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

Evaluation of Surgical Skills Using Machine Learning and Interpretation of Results with Explainable AI in Practical Laparoscopic Surgery 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 facilitate efficient laparoscopic surgical education, a system was developed that utilizes machine learning to classify surgical skill levels—novice, intermediate, and expert—based on the motion dynamics of surgical instruments. This system not only categorizes surgical proficiency but also incorporates SHAP, an Explainable AI technique, to provide insights into the rationale behind each classification result. For the machine learning dataset, the movements of four surgical instruments were recorded using a motion capture (mocap) system during total nephrectomy training sessions conducted on 46 cadaveric specimens prepared for laparoscopic surgery. 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). Surgical skill analysis was performed separately for each phase. Surgeons were categorized into three groups based on their prior experience with laparoscopic procedures: novices (0–9 cases), intermediates (10–49 cases), and experts (50 or more cases). A total of 111 features were extracted from the instrument motion data for each phase, and comparative analyses were conducted across the three groups. Multiple machine learning approaches—including Support Vector Machine (SVM), Principal Component Analysis followed by SVM (PCA-SVM), and Random Forest—were employed to develop models for classifying surgeons into three distinct skill levels. The classification performance of these models was subsequently validated. The results revealed that features related to efficiency and speed significantly contributed to differences in surgical skill levels. The developed system enables quantitative comparison and visualization of specific instrument characteristics. This system contributes to intelligent integration of surgical education and Explainable AI, providing actionable feedback for skill improvement.**

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.

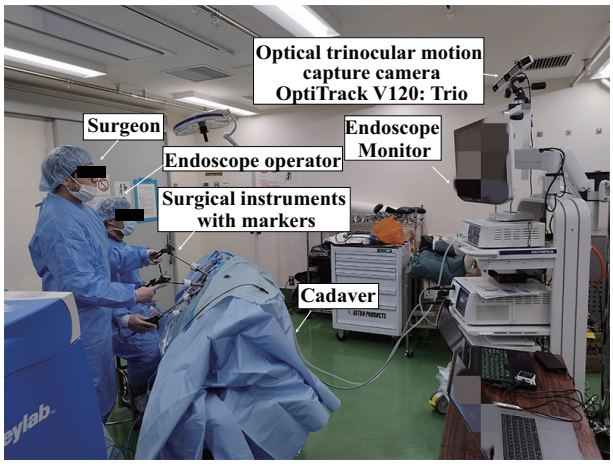


Fig. 1: Laparoscopic surgical training using cadaver.

I. Introduction

Laparoscopic surgery, in which surgical instruments are inserted through multiple small incisions in the abdomen to remove lesions, is expected to shorten recovery time and reduce postoperative pain and infection. This is because it requires fewer incisions and imposes less physical burden on the patient compared to conventional open abdominal surgery. However, laparoscopic surgery demands expert surgical skills due to several challenges: limited freedom in instrument manipulation, difficulty in perceiving the force applied by instruments, a restricted operative field, and the challenge of interpreting three-dimensional anatomical structures from two-dimensional endoscopic images, as illustrated in Fig. 1.

Research on the quantification of surgical skills has been conducted by both domestic and international institutions

∗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 training (CST) sessions were conducted by medical doctors from the Department of Urology, Graduate School of Medicine, Hokkaido University. Instrument motion tracking during the CST sessions was performed in accordance with the *Guidelines for Cadaveric Autopsies in Clinical Medical Education and Research*. All procedures were carried out in a 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

[1]–[4]. For example, Kowalewski et al. evaluated surgical skill by measuring hand movements during a suturing task using a plastic suture pad [3], while Oropesa et al. applied machine learning to assess skill based on forceps behavior in a peg transfer task [4]. However, these studies focused on single tasks performed in simulated environments using organ models, which differ significantly from actual surgical conditions. Although some studies have used animal organs in simulated surgeries [5], [6], they were limited to simple analyses of instrument speed and trajectory.

The authors previously developed a practical laparoscopic training model using porcine organs for (1) tissue dissection and vascular handling around major blood vessels, and (2) suturing of the renal parenchyma [7]. By integrating a laparoscopic motion measurement system into this model, they demonstrated accurate measurement capabilities in complex training tasks involving instrument exchanges. Furthermore, a machine learning-based evaluation method was developed to enable precise skill assessment in these complex tasks [8].

