2026 IEEE/SICE International Symposium on System Integration (SII)
January 11-14, 2026. Cancún, México

Motion Capture and Machine Learning-Based Evaluation of Surgical Skills in Laparoscopic Cadaver Training*

Lingbo Yan, Takashige Abe, Koki Ebina, Masafumi Kon, Madoka Higuchi, Kiyohiko Hotta, Jun Furumido, Naoya Iwahara, Shunsuke Komizunai, Teppei Tsujita, Kazuya Sase, Xiaoshuai Chen, Hiroshi Kikuchi, Haruka Miyata, Ryuji Matsumoto, Takahiro Osawa, Sachiyo Murai, Toshiaki Shichinohe, Soichi Murakami, Taku Senoo, Masahiko Watanabe, Atsushi Konno, *Member, IEEE*

***Abstract*— To promote efficient laparoscopic surgery education, a system utilizing machine learning has been developed to quantify surgical skill levels based on the movement of surgical instruments. In this system, the movements of surgical instruments operated by surgeons during laparoscopic cadaver surgery training are recorded using an optical motion capture system, and kinematic features are extracted from the recorded data. These extracted features are then used as input for machine learning models, with expert-evaluated scores—based on the Global Operative Assessment of Laparoscopic Skills (GOALS)—serving as the training data. The entire nephrectomy procedure was divided into three distinct processes: colon mobilization (Process 1), renal vascular dissection (Process 2), and incision and removal of the remaining tissues (Process 3). In this study, interpretable kinematic features were extracted from instrument movements during the colon mobilization phase (Phase 1). These features were used to train three regression models: Principal Component Analysis followed by Support Vector Regression (PCA-SVR), Partial Least Squares regression (PLS), and Ridge Regression. The models aimed to predict GOALS scores across five key domains: depth perception, bimanual dexterity, efficiency, tissue handling, and autonomy. Model performance was evaluated using 5-fold nested cross-validation repeated 100 times. Among the models, Ridge Regression consistently demonstrated high accuracy, with median mean absolute errors (MAEs) below 0.82 in most domains. This system is expected to contribute to more effective surgical education by providing multidimensional, objective feedback on surgical performance.**

I. Introduction

Laparoscopic surgery offers significant advantages over open procedures, including smaller incisions, reduced postoperative pain, lower risk of infection, and faster patient recovery [1]. However, mastering laparoscopic techniques remains challenging due to limited instrument mobility, two-dimensional visualization, and the absence of tactile feedback [2].

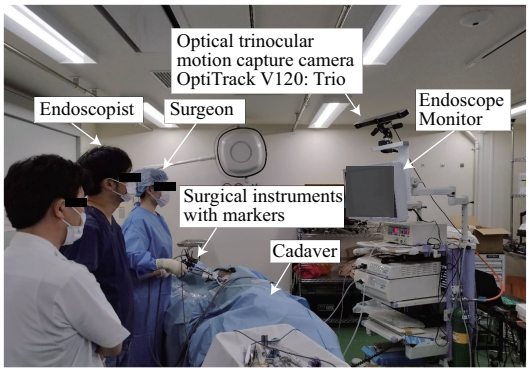


Fig. 1: Laparoscopic cadaver surgical training (CST).

To address these challenges, objective assessment tools have been developed based on motion analysis of surgical instruments. Early studies employed animal organs or physical models to train and evaluate psychomotor skills [3], [4], and validated frameworks have demonstrated that temporal and spatial motion features correlate with surgical experience [5], [6]. Several research groups have implemented machine learning algorithms to automatically classify skill levels [7], [8], with support vector machines (SVMs) being widely adopted due to their strong theoretical foundation [6], [7]. Our previous work further introduced the use of Global Operative Assessment of Laparoscopic Skills (GOALS) in wet lab training to provide real-time, onsite feedback [9].

