Surgical Endoscopy (2019) 33:3732–3740
https://doi.org/10.1007/s00464-019-06667-4

Sensor-based machine learning for workflow detection and as key to detect expert level in laparoscopic suturing and knot-tying

Karl-Friedrich Kowalewski[1] · Carly R. Garrow[1] · Mona W. Schmidt[1] · Laura Benner[2] · Beat P. Müller-Stich[1] · Felix Nickel[1]

Received: 28 October 2018 / Accepted: 17 January 2019 / Published online: 21 February 2019

Abstract

Introduction The most common way of assessing surgical performance is by expert raters to view a surgical task and rate a trainee's performance. However, there is huge potential for automated skill assessment and workflow analysis using modern technology. The aim of the present study was to evaluate machine learning (ML) algorithms using the data of a Myo armband as a sensor device for skills level assessment and phase detection in laparoscopic training.

Materials and methods Participants of three experience levels in laparoscopy performed a suturing and knot tying task on silicon models. Experts rated performance using Objective Structured Assessment of Surgical Skills (OSATS). Participants wore Myo armbands (Thalmic Labs™, Ontario, Canada) to record acceleration, angular velocity, orientation, and Euler orientation. ML algorithms (decision forest, neural networks, boosted decision tree) were compared for skill level assessment and phase detection.

Results 28 participants (8 beginner, 10 intermediate, 10 expert) were included, and 99 knots were available for analysis. A neural network regression model had the lowest mean absolute error in predicting OSATS score (3.7 ± 0.6 points, $r^2 = 0.03 \pm 0.81$; OSATS min.-max.: 4–37 points). An ensemble of binary-class neural networks yielded the highest accuracy in predicting skill level (beginners: 82.2% correctly identified, intermediate: 3.0%, experts: 79.5%) whereas standard statistical analysis failed to discriminate between skill levels. Phase detection on raw data showed the best results with a multi-class decision jungle (average 16% correctly identified), but improved to 43% average accuracy with two-class boosted decision trees after Dynamic time warping (DTW) application.

Conclusion Modern machine learning algorithms aid in interpreting complex surgical motion data, even when standard analysis fails. Dynamic time warping offers the potential to process and compare surgical motion data in order to allow automated surgical workflow detection. However, further research is needed to interpret and standardize available data and improve sensor accuracy.

Keywords Myo armband · Machine learning · Neural networks · Laparoscopy · Surgical education · Electromyography · Skill assessment · Workflow analysis · Artificial intelligence · Laparoscopic training

The benefits of laparoscopic surgery for patients have been extensively studied in the literature and range from smaller incisions to faster healing and less pain [1–4]. However, challenges still remain for surgeons, as laparoscopic surgery has a prolonged learning curve, requiring longer training periods [5–7]. Therefore, laparoscopic surgery requires safe and realistic training environments for trainees to practice before performing surgery on an actual patient. The most common method for assessing surgical performance requires expert raters to view a surgery or training task and rate a trainee's performance based on global or procedure-specific checklists [5, 8]. However, expert rating is expensive,

Karl-Friedrich Kowalewski and Carly R. Garrow contributed equally to the manuscript.

✉ Felix Nickel
felix.nickel@med.uni-heidelberg.de

[1] Department of General, Visceral, and Transplantation Surgery, University of Heidelberg, Im Neuenheimer Feld 110, 69120 Heidelberg, Germany

[2] Department of Medical Biometry and Informatics, University of Heidelberg, Im Neuenheimer Feld 130.3, 69120 Heidelberg, Germany

subjective and time intensive [9], and it can result in delayed feedback to the trainee, hampering trainee learning.

The use of technology and artificial intelligence in medicine and, more specifically, surgery, has grown tremendously in the past decade. Various research groups have attempted to use automatic data analysis to detect surgical phase [10–13] and classify surgeon skill level [14–17], with varying degrees of success. However, difficulties still remain, as some methods of data collection, such as sensors, cannot be easily implemented in the OR due to patient safety and sterility reasons. Likewise, some machine learning (ML) algorithms, a subsection of artificial intelligence, require a large amount of computational power and can take a large amount of time and data to train [9].

