This full text paper was peer-reviewed at the direction of IEEE Instrumentation and Measurement Society prior to the acceptance and publication.

Step-Wise Feature Selection for Surgical Gesture Recognition

Alec Cotton[1], Kade MacWilliams[1], James R. Green[1], Ahmed Nasr[2], Georges Azzie[3], and Carlos Rossa[1]
[1]*Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada*
[2]*Division of Paediatric Surgery, Children's Hospital of Eastern Ontario, Ottawa, ON, Canada*
[3]*Department of Surgery, Hospital for Sick Children, Toronto, ON, Canada*

***Abstract*—Laparoscopic surgical procedures often enable faster recovery, and less patient trauma when compared with open surgery. However, laparoscopic surgery has a steep learning curve, requiring surgeons to master complex stroke mechanics. Laparoscopic simulators are used to practice procedures and surgical movements in a safe and controlled environment. Modern simulators and robotic platforms can collect large volumes of kinematic data for post-processing and objective skill assessment. However, much of these data may be redundant, and identifying the key variables that predict performance would allow these simulators to provide more concise and focused feedback to trainees. In this paper, we apply stepwise feature reduction to determine the minimal set of kinematic dimensions required for accurate Global Rating Scale (GRS) regression and gesture classification using the JHU-ISI Gesture and Skill Assessment Working Set (JIGSAW) robotic surgical dataset. The results show that, from 76 original kinematic variables, only a small subset is sufficient to capture patterns indicative of surgical actions and expertise level. These results could allow simulators to provide streamlined actionable feedback while reducing dependence on expert supervision.**

I. Introduction

Laparoscopy is a form of minimally invasive surgery in which operations are performed through small incisions using long, thin instruments and an internal camera [1]. Compared with open surgery, laparoscopy results in less tissue trauma, shorter hospital stays, and reduced recovery times for patients [1], [2], [3]. However, laparoscopy has a steep learning curve. Procedures are performed using a 2D video feed with a limited field of view, leading to reduced depth perception. Surgeons must practise complex task sequences to develop 3D spatial awareness, to refine instrument handling, and to become proficient in stroke mechanics [4], [5].

Several laparoscopy training methods have been developed, including cadaveric training, intraoperative experience, and simulation-based systems [6]. A basic laparoscopy simulator consists of a box through which surgical tools are inserted and an internal camera that provides video feedback as the trainee performs training drills such as a bean drop and needle passing. These simulators offer a safe environment in which to practise, allowing surgeons to repeat primitive surgical gestures and tasks. Training with box simulators has been shown to improve overall performance after 30-35 repetitions [7]. However, these simulators often require expert supervision, as trainees typically require guidance to learn how to perform training drills correctly.

Advanced surgical simulators and robotic platforms have sensors that track tool motion and collect data to assess performance and provide feedback relative to expert benchmarks. On-board models that can automatically evaluate performance and provide guidance can allow trainees to practise without one-on-one supervision. However, providing feedback across a multitude of performance metrics can be overwhelming for the trainees and may not translate into truly actionable feedback. For example, the daVinci Surgical Simulator measures and reports metrics over 76 kinematic variables [8]. Feedback on all these variables simultaneously is not interpretable, creates cognitive overload, and reduces the effectiveness of a training session. Such high-dimensional data also pose the risk of emphasizing irrelevant metrics that do not provide noticeable improvement in surgical performance [9].

Focusing on only essential variables would reduce the amount of information that a trainee must process, and offer more specific and actionable feedback. This would allow a trainee to focus training to factors that have been shown to differentiate novice proficiencies from expert proficiencies.

Dimensionality reduction can be employed to focus feedback on truly discriminative variables. Principal component analysis (PCA) has been used extensively for kinematic dimension reduction [10], but fails to preserve the original dimensions of the kinematic data, and rather constructs new components that may not be intuitively interpretable for trainees. Dimensional reduction should instead be guided by identifying variables that are predictive of quantifiable surgical behaviours and proficiency metrics. One way is to determine which kinematic variables best discriminate between distinct surgical gestures or skill levels.