Building on previous studies, this research proposes an objective feedback system based on motion analysis and regression modeling. Detailed motion features were analyzed during laparoscopic nephrectomy performed in Thiel-

Lingbo Yan, Koki Ebina, Taku Senoo, Atsushi Konno are with the Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan. llingboyan@gmail.com, konno@ssi.ist.hokudai.ac.jp

Takashige Abe, Masafumi Kon, Madoka Higuchi, Kiyohiko Hotta, Naoya Iwahara, Hiroshi Kikuchi, Haruka Miyata, Ryuji Matsumoto, Takahiro Osawa, Sachiyo Murai, Toshiaki Shichinohe, Soichi Murakami, Masahiko Watanabe are with the Graduate School of Medicine, Hokkaido University, Sapporo, Japan

Jun Furumido is with Department of Urology, Asahikawa Kousei Hospital, Asahikawa, Japan.

Shunsuke Komizunai is with Faculty of Engineering and Design, Kagawa University, Takamatsu, Japan.

Teppei Tsujita is with Department of Mechanical Engineering, National Defense Academy of Japan, Yokosuka, Japan.

Kazuya Sase is with Faculty of Engineering, Tohoku Gakuin University, Sendai, Japan.

Xiaoshuai Chen is with the Graduate School of Science and Technology, Hirosaki University, Hirosaki, Japan.

∗This study was approved by the Medical Ethics Committee of the Faculty of Medicine, Hokkaido University (IRB No. 20-027), for motion capture-based analysis of cadaveric laparoscopic training using the Thiel embalming technique. The cadaver surgical trainings (CSTs) were conducted by medical doctors from the Department of Urology, Graduate School of Medicine, Hokkaido University. Instrument motion trackings during the CSTs were performed in accordance with the *Guidelines for Cadaveric Autopsies in Clinical Medical Education and Research*. All procedures were carried out in the clinical autopsy facility, where proper respect and dignity were consistently maintained for the cadavers throughout the training.

978-1-6654-5784-2/26/$31.00 ©2026 IEEE

TABLE I: Summary of Surgeons' Backgrounds

Category	Data
Total Surgeons	36
Age (Median, range)	32 (27-52)
Sex (Male/Female)	30/6
Experience of Laparoscopic Surgery	0-9 (novices, n=9) 10-49 (intermediartes, n=19) 50-499 (experts, n=8)
Dominant Hand (Right/Left)	35/1

embalmed cadaver training, and three regression models were investigated—Principal Component Analysis followed by Support Vector Regression (PCA-SVR), Partial Least Squares regression (PLS), and Ridge Regression—for predicting surgical skill competency.

The aim is to provide interpretable and personalized feedback to trainees, enhancing their understanding of performance across dimensions such as autonomy, depth perception, and tissue handling.

Figure 1 presents a representative scene from a CST session, in which surgeons performed laparoscopic radical nephrectomy via a transperitoneal approach on Thiel-embalmed cadavers.

Utilizing motion metrics and machine learning, this study focuses on predicting scores from the GOALS, a validated tool for evaluating laparoscopic surgical performance across five domains: depth perception, bimanual dexterity, efficiency, tissue handling, and autonomy [10]. Each domain is rated on a 5-point Likert scale, resulting in a total score ranging from 5 to 25.

II. Methods

A. Dataset and Experimental Setup

To enable structured analysis of laparoscopic radical procedures, each operation was segmented into three major Processes based on anatomical and procedural landmarks:

- **Process 1: Colon mobilization**
  In left nephrectomy, this includes mobilization of the descending colon, pancreas, and spleen.
  In right nephrectomy, it includes mobilization of the ascending colon and duodenum.
- **Process 2: Renal vascular dissection**
  Identification, clipping, and transection of renal arteries and veins.
- **Process 3: Incision and removal of remaining tissues**
  Incision and removal of the remaining perirenal tissues as the final process.

For each surgical process, GOALS scores were assessed by an expert surgeon in a blinded manner. In this study, we focused on Process 1 (colon mobilization), as it involves extensive tissue dissection and is considered a representative task for evaluating laparoscopic surgical skill competency. Accordingly, we began by developing and evaluating machine learning regression models to predict GOALS scores specifically for Process 1.

A total of 39 surgeons completed 51 training sessions. Among them, four urologists with caseloads of 0–9 procedures (novices), two with 10–49 procedures (intermediates), and two with more than 49 procedures (experts) participated twice. Additionally, two novice urologists (0–9 procedures) participated three times. After excluding five cases due to video recording failure (n = 3), inadequate embalming conditions (n = 1), and cases in which the mentor performed the majority of the procedure (n = 1), 46 motion capture datasets from 36 surgeons were included in the final analysis.

