For example, there was not any assistant for the subjects and the subjects performed operations with the Da Vinci surgical system using only a suture and a needle. In [1], one can see that "Instrument handling"has been replaced by "Suture/Needle handling". There is no more "Knowledge of specific procedure" and "Knowledge of instruments" but instead, one can find "Overall performance" and "Quality of final product". Even the scores descriptions differ from the original OSATS global rating scale. Interpretations of 1, 3 and 5 OSATS scores can be read in Table 1. However, JIGSAWS surgeons did not give interpretation for 2 and 4 scores despite being accorded to several trainees. The meta files contained in the JIGSAWS database show the obtained scores by each subject on each criterion of the modified OSATS criteria regarding to the three surgical tasks: Knot-Tying, Needle-Passing and Suturing.

Element	Rating scale
Respect for tissue	1-Frequently used unnecessary force on tissue; 3- Careful tissue handling but occasionally caused inadvertent damage; 5-Consistent appropriate tissue handling;
Suture/needle handling	1- Awkward and unsure with repeated entanglement and poor knot tying; 3-Majority of knots placed correctly with appropriate tension; 5-Excellent suture control
Time and motion	1-Made unnecessary moves; 3-Efficient time/motion but some unnecessary moves; 5-Clear economy of movement and maximum efficiency
Flow of operation	1-Frequently interrupted flow to discuss the next move; 3-Demonstrated some forward planning and reasonable procedure progression; 5-Obviously planned course of operation with efficient transitions between moves;
Overall performance	1-Very poor; 3-Competent; 5-Clearly Superior;
Quality of final product	1-Very poor; 3-Competent; 5-Clearly Superior;

Table 1 OSATS surgical skills annotations provided by the JIGSAWS expert surgeons.

5 Experimental setup

5.1 Training and testing steps

Our model has been trained and tested with different sets of the sub-sequences obtained after executing the data augmentation process (described in the main manuscript in Section 3.3). To get the most robust model for each architecture, we adopted three distinct validation schemes:A random training/validation/test data splitting that we will call *Random Data Splitting*, the famous *Leave-One-Supertrial-Out (LOSO)* cross validation scheme and a custom scheme that we will call *Leave-One-Grade-Out (LOGO)*.

Random Data Splitting: This first validation scheme consists in dividing the set containing all the crops of kinematic data into three distinct subsets: Training data, validation data and test data. The splitting is performed randomly by choosing a certain percentage for each subset. This scheme is performed after extracting sub-sequences from all the kinematic data with the data augmentation process.

Leave-One-Supertrial-Out (LOSO): As described previously, the JIGSAWS dataset comprises 5 trials for each subject performing one of the three surgical tasks. The LOSO scheme consists in leaving a folder containing one supertrial ,"*i*", of each subject for testing the network while the remaining trials are used for the training step. This process is repeated five times in five folders in order to test the robustness of the model with each trial. At the end, the model that delivered the best performance with any of the five trials is saved. The five folders are set before augmenting the data (entire trials).

Leave-One-Grade-Out (LOGO): This third validation scheme is a custom scheme we came up with for this work. It consists in choosing one grade (or score) of each OSATS criterion for each surgical task and leaving it for the test step. The purpose of this scheme is to test the network with a grade that is unseen during the training step. Thus, this validation method focuses on the outputs (scores), unlike the two aforementioned validation schemes that focus on the inputs (kinematic data). Unlike the two previously described validation schemes, the LOGO scheme shows the network effectiveness and how it would behave in real world when confronted to new data with previously unknown outputs. We believe that it proves robustness of our network more efficiently. The process is done before augmenting the data.

We have a trained and tested individual models built according to the CNN+BiLSTM architecture, on the basis of each validation scheme, for each surgical task and for each OSATS criterion.

