Deep Learning with Convolutional Neural Network for Objective Skill Evaluation in Robot-assisted Surgery

Ziheng Wang · Ann Majewicz Fey

Received: 10 January 2018 / Accepted: date

Abstract

Purpose: With the advent of robot-assisted surgery, the role of data-driven approaches to integrate statistics and machine learning is growing rapidly with prominent interests in objective surgical skill assessment. However, most existing work requires translating robot motion kinematics into intermediate features or gesture segments that are expensive to extract, lack efficiency, and require significant domain-specific knowledge.

Methods: We propose an analytical deep learning framework for skill assessment in surgical training. A deep convolutional neural network is implemented to map multivariate time series data of the motion kinematics to individual skill levels.

Results: We perform experiments on the public minimally invasive surgical robotic dataset, JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS). Our proposed learning model achieved a competitive accuracy of 94.1%, 90.3%, and 86.8%, in the standard training tasks: Suturing, Needle-passing, and Knot-tying, respectively. Without the need of engineered features or carefully-tuned gesture segmentation, our model can successfully decode skill information from raw motion profiles via end-to-end learning. Meanwhile, the proposed model is able to reliably interpret skills within 1-3 second window, without needing an observation of entire training trial.

Conclusion: This study highlights the potentials of deep architectures for an proficient online skill assessment in modern surgical training.

Keywords Surgical robotics, surgical skill evaluation, motion analysis, deep learning, convolutional neural network

Ziheng Wang
Department of Mechanical Engineering, The University of Texas at Dallas, Richardson, TX 75080, USA. E-mail: zihengwang@utdallas.edu

Ann Majewicz Fey
Department of Mechanical Engineering, The University of Texas at Dallas, Richardson, TX 75080, USA. Department of Surgery, UT Southwestern Medical Center, Dallas, TX 75390, USA.
1 INTRODUCTION

Due to the prominent demand for both quality and safety in surgery, it is essential for surgeon trainees to achieve required proficiency levels before operating on patients [1]. An absence of adequate training can significantly compromise the clinical outcome, which has been shown in numerous studies [2-4]. Effective training and reliable methods to assess surgical skills are thus critical in supporting trainees in technical skill acquisition [5]. Simultaneously, current surgical training is undergoing significant changes with a rapid uptake of minimally invasive robot-assisted surgery. However, despite advances of surgical technology, most assessments of trainee skills are still performed via outcome-based analysis [6], structured checklists, and rating scales [7-9]. Such assessment requires large amounts of expert monitoring and manual ratings, and can be inconsistent due to biases in human interpretations [10]. Considering the increasing attention to the efficiency and effectiveness of assessment and targeted feedback, conventional methods are no longer adequate in advanced surgery settings [11].

Modern robot-assisted surgical systems are able to collect a large amount of sensory data from surgical robots or simulators [12]. This high volume data could reveal valuable information related to the skills and proficiencies of the operator. However, analyzing such complex surgical data can be challenging. Specifically, surgical motion profiles by nature are nonlinear, non-stationary stochastic processes [13, 14] with large variability, both throughout a procedure, as well within repetitions of the same type of surgical task (e.g., suture throws) [15]. In addition, the high dimensionality of the data creates an additional challenge for accurate and robust skill assessments [10]. Further, although several surgical assessment methods have been developed, methods to autonomously coach the trainee are lacking. Towards this aim, there is a great need to develop techniques for quicker and more effective surgical skill acquisition [11, 16]. In this paper, we are particularly interested in online skill assessment methods that could pave the way for autonomous surgical coaching.

1.1 Previous Approaches in Objective Skill Assessment

Different objective skill assessment techniques have been reported in the literature [16]. Current approaches with a focus on surgical motions can be divided into two main categories: 1) descriptive statistic analysis, and 2) predictive modeling-based methods. Descriptive statistic analysis aims to compute features from motion observations to quantitatively describe skill levels. Specifically, summary features, such as movement time [17, 19], path length [17], motion jerk [18], curvature [17], etc., are widely used and have shown to have high correlations with surgical skills. Other novel measures of motion, such as energy expenditure [20], semantic labels [21], tool orientation [22], force [19], etc., can also provide discriminative information in measuring skills. However, this approach involves manual feature engineering, requires task-specific knowledge and significant effort to design optimal skill metrics [23]. In fact, defining the best metrics to capture adequate information and be generalized enough to apply across different types of surgery or groups of surgeons remains an open problem [16, 17, 23, 25].
In contrast to descriptive analysis, predictive modeling-based methods aim to predict surgical skills from motion data. This method can be further categorized into 1) descriptive, and 2) generative modeling. In descriptive modeling, models are learnt by transforming raw motion data to intermediate interpretations and summary features. Coupled with advanced feature selection, these pre-defined representations are subsequently fed into learning models as an input for skill assessment. In the literature, machine learning (ML) algorithms are commonly explored for modeling, such as k-nearest neighbors (kNN), logistic regression (LR), support vector machines (SVM), and linear discriminant analysis (LDA). Such algorithms yielded a skill predictive accuracy between 61.1% and 95.2% [26–29]. Forestier et al. developed a novel vector space model (VSM) to assess skills via learning from the bag of word, a collection of discretized local features (strings) obtained from motion data [30]. In [31], Brown et al. explored an ensemble approach, which combines multiple ML algorithms for modeling, and was able to predict rating scores with moderate accuracies (51.7% to 75.0%). More recently, Zia et al. utilized nearest neighbor (NN) classifiers with a novel feature fusion (texture-, frequency- and entropy-based features) and further improved skill assessment with accuracy ranging from 99.7% to 100% [32]. Although the descriptive modeling-based approaches show their validity in revealing skill patterns and underlying operation structures, the model accuracy and validity are typically limited by the quality of extracted features. Considering the complex nature of surgical motion profiles, critical information has the potential to be discarded within the feature extraction and selection process. Alternatively, in generative modeling, temporal motion data are usually segmented into a series of predefined rudimentary gestures for certain surgical tasks. Using generative modeling algorithms, such as Hidden Markov Model (HMM) and its variants, several class-specific skill models were trained for each level and achieved accuracy ranging from 94.4% to 100% [15, 33]. However, the segmentation of surgical gestures from surgeon motions can be a strenuous process. HMM models usually require large amounts of time and computational effort for parameter tuning and model development. Further, one typical deficiency is that the skill assessment is obtained at the global task level, i.e., at the end of each operation. It requires an entire observation for each trial. This drawback potentially undermines the goal for an efficient online surgical skill assessment.

