



HOKKAIDO UNIVERSITY

Title	Evaluation of Laparoscopic Surgical Skill Using Surgical Instrument Motion Analysis and Machine Learning
Author(s)	晏, 凌波
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第17045号
Issue Date	2026-06-30
DOI	https://doi.org/10.14943/doctoral.k17045
Doc URL	https://hdl.handle.net/2115/100098
Type	doctoral thesis
File Information	Lingbo_Yan.pdf, 全文



Doctoral Dissertation

**Evaluation of Laparoscopic Surgical Skill Using Surgical Instrument
Motion Analysis and Machine Learning
(手術器具動作解析と機械学習による腹腔鏡手術技量評価)**

Lingbo Yan (晏 凌波)



HOKKAIDO
UNIVERSITY

May, 2026

Course of Systems Science and Informatics
Graduate School of Information Science and Technology
Hokkaido University

Doctoral Dissertation
submitted to Graduate School of Information Science and Technology,
Hokkaido University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy.

Lingbo Yan (晏 凌波)

Dissertation Committee: Professor Atsushi Konno (Chief examiner)
Professor Takayuki Tanaka
Professor Natsuki Miyata
Professor Takashige Abe

Copyright © 2026 by Lingbo Yan. All rights reserved.

Evaluation of Laparoscopic Surgical Skill Using Surgical Instrument Motion Analysis and Machine Learning (手術器具動作解析と機械学習による腹腔鏡手術技量評価)*

Lingbo Yan (晏 凌波)

Abstract

Laparoscopic surgery requires refined instrument manipulation, stable visual control, and high-level bimanual coordination, making objective and efficient assessment of technical skills essential for contemporary surgical training. While multiple frameworks have been proposed to quantify surgical performance, most lack the ability to capture continuous, fine-grained instrument motion under realistic anatomical conditions. To address this gap, the present study applied and extended a previously validated optical motion-capture (MoCap) methodology—originally established in porcine wet-lab training—to the more anatomically complex environment of cadaveric laparoscopic nephrectomy. This transition allowed examination of the feasibility, accuracy, and educational applicability of MoCap-based motion analysis in a high-fidelity training model that closely approximates clinical operative settings.

Rigid-body marker clusters were attached to standard laparoscopic instruments, enabling three-dimensional tracking of instrument trajectories throughout the operative workflow without interfering with surgical tasks or ergonomics. Building upon this validated platform, motion data were collected from 46 cadaveric laparoscopic nephrectomy procedures performed by 36 surgeons with varying levels of experience.

Participants were stratified into 3 tiers including novices (0–9 cases), intermediates (10–49 cases), and experts (≥ 50 cases) according to previous surgical experiences. The procedure was subdivided into 3 phases: phase 1 = colon mobilization, phase 2 = renal vascular dissection and division, and phase 3 = liberation from remaining tissues.

A comprehensive set of motion features—including velocity-, acceleration-, spatial distribution-, and path-efficiency-related metrics—was extracted to characterize differences in operative technique across skill levels.

Statistical analyses revealed skill-associated trends in multiple motion domains, supporting the validity of MoCap-derived metrics as indicators of surgical expertise. Machine-learning classification models trained on these features were able to differentiate novice, intermediate, and expert surgeons defined by prior laparoscopic operative volume, demonstrating the potential for automated expertise stratification. Furthermore, regression models were constructed to estimate GOALS (Global Operative Assessment of Laparoscopic Skills) scores based on expert evaluations, allowing quantitative evaluation of surgical performance with high interpretability. These findings illustrate that combining MoCap with data-driven learning methods provides a powerful framework for evaluating surgical proficiency in high-fidelity training settings.

Although exploratory telementoring experiments were conducted in parallel as an independent research line, these did not integrate MoCap data in the present study. Nonetheless, the ability to

*Doctoral Dissertation, Course of Systems Science and Informatics, Graduate School of Information Science and Technology, Hokkaido University, SSI-DT46235022, May 3, 2026.

share spatial motion information in virtual environments represents a compelling future direction for remote surgical support and education. Overall, this thesis establishes a rigorously validated, motion-based analytical framework for objective assessment of laparoscopic skill in both wet-lab and cadaveric environments, and provides foundational insights for future extensions toward clinical training, simulation development, and technologically enhanced surgical education.

Keywords: motion capture; laparoscopic surgery; cadaveric training; surgical skill assessment; machine learning

Contents

1	Introduction	1
1.1	Background	1
1.2	Limitations of Existing Surgical Skill Assessment Frameworks	2
1.2.1	Common limitations among existing frameworks	3
1.3	Japanese Context: ESSQ and Its Constraints	3
1.4	Need for Objective and Data-Driven Assessment	4
1.5	Cadaver and Animal Models in Laparoscopic Education	4
1.5.1	Overview of High-Fidelity Training Environments	4
1.5.2	Cadaver-Based Training and the Thiel Embalming Method	4
1.5.3	Animal Organ Models and Wet-Lab Training	6
1.5.4	Need for Objective Assessment in High-Fidelity Environments	6
1.5.5	Tele-Mentoring and Early-Stage Feasibility Evaluation	7
1.6	Limitations of Motion-Analysis-Based Skill Assessment and the Need for Expanded Orientation Features	8
1.7	Research Aims	8
1.8	Structure of This Dissertation	9
2	Measurement system	11
2.1	Optical motion capture system	11
2.1.1	Hardware configuration	11
2.1.2	Marker design for laparoscopic instruments	11
2.1.3	Calibration procedure and definition of the global coordinate system	12
2.1.4	Measurement accuracy and sampling performance	14
2.1.5	Limitations of optical tracking in surgical environments	15
2.2	Experimental environment for cadaveric laparoscopic training	15
2.2.1	Experimental setup and operating room configuration	15
2.2.2	Surgical task: laparoscopic nephrectomy	16
2.2.3	Participant profiles	17
2.2.4	Instruments tracked in cadaver experiments	17
2.2.5	Data recording and reconstruction protocol	18
2.2.6	Post-training questionnaire and surgeon grouping	19
2.3	Real surgery: Motion data acquisition in a clinical operating room	19
2.3.1	Marker-attached laparoscopic instruments	20
2.3.2	Operating-room configuration	21
2.3.3	Role of real-surgery data	21

2.4	Wet-Lab (Porcine Organ) Experiments	22
2.4.1	Experimental Setup	22
2.4.2	Task Design and Experimental Workflow	22
2.4.3	Motion Capture Procedure	26
2.5	Data Synchronization and Management	26
2.5.1	Synchronization procedure in cadaveric experiments	26
2.5.2	Synchronization in Real Laparoscopic Surgery	27
2.5.3	Synchronization procedure in wet-lab experiments	27
2.5.4	File structure and dataset standardization	27
2.5.5	Rationale and implications for analysis	27
2.6	Tele-mentoring and Remote Surgical Support System	28
2.6.1	System architecture	28
2.6.2	Data collected in the tele-surgery project	28
2.6.3	Relevance to this thesis	29
2.7	Ethical Approval	29
2.7.1	Cadaveric laparoscopic motion capture	29
2.7.2	Porcine wet-lab experiments	30
2.7.3	Real-surgery motion measurement	30
2.7.4	Telementoring studies	30
2.7.5	Data protection and anonymization	30
2.8	Summary	30
3	Feature Extraction and Statistical Analysis	31
3.1	Overview of the Feature Extraction Pipeline	31
3.2	Feature Extraction	32
3.2.1	Temporal and Spatial Workspace Features	32
3.2.2	Path-Based Metrics	33
3.2.3	Speed-, Acceleration-, and Jerk-Based Metrics	33
3.2.4	Velocity Distribution Metrics	33
3.2.5	Orientation Metrics	34
3.2.6	Bimanual Coordination Metrics	34
3.2.7	Summary	34
3.3	Feature Pre-processing	34
3.3.1	Handling missing values	35
3.3.2	Temporal smoothing	35
3.3.3	Robust normalization	35
3.3.4	Outlier handling	35
3.3.5	Statistical feature filtering	35
3.3.6	Summary	35
3.4	Statistical Analysis	35
3.4.1	Group definitions	36
3.4.2	Univariate statistical testing	36
3.5	Data Quality and Descriptive Visualization	36
3.5.1	Measurement rate of motion capture	36
3.5.2	Summary of screened feature set	37
3.5.3	Definition of the Region of Interest (ROI)	37
3.5.4	Visualization of instrument usage timeline	38

3.5.5	Three-dimensional trajectory visualization	38
3.6	Skill-related feature analysis	41
3.6.1	Participants' backgrounds	41
3.6.2	Comparison of representative kinematic features	41
3.6.3	Summary and implications	42
3.7	Surgical Phase Segmentation	44
3.7.1	Definition of the three phases	44
3.7.2	Purpose of phase-based skill evaluation	44
3.8	Summary	44
4	Machine-Learning Based Surgical Skill Classification	47
4.1	Principal Component Analysis	48
4.1.1	PCA for three-group classification	48
4.1.2	PCA for ESSQ-based binary classification	49
4.1.3	Summary	50
4.2	System Construction for Machine-Learning-based Skill Classification	54
4.2.1	Overview of the analytical workflow	54
4.2.2	Advanced workflow with SHAP interpretability	54
4.2.3	Rationale for pipeline design	55
4.3	Classification Results	57
4.3.1	Three-class classification results	57
4.3.2	ESSQ classification results	57
4.3.3	Confusion matrix analysis	59
4.3.4	Summary of classification performance	59
4.4	SHAP-based Interpretation of Classification Models	63
4.4.1	SHAP analysis for Phase 1	63
4.4.2	SHAP analysis for Phase 2	63
4.4.3	SHAP analysis for Phase 3	63
4.4.4	Summary of SHAP-based interpretation	64
4.5	Summary	65
5	Regression Modeling for Continuous Surgical Skill Assessment	67
5.1	GOALS rating and inter-rater reliability	67
5.1.1	GOALS rating procedure	67
5.1.2	Statistical analysis of inter-rater reliability	68
5.1.3	Results of inter-rater reliability	68
5.2	Regression model construction for GOALS score prediction	69
5.2.1	Regression algorithms	69
5.2.2	Data preprocessing	69
5.2.3	Overall Regression-Model Framework	69
5.3	Regression analysis for GOALS scores in Phase 1	70
5.3.1	Model accuracy (MAE)	71
5.3.2	Comparison of predicted and actual scores	73
5.3.3	Item-wise prediction performance	73
5.3.4	Summary of Phase 1	73
5.3.5	Regression performance in Phase 2	73
5.3.6	Regression performance in Phase 3	77

5.3.7	Correlation between Expert Ratings and Model Predictions	81
5.4	Summary	84
6	Educational Applications and Integration Into Tele-mentoring Systems	85
6.1	Real Surgery Experiments	85
6.1.1	Measurement System and Rigid-Body Marker Attachment	85
6.1.2	Trajectory Recording and Preprocessing	85
6.1.3	Tracking Success Rate Across the Six Surgical Cases	86
6.1.4	Surgeon Questionnaire on Marker Interference	86
6.2	Wet-Lab (Porcine) Experiments	89
6.2.1	Overview of the Wet-Lab Environment	89
6.2.2	On-Site Skill Assessment by Expert Surgeons	89
6.2.3	Integrated Feedback Interface	89
6.2.4	Summary	89
6.3	Remote Surgery Experiments	90
6.3.1	Motivation and Background	90
6.3.2	Remote Multi-View Surgical Streaming System	90
6.3.3	Measurement of Communication Latency	91
6.3.4	Remote Surgical Guidance Trial	91
6.3.5	Technical Considerations and Limitations	92
6.3.6	Future Directions	92
6.4	Summary	92
7	Conclusions	95
7.1	Summary	95
7.2	Future Work	96
	Acknowledgements	99
	Bibliography	101

List of Figures

2.1	Overview of OptiTrack V120:Trio camera system [39]. The device integrates three infrared cameras into a rigid housing, enabling 120 Hz marker-based tracking in a compact surgical workspace.	12
2.2	OptiTrack rigid L-frame used for calibration. The L-shaped frame is placed in the XZ plane, where the X axis represents the depth direction of laparoscopic manipulation, Y is the vertical axis, and Z serves as an in-plane reference axis (cadaver head–foot direction in the cadaver experiment).	13
2.3	Overview of the cadaveric laparoscopic training environment. The OptiTrack V120:Trio motion capture camera is mounted above the operating field, tracking laparoscopic instruments equipped with reflective markers. The setup includes the surgeon, endoscope operator, endoscopic monitor, and a Thiel-embalmed cadaver.	16
2.4	Surgical instruments used in the cadaveric laparoscopic training. From left to right and top to bottom: Grasping forceps (GF), Scissor forceps (SF), Energy device (LigaSure, ED), Clip applier (CA), Suction tube (ST), and Right-angle forceps (RF). These abbreviations (GF, SF, ED, CA, ST, RF) are used consistently in the feature extraction and analysis throughout this dissertation.	18
2.5	Setup of the motion capture system in real clinical laparoscopic nephrectomy at Hokkaido University Hospital.	20
2.6	Sterilizable marker-attached laparoscopic instruments used for real surgical motion capture: Grasping Forceps (GF), Scissor Forceps (SF), Energy Device (Liga-Sure, ED), and Clip Applier (CA). All instruments were equipped with rigid-body marker clusters designed to withstand sterilization and avoid interference with intraoperative manipulation.	21
2.7	Wet-lab experimental setup used at the Simulation Center of Hokkaido University Hospital. Participants performed the tasks in a standing position while using an HD laparoscopic tower and a box trainer equipped with porcine organs. The OptiTrack V120 Trio optical motion capture system was positioned in front of the trainer to record 3D instrument trajectories throughout each task.	23
2.8	Three-port configuration of the box trainer used in the porcine experiments. All ports were 12 mm in diameter, consisting of one central port for the endoscopic camera and two lateral ports positioned at clinically relevant angles for instrument insertion.	23

2.9	Rigid-body markers attached to the needle holders used in Task 2. (A) Needle holder used in the left (assisting) hand. (B) Needle holder used in the right (dominant) hand. These instruments were newly introduced only for the suturing task in the porcine wet-lab experiment and differ from the grasping and dissecting forceps used in the cadaveric nephrectomy training. Each needle holder was equipped with a four-marker rigid-body cluster near the handle, allowing 6-DoF motion tracking with the OptiTrack V120 Trio system under the same coordinate definition as in the cadaver experiments.	24
2.10	Porcine organ fixation within the box trainer. The tissues were secured to a stainless-steel grid using elastic wires, providing a stable platform for standardized dissection tasks.	24
2.11	Laparoscopic suturing task using porcine organs in the wet-lab environment (Task 2). Needle driving and intracorporeal knot tying were performed using marker-attached needle holders. Motion data were recorded to obtain instrument tip position and orientation during suturing.	25
2.12	Remote mentoring workstation at Hokkaido University (mentor side). The instructor monitors multiple synchronized video streams (endoscopic view, room-side view, and console/robotic view) during the tele-mentoring session with Obihiro Kosei Hospital.	29
3.1	Overview of the motion-capture feature extraction pipeline. All datasets (cadaveric, real surgery, and wet-lab) were processed using the same standardized procedure.	32
3.2	Representative data of instrument usage timeline during the cadaveric nephrectomy session. Each horizontal bar indicates an active period of a specific instrument, illustrating the overall workflow and tool transitions. Brief and clearly identifiable misclassification segments were manually removed following verification with synchronized video.	37
3.3	Representative data of three-dimensional reconstructed trajectories of six instruments during the cadaveric nephrectomy. Each subplot visualizes the cleaned, interpolated, and outlier-removed motion of one instrument, with trocar positions provided as spatial reference.	39
3.4	Overview of ROI definition methods. (a) ROI definition based on recorded port coordinates. (b) ROI definition based on anatomical markers when port coordinates were unavailable.	40
3.5	Boxplots of key kinematic skill indices across three experience levels. (Left) Operation time (OT); (Center) Low-velocity ratio of scissor forceps (DVL); (Right) Path length of grasping forceps (PL). Statistical differences were evaluated using Kruskal–Wallis test.	43
4.1	PCA loading heatmaps for three-group classification: (a) whole phase, (b) Phase 1, (c) Phase 2, and (d) Phase 3.	52
4.2	PCA loading heatmaps for ESSQ-based binary classification: (a) Whole Phase, (b) Phase 1, (c) Phase 2, and (d) Phase 3.	53
4.3	Overview of the machine-learning analysis pipeline. The workflow integrates pre-processing, feature extraction, statistical filtering, PCA-based dimensionality reduction, classifier training, and nested cross-validation	56

4.4	Three-class classification accuracies for the whole phase and phases 1–3. Random Forest showed significantly higher performance than SVM and PCA-SVM in most comparisons. Statistical significance was evaluated using the Kruskal–Wallis test.	58
4.5	ESSQ classification accuracy for the whole phase and phases 1–3. Overall accuracy was substantially higher than in the three-class task. Random Forest consistently achieved the highest accuracy.	61
4.6	Confusion matrices for phases 1–3.	62
4.7	SHAP summary plot for Phase 1 using the PCA-SVM classifier. Positive SHAP values contribute to predicting higher skill classes.	64
4.8	SHAP summary plot for Phase 2 using the PCA-SVM classifier.	65
4.9	SHAP summary plot for Phase 3 using the PCA-SVM classifier.	66
5.1	Agreement between two expert raters for GOALS total scores in each phase. Each point represents one cadaveric nephrectomy trial, and the dashed line indicates perfect agreement ($y = x$).	68
5.2	Overall regression-model pipeline. Feature preprocessing, robust normalization, and selection of significant kinematic indices.	70
5.3	Predicted versus actual GOALS total scores for Ridge regression (Phase 1). The dashed red line represents the identity line.	74
5.4	Prediction performance of six GOALS items using Ridge, PLS, and PCA-SVR models (Phase 1).	75
5.5	Relationship between predicted and actual GOALS total scores for the Ridge regression model in Phase 2. Each boxplot shows the distribution of predicted scores over the repeated cross-validation runs for a given trial; the red dashed line indicates the identity line.	78
5.6	Prediction performance of GOALS items using Ridge, PLS, and PCA-SVR models (Phase 2).	79
5.7	Relationship between predicted and actual GOALS total scores for the Ridge regression model in Phase 3. Compared with Phases 1 and 2, the boxplots are more tightly clustered around the identity line, indicating improved prediction accuracy.	82
5.8	Prediction performance of GOALS items using Ridge, PLS, and PCA-SVR models (Phase 3).	83
6.1	Sterilized rigid-body marker attached to a laparoscopic instrument during a real donor nephrectomy. The rigid body was fabricated using autoclavable guide resin and attached using sterile bands.	86
6.2	Example 3D trajectories recorded during laparoscopic donor nephrectomy (Case 4). Left: grasping forceps; center: energy device; right: scissor forceps. Red indicates measured points inside the ROI, orange indicates interpolated points, and gray denotes out-of-ROI samples.	87
6.3	Representative instrument usage timeline for a full nephrectomy. Usage intervals are shown for grasping forceps (Karl-Storz), LigaSure, and scissor forceps (Olympus). Duplicate usage time was zero, indicating consistent identification of each instrument.	88
6.4	Survey results assessing perceived interference caused by the instrument-mounted rigid-body markers. Scores were evaluated on a five-point Likert scale (1 = no interference; 5 = strong interference).	88

6.5	Example of the integrated feedback interface used in the wet-lab porcine experiments. The interface displays GOALS assessment results, PCA score comparison, velocity symmetry metrics, expert feedback, and quantitative indices extracted for each instrument.	90
6.6	Example of the remote surgical streaming interface used in the pilot experiment. Multiple video sources—including laparoscopic view, robotic surgical view, and room-view camera—were transmitted simultaneously using the Zao Cloud View system®. Video compression artifacts and temporary pixelation were observed during network congestion, highlighting the impact of bandwidth and latency on remote surgical observation.	91

List of Tables

1.1	Comparison of major surgical skill assessment frameworks.	5
2.1	Specification of OptiTrack V120:Trio [39].	12
3.1	Measurement rate of surgical instruments. Values are expressed as median (1st quartile–3rd quartile).	37
4.1	Classification accuracies for three-class models across whole phase and phases 1–3. Values represent median accuracy with interquartile range (Q1–Q3).	59
4.2	Classification accuracies for ESSQ qualification prediction (two-class). Values represent median accuracy with interquartile range (Q1–Q3).	60
4.3	Classification accuracies for ESSQ qualification prediction (two-class). Values represent median accuracy with interquartile range (Q1–Q3).	61
5.1	Comparison of MAE among PCA-SVR, PLS, and Ridge regression models in Phase 1. Values indicate median MAE. Statistical comparisons were performed using the Kruskal–Wallis test with post-hoc pairwise testing.	72
5.2	Comparison of MAE among PCA-SVR, PLS, and Ridge regression models in Phase 2. Values indicate median MAE across all cases. Statistical comparisons were performed using the Kruskal–Wallis test with post-hoc pairwise testing. . .	76
5.3	Comparison of MAE among PCA-SVR, PLS, and Ridge regression models in Phase 3. Values indicate median MAE across all cases. Statistical comparisons were performed using the Kruskal–Wallis test with post-hoc pairwise testing. . .	80
5.4	Spearman correlation (ρ) between expert scores and Ridge regression predictions.	81
6.1	Tracking success rates across six completed surgical cases.	86
6.2	Average perceived-interference score for each instrument.	87

Chapter 1

Introduction

1.1 Background

Over the past three decades, minimally invasive surgery (MIS) has become a central pillar of abdominal and pelvic procedures. Compared with traditional open surgery, laparoscopic approaches offer reduced postoperative pain, shorter hospital stays, faster recovery, and superior cosmetic outcomes. These benefits have driven the widespread adoption of laparoscopic surgery across multiple specialties, including general surgery, gynecology, and urology [1, 2]. Concurrently, this transition has introduced new demands for surgical education, as the psychomotor requirements of laparoscopic surgery differ fundamentally from those of open procedures.