5.2 Implementation

We used Keras, a deep learning Python library to implement our network algorithm and train it from scratch. The parameters of the layers of the networks are initialized with the Glorot Uniform initialization method. We employed the optimizer Adam for all the networks with a learning rate of 0.001 while adopting the Random Data Split validation scheme. The learning rate has been decreased to 0.0001 while employing the two remaining validation schemes. The exponential decay rates of the 1st and 2nd moment estimates are set to 0.9 and 0.999, respectively. The gradient descent updates are performed with the mini batch learning method using a mini batch size of 32. To avoid overfitting, we added a $L2$ regularization with a value of 0.01 as well as input and hidden noise layers. The goal is to minimize the loss function which is a Mean Squared Error function between the real OSATS values and the predicted OSATS values. The size of sliding window is set to 60 with a step size of 30 when running the Random Split Data validation scheme. We increased the sliding window size to 120 while keeping the same step size of 30 when running the LOSO and the LOGO validation schemes. Theses values have been chosen after several experiments based on trial-and-error. To obtain the best model with the lowest error possible while training each network, we included an Early Stopping callback in the training function. This callback allows to stop the training if an arbitrary number of epochs is exceeded without any learning improvement (also called as *patience epochs*). At the end, the weights that brought the lowest validation loss are saved and returned to build the final and the best model.

5.3 Model evaluation metrics

We compute Spearman's rank correlation coefficients to highlight correlations between the ground truth OSATS scores and the predicted OSATS scores. The Spearman rank correlation coefficient (often symbolized by ρ) is a non-parametric measure of rank correlation. It assesses how well the relationship between two variables can be described using a monotonic function. Intuitively, the Spearman correlation between two distributions will be high (close to 1) when observations have a similar rank and low when having a dissimilar rank (close to -1). In addition we compute 3 statistical parameters: median, standard deviation and mean while evaluating the model on the basis of the LOGO scheme.

6 Training parameters study

In this section, we provide an experimental study involving a comparison between regression results obtained using two neural network architectures and by affecting different values to training batch size and dropout in the convolutional layers. The first architecture is the proposed CNN+BiLSTM and the second one is a variation of the latter where we put a GRU recurrent network instead of a LSTM (CNN+BiGRU) and we kept same number of units

(16) and same hyperparameters described in the main manuscript, subsection 3.2. The structural difference between these two recurrent networks lies on the number of gates: LSTM has three gates named *input, output* and *forget* gates, whereas GRU has only two gates named *reset* and *update* gates. Moreover, GRUs and LSTMs deal differently with the vanishing gradient problem which is encountered with the classic RNN. Here are the main points comparing the two:

- The GRU unit controls the flow of information like the LSTM unit, but without having to use a memory unit. It just exposes the full hidden content without any control.
- GRUs are computationally more efficient since they have a less complex structure. Thus, they train faster but are as effective as LSTMS in making predictions on sequence data.

For a detailed description, this paper [30] explains this efficiently.

We trained and tested models on the basis of the two architectures following the same data augmentation method described in the main manuscript (subsection 3.3) for all surgical tasks and all OSATS criteria. Learning rate, weights initialization, L2 regularization values and other implementation details described in subsection 5.2 stayed the same as well. LOSO validation scheme has been adopted while testing every model. Spearman rank correlation coefficient (ρ) has been used to compute correlation between predicted and target outputs. Regression results are reported in 2 and 3 for the CNN+BiLSTM and the CNN+BiGRU, respectively. For the batch size, we have chosen to double it twice starting from 32. Dropout values have been set to 25%, 50% and 75% as these values are common in the deep learning community. Each batch size is tested alongside a single dropout value. We could not perform tests for all batch sizes alongside all dropout values because of the lack of computational resources.