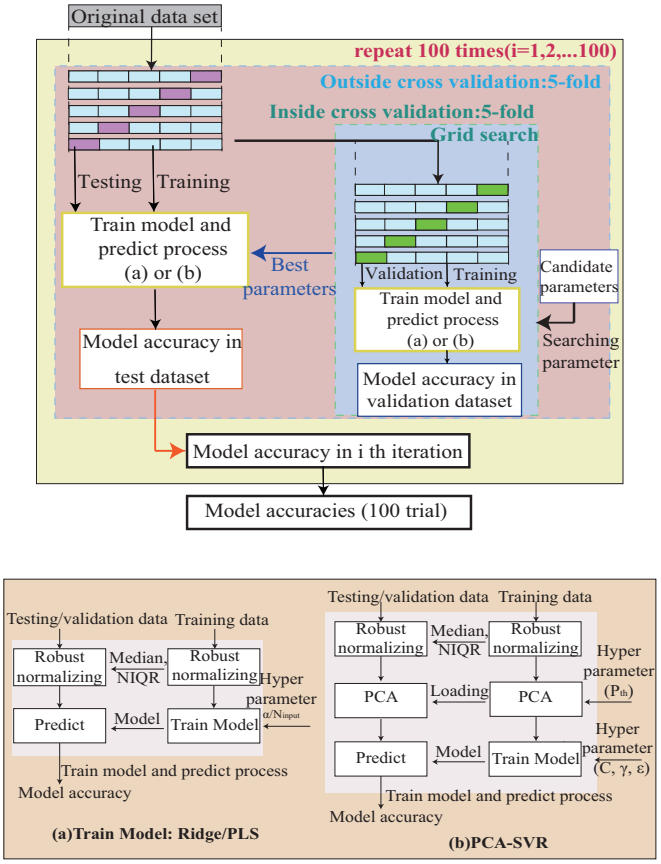


Fig. 2: The nested cross-validation workflow for regression model training and evaluation.

The background and surgical experience of the 36 surgeons are summarized in Table I.

To track the motion of surgical instruments, retroreflective markers were attached near the base of the forceps. The 3D position and orientation of each instrument were captured using an optical motion capture system (OptiTrack V120:Trio) at 120 Hz. Two types of forceps were analyzed in Process 1: grasping forceps and scissors forceps.

The expert-evaluated GOALS scores ranged from 5 to 25 (median 16), approximately following a normal distribution.

B. Feature Selection

To ensure the biological validity of the input variables, a feature selection process was conducted based on expert consensus and statistical correlation. Specifically, Spearman's rank correlation coefficients were calculated between each motion feature and the GOALS score evaluated by the expert laparoscopic surgeon (TA) as mentioned above.

TABLE II: Selected features associated with each GOALS domain after correlation analysis and expert screening

GOALS Domain	Selected Features
Autonomy	• **Overall task**: Operating time • **Scissors forceps**: Average velocity (v), acceleration (a), jerk (j), percentage of low velocity (DVL) ($0.5 \leq v < 2.0$ cm/s), high velocity (DVH) ($5.0 \leq v < 12.0$ cm/s)
Depth Perception	• **Grasping forceps**: Total path length in depth direction • **Scissors forceps**: Average velocity in depth direction
Bimanual Dexterity	• **Overall task**: Operating time • **Grasping forceps**: Using time, path length, sum of roll axis attitude changes, sum of pitch-yaw attitude changes • **Scissors forceps**: Using time
Efficiency	• **Overall task**: Operating time • **Grasping forceps**: Using time, path length, total path length in depth direction, sum of roll axis attitude changes, sum of pitch-yaw attitude changes • **Scissors forceps**: Average velocity (v), acceleration (a), jerk (j), using time, sum of roll axis attitude changes, percentage of low velocity (DVL) ($0.5 \leq v < 2.0$ cm/s), high velocity (DVH) ($5.0 \leq v < 12.0$ cm/s)
Tissue Handling	• **Grasping forceps**: Path length, sum of roll axis attitude changes, sum of pitch-yaw attitude changes

Only features showing a meaningful monotonic relationship—defined as having an absolute correlation coefficient of 0.4 or greater—were considered candidates for regression. From this subset, clinically interpretable features relevant to each individual GOALS domain (e.g., autonomy, depth perception, efficiency) were manually selected based on the discussion between TA and KA (TA: expert surgeon; KA: motion analysis engineer). The following motion metrics (Table II) exhibited a Spearman correlation coefficient $|r| > 0.4$ were selected for model training.

