International Journal of Computer Assisted Radiology and Surgery (2022) 17:1039–1048
https://doi.org/10.1007/s11548-022-02581-8

ORIGINAL ARTICLE

Ranking surgical skills using an attention-enhanced Siamese network with piecewise aggregated kinematic data

Burçin Buket Oğul[1,2] · Matthias Gilgien[3,4] · Suat Özdemir[1]

Received: 10 November 2021 / Accepted: 16 February 2022 / Published online: 14 March 2022

Abstract

Purpose Surgical skill assessment using computerized methods is considered to be a promising direction in objective performance evaluation and expert training. In a typical architecture for computerized skill assessment, a classification system is asked to assign a query action to a predefined category that determines the surgical skill level. Since such systems are still trained by manual, potentially inconsistent annotations, an attempt to categorize the skill level can be biased by potentially scarce or skew training data.

Methods We approach the skill assessment problem as a pairwise ranking task where we compare two input actions to identify better surgical performance. We propose a model that takes two kinematic motion data acquired from robot-assisted surgery sensors and report the probability of a query sample having a better skill than a reference one. The model is an attention-enhanced Siamese Long Short-Term Memory Network fed by piecewise aggregate approximation of kinematic data.

Results The proposed model can achieve higher accuracy than existing models for pairwise ranking in a common dataset. It can also outperform existing regression models when applied in their experimental setup. The model is further shown to be accurate in individual progress monitoring with a new dataset, which will serve as a strong baseline.

Conclusion This relative assessment approach may overcome the limitations of having consistent annotations to define skill levels and provide a more interpretable means for objective skill assessment. Moreover, the model allows monitoring the skill development of individuals by comparing two activities at different time points.

Keywords Robot-assisted surgery · Skill assessment · Attention-enhanced Siamese networks · Assessment of surgical skills

Introduction

Assessment of surgical skills may have three main objectives: (1) choosing appropriate surgeons for a specific operation, (2) examining current performance of candidate surgeons before credentialing, and (3) monitoring the progress of surgeon's skills during training activities. These assessment activities are usually performed manually in an operation room under supervision and feedback of expert surgeons. Manual assessment of surgical skills by individuals may lead to misinterpretations of the skill performance and hence lead to suboptimal training and organization of the surgical activities. Some structured methods such as Objective Structured Assessment of Technical Skills (OSATS [1]) have been employed to minimize the effect of the subjective nature of expert intervention. However, the process needs improvements to increase its efficiency since the application of these techniques still require significant effort of multiple experts

✉ Burçin Buket Oğul
buket.ogul@cankaya.edu.tr

Matthias Gilgien
matthias.gilgien@nih.no

Suat Özdemir
ozdemir@cs.hacettepe.edu.tr

[1] Department of Computer Engineering, Hacettepe University, Ankara, Turkey

[2] Department of Software Engineering, Çankaya University, Ankara, Turkey

[3] Department of Physical Performance, Norwegian School of Sport Sciences, Oslo, Norway

[4] Center of Alpine Sports Biomechanics, Engadin Health and Innovation Foundation, Samedan, Switzerland

Table 1 Methods for computerized assessment of surgical skills

References	Data type	Task	Dataset
Fard et al. [4]	Kinematic	Classification	JIGSAWS
Fawas et al. [7]	Kinematic	Classification	JIGSAWS
Wang and Fey [6]	Kinematic	Classification	JIGSAWS
Zia and Essa [5]	Kinematic	Regression	JIGSAWS
Dougthy et al. [15]	Video	Ranking	JIGSAWS
Fawas et al. [7]	Kinematic	Regression	JIGSAWS
Funke et al. [9]	Video	Classification	JIGSAWS
Nguyen et al. [10]	Kinematic	Classification	JIGSAWS
Li et al. [16]	Video	Ranking	JIGSAWS
Ogul et al. [17]	Kinematic	Ranking	JIGSAWS
Zhang et al. [8]	Kinematic	Classification	JIGSAWS
Kelly et al. [12]	Kinematic	Classification	In-house
Lavanchy et al. [13]	Video	Classification	In-house
Perez-Escamirosa et al. [14]	Video	Classification	In-house
This study	Kinematic	Ranking Regression Monitoring	JIGSAWS, ROSMA

over a long time period [2]. Considering the fact that evaluation of the candidates by senior surgeons has certain cost, there is an increasing need for alternative or complementary computerized assessment systems.

We have recently witnessed a significant attempt to computerize surgical skill assessment using machine learning algorithms [3]. Robot-assisted surgery helps this effort by providing data in different forms, such as kinematic sensor measurements derived from robot arms and video recording of a surgical action performed by an operator. An overview of recent methods for computerized skill assessment using machine learning is given in Table 1.

