Enhancing Skill Assessment of Autonomous Robot-Assisted Minimally Invasive Surgery: A Comprehensive Analysis of Global and Gesture-Level Techniques applied on the JIGSAWS Dataset

Eszter Lukács[1], Renáta Levendovics[1,2,3], and Tamás Haidegger[1,3]

[1]Antal Bejczy Center for Intelligent Robotics, University Research and Innovation Center, Óbuda University, H-1034 Budapest, Hungary
[2]Doctoral School of Applied Informatics and Applied Mathematics, Óbuda University, H-1034 Budapest, Hungary
[3]Austrian Center for Medical Innovation and Technology, 2700 Wiener Neustadt, Austria

{eszter.lukacs, renata.levendovics, haidegger}@irob.uni-obuda.hu

Abstract:

Improved surgical skills play a crucial role in ensuring optimal patient outcomes. Traditional methods for skill assessment include self-rating questionnaires and expert evaluations, but these approaches are prone to bias and require substantial qualified human resources. The emergence of Surgical Data Science (SDS) offers a promising avenue for automating skill assessment, leveraging data science techniques to capture, organize, analyze, and model surgical data. In this paper, kinematic data was employed from the JIGSAWS – which is the only skill-annotated Robot-Assisted Minimally Invasive Surgery (RAMIS) dataset – to classify surgeons into novice and experienced groups, using various classification methods (Decision Tree, k-Nearest Neighbors, Support Vector Machines, Logistic Regression, Dynamic Time Warping, and 1D Convolutional Neural Network). The research encompasses a thorough analysis of parameter tuning and dimensional reduction techniques with the aim of establishing a universal benchmark for data classification. The surgical training tasks of suturing, knot-tying and needle-passing consistently achieved 100 % accuracy. The accuracy attained during surgical gesture analysis often exceeded the overall accuracy of the global assessment of the dataset.

Keywords: Surgical Skill Assessment, Robot-Assisted Minimally Invasive Surgery, JIGSAWS, Decision Tree, k-Nearest Neighbors, Support Vector Machine, Logistic Regression, Dynamic Time Warping, 1D Convolutional Neural Network, Approximate Entropy, Mutual Information

1 Introduction

The transition from open access surgery to Minimally Invasive Surgery (MIS) marked a significant paradigm shift in medicine during the latter half of the 20th century [1]. The benefits of MIS have been shown during the recent pandemic as well [2]. Despite the evident advantages of MIS, such as reduced recovery time, smaller incisions and decreased blood loss, it presents challenges to surgeons due to the limited field of view, lack of depth perception caused by the 2D endoscopic camera, and the intricate manipulation of the laparoscopic tools [3, 4]. Successful performance of MIS necessitates extensive training besides continuous feedback on skills is crucial given the complexity of the procedures. While RAMIS is typically considered a costly technology add-on to surgery, partially it has been seen as an initial component towards sustainable and accessible healthcare [5]. More recently, RAMIS has been presented as method to support ethically aligned design in digital health devices [6].

The proficiency and knowledge of surgeons directly impact patient outcomes, reflecting years of training, supervisory evaluation and clinical experience [7]. Surgeons possess technical skills, non-technical skills and different level of experience. The objective assessment of technical skills, such as tool handling, bimanual dexterity and procedural flow, has been extensively researched. However, non-technical skills, including situation awareness, stress management, decision-making, among others, are similarly important in relation to patient outcomes [8–10]. Workload management, which quantifies the effort required to perform a task, may also exhibit a strong correlation with non-technical skills. Unfortunately, in many regions even in developed countries, surgical skill training and assessment are not yet integrated into routine clinical practices [11].

Surgical skill assessment can be performed using self-rating questionnaires, where participants evaluate their own performance, or standardized expert rating techniques, where a panel of experts (typically 8–10 surgeons) assesses surgical procedures or training based on video recordings. However, both of these approaches can be inherently biased and require significant qualified human resources. The ultimate goal is to automate surgical skill assessment, as it offers objectivity, but the technical aspects and critical surgical features are still subjects of intensive research. Surgical Data Science (SDS) aims to enhance the quality of interventional healthcare by employing data science techniques for data capture, organization, analysis and modeling [12]. SDS techniques enable automated surgical skill assessment and allow verification of key skills through correlations between them.

Robot-Assisted Minimally Invasive Surgery (RAMIS) refers to a surgical technique, in which laparoscopic tools controlled by a human operator at a console are applied remotely by a robot. Teleoperation provides benefits, such as tremor filtering, 3D endoscopic vision, ergonomic advantages, rescaled motion and improved tool handling. The assessment of RAMIS skills has been extensively studied due to the availability of sensory data, including robot kinematic data and 3D video endoscopy. RAMIS skill assessment closely relates to MIS, as the key factors encompass both technical and non-technical skills.

In this paper, our aim was to classify surgeons into novice and expert groups based on kinematic data collected by the da Vinci Surgical System (dVSS, Intuitive Surgical Inc., Sunnyvale, CA). The study included a thorough analysis to investigate the potential benefits of dataset partitioning, parameter optimization, and dimensionality reduction in improving the performance of classification algorithms. Several classifiers were employed to analyze the data, including non-time series methods such as Decision Tree (DT), k-Nearest Neighbors (k-NN), Support Vector Machines (SVM) and Logistic Regression (LR). Additionally, to compare the performance of non-time series classifiers with time series techniques, Dynamic Time Warping (DTW) and 1D Convolutional Neural Network (1D CNN) models were evaluated. The accuracy achieved during gesture (important and distinctive movement during the surgical tasks) analysis often exceeded the overall accuracy of the global assessment of the dataset. One particular gesture, ("using right hand to help tighten suture"), showed the highest performance, indicating its significant role in classifying surgeons' skills. The surgical tasks of suturing, knot-tying and needle-passing were achieved 100 % accuracy.

2 State of the Art

Automated skill assessment plays a crucial role in evaluating the proficiency of surgeons [13]. In this study, authors investigated the advantages and the challenges associated with utilizing the entire dataset or the gesture-divided dataset obtained from the RAMIS system, which provides kinematic and video data. The laparoscopic camera provides easy access to video data in robotic surgery, obviating the need for sensor-based kinematic data collection [14]. However, the multidimensional and considerably complex nature of video data has resulted in its relatively limited utilization compared to kinematic data [14].

Most automated skill assessment approaches employ Hidden Markov Models (HMMs), created on the basis of internal features extracted from kinematic and video data [15]. However, the training process of HMM methods can be extremely time-consuming compare to other approaches. In order to alleviate this limitation, Frad et al. introduced the application of the DTW and the k-NN algorithm on the JIGSAWS dataset, resulting in the development of an automated and personalized RAMIS gesture training system. The system achieved notable results with an accuracy of 80.49 % for suturing, 70.12 % for needle-passing, and 85.14 % for knot-tying, assessed using the Leave-One-User-Out (LOUO) cross-validation method [16].