1.2 Proposed Approach

Deep learning, also referred to as deep structured learning, is a set of learning methods that allow a machine to automatically process and learn from input data via hierarchical layers from low to high levels [34, 35]. These algorithms perform feature self-learning to progressively discover abstract representations during the training process. Due to its superiority in complex pattern recognition, this approach dramatically improves the state of the art. Currently, deep learning models have achieved success in strategic games [36], speech recognition [37], medical imaging [38], health informatics [39], and more. In the study of robotic surgical training, DiPietro et al. first apply deep learning based on Recurrent Neural Networks for gesture and high-level task recognition [40]. Still, relatively little work has been done to explore deep learning approaches for surgical skill assessment.
In this paper, we introduce and evaluate the applicability of deep learning for a proficient surgical skill assessment. Specifically, a novel analytical framework with deep surgical skill model is proposed to directly process multivariate time series via an automatic learning. We hypothesize the learning-based approach could help to explore the intrinsic motion characteristics for decoding skills and promote an optimal performance in online skill assessment systems. Fig 1 shows the end-to-end pipeline framework. Without performing manual feature extraction and selection, latent feature learning is automatically employed on multivariate motion data and directly outputs classifications. To validate our approach, we conduct experiments on the public robotic surgery dataset, JIGSAW [41], in analysis of three independent training tasks: Suturing (SU), Needle-passing (NP), and Knot-tying (KT). To the best of our knowledge, it is the first study to employ a deep architecture for an objective surgical skill analysis. The main contributions of this paper can be summarized as:

- An novel end-to-end analytical framework with deep learning for skill assessment based on high-level analysis of surgical motion.
- Experimental evaluation of our proposed deep skill model.
- Application of data augmentation leveraging the limitation of small-scale JIGSAWS dataset, and exploration of validation schemes applicable for deep-learning-based development.

In the remainder of this paper we first present our proposed approach and implementation details in Section 2. We then conduct experiments on JIGSAW dataset to validate the model in Section 3. Data pre-processing, training, and evaluation approaches are given. Then, we present our results in Section 4 and discussions in Section 5. Last, we conclude this paper in Section 6.

2 DEEP SURGICAL SKILL CLASSIFICATION MODEL

Our deep learning model for surgical skill assessment is motivated from studies in multiple domains [35, 42, 43]. In this section, we introduce a deep architecture using Convolutional Neural Network (CNN) to assess surgical skills from an end-to-end classification.

2.1 Problem Formulation

Here, the assessment of surgical skills is formalized as a supervised three-class classification problem, where the input is multivariate time series (MTS) of motion kinematics measured from surgical robot end-effectors, X, and the output is the predicted labels representing corresponding expertise levels of trainees, which can be one-hot encoded as y ∈ {1 : “Novice”, 2 : “Intermediate”, 3 : “Expert”}. Typically, ground-truth skill labels are acquired from expert ratings, crowdsourcing, or self-reporting experience. The objective cost function for training the network is defined as a multinomial cross-entropy cost, J, as shown in Eq. 1.
An end-to-end framework for online skill assessment in robot-assisted minimally-invasive surgery. The framework utilizes window sequences of multivariate motion data as an input, recorded from robot end-effectors, and outputs a discriminative assessment of surgical skills via a deep learning architecture.

\[ J(\theta) = -\sum_{i=1}^{m} \sum_{k=1}^{K} \mathbb{1}\{y^{(i)} = k\} \log p(y^{(i)} = k|x^{(i)}; \theta) \]  

where \( m \) is the total number of training examples, \( K \) is the class number, \( K = 3 \), and \( p(y^{(i)} = k|x^{(i)}; \theta) \) is the conditional likelihood that the predicted label \( y^{(i)} \) on a single training example \( x^{(i)} \) is assigned to class \( k \in K \), given specific trained model parameters \( \theta \).

2.2 Model Architecture

The architecture of the proposed neural network consists of five types of layers: convolutional layer, pooling layer, flatten layer, fully-connected layer and softmax layer. Fig. 2 shows a 10-layer working architecture and parameter settings used in the network. Note that, the length of the network is chosen after trial-and-error from the training/validation procedure.

The network takes the slide of length \( W \) from \( C \)-channel sensory measurement as input, which is a \( W \times C \) matrix, where \( C \) is the number of channels, or dimensions, of input time series. Then, input samples are first processed by three convolution-pooling (Conv-pool) stages, where each stage consists of a convolution layer and a
Fig. 2: Illustrations of the proposed deep architecture using a 10-layer convolutional neural network. The window width $W$ used in this example is 60. Starting from the inputs, this model consists of three Conv-pool stages with a convolution and max-pooling each, one flatten layer, two fully-connected layers, and one softmax layer for outputs. Note that the max-pooling dropout (with probability of 20%) and fully-connected dropout (with probability of 50%) is applied during training.

max-pooling layer. Each convolution layer has different numbers of kernels with the size of 2 and each kernel is convoluted with the input matrix of the layer with a stride of 1. Specifically, the first convolution ($Conv_1$) filters the $W \times 38$ input matrix with 38 kernels; the second convolution with 76 kernels ($Conv_2$) will filter the corresponding output matrix of previous layer; and the third convolutional layer ($Conv_3$) filters with 152 kernels. To reduce the dimensionality of the feature maps and avoid overfitting, corresponding connections of each convolution are followed by a max-pooling operation. The max-pooling operations take the output of convolution layer as input, and downsample the extracted feature maps, where each local input patch is replaced by taking the maximum value of each channel over the patch. The size of max-pooling is set as 2 with a stride of 2. In this network, we use the rectified linear unit (ReLU) as the activation function to add nonlinearity in all convolutional layers and fully-connected layers [44]. Finally, we apply a softmax logistic regression to produce a distribution of probability over three classes for the output layer.