A correlation threshold of $|r| > 0.4$ was chosen to ensure sufficient feature coverage across GOALS domains while maintaining statistical relevance. Preliminary tests with $|r| > 0.5$ reduced the number of features and slightly degraded performance.

C. Feature Extraction and Preprocessing

A variety of kinematic parameters were derived from the tracked instrument trajectories, such as path length, speed profile, acceleration, and the sum of angular changes. The instantaneous velocity $v(t)$ of each marker was calculated as the first derivative of its position vector $p(t)$, and its magnitude $\|v(t)\|$ was used to compute average speed, acceleration, and jerk:

$$v(t) = \frac{dp(t)}{dt}, \qquad a(t) = \frac{dv(t)}{dt}, \qquad j(t) = \frac{da(t)}{dt}. \quad (1)$$

The total path length (PL) of each forceps trajectory was calculated as:

$$\mathrm{PL} = \sum_{i=2}^{N} \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2 + (z_i - z_{i-1})^2}, \quad (2)$$

where x_i, y_i, and z_i denote the tip positions of the instrument tip in frame i. This metric represents the total travel distance of the surgical instrument during each task.

The *average speed* was defined as the time-averaged magnitude of velocity, while acceleration and jerk quantified motion smoothness. Path length was obtained by integrating $\|v(t)\|$ over time, and sum of the angular motion (the sum of roll, the sum of pitch and yaw) was computed from quaternion-derived instrument orientations. Idle intervals were identified when $\|v(t)\| < 0.5\,\mathrm{cm/s}$.

All features were normalized using the robust Z-score:

$$z_i = \frac{x_i - \mathrm{median}(x)}{\mathrm{NIQR}(x)}, \qquad \text{where } \mathrm{NIQR}(x) = \frac{\mathrm{IQR}(x)}{1.3490}. \quad (3)$$

This corrected definition ensures statistical equivalence with the standard deviation of a normally distributed variable.

D. Regression Modeling

As illustrated in Fig. 2, the core regression pipeline employed in this study integrates robust normalization, dimensionality reduction, and nonlinear regression to predict continuous skill scores based on instrument motion features.