Laparoscopic surgery requires surgeons to manipulate long instruments inserted through trocars while relying on a two-dimensional video display. This configuration introduces several technical challenges: loss of natural depth perception, counterintuitive fulcrum effects at the abdominal wall, a restricted range of motion, and reduced haptic feedback [3]. To navigate these constraints, surgeons must develop refined eye–hand coordination, bimanual dexterity, and precise control of instrument tips within a confined working space [4, 5]. These requirements are especially pronounced in complex procedures such as laparoscopic nephrectomy, where safe dissection of the kidney and its surrounding structures demands both advanced technical skill and procedural judgment.

Despite advances in instrumentation and imaging technology, the learning curve for laparoscopic surgery remains steep. Large clinical series have shown that complication rates, operation times, and conversion rates are higher during the early phase of a surgeon’s laparoscopic experience, even at high-volume centers [1]. In response, professional societies and national boards have introduced qualification systems and minimum case requirements to ensure that only adequately trained surgeons perform advanced laparoscopic procedures [6]. These developments reflect a broader shift: technical proficiency in MIS is now regarded as a core competency that must be systematically cultivated and objectively assessed [7].

However, opportunities for experiential learning in the operating room (OR) are increasingly constrained. Patient safety is a primary concern, and institutions are less willing to tolerate prolonged procedures or avoidable complications for training purposes. Meanwhile, duty hour regulations and work–life balance considerations have reduced the time residents can spend in the OR. Consequently, a gap has emerged between the complexity of laparoscopic procedures and the practical exposure that trainees can realistically obtain through routine clinical work.

Simulation-based training has therefore become an important complement outside OR experience [2]. A wide range of platforms is now available, from simple box trainers to high-fidelity vir-

tual reality (VR) simulators. Randomized trials and prospective studies have shown that structured simulator training can improve operative performance in both basic and advanced tasks [8, 9, 10]. Commercial systems such as LapMentor® provide standardized curricula covering camera navigation, fundamental laparoscopic tasks, and simulated full procedures [11]. These tools have contributed significantly to the standardization of early-phase laparoscopic skill acquisition.

Nevertheless, current simulation environments do not fully replicate the demands of real clinical operations. Box trainers are cost-effective and easy to use but lack realistic tissue properties and anatomical context. VR simulators offer controlled environments and objective metrics, yet limitations remain in haptic fidelity, anatomical realism, and representation of full procedural workflows [12]. Even high-fidelity systems typically focus on short segments of an operation rather than modeling the entire workflow from port placement to specimen extraction. As a result, notable discrepancies persist between performance in simulation and performance in actual surgery.

These limitations raise two central challenges for surgical education. First, how can training environments better reproduce the visual, tactile, and procedural complexity of real laparoscopic surgery. Second, how can technical proficiency be evaluated objectively and reproducibly in such environments, beyond subjective impressions or coarse global rating scales. A substantial body of research has explored the use of motion tracking, force measurement, and video analysis to quantify psychomotor performance in bench models, animal organs, and VR simulators [13, 14, 15, 16, 17]. These studies have demonstrated that motion- and sensor-based metrics can distinguish between novices and experts and can be used to monitor learning curves in simulation settings [18, 19, 20].

However, most prior work has been conducted in simplified or highly constrained experimental tasks—such as peg transfer, isolated suturing, or single-step wet-lab procedures [21, 22, 23, 14]. Although such studies are valuable for analyzing fundamental psychomotor skills, their findings may not generalize to near-clinical, multi-stage laparoscopic operations. Full procedures involve dynamic anatomy, shifting camera viewpoints, multiple instruments, and context-dependent strategy modulation. Evaluating performance under these conditions requires analytical frameworks capable of handling long-duration workflows, staged task structures, and diverse patterns of instrument interaction while still providing interpretable results [7].

In summary, the rise of laparoscopic surgery has created an urgent need for training and assessment paradigms that move beyond traditional apprenticeship models. Simulation-based education has become indispensable, yet current systems do not fully capture the complexity of real operations. At the same time, motion-based metrics have emerged as powerful tools for objective skill assessment but have been applied primarily to simplified tasks. Bridging these gaps—through realistic training environments and robust, interpretable, and clinically relevant evaluation frameworks—remains a central challenge in contemporary laparoscopic education [24]. This challenge forms the broader educational and clinical background against which the present dissertation is positioned.

1.2 Limitations of Existing Surgical Skill Assessment Frameworks

Surgical skill assessment has become a core component of modern surgical education as the complexity of minimally invasive procedures continues to increase. Although apprenticeship-based training remains foundational, contemporary surgical practice requires standardized and reproducible assessment systems to ensure patient safety and objectively evaluate trainee progression. Over recent decades, several structured frameworks have been introduced internationally, including

the Objective Structured Assessment of Technical Skills (OSATS) [26], the Global Operative Assessment of Laparoscopic Skills (GOALS) [27], and the Global Evaluative Assessment of Robotic Skills (GEARS) [28]. Each framework has contributed significantly to the evolution of surgical education, yet all possess important limitations. Table 1.1 summarizes a structured comparison of major assessment systems.

OSATS and its limitations: OSATS is one of the earliest standardized assessment frameworks and is based on global rating scales (GRS) that evaluate fundamental surgical skills [26]. The strengths of OSATS include its broad applicability and ease of implementation. However, the assessment remains inherently subjective, as the scoring heavily depends on the evaluator’s experience and judgment.

GOALS and the constraints of laparoscopic rating scales: GOALS [27] was developed specifically for laparoscopic surgery and evaluates performance in five domains: depth perception, bimanual dexterity, efficiency, tissue handling, and autonomy. Although GOALS demonstrates reasonable construct validity, its reliance on visual judgment introduces variability among evaluators. Prior studies report only moderate inter-rater reliability, particularly in the efficiency and tissue-handling domains [21, 22]. Furthermore, GOALS ratings are typically based on two-dimensional operative videos, which cannot capture subtle three-dimensional motion or fine motor control.

GEARS for robotic surgery: GEARS extends structured evaluation to robotic surgery and incorporates domains specific to telemanipulation, such as console control and instrument articulation [28]. While GEARS is well suited for robotic systems, it remains a subjective visual assessment framework and is not directly transferable to laparoscopic environments.

1.2.1 Common limitations among existing frameworks

Despite their widespread adoption, OSATS, GOALS, and GEARS share several intrinsic limitations:

- **Subjectivity and rater variability:** Evaluations depend heavily on individual assessors’ experience and interpretation, resulting in inconsistent ratings across evaluators and institutions.
- **Limited sensitivity to complex surgical tasks:** These systems were largely validated on basic tasks and often fail to capture performance in multi-stage laparoscopic procedures.
- **Lack of quantitative motion information:** Visual assessment cannot reliably quantify trajectory smoothness, three-dimensional path efficiency, rotational control, or bimanual coordination.
- **High time and labor burden:** Reviewing full-length operative videos requires substantial expert time, limiting scalability in large educational programs.

1.3 Japanese Context: ESSQ and Its Constraints

In Japan, surgical skill assessment has taken a particularly rigorous form. The Endoscopic Surgical Skill Qualification System (ESSQ), established by the Japanese Society of Endourology, is one of the most demanding national credentialing systems for laparoscopic surgery [6]. Applicants submit unedited operative videos that are evaluated by multiple certified experts, and only a minority achieve qualification. ESSQ has played a major role in standardizing the quality of advanced laparoscopic surgery nationwide.

However, ESSQ evaluations rely predominantly on subjective video-based expert review, leading to several constraints:

- **Subjective interpretation:** Scoring depends strongly on individual evaluators, and inter-rater reliability is not always quantified.
- **Case-context sensitivity:** Variations in patient anatomy, case difficulty, or assistant performance may influence assessments.
- **Limited constructive feedback:** ESSQ does not provide objective metrics indicating how trainees should improve or which aspects of instrument control require refinement.
- **Single-case bias:** Qualification is based on a single submitted case, which may not reflect a surgeon's overall performance level.

1.4 Need for Objective and Data-Driven Assessment

The limitations shared by existing international frameworks and the ESSQ highlight the need for objective, reproducible, and fine-grained evaluation methods. Recent studies have introduced motion-analysis-based assessment approaches using metrics such as path length, velocity profiles, acceleration, jerk, and bimanual symmetry [13, 15, 16]. These metrics offer quantitative and unbiased insights into technical performance.

However, most prior studies on motion analysis have been limited to simplified experimental environments—box trainers, VR simulators, or isolated wet-lab tasks [25, 23]. Consequently, it remains unclear whether quantitative motion metrics can reliably capture skill differences in near-clinical or full-procedure environments such as Thiel-embalmed cadaver surgery or real laparoscopic operations. Furthermore, no prior study has systematically examined whether motion-based features can distinguish ESSQ-certified from non-certified surgeons, an important question within the Japanese surgical education system.

These gaps motivate the present dissertation, which integrates optical motion capture, multi-rater GOALS scoring, ESSQ-based classification, and machine learning to establish a comprehensive, data-driven framework for objective laparoscopic skill evaluation.

1.5 Cadaver and Animal Models in Laparoscopic Education

1.5.1 Overview of High-Fidelity Training Environments

As laparoscopic surgery has become increasingly complex, there is growing demand for high-fidelity training environments that more closely reproduce real operative conditions than box trainers or virtual reality simulators alone. In particular, training for advanced procedures such as laparoscopic nephrectomy requires realistic anatomy, tissue handling, and intra-abdominal ergonomics that are difficult to replicate with simple task models. To address this need, many institutions have introduced cadaver-based and animal-based training programs as intermediate steps between simulation and live surgery [29, 30, 31, 32].

1.5.2 Cadaver-Based Training and the Thiel Embalming Method

Cadaver dissection has long played a central role in anatomical education and, in recent decades, has also been adopted for surgical skills training. Among various preservation techniques, the

Table 1.1: Comparison of major surgical skill assessment frameworks.

	OSATS [26]	GOALS [27]	GEARS [28]	ESSQ [6]
Country /organization	Canada / Royal College of Physicians and Sur- geons	Canada / McGill University group	USA / robotic surgery cen- ters	Japan / Japanese Society of Endourology
Target modality	Open and general surgery	Laparoscopic surgery	Robot-assisted surgery (da Vinci)	Advanced laparo- scopic procedures (urology)
Assessment type	global rating scale	Global rating scale (5 domains)	Global rating scale (robot-specific)	National video- based qualification system
Main strengths	Simple; widely applicable; structured basic eval- uation.	Tailored to MIS; validated for laparoscopic tasks.	Includes telemanipu- lation criteria; standardized.	Rigorous national credentialing; multi-expert review.
Main limitations	Subjective; limited sensitivity to motion efficiency.	Moderate reliability; limited for complex work- flows.	Subjective; domain variation; not for laparoscopy.	Single-case evalua- tion; no quantitative met- rics.

Thiel embalming method, first reported by Thiel in 1992, provides soft, flexible tissues with near-physiological joint mobility and realistic coloration [33]. These properties make Thiel-fixed cadavers particularly suitable for minimally invasive surgical training, where tactile feedback, organ consistency, and spatial relationships should approximate *in vivo* conditions.

Compared with conventional formalin-fixed cadavers, Thiel-embalmed cadavers offer several advantages for laparoscopic training:

- maintained tissue pliability, allowing realistic dissection planes and blunt separation;
- preserved joint mobility, facilitating positioning and port placement similar to clinical practice;
- relatively natural coloration, which supports visual recognition of vessels, organs, and fascial layers.

These characteristics enable near-clinical rehearsal of full procedures, including trocar insertion, exposure, dissection of major vessels, and specimen retrieval. For urologic procedures such as laparoscopic nephrectomy, Thiel cadavers provide an anatomically accurate environment in which surgeons can practice critical steps such as hilar dissection and control of renal vessels.

In Japan, cadaver-based surgical training was historically limited by legal and ethical constraints. The publication of the “Guidelines for cadaver dissection in education and research of clinical medicine” by the Japan Surgical Society and the Japanese Association of Anatomists [34], followed by the establishment of institutional frameworks, has gradually expanded the use of cadavers for clinical skills training. Kobayashi and Nudeshima reported that, compared with Western countries, Japan has only recently begun to establish systematic cadaver-based surgical training programs [31]. Nevertheless, Thiel-embalmed cadavers are now increasingly utilized for advanced laparoscopic and robotic training courses, particularly in tertiary centers and university hospitals.

Despite these advances, cadaver-based training has important limitations. Cadavers do not exhibit bleeding, hemodynamic changes, or physiological responses, and the absence of perfused circulation restricts the realism of hemostatic maneuvers. Access to cadaver facilities is also constrained by cost, infrastructure, and legal as well as ethical requirements. As a result, cadaver training is typically reserved for short courses or workshops rather than integrated into routine daily practice. Furthermore, most current cadaver programs rely on qualitative feedback from instructors instead of objective, quantitative assessment of technical performance.

1.5.3 Animal Organ Models and Wet-Lab Training

Animal organ models and in vivo animal surgery have also been widely used in laparoscopic education. Porcine models, in particular, closely mimic human abdominal anatomy in terms of organ size, vessel configuration, and tissue properties [29]. Wet-lab training using porcine organs allows trainees to practice tissue grasping, dissection, clipping, and suturing under realistic tactile and visual conditions, and provides an important bridge between dry-lab simulators and cadaver or clinical operations.

Higuchi *et al.* developed and validated a porcine organ model for essential laparoscopic skills training, demonstrating that structured wet-lab training can improve performance and support objective assessment [32]. Such models are particularly useful for repetitive practice, because they are more accessible and less costly than cadaver courses and can be integrated into regular educational curricula.

However, animal-based training raises ethical considerations and regulatory challenges [30]. The use of live animals is subject to strict institutional and national guidelines, and public opinion increasingly favors reduction and refinement of animal experiments. Even when ex vivo organs are used, issues of procurement, storage, and disposal must be addressed. In addition, anatomical differences between animals and humans—especially in vascular branching patterns and pelvic anatomy—limit the direct transferability of certain skills.

1.5.4 Need for Objective Assessment in High-Fidelity Environments

Cadaver-based and animal-based training environments provide higher fidelity than box trainers or purely virtual simulators and play complementary roles in the educational continuum. Cadaver models offer realistic human anatomy for full-procedure rehearsal, whereas porcine organ models support repeated practice of key technical components under near-real tissue conditions. Yet, in most programs, performance in these settings is still evaluated subjectively, often using global rating scales such as GOALS, checklists, or informal faculty feedback.

Previous work from our group has demonstrated that motion-analysis-based assessment in porcine organ training can objectively distinguish between different skill levels and track learning curves over time [15, 16, 32]. Nevertheless, it remains unclear whether similar objective metrics can be applied to more complex environments such as Thiel cadaver surgery, or whether such metrics can be linked to established benchmarks of clinical competence (e.g., ESSQ certification).

In this dissertation, Thiel-embalmed cadavers are used as a high-fidelity, near-clinical platform for laparoscopic nephrectomy, in combination with an optical motion capture system to record instrument trajectories throughout the procedure. Porcine organ wet-lab training data are also incorporated to examine the generalizability of motion-derived features across different training environments. All cadaver experiments were conducted under institutional ethical approval (Approval No. 20–027 and CERIA022–003), with appropriate consideration for donor consent and educational use. By integrating cadaver-based, animal-based, and clinical motion data, the present

work aims to bridge the gap between realistic surgical training environments and objective, data-driven assessment methods, thereby laying the groundwork for a comprehensive motion-analysis- and machine-learning-based skill evaluation framework with high interpretability.

1.5.5 Tele-Mentoring and Early-Stage Feasibility Evaluation

With the increasing complexity of laparoscopic and robotic surgery, tele-mentoring has gained attention as a means to support surgical education and reduce geographic disparities in clinical expertise. Tele-mentoring refers to remote experts observing the surgical field in real time and providing verbal guidance to the on-site surgeon, without direct manipulation of instruments. Unlike telesurgery, which requires fully robotic teleoperation, tele-mentoring focuses on educational support and clinical decision-making, and its feasibility depends heavily on the stability and latency of audiovisual transmission.

In this research project, tele-mentoring is conducted at an early technical-feasibility stage. Experienced surgeons at Hokkaido University remotely observe procedures performed at Obihiro Kosei Hospital through the *Zao Cloud View* system®(Soliton Systems), which enables simultaneous transmission of multiple video streams, including the endoscopic view, the robotic console view, and a room-side camera. Although real-time guidance is provided, the primary aim of the present study is not to evaluate instructional outcomes, but rather to quantify the performance of the underlying communication system.

To assess whether the tele-mentoring environment satisfies technical requirements for practical implementation, we conducted a series of network-performance measurements, including:

- **End-to-end latency measurements:** timestamping video frames at both sending and receiving sites to quantify total transmission delay;
- **In-network latency measurements:** evaluating baseline delay and jitter within the 5G and MEC (Multi-access Edge Computing) infrastructure using ICMP ping and related tools;
- **Multi-view synchronization assessment:** determining whether endoscopic, console, and room-side videos remain temporally aligned under varying network conditions.

These measurements aim to verify whether the system can provide acceptable latency, stable visual fidelity, and reliable multi-stream synchronization. The tele-mentoring work in this dissertation should therefore be regarded as an early-stage prototype validation of a remote surgical-support system, focusing on latency and transmission stability rather than educational or clinical effectiveness.