A new technology that enables intraoperative assessment of hand and forearm motion parameters is the Myo™ armband (Thalmic Labs Inc., Kitchener, Ontario, Canada). The Myo is a commercially available armband that contains an inertial measurement unit, which measures the acceleration, angular velocity, and orientation of the arm, as well as calculates the Euler orientation from these metrics. Angular velocity measures the speed at which the arm is rotating about the *x, y*, and *z* axes, while orientation provides the absolute direction the arm is positioned in space. Euler orientation is a second method for describing an object's orientation and represents the current angle of the arm with respect to the *x, y*, and *z* axes, which are defined as roll, pitch, and yaw. It additionally contains eight electromyographic (EMG) sensors, which read and record the electrical activity of the arm muscles. Because of its open application programming interface, developers have investigated using the Myo for a variety of applications, such as rehabilitation therapy [18–20] or hand hygiene training [21]. In the broader field of surgery, the Myo has been proposed for maneuvering 3D image-guided surgery systems [22, 23], but has not been investigated for surgical training or skill assessment purposes.

It is hypothesized that recordings from the Myo armband, in combination with ML, can be used to distinguish skill level and recognize the phases of a laparoscopic suturing and knot tying task. Using the Myo to perform these tasks has the potential to relieve time constraints for expert surgeons, deliver immediate performance feedback to trainees, and provide standardized and objective skill assessment. In addition, it can be easily slipped on and off and can be worn under the sterile gowns in the operating room (OR), eliminating problems with the implementation of new technology in the OR.

The aim of the present study was to evaluate the use of ML in combination with arm motion data recorded with the Myo during a suturing and knot-tying task to (1) distinguish skill level and (2) identify the procedural phase. Secondly, the study aimed to test the feasibility of the Myo as a training tool in its current stage.

Methods

Study design

This study was carried out at the Department of General, Visceral and Transplantation Surgery at Heidelberg University, Germany. Before starting, all participants received detailed information about the study and its purpose, and informed consent was obtained. The study previously received local ethics committee approval (S 334/2011, Amendment 07/2015). Participants were divided in advance into one of three skill level groups (beginner: no experience in laparoscopy, intermediate: < 50 knots performed, or expert: > 50 knots performed or attending surgeons) based on their previous experience in the OR.

Data collection

Each participant tied four square knots in a laparoscopic box trainer while wearing a Myo armband on each arm. Each knot was tied according to the modified standardized checklist as described by Romero et al. [24] and first introduced by Munz et al. [25] on a custom-made silicon suture pad developed in-house. Before the first knot, all participants watched an introduction video twice in order to standardize surgical technique. All of the parameters listed in Table 1 were collected from the Myo armbands. The endoscopic video, as well as instrument coordinates and time stamp from an NDI Polaris camera (Northern Digital Inc., Waterloo, Ontario, Canada), were recorded and synchronized for each knot. Thus, at each time point, a video image, instrument coordinates, time stamp, and the Myo data were captured. This allowed the assignment of a surgical task (by the video image) to each time point (Polaris time stamp). Eventually, the time stamps of the Polaris and the Myo were matched. By doing so, the Myo data and actual surgical phase could be matched more accurately compared to other approaches (e.g. observing the task and noting time points in real time). The endoscopic video was manually annotated into phases by frame number

Table 1 List of data collected from the Myo armband

Data type	Direction
Acceleration	x, y, z
Angular velocity	x, y, z
Orientation	x, y, z, w
Euler orientation	Roll, pitch, yaw

Table 2 List of phases identified for suturing and knot-tying task

Knot tying phases	
1. Grasp needle	8. Create the anti-C loop
2. Drive needle 1	9. Wrap the suture 2
3. Drive needle 2	10. Tighten the knot 2
4. Create C loop 1	11. Create the C loop 2
5. Wrap the suture 1	12. Wrap the suture 3
6. Grasp end of suture 1	13. Tighten the knot 3
7. Tighten the knot 1	

(Table 2), as established in a previous study [26]. Unless otherwise specified, all data analysis was completed in Matlab (MathWorks, Natick, Massachusetts, USA).

Expert rating

Each knot was rated by an expert rater blinded to the trainee's experience. The score, a modified version of the OSATS originally introduced by Chang et al., was used as a performance measure [8, 27].

Skill level discrimination

To assess skill level, each of the parameters given in Table 1 were averaged for each knot, based on the methods from Brown et al. [28]. Calculating the average rather than using the entire time series removes the bias of time, as beginners generally tended to take longer to tie a knot than experts, while simplifying the computations significantly as compared to other methods for removing time bias, such as dynamic time warping (DTW) or hidden Markov models. Thus, each knot had 13 averages (for each of the Table 1 parameters) included as inputs.

Phase detection

To predict the knot tying phase, the time series of each parameter were first split into phases based on the video annotations. Then, the average of each phase's time series was calculated for each knot. As described below, the average was calculated from both the raw recorded data and time warped data.