In one of the earliest studies, kinematic data collected during robot-assisted surgery were used to predict the expertise level of the surgeon [4]. A set of hand-crafted features were extracted from surgery action and fed into three different supervised classifiers (k-Nearest Neighbour, Support Vector Machine (SVM) and Linear Regression) for classification of surgeons into either "expert", "intermediate" or "novice" levels. The authors employed several kinematic features including task completion time, path length, depth perception, speed, motion smoothness, curvature, turning angle and tortuosity to build the model. In a similar work [5], the authors used different time and frequency domain features of kinematic data, which were obtained through sequential motion texture, discrete Fourier transform, discrete cosine transform and approximate entropy analysis to train a linear SVM model. In addition to classification, i.e. assigning objects into predefined skill labels, they also considered to predict the level of skills by running the SVM in a regression setup. Wang and Fey [6] proposed a deep learning architecture based on Convolutional Neural Networks (CNN) that can automatically extract relevant features and classify the expertise level using a fully-connected layer at the end. Similar architectures were used by Fawas et al. [7] and Zhang et al. [8] with slight modifications in layer organizations. Funke et al. [9] used video recordings of surgery actions instead of motion kinematics to feed a 3D CNN with the same objective (ternary classification). CNN was combined with Long Short-Term Memory (LSTM) model to analyze kinetic data for classification [10]. These studies reported very high classification accuracy, up to 100% for some surgery actions, in a public benchmark dataset for human gesture and skill assessment from surgical activity, called JIGSAWS [11]. The performance of conventional machine learning methods with hand-crafted features was recently re-evaluated in a larger in-house dataset [14], where they determined that an average accuracy of 91.5% can be achieved in binary classification of skill. The LSTM model was shown to be accurate in binary skill classification ("expert" or "novice") from kinematic signals in a private dataset [12]. The ability of CNN applied on video recordings was further assessed in another study with an in-house dataset [13]. However, they reported that the accuracy diminished from 86 to 70% when they increased the number of skill categories from two to five.

The major problem with these performance assessment systems is their limited ability to predict a fixed number of predefined, possibly inconsistent, categories for skill levels. As reported by Lavanchy et al. [13], they are unable to model skill levels between these predefined categories. Recalling the three main objectives for surgical skill assessment, discussed at the beginning of the text, i.e. (1) choosing appropriate surgeon, (2) examining current performance of surgeons, and (3) monitoring the progress of a surgeon, the classification approach may support partially the second objective. However, it fails to provide an accurate solution for first and third tasks since the number of categories representing skill levels is not sufficient to model precise comparison of actions. Regression can be considered as a possible solution in general. However, in small dataset scenarios, where continuous labels representing skill levels are too sparse, it is not easy to provide generalizable models for exact value predictions. Two previous approaches for this [5, 7] indeed reported very low correlations between predicted and actual skill levels.

The skill assessment problem was recently considered as a task of learning to rank video recordings [15, 16] instead of assigning them into predefined labels. These studies aimed to build generic models with wide applicability of skill determination in any domain, but algorithms were also tested for surgical skill assessment with the JIGSAW dataset. First, the study introduced a two-stream Temporal Segment Network to capture both the type and quality of actions [16]. Second, the study integrated an attention pooling and temporal aggregation mechanism to a two-stream CNN model [16]. Skill assessments through video recordings have two main limitations. First, video data processing is time and resource inefficient, which makes it difficult to run the algorithms in conventional personal computers. Second, video can record the actions in two dimensions, if only one camera is used. This is unfortunate since tracking of trajectories and velocities can only be measured in two dimensions and important information of surgical skills is lost, if the third dimension is lacking.

It has been shown in many studies that the use of motion characteristics obtained from kinematic sensors is promising to be used in medical practice. In the earlier study of Lin et al., [21], it was shown that the tool motion of an experienced surgeon has more clearly defined features than that of less experienced surgeon while performing the same task using da Vinci Surgical System. Fard et al. [4] showed that the kinematic data from the same system is able to provide direct measures of motions, such as path length, depth perception, speed, motion smoothness, curvature, turning angle and tortuosity, which are highly representative for modeling surgeon's ability. According to a recent systematic literature review by Castillo-Sagura et al., [22] tool motion data has been used in 59 of 101 papers identified for objective assessment of surgical skills. Common indicators used in these studies are organized into five types, position, velocity, acceleration, orientation and force. Experimental findings by many studies have shown that all these indicators can be captured by kinematic sensors [23], Table 1). It is even possible to evaluate the smooth motion that is normally violated by jerky motion, tremor, and hesitant motion by incorporating the effect of motion in both time and frequency domains [24].

In our earlier study, we offered to use three-dimensional kinematic data instead of two dimensional motion data from one camera video recording setup to develop a model for rank-based assessment of skills for robot-assisted surgery to overcome current limitations [17]. The preliminary version of the model was based on a Siamese LSTM network fed by two multi-variate time-series kinematic datasets to be compared. The model does not use any direct features, but instead, it uses raw motion signals to extract deep features to represent pairwise ranks.

In this study, we extend our previous work [17] in three ways. First, the model is significantly enhanced by adapting an attention mechanism to the LSTM, and a processing step, which calculates the Piecewise Aggregate Approximation (PAA) of input kinematic data to ease parameter optimization of the whole Siamese network. We show that these enhancements significantly improve the prediction accuracy. Second, we offer an approach that uses pairwise ranks of a query action against a set of reference actions as features to train a regression model. This allows the pairwise ranking model to be turned into an exact skill prediction model when needed. Third, we demonstrate that our model can serve as solution for the third objective of skill assessment, i.e. monitoring of surgeon's own progress. To the best of our knowledge, this is the first study that reports an empirical result in that respect.

The new model was first tested on the JIGSAWS dataset to compare it with previous methods. According to the results, our model can significantly improve the state-of-the-art in both ranking and regression tasks for computerized surgical skill assessments. Further, the model was evaluated for monitoring tasks in a larger and more recent dataset, called ROSMA [18]. The results show that our model can achieve reasonably good accuracy.