2.3 Implementation

To implement the proposed architecture, the deep learning skill model is trained from scratch, which does not require any pre-trained model. The network algorithm is implemented using Keras library with Tensorflow backend based on Python 3.6. We first initialize parameters at each layer using the Xavier initialization method [35], where biases are initialized as zeros, the weights at each layer are initialized from a Gaussian distribution with mean 0 and a variance of $1/N$, where $N$ specifies the number of neurons in the previous layer.
During the optimization, our network is trained end-to-end by minimizing the multinomial cross-entropy cost between the predicted and ground-truth labels, as defined in Eq. 2 at the learning rate, $\epsilon$, of 0.0001. To train the net efficiently, we run mini-batch updates of gradient descent, which calculate network parameters on a subset of the training data at each iteration [45]. The size of mini batches is set to 600. A total of 300 epochs for training were run in this work. The network parameters are optimized by an Adam solver [46], which computes adaptive learning rates for each neuron parameter via estimates of first and second moment of the gradients. The exponential decay rates of the first and second moment estimates are set to 0.9 and 0.999, respectively. Also, to achieve better generalization and model performance, we apply a stochastic dropout regularization to our neural network during training time. Components of outputs from specific layers of networks are randomly dropped out at a specific probability [47]. This method has proven its effectiveness to reduce over-fitting in complex deep learning models [48]. In this study, we implement two strategies of dropout: one is the max-pooling dropout on the layers of max-pooling after ReLU non-linearity; another regularization is the fully-connected dropout on the fully-connected layers. The probabilities of dropout for the max-pooling and fully-connected dropout are set at 0.2 and 0.5, respectively. As mentioned above, the hyper-parameters used for CNN implementation include the learning rate, mini-batch size, epoch, number of filters, stride and size of kernel, and dropout rates in the max-pooling and fully-connected layers. These hyper-parameters are chosen and fine-tuned by employing the validation set, which is split from training data. We save the best model, as evaluated on validation data, in order to obtain an optimal prediction performance.

3 EXPERIMENT SETUP

3.1 Dataset

Our dataset comes from the JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS), the only public-available minimally invasive surgical database, which is collected from the da Vinci tele-robotic surgical system [41, 49].

The da Vinci robot is comprised of two master tool manipulators (MTMs) on left and right sides, two patient-sides slave manipulators (PSMs), and an endoscopic camera arm. Robot motion data are captured (sampling frequency 30 Hz) as multivariate time series with 19 measurements for each end-effector: tool tip Cartesian positions ($x, y, z$), rotations (denoted by a $3 \times 3$ matrix $R$), linear velocities ($v_x, v_y, v_z$), angular velocities ($\omega'_x, \omega'_y, \omega'_z$), and the gripper angle $\theta$. Details of the JIGSAWS kinematic motion data are summarized in Table 1.

The dataset contains recordings from eight surgeons with varying robotic surgical experience. Each surgeon performed three different training tasks, namely, Suturing (SU), Knot-tying (KT), and Needle-passing (NP), and repeated each task in five times. All three tasks are typically standard components in surgical skill training curricula [41]. An illustration of the three operation tasks is shown in Fig. 3. The two ways in which skill levels are reported in this data set are: (1) self-reported labels
Table 1: Variables of sensory signals from end-effectors of da Vinci robot. These variables are captured as multivariate time series data in each surgical operation trial.

| End-effector Category         | Description                                      | Variables                                                                 | Channels |
|-------------------------------|--------------------------------------------------|---------------------------------------------------------------------------|----------|
| Master Tool Manipulator (MTM) |                                                                                             | $x, y, z, R \in \mathbb{R}^{3 \times 3}$, $v_x, v_y, v_z, \omega_x', \omega_y', \omega_z'$, $\alpha$ | $19 \times 2$ |
| Patient-side manipulator (PSM)|                                                                                             | $x, y, z, R \in \mathbb{R}^{3 \times 3}$, $v_x, v_y, v_z, \omega_x', \omega_y', \omega_z'$, $\alpha$ | $19 \times 2$ |

Fig. 3: Snapshots of operation tasks during robot-assisted minimally invasive surgical training. The operations are implemented using the da Vinci robot and are reported in JIGSAWS [41]: (A) Suturing, (B) Needle-passing, (C) Knot-tying.

Based on practice hours with expert reporting greater than 100 hours, intermediate between 10-100 hours, and novice reporting less than 10 hours of total surgical robotic operation time, and (2) the modified global rating scale (GRS), manually annotated by an experienced surgeon. In this study, we use the self-reported skill level labels as the ground-truth skill levels for each trial.

3.2 Data Preparation & Inputs

**Z-normalization** Due to differences in the scaling ranges and offset effects of the sensory data, the data fed into the neural network are first normalized with a $z$-normalization process. Each channel of raw data, $x$, is normalized individually as $z = \frac{x - \mu}{\alpha}$, where $\mu$ and $\alpha$ are the mean and standard deviation of vector $x$. This normalization process can be performed online by feeding the network with the batch of sensory data.