Data analysis and testing of ML algorithms for skill level discrimination and phase detection were two separate tasks performed in this study. Thus, it is important to note that inaccuracies in one would not have affected the outcome of the other analysis.

Dynamic time warping

DTW, originally introduced for speech recognition applications, has been widely used since its introduction to match together time series that demonstrate similar patterns over varying lengths of time [29]. In order to appropriately compare these data series, the motion data of each time series must be matched to each other based on its similarities, and the bias of time must be removed (e.g. a certain movement of a beginner takes longer than the same movement of an expert, making direct comparison difficult). DTW is an algorithm capable of stretching or compressing the data to achieve comparability. Therefore, DTW was applied to the data using a modified version of the Matlab DTW code provided by Quan Wang [30]. It was determined that DTW should be applied to each phase separately to prevent the possibility of one phase being warped to another phase. Therefore, the data was first split into its 13 respective phases based on the manual annotations. Because DTW can only be used to compare two time series at once, the longest time series in each phase was then identified and used as a reference to which all other time series of the same phase were warped.

Machine learning

Microsoft Azure Machine Learning Studio (Microsoft Corporation, Redmond, Washington, USA) was used to compare various ML methods. The online program uses a graphic-based programming interface and already has standard ML algorithms built into the program. The user can specify the desired ML algorithm as well as its traits.

Because two different methods were available to describe skill level (OSATS score or skill level group), both regression and classification ML methods were tested. Regression ML methods were used to predict a range of real numbers based on the input data. In this case, regression ML was used to predict a participant's OSATS score using the Myo data as input information. On the other hand, classification ML predicts a participant's group based on the input data. In this case, the prediction was whether a participant belonged in the beginner, intermediate, or expert group.

The Microsoft Azure program functioned as follows: for each ML model, the parameters were first tuned, and then cross validation was used to train, test, and validate the model. In each case the data was randomly separated into tenfold, meaning that onefold would equal one-tenth of the data. The model was trained and tested on nine of the tenfolds. Once the best model was achieved based on the data in these ninefolds, it was then validated, or tested again, on the last fold of data. This process was repeated, using each of the tenfolds as the validation fold once. The results achieved for each of the validation folds were averaged together to get the

average performance of the model on the data. Thus, when the motion data is input into Microsoft Azure and fed into a specific algorithm, it finds the best model for the data, and outputs the model's average performance.

Statistical comparison

Statistical analysis was performed to provide a baseline comparison of the data by an employee of the Department of Medical Biometry and Informatics at Heidelberg University who was otherwise not involved in the study. An ANOVA repeated measurements test was used, which demonstrated if a significant difference existed within one of the Myo parameter categories motion categories. If statistical significance was found, this was further investigated with Post-Hoc tests using Tukey–Kramer assessment to determine between which groups the difference existed.

Results

28 participants (8 beginner, 10 intermediate, 10 expert) took part in the study, tying a total of 112 knots. 13 of the 112 knots were removed from the analysis due to faulty recording ($n=11$) or ruptured suture pad during the knot tying process ($n=2$), leaving 99 knots available for analysis. Ground truth OSATS score (provided by the expert raters) was 21.3 ± 4.3, 27.5 ± 3.7 and 33.6 ± 1.8 for beginners, intermediates and experts, respectively.

Skill level discrimination

Various regression ML methods were tested to predict a participant's OSATS score based on their input data. Results were investigated both with and without the knot number included in the input data to investigate whether the ML algorithm performed better given the information of whether the knot was, for example, the participant's 1st or 4th. After tuning and cross validation, the findings presented by Microsoft Azure showed that a neural network regression model had the lowest mean absolute error when compared to other ML regression models, at 3.7 ± 0.6 OSATS points ($R^2 = 0.03 \pm 0.81$; min.-max. OSATS: 4–37 points). For example, to interpret this error, a real OSATS score of 20 would be on average predicted to be 23.7 (or 16.3) by the neural network regression model. The coefficient of determination, or R^2 value, describes how well the line of best fit actually fits the data. A value of 100% would be the perfect fit, which means the best fit formula (and its respective graph/curve) would contain all data points which were actually measured, and 0% would be no fit (no data point on the best fit curve). When knot information was not included, the mean absolute error was only minimally larger (3.82 ± 0.78 OSATS points, $R^2 = 0.07 \pm 0.75$). A plot demonstrating real OSATS scores versus those predicted by the neural network with good correlation can be found in Fig. 1. Error results of the neural network regression model, as compared to other ML algorithms, can be found in Table 3.