Methods

Pairwise ranking model

The surgical skill assessment problem is considered as a pairwise comparison task. We compare a query surgical action (m) with a reference action (n) to infer if the query is performed better than the reference. Semantically, the reference may refer to a previous action of the same surgeon to monitor the skill improvement, or to an action performed by another surgeon to make a skill comparison for better assignment to a surgery. While the model is formally the same, it can be used in any semantic model based on how the model parameters are trained from available data.

The kinematic data of two actions with length K and L are denoted by $\boldsymbol{x^m} = x_1^m x_2^m \ldots x_K^m$ and $\boldsymbol{x^n}$=$x_1^n x_2^n \ldots x_L^n$, respectively, x_i^m refers to a set of kinematics measurements at time i. A kinematic measurement can be position, angular velocity, gripper angle or any other motion-specific identifier of a particular hand at a given time point.

Rank-based assessment can be defined as determining which surgical action is performed with better skill. The output of the model is referred by p_{mn}, which is interpreted as the probability of the query surgical action being performed better than the reference;

$$p_{mn} = \begin{cases} 1 & m\ performs\ better\ than\ n \\ 0.5 & m\ and\ n\ show\ equal\ performance \\ 0 & n\ performs\ better\ than\ m \end{cases} \qquad (1)$$

Next, the goal is to train a model that minimizes the probabilistic loss in a set of samples annotated by experts. The model assumes that the annotations of the exact skill levels are not provided but all pairs are labeled by their pairwise rank for their surgical skills by experts.

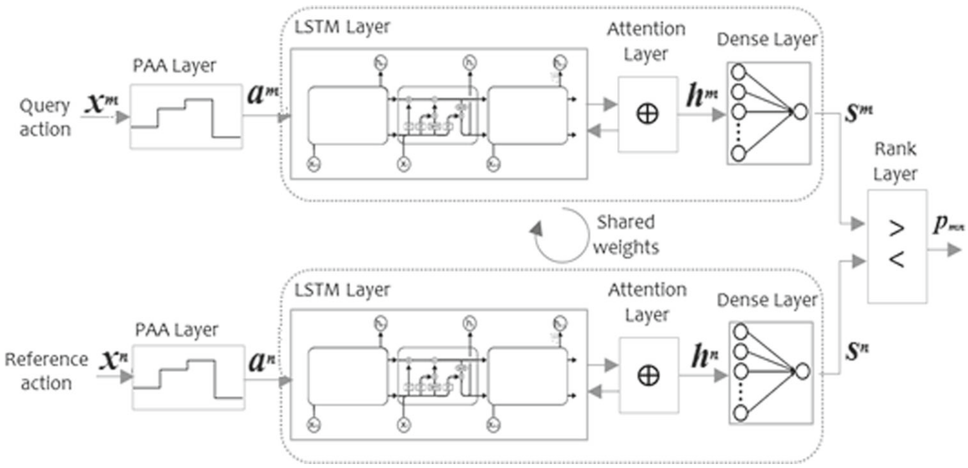


Fig. 1 General framework for pairwise ranking of surgery actions

The general framework that we introduce is based on a Siamese network of attention-enhanced LSTM integrated with a probabilistic ranking layer. The framework involves an essential pre-processing step for kinematic input based on PAA. (Fig. 1).

Piecewise aggregation of kinematic data

The action model based on attention-enhanced LSTM has an excessive number of parameters to be optimized in training phase (See 2.3). On the other hand, the kinematic data in our problem has a high dimensionality as opposed to the small number of samples in available datasets. This will lead to slow and insufficient learning of the model parameters in the proposed framework. To overcome this issue, we offer a pre-processing step based on PAA to reduce the dimensionality of the input signal while preserving the content that is representative for the skill level. PAA approximates a one-dimensional time-series kinematic signal x of length p into a of arbitrary length $q < p$, where each a_i is calculated by;

$$\boldsymbol{a}_i = \frac{\boldsymbol{q}}{\boldsymbol{p}}\sum_{j=\frac{p}{q}(\boldsymbol{i}-1)+1}^{\left(\frac{p}{q}\right)\boldsymbol{i}} \boldsymbol{x}_j \tag{2}$$

This approximation results with the reduction of the dimensionality of the kinematic signal by splitting it into equal-sized segments which are calculated by taking the average values in each segment. We apply PAA for each motion variable independently to get a smoother multi-variate signal at the input of the Siamese network.