Establishing such a technical foundation is essential before integrating objective motion analysis and interpretable, machine-learning-based skill assessment into future tele-mentoring platforms. Importantly, no motion-capture data from tele-mentoring sessions were included in any statistical analysis, Principal Component Analysis (PCA), machine-learning classification, regression models, or SHapley Additive exPlanations (SHAP) interpretation in this thesis. The tele-mentoring component is presented solely as a technological feasibility study and not as part of the motion-analysis dataset.

1.6 Limitations of Motion-Analysis–Based Skill Assessment and the Need for Expanded Orientation Features

Motion analysis has emerged as a promising methodology for objective surgical skill evaluation in minimally invasive surgery. Previous studies have shown that kinematic metrics such as path length, velocity, acceleration, and smoothness can distinguish novice from expert surgeons [13, 15, 16, 25, 23]. Several reviews have summarized the development of objective, motion-based surgical skill assessment frameworks and associated metrics [20].

However, existing motion-based approaches still have important limitations. Most prior studies rely on simplified metrics such as path length and velocity, which capture global efficiency but fail to reflect fine psychomotor control, including rotational stability, wrist-level adjustments, and multi-degree-of-freedom coordination.

In addition, orientation-related features are typically represented using mean roll, pitch, and yaw values. Such representations smooth out cumulative rotational behavior and do not adequately capture instability or coupled motions, particularly the interaction between pitch and yaw in laparoscopic manipulation.

Furthermore, applying motion analysis in realistic surgical environments introduces practical challenges, including occlusion, limited camera placement, and reconstruction variability, which can affect data completeness and reliability.

To address these limitations, the present study introduces expanded orientation features derived from the sum of absolute angular displacements between consecutive frames. Two metrics were defined:

- **AL-Roll:** total accumulated rotation about the longitudinal axis;
- **AL-PitchYaw:** combined accumulated displacement in pitch and yaw, representing their natural coupling during laparoscopic manipulation.

These features enable more sensitive quantification of rotational behavior compared to conventional mean-angle representations. By capturing cumulative angular changes, they improve the descriptive capability of motion features for machine-learning–based classification and GOALS score prediction.

Thus, the proposed feature design addresses both conceptual limitations of prior motion descriptors and practical constraints encountered in realistic surgical settings.

1.7 Research Aims

Despite the growing interest in objective surgical skill assessment, current methods remain insufficient for evaluating performance in complex laparoscopic procedures. Most existing systems rely on simplified tasks, limited motion features, or subjective rating scales, making it difficult to assess fine psychomotor skills or to generalize across different training environments. Furthermore, the consistency of skill-related motion features across animal models, cadaveric procedures, and clinical settings has not been systematically explored.

The overarching goal of this dissertation is therefore

to establish a comprehensive framework for objective evaluation of laparoscopic surgical skill across multiple training environments, based on high-resolution motion analysis and interpretable machine-learning models.

The specific aims are as follows:

1. To develop and validate a motion measurement system suitable for complex laparoscopic procedures.

High-precision optical motion capture is applied to Thiel-embalmed cadaveric laparoscopic training. Improvements in measurement rate and three-dimensional reconstruction are evaluated using Motive 3.0.3 (OptiTrack, USA), an optical motion-capture software used for tracking rigid-body markers in three-dimensional space. The procedures are manually segmented into Phase 1–3, corresponding to major surgical phases (e.g., exposure, hilar dissection, and kidney mobilization) based on anatomical and procedural landmarks, as a basis for future AI-based automatic video segmentation.

2. To construct a high-resolution feature set for characterizing surgical motion.

Kinematic features are extracted, including position-, velocity-, acceleration-, and complexity-based metrics. Cumulative orientation features (AL-Roll and AL-PitchYaw) are proposed to overcome limitations of traditional angle-based metrics.

3. To analyze skill-group differences using statistical methods.

Kruskal–Wallis tests, principal component analysis (PCA), and related methods are used to identify features that contribute most to differences among experience levels.

4. To develop machine-learning models for skill classification and GOALS score prediction.

Classification models (SVM, PCA-SVM, Random Forest) and regression models (PCA-SVR, Ridge, PLS) are constructed. Nested cross-validation is used to evaluate model generalizability.

5. To establish an interpretable assessment system using SHAP.

SHAP is used to quantify feature contributions to the machine-learning models, providing an explanatory AI layer while keeping the underlying models transparent and feature-based.

6. To discuss the educational implications of motion analysis in surgical training.

Motion data obtained from wet-lab and cadaveric settings are used to illustrate possible applications for objective feedback and skill assessment. Tele-mentoring experiments conducted independently are mentioned only as potential directions for future development and are not analyzed within the scope of this thesis.

Through these aims, this dissertation seeks to provide a scientifically grounded, scalable, and clinically meaningful motion-analysis– and machine-learning–based framework for objective laparoscopic skill assessment, with an AI-supported explanatory layer, contributing to the digital transformation of surgical education.

1.8 Structure of This Dissertation

This dissertation consists of seven chapters, organized as follows:

- **Chapter 1: Introduction**

Reviews the background of laparoscopic surgical education, limitations of traditional skill assessment methods, high-fidelity training environments (cadavers and animal models), recent advances in motion analysis, and outlines the aims of this study.

- **Chapter 2: Measurement system and data acquisition**

Describes the optical motion capture system, marker design, calibration procedures, reconstruction improvements (Motive 3.0.3), and data acquisition in Thiel cadaveric nephrectomy and porcine wet-lab training. Ethical approval procedures are also presented.

- **Chapter 3: Feature extraction and statistical analysis**

Defines kinematic features, including positional, velocity-based, acceleration-based, and time-series complexity metrics. Cumulative orientation features (AL-Roll, AL-PitchYaw) are introduced. Statistical comparisons among skill levels are presented.

- **Chapter 4: Machine-learning-based skill evaluation**

Presents classification (SVM, PCA-SVM, Random Forest) and regression models (PCA-SVR, Ridge, PLS) for skill assessment. Model performance is evaluated with nested cross-validation. SHAP analysis is used for model interpretability.

- **Chapter 5: GOALS scoring system and its application**

Explains the GOALS structure, rating procedures, and inter-rater comparisons in this study. Describes how GOALS scores were adapted, evaluated, and validated in conjunction with motion metrics.

- **Chapter 6: Educational applications and tele-mentoring integration**

Demonstrates AI-supported feedback analysis in porcine wet-lab training and discusses model applicability across training stages. Describes the early-phase tele-mentoring system and network latency evaluation, exploring future integration with motion-analysis-based assessment.

- **Chapter 7: Conclusion**

Summarizes the research outcomes and discusses future directions for clinical application and the digital transformation of surgical education.

Chapter 2

Measurement system and data acquisition

2.1 Optical motion capture system

An optical motion capture system (OptiTrack V120:Trio®; NaturalPoint Inc., USA) was used to record the three-dimensional trajectories of laparoscopic instruments during Thiel-embalmed cadaveric nephrectomy and porcine wet-lab training. The system integrates three infrared cameras into a compact rigid frame, enabling marker-based tracking at 120 Hz with sub-millimeter precision within a calibrated measurement volume.

2.1.1 Hardware configuration

The V120:Trio (Fig. 2.1) consists of three global-shutter infrared cameras fixed within a single housing, ensuring geometric stability and eliminating inter-camera drift during long-duration surgical recordings. Each camera is equipped with integrated infrared illumination, allowing robust marker detection even under intense laparoscopic lighting. Real-time marker detection and three-dimensional reconstruction were performed using Motive (NaturalPoint Inc.).

The measurement volume was configured to cover the primary operative workspace: approximately $1.0\text{ m} \times 0.8\text{ m} \times 0.8\text{ m}$, which was sufficient to capture the full range of laparoscopic instrument motion in both cadaveric nephrectomy and porcine organ tasks.

2.1.2 Marker design for laparoscopic instruments

Reflective spherical markers (diameter: 6.4 mm) were attached to the proximal portion of the laparoscopic instruments. Because metallic instrument shafts produce strong specular reflections, markers were mounted on a matte-black attachment fixture to enhance contrast between unintended reflections and true infrared marker signals.

A rigid-body marker cluster composed of three non-collinear markers was designed to enable six-degrees-of-freedom (6-DOF) reconstruction of each instrument. The cluster's local coordinate frame was aligned such that its longitudinal axis matched the instrument shaft direction, facilitating the computation of orientation-based metrics such as roll, pitch, and yaw.

Marker placement was optimized to minimize interference with trocar insertion and to avoid collisions with tissues or surrounding equipment.



Figure 2.1: Overview of OptiTrack V120:Trio camera system [39]. The device integrates three infrared cameras into a rigid housing, enabling 120 Hz marker-based tracking in a compact surgical workspace.

Table 2.1: Specification of OptiTrack V120:Trio [39].

Parameter	Value	Unit
Frame rate	120	(fps)
Image resolution	640×480	
Latency	8.3	(ms)
Shutter type	Global	
Horizontal FOV	47	(°)
Vertical FOV	43	(°)
Minimum shutter speed	20	(μ s)

2.1.3 Calibration procedure and definition of the global coordinate system

Calibration was performed using the official OptiTrack three-dimensional calibration frame (rigid L-frame) (Fig. 2.2). The frame consists of two rigid orthogonal bars equipped with three reflective markers, providing a stable reference geometry without requiring wand-based calibration.

(1) Unified global coordinate system for all experiments To maintain consistency between cadaveric nephrectomy and porcine organ experiments, a unified global coordinate system was constructed. Based on the orientation of the L-frame, the axes were defined as follows:

- **Y axis (vertical axis):** perpendicular to the operating table, pointing upward.
- **XZ plane:** aligned with the cadaver or porcine organ working surface.
- **X axis (depth axis):** in-out depth direction of laparoscopic manipulation; used as a primary analytic direction across all datasets.
- **Z axis (in-plane reference axis):** orthogonal to the X axis within the XZ plane, used as an in-plane spatial reference.

Although the same geometric coordinate frame was applied to all recordings, the interpretation of the Z axis differed between experimental environments.

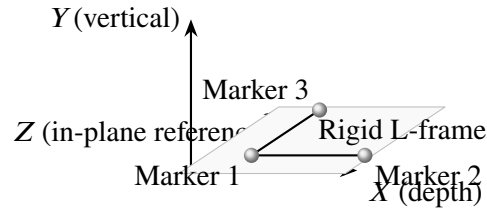


Figure 2.2: OptiTrack rigid L-frame used for calibration. The L-shaped frame is placed in the XZ plane, where the X axis represents the depth direction of laparoscopic manipulation, Y is the vertical axis, and Z serves as an in-plane reference axis (cadaver head–foot direction in the cadaver experiment).

(2) Interpretation in cadaveric nephrectomy During Thiel-embalmed cadaveric nephrectomy, the L-frame was aligned parallel to the cadaver body plane, allowing:

- the Z axis to represent the cadaver’s head–foot (longitudinal) direction.

This enabled anatomically meaningful interpretation, such as instrument motion toward the renal hilum.

(3) Interpretation in porcine wet-lab training

- the Z axis did not represent an anatomical direction and was used only as a spatial reference axis;
- kinematic analyses primarily used the X (depth) and Y (vertical) axes.

This ensured that datasets from different environments remained comparable without requiring additional coordinate transformations.

(4) Calibration steps in Motive Calibration was performed in Motive (version 2.1.1) using the following steps:

1. Place the L-frame parallel to the working surface.
2. Allow automatic detection of the three reflective markers.
3. Register the L-frame as the reference rigid body.
4. Establish the global coordinates based on the fixed geometry.
5. Verify alignment of the X , Y , and Z axes with the intended surgical directions.
6. Lock the coordinate system before data acquisition.

2.1.4 Measurement accuracy and sampling performance

The OptiTrack V120:Trio provides a nominal positional accuracy below 0.5 mm and an angular accuracy of approximately 0.5°. Under ideal optical conditions, the system is capable of stable rigid-body tracking at 120 Hz. However, in cadaveric laparoscopic training and porcine organ experiments, tracking performance can be affected by multiple factors, including instrument orientation, temporary occlusion by tissue or the operator’s hands, reflective surfaces, and variations in ambient lighting.

Consequently, although the nominal sampling frequency is fixed at 120 Hz, the effective measurement rate varied across measurement days and instruments, primarily due to occlusion and tracking loss. These were caused by factors such as the position of the laparoscopic camera, instrument orientation during manipulation, and occlusion by auxiliary instruments (e.g., assisting sticks), which occasionally prevented marker detection. In such cases, missing frames were supplemented by interpolation.

The measurement success rate was defined as the proportion of raw frames among total frames, where total frames consist of raw frames and interpolated frames.

$$\text{Measurement Success Rate} = \frac{N_{\text{raw}}}{N_{\text{total}}}$$

where N_{raw} denotes the number of frames obtained directly from Motive, and N_{total} denotes the total number of frames, including both raw and interpolated frames.

All trajectories were recorded at 120 Hz and exported from Motive. Short gaps and high-frequency noise were corrected using a third-order Savitzky–Golay filter (window length: 121 samples, corresponding to approximately 1.0 s). Frames requiring smoothing or interpolation were treated as interpolated frames.

Loss of raw frames primarily resulted from two factors: (1) occlusion, where instruments left the camera field of view or were temporarily covered by tissue or other instruments, and (2) orientation-dependent tracking failures, where rigid-body pose estimation became unstable due to unfavorable marker configurations.

A comprehensive summary indicates the following characteristics:

- For the primary working instruments, the reconstructed measurement rate remained high overall but differed slightly by instrument: 87.4–99.8% for the KARL STORZ forceps, 81.2–100.0% for the LigaSure device, and 90.4–99.8% for the OLYMPUS forceps.
- For auxiliary tools (SUCTION and HEM-O-LOK), the reconstructed measurement rate showed greater variability, occasionally dropping to 70–80% due to their peripheral positioning or short usage duration.
- Similar fluctuations were observed for an additional instrument with external markers (Right-angle Forceps), especially in early sessions, where the reconstructed measurement rate occasionally fell below 50%.

These reconstructed values were used for feature extraction (Chapter 3) and machine-learning analysis (Chapter 4), whereas raw measurement rates were used only for preliminary inspection.

To improve the stability of subsequent feature extraction and statistical analysis, all raw motion data were reprocessed using Motive 3.0.3. Compared with the acquisition version (Motive 2.1.1), Motive 3.0.3 provides enhanced algorithms in three major components:

- **Reconstruct:** geometric completion of missing three-dimensional marker positions using multi-camera constraints;
- **Auto-label:** automatic recovery of marker labels across long recordings, reducing the need for manual correction;
- **Solve:** rigid-body pose estimation with improved robustness, yielding smoother six-DOF trajectories during short tracking interruptions.

After reconstruction, the measurement rate improved for most instruments, particularly during deep dissection phases and complex multi-instrument interactions where occlusion is common. Although this study does not treat measurement rate improvement as a quantitative outcome, visual inspection and preliminary comparisons indicated that reconstructed trajectories exhibited substantially better continuity and pose stability. These improvements provide a more reliable data foundation for the feature extraction presented in Chapter 3 and the machine-learning analyses in Chapter 4.

2.1.5 Limitations of optical tracking in surgical environments

As discussed in Chapter 1, optical tracking faces multiple constraints in high-fidelity surgical settings:

- occlusion by anatomical structures, trocars, hands, or instruments;
- specular reflections from metallic shafts under laparoscopic lighting;
- marker displacement caused by tissue contact or trocar friction;
- limited space for multi-camera placement in cadaver rooms or wet labs.

These considerations guided the development of more robust motion features (discussed in Chapter 3) that tolerate transient dropout and reconstruction variability.

2.2 Experimental environment for cadaveric laparoscopic training

All cadaveric laparoscopic nephrectomy procedures were performed at the Cadaveric Surgical Training Laboratory, Hokkaido University Hospital, Sapporo, Japan, using human cadavers preserved with the Thiel embalming method. Thiel-preserved cadavers offer soft-tissue elasticity, color tone, and joint flexibility that closely resemble those of living human tissue, making them suitable for realistic laparoscopic training and motion analysis. The use of cadaveric specimens in this study was approved by the Ethics Committee of Hokkaido University Hospital (Approval No. 20–027 and CERIA022–003).

2.2.1 Experimental setup and operating room configuration

Each training was performed at the cadaveric surgical training laboratory, Hokkaido University Hospital, Sapporo, Japan. A standard laparoscopic tower (camera system, light source, insufflator, and electrosurgical generator) was set up next to the operating table. The operating field, instrument ports, and camera ports were determined following standard clinical nephrectomy procedures.

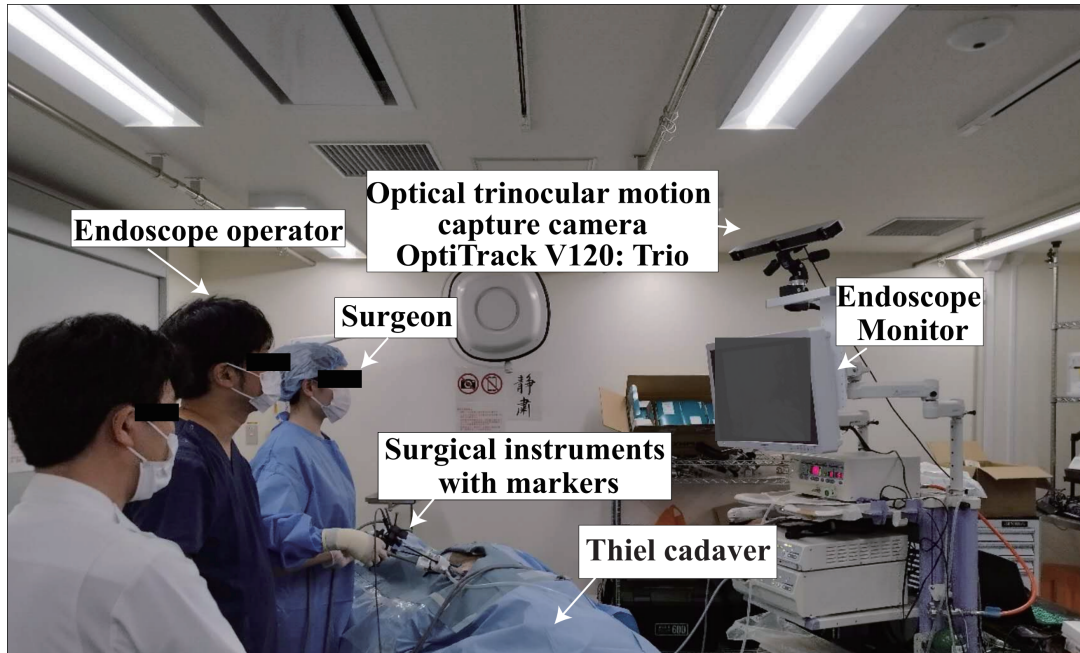


Figure 2.3: Overview of the cadaveric laparoscopic training environment. The OptiTrack V120:Trio motion capture camera is mounted above the operating field, tracking laparoscopic instruments equipped with reflective markers. The setup includes the surgeon, endoscope operator, endoscopic monitor, and a Thiel-embalmed cadaver.

The OptiTrack V120:Trio was installed above the operating field as described in Chapter 2, with its tracking volume covering the entire region where laparoscopic instruments were manipulated. The coordinate system established during calibration (X: depth direction, Y: vertical direction, Z: cadaver longitudinal axis) was kept consistent across all experimental days.