Classification ML methods were also investigated to predict skill level group. The best model found after tuning and cross validation was an ensemble of binary-class neural networks, with an average prediction accuracy of 0.704, meaning that the real skill level group was correctly predicted

Table 3 Comparison of results between various ML algorithms for OSATS score prediction

ML algorithm type	Mean error	r^2
Decision forest	4.45 ± 0.75	-0.2 ± 0.69
Neural networks	3.71 ± 0.64	0.03 ± 0.81
Boosted decision tree	4.43 ± 0.60	0.19 ± 0.46

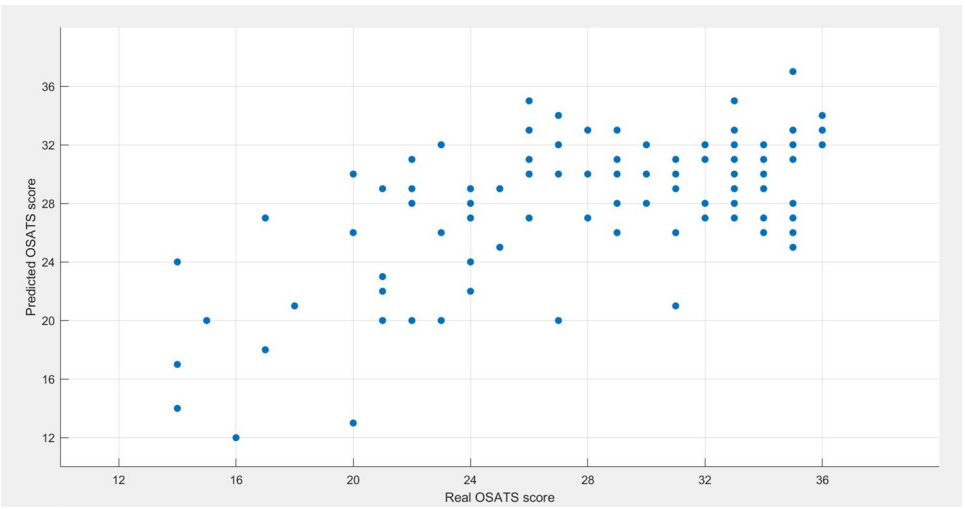

Fig. 1 Plot of real versus predicted OSATS scores from neural network regression model. Plot of real OSATS (from expert raters) score versus OSATS scores predicted by the neural network regression model. Results are from cross-validation. The correlation was good with $r=0.62$ and $p<0.001$ as calculated in Excel

70% of the time. Whether knot information was included or not did have a small effect on the prediction accuracy, precision, and sensitivity for intermediates and experts; namely, when knot information was included, intermediate prediction accuracy improved slightly, while expert accuracy decreased slightly. Results can be found in Fig. 2, and a comparison with other ML algorithms can be found in Table 4.

Phase detection

Phase detection on raw data showed the best results with a multi-class decision jungle (comparison with other ML methods not shown). As shown in Fig. 3, the average accuracy was only 16%, and only 1 phase achieved a correct identification rate over 50%. The best model on warped data was an ensemble of two-class boosted decision trees. With DTW, the average accuracy increased to 43%, and 5 out of 13 phases had an identification rate of 50% or higher (Fig. 3, Part B). Three of the phases (Drive the needle 2, Grasp the suture 1, Tighten the knot), were correctly identified more than 60% of the time. Whether the knot number was included in the input data only made a minimal difference in the phase detection results (data not shown). Comparison with other ML methods can be found in Table 5. All data shown are results after parameter tuning and cross validation.

Standard statistical analysis did not find a significant difference between the motion parameters of any of the three groups, and thus the data could not be used to statistically determine a participant's skill level (data not shown).

Table 4 Comparison of results between various ML algorithms for skill level group prediction

ML algorithm type	Accuracy	Precision	Recall
Decision jungle	0.62	0.43	0.43
Neural network	0.70	0.56	0.56
Support vector	0.60	0.39	0.39
Boosted decision tree	0.66	0.56	0.56

Each algorithm type tested for skill level prediction was an ensemble of two-class classification algorithms

Discussion

The present study aimed to predict skill level and identify phases during a suturing and knot tying task using data from the Myo armband, which uses EMG and other sensors together with artificial intelligence. Traditional statistical test yielded no difference between experience levels. However, the application of modern ML algorithms for classification using an ensemble of binary class neural networks proved to be an effective approach at determining skill level using the Myo data, while regression ML was less successful. Additionally, phase recognition algorithms demonstrated improved results with boosted decision trees after DTW was performed as a means to improve comparability between data.