As it can be seen, both architectures have shown good regression correlation results for the three tasks and the six OSATS criteria while affecting 32 and 25% values for the batch size and the dropout regularization, respectively. Increasing dropout to 50% and and the batch size to 64 led to a significant decrease in the quality of scores prediction for the CNN+BiLSTM architecture while the CNN+BiGRU suffered only from a slight decrease in performance after increasing the values of the two training parameters. Finally, increasing batch size and dropout to 128 and 75% respectively led to a near decorrelation between predicted and true outputs concerning the CNN+BiLSTM while the CNN+BiGRU behaves better but nevertheless showed a huge drop in prediction quality.

We conclude that training both architectures with the smallest batch size and the lowest percentage of dropout led to the best OSATS scores prediction

quality. To support this study, we cite this work [31] where *Ibrahem Kandel et al.* conducted a study about the effect of batch size on the performance of CNNs in medical image classification using a VGG16 network and training it with different batch sizes. Their results have shown that the network that has been trained with the lowest batch size provided the best accuracy and thus the best generalization. On the other hand, dropout regularization has been introduced in the deep learning field to prevent overfitting and improve robustness and generalization of the neural network on validation and test data. But, after a certain threshold percentage, the network may underfit. This may or may not happen depending on the network architecture complexity and the training dataset size and nature.

	Knot-Tying			Needle-Passing			Suturing		
Batch size	32	64	128	32	64	128	32	64	128
Dropout	25%	50%	75%	25%	50%	75%	25%	50%	75%
Respect for tissue	0.76	0.49	0.01	0.50	0.08	0.12	0.50	0.33	0.17
Suture/Needle handling	0.77	0.47	0.04	0.81	0.58	0.24	0.74	0.74	0.01
Time and motion	0.85	0.54	0.21	0.83	0.67	0.68	0.69	0.67	0.11
Flow of operation	0.76	0.42	0.13	0.55	0.22	0.01	0.64	0.58	0.40
Overall performance	0.89	0.72	0.10	0.56	0.40	0.23	0.79	0.58	0.36
Quality of final product	0.79	0.06	0.07	0.31	0.37	0.16	0.80	0.67	0.27
Mean on all criteria	0.80	0.45	0.02	0.59	0.38	0.24	0.69	0.59	0.22

Table 2 Correlation results obtained using the CNN+BiLSTM architecture by varying training batch sizes and dropout values.

	Knot-Tying			Needle-Passing			Suturing		
Batch size	32	64	128	32	64	128	32	64	128
Dropout	25%	50%	75%	25%	50%	75%	25%	50%	75%
Respect for tissue	0.82	0.66	0.15	0.38	0.43	0.19	0.54	0.52	0.36
Suture/needle handling	0.84	0.67	0.23	0.80	0.78	0.44	0.77	0.77	0.58
Time and motion	0.85	0.76	0.33	0.86	0.85	0.44	0.70	0.49	0.42
Flow of operation	0.78	0.67	0.65	0.63	0.64	0.62	0.65	0.66	0.67
Overall performance	0.88	0.78	0.44	0.60	0.60	0.22	0.74	0.66	0.29
Quality of final product	0.62	0.64	0.12	0.23	0.25	0.05	0.71	0.68	0.55
Mean on all criteria	0.80	0.70	0.32	0.58	0.60	0.33	0.69	0.63	0.48

Table 3 Correlation results obtained using the CNN+BiGRU architecture by varying training batch sizes and dropout values.

7 Regression versus classification

In this section, we present results obtained through a classification method using OSATS scores. It consists in considering OSATS outputs as categorical instead of considering it as a continuous scale. The goal of this experiment is to determine which approach leads to the best results and thus, gives the best and most meaningful feedback to the trainee. To compare this approach to the

proposed regression approach appropriately, we used the same CNN+BiLSTM architecture with the same hyperparameters described in subsection 3.2 in the main manuscript. The same implementation details described in subsection 5.3 have been adopted as well. The difference lies on the following: For this classification approach, we used the categorical Crossentropy loss function instead of the mean squared error loss function. In addition, desired outputs have been one-hot encoded and we used a Softmax activation function in the last layer of the network architecture to compute the predicted output on a one-hot encoded 6 elements array:

- Score 0 becomes $[1, 0, 0, 0, 0, 0]$
- Score 1 becomes $[0, 1, 0, 0, 0, 0]$
- Score 2 becomes $[0, 0, 1, 0, 0, 0]$
- Score 3 becomes $[0, 0, 0, 1, 0, 0]$
- Score 4 becomes $[0, 0, 0, 0, 1, 0]$
- Score 5 becomes $[0, 0, 0, 0, 0, 1]$

To evaluate performance of a network at predicting categorical outputs, metrics as accuracy, precision or F1-score are usually used. In our case, we are interested in the comparison with the regression approach. Thus, we computed the Spearman rank correlation coefficient between predicted and target outputs. To do so, we performed the reverse of one-hot encoding for each predicted outputs and went back to original score by considering the index of the maximum value in the one-hot encoded array (For example, if a predicted array is $[0.012, 0.014, 0.861, 0.001, 0.002, 0.110]$, the maximum value has the index 2 and the output would be then classified as 2. After that, we are able to compute Spearman (ρ) correlation coefficients between predicted and target outputs for the classification method. To provide a comparison on the same basis than the regression approach, we used the LOSO validation scheme.

Results are reported in Table 7: For each surgical task, the left column represents the computed Spearman coefficients for each OSATS criterion of the regression approach, while the right one represents the computed ρ coefficients of the classification approach. Clearly, ρ coefficients obtained through the pro-posed regression approach are superior to ρ coefficients obtained through the classification approach on the three surgical tasks. This means that predicted and target outputs are more correlated by adopting the regression approach than by adopting the classification approach. We can conclude that the pro-posed regression method provides a more meaningful and meticulous feedback to surgical trainees.

	Knot-Tying		Needle-Passing		Suturing	
Respect for tissue (RFT)	0.83	0.66	0.49	0.46	0.46	0.38
Suture/Needle handling (SNH)	0.82	0.62	0.79	0.67	0.75	0.54
Time and motion (TM)	0.87	0.79	0.85	0.71	0.68	0.54
Flow of operation (FO)	0.76	0.63	0.58	0.59	0.62	0.54
Overall performance (OP)	0.89	0.71	0.58	0.46	0.71	0.57
Quality of final product (QFP)	0.75	0.50	0.31	0.28	0.67	0.56

Table 4 Comparative table showing correlation results obtained for both classification and regression methods. For each task, the left column shows the Spearman correlation results obtained with the proposed regression method while the right one shows the Spearman correlation results obtained with the classification method. (CNN+BiLSTM and LOSO validation scheme)

References

[1] 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

[2] 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

[3] 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

[4] 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

[5] 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

[6] Twinanda A, Shehata S, Mutter D, (2016) Endonet: A deep architecture for recognition tasks on laparoscopic videos. IEEE Transactions on Medical Imaging 36. https://doi.org/10.1109/TMI.2016.2593957

[7] Lea C, Reiter A, Vidal R, (2016) In: Computer Vision – ECCV 2016Leibe B, Matas J, Sebe N, (eds) Segmental spatiotemporal cnns for fine-grained action segmentation. Springer International Publishing, Cham, pp 36–52

[8] Sahu M, Mukhopadhyay A, Szengel A, (2016) Tool and phase recognition using contextual CNN features. CoRR abs/1610.08854. URL http://arxiv.org/abs/1610.08854, https://arxiv.org/abs/arXiv:1610.08854

[9] Zhang D, Wang R, Lo B (2021) Surgical gesture recognition based on bidirectional multi-layer independently rnn with explainable spatial feature extraction. In: 2021 IEEE International Conference on Robotics and Automation (ICRA), IEEE, pp 1350–1356

[10] DiPietro R, Hager GD (2018) Unsupervised learning for surgical motion by learning to predict the future. In: International conference on medical image computing and computer-assisted intervention, Springer, pp 281–288