2.2.2 Surgical task: laparoscopic nephrectomy

The cadaveric laparoscopic nephrectomy procedures followed the standardized steps:

1. Initial camera port placement using the Hasson technique followed by creation of pneumoperitoneum.
2. Placement of three additional ports.
3. Mobilization of the ascending colon and duodenum (right-sided cases) or the descending colon, spleen, and pancreas (left-sided cases).
4. Dissection of the plane between the psoas muscle and the ureter/lower pole of the kidney.
5. Elevation of the kidney and identification of the renal hilum.
6. Identification and skeletonization of the renal artery and vein.
7. Clipping with Hem-o-Lok and division of the renal artery and vein.
8. Mobilization of the upper pole of the kidney.

9. Dissection of the remaining attachments except the ureter.

These steps involve complex bimanual instrument manipulation, tissue traction, energy-device activation, and precise dissection—all of which produce rich kinematic signatures suitable for quantitative motion analysis.

2.2.3 Participant profiles

A total of 39 participants performed 51 training sessions. Following the exclusion of 5 cases due to video recording failure ($n=3$), insufficient embalming condition ($n=1$), and in which the mentor completed most procedure ($n=1$), 46 Mocap datasets from the 36 participants were available for the current analysis. Participants were categorized into three experience groups based on their prior laparoscopic operative volume:

- (N) Novices (0–9 operations): 17 subjects
- (I) Intermediates (10–49 operations): 19 subjects
- (E) Experts (≥ 50 operations): 10 subjects

Although the number of individual participants was 36, several surgeons completed the CST more than once. Specifically, four novice surgeons and two expert surgeons participated twice, and two additional novice surgeons participated three times. These repeated participations resulted in a total of 46 recorded laparoscopic nephrectomy procedures.

This multi-level structure enabled cross-sectional comparison of motion patterns and served as the basis for skill classification and regression tasks presented in Chapters 3 and 4.

2.2.4 Instruments tracked in cadaver experiments

Across all 46 recorded nephrectomy procedures, motion data were consistently captured from six laparoscopic instruments, each equipped with a rigid-body marker cluster. These instruments correspond to those shown in Fig. 2.4. Among these instruments, the following 4 instruments were utilized in the feature extraction, PCA, and machine-learning analyses.

- **Grasping forceps (GF):** a left-hand assisting instrument used for traction, counter-traction, and maintaining exposure (commercially, typically a KARL-STORZ laparoscopic forceps).
- **Scissor forceps (SF):** a right-hand dominant working instrument used mainly for dissection and tissue manipulation (e.g., an OLYMPUS laparoscopic forceps with a cutting tip).
- **Energy device (ED):** a right-hand LigaSure device used for vessel sealing and tissue division during hilar dissection.
- **Clip applier (CA):** a HEM-O-LOK applicator used to apply polymer clips to blood vessels and other structures requiring secure ligation.

Because the two additional instruments described below appeared only in selected procedures, they were excluded in the present analysis,:

- **Suction tube (ST):** a suction instrument intermittently used for smoke evacuation and fluid removal.

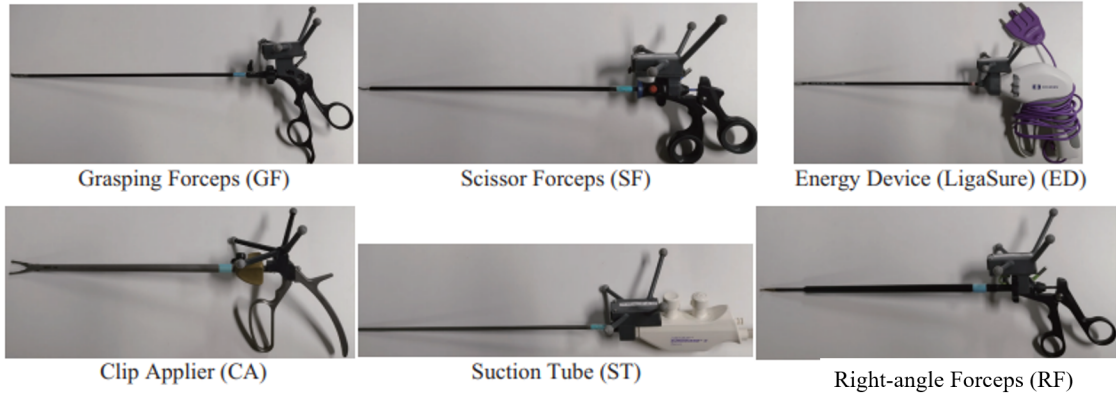


Figure 2.4: Surgical instruments used in the cadaveric laparoscopic training. From left to right and top to bottom: Grasping forceps (GF), Scissor forceps (SF), Energy device (LigaSure, ED), Clip applier (CA), Suction tube (ST), and Right-angle forceps (RF). These abbreviations (GF, SF, ED, CA, ST, RF) are used consistently in the feature extraction and analysis throughout this dissertation.

- **Right-angle forceps (RF):** a curved, right-angle dissecting forceps with external reflective markers, used for blunt dissection in specific cases.

The abbreviations GF, SF, ED, CA, ST, and RF are used consistently throughout this dissertation.

2.2.5 Data recording and reconstruction protocol

All cadaveric nephrectomy procedures were recorded continuously from initial trocar insertion to freeing the kidney completely from surrounding tissues other than uretermobilization. Three categories of data streams were collected in synchrony throughout each session:

- 3D motion data (120 Hz) recorded with Motive 2.1.1, selected for its stable real-time acquisition during long-duration procedures;
- Laparoscopic video (HD), recorded concurrently with the motion-capture data;
- Overhead room-side video, showing hand motion, trocar position, and the overall operative environment.

Two-stage data processing workflow. Although Motive 2.1.1 provides stable real-time capture, its reconstruction robustness is limited in deep-cavity surgical environments. Therefore all raw TAK/CSV data were exported and reprocessed offline using Motive 3.0.3. The offline pipeline included:

- re-running the updated Reconstruct algorithm to recover missing marker positions,
- applying the improved Auto-label routine with refined rigid-body templates,
- executing Solve to recover rigid-body orientation during short periods of partial marker visibility,

- applying a third-order Savitzky–Golay filter (window length 121) to correct short gaps and suppress high-frequency noise.

This two-stage workflow substantially improved the effective measurement success rate and reduced marker dropouts caused by occlusion or orientation-dependent visibility loss. The resulting trajectories form a high-quality basis for feature extraction, PCA, classification, regression, and SHAP analysis in Chapters 3 and 4.

All datasets were managed using anonymized participant identifiers and date-based codes (YYYYMMDD). A total of 46 complete laparoscopic nephrectomy procedures were recorded, forming one of the largest optical-motion-based datasets for near-clinical laparoscopic training reported to date.

2.2.6 Post-training questionnaire and surgeon grouping

Immediately after each cadaveric training session, all participating surgeons completed a structured, web-based questionnaire (SurveyMonkey, Momentive Inc., San Mateo, CA, USA). The questionnaire collected background information and evaluated participants’ subjective impressions of both the cadaveric training course and the motion-measurement system.

The first section collected demographic and professional information, including dominant hand, sex, age, year of graduation, years of urologic experience, and certification status (*urology specialist, urology training instructor*). Participants also reported whether they held certification in the *Endoscopic Surgical Skill Qualification System* (ESSQ) for laparoscopic urology. Case-volume information was collected, including the number of laparoscopic procedures performed as primary surgeon (all procedures) and the total number of laparoscopic nephrectomies.

The second section of the questionnaire focused on educational aspects of the course. Participants rated their understanding of perirenal and retroperitoneal membrane anatomy before and after the training, and evaluated the similarity of Thiel cadavers to living tissue (e.g., port insertion, color and consistency of vessels and organs). Several items were rated on a 5-point Likert scale (1: “strongly disagree” to 5: “strongly agree”).

Overall procedural workload was evaluated using the NASA Task Load Index (NASA-TLX) [40], which provides sub-scores for mental demand, physical demand, temporal demand, performance, effort, and frustration, as well as an overall composite workload score.

A specific item assessed whether reflective markers interfered with forceps manipulation. Surgeons rated the statement:

“The markers attached to the laparoscopic instruments disturbed my manipulation of the forceps.”

Responses were analyzed descriptively to confirm that marker attachment did not significantly hinder surgical handling, thereby supporting the ecological validity of the recorded motion data.

2.3 Real surgery: Motion data acquisition in a clinical operating room

To evaluate the feasibility of optical motion capture in a clinical setting, instrument-trajectory data were recorded during laparoscopic procedures performed in the Department of Urology at Hokkaido University Hospital. Unlike the cadaveric and porcine experiments described earlier,

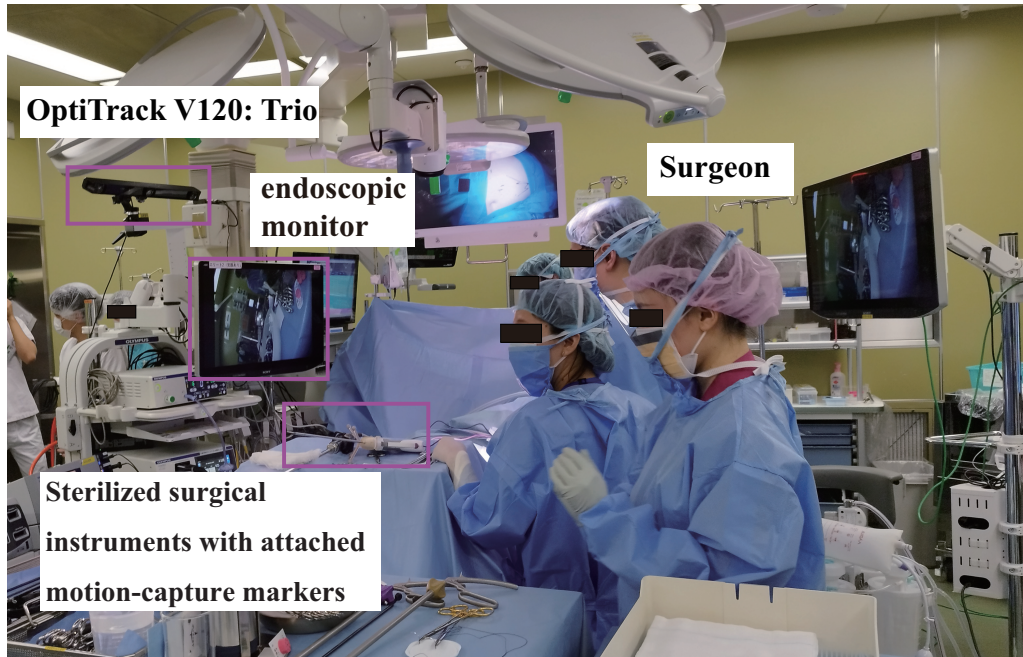


Figure 2.5: Setup of the motion capture system in real clinical laparoscopic nephrectomy at Hokkaido University Hospital.

these recordings were obtained under genuine clinical workflows, requiring strict consideration of sterilization compatibility, marker stability, and non-interference with surgical manipulation.

This clinical motion-capture study was conducted under a separate institutional review board (IRB) approval: Protocol No. 024-0363, Hokkaido University Hospital. All procedures for device sterilization, intraoperative deployment, and handling of clinical metadata were carried out in accordance with this approval.

2.3.1 Marker-attached laparoscopic instruments

Four commonly used laparoscopic instruments were equipped with sterilizable, autoclavable rigid-body marker clusters:

- **Grasping Forceps (GF):** used for traction, counter-traction, and maintaining exposure of the operative field.
- **Scissor Forceps (SF):** used for soft-tissue dissection and fine cutting.
- **Energy Device (LigaSure, ED):** used for vessel sealing and energy-based tissue division.
- **Clip Applier (CA):** used for vascular clipping. A sterilizable marker-mounting structure specifically designed for this project was prototyped and validated prior to deployment.

All marker fixtures were fabricated from sterilization-safe materials and were autoclaved or gas-sterilized before each operation. Mounting positions were carefully chosen to avoid interference with the surgeon's grip, ergonomic handling, or trocar insertion angles.

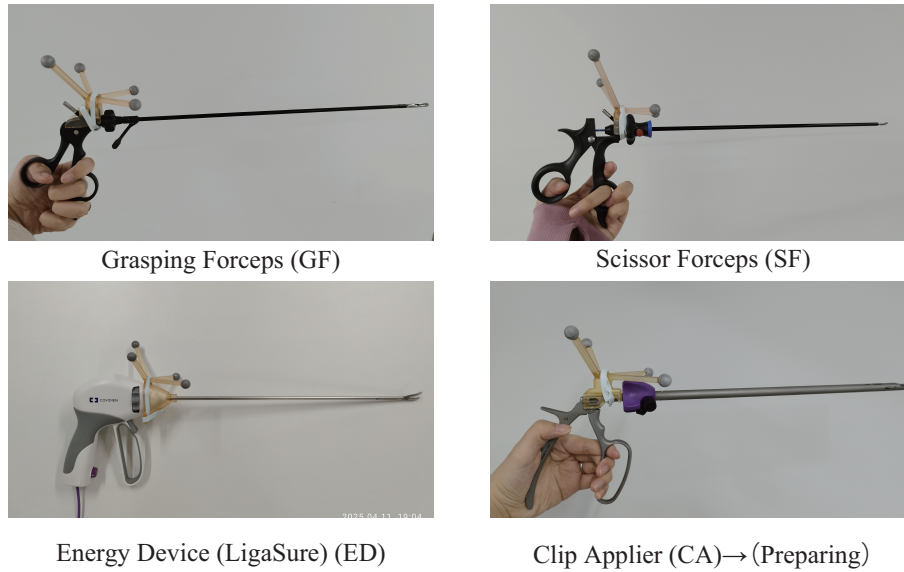


Figure 2.6: Sterilizable marker-attached laparoscopic instruments used for real surgical motion capture: Grasping Forceps (GF), Scissor Forceps (SF), Energy Device (LigaSure, ED), and Clip Applier (CA). All instruments were equipped with rigid-body marker clusters designed to withstand sterilization and avoid interference with intraoperative manipulation.

2.3.2 Operating-room configuration

An OptiTrack V120:Trio system was installed adjacent to the endoscopic monitor to maximize visibility of the instrument shafts and handles while maintaining full compliance with the sterile drape layout. Compared with controlled laboratory or cadaveric setups, the clinical environment introduced several additional sources of occlusion:

- temporary blockage by the surgeon’s or assistant’s hands,
- out-of-plane rotations that caused all markers to disappear simultaneously.

Despite these challenges, multiple high-quality motion segments suitable for feature extraction and cross-environment comparison were successfully obtained.

2.3.3 Role of real-surgery data

The clinical recordings collected in this section were used to:

- assess the practical feasibility of optical motion capture in an active clinical workflow;
- compare the noise characteristics, occlusion patterns, and trajectory smoothness with those observed in cadaveric and porcine environments;
- perform preliminary checks of feature robustness and cross-environment consistency in later statistical analyses.

2.4 Wet-Lab (Porcine Organ) Experiments

In addition to the Thiel cadaveric nephrectomy experiments, a series of wet-lab training sessions using porcine organs were conducted to verify the robustness and reproducibility of the motion-capture workflow under simplified experimental conditions. Compared with the cadaveric environment, wet-lab tasks involve fewer occlusions, shorter task durations, and more constrained instrument trajectories, providing an ideal platform for confirming marker visibility, testing rigid-body configurations, and validating the consistency of feature extraction.

The wet-lab dataset was not intended for full-scale skill assessment or machine-learning analysis. Instead, its role in this dissertation is threefold: (1) to confirm that the recording and export procedures using OptiTrack V120:Trio function reliably in a low-occlusion environment; (2) to evaluate the stability of feature computations across different instrument types, including needle-type forceps used only in wet-lab tasks; and (3) to support cross-environment comparisons of motion characteristics at a descriptive level.

The following subsections describe the wet-lab setup, task definitions, instrument configurations, and motion-capture procedure.

2.4.1 Experimental Setup

The porcine wet-lab experiments were conducted at the Simulation Center of Hokkaido University Hospital, using a custom box-trainer platform integrated with an optical motion capture system and standard laparoscopic instruments (Fig. 2.7). The goal of this setup was to reproduce the ergonomic constraints and visual conditions of real laparoscopic surgery while enabling controlled, repeatable tissue-handling tasks.

A full laparoscopic tower was used, consisting of an HD endoscopic camera, light source, insufflator, and monitor. Participants performed all tasks in a standing position, following the posture typically adopted in clinical laparoscopic procedures. The box trainer was placed on a height-adjustable table to standardize ergonomic conditions across participants.

For motion capture, an OptiTrack V120 Trio system (three-camera IR unit) was mounted on a dedicated tripod approximately 1.5–2.0 m in front of the box trainer. The optical axis was aligned toward the trocar plane to maximize the visible working volume and ensure consistent marker visibility during instrument manipulation. All laparoscopic instruments were equipped with custom rigid-body marker clusters identical to those used in the Thiel cadaver experiments, enabling consistent kinematic feature extraction across environments.

The box trainer employed a three-port configuration (Fig. 2.8), consisting of three 12 mm ports: one central port for the endoscopic camera and two lateral ports positioned at clinically relevant angles for instrument insertion. Inside the trainer, porcine organs were secured to a stainless-steel grid using elastic wires, providing a stable platform for standardized dissection tasks.

This experimental setup provided a high-fidelity wet-lab environment that captured essential psychomotor characteristics of laparoscopic surgery, while supporting precise motion capture for the quantitative analyses presented in subsequent chapters.

2.4.2 Task Design and Experimental Workflow

All porcine wet-lab experiments were conducted in the Simulation Center of Hokkaido University Hospital. Participants performed the tasks in a standing clinical posture, using a standard laparoscopic tower and porcine abdominal organs placed in a dedicated box trainer. The experimental



Figure 2.7: Wet-lab experimental setup used at the Simulation Center of Hokkaido University Hospital. Participants performed the tasks in a standing position while using an HD laparoscopic tower and a box trainer equipped with porcine organs. The OptiTrack V120 Trio optical motion capture system was positioned in front of the trainer to record 3D instrument trajectories throughout each task.



Figure 2.8: Three-port configuration of the box trainer used in the porcine experiments. All ports were 12 mm in diameter, consisting of one central port for the endoscopic camera and two lateral ports positioned at clinically relevant angles for instrument insertion.

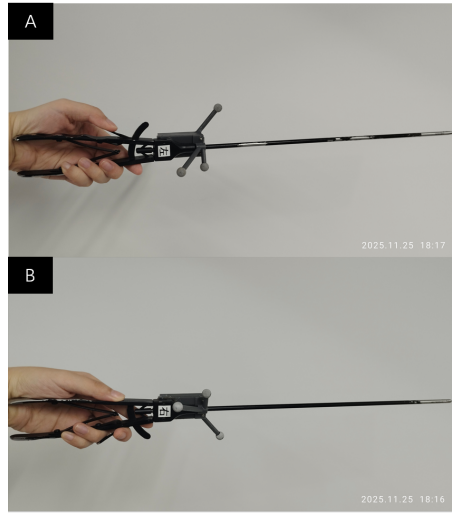


Figure 2.9: Rigid-body markers attached to the needle holders used in Task 2. (A) Needle holder used in the left (assisting) hand. (B) Needle holder used in the right (dominant) hand. These instruments were newly introduced only for the suturing task in the porcine wet-lab experiment and differ from the grasping and dissecting forceps used in the cadaveric nephrectomy training. Each needle holder was equipped with a four-marker rigid-body cluster near the handle, allowing 6-DoF motion tracking with the OptiTrack V120 Trio system under the same coordinate definition as in the cadaver experiments.