Skill assessment

For prediction of the OSATS score, a neural network regression model demonstrated the best results and led to the lowest average score prediction error, with an error of 3.71 OSATS points, or about 10% of the score. While this

A With Knot Info

Actual Class \ Predicted Class	Beg	Int	Exp
Beg	85.2%	3.7%	11.1%
Int	33.3%	12.1%	54.4%
Exp	17.9%	10.3%	71.8%

B Without Knot Info

Actual Class \ Predicted Class	Beg	Int	Exp
Beg	85.2%	3.7%	11.1%
Int	39.4%	3.0%	57.6%
Exp	12.8%	7.7%	79.5%

C

	Avg Accuracy	Precision Beg	Precision Int	Precision Exp	Sensitivity Beg	Sensitivity Int	Sensitivity Exp
With Knot Info	0.704	0.57 ± 0.32	0.25 ± 0.10	0.48 ± 0.29	0.76 ± 0.36	0.10 ± 0.17	0.7 ± 0.35
W/o Knot Info	0.704	0.62 ± 0.32	0.10 ± 0.31	0.48 ± 0.29	0.83 ± 0.33	0.03 ± 0.08	0.77 ± 0.35

Fig. 2 Results of binary-class neural network classification model for skill level group prediction. The darker the shade of green, the higher the phase prediction accuracy. **A** Best confusion matrix results for class prediction with knot number included and **B** without knot number included. **C** Average accuracy, precision, and sensitivity reported when knot information is and is not included. (Color figure online)

A Without DTW

Actual Phase \ Predicted Phase	1	2	3	4	5	6	7	8	9	10	11	12	13
1	5	18	2	10	8	8	6	8	6	1	11	5	12
2	3	59	7	3	4		5		6	2	1	10	
3	7	49	4	1	8	1	11	5	1	2	3	8	
4	5	15	6	9	5	2	8	8	5	1	13	12	11
5	5	30	7	5	10	3	4	3	6		10	14	3
6	4	10	4	1	1	47		11		1	6	3	12
7	6	40	8	3	4	2	9	6	7	2	1	6	6
8	8	20	2	7	1	9	4	10	1	3	11	6	18
9	4	26	3	4	5	5	4	4	10	1	11	21	2
10	9	38	9	6	3	1	6	5	3		6	6	8
11	5	11	6	9	10	7	3	11	12	4	12	7	3
12	3	19	4	6	15	4	2	2	14	2	8	17	4
13	4	9	2	3	3	9	10	7	2	3	9	11	28

B With DTW

Actual Phase \ Predicted Phase	1	2	3	4	5	6	7	8	9	10	11	12	13
1	30	9	7	4	11	4	4	11	4	5	1	1	9
2	12	59	6	1			5	6	1	9		1	
3	6	5	66	2	3	1	3			3		7	4
4	5	7	5	33	9	11	2	8	7	3	8	2	
5	4	1	4	5	42	2	4	4	19	10	2	3	
6	1			8	2	61	7	5	2	3	6	2	3
7	5	4	4		3	1	61	5		4		5	8
8	8	9	3	14	4	10	11	13	6	11	4	4	3
9	1	1		2	15	2	3	1	56	5	6	7	1
10	4	11	3	3	14		6	13	7	26	6	7	
11	2	2	4	7	4	6	4	3	12	2	24	23	7
12	1		8	1	8	1	5	6	7	3	17	38	5
13	4	1	9		1	3	16	6			6	5	49

C	Accuracy	Precision	Sensitivity
No DTW	0.16	0.16	0.16
With DTW	0.43	0.43	0.43

Fig. 3 Results of classification machine learning for phase identification. The darker the shade of green, the higher the phase prediction accuracy (in %). **A** Best confusion matrix results for phase prediction before dynamic time warping with the multi-class decision jungle. **B** Best confusion matrix results for phase prediction after dynamic time warping with an ensemble of two class boosted decision trees. **C** Average accuracy, precision, and recall for phase identification both before and after dynamic time warping. (Color figure online)

error seems acceptable, the coefficient of determination, or R^2 value, which describes how well the line of best fit actually fits the data, was less than 10%, regardless of whether the knot information was included or not. It is worth mentioning that the boosted decision tree regression model demonstrates a higher R^2 value than the neural network model, as seen in Table 3. However, because the R^2 value is still relatively low and the error is higher than the neural network model, the neural network model was chosen as the best-performing regression model in the present study.