Many models, such as dynamic time warping, hidden Markov models, convolutional neural networks, and recurrent

The first two authors contributed equally and share first authorship. Email: {aleccotton, kademacwilliams}@carleton.ca (A. Cotton, K. MacWilliams); jrgreen@sce.carleton.ca (J. Green); anasr@cheo.on.ca (A. Nasr), georges.azzie@sickkids.ca (G. Azzie), carlos.rossa@carleton.ca (C. Rossa).

This research is supported by the Natural Sciences and Engineering Research Council - Canadian Graduate Scholarship (Master's), and the Children's Hospital Academic Medical Organization (CHAMO) innovation fund.

Cette recherche est appuyée par le Conseil de recherches en sciences naturelles et en génie - Bourse d'études supérieures du Canada (Maîtrise) et par les fonds d'innovation CHAMO.

neural networks, use kinematic data to identify the specific surgical gesture the trainee attempting to perform [11], [12], [13], [14]. Once the surgical gesture is recognised, these models can evaluate how well the trainee performs the gesture to determine their proficiency level. Common methods of surgical skill assessment include skill level labels (expert, intermediate, novice) [15], safety-based metrics such as the Surgical Apgar Score [16], Objective Structure Assessment of Technical Skills [17], and the Global Rating System (GRS) [8], [17]. In the latter, an expert surgeon examines a video recording of the trainee and provides a rating on a scale from 1 to 5, where a score of 5 indicates proficiency, in each following criteria [8], [17]:

- Respect for tissue: excess force *vs* appropriate handling
- Suture handling: poor tying *vs* accurate suture
- Time and motion: efficient *vs* inefficient motion and time
- Flow of operation: interrupted *vs* continuous motion
- Operation flow/planning: hesitations *vs* forward planning
- Procedure knowledge: level of familiarity with procedure

As with surgical gesture recognition algorithms, GRS scores can also be determined from kinematic data using regression algorithms [18], [19].

In this paper, we use stepwise feature reduction to examine a high-dimensional set of kinematic variables to determine the subset of variables that can predict surgical gestures and GRS scores. In doing so, we assume that these kinematic variables most strongly determine surgical proficiency. In the context of a surgical simulator, reducing the feedback provided to the trainee to this subset of variables can reduce cognitive load and ensure that they receive clear and actionable feedback on those aspects of performance that matter most. The proposed algorithm iteratively evaluates the contribution of each kinematic variable in accurately predicting surgical gesture and GRS score, using respectively a support vector machine classifier and a support vector regression model. At each step, the algorithm retains variables that improve classification and regression scores while discarding those that provide minimal or redundant information. The iterative process continues until removing variables reduces prediction accuracy.

This paper is structured as follows. Section II defines statistical features extracted from a data set of temporal variables, which serve as the input to the GRS regression model and the gesture classification algorithm. Section III describes the iterative feature selection algorithm. The proposed method is validated on the JIGSAW dataset using the metrics described in Section IV-A. The results reported in Section V suggest that GRS score regression and gesture classification accuracy can be largely maintained with a minimal subset of kinematic variables. This indicates that after identifying these key variables a compact subset is sufficient to accurately predict performance, providing a more concise and actionable basis for providing actionable feedback in laparoscopy simulators.

II. Classification and Regression Models

This section provides a formalism to encode time series data into uniform-length feature vectors for subsequent application of the stepwise feature selection algorithm. Given a dataset of n samples, where each entry i in the dataset is a d-dimensional discrete time series with the same length T_i, we denote the data in sample i as

$$\mathbf{X}^{(i)} = [\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \ldots, \mathbf{x}_{T_i}^{(i)}]^T \in \mathbb{R}^{T_i \times d} \tag{1}$$

where $\mathbf{x}_t^{(i)} \in \mathbb{R}^{1\times d}$ represents a measurement of all d variables taken at time t.