Another approach to automatic performance evaluation of surgeons involves the utilization of various machine learning algorithms such as k-NN, LR, and SVM. Fard et al. employed these classification methods along with eight significant movement features (task completion time, path length, depth perception, speed, motion smoothness, curvature, turning angle, tortuosity) to classify surgeons into expert and novice groups. The proposed framework achieved an accuracy of 89.9 % for suturing and 82.3 % for knot-tying using the Leave-One-Supertrial-

Out (LOSO) cross-validation method [17].

Zia et al. achieved outstanding results, with nearly perfect accuracy for almost every task, except for knot-tying, which achieved an accuracy of 99.9 % using the LOSO validation method. In their study, holistic features, such as Approximate Entropy (ApEn) were employed to assess the skill of surgeons. The dimensionality-reduced data was processed using Principal Component Analysis (PCA) before applying the k-NN classifier. These findings highlight the superiority of their method over traditional HMM-based approaches, emphasizing its potential for accurate and reliable surgical skill evaluation [18].

3 Materials and methods

3.1 JIGSAWS dataset

The JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) was collected with the research version of the dVSS [19]. This dataset contains kinematic and video data from 8 participants (B, C, D, E, F, G, H, I) performing 3 different robotic surgical tasks: suturing, knot-tying and needle-passing. All participants were right-handed and their levels of surgical experience varied.

The dataset includes annotations for surgical skill assessment, as well as annotations for specific gestures. The participants' technical skill level was assessed using a Global Rating Scale (GRS). The GRS is derived from modified Objective Structured Assessments of Technical Skills (OSATS), this is a method that incorporates six criteria (concerning tissue, suture/needle handling, time and motion, flow of operation, overall performance, quality of final product) to measure the performance [20]. Each criterion is scored on a scale ranging from 1 to 5, where 5 is the best. The GRS combines the scores obtained in these six criteria to provide an overall assessment of the participants' technical skill.

Moreover, to further analyze the participants' performance, the dataset annotations includes 15 gestures that are summarized in Table 1.

Table 1
Gesture descriptions of RAMIS training sessions in the JIGSAWS dataset [19].

Gesture index	Gesture description
G1	Reaching for needle with right hand
G2	Positioning needle
G3	Pushing needle through tissue
G4	Transferring needle from left to right
G5	Moving to center with needle in grip
G6	Pulling suture with left hand
G7	Pulling suture with right hand

G8	Orienting needle
G9	Using right hand to help tighten suture
G10	Loosening more suture
G11	Dropping suture at end and moving to end points
G12	Reaching for needle with left hand
G13	Making C loop around right hand
G14	Reaching for suture with right hand
G15	Pulling suture with both hands

The specific gestures that were involved in each task are as follows: at knot-tying: G1, G11, G12, G13, G14, G15; at needle-passing: G1, G2, G3, G4, G5, G6, G8, G11; at suturing: G1, G2, G3, G4, G5, G6, G8, G9, G11 (excluding G10). The number of gestures varies not only between different surgical tasks but also across trials within the same task. Some gestures may appear multiple times in one trial, but may be absent at another trial of the same task.

3.2 Hardware and software environment

The classification and evaluation of the properly formatted kinematic data was performed utilizing the Python programming language within Jupyter Notebook. Several libraries, including numpy (version 1.21.4), and pandas (version 1.2.4) were utilized in the implementation process.

The time series data was transformed into a suitable single value format with ApEn for non-time series classification algorithms using the 'antropy' library (version 0.1.4). After the transformation, various classification algorithms were applied to the data. For the implementation of the classification algorithms and the cross-validation techniques, 'sklearn' (version 0.24.1), sktime (version 0.13.4), tensorflow (version 2.9.1) and keras (version 2.9.0) libraries were employed during the implementation.

3.3 Classification methods

The surgeons were classified into two groups (novice, experienced) based on the kinematic data derived from the dVSS. Several different classification methods were utilized during the classification process, most of which required the transformation of the time series kinematic data.

3.3.1 Data preparation

To assess the skill level of participants (Y = features), a binary conversion of the GRS was performed. If a surgeon scored 16 or more out of a total of 30 points on the GRS, he was categorized as experienced (assigned the value 1). If a surgeon obtained a score lower than 16, he was categorized as a novice (assigned the

value 0). This binary format provided a clear distinction between the two skill levels based on the GRS scores.

The data for non-time series classification algorithms was standardized in order to assess the variations in performance between the original and the standardized datasets. During the standardization process, the mean of the features is removed, and the data is centered around a specific value. In some cases, the utilization of standardized data resulted in higher accuracy compared to the original data. This implies that the standardization process may enhance the classification performance, potentially enhancing the algorithms' ability to distinguish between different classes or categories within the dataset. This also suggests that standardization can be a beneficial preprocessing step for non-time series classification algorithms, as it can potentially lead to better overall results. However, the Decision Tree algorithm did not exhibit the same improvement with standardization as other algorithms.

3.3.2 Non-time series classification algorithms

In order to explore the potential benefits of dividing the dataset into intervals based on gestures, four non-time series classification algorithms were applied to the transformed data: Decision Tree (DT), k-Nearest Neighbors (k-NN), Support Vector Machines (SVM) and Logistic Regression (LR). By dividing the dataset into intervals, the aim was to analyze whether this partitioning approach could lead to higher accuracy than utilizing the full dataset. To maximize the accuracy on both the full and the divided datasets, parameter tuning was implemented for the classification algorithms.

DT constructs a classification model in the form of a tree structure, where paths can be represented as "if-then" rules [21]. This algorithm was evaluated by varying the values of the following parameters: criterion with options 'gini' and 'entropy', max_depth (maximum depth of the tree) with options None and integer values from 1 to 10, and max_features (maximum number of features to consider for each split) with options None, 'sqrt', 'log2', and decimal values ranging from 0.1 to 1. Additionally, the splitter parameter (strategy used for splitting each node) was tested with 'best' and 'random' options.

The SVM algorithm constructs a hyperplane in a high-dimensional feature space, with the goal of maximizing the margin between two classes of data points [17]. The parameter tuning for this classifier involved testing different values for the C parameter (penalty parameter) with options 1000, 500, 100, 50, 10, 5, 1, 0.5, and 0.1. The gamma parameter (kernel coefficient) was tested with 'auto', 'scale', and 1, 0.5, 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001 values. The kernel parameter was tested with 'poly', 'rbf', 'sigmoid', and 'linear' options.

