Received: 2 October 2016 | Revised: 1 June 2017 | Accepted: 2 June 2017

DOI: 10.1002/rcs.1850

WILEY The International Journal of Medical Robotics and Computer Assisted Surgery

ORIGINAL ARTICLE

Automated robot-assisted surgical skill evaluation: Predictive analytics approach

Mahtab J. Fard[1] | Sattar Ameri[1] | R. Darin Ellis[1] | Ratna B. Chinnam[1] | Abhilash K. Pandya[2] | Michael D. Klein[3]

[1] Department of Industrial and Systems Engineering, Wayne State University, Detroit, Michigan, USA

[2] Department of Electrical and Computer Engineering, Wayne State University, Detroit, Michigan, USA

[3] Department of Surgery, Wayne State University School of Medicine and Pediatric Surgery, Children's Hospital of Michigan, Detroit, Michigan, USA

Correspondence
Mahtab J. Fard, 4815 Fourth St., Detroit, MI 48202, USA.
Email: fard@wayne.edu

Abstract

Background: Surgical skill assessment has predominantly been a subjective task. Recently, technological advances such as robot-assisted surgery have created great opportunities for objective surgical evaluation. In this paper, we introduce a predictive framework for objective skill assessment based on movement trajectory data. Our aim is to build a classification framework to automatically evaluate the performance of surgeons with different levels of expertise.

Methods: Eight global movement features are extracted from movement trajectory data captured by a da Vinci robot for surgeons with two levels of expertise – novice and expert. Three classification methods – *k*-nearest neighbours, logistic regression and support vector machines – are applied.

Results: The result shows that the proposed framework can classify surgeons' expertise as novice or expert with an accuracy of 82.3% for knot tying and 89.9% for a suturing task.

Conclusion: This study demonstrates and evaluates the ability of machine learning methods to automatically classify expert and novice surgeons using global movement features.

KEYWORDS

automated skill evaluation, global movement features, machine learning, robot-assisted surgery, skill assessment, surgeon dexterity

1 | INTRODUCTION

Despite advances in computer systems and simulation methods, surgical training is still based on direct observation involving expert surgeons.[1] These methods are limited by a lack of consistency, reliability and efficiency due to the subjective nature of human observation.[2] As the medical profession faces demands for greater accountability and patient safety, there is a critical need for to develop consistent and reliable methods in order to evaluate clinical performance objectively. Surgical training is undergoing a paradigm shift and the clinical competence of practising surgeons is a matter of growing public concern. More emphasis is being placed on competency-based training and earlier development of technical skills for new surgeons. Hence, training is based not only on the total time spent or subjective evaluation, but also on dexterity in various skills. Therefore, Objective Structured Assessment of Technical Skills (OSATS)[3] is being subjected to more structured techniques, which are still subjective and require the presence of an expert surgeon.

New technological innovations such as robotic surgery create great opportunities for automated objective skill assessment and a prompt feedback system, which was not previously available. Robotic surgical devices such as da Vinci (Intuitive Surgical, Sunnyvale, CA)[4] record motion and video data, enabling development of computational models to analyze surgical skills through data-driven approaches.[5] Techniques such as data mining and machine learning are likely to have a huge impact on ongoing studies of clinical decision support.[6] The ability of machine learning methods to uncover concealed patterns in a large dataset, such as kinematic and video data, offers the possibility to better understand and model surgical data in order to evaluate surgeons' skill and individualized training.[7] The key step is to extract meaningful features from quantitative motion data that explain the underlying pattern of surgeons' dexterity.

In this paper, we extend our previous work[8] by introducing new features to quantify smoothness, variability and complexity of the

The manuscript has been read and approved by all the authors.

surgical flow. The key differentiation between the proposed method and existing work is its accuracy, robustness and time efficiency. The proposed method also has the ability to automatically evaluate a surgeon's skill and provide prompt feedback to a trainee by comparing their surgical skills with other surgeons' using the comprehensive dataset. To our knowledge, this is the first time that such a study has been conducted in the area of robotic surgery skill evaluation using more advanced techniques such as machine learning methods. The proposed framework has the important characteristics of objective skill evaluation such as repeatability, stability and clinical relevance.

The rest of the paper is organized as follows. In Section 2, we provide background on methods of surgical skill evaluation. In Section 3, we present details of experimental methods and the proposed skill assessment framework. We demonstrate and discuss the performance of our method on real-world robotic surgery data in Sections 4 and 5. Finally, in Section 6, we conclude our paper with a summary of the main results and directions for possible future research.

2 | BACKGROUND

Surgical skill evaluation has been traditionally conducted by an expert observer via direct observation. This procedure is very time-consuming and can be unreliable.[9] Therefore, a range of structure-based techniques such as OSATS[3] and the Global Assessment of Laparoscopy Skills (GOALS)[10] have been introduced and validated. Using these methods of evaluation, the trainees perform standardized surgical tasks while an expert surgeon evaluates their performance. The expert surgeon gives scores to surgical trainees based on predefined criteria such as flow of surgery, time to motion, efficiency, tissue handling and overall performance by observing the surgical procedure in person or watching a recorded video of the operation.[11] However, these methods are threatened by a lack of consistency, reliability and efficiency due to the subjective nature of the expert's intervention. With the advent of minimally invasive surgery (MIS) and robot-assisted minimally invasive surgery (RMIS), the need for automated methods of objective surgical assessment is even more pressing because they require surgeons to undergo a much longer and more difficult training and pose new challenges for surgical training.[12]

Evaluation of surgical skills can be performed by utilizing two different modalities: decomposing surgical tasks into predefined surgical gestures, and comparing the experts' and novices' gestures to assess surgical skill or evaluate a surgeon's overall performance by defining competency metrics.[13-15] The former approach has been studied thoroughly for MIS skill evaluation using techniques such as hierarchical decomposition of surgical tasks,[16] the Hidden Markov Model (HMM)[17-19] and multivariate autoregressive modelling.[20] For RMIS, the HMM[21,22] and descriptive structure method[23] have been developed to assess surgeons' skill. Although these methods have the ability to find the structure underlying MIS or RMIS tasks, they are context-based and suffer from requiring a large number of training samples and complex parameter tuning, causing a lack of robustness in the results.[24]

While the first approach focuses on skill evaluation at a more granular level, the second approach, in contrast, evaluates overall surgical skill using global movement features (GMFs), and this is easier to implement and interpret.[25] Metrics such as operation time, speed, number of hand movements,[26] force and torque signatures,[27,28] path length and motion smoothness[25,29] have been used to identify the relation between the global features and basic psychomotor skills of experts and novices during MIS. However, the development of a quantitative method to automatically recognize a surgeon's skill level in robotic surgery has always lagged behind. Previous work built the foundation of objective surgical skill assessment,[30] but the current state-of-the-art has a few shortcomings. First, it mostly focuses on descriptive statistics, which is not an adequate measurement of proficiency to classify surgeons by their skill level. Second, robot-assisted surgery is a completely different technique than laparoscopic surgery, requiring new methods of training and evaluation. Finally, the ability to capture all movements during robotic surgery opens up new opportunities for automated surgical skill assessment, real-time feedback and individualized training by developing more advanced techniques such as machine learning algorithms (31).