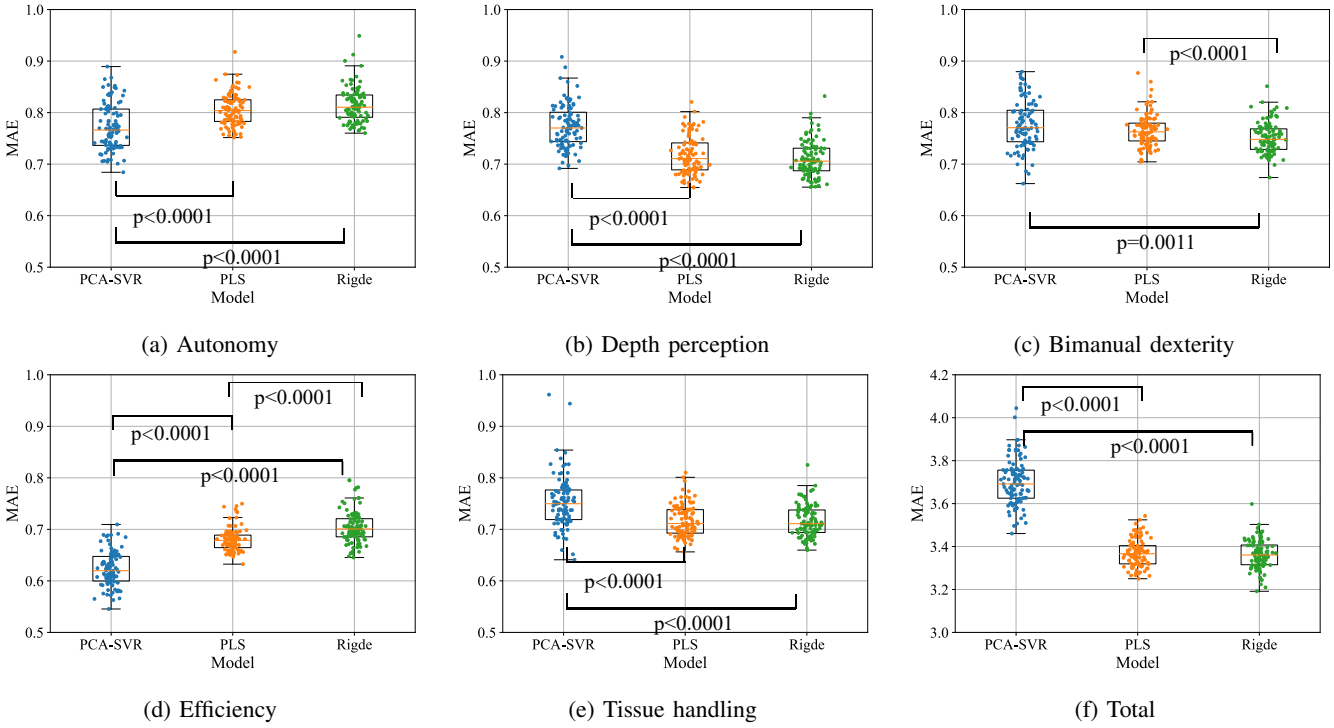
The primary model—PCA-SVR—consists of three steps: first, motion features are normalized using a robust Z-score method that scales data based on median and interquartile range, reducing the influence of outliers. Second, Principal Component Analysis (PCA) is applied for dimensionality reduction, retaining components that explain between 80% and 95% of the total variance. Third, a Support Vector Regression (SVR) model with a radial basis function (RBF) kernel is trained on the transformed data. Hyperparameters, including the regularization parameter C, kernel coefficient γ, and ϵ-insensitive margin width, are optimized using a grid search on exponential scales.

To benchmark this approach, two additional models were evaluated. The first is Partial Least Squares (PLS) regression, which simultaneously reduces dimensionality and performs regression by maximizing the covariance between input features and target scores. The second is Ridge regression, a linear model that applies L2 regularization to mitigate overfitting. Both PLS and Ridge were trained on the same robustly normalized input features, without PCA preprocessing.

All models were evaluated using nested 5-fold cross-validation repeated 100 times. In each iteration, the outer loop partitioned the dataset for performance evaluation, while the inner loop conducted hyperparameter tuning using grid search. This process ensured that model selection and error estimation remained unbiased. The primary evaluation metric

TABLE III: Group and Pairwise Comparison of MAE for Each GOALS Metric (Median MAE in Parentheses)

Metric	Median MAE (PCA-SVR / PLS / Ridge)	Friedman	PCA-SVR vs Ridge	PCA-SVR vs PLS	Ridge vs PLS
Autonomy (a)	(0.7662 / 0.8039 / 0.8105)	$<$ **0.0001**	$<$ **0.0001**	$<$ **0.0001**	0.0761
Depth perception (b)	(0.7701 / 0.7107 / 0.7058)	$<$ **0.0001**	$<$ **0.0001**	$<$ **0.0001**	0.1705
Bimanual dexterity (c)	(0.7706 / 0.7627 / 0.7481)	$<$ **0.0001**	$<$ **0.0001**	0.0536	**0.0011**
Efficiency (d)	(0.6198 / 0.6787 / 0.7007)	$<$ **0.0001**	$<$ **0.0001**	$<$ **0.0001**	$<$ **0.0001**
Tissue handling (e)	(0.7498 / 0.7112 / 0.7110)	$<$ **0.0001**	$<$ **0.0001**	$<$ **0.0001**	0.9474
Total (f)	(3.6915 / 3.3665 / 3.3604)	$<$ **0.0001**	$<$ **0.0001**	$<$ **0.0001**	0.6976



(a) Autonomy (b) Depth perception (c) Bimanual dexterity

(d) Efficiency (e) Tissue handling (f) Total

Fig. 3: Boxplot comparison of prediction accuracy across three regression models (PCA-SVR, PLS, Ridge) for six surgical skill evaluation metrics.

was Mean Absolute Error (MAE) between predicted scores and expert-assigned GOALS ratings.