In comparison, the (skill level) classification model showed better results than the regression model. The best-performing model was an ensemble of binary-class neural networks, and the model without knot information included seemed to provide slightly better results. The model performed very well in identifying beginners and experts, with

Table 5 Comparison of results between various ML algorithms for phase prediction

ML algorithm type	Accuracy	Precision	Sensitivity
Decision jungle	0.4	0.4	0.4
Neural network	0.18	0.18	0.18
Support vector	0.21	0.21	0.21
Boosted decision tree	0.43	0.43	0.43

Each algorithm type tested for skill level prediction was an ensemble of two-class classification algorithms

a prediction rate equal to or greater than 80%. However, the model had limitations in identifying intermediates. As indicated in Fig. 2, it can be concluded that the model would rather place the intermediates into either the beginner or expert category. A review of other publications in the literature has shown that many chose to differentiate skill level between only novices and experts, though these two groups have varying definitions in these papers [14, 31, 32]. Rosen et al. was successful in using hidden Markov models to differentiate between 4 different skill levels, though the authors used force/torque readings from laparoscopic instruments [33]. Oropesa et al. compared 2D and 3D motion metrics extracted from laparoscopic video for novices, residents, and experts, finding that significant differences existed only between novices and the two other groups [34]. Thus, while defining an intermediate level based on this study's measured parameters proved difficult for the algorithm, it has also been problematic in other studies and in practice has no clear definitions. In addition, neural networks algorithms are not fully transparent on how they work, so it is not clear exactly which characteristics the neural networks used to identify skill level in this case. Thus, the results found in the present paper seem to be reasonable and in-line with expectations set forth from previous studies.

Another factor worth discussing is the precision and sensitivity of the classification ML algorithm. The sensitivity, or the proportion of correct guesses for a group over the real number of participants in that group, was fairly good for beginners and experts, as a high proportion of the participants in these groups were correctly identified. The precision, or the number of correct guesses for a group divided by the total number of guesses in that group, also showed better results for novices and experts compared to intermediates, though the absolute percentages were lower as compared to sensitivity. However, this is understandable, as the algorithm tended to predict intermediates in either the expert or beginner group, causing precision to decrease. As intermediates were defined as "the group between the beginners and experts," it is easy to understand the algorithm's limitation. Interestingly, the traditional statistical testing with ANOVA was not able to discriminate between experience levels at all, probably due to the moderate sample size. In contrast, ML might consider aspects that are too complex for common established methods. To summarize, ML methods for regression and classification of raw data obtained with the Myo armband show promising results for skill assessment. These methods seem to be superior to traditional statistical test methods. The combination of Myo's contact- and line-of-sight-free tracking with modern ML algorithms can thus be used to monitor surgeon's skill level and progress.

Phase detection

Before DTW, the average phase identification was only at 16%. DTW did improve identification by 27%, leading to an average accuracy of 43%. Five of the 13 phases, however, did show an identification rate above 50%, and three of the phases (Drive the needle 2, Grasp the suture 1, Tighten the knot), were correctly identified more than 60% of the time. Before DTW, the algorithm predicted a majority of the phases as Phase 2, although it is unclear why. DTW seemed to remove this effect, improving accuracy for the other phases. Interestingly, after DTW, the algorithm did not seem to wrongly classify phases that, at least to the human eye, looked very similar, such as Phases 2 and 3 or Phases 4 and 7. However, prediction accuracy still needs to be improved to a higher level. In the present study, the Myo measurements from each time point were first collected as a time series and then averaged, likely removing much of the motion information necessary to accurately detect a phase. A promising approach to improve phase detection would be to investigate the vector of measurements over time of the Myo parameter signals. Other authors have been successful in using similar feature vectors collected from instrument use in combination with DTW to perform phase recognition [10, 12].

The knot number was also investigated as a possible piece of information that could help improve ML prediction results. In both the skill level discrimination and phase detection results, whether the knot number was included seemed to make hardly any difference in prediction results, helping to confirm that each of the knots tied by a single participant did not differ wildly in their motion parameters, as may have been suggested by the statistical analysis.