Machine learning techniques have risen to prominence in many fields because of their advantages over traditional statistical methods, such as robustness, better predictive ability and higher tolerance of any violation of assumptions (e.g. normality or undependability of data),[31,32] but it is only recently that these methods have been considered to analyze RMIS tasks.[7,33] Thus, the development of quantitative classification techniques that can automatically and accurately evaluate surgical skills needs to be investigated. In this paper, we address the limitation of previous methods by introducing a skill assessment framework using global movement features and machine learning algorithms to automatically classify surgeons based on their expertise.

3 | MATERIALS AND METHODS

We implemented our model on real robot-assisted minimally invasive surgery data presented in.[34] This comprises eight surgeons with varying experience of robotic surgery who performed five trials of different surgical tasks including suturing and knot tying (Figure 1). We analyze kinematic data captured using the da Vinci robot for both right and left hands, which results in total of 160 data points. Data include a global rating score (GRS) for each trial. This score is from 1 to 5 for each of six criteria: respect for tissue, suture or needle handling, time and motion, flow of operation, overall performance and quality of final product. Therefore, the total score ranges between 6 and 30 for each surgical trial. Using the GRS for the knot tying and suturing tasks, thresholds of 15 and 19 are used respectively to divide surgeons into two skill levels: experts and novices. Hence, we try to solve a binary classification problem, which is very common in the data mining community.

The aim of this paper is to develop a predictive method for objective skill assessment based on the movement trajectory of the surgical robot arms. For this purpose, we quantify the surgical task by extracting GMFs from the raw motion data for each task. Based on the extracted features, different classifiers, including *k*-nearest neighbour, logistic regression and support vector machines have been applied. The classifier with the highest accuracy can be used to automatically predict the skill level of a surgeon.

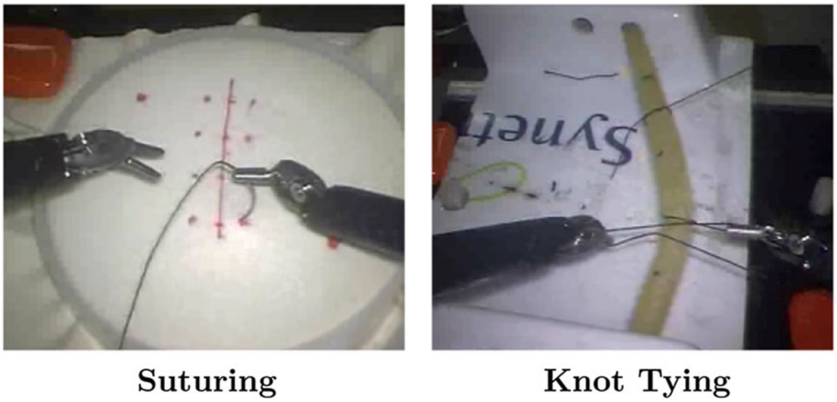

FIGURE 1 Snapshot of the two fundamental RMIS tasks

3.1 | Global movement features

Surgical tasks have different characteristics, such as smoothness, straightness or response orientation, which determine competence while relying only on instrument motion.[24] For instance, studies have shown that the tool motion of an experienced surgeon has more clearly defined features than that of a less experienced surgeon while performing the same task.[35] Figure 2 illustrates Cartesian position plots of an expert and a novice surgeon doing suturing on the da Vinci surgical robot.

In order to transform characteristics of a surgical task into quantitative metrics, we applied kinematic analysis theory that has been successfully used in previous works to study psychomotor skills.[25] Metrics such as task completion time, length of path, depth perception and velocity can show several aspects of a surgeon's dexterity. However, other aspects such as smoothness, curvature, torsion and complexity of the motion need to be quantified. In the following, we explain the six important global movement features from the clinical point of view and introduce our two new features, which measure the average turning angle and tortuosity of the task.

- **Task completion time:** defined as the total time required to complete the task, measured in seconds from the moment the surgeon starts to move the instrument until (s)he finishes the task and releases the robot device.
- **Path length:** the length of the curve described by the tip of the instrument while performing the task (in cm). We calculate it using the sum of the Euclidean distance between all consecutive pairs of points based on the (x, y, z) trajectory data captured during the surgery.
- **Depth perception:** the total distance travelled by the instrument along its axis (in cm).
- **Speed:** this can be defined as the magnitude of velocity and calculated as the rate of change of position from the previous time step as $\frac{\mathrm{dis}(p_i,\ p_{i-1})}{\Delta t_i}$ (in cm s^{-1}), where $\mathrm{dis}(p_i, p_{i-1})$ can be calculated as the