To extract feature representations suitable for machine learning algorithms, we compute statistical moments for each kinematic dimension $j \in \{1, 2, \ldots, d\}$ independently [20]. The mean of dimension j is:

$$\mu_j^{(i)} = \frac{1}{T_i}\sum_{t=1}^{T_i} x_{t,j}^{(i)}, \tag{2}$$

while the variance is defined as

$$\sigma_j^{(i)2} = \frac{1}{T_i}\sum_{t=1}^{T_i}\left(x_{t,j}^{(i)} - \mu_j^{(i)}\right)^2, \tag{3}$$

the skewness as

$$\gamma_j^{(i)} = \frac{1}{T_i}\frac{\sum_{t=1}^{T_i}\left(x_{t,j}^{(i)} - \mu_j^{(i)}\right)^3}{\left(\sigma_j^{(i)2}\right)^{3/2}}, \tag{4}$$

and finally the kurtosis is calculated as

$$\kappa_j^{(i)} = \frac{1}{T_i}\frac{\sum_{t=1}^{T_i}\left(x_{t,j}^{(i)} - \mu_j^{(i)}\right)^4}{\left(\sigma_j^{(i)2}\right)^{2}} - 3. \tag{5}$$

The resulting feature vector for sample i is constructed by concatenating these statistics across all d dimensions:

$$\mathbf{f}^{(i)} = \left[\mu_1^{(i)}, \sigma_1^{(i)2}, \gamma_1^{(i)}, \kappa_1^{(i)}, \ldots, \mu_d^{(i)}, \sigma_d^{(i)2}, \gamma_d^{(i)}, \kappa_d^{(i)}\right]^T \in \mathbb{R}^{4d} \tag{6}$$

Applying this encoding to each sample in the dataset transforms a set of time series (of potentially differing lengths) into a set of feature vectors of equal length.

Given a labelled training data $\{(\mathbf{f}^{(i)}, y^{(i)})\}_{i=1}^n$ where each entry has either a surgical gesture label or a GRS score, we employ a Support Vector Machine (SVM) with radial basis function kernel to classify new samples. Let $y^{(i)} \in \{1, 2, \ldots, k\}$ be class labels, the decision function for a new instance $\mathbf{f}$ is:

$$h(\mathbf{f}) = \text{sign}\left(\sum_{i=1}^{n}\alpha_i y^{(i)} K(\mathbf{f}^{(i)}, \mathbf{f}) + b\right) \tag{7}$$

where

$$K(\mathbf{f}^{(i)}, \mathbf{f}) = e^{\left(-\gamma\|\mathbf{f}^{(i)} - \mathbf{f}\|^2\right)} \tag{8}$$

is the radial basis function kernel, α_i are Lagrange multipliers obtained by solving the dual optimization problem, and b is the bias term.

For GRS score regression, we employ Support Vector Regression. The optimization problem seeks to find a function

$f(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b$ that deviates at most by ϵ from the target values $y^{(i)}$ while minimizing:

$$\min_{\mathbf{w},b} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \right\} \tag{9}$$

subjected to

$$\begin{aligned} & y^{(i)} - \mathbf{w}^\top\phi(\mathbf{f}^{(i)}) - b \leq \epsilon + \xi_i \\ & \mathbf{w}^\top\phi(\mathbf{f}^{(i)}) + b - y^{(i)} \leq \epsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0, \quad i = 1, \ldots, n. \end{aligned} \tag{10}$$

where ϕ maps features to a higher-dimensional space via the kernel trick, C controls the trade-off between model complexity and training error, and ξ_i, ξ_i^* are slack variables.

III. Stepwise Feature Selection Algorithm

Backwards stepwise feature selection is used to identify a reduced set of kinematic dimensions that maintains predictive performance. The method operates on the encoded feature matrix $F \in \mathbb{R}^{n\times 4d}$, where each of the d original kinematic dimensions contributes 4 derived features. Let S_0 denote the initial set of available kinematic dimensions. At iteration k, the subset of dimensions under consideration is S_k. The model described in the previous section is trained using only the features associated with the dimensions included in S_k, and its validation error is evaluated on a held-out dataset. This validation error is denoted by $E(S_k)$ and corresponds to classification error for classification tasks or to mean absolute error for regression tasks. At each iteration, a dimension j is temporarily removed from S_k to form the reduced set

$$S_k^{(j)} = S_k \setminus \{j\}. \tag{11}$$

A model is trained using the features associated with $S_k^{(j)}$, and the corresponding validation error $E(S_k^{(j)})$ is computed. The dimension whose removal results in the smallest validation error is identified as

$$j^* = \arg\min_{j\in S_k} E(S_k^{(j)}). \tag{12}$$

The candidate set is then updated according to

$$S_{k+1} = S_k \setminus \{j^*\}. \tag{13}$$

This elimination process is repeated until a single kinematic dimension remains. During the procedure, the validation error at each iteration is recorded. The final selected feature set is defined as the subset encountered during the elimination process that achieves the minimum validation error, namely

$$S^* = \arg\min_k E(S_k). \tag{14}$$

This procedure produces an ordered sequence of progressively lower-dimensional models and enables identification of the smallest kinematic subspace that preserves predictive accuracy.