**Data Augmentation** One challenge for developing a robust skill model with our approach comes from the lack of large-scale data samples in JIGSAWS, where the number of labeled samples is only 40 in total (8 subjects with 5 trial repetitions) for each surgical task. Generally, deep learning might suffer from overfitting if the size of available dataset is limited [54]. To overcome this problem, data augmentation is introduced to prevent overfitting and improve generalization of the deep learning model. This has been seen so far mostly in image recognition, where several methods, such as scaling, cropping, and rotating are used [50][51].
Algorithm 1 Sliding-window Cropping Algorithm

INPUT: raw time series $X$, stepSize $L$, windowWidth $W$
OUTPUT: sub-sequences $s = \text{SlidingWindow}(X, L, W)$

1: initialization $m := 0$, $n := 0$
2: $s := \text{empty}$
3: while $m + W \leq \text{length}(X)$ do
4: \hspace{1em} $x[n] := X[m : (m + W - 1)]$
5: \hspace{1em} $m := m + L$, $n := n + 1$
6: end while
7: return sub-sequences $s$

In this study, to support the network in learning, we adapted the augmentation strategy and introduced a two-step augmentation process before inputting data into our network. First, followed by z-normalization, we viewed and separated the surgical motion data from master (MTMs) and patient-side manipulators (PSMs) as two distinct sample instances, while the class labels for each trial were preserved. This procedure is also appropriate in cases where the MTMs and PSMs are not directly correlated (e.g., position scaling, or other differences in robot control terms). Then, we carried out a label-preserving cropping with a sliding window, where the motion sub-sequences were extracted using crops, i.e., sliding a fixed-size window within the trial. The annotation for each window is identical to the class label of original trial, from which the sub-sequences are extracted. One advantage of this approach is that it leads to larger-scale sets for the robust training and testing of the network. Also, this technique allows us to format time series as equal-length inputs, regardless of the varied lengths of original data. The pseudo-code of sliding-window cropping algorithm is shown in Algorithm 1, where $X$ is the input motion data, $s$ is the output crops (i.e., sub-sequences), $W$ is the sliding window size and $L$ is the step size. After experimenting based on trial-and-error, we chose a window size $W = 60$ and a step size $L = 30$ in this work. Overall, by applying the aforementioned data augmentation process on the original dataset, it resulted in 6290, 6780, and 3542 crops for Suturing, Needle-passing, and Knot-tying, respectively. All of these crops are new data samples for the network. The overall numbers of obtained crops are different since original recording lengths are varied across each trial in JIGSAWS. As a result, we obtained the total sample trials with the size of 6290, 6780, and 3542 for three tasks, respectively, according to the selected setting.

3.3 Training & Testing

To validate the model classification, we adopt two individual validation schemes in this work: Leave-one-supertrial-out (LOSO) and Hold-out. The objective of the comparison is to search for the best validation strategy suitable for system development in the case of deep learning. Based on each cross-validation setting, we train and test
a surgical skill model for each surgical task, Suturing (SU), Knot-tying (KT), and Needle-passing (NP).

**Leave-one-supertrial-out (LOSO) cross-validation:** This technique involves repetitively leaving out a single subset for testing in multiple partitions. Specifically, a supertrial, i, defined as a subset of examples combining the i-th trials from all subjects for a given surgical task [41], is left out for testing, while the union of remaining examples is used for training. This process is repeated in five folds where each fold consists of each one of the five supertrials. The average of all five-fold performance measures (see Section 3.4 for definitions) in each test set is reported and gives an aggregated classification result. As a widely-used validation strategy, the LOSO cross-validation shows its value in evaluating the robustness of a method for skill assessment.

**Hold-out:** Different from the LOSO cross-validation, the Hold-out strategy is implemented by conducting a train/test split once, which is normally adopted in deep learning models when large datasets are presented. In this work, one single subset consisting of one of the five trials from each surgeon, for a given surgical task, is left out throughout the training and used as a hold-out for the purpose of testing. Also, to reduce the bias and avoid potential overfitting, we randomly select a trial out of the five repetitions for each subject.

### 3.4 Modeling Performance Measures

To compare the model performance, classifications are evaluated regarding four common metrics (Eq. 2) [49, 55, 56]: the average accuracy – ratio between the sum of correct predictions and the total number of predictions; precision – ratio of correct positive predictions ($T_p$) and the total positive results predicted by the classifier ($T_p + F_p$); recall – ratio of positive predictions ($T_p$) and the total positive results in the ground-truth ($T_p + F_n$); and $f1$-score – a weighted harmonic average between precision and recall.

\[
\text{precision} = \frac{T_p}{T_p + F_p}
\]
\[
\text{recall} = \frac{T_p}{T_p + F_n}
\]
\[
\text{f1-score} = \frac{2 \times (\text{recall} \times \text{precision})}{\text{recall} + \text{precision}}
\]  

where $T_p$ and $F_p$ are the numbers of true positives and false positives, $T_n$ and $F_n$ are the numbers of true negatives and false negatives, for a specific class.

In order to assess the computing effort involved in model classification, we measure the running time of skill models to classify all samples in the entire testing set. In the LOSO scheme, the running time is measured as the average value from the five-fold cross-validation.
4 RESULTS

Fig. 4 (A) and (B) give a visual intuition of modeling performance and confusion matrices of three-class skill classifications, which are evaluated on the testing set under the LOSO scheme. We compare our results to the state-of-the-art skill models in Table 2. It is important to mention that in order to obtain a valid benchmarking analysis, the classifiers investigated in this study are selected among the skill assessment from JIGSAW motion data and evaluated using the same LOSO validation. Upon the per-window classification basis, our model achieved accuracies of 93.4%, 89.8%, and 84.9% in tasks of Suturing, Needle-passing and Knot-tying, respectively, using a crop with 2-second duration containing 60 time points ($W = 60$). Highest accuracies reported in the literature range from 99.9% to 100% via a descriptive model using novel entropy features. Nevertheless, the deep skill model outperforms $k$-nearest neighbors ($k$-NN), logistic regression (LR), and support vector machine (SVM), with the accuracy improvements ranging from 3.89% to 23.87% in Suturing, and 3.16% to 12.60% in Knot-tying.