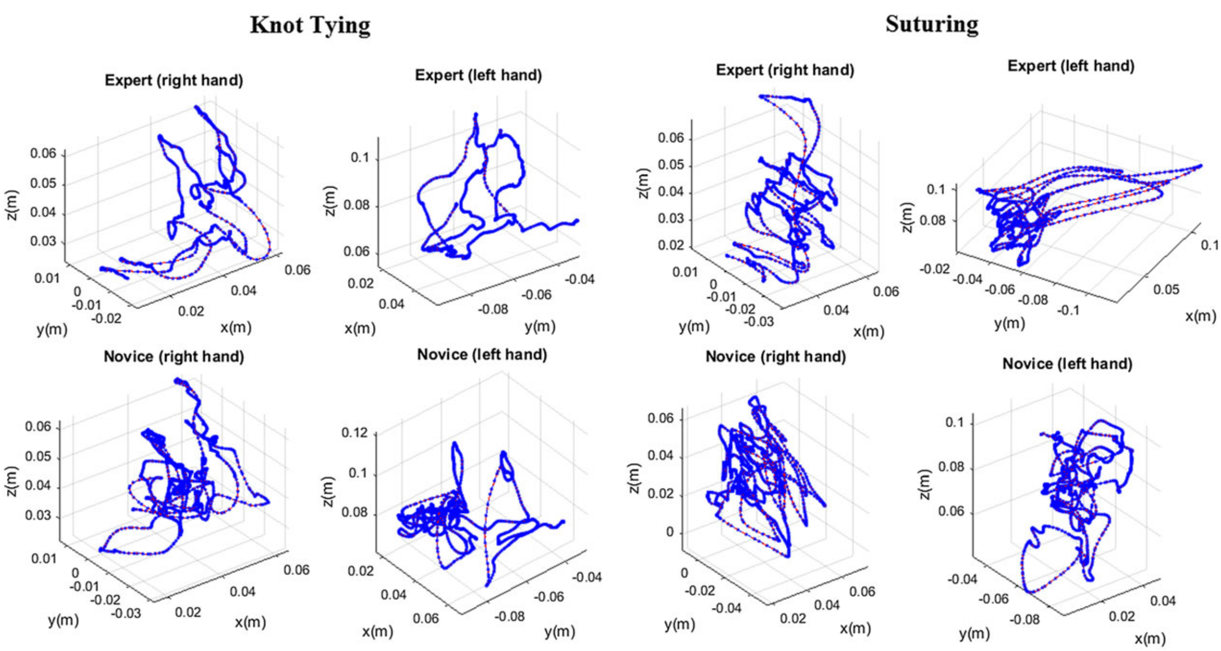

FIGURE 2 Illustration of the 3D Cartesian trajectory (blue line) where red arrows show the movement direction for an expert and a novice surgeon doing knot tying and suturing on the da Vinci surgical robot

Euclidean distance between the ith point and the (i – 1)th point. Given that the time difference between two consecutive frames is constant, Δt_i is equal to 1.

- **Motion smoothness:** a measure of the rhythmic pattern of acceleration and deceleration. Smoothness has most often been based on minimizing jerk, the third time derivative of position, which represents a change in acceleration (in cm s^{-3}).
- **Curvature:** a measure of the straightness of the path, calculated at each point from the following equation (25):

$$\kappa_i = \frac{v_i \times a_i}{v_i^3} \tag{1}$$

where v_i and a_i are the instantaneous velocity and acceleration of the instrument tips respectively, which can be calculated directly by computing the first and second derivatives of the positions of the instrument tips. For straight and smooth movement, the mean of curvature is close to zero, while larger values indicate curved and jerky movements.

- **Turning angle (TA):** calculated as the direction of the movement with regard to the previous and next time steps. It can be defined as

$$\mathrm{TA}_i = \theta_i = \arccos\left(\frac{u_{i-1} \bullet u_{i+1}}{\|u_{i-1}\|\|u_{i+1}\|}\right) \tag{2}$$

where $u_{i-1}u_{i-1}$ is the vector from p_{i-1} to p_i and u_{i+1} is the vector from p_i to p_{i+1} as shown in Figure 3.

- **Tortuosity:** a property of a curve being twisted or having many turns. It has been used successfully in a variety of research such as analyzing animal paths,[36] evaluating the performance of human–robot interaction[37] or distinguishing cognitive impairment through walking behavior.[38] In this paper, tortuosity is introduced as a new metric, to measure the complexity of a path or, in other words, the variability in movement path of a surgical instrument during robotic surgery. Tortuosity can be quantified using the fractal dimension (F)[39] where the ratio of a pattern changes with respect to the measurement scale, providing a statistical index for variability. More specifically, the length of the path is measured by walking a pair of dividers of a certain size along it (see Figure 4).

Tortuosity can be derived from the linear relationship between the logarithm of total distance (D) and the logarithm of the currently employed measuring scale (S) based on the knowledge that the total length is highly dependent on the scale adopted,[36] as follows:

$$\log(D_i) = b + a\log(S_i) \quad i \in [1,\ 2, \ldots, n] \tag{3}$$

where n is the number of different measurement scales employed to calculate the total distance of a trajectory, b is the regression intercept and a is the slope of the regression line. A regression model can be constructed from the (D_iD_i, S_i) pairs, and F is then calculated as $F = 1 + a$. The fractal dimension for movement paths lies between 1 and 2: F is 1 when the path is straight and 2 when the path is so tortuous that it occupies whole plane (Brownian motion). As an example, the path of an expert's right hand in Figure 2 has a tortuosity of 1.14 while that of a novice's right hand is 1.67. In order to make this metric more robust to different measuring scales, the average F value for different measuring scales is computed.[40]

We extract GMFs for both hands using the Cartesian positions of right and left patient-side manipulator end-effectors of da Vinci arms. More precisely, we only need (x, y, z) trajectory data for both hands to derive the proposed metrics. Speed, motion smoothness, turning angle and curvature are temporal features, which are calculated for each data point. The mean and standard deviation of these features are derived for each trial. On the other hand, remaining features such as task completion time, path length, depth perception and tortuosity have only one value for each trial. For instance, we measure the time needed to complete a surgical trial from start to finish and report it as task completion time. Finally, a total of 23 global movement features are derived from each trajectory: 12 spatial characteristics of tool tip movement (including path length, depth perception, mean and standard deviation of speed and motion smoothness for each hand) and 10 features that capture the curvature and torsion of movement (including tortuosity, mean and standard deviation of turning angle and curvature for each hand) and time to complete the task.

3.2 | Surgical skill classification

Features extracted in the previous section are used to quantify the movement pattern of surgeons. Our aim is to build a classification model to differentiate between surgeons with different levels of expertise while doing RMIS tasks. Surgeons are categorized into two