IV. Dataset and Experimental Validation

The proposed method is evaluated on the JHU-ISI Gesture and Skill Assessment Working Set (JIGSAW) dataset [21].

TABLE I: List of surgical gestures, and gestures in each task in the JIGSAW dataset

Gesture	Action	Samples
G1	Right hand reaches for needle	78
G2	Positioning needle	283
G3	Pushing needle through tissue	275
G4	Left/right hand needle transfer	202
G5	Moving needle to centre	68
G6	Left hand pulls suture	275
G8	Orienting needle	76
G9	Right hand tightens suture	25
G10	Loosening suture	5
G11	Dropping suture/moving to end	100
G12	Left hand reaches for needle	70
G13	Right hand making loop	75
G14	Right hand reaches for suture	98
G15	Pulling suture	73

Task	Gestures in Task	
Suturing (ST)	G1 to G6, and G8 to G11	40
Knot-tying (NT)	G1, and G11 to G15	38
Needle passing (NP)	G1 to G6, and G8, and G11	40

The dataset contains kinematic recordings collected from three robotic surgical training tasks performed on the da Vinci surgical system by surgeons with varying levels of expertise. The dataset comprises 15 primitive different surgical gestures segmented from three surgical tasks, namely suturing (ST), knot-tying (NT), and needle passing (NP). These tasks capture a broad range of dexterous manipulation behaviours commonly used in robotic-assisted surgery. A list of gesture labels and the gestures in each surgical task can be seen in Table I.

The dataset provides 76 time series corresponding to the variables listed in Table II, recorded from the master manipulators and slave robots at a sampling rate of 30 Hz. In addition to kinematic data, the dataset includes manually annotated gesture labels that segment each task trial into predefined surgical gestures, as well as GRS scores assigned to each trial by expert surgeons. The feature encoding strategy defined in Section II transforms these variable-length motion sequences into fixed-dimensional representations while preserving the statistical characteristics of surgical motion.

A. Experimental Design

Gesture classification and GRS score regression accuracy for different combinations of kinematic variables are evaluated using both leave-one-user-out (LOUO) and leave-one-supertrial-out (LOSO) validation. In LOUO, all trials from a single surgeon are held out for testing while the remaining surgeons' data are used for training. This process is repeated for each surgeon in the dataset, and performance metrics are aggregated across all held-out subjects. This method evaluates the model's ability to generalize across different surgical styles and skill levels. Under LOSO, data from one surgeon is