To further assess the capability of our proposed approach for online skill decoding, we repeat our experiment with different sizes of sliding window: $W_1 = 30$, $W_2 = 60$ and $W_3 = 90$. Modeling performance of window sizes together with the average running time taken for classification is reported in Table 3. The results show that our deep learning skill model can offer advantages over traditional approaches with highly time-efficient skill classification on the per-window basis, without the full observation of surgical motion for each trial (per-trial basis). Also, a higher average accuracy can be found with an increase of sliding window size. Specifically, the 3-second sliding window containing 90 time steps ($W_3 = 90$) can obtain better results compared to 2-second window ($W_2 = 60$), with average accuracy improvements of 0.75% in Suturing, 0.56% in Needle-passing and 2.38% in Knot-tying, respectively.

Furthermore, in order to characterize the roles of two validation schemes, we repeat the above modeling process with the Hold-out strategy. Table 3 shows the classification results under LOSO cross-validation and Hold-out schemes.

5 DISCUSSION

Recent trends in robot-assisted surgery have promoted a great need for proficient approaches for objective skill assessment [11]. Although several analytical techniques have been developed, efficiently measuring surgical skills from complex surgical data still remains an open problem. In this paper, our primary goal is to introduce and evaluate the applicability of a novel deep learning approach towards online surgical skill assessment. Compared to conventional approaches, our proposed deep learning model reduced dependency on the complex feature design or carefully-tuned gesture segmentation. Overall, a deep learning skill model, with appropriate design choices, yielded competitive performance in both accuracy and time efficiency.
Fig. 4: Performance measures and confusion matrices in LOSO validation using the proposed deep skill model in three surgical training tasks. (A) Measures are reported regarding precision, recall, and f1-score for each skill level. (B) Confusion matrices of classification results. Element value \((i, j)\) and color represent the probability of predicted skill label \(j\), given the self-proclaimed skill label \(i\), where \(i, j \in \{1: \text{"Novice"}, 2: \text{"Intermediate"}, 3: \text{"Expert"}\}\). The diagonal corresponds to correct predictions.

5.1 Validity of our deep learning model for objective skill assessment

For results shown in Fig. 4 (A) and (B), we note that both Suturing and Needle-passing are associated with better results than Knot-tying. The majority of mistakes are made in the Knot-tying task when discriminating between Novice and Intermediate. Specifically, as shown in Fig. 4(B), the distribution across Intermediate is pronounced with the probability of 0.34 being misclassified as Novice. It indicates a slightly higher challenge to recognize Intermediate, which is likely to be confused with the Novice. This mistake made by the network can be explained as there exist minor differences of skill levels between Intermediate and Expert in the JIGSAW dataset [30]. It should also be considered that the self-reported skill labels based on hours spent in robot operations might not be strong enough as the ground-truth knowledge of expertise in this study. More accurate class labels relative to true surgeon skills might help to reduce the existing misclassifications in our models.

As shown in Table 2, better classification accuracy can be achieved by a few existing methods using generative modeling and descriptive modeling. Specifically, a generative model, sparse HMM (S-HMM), is able to give high predictive accuracy.
Table 2: Comparative evaluation of existing techniques employed for skill assessment using motion data from the JIGSAWS. We benchmark the results in terms of accuracy in the LOSO cross-validation. Models conducting classification on the trial level are categorized as *per-trial basis*.

| Author, Year | Metric Extraction | Validation Scheme | Method | SU   | NP   | KT   | Characteristics                                                                 |
|--------------|-------------------|-------------------|--------|------|------|------|--------------------------------------------------------------------------------|
| Lingling 2012 [33] | local gesture segments | LOSO | S-HMM | 97.4 | 96.2 | 94.4 | • generic modeling  
• segment-based  
• per-trial basis |
| Forestier 2017 [30] | bag of words features | LOSO | VSM  | 89.7 | 96.3 | 61.1 | • descriptive modeling  
• feature-based  
• per-trial basis |
| Zia 2018 [32] | entropy features | LOSO | NN   | 100  | 99.9 | 100 | • descriptive modeling  
• feature-based  
• per-trial basis |
| Fard 2017 [28] | basic motion features | LOSO | k-NN | 89.7 | -    | 82.1 | • descriptive modeling  
• feature-based  
• two-class skill only  
• per-trial basis |
| Current study | N/A | LOSO | CNN  | 93.4 | 89.8 | 84.9 | • deep learning modeling  
• no manual feature  
• per-window basis  
• end-to-end analysis |

ranging from 94.4% to 97.4%. This result might benefit from a precise description of motion structures and pre-defined gestures in each task. However, such an approach requires prerequisite segmentation of motion sequences, as well as different complex class-specific models for each skill level [33]. Second, descriptive models sometimes may be superior to provide highly accurate results, such as the use of novel entropy features. However, the deficiency is that significant domain-specific knowledge and development is required to define the most informative features, which directly affect the assessment performance. This deficiency could also explain why there exists a larger variance in accuracy between other studies (61.1%-100%), as shown in Table 2, which are sensitive to the choice of extracted representations.

Another attention of our analysis is focused on the optimal sliding windows needed to render an efficient assessment. The duration of time steps in each window should roughly correspond to the minimum time required to decode skills from input signals. Usually, technical skill is assessed at the trial level; however, a quicker and more efficient acquisition may enable immediate feedback to the trainee, possibly improving learning outcomes. Overall, our findings suggest that the per-window-based classification in this work is well-applicable for online settings. Smaller window size can allow for a faster running speed and less delay due to the light-weight computing expense. In contrast, an larger window size implies an increase of delay due to larger
Table 3: Summary table showing classification performance based on different validation schemes and sliding windows. Window size is set as $W_1 = 30$, $W_2 = 60$ and $W_3 = 90$. Bold numbers denote best classification results regarding $f_1$-score and accuracy. Running time quantifies the computing effort involved in classification.