In contrast to simulated tasks, actual surgical procedures are more complex [9], and previous studies may have overlooked important features relevant to skill evaluation. Additionally, because laparoscopic surgery involves multiple distinct procedures, evaluating the entire surgery as a single unit is inappropriate. Instead, each procedure should be analyzed separately. Therefore, in this study, laparoscopic nephrectomy training using cadavers—closely resembling actual surgery—was divided into three major processes based on expert urologists' advice: colon mobilization (process 1), renal vascular dissection (process 2), and incision and removal of remaining tissues (process 3).

Using dynamic motion data of surgical instruments recorded during cadaver-based laparoscopic training (Fig. 1), each surgical process was classified individually. Differences in surgical skill among novices, intermediates, and experts were analyzed by extracting motion features and applying principal component analysis. Based on the data obtained from these training sessions and subsequent machine learning analysis, the study identified distinguishing characteristics among the three skill levels.

In this paper, SHapley Additive exPlanations (SHAP) [10], a method of Explainable AI, was used to interpret the machine learning models applied to skill evaluation in nephrectomy cadaver surgical trainings (CSTs). We report the results of the motion measurement experiments and the machine learning analysis using SHAP. Unlike previous simulation-based studies, our system provides explainable, process-specific, and quantitatively interpretable assessment in cadaveric laparoscopic training.

II. The Measurement Experiment

A. Experimental Setup

The authors developed a system to measure the movements of all surgical instruments used in simulated surgery by attaching uniquely arranged infrared reflective markers to each instrument [11]. In this study, four types of surgical instruments—scissors forceps, grasping forceps, energy devices, and clip applier forceps—were equipped with infrared reflective markers in distinct configurations. Instrument motion was captured using an optical triocular motion capture camera (OptiTrack V120: Trio). The median success rate of kinematic measurements exceeded 90% for all instruments.

B. Subjects and Tasks

The Department of Nephrology and Urological Surgery at Hokkaido University Graduate School of Medicine regularly conducts cadaver surgical training (CST) using donated cadavers. In this experiment, laparoscopic radical nephrectomy (left and right, transperitoneal approach) was used as the measurement target. Four types of surgical instruments were used in the CST: grasping forceps (GF), scissors forceps (SF), clip applier forceps (CA), and an energy device (ED), each equipped with infrared reflective markers for individual recognition. Forty-six subjects were analyzed according to the number of laparoscopic surgeries they had performed in the past.

- (N) Novices (number of operations: 0-9): 17 subjects
- (I) Intermediates (10-49 operations): 19 subjects
- (E) Experts (more than 50 operations): 10 subjects

Although the actual number of subjects was 36, some participants performed CST more than once: four novice and two expert surgeons participated twice, and two novice surgeons participated three times. Since the measurement data contained significant noise and abnormal values due to misrecognition of surgical instruments, abnormal values were excluded, missing values were interpolated, and smoothing was applied [12].

C. Skill Evaluation by Surgical Procedure

Laparoscopic surgery consists of several surgical processes, and the surgical tools used and the surgical skills required differ depending on the surgical process. In this study, the surgical process is divided into the following three steps.