The optimal hyperparameters ($C = 10^3$, $\gamma = 0.1$, and $\varepsilon = 0.01$) were determined by grid search within the inner cross-validation loop.

E. Evaluation Metrics

The prediction accuracy of each regression model was evaluated using the Mean Absolute Error (MAE) between predicted scores and expert ratings. In addition to numerical error metrics, visual comparison was performed using boxplots across the three experience groups. Statistical significance of prediction differences was tested using non-parametric tests, and p-values were reported to indicate group-level variation in skill estimation.

III. Results

The predictive accuracy of the three regression models—PCA-SVR, PLS, and Ridge regression—was compared using the mean absolute error (MAE) over 100 repetitions of nested 5-fold cross-validation. The results are shown in Table III. Ridge regression exhibited the best overall performance for most GOALS domains, while PCA-SVR achieved the lowest error for Autonomy and Efficiency.

Among the five GOALS domains, Autonomy (Fig. 3 (a)) and Efficiency (Fig. 3 (d)), the PCA-SVR performs better. Ridge regression showed the lowest median MAE for Bimanual Dexterity (Fig. 3 (c)). And in Depth Perception (Fig. 3 (b)) and Tissue Handling (Fig. 3 (e)), the differences across the ridge model and PLS model were minimal, with the ridge model performing slightly better.

For overall performance (Fig. 3 (f)), Ridge and PLS showed comparable results, with Ridge having slightly higher accuracy. These findings indicate that while Ridge regression offers stable and high accuracy across most domains, PCA-SVR may still be advantageous in certain skill dimensions that involve more nuanced instrument control patterns.

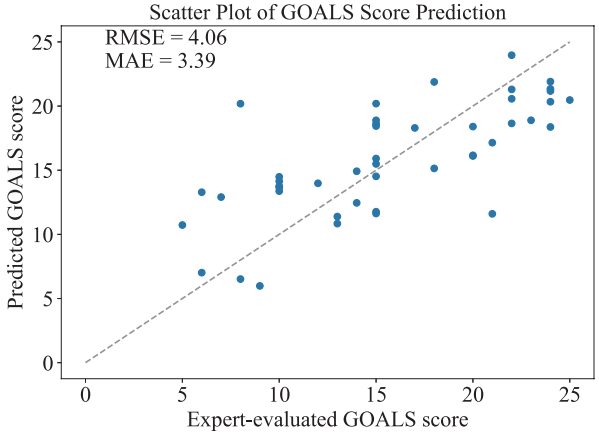


Fig. 4: Scatter between expert-rated and predicted total GOALS (Process 1). Dashed line shows $y = x$. Ridge regression: RMSE $= 4.06$, MAE $= 3.39$.

In addition to the domain-wise analysis, the regression model was applied to predict the total GOALS score using motion features extracted from Process 1. The resulting scatter plot in Fig. 4 illustrates the relationship between the predicted and expert-evaluated scores in Ridge regression with the best performance in prior evaluation. The prediction achieved a root mean squared error (RMSE) of 4.06 and a mean absolute error (MAE) of 3.39. Most points were distributed near the diagonal line ($y = x$), indicating good agreement and demonstrating the potential of motion-based modeling for comprehensive surgical skill assessment in cadaver-based laparoscopic training.

IV. Discussion

This study demonstrated that objective feedback systems using motion-derived features can provide reliable prediction of surgical skill scores based on the GOALS framework. The best-performing model (Ridge regression) achieved a median MAE of less than 0.82 for multiple domains, indicating high agreement with human expert ratings. Compared to binary classification (expert vs novice) in previous studies [6], [7], our regression-based approach allows more granular and personalized feedback.

The analysis also revealed that certain domains such as Efficiency and Depth Perception were more predictable, possibly due to their higher correlation with time-based and kinematic features. On the other hand, Autonomy had weaker associations, likely due to its subjective nature and dependence on holistic surgical judgment.