| Task          | Validation Scheme | Window Size | $f_1$-score | Accuracy | Running Time (ms) |
|---------------|-------------------|-------------|-------------|----------|------------------|
| **Suturing**  |                   |             |             |          |                  |
|               | LOSO              | $W_1$       | 0.94        | 0.95     | 0.930           |
|               |                   | $W_2$       | 0.94        | 0.93     | 185.40           |
|               |                   | $W_3$       | 0.95        | 0.96     | 247.01           |
| Hold-out      |                   | $W_1$       | 0.98        | 0.94     | 247.01           |
|               |                   | $W_2$       | 0.99        | 0.96     | 98.10            |
|               |                   | $W_3$       | 0.99        | 0.97     | 146.40           |
| **Needle-passing** |             |             |             |          |                  |
| LOSO          |                   | $W_1$       | 0.95        | 0.73     | 0.889           |
|               |                   | $W_2$       | 0.95        | 0.75     | 153.36           |
|               |                   | $W_3$       | 0.96        | 0.89     | 194.98           |
| Hold-out      |                   | $W_1$       | 0.97        | 0.80     | 194.98           |
|               |                   | $W_2$       | 0.98        | 0.81     | 169.72           |
|               |                   | $W_3$       | 0.98        | 0.94     | 207.12           |
| **Knot-tying** |             |             |             |          |                  |
| LOSO          |                   | $W_1$       | 0.90        | 0.85     | 214.14           |
|               |                   | $W_2$       | 0.90        | 0.85     | 248.03           |
|               |                   | $W_3$       | 0.92        | 0.91     | 147.38           |
| Hold-out      |                   | $W_1$       | 0.98        | 0.94     | 147.38           |
|               |                   | $W_2$       | 0.88        | 0.92     | 74.5             |
|               |                   | $W_3$       | 0.35        | 0.91     | 139.39           |

network complexity and higher computing effort involved in decoding. Specifically, as shown in Table 3 within the LOSO validation scheme, the network can classify the entire testing dataset within 133.88 ms for $W_1$ and 172.87 ms running time for $W_2$, while it required 214.14 ms running time for $W_3$ to classify the samples. However, it is important to mention that given an increase of window sizes, a higher accuracy can be achieved. In particular, there seems to be more gains in the Knot-tying analysis, where the highest 2.24% accuracy improvement was obtained from $W_2$ to $W_3$. This result might be due to the fact that more information of motion dynamics are contained in larger crops, thus allowing for an improved decoding accuracy. We suggest that this trade-off between decoding accuracy and time efficiency could be a factor of interest in online settings.

5.2 Comparison of Validation Schemes

We investigated the validity of two different validation schemes for skill modeling. In this case, the differences between both are non-trivial in the deep learning development. Noticeably, LOSO cross-validation gives a reliable estimate of system performance. However, the Hold-out scheme, which uses a random subset of trials as a hold-out, demonstrates relatively larger variances among results. This result can be explained by the differences among these randomly selected examples in the Hold-out validation. Nevertheless, the Hold-out shows consistency with the results.
in LOSO scheme across different tasks and window sizes, as shown in Table 3. It is important to note that given a large dataset, the LOSO cross-validation might be less efficient for model assessment. In this scenario, the computing load in LOSO modeling has been largely increased, which may not be suitable for complex deep architectures. However, the Hold-out only needs to run once and is less computationally expensive in modeling.

5.3 Limitations

Despite the progress in present work, there still exist some limitations of deep learning models towards a proficient online skill assessment. First, as performance of supervised deep learning relies heavily on the labeled samples, the primary concern in this study lies with the JIGSAWS dataset and the lack of strong ground-truth labels of skill levels. It is important to note that there is a lack of consensus in the ground-truth annotation of skills. In [28], skill labels were annotated based on a cut-off value of GRS, though no commonly accepted cutoff exists. For future work, stronger ground-truth knowledge of surgeon expertise with a refined annotation approach may further improve the overall skill assessment [57, 58]. We will search for a detailed optimization of our deep architecture, parameter settings and augmentation strategies to better handle motion time-series data and improve the performance further. In addition, it would be interesting to investigate a visualization of deep hierarchical representations to better understand hidden motion patterns in surgical skills.

6 CONCLUSION

The primary contributions of this study are: (1) a novel data-driven deep architecture for an active classification of surgical skill via end-to-end learnings, (2) an insight in accuracy and time efficiency improvements for online skill assessment, and (3) application of data augmentation and exploration of validation schemes feasible for deep skill modeling. Taking advantage of recent technique advances, our approach has several desirable proprieties and is extremely valuable for online skill assessment. First, a key benefit is an end-to-end skill decoding, learning abstract representations of surgery motion data with automatic recognitions of skill levels. Without a priori dependency on engineered features or segmentation, the proposed model achieved comparable results to previously reported methods. It yielded highly competitive time efficiency given relatively small crops (1−3 second window with 30−90 timesteps), which were computationally feasible for online assessment and immediate feedback in training. Furthermore, we demonstrated that an improvement of modeling performance could be achieved by the optimization of design choices. An appropriate window size could provide better results in Knot-tying with a 2.24% accuracy increase. Also, the development of deep skill models might benefit from the Hold-out strategy, which requires less computing effort than the LOSO cross-validation, especially in the case where large datasets are involved.
Overall, the ability to automatically learn abstract representations from raw sensory data with high predictive accuracy and fast processing speed, makes our approach well-suited for online objective skill assessment. The proposed deep model can be integrated into the pipeline of robot-assisted surgical systems and could allow for immediate feedback in training.

Acknowledgment

This work is supported by National Science Foundation (NSF#1464432).

Conflict of interest

The authors, Ziheng Wang and Ann Majewicz Fey, declare that they have no conflict of interest.

Ethical approval

For this type of study formal consent is not required.

Informed consent

This articles does not contain patient data.