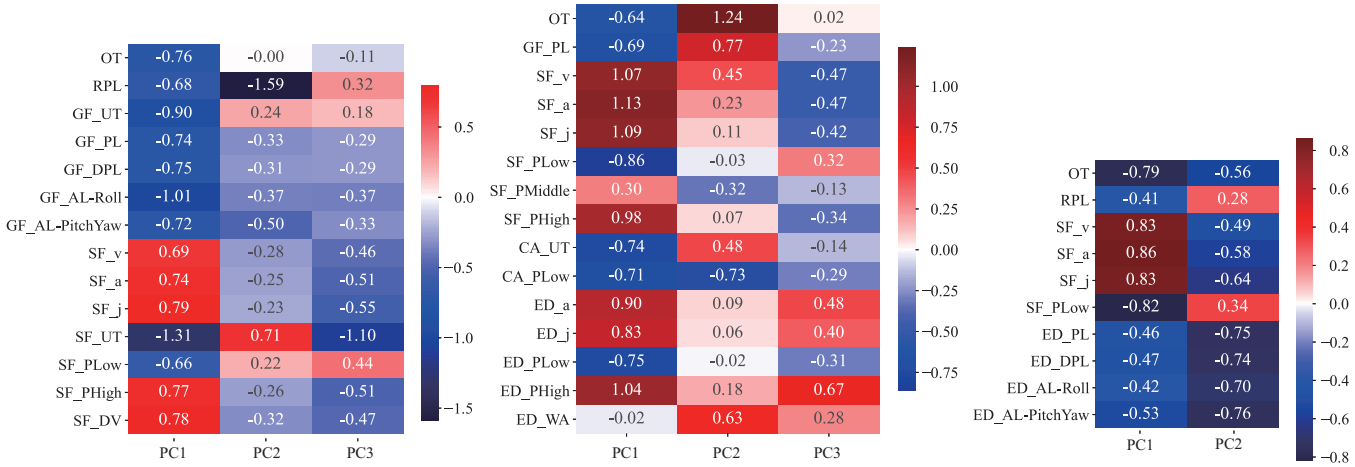
- **Process 1: Colon mobilization**
  In left nephrectomy, this includes mobilization of the descending colon, pancreas, and spleen.
  In right nephrectomy, it includes mobilization of the ascending colon and duodenum.
- **Process 2: Renal vascular dissection**
  Identification, clipping, and transection 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.

In this study, the movements of the surgical instruments for each process 1, 2, and 3, were analyzed and machine learning was to determine the surgeon's skills level.

III. Feature Analysis for Skill Evaluation

A. Motion Metrics

From the movements of the surgical instruments measured in the nephrectomy cadaver training, the following features



(a) Process 1 (b) Process 2 (c) Process 3

Fig. 2: PCA loading heatmaps for each surgical process. The color scale represents the loading weight of each feature on the principal components.

were calculated with reference to the surgical skill evaluation [8] conducted by the authors using porcine organs.

(1) Overall task: operating time (OT), the ratio of the path length of a surgical instrument in both hands (RPL), and using time for each instrument (UT).
(2) Quickness features: average velocity (v), acceleration (a), jerk (j), mean velocity in depth direction (Vd).
(3) Percentage of distribution of velocity range: idle (DVI) (< 0.5 cm/s), low velocity (DVL) ($0.5 \leq v < 2.0$ cm/s), medium velocity (DVM) ($2.0 \leq v < 5.0$ cm/s), high velocity (DVH) ($5.0 \leq v < 12.0$ cm/s), very high velocity (DVVH) (≥ 12.0 cm/s).
(4) Efficiency feature: path length (PL), total path length in depth direction (PLd).
(5) The sum of changes in roll axis attitude angle (AVR), sum of Pitch-Yaw attitude change angles (AVPY).

A total of 111 features were extracted. The Kruskal-Wallis test was used to determine if there were significant differences ($p < 0.05$) among the three groups of novice (N), intermediate (I) and expert (E) surgeons. The principal component analysis is shown in Fig. 2.

B. Machine Learning Model

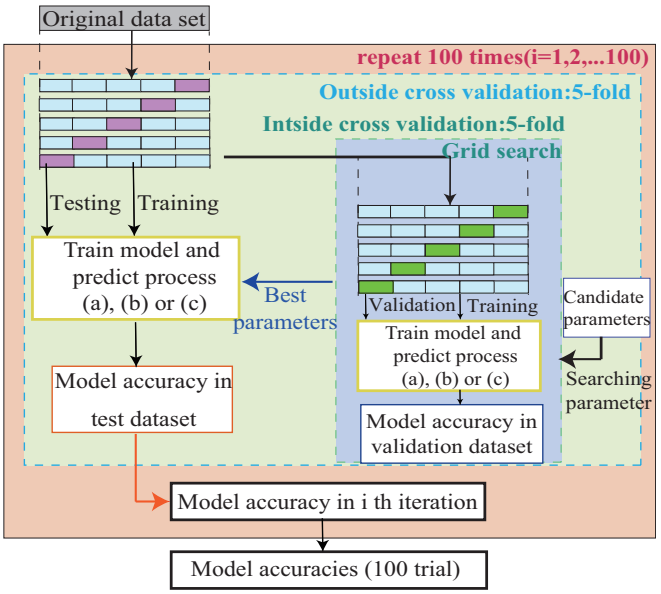
Features were standardized using robust Z-scores. Three models were trained using nested repeated k-fold cross-validation:

- Support Vector Machine (SVM)
- PCA-SVM (PCA followed by SVM)
- Random Forest

The classification accuracy of each method was compared.