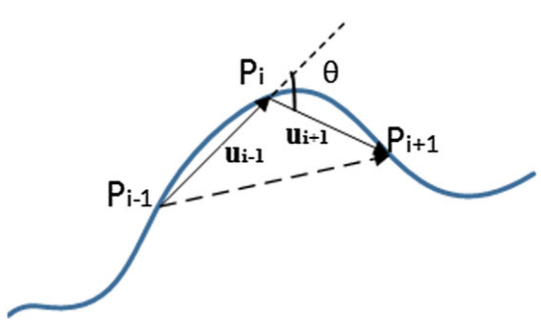

FIGURE 3 Illustrative example for computing movement trajectory features

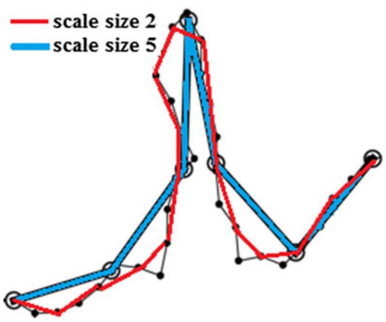

FIGURE 4 Example for measuring tortuosity by walking a pair of dividers of a certain size along the path

skill levels – expert and novice – resulting in a binary classification problem that can be solved by applying machine learning algorithms. Although there is no particular rule for choosing the best classification method there are various aspects to take into consideration.[41] Criteria such as the number of training examples, dimensionality of the feature space, independence of the features from each other, linear or nonlinear dependence between features and target, and overfitting play an important role when different classification methods are applied. Aiming for interpretability of the method, practical use and characteristic of the robotic surgery dataset, we applied and compared three frequently used and well-suited machine learning techniques: *k*-nearest neighbor (*k*-NN),[42] logistic regression (44) and support vector machine.[43]

3.2.1 | *k*-nearest neighbor (*k*-NN)

The first classifier that we used is *k*-nearest neighbor. The principle of this technique is to predict the label for the new point based on the closest distance to a predefined number (k) of training samples. The *k*-NN classifier is instance-based learning where instead of constructing a general model, it simply stores instances of training data. Therefore, it is a non-parametric classifier that does not rely on any assumptions about the underlying distribution of the data. This is a very important characteristic since most practical data do not conform to typical theoretical assumptions. During the classification phase, the majority of the k nearest neighbors for each point are computed. Thus, the label for the query point is assigned based on the most representatives within the nearest neighbors of the points. We examined different k and the best results obtained with $k = 3$.

3.2.2 | Logistic regression (LR)

One of the well-established statistical models is logistic regression where the dependent variable is categorical. Logistic regression is quite robust to noise in the dataset and avoids overfitting. In this model, logit transformation of a linear combination of features is used to resolve a binary classification problem.[47] Formally, the logistic regression model can be formalized as

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta x)}} \tag{4}$$

where β_0 is the intercept (often labelled the constant), β is the coefficient for the corresponding x feature and $p(x)$ is the probability of belonging to one of the classes.

3.2.3 | Support vector machine (SVM)

Support vector machine (SVM) constructs a hyperplane and tries to maximize the margin that separates two classes of data shows as $2/\|\vec{\omega}\|$ (Figure 5). The ability to learn the nonlinear separable function by mapping the data to a higher dimensional space makes this classifier unbeatable for some problems. Linear SVM can be formalized as

$$\begin{aligned} &\text{maximize} \frac{2}{\|\vec{\omega}\|} \\ &\text{s.t.} \quad y_i(\vec{w} \bullet \vec{x_i} + b) \geq 1, \quad \forall i = 1, \ldots, n \end{aligned} \tag{5}$$

where y_i is the class label for the ith datum. In order to solve the nonlinear classification problem, SVM uses a kernel transformation. The radial basis function (RBF) is one of the most popular kernel functions used in SVM,[44] and is defined as

$$K(x_i, x_j) = e^{-\left(-\gamma\|x_i - x_j\|^2\right)} \tag{6}$$

where γ controls the width of the function. The γ parameters can be seen as the inverse of the radius of influence of samples selected by the model as support vectors. If γ is too large, the radius of the area of influence of the support vectors only includes the support vector itself. For a very small value of γ, the model is too constrained and cannot capture the complexity or 'shape' of the data. The region of influence of any selected support vector would include the whole training set. It is suggested to choose γ as the inverse of the number of features: in this study we set $\gamma = 0.1$.

Another important parameter in the SVM algorithm is C, the penalty associated with instances that either are misclassified or violate the maximal margin. Therefore, equation 5 can be rewritten as

$$\begin{aligned} &\text{maximize} \frac{2}{\|\vec{\omega}\|} + C\sum_{i=1}^{n} \xi_i \\ &\text{s.t.} \quad y_i\left(\vec{w}.\vec{\phi}(x_i) + b\right) \geq 1 - \xi_i, \quad \forall i = 1, \ldots, n \\ &\xi_i \geq 1, \qquad \forall i = 1, \ldots, n \end{aligned} \tag{7}$$

where $\phi(x_i)^t \bullet \phi(x_j) = K(x_i, x_j)$. Here ξ_i is the smallest non-negative number satisfying $y_i(\vec{w} \bullet \vec{x_i} + b) \geq 1$ and C is a regularization term, which provides a way to control overfitting. If C becomes large, it is unattractive to respect the data at the cost of reducing the geometric margin, and on the other hand, when it is small, it is easy to account for some data points with the use of slack variables and to have a fat margin placed so that it models the bulk of the data. In this study we set $C = 1$.

3.3 | Performance evaluation

Classifier validation was conducted using two model validation schemas as suggested in[34]: leave-one-super-trial-out (LOSO), where one trial for each of the surgeons is left out for testing, and leave-one-user-out (LOUO), where all the trials from one surgeon are left out for testing. While the first validation method evaluates the robustness of a method for repetition of a task, the second schema evaluates the robustness of a method when a subject is not previously seen in

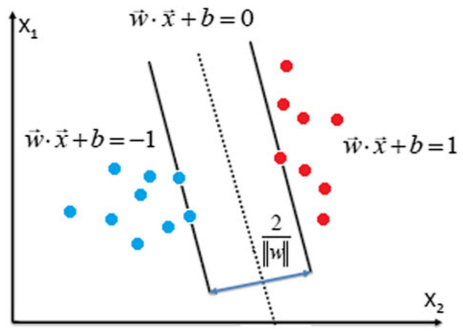

FIGURE 5 Illustration of the support vector machine (SVM) method for a binary classification with two features

the training data. The performance of the different classification methods was determined by classification accuracy, which is expressed as the percentage of surgeons whose skill level is correctly classified. We implement the three classification algorithms using the Weka software,[45] one of the best known open-source toolkits for data mining and machine learning.

4 | RESULTS

We start our skill assessment analysis by providing some exploratory statistics. Table 1 summarizes the mean and standard deviation for the eight GMFs extracted from RMIS motion trajectory data. From the table we observe that the basic descriptive statistic might not be an adequate measurement of proficiency for recognition of surgical skills. Box plots for each task are shown in Figures 6 and 7. The plots show that for some features such as turning angle and curvature in knot tying or tortuosity and smoothness in suturing, novices and experts can be distinguished. Table 2 shows 10 features most relevant to the skill level of surgeons performing RMIS tasks. Features are selected based on statistical significance ($p < 0.05$) and sorted according to Spearman's correlation coefficient ρ.