LR is used to analyze the relationship between multiple independent variables and the probability of a binary outcome, typically ranging between 0 and 1 [17]. In the context of LR, parameter tuning is a crucial step in optimizing the model's performance. One of the important parameters to consider is the C parameter (inverse of regularization strength) that was tested for optimization with values 0.1, 0.5, 1, 5, 10, 50, 100, 500 and 1000. To ensure convergence of the LR

algorithm, the maximum iteration parameter (max_iter) was set to a higher value of 10000. The penalty parameter was tested with 'l2'and 'l1'. The solver was tested with 'newton-cg', 'lbfgs', 'sag', 'liblinear' and 'saga'.

One observed challenge in the performance of these three non-time series classifiers was the inconsistency in achieving the same accuracy when the test was rerun. To address this task and ensure reproducibility of the results, the random_state parameter was set to a fixed value of 0. By doing so, the classifiers were provided with a fix random seed, thereby facilitating the replication of the results.

Furthermore, the k-NN algorithm that predicts the class label of a given data point based on the class labels of its k most similar neighbors [17] was tuned by testing different metric parameters (distance metric) with 'euclidean', 'manhattan', and 'minkowski' options. The n_neighbors parameter determines the number of neighbors to consider. It was observed to vary based on the value of K in the K-Fold cross-validation (K-Fold cv) method used during evaluation. The weights parameter was tested with 'uniform' and 'distance' to determine the weight of neighbors in the voting process.

To ensure that k-NN always uses all possible neighbors, a specific formula is used. This formula involves the length of the feature set (Y) divided by the chosen K (folds) value. The resulting quotient serves as a crucial factor at the determination of the maximum number of neighbors (max_neighbors) for analysis. If the quotient is an integer, it should be substracted from the length of the feature set, thereby determining the maximum number of neighbors. However, if the quotient is a floating-point number, it is converted to an integer and incremented by one before being subtracted from the length of the feature set.

For instance, consider a scenario where the length of the feature set is 10 and the K value is 2. In this case, the quotient is 10 / 2 = 5, which is an integer. Consequently, the data is divided into two groups of 5 data points each, with one group reserved for testing in the cross-validation process. This results in a maximum of 10 - 5 = 5 neighbors available for analysis in the training data. In contrast, if the K value is 3, the quotient becomes 10 / 3 = 3.33. Here, the data is partitioned into three groups, with approximate sizes of 3, 3 and 4 (the order does not matter). If the validation process selects the last group for testing, the available neighbors for analysis in the training data would be limited to 10 - (int(3.33) + 1) = 6.

After the parameter tuning, a mutual information-based dimensionality reduction technique was used on the classification algorithms. This technique quantifies the degree of dependence between the features, higher values indicating greater dependency. The objective was to leverage this information to reduce the dimensionality of the data using a straightforward equation. This equation involved iteration over a range determined by the minimum and maximum dependency values obtained from the mutual information method, with a step size of 0.01.

3.3.3 Time series classification algorithms

In order to analyze the original time series data, two classification algorithms, namely Dynamic Time Warping (DTW) and 1D Convolutional Neural Network (1D CNN), were employed. However, prior to constructing the models using these algorithms, specific data preparation steps were undertaken.

The DTW measures the similarity between two sequences such as time series that do not align perfectly in terms of time [16]. This classifier requires time series data to be of equal length for classification purposes. In this study, three different data transformation methods were applied to address this requirement. The first one was the maximum length conversion, in this method all time series data within each task were converted to the maximum length observed across all trials in that specific task. This involved padding shorter trials with zeros until their length matched the maximum length. In the average length conversion method, the time series data for each surgical task were converted to the average length. Shorter trials were padded with zeros, longer trials were truncated to match the average length. In the minimum length conversion, the time series data for each surgical task were converted to the minimum length observed across all trials in one specific surgical task. This method involved truncating all trials to match the minimum length. The data loss differed between each trial of the surgeons. By applying these conversions, the time series data were adjusted to a common length, allowing the DTW classifier to effectively classify the surgeons into two groups.

The CNN can be applied to extract spatial features directly [22], but it required additional data transformation for the time series data beyond the conversion methods that were used in the previous classifier. The data underwent further preprocessing to ensure its compatibility with the neural network architecture. This classifier utilized three consecutive 1D convolutional layers (conv1D) in conjunction with BatchNormalization layers. The conv1D layers were configured to employ 64 filters and a kernel size of 3, while applying ”same” padding. The purpose of the BatchNormalization layers was to normalize the activations of each conv1D layer, improving the model’s performance and training efficiency. Following the conv1D and BatchNormalization layers, the output was fed into a GlobalAveragePooling1D layer. Subsequently, the output of the GlobalAveragePooling1D layer was passed through a Dense layer with softmax, which produced the final classification output.

The model was trained using a configuration that included 1000 epochs (EarlyStopping after 50), a batch size of 32 and the Adam optimizer. To evaluate the performance of the model, a LOOCV approach was employed. Due to the inherent variability in the training process of the model, the obtained accuracies differed across iterations. In order to obtain more reliable and robust results, the model was trained five times, and the average accuracy of the five runs was calculated. This averaging process helped to mitigate the impact of the inconsistent weights and training outcomes on the final evaluation. By considering the average accuracy, a more representative performance measure of the model was obtained.

To evaluate not only the variances in the conversion methods but also the impact of different activation functions, two functions were utilized in the model. The first activation function employed was the Rectified Linear Unit (ReLU), which sets all negative values to zero while leaving positive values unchanged. The second activation function used was the Exponential Linear Unit (ELU), which applies an exponential function to both negative and positive inputs.

3.4 Validation methods

The classification methods that were used to categorize surgeons into novice and experienced groups were validated using various techniques. The two main methods were the LOOCV and the simple K-Fold cv method.

In order to mitigate the potential inaccuracies in cross-validation calculations, a constraint was imposed on the range of possible K values. This constraint was necessary due to the inherent limitations of the method, which cannot operate with values lower than 2 (it is necessary because of the train and test splitting). Furthermore, it has been observed that the method produces unreliable results when the K value exceeds the number of the least populated class. Therefore, to ensure reliable and meaningful outcomes, the range of acceptable K values was restricted based on these considerations.

By employing these two validation methods, the classification methods' accuracy could be examined. The simple K-Fold cv provided an assessment by testing the model on different, non-overlapping groups of data, while the LOOCV method ensured that every individual trial was evaluated as a test case.

4 Results

The results of the applied non-time series methods on the entire dataset are presented in Figure 1. The accuracies were attained by firstly using ApEn on the kinematic based time series dataset to transformed it into a usable format, than an excessive parameter tuning was performed to achieve higher results and lastly MI dimensionality reduction method was utilized to reduce the time of the training by reducing the features.