Grid search was used to tune hyperparameters such as depth, estimators, kernel coefficients, and PCA thresholds. In this study, machine learning was performed using the forceps behavior features described in the previous section to evaluate each surgeon's skill. The input features were standardized using the robust Z-score, as shown in the following equation:

$$z_i = \frac{x_i - x_m}{\mathrm{NIQR}}, \tag{1}$$



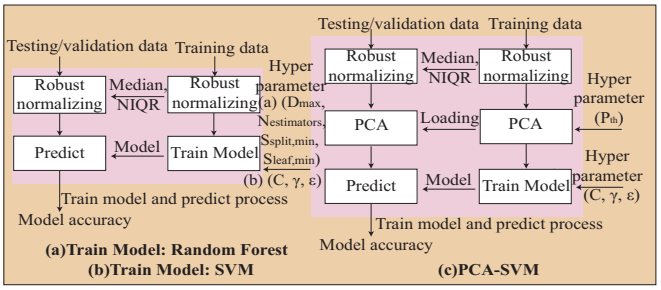


Fig. 3: (a)–(c) correspond to the three models (Random Forest, SVM, PCA-SVM) in each process

TABLE I: Machine Learning Model Discrimination Accuracy by Surgical Procedure (%)

Model	Accuracy (%)
P1: Random Forest	74.22 (71.78–76.39)
P1: SVM	76.22 (74.00–78.22)
P1: PCA-SVM	76.00 (73.56–78.22)
P2: Random Forest	69.78 (67.33–73.72)
P2: SVM	63.33 (60.89–67.56)
P2: PCA-SVM	60.78 (56.89–64.61)
P3: Random Forest	60.67 (56.67–63.06)
P3: SVM	54.22 (50.22–56.67)
P3: PCA-SVM	50.00 (47.56–55.72)

where z_i is the robust Z-score, x_i is the raw value, x_m is the median of the data, and NIQR is the normalized interquartile range, calculated as:

$$\mathrm{NIQR} = 0.7414 \cdot \mathrm{IQR}. \tag{2}$$

Here, IQR denotes the interquartile range.

In each method, the dataset was divided into two parts: one for training and one for testing. Model evaluation was conducted using nested and repeated k-fold cross-validation. In this study, $k = 5$, as shown in Fig. 3.

Since both Random Forest and SVM require hyperparameter tuning, a grid search was performed to find the optimal combination. The hyperparameters are defined as follows:

Random Forest:

- Number of estimators ($N_{\text{estimators}}$): 50, 100, 200
- Maximum depth ($D_{\max}$): None, 10, 20
- Minimum samples to split ($S_{\text{split,min}}$): 2, 5, 10
- Minimum samples per leaf ($S_{\text{leaf,min}}$): 1, 2, 4

SVM:

- Regularization parameter (C)
- RBF kernel parameter (γ)

PCA-SVM:

- Cumulative contribution threshold (P_{th})

Simply presenting CST subjects with novice, intermediate, and expert skill levels does not help them improve, as the reasons behind their classification are unclear. Providing explanations for such evaluations enhances the learning effect. SHAP [10] is a Python library that calculates Shapley values for machine learning models to determine feature importance and provide model interpretability.

The Shapley values were calculated 100 times, and the median value was used in the analysis. They are listed in order of their contribution to the evaluation (i.e., highest Shapley value). In other words, the higher the Shapley value, the greater the contribution to the evaluation of surgical skill.