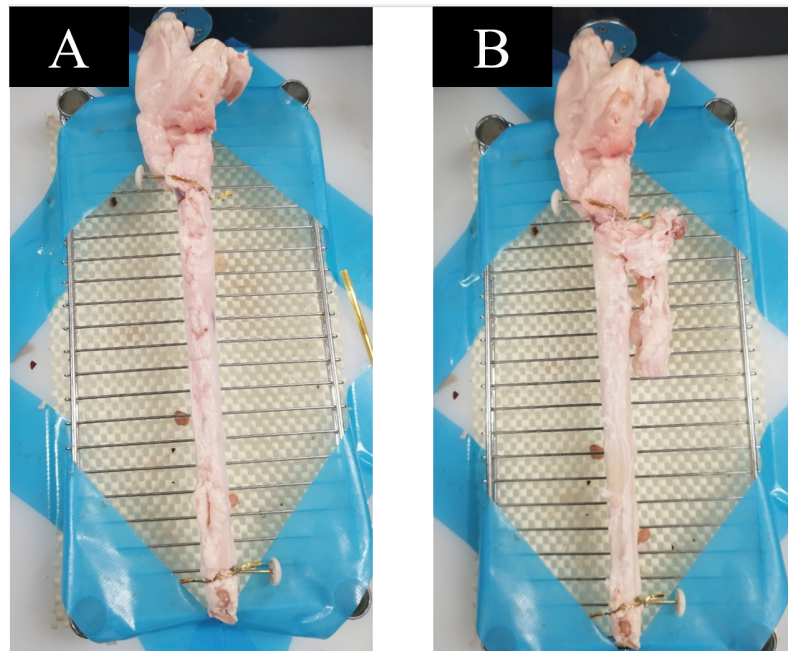


Figure 2.10: Porcine organ fixation within the box trainer. The tissues were secured to a stainless-steel grid using elastic wires, providing a stable platform for standardized dissection tasks.

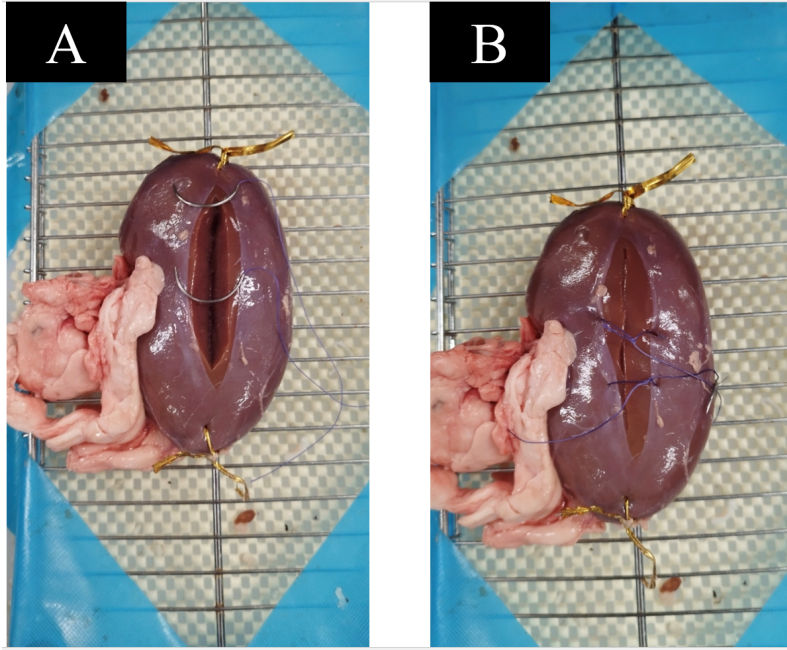


Figure 2.11: Laparoscopic suturing task using porcine organs in the wet-lab environment (Task 2). Needle driving and intracorporeal knot tying were performed using marker-attached needle holders. Motion data were recorded to obtain instrument tip position and orientation during suturing.

protocol was designed to reproduce key psychomotor components of laparoscopic surgery while maintaining stable motion capture using the OptiTrack V120 Trio system.

Overall task composition. Each participant completed two predefined tasks representing fundamental psychomotor components of laparoscopic dissection and suturing:

- **Task 1: Dissection task.** This task involved soft-tissue dissection and pedicle exposure using three standard laparoscopic instruments: grasping forceps, scissor forceps, and a vascular clip applicator (Hem-o-lok type). The grasping forceps were used for traction and counter-traction, the scissors for tissue division, and the clip applicator for placing polymer locking clips. The tip position and orientation were calculated from the attached rigid-body markers.
- **Task 2: Suturing task** This task involved needle handling, needle driving, and intracorporeal knot tying using a pair of laparoscopic needle holders (right and left). A single marker cluster was attached to the proximal sheath of each instrument, allowing computation of tip motion without independent sheath rotation or gripper-opening measurements. The metallic surfaces of the needle holders were coated with matte black paint to prevent infrared reflection and ensure reliable marker detection.

Experimental workflow. All participants followed an identical workflow:

1. **Synchronized start** of V120 Trio motion capture and laparoscopic video.

2. **Task execution** following standardized instructions to ensure equivalent experimental conditions.
3. **Motion data acquisition** at 120 Hz with simultaneous HD video.
4. **Environment reset** before the next participant.

Rationale of the two-task design. Task 1 provides motion data containing traction, dissection, and instrument rotation typical of laparoscopic tissue manipulation, while Task 2 captures fine bimanual control and complex rotational movements essential for suturing. Together, these tasks represent a wide spectrum of laparoscopic psychomotor skills and complement the cadaveric nephrectomy dataset described in Section 2.2.

Representative actions of Task 1 and Task 2 are shown in Figs. 2.10 and 2.11, and the needle holders used in Task 2 are illustrated in Fig. 2.9.

2.4.3 Motion Capture Procedure

Motion data for both tasks were recorded at 120 Hz using the OptiTrack V120:Trio optical motion capture system. Because the wet-lab environment involved fewer occlusions and simpler tool trajectories than Thiel cadaver nephrectomy, a single-camera-set configuration provided stable and reliable tracking throughout the sessions.

Wet-lab motion datasets were used exclusively for:

- verifying system operation and marker visibility,
- checking the consistency of motion-feature computations,
- validating the instrument-specific marker rigs.

Importantly, wet-lab motion data were not included in the statistical analysis, PCA, machine-learning classification, regression, or SHAP interpretation presented in Chapters 3 and 4, due to differences in task structure and workflow.

2.5 Data Synchronization and Management

Accurate synchronization among multimodal data streams is essential for integrating motion-capture trajectories with corresponding operative visual information. In this study, synchronization procedures were applied for the cadaveric nephrectomy experiments, where three independent recording sources were used: (1) optical motion capture trajectories, (2) laparoscopic endoscopic video, and (3) an external room-side video camera.

To avoid redundancy, synchronization details are described only once here.

2.5.1 Synchronization procedure in cadaveric experiments

Because the OptiTrack V120:Trio system, the laparoscopic imaging tower, and the external video camera do not share a common hardware clock, synchronization was performed using a manual start protocol. At the beginning of each nephrectomy session, the operator sequentially initiated recording on all systems:

- Motive software (OptiTrack) for motion-capture logging (120 Hz),

- the laparoscopic camera recording console,
- the external room-side video camera.

Post hoc temporal alignment was performed based on shared initial timestamps and identifiable procedural events across all modalities.

2.5.2 Synchronization in Real Laparoscopic Surgery

Real laparoscopic nephrectomy performed at Hokkaido University Hospital used two data sources: the V120:Trio motion-capture system and the laparoscopic endoscopic video.

Temporal alignment was performed using identifiable procedural events visible in both modalities (e.g., instrument insertion and major dissection phases).

This approach enabled reliable post-hoc synchronization despite the absence of hardware triggering.

Because the clinical environment involved relatively stable endoscopic imaging conditions, temporal alignment across datasets was considered sufficiently robust for subsequent analysis.

2.5.3 Synchronization procedure in wet-lab experiments

Wet-lab porcine experiments employed only two data sources: the motion-capture system and the laparoscopic endoscopic video. As in the cadaveric setting, the operator manually initiated both recordings in synchrony. The absence of a room-side camera simplified post-hoc alignment. Moreover, wet-lab tasks were shorter in duration and contained fewer occlusions, making temporal matching highly stable and reliable.

2.5.4 File structure and dataset standardization

To support reproducibility, all data streams were organized with a consistent folder structure:

- Mocap/ – Raw and reconstructed CSV files exported from Motive.
- Videos/Endoscope/ – Laparoscopic recordings for each session.
- Videos/External/ – External camera videos (cadaver experiments only).
- Notes/ – Operator notes on occlusion events, camera changes, and anatomical difficulties.

Each session was assigned a unique anonymized identifier (e.g., 20230112_CadaverL), shared across motion and video files. This ensured that motion trajectories, videos, and metadata could be reliably cross-referenced during feature extraction and later analyses.

2.5.5 Rationale and implications for analysis

Although manual synchronization is less precise than hardware triggering, it provides sufficient alignment for the objectives of this dissertation. The machine-learning analyses presented in Chapters 3 and 4 rely solely on motion-capture trajectories, which preserve high temporal fidelity at 120 Hz. Videos were used only for qualitative review, GOALS rating, and annotation of instrument usage phases. As such, sub-second synchronization accuracy was adequate and did not affect statistical analysis, PCA, classification, or regression performance.

This synchronization framework also supports future extensions such as automated video–motion alignment, workflow segmentation, or integration into remote tele-mentoring platforms described in Chapter 6.

2.6 Tele-mentoring and Remote Surgical Support System

In addition to the on-site cadaveric and porcine experiments, this study also contributed to a collaborative project on tele-mentoring and remote surgical assistance between Hokkaido University Hospital (HUH) and Obihiro Kousei Hospital.

This component is presented as a supplementary technical study and is not included in the motion-analysis dataset or machine-learning framework of this dissertation.

2.6.1 System architecture

A bidirectional communication framework was established using high-bandwidth network infrastructure supported by a Multi-access Edge Computing (MEC) platform. The system consisted of:

- **HUH (mentor side):** senior laparoscopic surgeons provided real-time feedback based on the endoscopic video streamed from the remote site.
- **Obihiro Kousei Hospital (operator side):** local surgeons performed laparoscopic or robotic procedures while receiving live guidance.
- **Endoscopic video transmission:** high-definition laparoscopic video was encoded and transmitted through the MEC-based low-latency network.
- **External monitoring cameras:** multiple viewpoints were used to confirm operative field, instrument movement, and surgeon positioning.
- **Delay-monitoring subsystem:** round-trip end-to-end latency was measured using synchronized LED–sensor triggers and timestamp logging, following the protocol described in the MEC technical documents.

2.6.2 Data collected in the tele-surgery project

No motion-capture data were collected during the tele-mentoring sessions in this study.

Instead, the contribution of this study to the collaborative project focused on:

- recording and analyzing end-to-end latency under various network conditions;
- monitoring MEC and 5G network stability during procedures;
- confirming multi-camera synchronization at both sites;
- providing technical support, including device restarts and troubleshooting during transient connection failures.

These activities are presented to demonstrate the feasibility of integrating motion-analysis frameworks with remote surgical-support systems, rather than as part of the core analytical dataset.



Figure 2.12: Remote mentoring workstation at Hokkaido University (mentor side). The instructor monitors multiple synchronized video streams (endoscopic view, room-side view, and console/robotic view) during the tele-mentoring session with Obihiro Kosei Hospital.

2.6.3 Relevance to this thesis

Although the remote surgery experiments do not contribute numerical motion-feature data, they provide contextual validation for the future integration of quantitative surgical skill assessment into tele-mentoring environments.

In particular, the latency data support future development of real-time feedback or predictive modeling based on motion patterns under constrained network conditions.

2.7 Ethical Approval

This dissertation includes studies conducted under four independent ethics approvals issued by Hokkaido University Hospital. These approvals cover cadaveric surgical training, porcine wet-lab experiments, real-surgery motion capture, and telementoring studies.

2.7.1 Cadaveric laparoscopic motion capture

Motion-capture measurements during cadaveric laparoscopic nephrectomy training were initially approved under protocol Med20–027 (2020/12/14). A revised and expanded version of this protocol was subsequently approved as CERIA022–003 (2022/08/17–2027/03/31) as a comprehensive review. This protocol did not require additional submission to the Faculty of Medicine ethics committee.

2.7.2 Porcine wet-lab experiments

Porcine laparoscopic training experiments were conducted under protocol 018–0257, approved on 2019/03/12. All experimental procedures complied with institutional guidelines for the use of biological specimens in surgical training.

2.7.3 Real-surgery motion measurement

Motion-capture data acquired during clinical laparoscopic surgery were approved under protocol 024–0363 (2024/11/27). This approval permitted the acquisition of instrument trajectories together with associated clinical metadata stored in the hospital’s secure database. Patient identifiers remained within the hospital system and were not exported to the research environment; all motion data used for analysis were fully anonymized in accordance with institutional data-protection policies.

2.7.4 Telementoring studies

Exploratory telementoring experiments were conducted under protocol 023–0006 (2023/04/06). These studies are referenced in this dissertation only to outline future applications of surgical motion analysis and were operationally separate from the motion-capture experiments.

2.7.5 Data protection and anonymization

All protocols complied with institutional and national guidelines for research involving human subjects, animal specimens, and cadaveric materials. Clinical identifiers were retained exclusively within hospital servers and were not accessible during analysis. The datasets used in this dissertation contained no patient-identifiable information or recognizable images.

2.8 Summary

This chapter described the measurement systems and data-acquisition environments used throughout this study. Optical motion capture was applied across cadaveric nephrectomy, real laparoscopic surgery, and porcine wet-lab tasks, with consistent marker configurations and standardized export formats. Each experimental environment required specific adaptations in camera placement, instrument preparation, and recording workflow.

In the cadaveric and real-surgery settings, motion capture was combined with laparoscopic endoscopic video, whereas the porcine wet-lab experiments used motion capture alone due to their simplified setup. For all datasets, instrument types, marker arrangements, and data-management procedures were kept uniform to ensure compatibility with the feature-extraction pipeline presented in Chapter 3.

Ethical considerations for cadaveric, surgical, and wet-lab experiments were addressed according to institutional and national guidelines. The resulting datasets form the basis for the quantitative motion analysis and machine-learning investigations in subsequent chapters.

Chapter 3

Feature Extraction and Statistical Analysis

3.1 Overview of the Feature Extraction Pipeline

This chapter focuses on the extraction of quantitative kinematic features from motion-capture data after preprocessing.

All motion-capture recordings were first exported from Motive as comma-separated value (CSV) files. Each rigid body—representing the left-hand instrument, right-hand instrument, and any additional tools—contained synchronized position and orientation streams at 120 Hz. The subsequent processing pipeline consisted of four major stages:

1. **Structural reconstruction and missing-frame correction:** For datasets recorded using Motive 2.1.1, rigid-body trajectories were reconstructed to restore frames lost due to occlusion or marker misidentification.
2. **Temporal smoothing and interpolation:** Position signals were smoothed using a Savitzky–Golay filter (third-order polynomial, 121-sample window). Missing samples were interpolated using linear interpolation, and the measurement rate was computed as the proportion of frames that did not require smoothing or interpolation.
3. **Segmentation into surgical phases:** Cadaveric nephrectomy procedures were manually divided into three procedural phases (Phase 1–3), allowing phase-specific feature analysis.
4. **Feature computation:** Kinematic features were extracted from the smoothed trajectories. These included linear and angular velocity, acceleration, jerk, motion-intensity distributions (idle/low/middle/high/very-high), path length (PL), direction-changed path length (DPL), directional variability (DV), working-area metrics, and angular metrics such as instrument roll and pitch–yaw changes. All computations were performed independently for each rigid body.

This standardized pipeline ensured that all datasets—regardless of differences in surgical environment, instrument type, or recording system—were processed under uniform criteria. The resulting features form the basis for the statistical analyses, PCA visualizations, and classification/regression models described in the subsequent sections of this chapter.

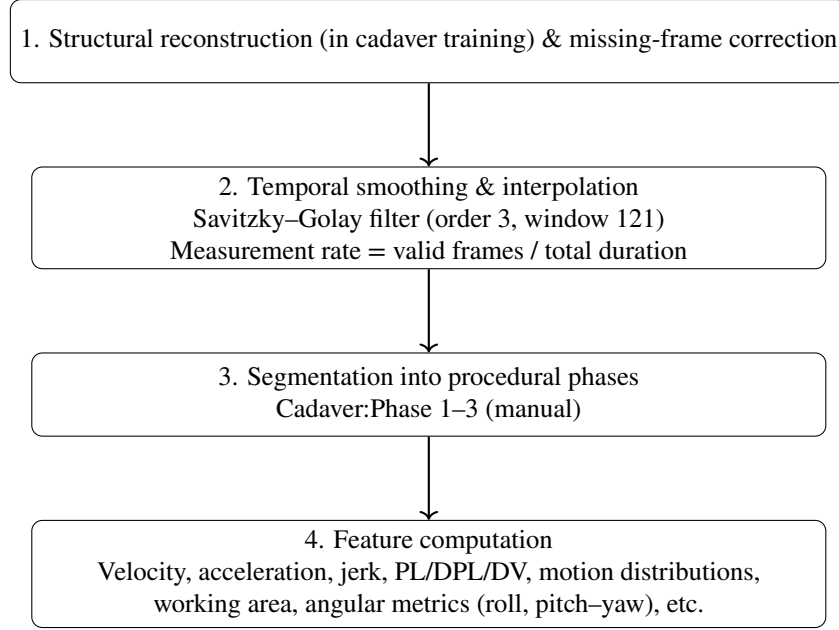


Figure 3.1: Overview of the motion-capture feature extraction pipeline. All datasets (cadaveric, real surgery, and wet-lab) were processed using the same standardized procedure.

3.2 Feature Extraction

A comprehensive set of kinematic features was extracted from the 120 Hz optical motion-capture data recorded during cadaveric laparoscopic nephrectomy. All trajectories were reconstructed and smoothed using a Savitzky–Golay filter (order 3, window size 121). For each instrument, 3D position and orientation data were transformed into a local coordinate system in which the x -axis corresponds to the depth direction (i.e., the surgical port axis). In all analyses, the local coordinate system of each instrument was defined based on the rigid-body orientation described in Chapter 2, where the longitudinal axis of the instrument is aligned with the local x -axis.

3.2.1 Temporal and Spatial Workspace Features

Operative Time (OT). OT represents the total duration required to complete the task:

$$OT = N \times \Delta t, \quad \Delta t = 1/120 \text{ s.}$$

Here, N is the number of valid motion-capture frames. OT is expressed in seconds.

Using Time (UT). UT quantifies the duration during which an instrument is actively used. A frame is classified as “active use” if the instrument tip is located within the predefined ROI and exhibits motion above a small movement threshold (or if the activation signal is ON for energy/clip devices):

$$UT = N_{\text{active}} \times \Delta t.$$

3.2.2 Path-Based Metrics

Path Length (PL). The total 3D distance traveled by the instrument tip:

$$PL = \sum_{i=1}^{N-1} \|\mathbf{p}_{i+1} - \mathbf{p}_i\|.$$

Trajectory segments outside the predefined region of interest (ROI) were excluded.