This work extends the application of Thiel cadaver-based training toward automatic feedback generation. Future work may include longitudinal tracking of trainee progress and incorporation of video-based features to improve prediction. We are now developing similar machine-learning models of Processes 2 and 3.

V. Conclusion

This study proposed an objective feedback system for laparoscopic surgical training using motion analysis and regression modeling. By extracting interpretable kinematic features from the colon mobilization process (Process 1) of Thiel-embalmed cadaver training, and predicting GOALS scores with regression models, the system enables continuous and personalized assessment of surgical performance.Among the evaluated models, Ridge regression achieved consistently high prediction accuracy, while PCA-SVR showed competitive performance in specific skill domains. Features such as path length, velocity, and forceps usage time were closely associated with GOALS dimensions, validating the utility of motion-derived metrics.This work extends our previous classification-based evaluation framework to a continuous regression-based system. Future developments may include integration with visual data, real-time feedback modules, and larger-scale training datasets to further improve feedback precision and educational value.

Acknowledgment

This work was supported by JSPS Grants-in-Aid for Scientific Research (C)(JP17K08897), (A)(JP18H04102), (B)(JP21H00893), (A)(23H00480), Grant-in-Aid for Challenging Research (Exploratory)(23K18486), JST SPRING under Grant Number JPMJSP2119 and AMED under Grant Number JP22vk0124006.

References

[1] J. W. Milsom, B. Bohm, K. A. Hammerhofer, M. Fazio, P. Steiger, and J. Elson, "Laparoscopic sigmoid colectomy: a prospective trial comparing open and laparoscopic techniques," *Surgery*, vol. 124, no. 4, pp. 565–572, 1998.

[2] R. M. Satava, "Virtual reality surgical simulator: the first steps," *Surgical Endoscopy*, vol. 7, no. 3, pp. 203–205, 1993.

[3] M. Higuchi, T. Abe, K. Hotta, et al., "Development and validation of a porcine organ model for training in essential laparoscopic surgical skills," *International Journal of Urology*, vol. 27, pp. 929–938, 2020.

[4] E. F. Hofstad, C. Vapenstad, L. E. Bo, T. Lango, E. Kuhry, and R. Marvik, "Psychomotor skills assessment by motion analysis in minimally invasive surgery on an animal organ," *Minimally Invasive Therapy & Allied Technologies*, vol. 26, pp. 240–248, 2017.

[5] M. K. Chmarra, S. Klein, J. C. de Winter, F. W. Jansen, and J. Dankelman, "Objective classification of residents based on their psychomotor laparoscopic skills," *Surgical Endoscopy*, vol. 24, pp. 1031–1039, 2010.

[6] I. Oropesa, P. Sanchez-Gonzalez, M. K. Chmarra, et al., "Supervised classification of psychomotor competence in minimally invasive surgery based on instruments motion analysis," *Surgical Endoscopy*, vol. 28, pp. 657–670, 2014.

[7] B. Allen, V. Nistor, E. Dutson, G. Carman, C. Lewis, and P. Faloutsos, "Support vector machines improve the accuracy of evaluation for the performance of laparoscopic training tasks," *Surgical Endoscopy*, vol. 24, pp. 170–178, 2010.

[8] F. Perez-Escamirosa, A. Alarcon-Paredes, G. A. Alonso-Silverio, et al., "Objective classification of psychomotor laparoscopic skills of surgeons based on three different approaches," *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, pp. 27–40, 2020.

[9] K. Ebina, T. Abe, K. Hotta, et al., "Objective evaluation of laparoscopic surgical skills in wet lab training based on motion analysis and machine learning," *Langenbeck's Archives of Surgery*, vol. 407, no. 5, pp. 2123–2132, 2022.

[10] M. C. Vassiliou, L. S. Feldman, C. G. Andrew, S. Bergman, K. Leffondre, D. Stanbridge, and G. M. Fried, "A global assessment tool for evaluation of intraoperative laparoscopic skills," *American Journal of Surgery*, vol. 190, no. 1, pp. 107–113, 2005.

[11] L. Yan, et al., "Validation and motion analyses of laparoscopic radical nephrectomy with Thiel-embalmed cadavers," *Current Problems in Surgery*, vol. 61, no. 10, 2024. DOI: 10.1016/j.cpsurg.2024.101559.