References

1. Roberts K. E., Bell R. L., Duffy A. J. (2006) Evolution of surgical skills training. World journal of gastroenterology: WJG 12(20):3219
2. Reznick R. K., MacRae H. (2006) Teaching surgical skills changes in the wind. New England Journal of Medicine 355(25):2664–2669
3. Aggarwal R., Mytton O. T., Derbrew M., Hananel D., Heydenburg M., Issenberg B., MacAulay C., Mancini M. E., Morimoto T., Soper N., Ziv A., Reznick R. (2010) Training and simulation for patient safety. BMJ Quality & Safety 19(Suppl 2):i34–i43
4. Birkmeyer J. D., Finks J. F., O’Reilly A., Oerline M., Carlin A. M., Nunn A. R., Dimick J., Banerjee M., Birkmeyer N. J. (2013) Surgical skill and complication rates after bariatric surgery. New England Journal of Medicine 369(15):1434–1442
5. Darzi A., Mackay S. (2001) Assessment of surgical competence. BMJ Quality & Safety 10(suppl 2):ii64–ii69
6. Bridgewater B., Grayson A. D., Jackson M., Brooks N., Grotte G. J., Keenan D. J., Millner R., Fabri B. M., Mark J. (2003) Surgeon specific mortality in adult cardiac surgery: comparison between crude and risk stratified data. British Medical Journal 327(7405):13–17
7. Goh A. C., Goldfarb D. W., Sander J. C., Miles B. J., Dunkin B. J. (2012) Global evaluative assessment of robotic skills: Validation of a clinical assessment tool to measure robotic surgical skills. The Journal of Urology 187(1):247–252
8. Aghazadeh M. A., Jayaratna I. S., Hung A. J., Pan M. M., Desai M. M., Gill I. S., Goh A. C. (2015) External validation of global evaluative assessment of robotic skills (gears). Surgical Endoscopy 29(11):3261–3266
9. Niitsu H., Hirabayashi N., Yoshimitsu M., Mimura T., Taomoto J., Sugiyama Y., Murakami S., Saeki S., Mukaida H., Takiyama W. (2013) Using the objective structured assessment of technical skills (osats) global rating scale to evaluate the skills of surgical trainees in the operating room. Surgery Today 43(3):271–275
10. Reiley C. E., Lin H. C., Yuh D. D., Hager G. D. (2011) Review of methods for objective surgical skill evaluation. Surgical Endoscopy 25(2):356–366
11. Vedula S. S., Ishii M., Hager G. D. (2017) Objective assessment of surgical technical skill and competency in the operating room. Annual review of biomedical engineering 19:301–325
12. Moustris G. P., Hiridis S. C., Deliparaschos K. M., Konstantinidis K. M. (2011) Evolution of autonomous and semi-autonomous robotic surgical systems: a review of the literature. The International Journal of Medical Robotics and Computer Assisted Surgery 7(4):375–392
13. Cheng C., Sa-Ngasoongsong A., Beyca O., Le T., Yang H., Kong Z., Bukkapatnam S. T. (2015) Time series forecasting for nonlinear and non-stationary processes: a review and comparative study. IIE Transactions 47(10):1053–1071
14. Klonowski W. (2009) Everything you wanted to ask about eeg but were afraid to get the right answer. Nonlinear Biomedical Physics 3(1):2
15. Reiley C. E., Hager G. D. (2009) Task versus subtask surgical skill evaluation of robotic minimally invasive surgery. In: International Conference on Medical Image Computing and Computer-assisted Intervention, Springer, pp. 435–442
16. Kassahun Y., Yu B., Tibebu A. T., Stoyanov D., Giannarou S., Metzen J. H., Vander Poorten E. (2016) Surgical robotics beyond enhanced dexterity instrumentation: a survey of machine learning techniques and their role in intelligent and autonomous surgical actions. International Journal of Computer Assisted Radiology and Surgery 11(4):553–568
17. Judkins T. N., Oleynikov D., Stergiou N. (2009) Objective evaluation of expert and novice performance during robotic surgical training tasks. Surgical endoscopy 23(3):590
18. Liang K., Xing Y., Li J., Wang S., Li A., Li J. (2018) Motion control skill assessment based on kinematic analysis of robotic end-effector movements. The International Journal of Medical Robotics and Computer Assisted Surgery 14(1):e1845–n/a, DOI 10.1002/rcs.1845
19. Trejos A. L., Patel R. V., Malthaner R. A., Schlachta C. M. (2014) Development of force-based metrics for skills assessment in minimally invasive surgery. Surgical Endoscopy 28(7):2106–2119
20. Poursartip B., LeBel M.-E., Patel R., Naish M., Trejos A. L. (2017) Analysis of energy-based metrics for laparoscopic skills assessment. IEEE Transactions on Biomedical Engineering
21. Ershad M., Koesters Z., Rege R., Majewicz A. (2016) Meaningful assessment of surgical expertise: Semantic labeling with data and crowds. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, pp. 508–515
22. Sharon Y., Lendvay T. S., Nisky I. (2017) Instrument orientation-based metrics for surgical skill evaluation in robot-assisted and open needle driving. arXiv preprint arXiv:170909452
23. Shackelford S., Bowyer M. (2017) Modern metrics for evaluating surgical technical skills. Current Surgery Reports 5(10):24
24. Fard M. J., Ameri S., Darin Ellis R., Chinnam R. B., Pandya A. K., Klein M. D. (2018) Automated robot-assisted surgical skill evaluation: Predictive analytics approach. The International Journal of Medical Robotics and Computer Assisted Surgery 14(1)
25. Stefanidis D., Scott D. J., Korndorffer Jr J. R. (2009) Do metrics matter? time versus motion tracking for performance assessment of proficiency-based laparoscopic skills training. Simulation in Healthcare 4(2):104–108
26. Chmarra M. K., Klein S., de Winter J. C., Jansen F.-W., Dankelman J. (2010) Objective classification of residents based on their psychomotor laparoscopic skills. Surgical Endoscopy 24(5):1031–1039
27. Vedula S. S., Malpani A., Ahmidi N., Khudanpur S., Hager G., Chen C. C. G. (2016) Task-level vs. segment-level quantitative metrics for surgical skill assessment. Journal of Surgical Education 73(3):482–489
28. Fard M. J., Ameri S., Darin Ellis R., Chinnam R. B., Pandya A. K., Klein M. D. (2018) Automated robot-assisted surgical skill evaluation: Predictive analytics approach. The International Journal of Medical Robotics and Computer Assisted Surgery 14(1):e1850–n/a, DOI 10.1002/rcs.1850
29. Poursartip B., LeBel M.-E., McCracken L. C., Escoto A., Patel R. V., Naish M. D., Trejos A. L. (2017) Energy-based metrics for arthroscopic skills assessment. Sensors 17(8):1808
30. Forestier G., Petitjean F., Senin P., Despinoy F., Jannin P. (2017) Discovering discriminative and interpretable patterns for surgical motion analysis. In: Conference on Artificial Intelligence in Medicine in Europe, Springer, pp. 136–145
31. Brown J. D., O’Brien C. E., Leung S. C., Dumon K. R., Lee D. I., Kuchenbecker K. J. (2017) Using contact forces and robot arm accelerations to automatically rate surgeon skill at peg transfer. IEEE Transactions on Biomedical Engineering 64(9):2263–2275
32. Zia A., Essa I. (2018) Automated surgical skill assessment in rmis training. International Journal of Computer Assisted Radiology and Surgery DOI 10.1007/s11548-018-1735-5
33. Tao L., Elhamifar E., Khudanpur S., Hager G. D., Vidal R. (2012) Sparse hidden markov models for surgical gesture classification and skill evaluation. In: IPCAI, Springer, pp. 167–177
34. LeCun Y., Bengio Y., Hinton G. (2015) Deep learning. Nature 521(7553):436–444
35. Schmidhuber J. (2015) Deep learning in neural networks: An overview. Neural Networks 61:85–117
36. Silver D., Huang A., Maddison C. J., Guez A., Sifre L., van den Driessche G., Schrittwieser J., Antonoglou I., Panneershelvam V., Lanctot M., Dieleman S., Grewe D., Nham J., Kalchbrenner N., Sutskever I., Lillicrap T., Leach M., Kavukcuoglu K., Graepel T., Hassabis D. (2016) Mastering the game of go with deep neural networks and tree search. Nature 529(7587):484–489