The x-axis of the box plots presents the different classification methods, while the y-axis indicates the achieved accuracy for each method with different K (folds) values. Therefore the accuracies in the diagrams were attained only using the K-Fold cross-validation method.

In most cases the Decision Tree (DT) classifier demonstrated the best accuracy out of the four classifiers. Notably, with the standardized and the original data of the suturing surgical task, it achieved a maximum score of 1.0, indicating a perfect classification outcome. There was only one case when another non-time series classifier, namely the k-Nearest Neighbors (k-NN) displayed better results than the DT classifier in classifying surgeons into two distinct groups. Further-

more this method displayed a relatively smaller number of outliers – represented as circles in the diagrams – compared to the other classifiers across the three surgical tasks. Across all surgical tasks, the maximum achieved accuracies were between 0.83 - 1.0, indicating a high level of accuracy in the classification results. Overall, these findings suggest that the non-time series algorithms, especially the DT classifier implementation for kinematic data based binary classification, have the potential to attain high results.

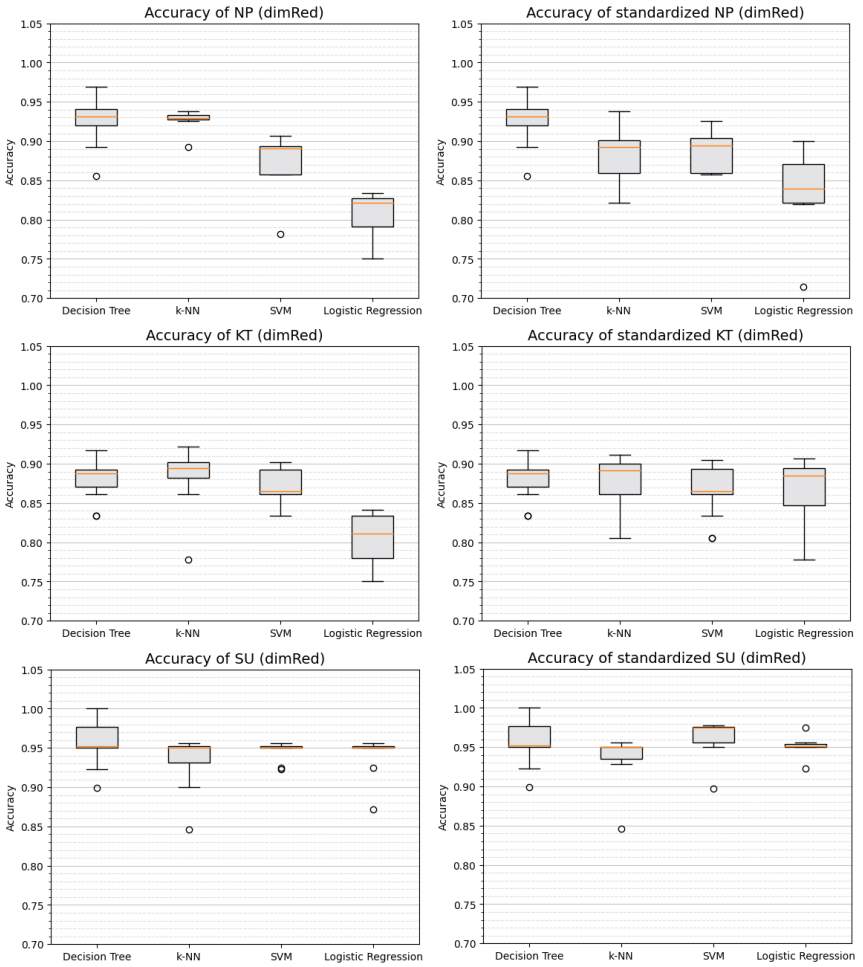

Figure 1
Box plot diagrams of the three surgical tasks, using only non-time series classification algorithms. The accuracies presented in the diagrams on the left were derived from the original dataset and those on the right were computed based on the standardized dataset. Used abbreviations: NP: needle-passing, KT: knot-tying, SU: suturing, k-NN: k-Nearest Neighbors, SVM: Support Vector Machines, dimRed: dimensionality reduction (MI).

The conducted analysis revealed that gesture classification yielded significantly higher accuracy compared to the classification performed on the entire dataset, as shown in Figure 2. The accuracy achieved by the classification models is represented on the y-axis, while the x-axis presents the usable gestures in the specified surgical task. In the diagrams each of the gestures have four shades of blue bars representing the four classification method that was also used for the entire dataset analysis. The accuracies are the outcomes of utilization of ApEn, parameter tuning and dimensional reduction. In contrast to Figure 1, Figure 2 shows the best possible result for the gestures, therefore the results also included LOOCV if a classifier attained higher result with this validation method. Notably, within the surgical tasks examined, a minimum of two gestures demonstrated impeccable accuracy, achieving a perfect score of 1.0. Among the gestures, G9 in the suturing task emerged as the top-performing gesture.

The best non-time series classification results, including parameter tuning and dimensionality reduction, are summarized in Table 2, Table 3 and Table 4. These tables present the best accuracies achieved by the classifiers using either the original or the standardized data. In cases where the result was obtained from the standardized data, it is denoted by an (s) in the corresponding cell. Conversely, if the result was derived from the original data, only the classifier name is listed under the Class column. Notably, the DT classifier consistently yielded the same result for both types of data, represented as (s, o).

The parameters presented in the tables follow the specific order introduced in the non-time series classification methods section. Each table provides a comprehensive overview of the model's performance, respectively, with the first accuracy value indicating the results obtained after parameter tuning, and the second accuracy (MI accuracy) encompasses the utilization of dimensionality reduction technique as well. The best accuracies achieved during the experiments are highlighted in bold text, reflecting the highest achieved accuracy scores.

The models also trained on the split data, which involves dividing the dataset based on gestures, achieved higher accuracy compared to the models trained on the full dataset. However, in the case of suturing, both the individual gestures and the full dataset achieved a perfect accuracy of 1.0. In Table 2, G4, G5, G6 and G8, gestures achieved the best possible results. In Table 3, G1 achieved the best result twice, while G15 achieved the best result three times. In Table 4, the full dataset, along with the gestures G2 and G9 achieved the best result, with G9 achieving the best result eleven times.

The consistent high performance and frequent occurrence of the G9 gesture in Table 4 indicate that this particular gesture possesses characteristics that enable the models to effectively differentiate and accurately classify it from other gestures within the dataset. In summary, extensive parameter tuning and dimensional reduction was performed on each non-time series classification algorithm to optimize their performance. This involved testing various combinations of parameters and features to determine the best settings for achieving higher accuracy.