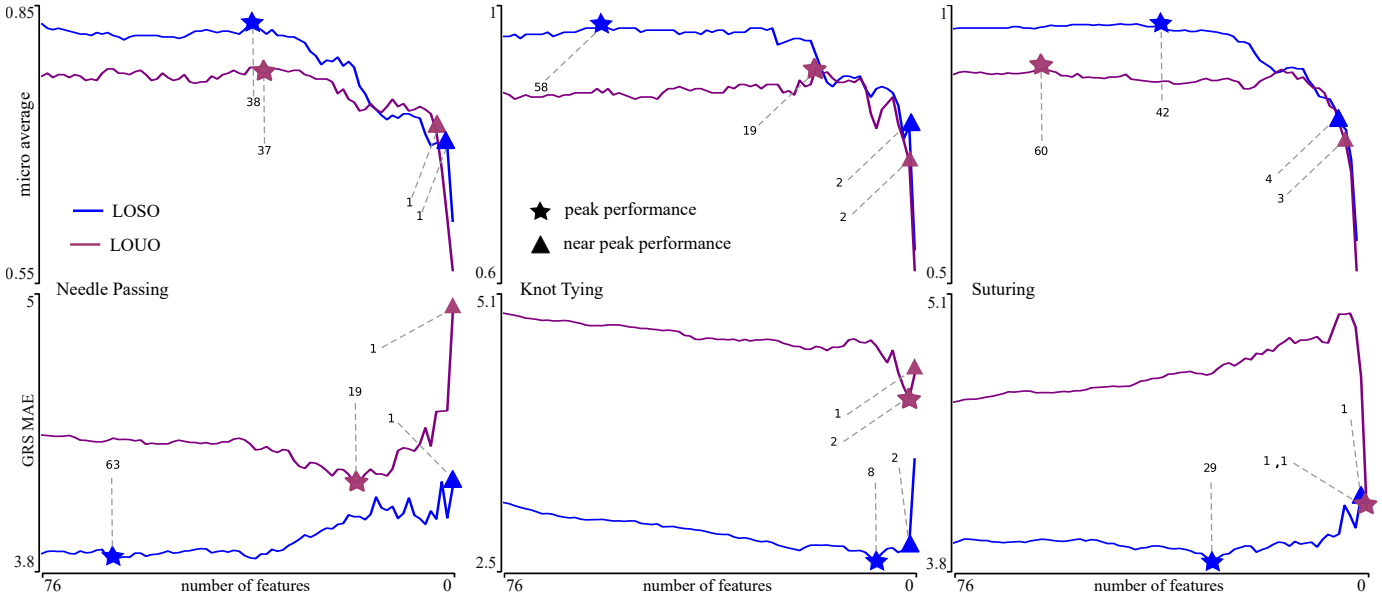

Fig. 1: Gesture classification accuracy for gestures in each surgical task (top row, higher is better), and GRS regression scores (bottom row, lower is better) as a function of the number of features retained step-wise feature reduction model. The star indicates the peak performance point, and the triangle the point with the fewest number of retained variables that maintains an accuracy between 75%-85% of peak accuracy.

TABLE II: List of kinematic features in the dataset and the number of dimensions (Dim).

	ID	Dim.	Description
master	1-3	3	Left tooltip position (x, y, z)
	4-12	9	Left tooltip rotation matrix
	13-15	3	Left tooltip translational velocity
	16-18	3	Left tooltip rotational velocity
	19	1	Left gripper angle
	20-38	19	Same as above but from right hand
slave	39-41	3	Left tooltip position
	42-50	9	Left tooltip rotation matrix
	51-53	3	Left tooltip translational velocity
	54-56	3	Left tooltip rotational velocity
	57	1	Left gripper angle
	58-76	19	Same as above but from right hand

isolated and evaluated independently with all available trials partitioned into cross-validation folds: in each fold, a subset of the surgeon's trials is held out for validation, while the remaining trials from that surgeon are used for training. This process is repeated until every trial from the selected surgeon has served as validation data once, and performance is averaged across folds.

Gesture classification accuracy is evaluated using micro-averaged accuracy, macro-averaged recall, and macro-averaged precision. The distribution of classification accuracy across cross-validation folds is modelled using a β-distribution, from which the mean performance and confidence intervals (CI) are derived. GRS regression is assessed using mean absolute error (MAE), and Spearman rank correlation (SRC) between predicted and ground-truth scores. All results include mean performance values computed across cross-validation folds, along with standard deviations to reflect variability. We compare feature subsets and examine the trade-off between model performance and feature dimensionality.

V. Results and Discussion

The top three panels of Fig. 1 summarize the micro average of gesture classification accuracy (higher is better) for all gestures in each of the three tasks as a function of the number of dimensions retained in the step wise algorithm under both LOUO and LOSO validations. Starting with the 76 kinematic variables, the algorithm selectively reduces the number of variables, re-trains and re-evaluates the models until only 1 feature is left. The classification accuracy remains mostly unchanged across the suturing and knot-tying tasks until around 10 features remain; below this threshold classification accuracy starts to decrease sharply. In needle passing, the model requires around 20 features to maintain peak classification accuracy; when 10 feature remain, the accuracy drops by about 4 percent points in both LOUO and LOSO.

The bottom three panels of Fig. 1 show the GRS MAE regression error (lower is better) for all gestures in each surgical task as a function of the number of kinematic dimensions. Similarly, the regression error does not vary substantially as the number of features decreases, within a margin of error for NP, KT, and ST of 0.24, 0.13, and 0.3 under, and 0.56, 0.24, and 0 under LOSO.