Modeling action: attention-enhanced LSTM

Both query and reference actions are pre-processed using PAA and given into different inputs of Siamese network. The pre-processed kinematic data, given in the form of multi-variate time-series, is used to feed an LSTM network at each stream:

$$h_t = LSTM(h_t - 1, a_t) \tag{3}$$

where a_t and h_t are the input vectors at time t, where the superscript defining the stream is ignored. The LSTM model is parameterized by output, input and forget gates, controlling the information flow within the recursive operation. Given i_t represents input gate, f_t represents forget gate and o_t represents the output gate at time point t, the following equations formally describe the LSTM function:

$$i_t = \sigma\left(W_i a_t + U_i h_{t-1} + b_i\right) \tag{4}$$

$$f_t = \sigma\left(W_f a_t + U_f h_{t-1} + b_f\right) \tag{5}$$

$$o_t = \sigma(W_o a_t + U_o h_{t-1} + b_o) \tag{6}$$

$$\tilde{c}_t = \tanh(W_c a_t + U_c h_{t-1} + b_c) \tag{7}$$

$$c_t = \sigma(i_t^{\circ}\tilde{c}_t + f_t^{\circ}c_{t-1}) \tag{8}$$

$$h_t = o_t^{\circ}\tanh(c_t) \tag{9}$$

Here, c_t is cell state and $\tilde{c}_t$ represents a candidate for cell state at t. W_x, U_x and b_x are weights and biases for gate x, respectively. Finally, σ refers to sigmoid function. At every time step t, LSTM outputs a hidden vector h_t that reflects the skill representation of the kinematic motion at time point t. In our application, we used a bidirectional version of LSTM [20] to allow the modeling of two-way temporal dependencies in actions.

The LSTM layer is enhanced by an attention mechanism, which helps maximizing the contribution of the relevant encoding context vectors and minimize those of irrelevant vectors while building the decoding context [25]. The attention layer that we implement uses an attention function to assign weight to each hidden state produced by the LSTM layer. The weighted distribution of hidden states is used as a new representation of input signals. We calculate an attention function for each hidden state h_t, $t = 1,\ldots,T$, as follows;

$$u_t = \tanh(W_s h_i + b) \tag{10}$$

where *Ws* is an attention hidden weight matrix and b is a bias parameter. From this function, softmax weights are calculated by;

$$\alpha_t = \frac{\exp(u_t)}{\sum_{t'=1}^{T} \exp(u_{t'})} \tag{11}$$

These are used to produce a context vector *c*, which will be forwarded to the next layer:

$$c = \sum\nolimits_{t=1}^{T} h_t \alpha_t \tag{12}$$

The attention-enhanced LSTM layer is followed by a fully connected layer fed by the vector of skill representation, c^m for any of the input *m*. This layer transforms skill representations of query and reference actions into scalars, s^m and s^n, to make them explicitly comparable.

Ranking loss

We adapt a probabilistic loss function for model learning, which was originally introduced to learn how to rank text objects using a gradient descent approach [19]. A probabilistic rank layer is built such that skill equivalence is taken into account. We denote the posterior probability distribution $P_{ij} = P(i > j)$, where › refers to the skill superiority of *i* to *j* and let $\overline{P}_{ij}$ be the desired target values for those posteriors, such that $\overline{P}_{ij} \in \{1, 0.5, 0\}$. The goal is then to minimize the distance between these two entities. We use a cross entropy cost function, C_{ij} to measure the closeness between two probability distributions, given by,

$$C_{ij} = -\overline{P}_{ij} o_{ij} + \log(1 + e^{o_{ij}}) \tag{13}$$

where $o_{ij} = (s^i - s^j)$, i.e. is the difference between rank orders of *i* and *j*, Then, the Siamese network parameters are inferred by minimizing this loss for all (i, j) trial pairs in the training data.

Results and discussion

Data

The performance of the entire model was evaluated in two different publicly available surgery data sets obtained from the da Vinci robot systems. They can provide both three-dimensional kinematic data and stereo video of surgery tasks. The kinematic data contain variables of both master and slave's left and right manipulators. The kinematic data for each sample is considered as a multi-variate time series, in which each variable corresponds to a different motion-specific parameter.

JIGSAW [11], is a common benchmark dataset in the field. It has surgical data collected from eight subjects with different skill levels performing three different surgical tasks. The tasks are 'throw suturing', 'needle passing', and 'knot tying' performed on benchtop training phantoms. The data consist of 76 motion variables collected at 30 Hz, including tooltip positions and orientation, linear and rotational velocities, and gripper angle. A trial is a part of the data set that corresponds to one subject performing one instance of a specific task. Each subject is categorized by a fixed expertise level but each trial may have a different skill score. This score is annotated using the global rating score.

ROSMA [18] was recently released to facilitate the research in the field. It contains more samples and longer actions compared with JIGSAWS. Twelve subjects operated the da Vinci Research Kit to perform three different surgery tasks: post and sleeve, pea on a peg and wire chaser. The twelve subjects attempted each of the surgical task 4–6 different times to a total of 207 trials. The obtained dataset includes all the kinematic and dynamic information provided by the da Vinci robot (both master and slave side). A board of human experts defined an objective performance scale by introducing penalty points for each surgery task. Then, each trial (subject + task) was given a score based on penalty points and completion time in seconds.

Using JIGSAW and ROSMA data, we performed experiments in three different evaluation setups for (1) pairwise ranking of different surgeons, (2) regression to predict the exact skill level, and (3) monitoring of individual skill. For setups 1 and 3, we identified pairwise ranking labels from the exact scores, which were not used in any stage of the proposed system later. Therefore, the model mimics the approach where all assessments were performed in a pairwise manner.