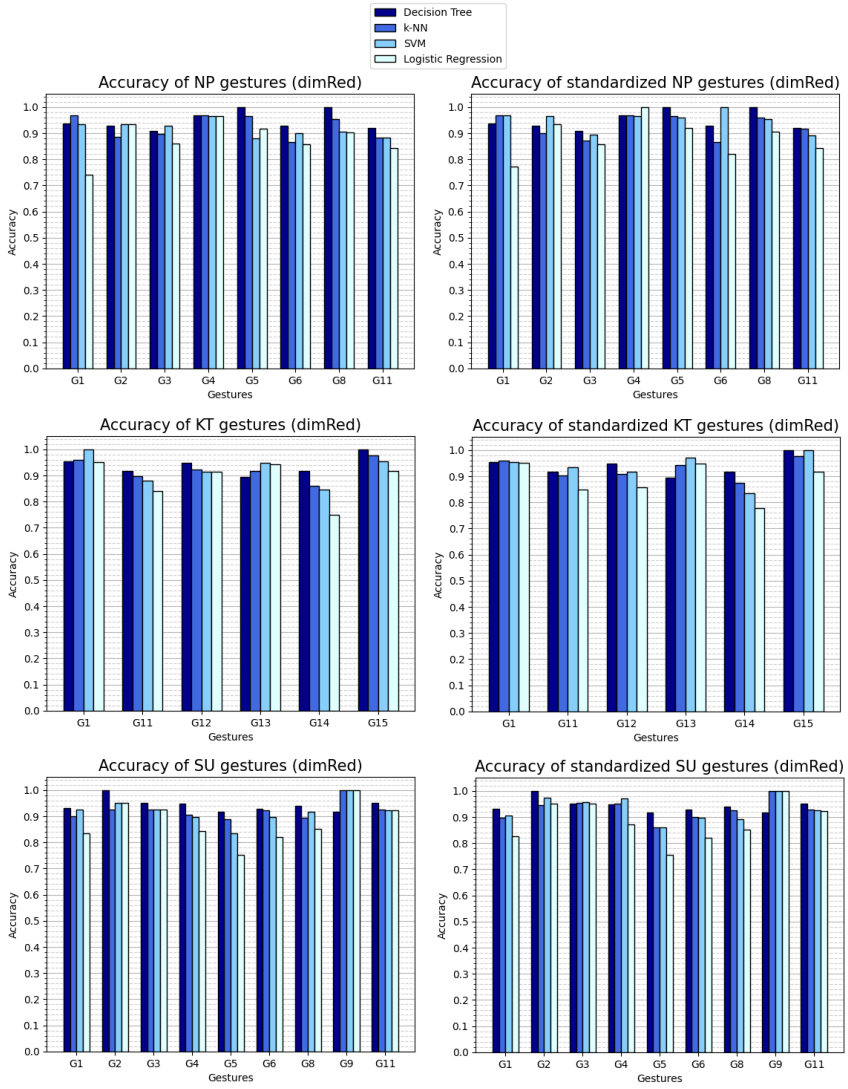

Figure 2
The best accuracy of gestures that can be used in needle-passing, knot-tying and suturing surgical tasks. These accuracies were obtained from either the original gesture dataset (left bar plots) or the standardized gesture dataset (right bar plots), considering the best accuracies. Each color represents a distinct non-time series classification algorithm.

Table 2
Best accuracies of needle-passing. Used abbreviations: Class: Classifier, MI: Mutual Information, K: Folds of the cross-validation, LOOCV: Leave-One-Out cross-validation, DT: Decision Tree, SVM: Support Vector Machines, k-NN: k-Nearest Neighbors, LR: Logistic Regression

Data	**Class**	**Accuracy**	**MI accuracy**	**Best parameters**	**MI score**
Full	DT (o, s)	K = 9: 0.851852	K = 8: **0.96875**	'entropy', None, 0.7, 'random', 0	> 0.08
G1	k-NN (s)	K = 7: 0.821429	K = 8: **0.96875**	'manhattan', 2, 'uniform'	> 0.06
G2	SVM (s)	LOOCV: 0.857143	LOOCV: **0.962963**	50, 0.1, 'sigmoid', 0	> 0.07
G3	SVM	LOOCV: 0.821429	LOOCV: **0.928571**	1000, 0.1, 'sigmoid', 0	> 0.03
G4	LR (s)	K = 8: 0.791667	K = 4: **1.0**	5, 10000, 'l2', 'newton-cg', 0	> 0.03
G5	DT (o, s)	K = 2: 0.833333	K = 6: **1.0**	'gini', None, 0.1, 'random', 0	> 0.28
G6	SVM (s)	K = 8: 0.854167	K = 7: **1.0**	10, 0.1, 'sigmoid', 0	> 0.01
G8	DT (o, s)	K = 6: 0.805556	K = 4: **1.0**	'entropy', None, 'sqrt', 'random', 0	> 0.16
G11	DT (o, s)	LOOCV: 0.84	LOOCV: **0.92**	'gini', 2, 0.8, 'random', 0	> 0.05

Table 3
Best accuracies of knot-tying. Used abbreviations: Class: Classifier, MI: Mutual Information, K: Folds of the cross-validation, LOOCV: Leave-One-Out cross-validation, DT: Decision Tree, SVM: Support Vector Machines, k-NN: k-Nearest Neighbors

Data	**Class**	**Accuracy**	**MI accuracy**	**Best parameters**	**MI score**
Full	k-NN	K = 14: 0.869048	K = 17: **0.921569**	'manhattan', 8, 'distance'	> 0.16
G1	SVM	LOOCV: 0.789474	LOOCV: **1.0**	1000, 1, 'sigmoid', 0	> 0.11
	SVM	K = 6: 0.805556	K = 4: **1.0**	500, 1, 'sigmoid', 0	> 0.11
G11	SVM (s)	K = 10: 0.825	K = 15: **0.933333**	0.5, 1, 'sigmoid', 0	> 0.03

G12	DT (o, s)	K = 10: 0.85	K = 16: **0.947917**	'gini', None, 0.5, 'random', 0	> 0.1
G13	SVM (s)	LOOCV: 0.888889	LOOCV: **0.972222**	0.5, 0.1, 'sigmoid', 0	> 0.03
	SVM (s)	K = 13: 0.923077	K = 4: **0.972222**	10, 0.05, 'sigmoid', 0	> 0.09
G14	DT (o, s)	LOOCV: 0.805556	LOOCV: **0.916667**	'gini', 5, 0.1, 'random', 0	> 0.1
G15	DT (o, s)	LOOCV: 0.888889	LOOCV: **1.0**	'gini', None, 0.5, 'random', 0	> 0.19
G15	DT (o, s)	K = 6: 0.944444	K = 12: **1.0**	'gini', 5, 0.5, 'random', 0	> 0.19
	SVM (s)	K = 9: 0.944444	K = 15: **1.0**	5, 0.05, 'sigmoid', 0	> 0.01