Table III summarizes the gesture classification metrics and the GRS regression error at peak performance (i.e., highest classification micro average and lowest regression MAE), and

TABLE III: Classification metrics at peak micro average (higher is better) and lowest MAE GRS regression error (lower is better) for gestures in needle passing (NP), knot-tying (NT), and suturing (ST) tasks under LOSO and LOUO validation. Metrics are also reported at near peak performance (75%-85% of peak) and the number of retained features.

	Peak performance						Near peak performance					
	LOSO			LOUO			LOSO			LOUO		
Metric	NP	KT	ST	NP	KT	ST	NP	KT	ST	NP	KT	ST
Gesture classification accuracy												
micro avg.	0.861	0.919	0.950	0.805	0.863	0.868	0.705	0.787	0.767	0.675	0.741	0.730
macro (μ)	0.643	0.893	0.827	0.566	0.791	0.681	0.480	0.758	0.611	0.430	0.634	0.530
macro (σ)	0.375	0.080	0.300	0.352	0.235	0.370	0.321	0.114	0.278	0.342	0.308	0.320
precision (μ)	0.677	0.944	0.861	0.663	0.901	0.708	0.509	0.774	0.627	0.490	0.701	0.589
precision (σ)	0.377	0.077	0.305	0.364	0.097	0.381	0.294	0.076	0.258	0.346	0.148	0.237
$\beta(\mu)$	0.864	0.919	0.952	0.818	0.856	0.871	0.708	0.787	0.770	0.696	0.738	0.734
$\beta(\sigma)$	0.040	0.024	0.029	0.110	0.076	0.068	0.055	0.087	0.041	0.102	0.062	0.068
β CI (low)	0.776	0.866	0.882	0.559	0.677	0.711	0.595	0.595	0.685	0.480	0.610	0.590
β CI (high)	0.932	0.960	0.991	0.973	0.969	0.971	0.809	0.928	0.846	0.873	0.849	0.855
# of features	38	58	42	37	19	60	2	2	4	3	2	3
GRS regression error												
MAE	3.87	2.59	3.82	4.13	4.19	4.07	4.11	2.72	4.12	4.69	4.43	4.07
SRC	0.31	0.80	0.53	-0.22	-0.02	0.60	0.06	0.75	0.28	-0.22	-0.18	0.60
# of features	63	8	29	19	2	1	1	2	1	1	1	1

the number of features retained by the step wise algorithm for each surgical task. These results are in line with state-of-the-art classification algorithms operating on the same dataset, which report micro averages in the range of 70-85 % [13], [22]. To further illustrate the trade-off between predictive performance and model dimensionality, the table also reports classification and GRS regression metrics for the lowest number of features that retain 75%-85% of peak performance (defined as 0.75-0.85 $\times$ peak micro-average for classification and 1.25-1.33 $\times$ minimum MAE for regression), together with the corresponding number of dimensions. As it can be seen, this operating point substantially reduces dimensionality to 1-4 features, while incurring only a modest performance loss but within ranges reported in the literature. The peak-performance and reduced-dimension operating points could be used in tandem; the former maximizes assessment accuracy, while the latter restricts the kinematic dimensions to a compact subset suitable for formative or summative feedback during simulator training.

Table IV and Table V list the most commonly retained features unique to each task at peak performance, and features common across all tasks. While the number and identity of retained variables at peak performance vary; however, several variables are common across all tasks. For gesture classification, the algorithm consistently retained 85%, 90%, and 91% micro accuracy for NP, KT, and ST under LOSO. The variables common to 80% of classification of the three tasks at peak performance seen in Table IV correspond to the left and right tool rotation and rotational velocity, and the left gripper angle. In more than 50% of GRS experiments, the prominent features seen in Table V are the master left and right positions as well as master and slave rotations and rotational velocities.

TABLE IV: Most common features retained in 100% or at least 80% of 8 folds of needle passing (NP), 7 folds of knot-tying (NT), and 8 folds of suturing (ST) tasks, along with features common (CM) across all tasks.