IV. Results and Discussion

A. Classification Accuracy

Figure 5 and Table I show the accuracy calculated for each of the repeated and nested cross-validation trials for the three machine learning methods used in this study: Random Forest, SVM, and PCA-SVM, based on the accuracy evaluation test conducted during CST. The Kruskal–Wallis test was conducted to evaluate differences among the three machine learning results shown in Fig. 5. If a significant difference was found, the Wilcoxon signed-rank sum test—a nonparametric test for paired samples—was performed for each pairwise comparison among the three results. The test results showed that SVM had significantly higher discrimination accuracy than Random Forest in Process 1, and a higher median accuracy than PCA-SVM, although the difference was not statistically significant. In Process 2, SVM had significantly higher discrimination accuracy than PCA-SVM and a higher median accuracy than Random Forest, though not significantly so. In Process 3, Random Forest had significantly higher discrimination accuracy than the other two methods. Overall, the discrimination accuracy in Process 1 was higher than in Processes 2 and 3, suggesting that this surgical process is more likely to reflect differences in skill based on surgical experience. Although 46 sessions were analyzed, some surgeons participated multiple times. This may introduce partial dependency within the dataset, which will be addressed in future multi-institutional studies.

B. Supplementary Analysis of Classification Performance

Although the overall classification accuracy ranged from approximately 60% to 76% depending on the surgical process, this level of performance is reasonable given the inherent variability in cadaveric procedures and the relatively limited dataset size. In particular, the lower accuracies observed in Processes 2 and 3 likely reflect the increased procedural complexity and inter-surgeon variability during vascular dissection and tissue removal phases.

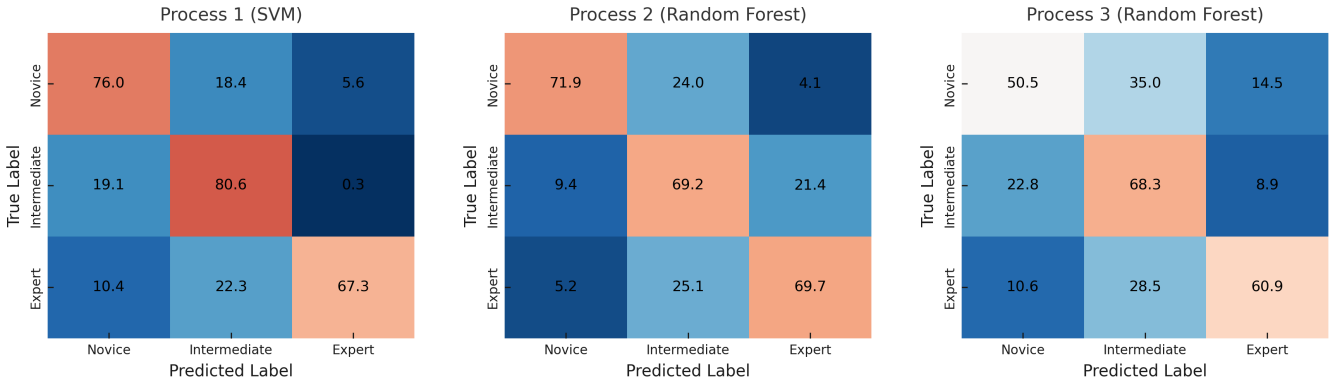
To further interpret the model behavior, confusion matrices were generated for each surgical process, as shown in Fig. 4. The results demonstrate that most misclassifications occurred between adjacent skill levels (*Novice* $\leftrightarrow$ *Intermediate*), while *Expert* surgeons were generally identified correctly. This indicates that the models successfully captured the continuous gradient of surgical proficiency rather than discrete category boundaries.

In addition, Process 1 exhibited the highest classification accuracy across all models, suggesting that motion efficiency and hand coordination features during colon mobilization are more discriminative of experience level. Conversely, the later phases involved greater variability in tool–tissue interaction and procedural strategy, which may obscure distinct motion patterns.

Future improvements will focus on (1) expanding the dataset to include more participants and institutions to mitigate class imbalance and improve generalizability, (2) incorporating temporal or sequence-based representations such as recurrent or attention-based models to capture dynamic motion patterns, and (3) integrating multimodal data—such as endoscopic video, force measurement, and eye-tracking information—to provide a more comprehensive and interpretable evaluation of surgical skill.