Depth Path Length (DPL). DPL quantifies motion along its sheath axis:

$$DPL = \sum_{i=1}^{N-1} |x_{i+1} - x_i|.$$

Depth Velocity (DV). The average velocity along its sheath axis:

$$DV = \frac{1}{T} \sum_{i=1}^{N-1} |x_{i+1} - x_i|.$$

3.2.3 Speed-, Acceleration-, and Jerk-Based Metrics

Mean velocity.

$$v_i = \frac{\|\mathbf{p}_{i+1} - \mathbf{p}_i\|}{\Delta t}.$$

Mean acceleration.

$$a_i = \frac{v_{i+1} - v_i}{\Delta t}.$$

Mean jerk.

$$j_i = \frac{a_{i+1} - a_i}{\Delta t}.$$

3.2.4 Velocity Distribution Metrics

Velocities were categorized into five ranges (cm/s):

- Idle: 0–0.5
- Low: 0.5–2.0
- Middle: 2.0–5.0
- High: 5.0–12.0
- Very high: ≥ 12

For each range:

$$\text{Ratio}_k = \frac{N_k}{N_{\text{total}}}.$$

3.2.5 Orientation Metrics

AL-Roll. Sum of absolute frame-to-frame roll-axis rotations. Both AL-Roll and AL-PitchYaw are expressed in degrees.

AL-PitchYaw. Combined pitch and yaw cumulative rotation:

$$\text{AL-PitchYaw} = \sum_i (|\Delta\theta_{\text{pitch},i}| + |\Delta\theta_{\text{yaw},i}|).$$

Working Area (WA). The spatial spread of instrument-tip motion in the plane perpendicular to the axis (X-Z plane) is computed using percentile ranges:

$$\text{WA} = (x_{0.975} - x_{0.025}) (z_{0.975} - z_{0.025}),$$

where $x_{0.975}$ and $x_{0.025}$ denote the 97.5th and 2.5th percentiles of the X-coordinate distribution, and similarly for z . WA is expressed in cm^2 . In the cadaveric experiments, the Z-axis corresponds to the head-foot (anatomical longitudinal) direction, whereas in the porcine environment it serves as a non-anatomical in-plane reference axis.

3.2.6 Bimanual Coordination Metrics

BD (Bimanual Dexterity).

$$\text{BD} = \frac{\text{Mean}(v_{\text{left}})}{\text{Mean}(v_{\text{right}})}.$$

Values greater than 1 indicate left-hand-dominant motion, whereas values less than 1 reflect right-hand dominance.

RPLB (Ratio of Path Length for Both Hands).

$$\text{RPLB} = \frac{\text{PL}_{\text{left}}}{\text{PL}_{\text{right,scissor}} + \text{PL}_{\text{right,energy}}}.$$

Values closer to 1 represent more balanced bimanual workloads between the assisting and working instruments.

Additional ratios. Left-right path-length ratios (e.g., PL_L/PL_R) were computed to assess bimanual workload balance.

3.2.7 Summary

The resulting feature set includes PL, DPL, DV, mean velocity, acceleration, jerk, five-stage velocity distributions, cumulative orientation metrics (AL-Roll, AL-PitchYaw), and bimanual indices (BD, RPLB, and left-right ratios), providing a multidimensional description of efficiency, smoothness, rotational control, and bilateral coordination.

3.3 Feature Pre-processing

Prior to statistical analysis and machine-learning modeling, all extracted kinematic features underwent a standardized pre-processing workflow to ensure comparability across procedures and participants.

3.3.1 Handling missing values

Missing frames arising from occlusion or reconstruction loss were identified. Short gaps (< 0.10 s) were linearly interpolated, whereas longer gaps remained missing and were excluded from feature computation.

3.3.2 Temporal smoothing

All trajectories were smoothed using a third-order Savitzky–Golay filter (window length: 121 samples). Velocity and acceleration were computed using central finite differences; jerk was obtained from filtered acceleration.

3.3.3 Robust normalization

Because many kinematic features exhibit non-Gaussian distributions and may contain outliers originating from reconstruction imperfections, all features were standardized using a robust Z-score (RZS). The RZS is defined as:

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

where x_i is the data value, x_m is the median, and $NIQR$ is the normalized interquartile range:

$$NIQR = 0.7414 \times IQR, \quad IQR = Q_3 - Q_1.$$

This normalization is less sensitive to extreme values than mean–standard deviation scaling and is therefore well suited to surgical motion data. Normalization was performed separately for each instrument type.

3.3.4 Outlier handling

Outliers beyond $Q1 - 3IQR$ or $Q3 + 3IQR$ were winsorized to the boundary limits. No trials were discarded.

3.3.5 Statistical feature filtering

Kruskal–Wallis tests (three groups) or Mann–Whitney U tests (two groups) were applied. Only features with $p < 0.05$ entered PCA or machine-learning models.

3.3.6 Summary

The complete workflow consisted of: reconstruction, smoothing, derivative calculation, interpolation, winsorization, normalization, and statistical filtering.

3.4 Statistical Analysis

Statistical analyses were conducted to identify motion features that showed significant differences among surgeon groups prior to dimensionality reduction and machine-learning classification. All analyses were performed exclusively on the cadaveric nephrectomy dataset.

3.4.1 Group definitions

Two grouping schemes were used:

- **Experience-based groups:** novice (0–9 surgeries), intermediate (10–49), expert (≥ 50).
- **Certification-based groups:** ESSQ-certified vs. non-certified.

3.4.2 Univariate statistical testing

Normality was assessed with the Shapiro–Wilk test. As most features were non-Gaussian, non-parametric tests were applied:

- Mann–Whitney U test for two-group comparisons (ESSQ vs. non-ESSQ);
- Kruskal–Wallis test for three-group comparisons, followed by pairwise Mann–Whitney U tests when appropriate.

A significance threshold of $p < 0.05$ was used. Multiple-comparison correction was not applied because this step served only as an initial feature-screening procedure prior to dimensionality reduction and cross-validated machine-learning analyses, which provide additional regularization against overfitting. Features failing to reach significance in either grouping were excluded from subsequent PCA and machine-learning analyses.

3.5 Data Quality and Descriptive Visualization

Before conducting feature extraction and machine-learning analyses, the quality and completeness of the cadaveric motion-capture data were evaluated. Descriptive visualizations were also generated to verify the consistency of instrument usage and to qualitatively inspect characteristic differences among instruments. These analyses ensure the reliability of the dataset and provide empirical support for the feature definitions introduced in Section 3.2.

3.5.1 Measurement rate of motion capture

The measurement rate was defined as the ratio between the number of valid motion-capture frames and the theoretical number of frames expected at 120 Hz. A frame was considered valid when all reflective markers forming a rigid body were simultaneously detected and successfully reconstructed by Motive. Incomplete frames caused by temporary occlusion, marker overlap, or instrument rotation leading to partial marker visibility were automatically discarded.

Table 3.1 summarizes the distribution of measurement rates for the six surgical instruments used in the cadaveric nephrectomy dataset. Overall, the motion-capture system demonstrated high reliability across all instruments, with median measurement rates exceeding 92% and several instruments—particularly the Clip Applier (CA), Right-angle Forceps (RF), and Suction Tube (ST)—achieving values above 98%. These results indicate that occlusion events were limited and that rigid-body recognition was stable throughout most procedures.

To ensure data quality for downstream analyses, trials with extremely low measurement rates were reviewed. All analyzed trials (46 procedures) provided sufficiently dense trajectories for feature extraction, task-level quantification, and machine-learning analyses presented in subsequent chapters.

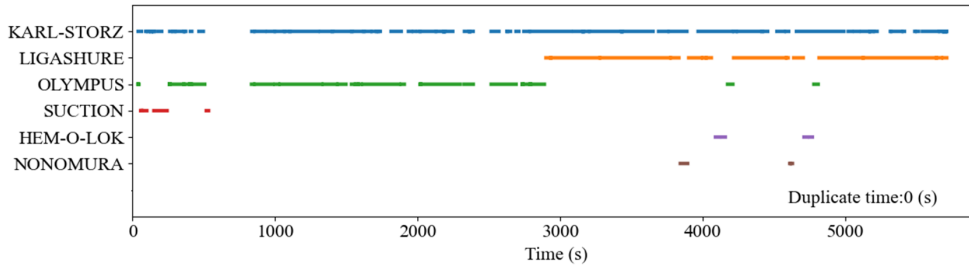


Figure 3.2: Representative data of instrument usage timeline during the cadaveric nephrectomy session. Each horizontal bar indicates an active period of a specific instrument, illustrating the overall workflow and tool transitions. Brief and clearly identifiable misclassification segments were manually removed following verification with synchronized video.

Table 3.1: Measurement rate of surgical instruments. Values are expressed as median (1st quartile–3rd quartile).

Surgical instrument	Measurement rate (%)
Grasping Forceps (GF)	92.74 (89.88–95.75)
Scissor Forceps (SF)	93.72 (88.91–96.36)
Clip Applier (CA)	98.87 (93.61–99.90)
Energy Device (ED)	94.22 (88.65–97.60)
Right-angle Forceps (RF)	98.40 (96.47–99.97)
Suction Tube (ST)	98.79 (96.48–99.79)

High measurement rates also confirm that smoothing and interpolation procedures—applied using a third-order Savitzky–Golay filter (window size: 121 frames)—contributed minimally to the effective trajectory length, preserving the temporal fidelity required for velocity-, acceleration-, and jerk-based kinematic features.

3.5.2 Summary of screened feature set

The remaining features after significance testing and correlation filtering formed the final input set for PCA, visualization, and classification models. These features and their importance are reported in Chapter 4.

3.5.3 Definition of the Region of Interest (ROI)

The Region of Interest (ROI) used for kinematic analysis was defined based on the spatial relationship between the cadaver body, surgical ports, and anatomical markers. Two definition methods were adopted depending on whether the port positions were recorded.

Figure 3.4(a) illustrates the standard ROI definition, which was used when the coordinates of all surgical ports were available. The ROI was modeled as a rectangular box aligned with the floor (X–Z plane) and oriented parallel to the anatomical reference line connecting the shoulder and knee (red dashed line). Its mediolateral width was set to $0.4l$, where l represents the shoulder–knee distance. The craniocaudal height and anteroposterior depth were fixed at 300 mm. The cranial

boundary of the ROI was positioned 50 mm below the level of the anterior superior iliac spine, while the caudal boundary was placed at 0.05 l distal to the most caudally located surgical port.

In six cases where port positions were not recorded, the alternative definition shown in Figure 3.4(b) was applied. The overall dimensions of the ROI (width, height, depth) remained unchanged, but the anchoring reference was shifted to anatomical markers. Specifically, the cranial boundary was defined as 0.15 l from the shoulder marker along the shoulder–knee line, ensuring a consistent anatomical frame of reference across all cadaver sessions.

This dual-definition strategy allowed consistent ROI construction while accounting for variability in recorded port information, enabling reliable comparison of instrument trajectories across all cases.

3.5.4 Visualization of instrument usage timeline

To validate the temporal annotation and verify consistency between the motion-capture data and intraoperative endoscopic video, instrument-usage timelines were created. Each colored bar indicates the period during which a specific instrument was actively used, as determined by synchronized endoscopic-video review.

During this process, brief and clearly identifiable misclassification segments (e.g., caused by instrument overlap or tracking errors) were manually removed based on synchronized video verification. These corrections were limited and treated as adjustments of obvious misrecognition rather than missing data.

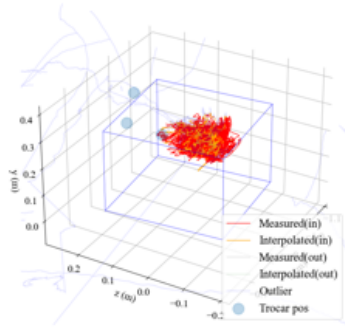
An example from a representative data is shown in Figure 3.2. Although these timelines were not directly used as input for machine-learning models, they served as an important diagnostic tool to confirm segmentation consistency and to ensure that each instrument’s active phase corresponded to its actual intraoperative usage.

3.5.5 Three-dimensional trajectory visualization

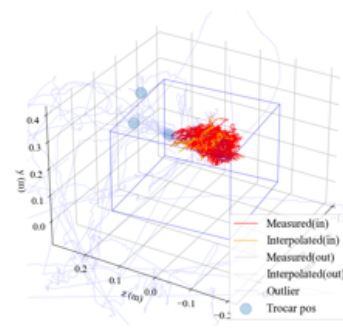
To further illustrate spatial characteristics of surgical manipulation, three-dimensional trajectories were visualized for all six cadaveric instruments: Grasping Forceps (GF), Scissor Forceps (SF), Energy Device (ED), Clip Applier (CA), Right-angle Forceps (RF), and Suction Tube (ST). Each trajectory plot displays the measured (red) and interpolated (yellow) segments, together with the trocar positions and the port-box boundary used during the cadaveric nephrectomy tasks.

A combined six-panel visualization is presented in Figure 3.3. These plots reveal clear differences in working area, depth motion, and manipulation strategy among instruments. For example, graspers and scissors exhibit compact motion volumes, whereas the Energy Device and Suction Tube occupy larger depth ranges. These qualitative patterns match the quantitative features introduced in Section 3.2 (e.g., working area, DPL, DV) and help validate their design.

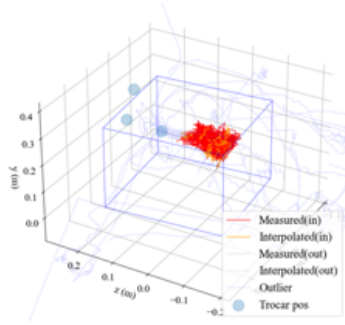
Overall, the descriptive visualizations in this section provide qualitative confirmation of data integrity and operational context. They support the validity of subsequent feature-based and machine-learning analyses in Chapters 4 and 5.



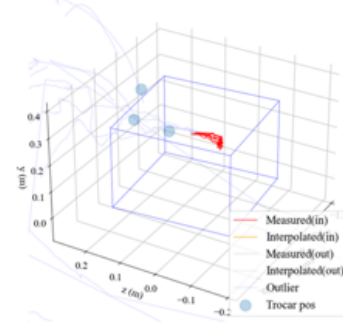
Grasping Forceps



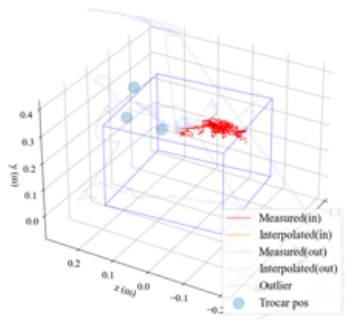
Scissor Forceps



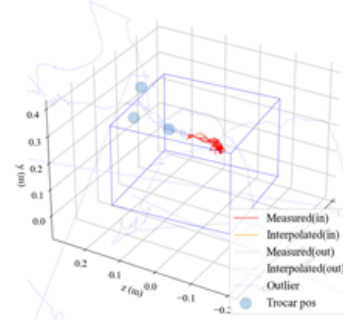
Energy Device



Clip Applier



Suction Tube



Depart right-angle Forceps

Figure 3.3: Representative data of three-dimensional reconstructed trajectories of six instruments during the cadaveric nephrectomy. Each subplot visualizes the cleaned, interpolated, and outlier-removed motion of one instrument, with trocar positions provided as spatial reference.

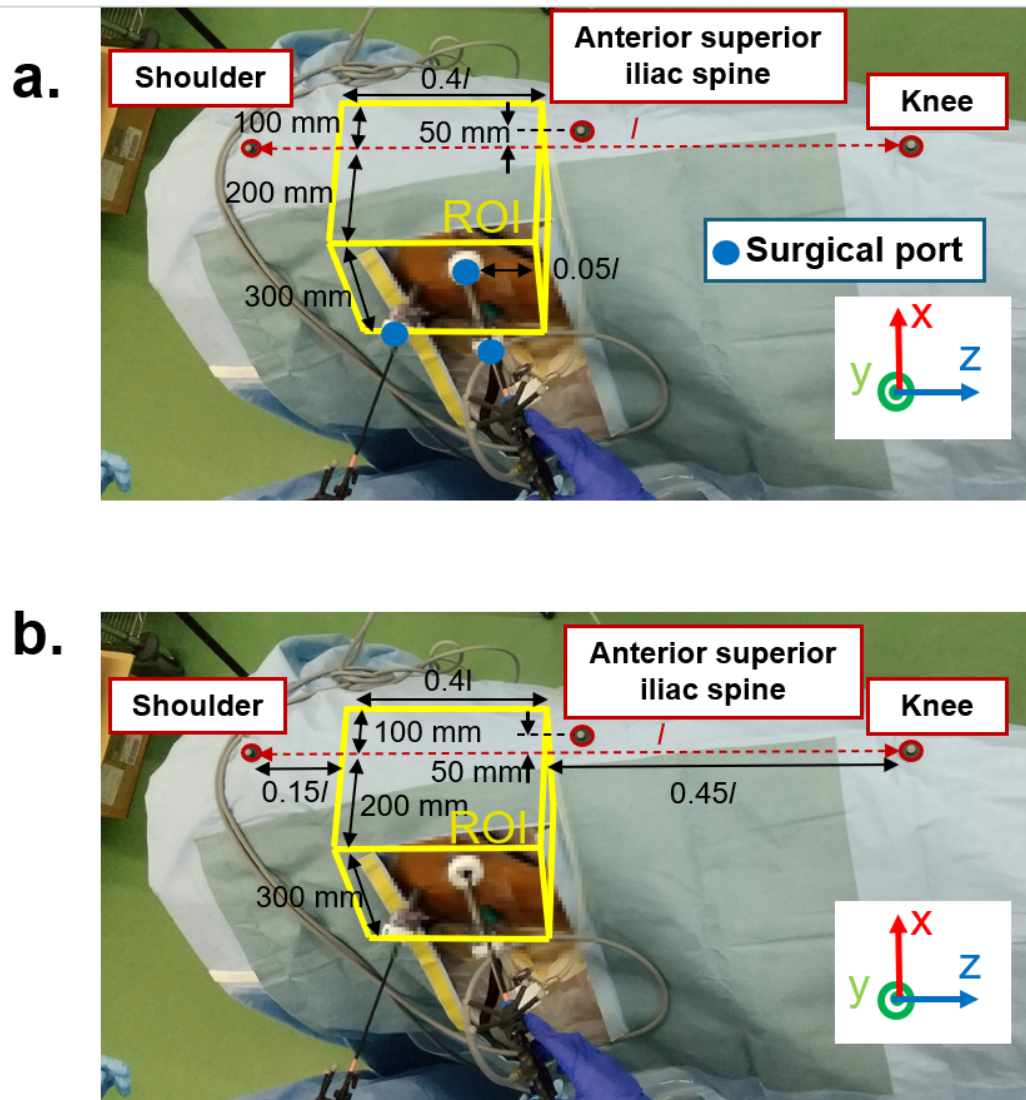


Figure 3.4: Overview of ROI definition methods. (a) ROI definition based on recorded port coordinates. (b) ROI definition based on anatomical markers when port coordinates were unavailable.

3.6 Skill-related feature analysis

To explore how fundamental kinematic features reflect surgical proficiency, statistical comparisons were performed across three experience groups: novice, intermediate, and expert. Significant differences were observed in both efficiency-related features (e.g., path length of the grasping forceps, scissor forceps, and energy device) and velocity-related features across experience groups (Kruskal–Wallis test, $p < 0.05$). These results were consistent with the trends previously reported in our porcine organ training model, confirming a similar relationship between surgical proficiency and kinematic indices.

3.6.1 Participants' backgrounds

A total of 36 surgeons participated in the cadaveric surgical training (CST) program, classified into three experience groups based on their prior laparoscopic operative volume: 17 novices (0–9 cases), 19 intermediates (10–49 cases), and 10 experts (≥ 50 cases). Several surgeons completed the CST more than once: four novices and two experts participated twice, and two novices participated three times. These repeated participations resulted in a total of 46 analyzed laparoscopic nephrectomy procedures.

Demographic information. The median age was 32 years (range: 27–52). The cohort consisted of 30 males and 6 females. Regarding handedness, 35 participants were right-handed and 1 was left-handed.