Table 4
Best accuracies of suturing. Used abbreviations: Class: Classifier, MI: Mutual Information, K: Folds of the cross-validation, LOOCV: Leave-One-Out cross-validation, DT: Decision Tree, SVM: Support Vector Machines, k-NN: k-Nearest Neighbors, LR: Logistic Regression

Data	**Class**	**Accuracy**	**MI accuracy**	**Best parameters**	**MI score**
Full	DT (o, s)	K = 9: 0.955556	K = 10: **1.0**	'entropy', None, 0.9, 'random', 0	> 0.12
G1	DT (o, s)	LOOCV: 0.827586	LOOCV: **0.931034**	'gini', None, 'sqrt', 'random', 0	> 0.09
G2	DT (o, s)	K = 9: 0.877778	K = 2: **1.0**	'gini', 6, 0.1, 'random', 0	> 0.07
G3	DT (o, s)	K = 4: **0.975**	K = 8: 0.95	'gini', 2, 0.9, 'random', 0	-
G4	SVM (s)	K = 9: 0.805556	K = 9: **0.972222**	100, 0.05, 'sigmoid', 0	> 0.02
G5	DT (o, s)	LOOCV: 0.861111	LOOCV: **0.916667**	'entropy', 3, 'sqrt', 'random', 0	> 0.01
G6	DT (o, s)	K = 11: 0.878788	K = 9: **0.927778**	'gini', 4, 'log2', 'random', 0	> 0.02
G8	DT (o, s)	LOOCV: 0.848485	LOOCV: **0.939394**	'gini', 2, None, 'random', 0	> 0.0
G9	k-NN	LOOCV:	LOOCV:	'manhattan', 1,	> 0.05

		0.555556	**1.0**	'uniform'	
	k-NN	K = 2: 0.8	K = 2: **1.0**	'manhattan', 1, 'uniform'	> 0.0
	k-NN (s)	K = 3: 0.777778	K = 2: **1.0**	'manhattan', 1, 'uniform'	> 0.01
	SVM	LOOCV: 0.888889	LOOCV: **1.0**	1000, 'scale', 'poly', 0	> 0.03
	SVM (s)	LOOCV: 0.888889	LOOCV: **1.0**	1000, 'auto', 'sigmoid', 0	> 0.13
G9	SVM	K = 2: 0.9	K = 2: **1.0**	500, 0.005, 'sigmoid', 0	> 0.0
	SVM (s)	K = 4: 0.875	K = 2: **1.0**	1000, 'auto', 'sigmoid', 0	> 0.13
	LR	LOOCV: 0.555556	LOOCV: **1.0**	500, 10000, 'l2', 'newton-cg', 0	> 0.05
	LR (s)	LOOCV: 0.777778	LOOCV: **1.0**	0.1, 10000, 'l2', 'liblinear', 0	> 0.1
	LR	K = 2: 0.9	K = 2: **1.0**	5, 10000, 'l2', 'newton-cg', 0	> 0.0
	LR (s)	K = 3: 0.777778	K = 2: **1.0**	0.1, 10000, 'l2', 'liblinear', 0	> 0.13
G11	DT (o, s)	K = 10: 0.875	K = 9: **0.95**	'gini', None, 0.3, 'best', 0	> 0.01

The result of the Dynamic Time Warping (DTW) analysis was visualized in Figure 3. The x-axis represents the number of neighbors used to build the model, while the y-axis represents the accuracy achieved by the DTW classifier when using the specified number of neighbors. It is important to note that the accuracy values presented were obtained only using 'uniform' as a weight parameter value and the validation was performed without using cross-validation. Instead, the dataset was split into a training and a test dataset, following an 80-20 % ratio. The lack of cross-validation can potentially result in an overly optimistic or pessimistic evaluation of the model, as it may not fully capture the model's performance across different variations of the data. The analysis of the results revealed a remarkable similarity between the three conversion methods across all three surgical tasks. This indicates that the choice of conversion method without cross-validation did not significantly impact the classification accuracy of the DTW classifier.

Upon applying cross-validation with various parameter values, incorporating an additional weight (distanced) and the equation for neighbors calculation, the observed similarity between the conversion methods was notably reduced. To

specifically compare the performance of the minimum, average and the maximum lengths conversion methods, a comparison analysis was conducted and visualized in Figure 4. The y-axis represents the accuracy achieved by the classifier when applying different K values. The x-axis represents the three conversion methods under comparison, namely the minimum, average and maximum lengths conversion methods. Based on the results presented in Figure 4, it can be concluded that across all three surgical tasks, the minimum lengths conversion method consistently outperformed the other lengths conversion methods, in terms of accuracy. These findings suggest that the minimum lengths conversion method was more effective in capturing the relevant patterns and features within the time series data, leading to improved classification performance.

The results obtained from applying DTW to three distinct surgical tasks are presented in Table 5. Only highest accuracies are included in the tables. The parameters listed in the tables refer to the maximum number of neighbors considered and the weight assigned during the model training phase ('uniform', 'distance'). In the NP and KT tasks, the minimum conversion method demonstrated better performance compared to the other two conversion methods. However, in the suturing task, the combination of the average and the maximum conversion method with LOOCV resulted in the highest accuracy.

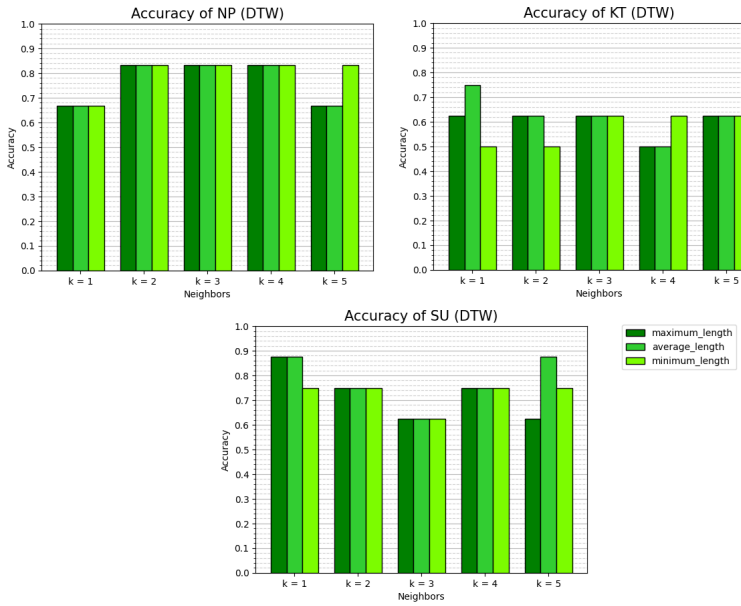

Figure 3
The best accuracies of the Dynamic Time Warping classification algorithm. Each color represents a conversion method that was used for transforming the time series data into equal length.