The results of performing three classification methods – *k*-NN, logistic regression and SVM – using spatially based features (S), curvature-based features (C) and a combination of the two based on two validation schemas – LOSO and LOUO – for knot tying and suturing are shown in Table 3. The best accuracy is obtained for the combination of all GMFs. Table 3 shows that for knot tying, the highest overall accuracy is 82.3% for LOSO and is 77.9% for LOUO. For suturing, the best overall accuracy is 89.9% in LOSO and 79.8% in LOUO.

Results also show that for knot tying, 86.4% of experts and 79.2% of novices are classified correctly in LOSO, while the classification accuracy for LOUO reduces respectively to 80.5% for experts and 75.3% for novices. For suturing, 95.2% of experts and 88.9% of novices are correctly classified using LOSO as a validation schema. For LOUO, we achieved accuracies of 81.2% for experts and 74.7% for novices.

TABLE 1 Mean and standard deviation of the eight extracted global movement features (GMFs) for both hands of expert and novice surgeons during knot tying and suturing tasks

Task		Path length (cm)	Depth perception (cm)	Speed (cm s^{-1})	Smoothness (10^{-7} cm s^{-3})	Turning angle (deg)	Curvature (10^{-4} cm^{-1})	Tortuosity	Time (s)
Knot tying	Experts	48.43 (8.65)	0.71 (0.13)	0.04 (0.013)	10.70 (8.66)	5.74 (1.49)	18.68 (6.33)	1.24 (0.05)	45.11 (11.44)
	Novices	50.37 (14.18)	0.73 (0.20)	0.02 (0.009)	8.50 (8.19)	3.86 (1.40)	11.01 (3.52)	1.25 (0.11)	69.86 (19.87)
Suturing	Experts	78.28 (23.49)	1.16 (0.41)	0.03 (0.011)	1.27 (1.23)	2.54 (0.50)	7.64 (2.34)	1.26 (0.04)	99.81 (20.82)
	Novices	75.99 (26.81)	1.14 (0.39)	0.02 (0.015)	3.30 (2.38)	2.59 (1.32)	6.70 (3.62)	1.33 (0.08)	126.84 (55.86)

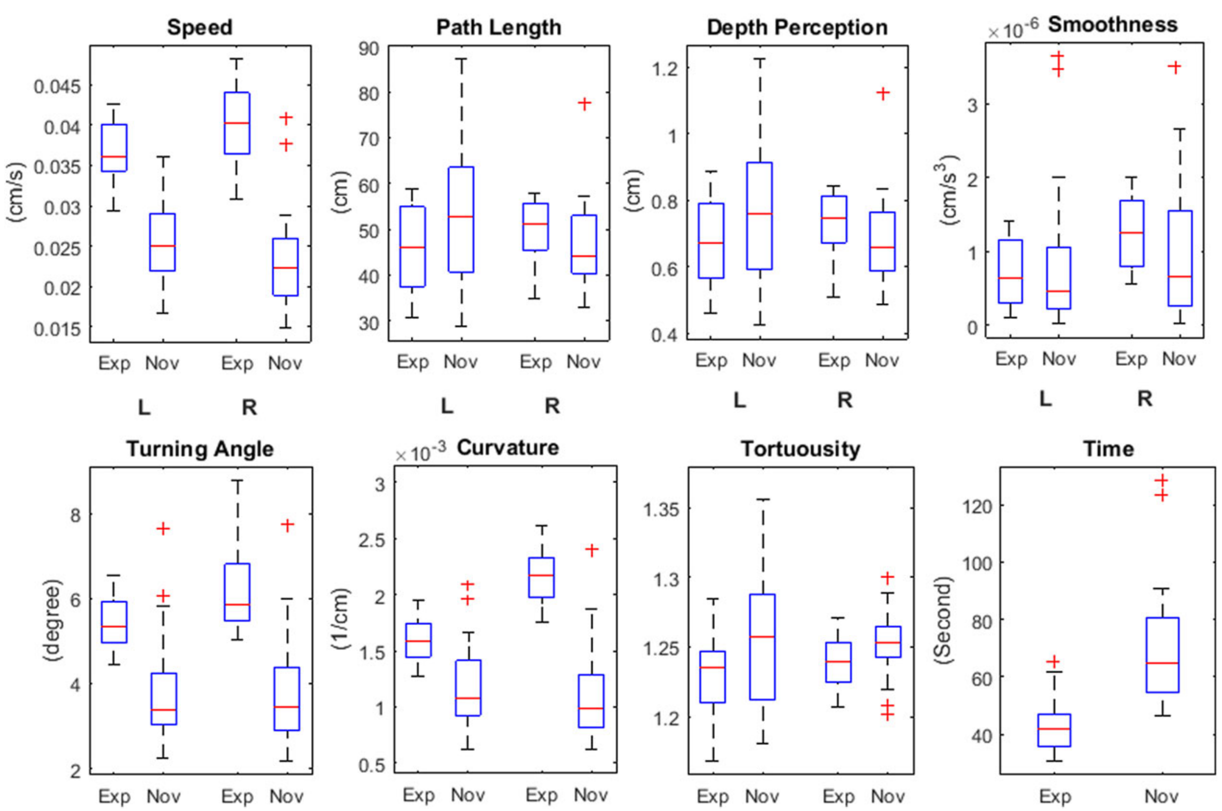

FIGURE 6 Box plots for experts (Exp) and novices (Nov) for eight GMFs during knot tying for the left hand (L) and right hand (R) of surgeons