Laparoscopic experience. Participants exhibited a wide range of laparoscopic experience, categorized by the number of previously performed laparoscopic surgeries:

- 0–9 procedures: $n = 9$
- 10–49 procedures: $n = 19$
- 50–99 procedures: $n = 1$
- 100–499 procedures: $n = 7$

Formal technical qualification. Seven surgeons held the Japanese Endoscopic Surgical Skill Qualification (ESSQ), whereas twenty-nine did not. ESSQ-certified surgeons were treated as validated experts in subsequent analyses.

Summary. Overall, the participant pool covered the full spectrum of laparoscopic proficiency—from novice to highly experienced surgeons—providing a representative basis for evaluating motion-related skill indices, principal-component structures, and machine-learning classification models in later sections.

3.6.2 Comparison of representative kinematic features

As shown in Figure 3.5, clear skill-dependent trends were observed. Experts completed the nephrectomy tasks significantly faster than novices and intermediates ($p < 0.01$). The proportion of low-velocity operation time (DVL) was highest among novices, indicating slower and less continuous

manipulation. Conversely, experts demonstrated a smaller DVL ratio, reflecting smoother forceps handling and better temporal continuity.

The total path length (PL) of the grasping forceps exhibited a similar pattern: experts showed significantly shorter and more efficient instrument trajectories ($p < 0.01$). This result suggests that experienced surgeons maintain a smaller and more controlled workspace, while novices often perform excessive or redundant movements.

3.6.3 Summary and implications

These results confirm that temporal and spatial kinematic indices effectively reflect technical proficiency in laparoscopic manipulation. Specifically:

- Operating time increases markedly in novice surgeons.
- Low-velocity ratio (DVL) decreases with experience, indicating improved smoothness and control.
- Path length (PL) shortens with experience, representing enhanced motion efficiency.

Such trends align with findings from previous motion-analysis studies, and provide empirical support for the kinematic feature set used in subsequent PCA and machine-learning evaluations.

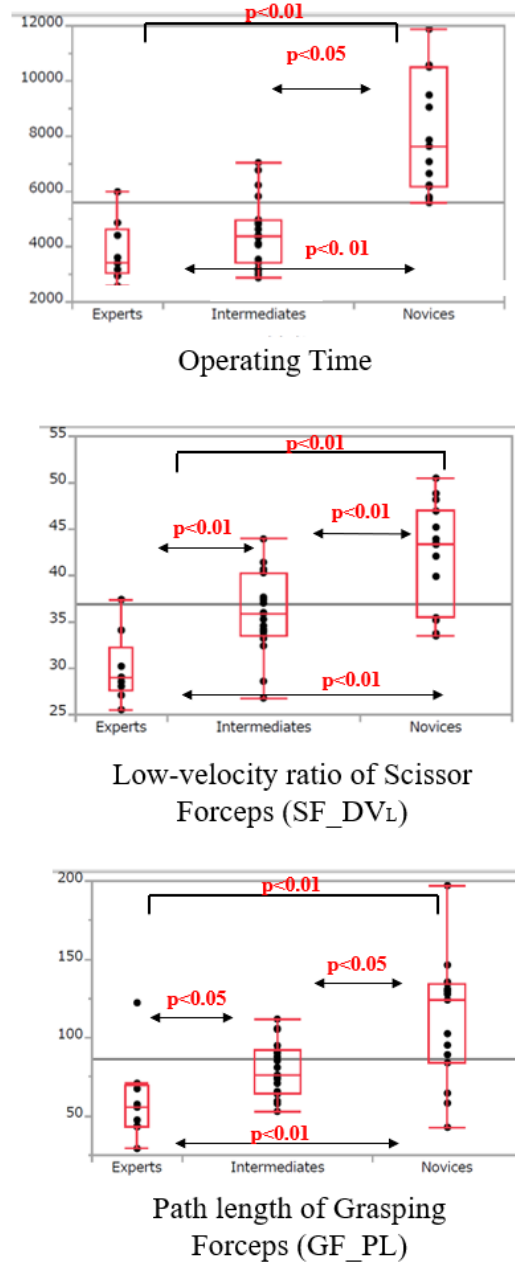


Figure 3.5: Boxplots of key kinematic skill indices across three experience levels. (Left) Operation time (OT); (Center) Low-velocity ratio of scissor forces (DVL); (Right) Path length of grasping forces (PL). Statistical differences were evaluated using Kruskal–Wallis test.

3.7 Surgical Phase Segmentation

Laparoscopic nephrectomy consists of multiple heterogeneous subtasks that require different psychomotor skills, visual–motor coordination, and instrument-handling patterns. Evaluating surgical performance using the entire procedure as a single undivided unit is therefore inappropriate, as it mixes fundamentally different actions into one metric. To obtain phase-specific and skill-sensitive indicators, the procedure was manually segmented into three phases based on the intraoperative workflow.

3.7.1 Definition of the three phases

The segmentation was performed by an experienced surgeon using synchronized endoscopic videos only.

For both left- and right-sided nephrectomy, the following three phases were defined:

- **Phase 1: Colon mobilization**
 - Left nephrectomy: mobilization of the descending colon with cranial retraction of the spleen.
 - Right nephrectomy: mobilization of the ascending colon and Kocherization of the duodenum.

This phase involves bowel mobilization and exposure of the operative field.

- **Phase 2: Renal vessel control** Identification and dissection of the renal artery and vein, followed by placement of vascular clips. Because cadaveric tissues do not bleed, hemostasis is not part of this phase.
- **Phase 3: Resection of residual perirenal tissue** culminating in complete mobilization of the kidney.

3.7.2 Purpose of phase-based skill evaluation

The present study evaluates technical skill in each of the three phases individually. phase-based segmentation enables clearer identification of which motion features—such as path length, velocity profiles, and depth motion—reflect proficiency within each subtask. This provides a more accurate and granular assessment of surgical performance compared with whole-procedure analysis.

3.8 Summary

In this chapter, we established a complete kinematic-processing framework for analyzing cadaveric laparoscopic nephrectomy procedures, beginning with motion data preprocessing and extending through feature extraction, data quality assessment, and preliminary skill-related evaluations. The workflow was designed to ensure that the resulting kinematic indices were both reliable and physiologically interpretable, enabling downstream analyses in subsequent chapters.

A series of preprocessing steps were implemented to correct for measurement noise, marker occlusion, and missing data inherent to optical motion capture. Robust smoothing, interpolation, and outlier removal ensured stable reconstruction of the instrument-tip trajectories. The Region of Interest (ROI) was then defined to standardize the spatial domain of analysis across all cadaver sessions. Because cadaver postures, port placements, and available reference markers varied between

procedures, two ROI-definition strategies were developed: one based on recorded port positions, and another based on anatomical landmarks such as the shoulder and anterior superior iliac spine. This dual-definition approach ensured geometric consistency of the workspace and enabled fair comparison of kinematic variables across trials.

Data quality was evaluated using measurement rate (the ratio of valid frames to the theoretical 120 Hz sampling count). All major instruments demonstrated high measurement rates, confirming that the reconstructed trajectories were sufficiently complete for quantitative analysis.

Using the cleaned trajectories, we extracted a comprehensive set of kinematic features that characterize temporal efficiency, velocity distribution, spatial economy of movement, and tool usage patterns. Skill-related statistical comparisons revealed clear differences among novice, intermediate, and expert surgeons. Experts completed the nephrectomy tasks with shorter operative time, smaller low-velocity ratios, and reduced path lengths, indicating smoother, more economical instrument control. These feature-based trends were consistent with previous motion-analysis studies and reproduced the tendencies observed in wet-lab porcine training experiments, supporting the generalizability of the proposed kinematic framework.

The findings of this chapter demonstrate that the constructed preprocessing pipeline, ROI standardization, and kinematic feature set provide a robust foundation for quantitative assessment of surgical skill. These validated features form the basis for the dimensionality reduction and machine-learning methods presented in Chapters 4 and 5. In particular, the PCA analysis in the next chapter leverages the feature space established here to clarify the dominant movement patterns associated with surgical expertise, while the classification models further evaluate the discriminative power of these kinematic indices. Overall, Chapter 3 establishes the methodological backbone for the subsequent analytical components of the dissertation.

Chapter 4

Machine-Learning Based Surgical Skill Classification

Building upon the kinematic feature extraction and descriptive analyses presented in Chapter 3, this chapter introduces the machine-learning framework developed to classify surgical skill levels during laparoscopic nephrectomy. While traditional skill assessment relies on expert observation and structured rating systems, motion-capture data provide an objective and quantitative representation of surgical performance. The goal of this chapter is to utilize these objective indices to construct robust classification models capable of distinguishing between different levels of surgical expertise.

The analyses focus on two complementary classification problems:

- **Three-group classification:** novice, intermediate, and expert.
- **Binary classification:** ESSQ-certified versus non-certified surgeons.

To support these tasks, we integrate dimensionality reduction using Principal Component Analysis (PCA), supervised learning models (Random Forest, Support Vector Machine, and PCA-SVM), and a nested and repeated 5-fold cross-validation framework to ensure generalizability and prevention of overfitting.

This chapter is organized as follows. Section 4.1 presents the PCA results and interprets the latent structure underlying the kinematic feature space. Section 4.2 describes the construction of the machine-learning models and the normalization and validation pipeline. Section 4.3 reports classification performance across phases and model types. Section 4.4 introduces SHAP-based model interpretation to identify the most influential motion features contributing to classification decisions.

Previous studies have applied machine-learning models to surgical motion features for automated skill assessment, primarily in box trainers, VR simulators, or robotic systems [13, 41, 42, 43]. In this chapter, we extend such approaches to Thiel-embalmed cadaveric laparoscopic nephrectomy using a dedicated MoCap-based feature set and phase-wise analysis.

The methods and results presented here form the foundation for the regression analysis of GOALS scores in Chapter 5.

4.1 Principal Component Analysis

PCA was performed to clarify the latent structure of the kinematic feature space and to identify the combinations of indices that most strongly characterize surgical skill. Because the original feature set included highly correlated temporal, spatial, and velocity–distribution metrics, PCA provided an effective approach to reduce dimensionality while retaining the major variance components relevant to technical proficiency.

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms a set of correlated variables into a smaller number of uncorrelated principal components. Each principal component represents a linear combination of the original features and captures a specific direction of variance in the data.

In this study, PCA was applied to the normalized kinematic feature set to identify dominant patterns of variation associated with surgical skill. The first few principal components (PC1–PC3) were analyzed, as they explain the majority of the variance and provide interpretable representations of underlying motor behavior.

Two PCA analyses were conducted: (1) three-group classification (novice, intermediate, expert), and (2) binary classification based on ESSQ certification. Separate PCA models were computed for the whole procedure and for each of the manually segmented subprocesses.

4.1.1 PCA for three-group classification

Figures 4.1a presents the loading heatmaps for whole phase, and figures 4.1b–4.1d present the loading heatmaps for Phase 1 and Phase 3. Across both tasks, the first three principal components (PC1–PC3) explained the majority of the variance beyond (75%), and consistent patterns emerged.

PC1: Motion efficiency and economy of movement. PC1 showed strong negative loadings for:

- Path length (PL) of all forceps;
- Depth-direction path length (DPL);
- Low-speed ratio (PLow);
- Task duration (OT).

These indices are known markers of inefficient motor control. Experts typically showed smaller PL, lower PLow, and shorter OT, meaning that a larger negative PC1 score corresponds to higher proficiency. Conversely, novices showed large positive contributions to PC1, reflecting redundant movements and poor temporal efficiency.

PC2: Velocity distribution balance and smoothness. PC2 was dominated by the velocity-distribution features (idle, low, middle, high, very-high zones). Intermediate surgeons tended to show moderate contributions across all ranges, while experts exhibited higher high-speed ratios and lower idle/low-speed components. PC2 thus captures smoothness and continuity of instrument manipulation.

PC3: Localized instrument-specific patterns. PC3 presented loadings specific to advanced manipulations:

- Pitch–yaw amplitude changes (AL–PitchYaw);
- Forceps-specific speed/acceleration;
- Deep-angle trajectory components.

These patterns often reflect short, decisive movements used by experts during critical steps (e.g., sealing, cutting, clip application). PC3 therefore represents higher-order control strategies rather than basic efficiency.

Interpretation of skill groups. When plotting PC1–PC2 scores, novices clustered strongly at high PC1 and low PC2, indicating slow, inefficient trajectories. Experts clustered at the opposite quadrant (low PC1, high PC2), reflecting rapid and economical manipulation. Intermediate participants were distributed between the two extremes, confirming that PCA captured meaningful skill gradients.

4.1.2 PCA for ESSQ-based binary classification

Figures 4.2a–4.2d display the PCA loadings for ESSQ-certified versus non-certified surgeons.

Compared with the three-group model, the ESSQ PCA emphasized slightly different dimensions:

PC1: High-level motion proficiency. PC1 contained:

- very-high-speed ratio (PVH);
- acceleration/jerk indices;
- minimal path length of the dominant-hand forceps.

ESSQ-certified surgeons typically demonstrated:

- shorter PL,
- smaller DPL,
- higher PVH and PH,
- and reduced idle/low-speed ratios.

Thus, PC1 in the ESSQ model maps directly to expert kinematic signatures.

PC2: Instrument-specific strategic patterns. PC2 in the ESSQ PCA highlighted:

- grasping forceps working area (WA),
- scissor forceps jerk,
- clip applier UT (using time),

- depth-direction velocities (DV).

This suggests that ESSQ surgeons tend to:

- work within a compact and controlled workspace,
- execute clean and purposeful scissor movements,
- and use the clip applier more efficiently.

PC3: Fine motor control and angular adjustments. PC3 showed meaningful contributions from:

- AL–PitchYaw,
- AL–Roll,
- middle-speed ratio (PMiddle).

These angular parameters are associated with subtle adjustments of the instrument tip—skills typically mastered by ESSQ-certified surgeons.

Interpretation for ESSQ skill classification. ESSQ-certified surgeons formed a compact cluster at lower PC1 (high efficiency) and higher PC2 (controlled workspace), whereas non-certified participants were widely dispersed. This indicates that ESSQ criteria correspond closely to motion-efficiency dimensions captured automatically by PCA.

Comparison between whole-procedure PCA and phase-wise PCA revealed both consistent and phase-specific characteristics of surgical motion.

Across all phases, PC1 consistently represented motion efficiency, with similar contributions from path length, low-speed ratio, and task duration, indicating that overall efficiency is a stable and global indicator of surgical skill.

In contrast, phase-wise PCA demonstrated variations in PC2 and PC3. In Phase 1 (colon mobilization), broader spatial movements and workspace-related features showed higher contributions, reflecting exposure and tissue mobilization. In Phase 2 (renal vessel control), depth-related motion and fine instrument control became more dominant, consistent with the precision required for vascular dissection and clipping. Phase 3 (perirenal tissue dissection) exhibited mixed characteristics, combining both efficiency and localized control patterns.

These findings suggest that while global skill differences are captured by common latent dimensions, phase-based analysis provides additional resolution to identify phase-specific motor strategies.

4.1.3 Summary

Across all tasks and classification settings, PCA consistently identified three latent dimensions of surgical manipulation:

1. **Motion efficiency** (PC1): trajectory length, temporal efficiency, and low-speed dominance.
2. **Smoothness and velocity distribution** (PC2): proportion of fast, continuous motion versus idle/low-speed manipulation.

3. **Higher-order control strategies** (PC3): instrument-specific angular adjustments and fine motor behaviors.

These components align with established skill assessment frameworks and match the descriptive analyses presented in Chapter 3. Therefore, PCA not only serves as a dimensionality reduction tool but also provides meaningful, interpretable structure that supports subsequent machine-learning classification in Section 4.3.

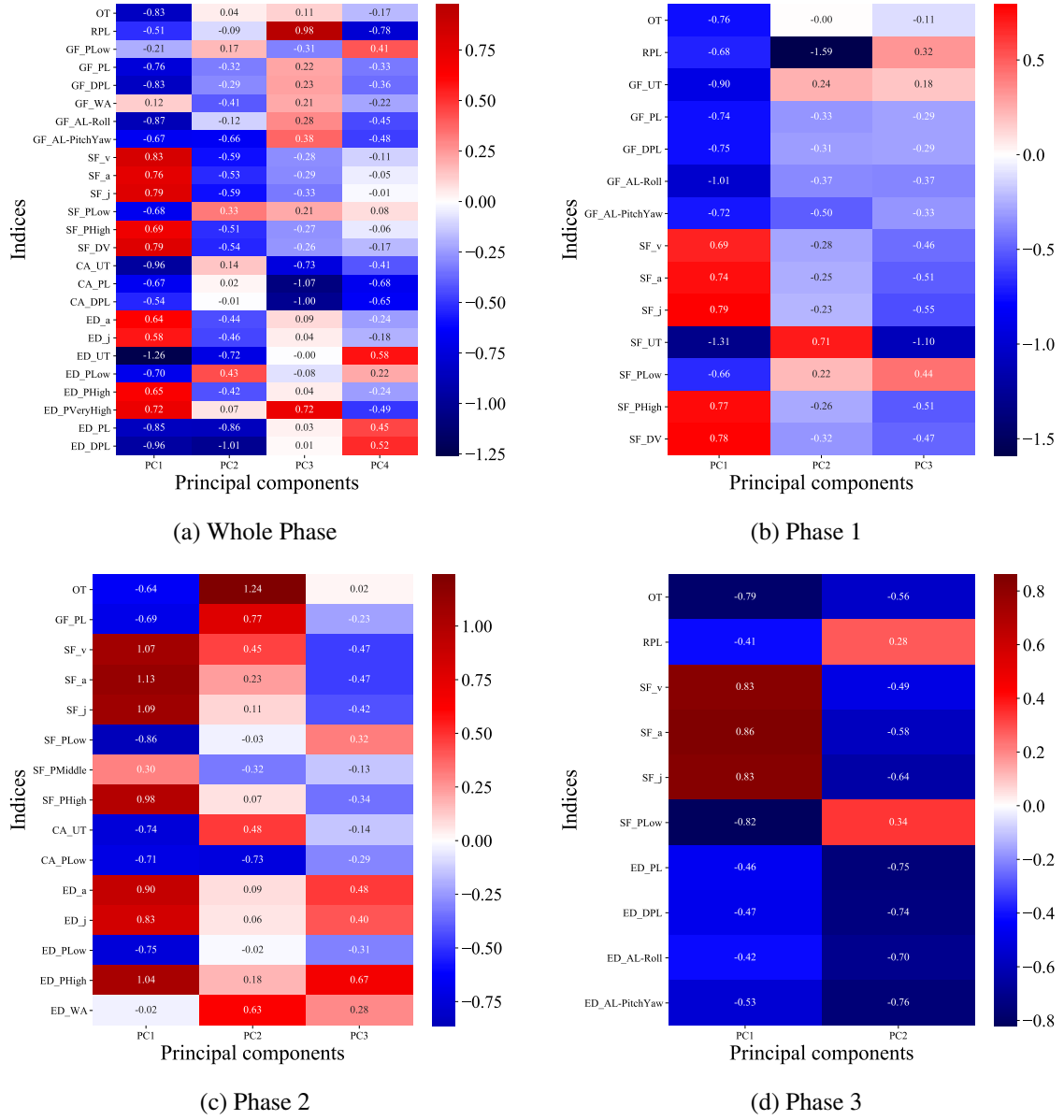
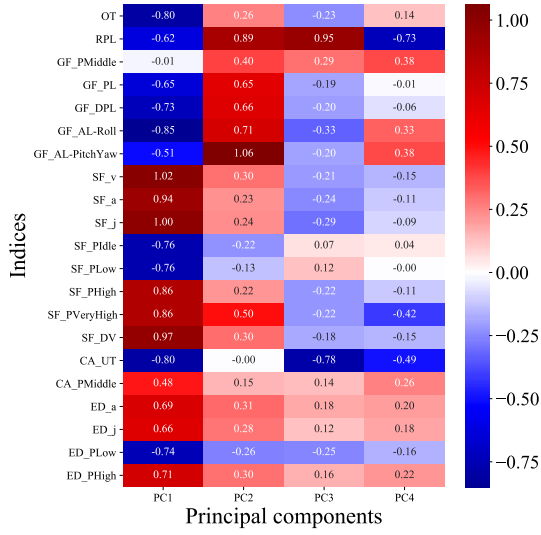
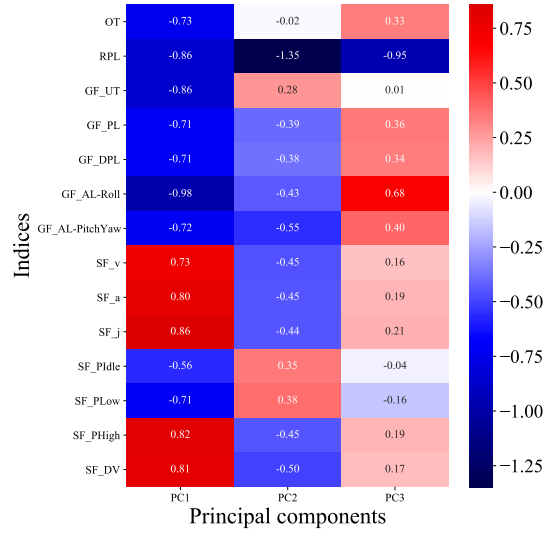


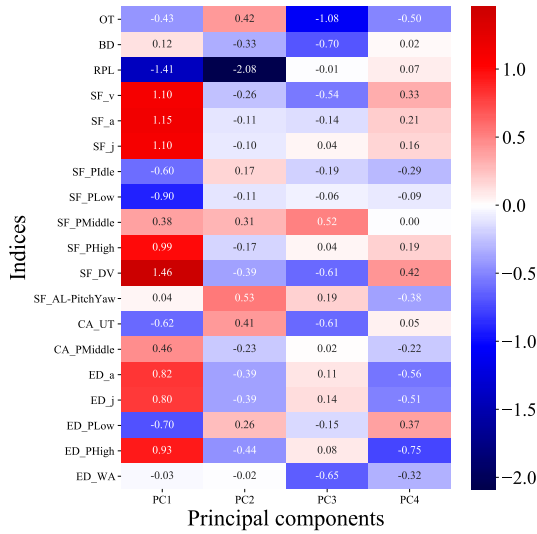
Figure 4.1: PCA loading heatmaps for three-group classification: (a) whole phase, (b) Phase 1, (c) Phase 2, and (d) Phase 3.