Evaluation 1: ranking

We aim first to evaluate our framework in a common setup to justify our own model parameters and to benchmark against current state-of-the-art for pairwise ranking. To this end, we

Table 2 Conditions for correct predictions of pairwise ranking

Ranking type	p_{mn}	Ground truth
Ternary	$\geq 0.5 + \varepsilon$	$m > n$
	$\geq 0.5 - \varepsilon$ and $< 0.5 + \varepsilon$	$m \equiv n$
	$< 0.5 - \varepsilon$	$m < n$
Binary	≥ 0.5	$m > n$
	< 0.5	$m < n$

Table 3 Results of pairwise ranking with the present framework

Surgery type	Ternary ranking (including skill equivalence)	Binary ranking (excluding skill equivalence)
	Acc	Acc
Knot tying	79.2	83.65
Needle passing	78.87	82.48
Suturing	69.29	72.89
AVG	75.8	79.67

built an experimental setup that performed a fourfold cross validation to evaluate the prediction performance. In this setup, the pairs between 3/4 of the surgery actions were used for training and the remaining pairs were used for testing. As suggested by [15], the folds were organized such that the test samples included both the pairs where neither action has been used in a pair for training and the pairs where the other action was used for training in a different pairing. This guarantees that all possible pairs were tested after four folds of an experiment. The model performance is discerned using pairwise ranking accuracy, which is the percentage of correctly ordered pairs, produced by each testing fold. This scheme reports two different accuracy results for the cases where the skill equivalence is considered and where it is not. When skill equivalence is considered, the accuracy gives the evaluation of ternary ranking performance. Otherwise, it evaluates the binary ranking. Table 2 lists the conditions of correct ordering of a pair (m,n) in binary and ternary cases. We used $\varepsilon = 0.01$ in our evaluations.

We applied our model for each surgery task separately to rank surgery actions by their skills. We used the following hyper-parameters for the learning step by a stochastic gradient descent algorithm: a learning rate of 0.001, a batch size of 2 and a unit size of 64 with single hidden layer. Table 3 discerns the accuracy for each task for ternary and binary ranking.

Figure 2 shows Receiver Operating Characteristic (ROC) curve for the proposed model when applied for binary pairwise ranking. The ROC curve depicts the performance of the model is also discerned when the attention layer is removed. The figure shows that the attention enhancement has a significant contribution for the prediction performance. Reported ranking accuracy decreased to 74.64% when attention mechanism is eliminated. The contribution PAA step is also shown in the figure. The PAA can boost the prediction accuracy around 74%.

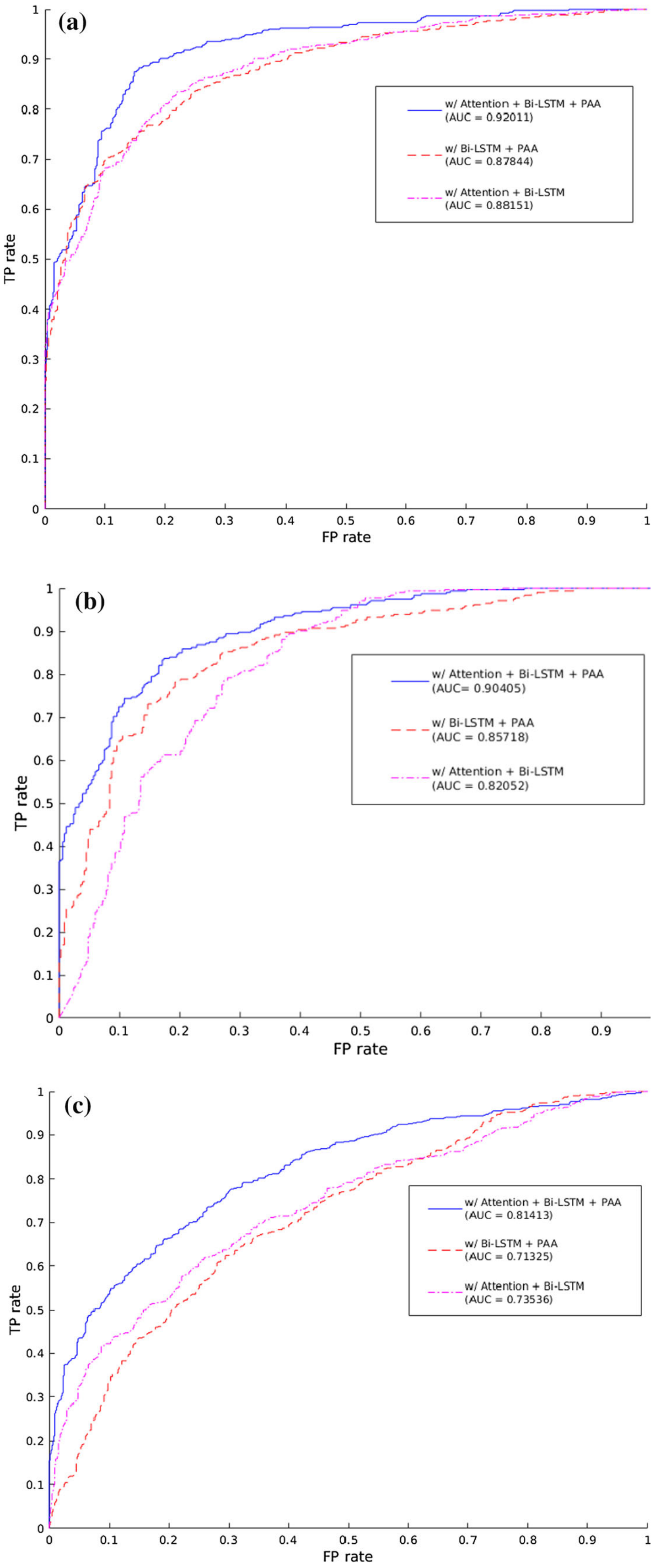
Although kinematic data is a multi-variate signal with so many sensory measurements, it involves two main characteristic channels. One represents the changes in the position of the arms and the other refers to varying velocity over time. To understand the contribution of these two characteristics, we run binary ranking experiments with positional features and velocity features separately. The experiments revealed that the binary ranking accuracies with positional characteristics are 77.33%, 74.99% and 71.55% for knot tying, needle passing and suturing, respectively. With velocity characteristics only, the model can achieve the accuracies of 71.95%, 67.88% and 66.84% for the same tasks. According to the results, positional features contribute more on ranking performance for all tasks, however, the integration of velocity features improves the final accuracy.