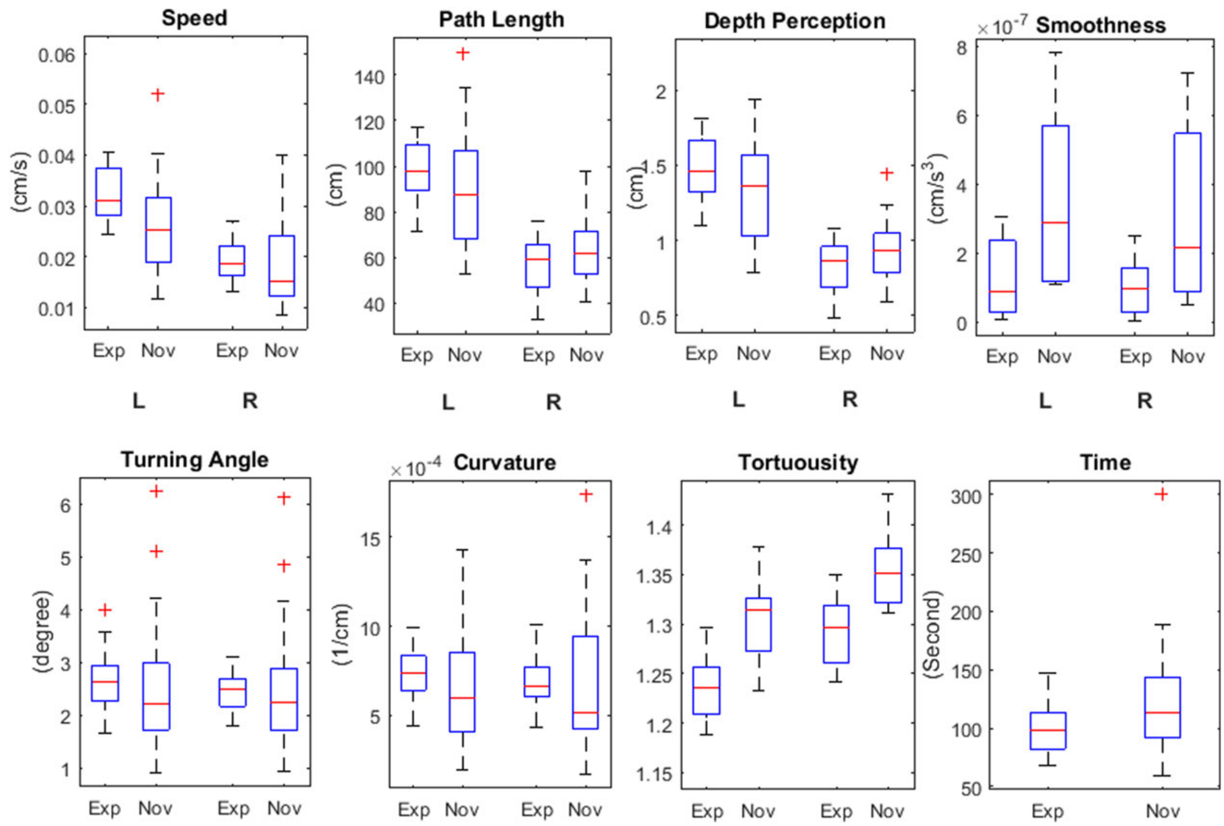

FIGURE 7 Box plots for experts (Exp) and novices (Nov) for eight GMFs during suturing for the left hand (L) and right hand (R) of surgeons

TABLE 2 Spearman's correlation coefficient |ρ| for the GMFs with the highest relevance to class label for different hands for knot tying and suturing

Knot tying		Suturing	
Feature, hand, stat	**|ρ|**	**Feature, hand, stat**	**|ρ|**
Turning angle, right, mean	0.45	Tortuosity, left, mean	0.47
Curvature, right, mean	0.43	Tortuosity, right, mean	0.44
Turning angle, left, mean	0.39	Time to complete task	0.43
Time to complete task	0.37	Curvature, right, mean	0.38
Speed, right, mean	0.29	Smoothness, left, mean	0.34
Tortuosity, right, mean	0.22	Smoothness, right, std	0.33
Speed, left, mean	0.19	Speed, left, std	0.28
Smoothness, right, std	0.17	Path length, left	0.27
Tortuosity, right, mean	0.14	Curvature, left, mean	0.21
Smoothness, left, std	0.12	Speed, right, std	0.15

5 | DISCUSSION

The results of descriptive analysis (Figures 6 and 7) illustrate several important aspects of surgical skill assessment. First, contrary to prior belief,[29] a higher (or equal) value of all global features describes a better performance. For instance, we observed that, on average, experts have a higher curvature than novices for both suturing and knot tying. This can be explained by looking at Figure 2, where an expert surgeon makes a decisive sharp turn with his left hand. This can be translated as the surgeon's skill in making the necessary curve in order to successfully finish the task. Also, for suturing, the path length of the left hand is longer for experts than for novices. This pattern may give the surgeon enough room for planning and performing further movements. In addition, although all the surgeons in this study are right-handed, Table 2 shows that features extracted from the non-dominant hand can be equally, if not more, important than those from the dominant hand. This is in complete agreement with the literature on skill acquisition where dexterity can be assessed based on non-dominant hand performance.[25]

From results shown in Table 3, the classification accuracy improves when a combination of spatial and curvature features are used. This is consistent with previous studies,[30] which emphasized that task completion time and distance travelled are insufficient to explain all aspects of surgical assessment. The results from this study clearly suggest that the proposed objective metrics for robotic surgery, such as curvature, turning angle and particularly tortuosity, can help further in distinguishing expert and novice surgeons. These features can be applied globally to RMIS tasks because they have the potential to identify additional aspects of a surgeon's dexterity that could not be quantified by task completion time and distance travelled alone.

Table 3 shows that the overall accuracy of skill classification decreases by 6% for knot tying and 11% for suturing when we switch from LOSO to LOUO schema. This suggests that surgeons with the same level of expertise perform knot tying in a more similar way than suturing. Furthermore, the best overall accuracy for LOSO is obtained either from logistic regression or *k*-NN, while SVM gives the best result for LOUO. Hence, a more sophisticated method such as SVM with a nonlinear kernel (e.g. RBF) is needed to assess the skill level of surgeons who are not previously seen in the training data. In other

TABLE 3 Classification accuracy for skill level evaluation in knot tying and suturing using *k*-NN, logistic regression (LR) and SVM for two validation schema (LOSO and LOUO) based on spatial motion features, curvature features and a combination of both (best accuracy is highlighted in bold and numbers are percentages)