37. Graves A., Mohamed A.-r., Hinton G. (2013) Speech recognition with deep recurrent neural networks. In: Acoustics, speech and signal processing (icassp), 2013 iee international conference on, IEEE, pp. 6645–6649

38. Esteva A., Kuprel B., Novoa R. A., Ko J., Swetter S. M., Blau H. M., Thrun S. (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542(7639):115–118

39. Rajpurkar P., Hannun A. Y., Haghpanahi M., Bourn C., Ng A. Y. (2017) Cardiologist-level arrhythmia detection with convolutional neural networks. arXiv preprint arXiv:170701836

40. DiPietro R., Lea C., Malpani A., Ahmidi N., Vedula S. S., Lee G. I., Lee M. R., Hager G. D. (2016) Recognizing surgical activities with recurrent neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, pp. 551–558

41. Gao Y., Vedula S. S., Reiley C. E., Ahmidi N., Varadarajan B., Lin H. C., Tao L., Zappella L., Béjar B., Yuh D. D., Chen C. C. G., Vidal R., Khudanpur S., Hager G. D. (2014) Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling

42. Längkvist M., Karlsson L., Loutfi A. (2014) A review of unsupervised feature learning and deep learning for time-series modeling. Pattern Recognition Letters 42:11–24

43. Gamboa J. C. B. (2017) Deep learning for time-series analysis. arXiv preprint arXiv:170101887

44. Nair V., Hinton G. E. (2010) Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 807–814

45. Li M., Zhang T., Chen Y., Smola A. J. (2014) Efficient mini-batch training for stochastic optimization. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp. 661–670

46. Kingma D., Ba J. (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980

47. Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R. (2014) Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research 15(1):1929–1958

48. Wu H., Gu X. (2015) Max-pooling dropout for regularization of convolutional neural networks. In: International Conference on Neural Information Processing, Springer, pp. 46–54

49. Ahmidi N., Tao L., Sefati S., Gao Y., Lea C., Haro B. B., Zappella L., Khudanpur S., Vidal R., Hager G. D. (2017) A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. IEEE Transactions on Biomedical Engineering 64(9):2025–2041
50. Krizhevsky A., Sutskever I., Hinton G. E. (2012) Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105
51. He K., Zhang X., Ren S., Sun J. (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778
52. Cui Z., Chen W., Chen Y. (2016) Multi-scale convolutional neural networks for time series classification. arXiv preprint arXiv:160306995
53. Le Guennec A., Malinowski S., Tavenard R. (2016) Data augmentation for time series classification using convolutional neural networks. In: ECML/PKDD Workshop on Advanced Analytics and Learning on Temporal Data
54. Um T. T., Pfister F. M., Pichler D., Endo S., Lang M., Hirche S., Fietzek U., Kulić D. (2017) Data augmentation of wearable sensor data for parkinsons disease monitoring using convolutional neural networks. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction, ACM, pp. 216–220
55. Sammut C., Webb G. I. (2011) Encyclopedia of machine learning. Springer Science & Business Media
56. Kumar R., Jog A., Malpani A., Vagvolgyi B., Yuh D., Nguyen H., Hager G., Chen C. C. G. (2012) Assessing system operation skills in robotic surgery trainees. The International Journal of Medical Robotics and Computer Assisted Surgery 8(1):118–124
57. Sun C., Shrivastava A., Singh S., Gupta A. (2017) Revisiting unreasonable effectiveness of data in deep learning era. In: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, pp. 843–852
58. Dockter R. L., Lendvay T. S., Sweet R. M., Kowalewski T. M. (2017) The minimally acceptable classification criterion for surgical skill: intent vectors and separability of raw motion data. International journal of computer assisted radiology and surgery 12(7):1151–1159