The present model was compared with three most relevant studies in the literature. Two of them used video data for skill ranking and tested their methods in the same dataset. The third study is our own preliminary model on kinematic data presented in [17]. Video-based methods work for only binary ranking cases since their loss function did not support the evaluation of equivalence in skills. They did not give accuracies separately for each task, but rather reported overall performance in surgery dataset. To make a comparison with these methods we ran our model with a subset of the original data in which the equally rated pairs were removed. We calculated the average of accuracies achieved with three surgery types.

The results are shown in Table 4. Our model can significantly outperform both video-based methods and the kinematic-based method in terms of pairwise ranking accuracy. Moreover, the present model built upon kinematic data reduces the computational resource requirements compared to approaches which use video recordings. Doughty et al. [15] reported that average running time to train a single fold is 18 h with NVIDIA TITANX GPU, whereas learning a fold in our model is conducted in less than an hour with a conventional CPU.

Table 5 shows the results of the same architecture on ROSMA dataset. This performance is also consistent with the results of pairwise rankings that we obtained in first dataset, which therefore constitutes a validation of our model in an independent dataset.

Fig. 2 ROC curves for binary ranking for surgical skill assessment for **a** knot tying, **b** needle passing, **c** suturing



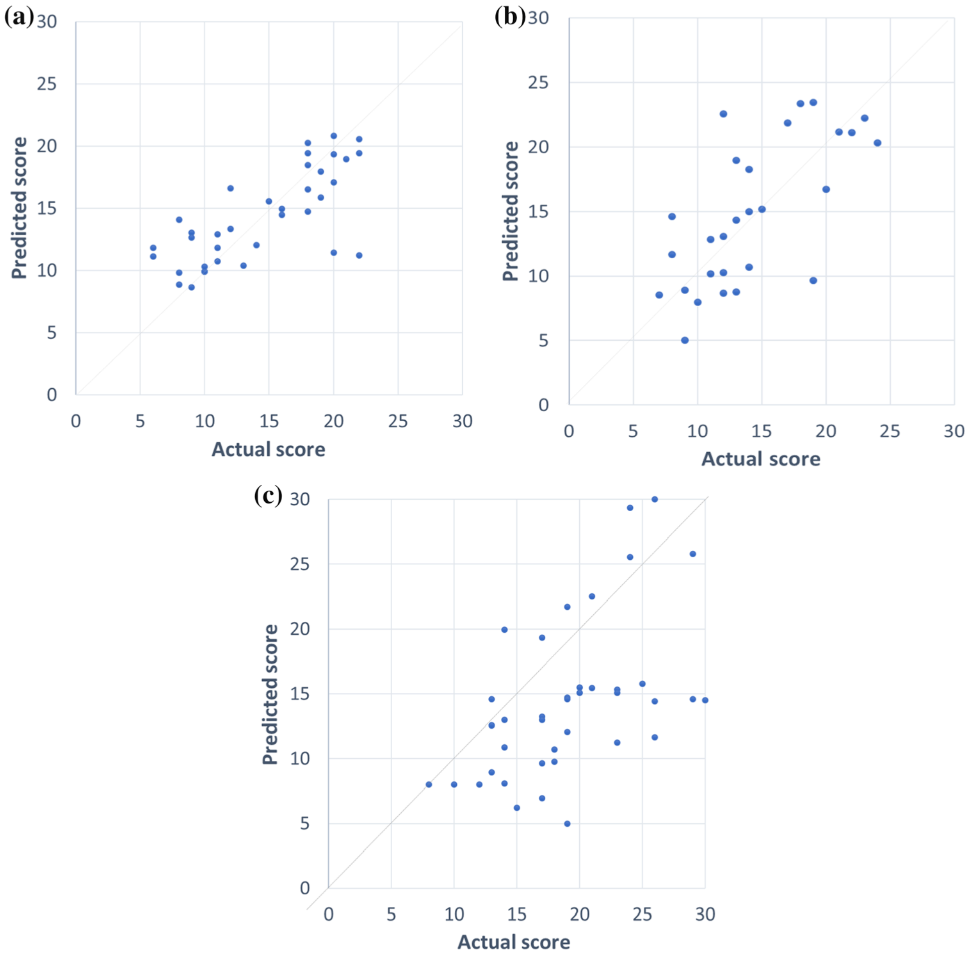


Fig. 3 Scatter plots for predicted skill scores vs actual scores for the tasks of **a** knot tying, **b** needle passing, and **c** suturing