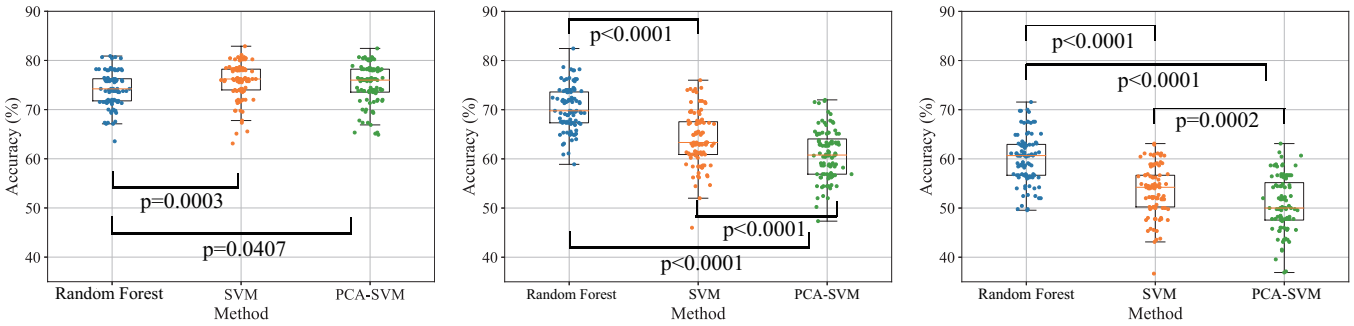
C. Discussion

This study demonstrated that machine learning combined with explainable AI (XAI) can objectively evaluate surgical



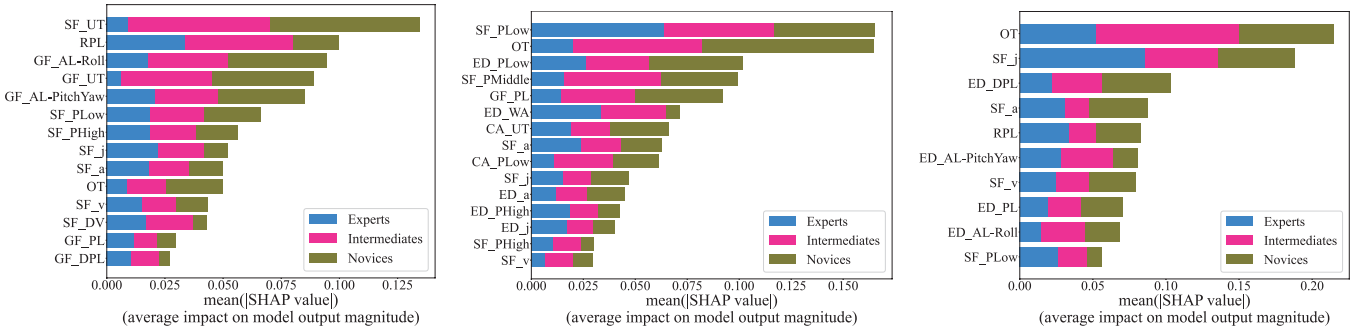
(a) Process 1 (b) Process 2 (c) Process 3

Fig. 4: Confusion matrices for classification of surgical skill levels in each process. Most errors occurred between *Novice* and *Intermediate* groups, indicating that the models captured gradual transitions in surgical proficiency rather than discrete categories.



(a) Process 1 (b) Process 2 (c) Process 3

Fig. 5: Box-and-whisker plots showing classification accuracy of each model for surgical Processes 1–3. Significant differences are indicated by p-values.



(a) Process 1 (b) Process 2 (c) Process 3

Fig. 6: SHAP summary plots of feature contributions in each surgical process. Each point represents a SHAP value for a feature in a single prediction.

skills in cadaveric laparoscopic training. By dividing the nephrectomy procedure into three processes, the system enabled process-specific assessment and clarified model reasoning through SHAP analysis.

The observed accuracy (60–76%) is moderate but consistent with prior motion-based studies. Lower accuracies in Processes 2 and 3 likely result from greater procedural complexity and inter-surgeon variability. Although some participants performed multiple sessions, which may have introduced partial dependency, future multi-institutional validation will improve generalizability.

The SHAP analysis provided interpretable feedback by visualizing the contribution of key motion features, such as the ratio of path length (*RPL*) and usage time of scissors forceps (*SF_UT*). Lower *RPL* and shorter *SF_UT* in experts indicate coordinated and efficient instrument handling, offer-

ing concrete educational insights.

Limitations include the use of a single procedure and reliance solely on motion data. Expanding to other surgical tasks and integrating multimodal information such as video, force, or gaze data would yield a more comprehensive skill evaluation. Ultimately, this system integrates hardware, software, and explainable feedback, representing a step toward intelligent, data-driven surgical education.