		KT	NP	ST	CM
100%	Feat. ID	59 68 71 72 74 75	40 48-50 52 58-61 66-69 71-74	69 75	56
	# feat.	7	18	3	1
$\geq$80%	Feat. ID	39 68-75 59 61 66	36-45 47-62 18 64-75	51 52	49 56 57 69 71 72 75
	# feat.	15	38	9	7
	Acc.	0.881	0.845	0.900	-

VI. Conclusion

The proposed stepwise feature reduction algorithm demonstrates that accurate GRS regression and gesture classification do not require the full set of kinematic variables in the JIGSAW dataset, indicating that many variables are redundant once task-relevant motion patterns are identified. The overlap of selected kinematic variables across different tasks is an indicator that a trainee's ability to reproduce motion patterns along these specific dimensions, relative to an expert surgeon, strongly influences their performance evaluation.

The results of gesture recognition and GRS score regression show that on average 42 kinematic variables yield peak classification performance, and 20 are required for the lowest

TABLE V: Common features (feat.) retained in 75%, 50%, and 25% of GRS regression in 8 folds of needle passing (NP), 7 folds of knot-tying (NT), and 8 folds of suturing (ST) tasks, along with features common (CM) across all tasks.

		KT	NP	ST	CM
$\geq$75%	Feat. ID	-	61	74	-
	# feat.	0	1	1	0
$\geq$50%	Feat. ID	4 6 20 45 63, 64	1 10 25 29 45 47 56 61 63 67	74	-
	# feat.	6	10	1	0
$\geq$25%	Feat. ID	4 6 64 20 7 1 3 8 45 63	1 10 29 61 67 25 45 56 47 63	74	57
	# feat.	10	10	10	1
	MAE	3.813	3.992	4.013	-
	SRC	0.347	−0.212	0.554	-

GRS regression error. However, when operating at 75%-85% of peak performance, these numbers are lowered respectively to 3 and 1 kinematic features. For gesture recognition the retained variables are the left slave rotational velocity about the x-axis, gripper angle, right slave velocity along y and z, and rotational velocity about the y-axis. For GRS regression, results suggest that the right slave rotational velocity about the x is a key performance indicator. The near-to-peak-performance subset offers comparable performance metrics to JIGSAW's classification benchmarks reported in [13].

This study was conducted using data from a robotic platform, but the proposed framework is transferable to conventional laparoscopy box trainers and other instrumented surgical platforms. By identifying the minimal set of kinematic variables required for performance assessment and motion recognition, this approach can guide the design of instrumented laparoscopic trainers, including the strategic selection and placement of sensors.

Providing feedback to trainees during surgical training over a large number of performance metrics of kinematic variables can become confusing and overwhelming, resulting in cognitive overload, and a negative impact on training performance. Reducing the dimensionality of kinematic data to only the most important predictors of surgical proficiency can allow instrumented surgical simulators powered by pattern recognition algorithm to provide more precise and actionable feedback to trainees. The results of this paper will guide our future work on laparoscopic training simulators that do not rely on expert supervision, offering trainees unlimited training opportunities with formative and summative, focused performance feedback.

REFERENCES

[1] A. Johnson, "Laparoscopic surgery," *The Lancet*, vol. 349, no. 9052, pp. 631–635, 1997.

[2] B. Jönsson and N. Zethraeus, "Costs and benefits of laparoscopic surgery—a review of the literature," *European Journal of Surgery*, vol. 166, no. S12, pp. 48–56, 2000.

[3] X. Li, J. Zhang, L. Sang, W. Zhang, Z. Chu, X. Li, and Y. Liu, "Laparoscopic versus conventional appendectomy – a meta-analysis of randomized controlled trials," *BMC Gastroenterology*, vol. 10, p. 129, 2010.

[4] S. Dreier, S. Mona, L. Konge, and F. Bjerrum, "Three-dimensional versus two-dimensional vision in laparoscopy: a systematic review," *Surgical Endoscopy*, vol. 30, no. 1, pp. 11–23, 2016.

[5] Y. Tanagho *et al.*, "2D versus 3D visualization: Impact on laparoscopic proficiency using the fundamentals of laparoscopic surgery skill set," *Journal of Laparoendoscopic & Advanced Surgical Techniques*, vol. 22, no. 9, pp. 865–870, 2012.