Table 4 Results of pairwise ranking excluding skill equivalence

Method	Action data	Surgery type	Accuracy (%)
Doughty et al. [15]	Video	–	70.1
Li et al. [16]	Video	–	73.1
Ogul et al. [17]	Kinematic	Knot tying	79.6
		Needle passing	77.5
		Suturing	63.5
		Average	73.5
Present study	Kinematic	Knot tying	83.7
		Needle passing	82.5
		Suturing	72.9
		Average	79.7

Evaluation 2: regression

We argue that the results of pairwise rankings can be used for prediction of the exact score of surgical skill. To do this, we offer a method which could translate a list of pairwise ranks into an exact score of skill level. The conventional way of regression involves extracting a number of features from input signals to represent the sample in a machine learning model. Instead, we use an empirical representation where each feature refers to the pairwise rank between the query sample and another sample from a reference list. A pairwise rank here refers to the probability of the query action being performed better than the corresponding sample in a reference list. Figure 3 shows the results of exact value predictions as the comparison of predicted scores against actual scores for each task.

Table 5 Results of pairwise ranking with present framework on ROSMA

Action	Acc
Wire chaser	75.6
Post and sleeve	75.1
Pee on a peg	74.9
AVG	75.2

Table 6 Comparison of regression models for surgical skill assessment in SCC

Method	Knot tying	Needle passing	Suturing
Zia and Essa [5]	0.66	0.45	0.59
Fawas et al. [7]	0.65	0.57	0.60
Present method (with actual ranks)	0.99	0.99	0.99
Present method (with predicted ranks)	0.71	0.65	0.59

In the regression setup, the performance of predictions was evaluated using Spearman's Correlation Coefficient (SCC) between actual and predicted values of skill levels, as suggested by [5, 7], two previous studies that adopted the idea of using regression for surgical skill assessment. We followed the same procedure to benchmark our method against these methods in the same dataset. SCC is a nonparametric metric that evaluates how well the relationship between two distributions can be described by a monotonic function. It is calculated by $1 - \frac{6\sum d_i}{n(n^2-1)}$, where d_i is the difference between the ranks of actual and predicted scores and n is the number of samples. Tenfold cross-validation was performed to measure the performance. The results are given in Table 6. This experiment validates that the pairwise ranking model could be turned into a regression model with increased performance. Pearson Correlation Coefficient were calculated as 0.74, 0.65 and 0.53 for knot tying, needle passing and suturing tasks, respectively.

The same framework was run with known pairwise ranks, instead of predicted ranks, to justify the idea that the ranks are appropriate features. As titled by "Present method (with actual ranks)" in the table, a regression performance can be achieved up to 0.99 in SCC with our model when we know the actual pairwise ranks.

Evaluation 3: monitoring

Our last objective is to demonstrate that the pairwise ranking model can be used for measuring the progress of a candidate surgeon during training activities. This demonstration is done using the ROSMA dataset, in which different trials are available from the same surgeon on the same surgery task. Instead of a typical k-fold cross-validation, we performed a leave-user-out (LUO) procedure for testing. In this procedure, the trials of one user (surgeon) are left out for prediction, while all other pairs of the remaining trials on the same surgery task are used for training. This was repeated 12 times for each surgeon independently. Final, accuracy was determined by averaging the pairwise ranking accuracy of these folds. We used the following hyper-parameters for the learning step by a stochastic gradient descent algorithm: a learning rate of 0.001, a batch size of 2 and a unit size of 64 with single hidden layer. According to Table 7, our model achieved 70% pairwise ranking accuracy.

Table 7 Performance of our method in individual progress monitoring

Action	Ranking accuracy (%)
	Present method
Wire chaser	73.9
Post and sleeve	66.7
Pee on a peg	69.4
Average	70.0

Conclusion

A novel framework for objective skill assessment for robot-assisted surgery using kinematic data was introduced, that shall be used for choosing, credentialing and monitoring of surgeons. The framework including an attention-enhanced Siamese network with PAA, and was based on pairwise ranking, instead of classification or regression. The model provides a more interpretable and reliable view of skill assessment. The experimental results justify that this model can achieve better accuracy than the state-of-the-art methods in both ranking and regression setups for surgical skill assessment. Relative assessment approach offered in this study may help to overcome the limitations caused by inconsistencies in subjective skill grading scales that are used to train such machine-learning-based systems. Compared to video-based solutions, the use of kinematic data reduces the demands on computational power and is therefore a more applicable alternative for the practical implementation in a hospital setting.

To our knowledge, this is the first study that has considered and experimented the task of individual progress monitoring for surgical skills from a computational perspective. We describe how our model can be used in this context and validate it empirically in a recent dataset. The empirical results are promising; these results will serve as a strong baseline for future studies in monitoring task.

One of the limitations of the current study is the fact that reported pairwise rankings may violate triangular consistency, which will result in an unidentifiable full ranking of all actions. Although this information is not always requested in real—life surgeon trainings, considering the consistency in full ranking in the loss function may improve the prediction accuracy of the model. This is left for future work. The need for further validation of the ROSMA dataset with deeper statistical analysis challenges another future study. Another limitation is related to the kinematic data. Although kinematic data has an advantage over video data in capturing three-dimensional motion information, kinematic data does not contain contextual and semantic information such as the smoothness and strength of the movement, and the interaction between tools and tissue. Therefore, it may be a future direction to integrate video and kinematic data for more accurate ranking predictions with the expense of increasing computational costs. As a result, the assessment of surgical skill needs further investigation to perform in an objective way. Current progress in kinematic sensor data analysis is considered as a powerful complementary tool to manual assessment. It is reasonable to suggest that assessing surgical skill requires multiple simultaneous assessments, including machine-learning-based decision support systems as offered in the present study.