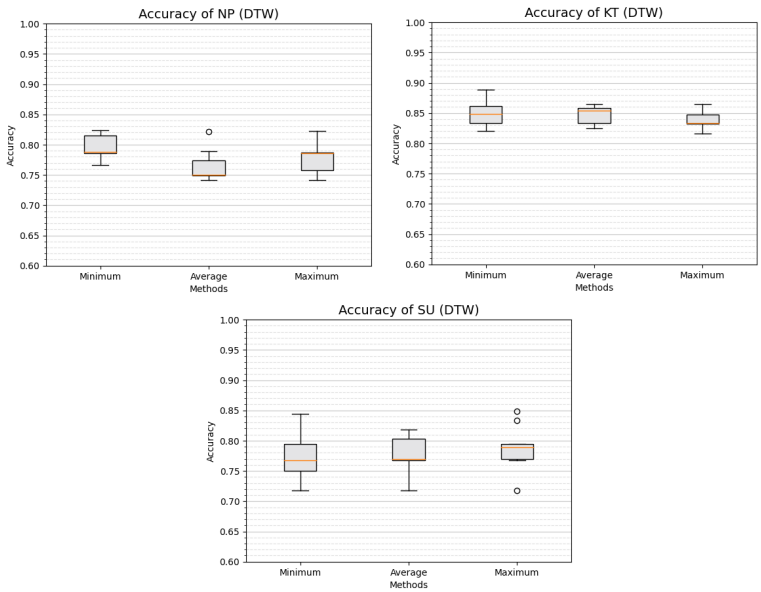

Figure 4
Box plot diagrams of the surgical tasks using Dynamic Time Warping classification algorithm. The different conversion methods (minimum, average, maximum) show slightly different results.

Table 5
Dynamic Time Warping accuracies. Used abbreviations: CV: cross-validation, K: Folds of the cross-validation, LOOCV: Leave-One-Out cross-validation, NP: needle-passing, KT: knot-tying, SU: suturing, min: minimum length, avg: average length, max: maximum length

Task	CV	Accuracy	Best parameters
NP	K = 9 (min)	**0.824074**	6, 'uniform'
	K = 2 (avg)	0.821429	5, 'uniform'
	K = 3 (max)	0.822222	14, 'distance'
KT	K = 2 (min)	**0.888889**	10, 'uniform'
	K = 16 (avg)	0.864583	1, 'uniform'
	K = 16 (max)	0.864583	4, 'uniform'
SU	K = 9 (min)	0.844444	2, 'uniform'
	LOOCV (avg)	**0.871795**	2, 'uniform'
	LOOCV (max)	**0.871795**	2, 'uniform'

The comparative results of the various 1D CNN (1D Convolutional Neural Network) approaches are presented in Table 6. Across most of the surgical tasks, the use of the ELU activation function, in conjunction with the minimum conversion method yielded superior performance compared to the other method combinations. Notably, the exception was observed in the suturing task, where the

average conversion method coupled with ReLU activation function achieved the highest accuracy.

These findings highlight the effectiveness of the ELU activation function for most surgical tasks, particularly when used in conjunction with the minimum conversion method. However, the overall accuracy achieved by the CNN approach was significantly lower compared to the other classification methods employed. This implies that further refinement of the CNN architecture, as well as exploration of additional features or teaching techniques, may be necessary to improve its performance and bridge the accuracy gap observed with other classification methods.

Table 6
Neural Network accuracies. Used abbreviations: ReLu: Rectified Linear activation function, ELU: Exponential Linear Unit, NP: needle-passing, KT: knot-tying, SU: suturing, min: minimum length, avg: average length

Task	**Method**	**Accuracy (ReLu)**	**Accuracy (ELU)**
NP	min	0.657	**0.771**
	avg	0.621	0.664
KT	min	0.761	**0.794**
	avg	0.772	0.778
SU	min	0.687	0.708
	avg	**0.723**	0.59

5 Discussion

The JIGSAWS dataset has some limitations. Primarily it contains a very little amount of data, due to the participation of only eight surgeons. Each of the participants repeated the three surgical task five times, therefore the dataset consist of 8*5 = 40 trials. Secondly, the dataset lacks trials performed by certain surgeons. Specifically, the dataset does not include the second trial conducted by surgeon H. Due to incomplete GRS evaluations, certain data instances were removed to ensure data consistency. The following specific data instances were excluded: from knot-tying: B005, H00, I004; from needle-passing: B005, E002, F002, F005, G001, G002, G003, G004, G005, H001, H003, I001; from suturing: H002. After the training of the models, the G10 gesture was removed from suturing, and the models were retrained because the cross-validation method could not work with the low amount of trial data of this gesture.

Performing parameter tuning with all the listed parameter values and subsequently conducting dimensionality reduction on the data can be computationally intensive. This is especially true when utilizing time series classifiers, as they inherently possess slower processing times. Because of the excessive need for computing resource, the classification tasks were implemented on a server

equipped with 64 CPU cores, a maximum available memory of 85GB, and a dedicated 10GB Nvidia A100 GPU.

The key features and functions have been implemented in the IROB-SAF GitHub repository (https://github.com/rlevendovics/irob-saf).

Table 7
Comparison between the results obtained in this study and those reported by another authors. All of the methods listed in the table were applied to the JIGSAWS dataset. Used abbreviations: NP: needle-passing, KT: knot-tying, SU: suturing, DTW: Dynamic Time Warping, DT: Decision Tree, SVM: Support Vector Machines, k-NN: k-Nearest Neighbors, LR: Logistic Regression, ApEn: Approximate Entropy, PCA: Principal Component Analysis, MI: Mutual Information

Author (Year)	**Method**	**NP**	**KT**	**SU**
Fard et al. (2016) [16]	DTW, k-NN	70.12 %	85.14 %	80.49 %
Zia et al. (2017) [18]	k-NN, ApEn, PCA	**100** %	99.99 %	**100** %
Fard et al. (2018) [17]	k-NN, LR, SVM	-	82.3 %	89.9 %
This study	DTW	82.4 %	88.89 %	87.18 %
This study	Neural Network	77.1 %	79.4 %	72.3 %
This study	DT, k-NN, LR, SVM, ApEn, MI	**100** %	**100** %	**100** %

Conclusions

The aim of the study was to compare algorithms capable of categorizing surgeons into two groups (novice and experienced) based on kinematic data recorded by the highly successful RAMIS robot system, the da Vinci. Zia et al. [18] achieved remarkable results on the JIGSAWS dataset, nearly attaining 1.0 accuracy for all three surgical tasks, that is showed in Table 7. In this study, similar classification methods were employed to evaluate the data. The non-time series classifiers demonstrated significantly superior performance compared to the time series classifiers (such as DTW and 1D CNN). In contrast to Zia et al., this study not only utilized the k-Nearest Neighbors (k-NN) method but also employed Decision Trees (DT), Support Vector Machines (SVM) and Logistic Regression (LR). To enhance the performance of the classification algorithms, the data underwent ApEn transformation, followed by a thorough examination of potential parameters and dimensionality reduction technique. Consequently, all surgical tasks achieved 100 % accuracy through this analysis.