Task	Validation schema	Features	NOVICES			EXPERTS			OVERALL		
			k-NN	LR	SVM	*k*-NN	LR	SVM	*k*-NN	LR	SVM
Knot tying	LOSO	S	72.2	72.2	63.3	77.8	72.2	70.9	75.1	72.2	62.6
		C	76.1	76.9	69.1	79.3	77.6	71.3	77.7	79.3	67.3
		S + C	75.3	**79.2**	71.1	**86.4**	85.4	77.7	82.1	**82.3**	75.4
	LOUO	S	65.7	66.0	65.1	66.2	68.2	74.2	66	67.1	69.6
		C	63	69.1	75.1	71.2	68.5	79.9	67.1	68.7	74.7
		S + C	69.5	68.7	**75.3**	76.3	71.6	**80.5**	72.9	70.2	**77.9**
Suturing	LOSO	S	66.7	72.2	65.0	85.7	85.7	67.9	76.9	79.5	65.3
		C	72.2	88.9	67.1	95.2	85.7	71.9	84.6	87.2	69.5
		S + C	83.3	**88.9**	69.3	**95.2**	90.5	78.7	89.7	**89.9**	75.4
	LOUO	S	63.9	66.9	64.2	68.3	73	70.5	66	69.7	67.1
		C	70.6	67.9	69.9	72.1	77.1	79.5	71.3	72.5	77.5
		S + C	68.7	69.7	**74.7**	75.0	78.9	**81.2**	71.9	74.4	**79.8**

words, SVM is more generalizable in the context of surgical skill evaluation than other classification methods. This result confirms the conclusion drawn from previous work in minimally invasive surgery[46] and has brought out the value of machine learning approaches such as SVM for more accurate surgical skill evaluation.

Our analyses also show that experts can be classified with higher accuracy than novices due to the stability (less variation) in the values of global movement features. It is also important to mention that the overall classification accuracy for suturing is higher than that for knot tying. This suggests that surgical skill levels are better distinguished in more complex tasks such as suturing. The reason could be the special characteristics of these tasks and also the need to follow specific procedure in order to finish them successfully. However, a larger dataset consisting of different surgical tasks and surgeons is needed to generalize this conclusion.

The time required to classify surgeons based on their skill using the proposed framework is only a few seconds. This stands in bold contrast with current state-of-the-art methods for surgical skill assessment, which have a very time-consuming process for parameter tuning and feature extraction. It suggests the potential of incorporating the proposed method for prompt feedback and evaluation of surgeons during training and individualized skill assessment while performing different robotic surgery tasks.

6 | CONCLUSION

In this paper, we have described the development of an automated objective skill assessment method based on global movement features extracted from movement trajectory data of the surgical robot arms. Previous attempts at objective surgical skill assessment have mostly been based on conventional statistical methods such as HMMs. However, robot-assisted surgical tasks have a specific complexity that cannot be modelled effectively unless more advanced methods are employed. Therefore, in this study we demonstrated the ability of machine learning methods to automatically distinguish between expert and novice performance during robotic surgery, where all movements are already digitized and available for analysis. It is generally accepted that the skill levels of surgeons vary and each surgical task has different levels of complexity. This complexity is captured not only through more sophisticated global features such as tortuosity, but also through more advanced machine learning methods to model the underlying pattern of surgical skill level. The results presented in this study could form a foundation for decision support tools that effectively, objectively and automatically evaluate a surgeon's dexterity and provide more personalized skill assessment and online feedback to trainees based on their performance. Furthermore, the proposed method can be applied on a more granular level of tasks in robot-assisted surgery, such as surgical gestures, to provide more insight into the surgeons' skill levels. Future research could focus on performing more validation studies with a larger number of participants. This would yield a larger training set that has the potential to improve the classification results.

AUTHOR JUSTIFICATIONS/CONTRIBUTIONS

Mahtab J. Fard: Worked on developing model, analysis of results and writing the manuscript

Sattar Ameri: Worked on developing model and analysis of results

R. Darin Ellis: Worked on analysis of results, and writing the manuscript

Ratna B. Chinnam: Worked on development of the model and analysis of results

Abhilash K. Pandya: Provided subject matter expertise on surgery for model development, analysis of results and implications for robotic interface design

Michael D. Klein: Provided subject matter expertise on surgery for model development, and analysis of results

REFERENCES

1. Grantcharov TP, Bardram L, Funch-Jensen P, Rosenberg J. Assessment of technical surgical skills. *Eur J Surg*. 2002;168(3):139-144.

2. Schout BMA, Hendrikx AJM, Scheele F, Bemelmans BLH, Scherpbier AJJA. Validation and implementation of surgical simulators: A critical review of present, past, and future. *Surg Endosc Other Interv Tech*. 2010;24(3):536-546.

3. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg*. 1997;84(2):273-278.
4. Guthart GS, Salisbury JK. The Intuitive telesurgery system: Overview and application. *Proc 2000 ICRA Millenn Conf IEEE Int Conf Robot Autom Symp Proc*. 2000;1:618-621.
5. Fard MJ, Pandya AK, Chinnam RB, Klein MD, Ellis RD. Distance-based time series classification approach for task recognition with application in surgical robot autonomy. *Int J Med Robotics Comput Assist Surg*. 2016. Available from: http://doi.wiley.com/10.1002/rcs.1766
6. Dreiseitl S, Binder M. Do physicians value decision support? A look at the effect of decision support systems on physician opinion. *Artif Intell Med*. 2005;33(1):25-30.
7. Kassahun Y, Yu B, Tibebu AT, et al. Surgical robotics beyond enhanced dexterity instrumentation: A survey of machine learning techniques and their role in intelligent and autonomous surgical actions. *Int J Comput Assist Radiol Surg*. 2016;11(4):553-568.
8. Fard MJ, Ameri S, Chinnam RB, Pandya AK, Klein MD, Ellis RD. Machine learning approach for skill evaluation in robotic-assisted surgery. In: *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2016 [Internet]*. San Francisco; 2016. Available from: http://arxiv.org/abs/1611.05136
9. Moorthy K, Munz Y, Sarker SK. Darzi A. *Clinical review: Objective assessment of technical skills in surgery*. 2003;1032-1037.
10. Gumbs AA, Hogle NJ, Fowler DL. Evaluation of resident laparoscopic performance using global operative assessment of laparoscopic skills. *J Am Coll Surg*. 2007;204(2):308-313.
11. Saggio G, Lazzaro A, Sbernini L, et al. Objective surgical skill assessment: An initial experience by means of a sensory glove paving the way to open surgery simulation? *J Surg Educ*. 2015 Sep;72(5):910-917.
12. Marescaux J, Leroy J, Gagner M, et al. Transatlantic robot-assisted telesurgery. *Nature*. 2001;413(6854):379-380.
13. Reiley CE, Lin HC, Yuh DD, Hager GD. Review of methods for objective surgical skill evaluation. *Surg Endosc*. 2011;25(2):356-366.
14. Fard MJ, Ameri S, Ellis RD. Toward personalized training and skill assessment in robotic minimally invasive surgery. In: *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2016 [Internet]*. San Francisco; 2016. Available from: http://arxiv.org/abs/1610.07245
15. Ahmidi N, Tao L, Sefati S, et al. A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. *IEEE Trans Biomed Eng*. 2017;1-1.
16. Mackenzie CL, Ibbotson JA, Cao CGL, Lomax AJ. Hierarchical decomposition of laparoscopic surgery: A human factors approach to investigating the operating room environment. *Minim Invasive Ther Allied Technol*. 2001;10(3):121-127.
17. Rosen J, Solazzo M, Hannaford B, Sinanan M. Task decomposition of laparoscopic surgery for objective evaluation of surgical residents' learning curve using hidden Markov model. *Comput Aided Surg*. 2002;7(1):49-61.
18. Leong JJH, Nicolaou M, Atallah L, Mylonas GP, Darzi AW, Yang G-Z. HMM assessment of quality of movement trajectory in laparoscopic surgery. *Comput Aided Surg*. 2007 Nov;12(6):335-346.
19. Megali G, Sinigaglia S, Tonet O, Dario P. Modelling and evaluation of surgical performance using hidden Markov models. *IEEE Trans Biomed Eng*. 2006;53(10):1911-1919.
20. Loukas C, Georgiou E. Multivariate autoregressive modeling of hand kinematics for laparoscopic skills assessment of surgical trainees. *IEEE Trans Biomed Eng*. 2011 Nov;58(11):3289-3297.
21. Varadarajan B, Reiley C, Lin H, Khudanpur S, Hager GD. Data-derived models for segmentation with application to surgical assessment and training. *Med Image Comput Comput Assist Interv*. 2009;12(1):426-434.
22. Tao L, Elhamifar E, Khudanpur S, Hager GD, Vidal R. Sparse hidden Markov models for surgical gesture classification and skill evaluation. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2012;7330 LNCS:167–77.
23. Ahmidi N, Poddar P, Jones JD, et al. Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty. *Int J Comput Assist Radiol Surg Springer Berlin Heidelberg*. 2015;10(6):981-991.
24. Mahtab JF. *Computational modeling approaches for task analysis in robotic-assisted surgery*. 2016.
25. Chmarra MK, Klein S, De Winter JCF, Jansen FW, Dankelman J. Objective classification of residents based on their psychomotor laparoscopic skills. *Surg Endosc Other Interv Tech*. 2010;24(5):1031-1039.
26. Datta V, Chang A, Mackay S, Darzi A. The relationship between motion analysis and surgical technical assessments. *Am J Surg*. 2002 Jul;184(1):70-73.
27. Rosen J, Hannaford B, Richards CG, Sinanan MN. Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills. *IEEE Trans Biomed Eng*. 2001;48(5):579-591.
28. Yamauchi Y, Yamashita J, Morikawa O, et al. *Surgical Skill Evaluation by Force Data for Endoscopic Sinus Surgery Training System*. Berlin Heidelberg: Springer; 2002:44-51.
29. Cotin S, Cotin S, Stylopoulos N, et al. Metrics for laparoscopic skills trainers: The weakest link! *MICCAI - Med Image Comput Comput Interv*. 2002;2488:35-43.
30. Judkins TN, Oleynikov D, Stergiou N. Objective evaluation of expert and novice performance during robotic surgical training tasks. *Surg Endosc*. 2009;23(3):590-597.
31. Murphy KP. *Machine Learning: A Probabilistic Perspective*. The MIT Press; 2012 Aug 24.
32. Fard MJ, Wang P, Chawla S, Reddy CK. A Bayesian perspective on early stage event prediction in longitudinal data. *IEEE Trans Knowl Data Eng*. 2016 Dec;28(12):3126-3139.
33. Fard MJ, Ameri S, Chinnam RB, Ellis RD. Soft boundary approach for unsupervised gesture segmentation in robotic-assisted surgery. *IEEE Robot Autom Lett IEEE*. 2017 Jan;2(1):171-178.
34. Gao Y, Vedula SS, Reiley CE, et al. JHU-ISI gesture and skill assessment working set (JIGSAWS): A surgical activity dataset for human motion modeling. *Model Monit Comput Assist Interv*. 2014;1-10.
35. Lin HC, Shafran I, Yuh D, Hager GD. Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions. *Comput Aided Surg*. 2006;11(5):220-230.
36. Nams VO. Using animal movement paths to measure response to spatial scale. *Oecologia*. 2005;143(2):179-188.
37. Voyles R, Bae J, Godzdanker R. The gestural joystick and the efficacy of the path tortuosity metric for human/robot interaction. Proc 8th Work Perform Metrics Intell Syst. 2008;91–97.
38. Kearns WD, Nams VO, Fozard JL. Tortuosity in movement paths is related to cognitive impairment. *Methods Inf Med*. 2010;49(6):592-598.
39. Mandelbrot B. How long is the coast of Britain? Statistical self-similarity and fractional dimension. *Science* (80-). 1967;156(3775):636-638.
40. Nams VO. Improving accuracy and precision in estimating fractal dimension of animal movement paths. *Acta Biotheor*. 2006;54(1):1-11.
41. Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? *J Mach Learn Res*. 2014;15:3133-3181.
42. Cover T, Hart P. Nearest neighbor pattern classification. *IEEE Trans Inf Theory*. 1967;13(1):21-27.
43. Saitta L. Support-vector networks. *Mach Learn*. 1995;20:273-297.
44. Vapnik V. *Statistical Learning Theory*. John Wiley & Sons; 1998.
45. Witten IH, Frank E. *Data Mining: Practical Machine Learning Tools and Techniques*. Second ed. Morgan Kaufmann Publishers Inc.; 2005 Jun 1.

46. Allen B, Nistor V, Dutson E, Carman G, Lewis C, Faloutsos P. Support vector machines improve the accuracy of evaluation for the performance of laparoscopic training tasks. *Surg Endosc*. 2010;24:170-178.

47. Kleinbaum DG, Klein M. *Logistic Regression*. New York: Springer; 2010.

How to cite this article: Fard MJ, Ameri S, Darin Ellis R, Chinnam RB, Pandya AK, Klein MD. Automated robot-assisted surgical skill evaluation: Predictive analytics approach. *Int J Med Robotics Comput Assist Surg*. 2018;14:e1850. https://doi.org/10.1002/rcs.1850