Acknowledgements Burçin Buket Oğul was financially supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under 2214-A—PhD Research Scholarship Program on Abroad.

Declarations

Conflict of interest The authors declare that there is no conflict of interest.

Ethical approval No human or animal study is involved.

References

1. Martin J, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, Brown M (1997) Objective structured assessment of technical skill (osats) for surgical residents. Br J Surg 84:273–278
2. van Hove PD, Tuijthof GJ, Verdaasdonk EG, Stassen LP, Dankelman J (2010) Objective assessment of technical surgical skills. Br J Surg 97(7):972–987
3. Rivas-Blanco I, Pérez-Del-Pulgar CJ, García-Morales I, Muñoz VF (2021) A review on deep learning in minimally invasive surgery. IEEE Access 9:48658–48678
4. Fard MJ, Ameri S, Darin Ellis R, Chinnam RB, Pandya AK, Klein MD (2018) Automated robot-assisted surgical skill evaluation: predictive analytics approach. Int J Med Robot Comput Assist Surg 14(1):e1850
5. Zia A, Essa I (2018) Automated surgical skill assessment in RMIS training. Int J Comput Assist Radiol Surg 13:731–739
6. Wang Z, Fey AM (2018) Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. Int J Comput Assist Radiol Surg 13:1959–1970
7. Fawaz H, Forestier G, Weber J et al (2019) (2019) Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks. Int J CARS 14:1611–1617
8. Zhang D et al (2020) Automatic microsurgical skill assessment based on cross-domain transfer learning. IEEE Robot Automat Lett 5(3):4148–4155
9. Funke I, Mees ST, Weitz J et al (2019) Video-based surgical skill assessment using 3D convolutional neural networks. Int J CARS 14:1217–1225
10. Nguyen XA, Ljuhar D, Pacilli M, Nataraja RM, Chauhan S (2019) Surgical skill levels: Classification and analysis using deep neural network model and motion signals. Comput Methods Programs Biomed 1(177):1–8
11. Gao Y, Vedula SS, Reiley CE, Ahmidi N, Varadarajan B, Lin HC, Tao L, Zappella L, Bejar B, Yuh DD et al (2014) JHU-ISI gesture and skill assessment working set (JIGSAWS): a surgical activity dataset for human motion modelling. Modeling and Monitoring of Computer Assisted Interventions (MICCAI) Workshop
12. Kelly JD, Petersen A, Lendvay TS et al (2020) Bidirectional long short-term memory for surgical skill classification of temporally segmented tasks. Int J CARS 15:2079–2088
13. Lavanchy JL, Zindel J, Kirtac K et al (2021) Automation of surgical skill assessment using a three-stage machine learning algorithm. Sci Rep 11:5197
14. Pérez-Escamirosa F, Alarcón-Paredes A, Alonso-Silverio GA et al (2020) Objective classification of psychomotor laparoscopic skills of surgeons based on three different approaches. Int J CARS 15:27–40
15. Doughty H, Damen D, Mayol-Cuevas W (2018) Who's better? who's best? pairwise deep ranking for skill determination. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR
16. Li Z, Huang Y, Cai M, Sato Y (2019) Manipulation-skill assessment from videos with spatial attention network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)
17. Oğul BB, Gilgien MF, Sahin PD (2019) Ranking robot-assisted surgery skills using kinematic sensors, In: 15th European Conference on Ambient Intelligence (AMI'19), pp 330–336
18. Rivas-Blanco I, Pérez-Del-Pulgar CJ, García-Morales I and Muñoz VF (2021) A surgical dataset from the da Vinci Research Kit for task automation and recognition, arXiv:2102.03643.
19. Burges CJ, Shaked T, Renshaw E, Lazier A, Deeds M, Hamilton N, Hullender G (2005) Learning to rank using gradient descent. In: International Conference in Machine Learning (ICML), pp 89–96
20. Graves A, Fernández S, Schmidhuber J (2005) Bidirectional LSTM networks for improved phoneme classification and recognition. Int Conf Artif Neural Netw (ICANN) 3697:799–804
21. Lin HC, Shafran I, Yuh D, Hager GD (2006) Towards automatic skill evaluation: detection and segmentation of robot-assisted surgical motions. Comput Aided Surg 11(5):220–230
22. Castillo-Segura P, Fernández-Panadero C, Alario-Hoyos C, Muñoz-Merino PJ, Kloos CD (2021) Objective and automated assessment of surgical technical skills with IoT systems: a systematic literature review. Artif Intell Med 112:102007
23. Mason JD, Ansell J, Warren N (2013) Torkington is motion analysis a valid tool for assessing laparoscopic skill? J Surg Endosc 27(5):1468–1477
24. Ghasemloonia A, Maddahi Y, Zareinia K, Lama S, Dort JC, Sutherland GR (2017) Surgical skill assessment using motion quality and smoothness. J Surg Educ 74(2):295–305
25. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473

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