The achieved accuracies could be further increased by incorporating a higher number of surgeons, thereby enhancing the robustness of classification outcomes.

This study only analyzed the DTW and the NN on the entire dataset. Given that non-time series algorithms attained higher accuracies on the splitted dataset, implementing these two classifier on the gesture level as well could be advantageous. The other two method (standardization, dimensionality reduction) could also enhance the accuracy for the time series algorithms. Beyond these considerations the achieve results have potential implications for the development of personalized training programs that target the specific deficiencies of individual surgeons.

Acknowledgment

This work has partially been supported by ACMIT (Austrian Center for Medical Innovation and Technology), which is funded within the scope of the COMET (Competence Centers for Excellent Technologies) program of the Austrian Government. T. Haidegger is a Consolidator Researcher, supported by the Distinguished Researcher program of Óbuda University.

References

[1] G. Fichtinger, J. Troccaz, and T. Haidegger. Image-guided interventional robotics: Lost in translation? *Proceedings of the IEEE*, 110(7):932–950, 2022.

[2] A. Khamis, J. Meng, J. Wang, A. T. Azar, E. Prestes, Á. Takács, I. J. Rudas, and T. Haidegger. Robotics and intelligent systems against a pandemic. *Acta Polytechnica Hungarica*, 18(5):13–35, 2021.

[3] T. Haidegger, S. Speidel, D. Stoyanov, and R. Satava. Robot-assisted minimally invasive surgery—surgical robotics in the data age. *Proceedings of the IEEE*, 110:835–846, 2022.

[4] T. D. Nagy and T. Haidegger. Performance and capability assessment in surgical subtask automation. *Sensors*, 22(7):2501, 2022.

[5] T. Haidegger, V. Mai, C. Mörch, D. Boesl, A. Jacobs, A. Khamis, L. Lach, B. Vanderborght, et al. Robotics: Enabler and inhibitor of the sustainable development goals. *Sustainable Production and Consumption*, 43:422–434, 2023.

[6] M. A. Houghtaling, S. R. Fiorini, N. Fabiano, P. J. Gonçalves, O. Ulgen, T. Haidegger, J. L. Carbonera, J. I. Olszewska, B. Page, Z. Murahwi, et al. Standardizing an ontology for ethically aligned robotic and autonomous systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2023.

[7] R. Nagyné Elek and T. Haidegger. Next in surgical data science: Autonomous non-technical skill assessment in minimally invasive surgery training. *Journal of Clinical Medicine*, 11(24):7533, 2022.

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

[9] J. Katona. Clean and dirty code comprehension by eye-tracking based evaluation using gp3 eye tracker. *Acta Polytechnica Hungarica*, 18(1):79–99, 2021.

[10] M. Koctúrová and J. Juhár. EEG-based speech activity detection. *Acta Polytechnica Hungarica*, 18(1):65–77, 2021.

[11] J. Chen, N. Cheng, G. Cacciamani, P. Oh, M. Lin-Brande, D. Remulla, I. S. Gill, and A. J. Hung. Objective assessment of robotic surgical technical skill: a systematic review. *The Journal of urology*, 201(3):461–469, 2019.

[12] D. A. Nagy, I. J. Rudas, and T. Haidegger. Surgical data science, an emerging field of medicine. In *Proc. of 2017 IEEE 30th Neumann Colloquium*, pages 59–64, 2017.

[13] R. Nagyné Elek and T. Haidegger. Robot-assisted minimally invasive surgical skill assessment—manual and automated platforms. *Acta Polytechnica Hungarica*, 16(8):141–169, 2019.

[14] I. Funke, S. T. Mees, J. Weitz, and S. Speidel. Video-based surgical skill assessment using 3d convolutional neural networks. *International journal of computer assisted radiology and surgery*, 14:1217–1225, 2019.

[15] C. E. Reiley and G. D. Hager. Task versus subtask surgical skill evaluation of robotic minimally invasive surgery. In G.-Z. Yang, D. Hawkes, D. Rueckert, A. Noble, and C. Taylor, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009*, pages 435–442, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[16] M. J. Fard, S. Ameri, and R. D. Ellis. Toward personalized training and skill assessment in robotic minimally invasive surgery. *arXiv preprint arXiv:1610.07245*, 2016.

[17] M. J. Fard, S. Ameri, R. Darin Ellis, R. B. Chinnam, A. K. Pandya, and M. D. Klein. Automated robot-assisted surgical skill evaluation: Predictive analytics approach. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 14(1):e1850, 2018.

[18] A. Zia and I. Essa. Automated surgical skill assessment in RMIS training. *International Journal of Computer Assisted Radiology and Surgery*, 13:731–739, 2017.

[19] Y. Gao, S. S. Vedula, C. E. Reiley, N. Ahmidi, B. Varadarajan, H. C. Lin, L. Tao, L. Zappella, B. B'ejar, D. D. Yuh, et al. The JHU-ISI gesture and skill assessment working set (JIGSAWS): A surgical activity dataset for human motion modeling. In *Modeling and Monitoring of Computer Assisted Interventions (M2CAI) – MICCAI Workshop*, 2014.

[20] H. Niitsu, N. Hirabayashi, M. Yoshimitsu, T. Mimura, J. Taomoto, Y. Sugiyama, S. Murakami, S. Saeki, H. Mukaida, and W. Takiyama. 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*, 43:271–275, 2013.

[21] Y.-Y. Song and L. Ying. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130, 2015.

[22] E. Yanik, X. Intes, U. Kruger, P. Yan, D. Diller, B. Van Voorst, B. Makled, J. Norfleet, and S. De. Deep neural networks for the assessment of surgical skills: A systematic review. *The Journal of Defense Modeling and Simulation*, 19(2):159–171, 2022.