[6] J.-U. Stolzenburg, M. C. Truss, R. Rabenalt, M. Do, T. Schwalenberg, P. F. Katsakiori, A. McNeill, and E. Liatsikos, "Training in laparoscopy," *EAU-EBU Update Series*, vol. 5, no. 2, pp. 53–62, 2007.

[7] D. J. Scott, P. C. Bergen, R. V. Rege, R. Laycock, S. T. Tesfay, R. J. Valentine, D. M. Euhus, D. R. Jeyarajah, W. M. Thompson, and D. B. Jones, "Laparoscopic training on bench models: better and more cost effective than operating room experience?" *Journal of the American College of Surgeons*, vol. 191, no. 3, pp. 272–283, 2000.

[8] Y. Gao *et al.*, "JHU-ISI gesture and skill assessment working set (JIGSAWS): A surgical activity dataset for human motion modeling," *Modeling and Monitoring of Comp. Assisted Interventions*, 2014.

[9] K. Narazaki, D. Oleynikov, and N. Stergiou, "Robotic surgery training and performance: identifying objective variables for quantifying the extent of proficiency," *Surgical Endoscopy And Other Interventional Techniques*, vol. 20, no. 1, pp. 96–103, 2006.

[10] F. Song, Z. Guo, and D. Mei, "Feature selection using principal component analysis," in *2010 international conference on system science, engineering design and manufacturing informatization*, vol. 1. IEEE, 2010, pp. 27–30.

[11] A. Cotton, B. Sainsbury, and C. Rossa, "Sliding adaptive dynamic time warping for segment matching in surgical simulation data," in *Inter.Conf. on Smart Multimedia*, 03 2024.

[12] B. van Amsterdam *et al.*, "Gesture recognition in robotic surgery: A review," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 6, pp. 2021–2035, 2021.

[13] N. Ahmidi *et al.*, "A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 9, pp. 2025–2041, 2017.

[14] K. MacWilliams, J. Green, A. Nasr, G. Azzie, and C. Rossa, "Optimizing sensor selection in laparoscopic simulators: Lessons learned in a robotic platform," in *2025 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*. IEEE, 2025, pp. 1–6.

[15] A. Nasr, B. Carrillo, J. T. Gerstle, and G. Azzie, "Motion analysis in the pediatric laparoscopic surgery (pls) simulator: validation and potential use in teaching and assessing surgical skills," *Journal of Pediatric Surgery*, vol. 49, no. 5, pp. 791–794, 2014.

[16] S. E. Regenbogen, R. T. Lancaster, S. R. Lipsitz, C. C. Greenberg, M. M. Hutter, and A. A. Gawande, "Does the surgical apgar score measure intraoperative performance?" *Annals of Surgery*, vol. 248, no. 2, pp. 320–328, August 2008.

[17] J. A. Martin, G. Regehr, R. Reznick, H. MacRae, J. Murnaghan, C. Hutchison, and M. Brown, "Objective structured assessment of technical skill (OSATS) for surgical residents," *British Journal of Surgery*, vol. 84, no. 2, pp. 273–278, Feb. 1997.

[18] J. Quarez, M. Modat, S. Ourselin, J. Shapey, and A. Granados, "ReCAP: Recursive Cross Attention Network for Pseudo-Label Generation in Robotic Surgical Skill Assessment," *arXiv preprint arXiv:2407.05180*, 2024, last revised 7 Jul 2025. [Online]. Available: https://arxiv.org/abs/2407.05180

[19] D. Anastasiou *et al.*, "Keep your eye on the best: Contrastive regression transformer for skill assessment in robotic surgery," *IEEE Robotics and Automation Letters*, vol. 8, no. 3, pp. 1755–1762, 2023.

[20] A. Nanopoulos *et al.*, "Feature-based classification of time-series data," *International Journal of Computer Research*, vol. 10, pp. 49–61, 01 2001.

[21] Y. Gao *et al.*, "Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling," in *MICCAI workshop: M2cai*, vol. 3, no. 2014, 2014, p. 3.

[22] D. Liu and T. Jiang, "Deep reinforcement learning for surgical gesture segmentation and classification," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2018, pp. 247–255.