D. SHAP-based Explanation

Figure 6 shows the SHAP summary plots for the three surgical processes. In Process 1 (colon mobilization), the time spent using the scissors forceps (SF) was the most influential feature for classification, and the ratio of the path length of a surgical instrument in both hands (RPL) indicated the efficiency of forceps usage. In Process 2 (renal vascular dissection), the low-velocity range ratio of the clip applier (CA) showed the strongest contribution, indicating that velocity was an important factor. In Process 3 (incision and resection of the remaining tissues), operative time was identified as the most informative feature. These results suggest that efficiency, velocity of right-hand forceps (SF and ED), and tool-specific behavioral metrics are key indicators of surgical proficiency.

V. Conclusions

The system represents an integration of hardware (optical motion capture), software (machine learning analysis), and feedback interface (SHAP-based visualization), aligning with the SII theme of intelligent system integration. In this study, we developed a practical evaluation system for laparoscopic surgical skills based on the experimental data obtained from cadaver training. The results were compared among several machine learning methods and discussed through accuracy validation. Overall, the discrimination performance of Process 1 was high, suggesting that the difference in skill by the number of surgical experiences is likely to appear in this process.

The contribution of input motion features to the classification of novice, intermediate, and expert surgeons by machine learning using SHAP, one of the Explainable AIs, was shared based on the motion measurement data of surgical instruments in 46 CST cases. Although deep learning is often used for such classification problems, neural networks, as typified by deep learning, are highly accurate but cannot explain why the decision was made in such a way. Therefore, in this study, we used machine learning models such as SVM and PCA, which can easily identify which feature values differentiate novices from experts, and visualized each feature value for evaluation using SHAP. The visualization of the important features of each surgical procedure will help the surgeons to understand what improvements can lead to the improvement of surgical skills and will lead to effective surgical education.

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] H. Egi, et al., "Objective assessment of endoscopic surgical skills by analyzing direction dependent dexterity using the Hiroshima university endoscopic surgical assessment device (HUESAD)," *Surgery Today*, vol. 38, pp. 705–710, 2008.

[2] M. K. Chmarra, et al., "TrEndo, a device for tracking minimally invasive surgical instruments in training setups," *Sensors and Actuators A*, vol. 126, pp. 328–334, 2006.

[3] K. F. Kowalewski, et al., "Sensor-based machine learning for workflow detection and as key to detect expert level in laparoscopic suturing and knot-tying," *Surgical Endoscopy*, vol. 33, pp. 3732–3740, 2019.

[4] I. Oropesa, et al., "Supervised classification of psychomotor competence in minimally invasive surgery based on instruments motion analysis," *Surgical Endoscopy*, vol. 28, no. 2, pp. 657–670, 2014.

[5] E. F. Hofstad, et al., "Psychomotor skills assessment by motion analysis in minimally invasive surgery on an animal organ," *Minimally Invasive Therapy & Allied Technologies*, vol. 26, no. 4, pp. 240–248, 2017.

[6] S. G. T. Smith, J. Torkington, T. J. Brown, N. J. Taffinder, and A. Darzi, "Motion analysis," *Surgical Endoscopy*, vol. 16, no. 4, pp. 640–645, 2002.

[7] M. Higuchi, et al., "Development and validation of a porcine organ model for training in essential laparoscopic surgical skills," *International Journal of Urology*, vol. 27, no. 10, pp. 929–938, 2020.

[8] K. Ebina, et al., "Automatic assessment of laparoscopic surgical skill competence based on motion metrics," *PLoS One*, vol. 17, no. 11, pp. 1–13, 2022.

[9] 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.

[10] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, vol. 30, pp. 4768–4777, 2017.

[11] K. Ebina, et al., "A surgical instrument motion measurement system for skill evaluation in practical laparoscopic surgery training," *PLoS One*, vol. 19, no. 6, e0305693, 2024.

[12] B. P. Rai, J. U. Stolzenburg, S. Healy, et al., "Preliminary validation of Thiel embalmed cadavers for laparoscopic radical nephrectomy," *Journal of Endourology*, vol. 29, pp. 595–603, 2015.