DTW seems to be a valuable method for standardizing complex sensor motion data in surgical training. This gains more importance as the amount of available data in healthcare and surgery increases exponentially [35, 36]. With current technology in the OR, almost all available information can be considered to be a sensor stream and can potentially be used for workflow analysis [37]. In the present study, the Myo armband was primarily used to evaluate surgical experience, but eventually also proved feasible as a sensor in terms of surgical phase detection. Likewise, the tested ML

algorithms can potentially be applied to all different kinds of sensor streams and are not limited to the Myo. The most promising approach would probably be to incorporate as many sensors as possible in order to account for inaccuracies of a single sensor and make analysis more robust.

Machine learning

Microsoft Azure Machine Learning Studio proved to be a useful and fast tool for using ML algorithms, without requiring hours of coding. As far as could be found in the literature, no other study in this area has used Azure as their ML platform. We found this tool to be especially useful for those working in the medical community who may not have a background in programming or computer science but have a good understanding of ML algorithms and would like to apply these techniques to their data. However, one limitation of Azure seems to be that the program cannot accept vectors as inputs. Therefore, it is not possible to use Azure in real-time or to investigate finer patterns within the data. This should rather be done in multidisciplinary teamwork where software developers work together with practicing surgeons.

Limitations and future work

One of the most difficult tasks was matching the raw data of the Myo armbands with its actual counterparts of the laparoscopic tasks. It is not possible to rule out that this transformation has led to some inaccuracies of the recorded data. However, by matching the Myo time stamp with the Polaris' optical tracking system, the potential error was reduced to a minimum In addition, the Myo armbands do not require calibration, which may negatively affected the data quality but could also be considered a strength, since a fast and easy setup is mandatory for a successful integration in the surgical workflow. The Microsoft Azure program prevented the use of entire motion vectors as inputs, which may have led to better phase prediction results. Therefore, future work should focus on using entire vectors from the collected Myo parameters to investigate phase recognition. The skill level classification can be further studied in terms of improving the discrimination of the intermediate group from the groups of experts and non-experts. Based on the results of the present study it would make sense to simply leave out the intermediate group to improve discrimination, however this might not fully represent clinical reality in surgical training. It could additionally be investigated if an OSATS cutoff should be defined and set, only above which someone is considered an expert, as suggested by Fard et al. [32]. Future skill and phase detection should be investigated using a Myo on both arms, as each arm's motions are complimentary to each other.

Conclusion

Modern machine learning algorithms aid in interpreting complex surgical motion data, even when standard analysis fails. Dynamic time warping offers the potential to process and compare surgical motion data in order to allow automated surgical workflow detection. The Myo was chosen as a data source and tested for its feasibility as an alternative for surgical skill assessment. However, further research is needed to interpret and standardize available data and improve sensor accuracy. Finally, the Myo is not yet usable as a stand-alone training tool without further development and data processing.

Compliance with ethical standards

Disclosure Felix Nickel reports receiving travel support for conference participation as well as equipment provided for laparoscopic surgery courses by KARL STORZ, Johnson & Johnson, Intuitive, and Medtronic. Karl-Friedrich Kowalewski, Carly R. Garrow, Mona W. Schmidt, Laura Benner and Beat Müller-Stich have no conflicts of interest or financial ties to disclose.

References

1. Delaney CP et al (2003) Case-matched comparison of clinical and financial outcome after laparoscopic or open colorectal surgery. Ann Surg 238(1):67
2. Reza M et al (2006) Systematic review of laparoscopic versus open surgery for colorectal cancer. Br J Surg 93(8):921–928
3. Nguyen KT et al (2011) Comparative benefits of laparoscopic vs open hepatic resection: a critical appraisal. Arch Surg 146(3):348–356
4. Shabanzadeh DM, Sørensen LT (2012) Laparoscopic surgery compared with open surgery decreases surgical site infection in obese patients: a systematic review and meta-analysis. Ann Surg 256(6):934–945
5. Vassiliou MC et al (2005) A global assessment tool for evaluation of intraoperative laparoscopic skills. Am J Surg 190(1):107–113
6. Miskovic D et al (2012) Learning curve and case selection in laparoscopic colorectal surgery: systematic review and international multicenter analysis of 4852 cases. Dis Colon Rectum 55(12):1300–1310
7. Nickel F et al (2016) Sequential learning of psychomotor and visuospatial skills for laparoscopic suturing and knot tying-a randomized controlled trial "The shoebox study" DRKS00008668. Langenbecks Arch Surg 401(6):893–901
8. Martin J et al (1997) Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg 84(2):273–278
9. Loukas C (2017) Video content analysis of surgical procedures. Surg Endosc 32:553
10. Ahmadi S-A et al (2006) Recovery of surgical workflow without explicit models. In: International Conference on medical image computing and computer-assisted intervention. Springer, New York