[11] DiPietro R, Ahmidi N, Malpani A, (2019) Segmenting and classifying activities in robot-assisted surgery with recurrent neural networks. International journal of computer assisted radiology and surgery 14(11):2005–2020

[12] Gurcan I, Nguyen HV (2019) Surgical activities recognition using multi-scale recurrent networks. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 2887–2891, https://doi.org/10.1109/ICASSP.2019.8683849

[13] Sbernini L, Quitadamo LR, Riillo F, (2018) Sensory-glove-based open surgery skill evaluation. IEEE Transactions on Human-Machine Systems 48(2):213–218. https://doi.org/10.1109/THMS.2017.2776603

[14] Menegozzo G, Dall'Alba D, Zandonà C, (2019) Surgical gesture recognition with time delay neural network based on kinematic data. In: 2019 International Symposium on Medical Robotics (ISMR), pp 1–7, https://doi.org/10.1109/ISMR.2019.8710178

[15] Sefati S, Cowan NJ, Vidal R (2015) Learning shared, discriminative dictionaries for surgical gesture segmentation and classification. In: MICCAI Workshop: M2CAI

[16] Cavallo F, Sinigaglia S, Megali G, (2014) Biomechanics–machine learning system for surgical gesture analysis and development of technologies for minimal access surgery. Surgical Innovation 21(5):504–512

[17] Tao L, Zappella L, Hager GD, (2013) Surgical gesture segmentation and recognition. In: International conference on medical image computing and computer-assisted intervention, Springer, pp 339–346

[18] Law H, Ghani K, Deng J (2017) Surgeon technical skill assessment using computer vision based analysis. In: Machine learning for healthcare conference, PMLR, pp 88–99

[19] Goh A, Goldfarb D, Sander J, (2011) Global evaluative assessment of robotic skills: Validation of a clinical assessment tool to measure robotic

surgical skills. The Journal of urology 187:247–52. https://doi.org/10.1016/j.juro.2011.09.032

[20] Jin A, Yeung S, Jopling J, (2018) Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) pp 691–699

[21] Funke I, Mees ST, Weitz J, (2019) Video-based surgical skill assessment using 3d convolutional neural networks. International journal of computer assisted radiology and surgery 14(7):1217–1225

[22] Doughty H, Damen D, Mayol-Cuevas W (2018) Who's better? who's best? pairwise deep ranking for skill determination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 6057–6066

[23] 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

[24] Wang Z, Fey AM (2018) Satr-dl: Improving surgical skill assessment and task recognition in robot-assisted surgery with deep neural networks. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, pp 1793–1796

[25] Kelly J, Petersen A, Lendvay T, (2020) Bidirectional long short-term memory for surgical skill classification of temporally segmented tasks. International Journal of Computer Assisted Radiology and Surgery 15:2079–2088. https://doi.org/10.1007/s11548-020-02269-x

[26] Kowalewski TM, Comstock B, Sweet R, (2016) Crowd-sourced assessment of technical skills for validation of basic laparoscopic urologic skills tasks. The Journal of urology 195(6):1859–1865

[27] Poddar P, Ahmidi N, Vedula SS, (2014) Automated objective surgical skill assessment in the operating room using unstructured tool motion. CoRR abs/1412.6163. URL http://arxiv.org/abs/1412.6163, https://arxiv.org/abs/arXiv:1412.6163

[28] Fard MJ, Ameri S, Chinnam RB, (2016) Machine learning approach for skill evaluation in robotic-assisted surgery. CoRR abs/1611.05136. URL http://arxiv.org/abs/1611.05136, https://arxiv.org/abs/arXiv:1611.05136

[29] 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

[30] Chung J, Gülçehre Ç, Cho K, (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR abs/1412.3555. URL http://arxiv.org/abs/1412.3555, https://arxiv.org/abs/1412.3555

[31] Kandel I, Castelli M (2020) The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. ICT express 6(4):312–315