11. Bardram JE et al (2011) Phase recognition during surgical procedures using embedded and body-worn sensors. IT University of Copenhagen, Copenhagen
12. Padoy N et al (2007) A boosted segmentation method for surgical workflow analysis. In: International Conference on medical image computing and computer-assisted intervention. Springer, New York
13. Katic D et al (2016) Bridging the gap between formal and experience-based knowledge for context-aware laparoscopy. Int J Comput Assist Radiol Surg 11(6):881–888
14. Rosen J et al (2001) Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills. IEEE Trans Biomed Eng 48(5):579–591
15. Reiley CE et al (2011) Review of methods for objective surgical skill evaluation. Surg Endosc 25(2):356–366
16. Spangenberg N et al (2017) Method for intra-surgical phase detection by using real-time medical device data. In: IEEE 30th International Symposium on computer-based medical systems
17. Ganni S et al (2018) A software-based tool for video motion tracking in the surgical skills assessment landscape. Surg Endosc 32(6):2994
18. Lin P-J, Chen HY (2018) Design and implement of a rehabilitation system with surface electromyography technology. In: 2018 IEEE International Conference on applied system invention (ICASI). IEEE
19. Ryser F et al (2017) Fully embedded myoelectric control for a wearable robotic hand orthosis. IEEE Int Conf Rehabil Robot 2017:615–621
20. Sathiyanarayanan M, Raja S (2016) Myo armband for physiotherapy healtchare: a case study using gesture recognition application
21. Kutafina E et al (2016) Wearable sensors for eLearning of manual tasks: using forearm EMG in hand hygiene training. Sensors (Basel) 16(8):1221
22. Jimenez DA et al (2016) Human-computer interaction for image guided surgery systems using physiological signals: application to deep brain stimulation surgery. In VII Latin American Congress on Biomedical Engineering CLAIB 2016, Bucaramanga, Santander, Colombia, October 26th–28th, 2017. Springer
23. Sanchez-Margallo FM et al (2017) Use of natural user interfaces for image navigation during laparoscopic surgery: initial experience. Minim Invasive Ther Allied Technol 26(5):253–261
24. Romero P et al (2014) Intracorporal suturing—driving license necessary? J Pediatr Surg 49(7):1138–1141
25. Munz Y et al (2007) Curriculum-based solo virtual reality training for laparoscopic intracorporeal knot tying: objective assessment of the transfer of skill from virtual reality to reality. Am J Surg 193(6):774–783
26. Kowalewski K-F et al (2016) Development and validation of a sensor- and expert model-based training system for laparoscopic surgery: the iSurgeon. Surg Endosc 31:2155
27. Chang OH et al (2015) Developing an objective structured assessment of technical skills for laparoscopic suturing and intracorporeal knot tying. J Surg Educ 73:258
28. Brown JD et al (2017) Using contact forces and robot arm accelerations to automatically rate surgeon skill at peg transfer. IEEE Trans Biomed Eng 64(9):2263–2275
29. Sakoe H, Chiba S (1978) Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans ASSP 26(1):43
30. Wang Q (2013) Dynamic Time Warping (DTW). MathWorks file exchange. https://www.mathworks.com/matlabcentral/fileexchange/43156-dynamic-time-warping-dtw
31. Ahmidi N et al (2010) Surgical task and skill classification from eye tracking and tool motion in minimally invasive surgery. Med Image Comput Comput Assist Interv 13(Pt 3):295–302
32. Fard MJ et al (2018) Automated robot-assisted surgical skill evaluation: predictive analytics approach. Int J Med Robot 14(1):e1850
33. Rosen J et al (2001) Objective laparoscopic skills assessments of surgical residents using Hidden Markov Models based on haptic information and tool/tissue interactions. Stud Health Technol Inform 81:417–423
34. Oropesa I et al (2013) EVA: laparoscopic instrument tracking based on endoscopic video analysis for psychomotor skills assessment. Surg Endosc 27(3):1029–1039
35. Murdoch TB, Detsky AS (2013) The inevitable application of big data to health care. Jama 309(13):1351–1352
36. Raghupathi W, Raghupathi V (2014) Big data analytics in healthcare: promise and potential. Health Inf Sci Syst 2(1):3
37. Kenngott HG et al (2016) Intelligent operating room suite: from passive medical devices to the self-thinking cognitive surgical assistant. Chirurg 87(12):1033–1038