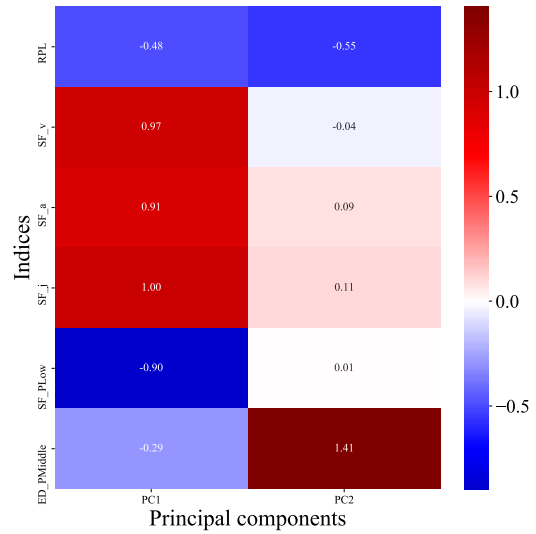
(a) Whole Phase (ESSQ)



(b) Phase 1 (ESSQ)



(c) Phase 2 (ESSQ)



(d) Phase 3 (ESSQ)

Figure 4.2: PCA loading heatmaps for ESSQ-based binary classification: (a) Whole Phase, (b) Phase 1, (c) Phase 2, and (d) Phase 3.

4.2 System Construction for Machine-Learning–based Skill Classification

To develop a robust and reproducible analytical pipeline for surgical skill classification, a fully standardized machine-learning framework was constructed. The system integrates all components from motion-capture preprocessing, feature extraction, feature selection, dimensionality reduction, and model training, to cross-validated performance evaluation. This unified workflow ensures that classification results are directly comparable across different analytical settings (three-class experience level classification and ESSQ-based two-class classification), and supports transparent interpretation in subsequent SHAP analysis.

4.2.1 Overview of the analytical workflow

Figure 4.3 presents an overview of the machine-learning pipeline (Model 1), which consists of the following components:

1. **Data preprocessing:** missing-value handling, measurement-rate filtering, and outlier removal using the robust Z-score method;
2. **Feature construction:** extraction of kinematic features from each instrument, including velocity/acceleration metrics, frequency-domain features, motion smoothness indices, complexity descriptors, and cumulative orientation metrics (AL-Roll and AL-PitchYaw);
3. **Statistical filtering:** Kruskal–Wallis tests (three-class) or Mann–Whitney U tests (ESSQ two-class) to pre-select significant features ($p < 0.05$);
4. **Dimensionality reduction:** principal component analysis (PCA) applied to standardized features, reducing redundancy and extracting orthogonal motion components;
5. **Model training:** three primary classifiers were used: Support Vector Machine (SVM), PCA-SVM, and Random Forest (RF);
6. **Performance evaluation:** nested cross-validation (inner-loop parameter search, outer-loop testing) to obtain unbiased accuracy estimates.

The pipeline was identically applied to both classification tasks: (i) novice–intermediate–expert (three-class), and (ii) ESSQ-certified vs. non-certified (two-class).

4.2.2 Advanced workflow with SHAP interpretability

To enhance interpretability, an extended pipeline was constructed. In this model, the best-performing classifier from Model 1 is re-trained on the full dataset, and SHAP (Shapley Additive exPlanations) values are computed to quantify the contribution of each feature to the predicted skill level.

This extended workflow provides two major benefits:

- **Pedagogical interpretability:** SHAP highlights which motion features (e.g., jerk complexity, AL-Roll, velocity irregularity, low-velocity duration) most strongly distinguish expert-like from novice-like motions.

- **Consistency evaluation:** SHAP-based rankings can be compared against PCA loading patterns and statistical significance results, helping confirm whether similar motion dimensions repeatedly emerge as key skill determinants.

Importantly, SHAP analysis was applied *only after* nested cross-validation confirmed classifier reliability, thereby avoiding overinterpretation of unstable models.

4.2.3 Rationale for pipeline design

This system was designed with the following objectives:

- **Reproducibility:** Standardized preprocessing and cross-validation ensure that classification performance does not depend on arbitrary parameters.
- **Generalizability:** Nested cross-validation and PCA reduce the risk of overfitting, enabling stable evaluation even with moderate-sized datasets.
- **Interpretability:** SHAP analysis bridges the gap between abstract feature vectors and concrete surgical motion behaviors, supporting future educational use.
- **Integration across experiments:** The same workflow is applicable to real surgery, cadaver training, and wet-lab models (for future extensions), enabling cross-environment comparison of motion characteristics.

Overall, this system provides a unified framework for all subsequent analyses in this dissertation, including PCA component interpretation (Section 4.1), classification accuracy evaluation (Section 4.3), and SHAP-based feature interpretation (Section 4.4).

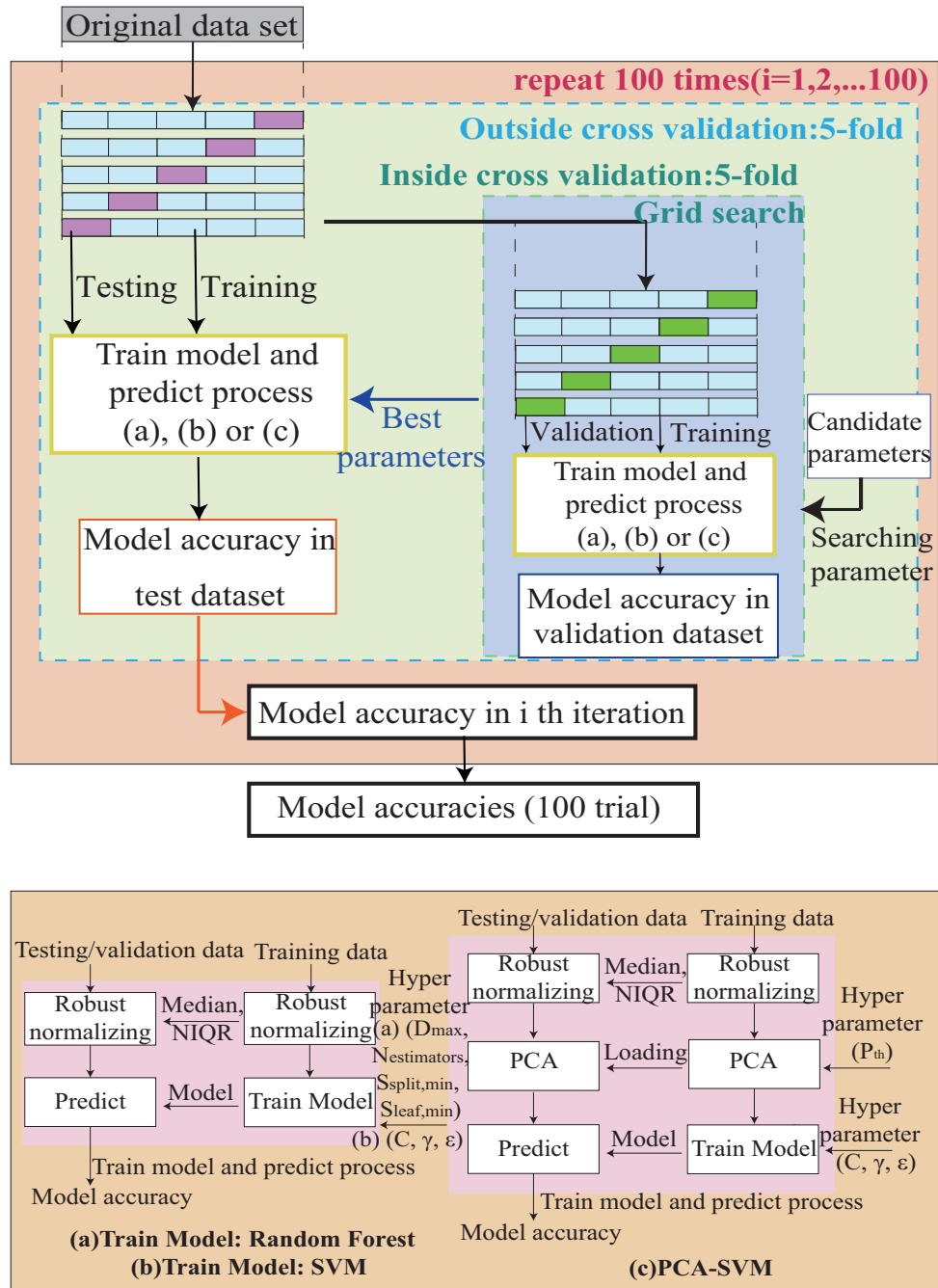


Figure 4.3: Overview of the machine-learning analysis pipeline. The workflow integrates preprocessing, feature extraction, statistical filtering, PCA-based dimensionality reduction, classifier training, and nested cross-validation

4.3 Classification Results

This section presents the performance of the machine-learning models for two prediction tasks: (i) three-class classification of surgical experience level (novice, intermediate, expert), and (ii) binary classification of the Endoscopic Surgical Skill Qualification (ESSQ). All models were evaluated using nested and repeated 5-fold cross-validation, ensuring unbiased estimation of generalization performance. Three models were compared: Random Forest, Support Vector Machine (SVM), and PCA-SVM. Classification accuracies are reported as median values with interquartile range (Q1–Q3).

4.3.1 Three-class classification results

Figure 4.4 summarizes the model accuracies for the whole phase and for Phases 1–3, respectively. Across all conditions, Random Forest consistently achieved the highest accuracy, followed by SVM and PCA-SVM. For the whole phase model, Random Forest reached a median accuracy of 76%, while SVM and PCA-SVM yielded comparable performance (75% and 74%, respectively).

Phase-wise evaluation revealed substantial differences among tasks. Phase 1 showed similar trends to the whole [phase, whereas phases 2 and 3 demonstrated significantly lower accuracies across all models. In particular, Phase 2 yielded the lowest performance, indicating that this segment contains complex or less discriminative motion patterns for experience classification. Statistical testing using the Kruskal–Wallis test followed by Dunn’s post-hoc correction confirmed significant differences between the three models in most phases, especially for phases 2 and 3 ($p < 0.0001$), as shown in Figure 4.4.

Table 4.1 provides the numerical values of the accuracies for each model. Random Forest performed consistently well across all participants and phases. In contrast, PCA-SVM tended to show lower accuracy, despite the dimensionality reduction effect of PCA. This suggests that the PCA-projected subspace may not preserve the discriminative structure required for experience-level classification.

To further clarify the relationship between whole-procedure and phase-wise analyses, classification performance was directly compared across these settings (Table 4.1).

The whole-procedure model achieved the highest and most stable accuracy across classifiers, with Random Forest reaching a median accuracy of 76.00%. In contrast, phase-wise models exhibited a clear decline in performance, particularly in phases 2 and 3.

Among the segmented analyses, Phase 1 showed performance comparable to the whole-procedure model, whereas Phases 2 and 3 demonstrated progressively lower accuracies across all models. In particular, Phase 2 yielded a median accuracy of 69.78% (Random Forest), indicating that this phase involves more complex or less discriminative motion patterns for experience classification.

These findings suggest that while whole-procedure analysis provides a more robust representation of overall surgical skill, phase-based segmentation offers complementary insight into phase-dependent variability in motor behavior.

4.3.2 ESSQ classification results

The ESSQ prediction task yielded substantially higher accuracies than the three-class experience model. Across the whole phase, Random Forest achieved a median accuracy of 89.11%, clearly outperforming SVM (84.89%) and PCA-SVM (86.78%). This trend was consistent across phases 1–3, as shown in Figure 4.5. The high accuracy reflects the strong discriminative power of kinematic indices for identifying surgeons who have achieved ESSQ certification.

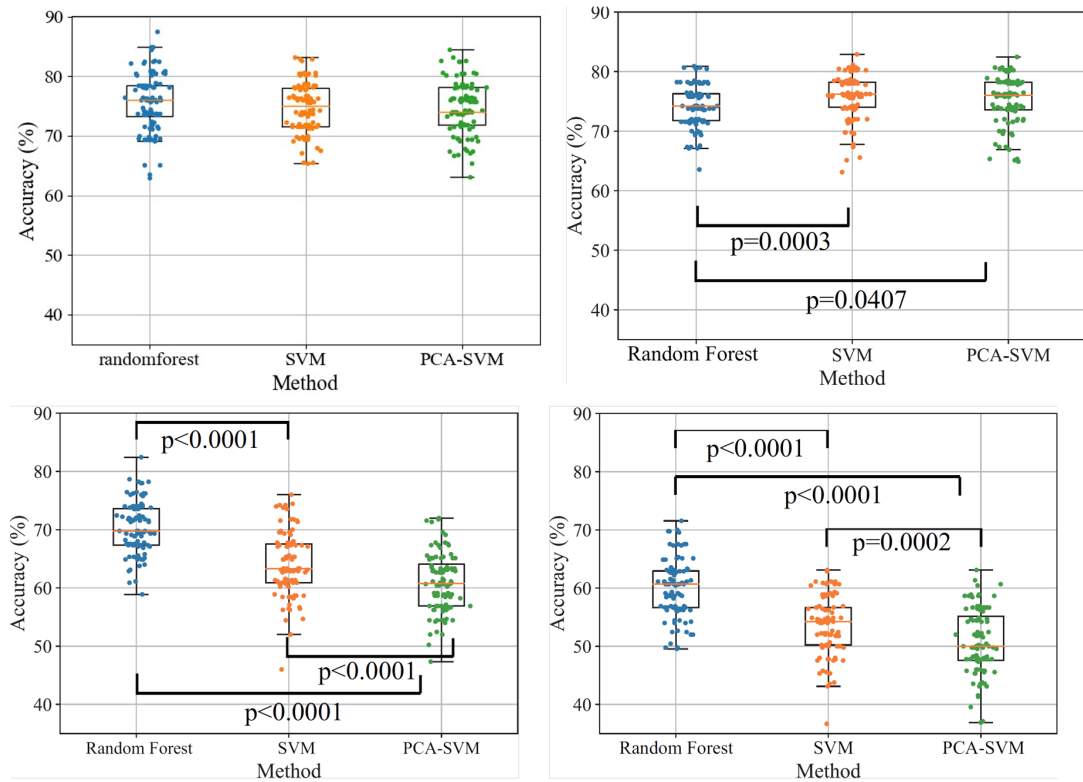


Figure 4.4: Three-class classification accuracies for the whole phase and phases 1–3. Random Forest showed significantly higher performance than SVM and PCA-SVM in most comparisons. Statistical significance was evaluated using the Kruskal–Wallis test.

Phase-wise results revealed that Phase 1 provided the highest classification accuracy, while phase 2 and 3 yielded slightly lower yet still robust performance. The Kruskal–Wallis test confirmed significant differences among models across nearly all phases ($p < 0.0001$). Random Forest again outperformed the other approaches, indicating superior robustness to nonlinearity and noise in surgical motion data.

Table 4.3 summarizes the accuracy statistics. Notably, the Random Forest model reached over 90% accuracy for Phase 1, suggesting strong alignment between motion characteristics and ESSQ competency. Even the lowest-performing groups (e.g., Phase 3) maintained median accuracies above 78%, highlighting the reliability of motion-based ESSQ prediction.

To further examine the relationship between whole-procedure and phase-wise analyses in ESSQ classification, performance was compared across all settings (Table 4.3).

The whole-procedure model achieved the highest and most stable performance, with Random Forest reaching a median accuracy of 89.11%. In contrast, phase-wise models showed slightly reduced but still robust accuracies across all classifiers.

Among the segmented analyses, Phase 1 exhibited the highest performance (91.11% with Random Forest), even exceeding the whole-procedure model, suggesting that early-stage surgical actions strongly reflect ESSQ-level proficiency. In comparison, Phases 2 and 3 showed lower accuracies (e.g., 85.22% and 78.44% for Random Forest), indicating increased variability or reduced discriminative power in later surgical phases.

These findings suggest that, unlike the three-class classification task, ESSQ prediction remains

Table 4.1: Classification accuracies for three-class models across whole phase and phases 1–3. Values represent median accuracy with interquartile range (Q1–Q3).

Participant	Model	Accuracy (%) (Q1–Q3)
Whole Phase	Random Forest	76.00 (72.72–78.44)
	SVM	75.00 (71.56–78.00)
	PCA-SVM	74.00 (71.78–78.22)
P1	Random Forest	74.22 (71.78–76.39)
	SVM	76.22 (74.00–78.22)
	PCA-SVM	76.00 (73.56–78.22)
P2	Random Forest	69.78 (67.33–73.72)
	SVM	63.33 (60.89–67.56)
	PCA-SVM	60.78 (56.89–64.61)
P3	Random Forest	60.67 (56.67–63.06)
	SVM	54.22 (50.22–56.67)
	PCA-SVM	50.00 (47.56–55.72)

highly robust across both whole-procedure and phase-wise analyses. While whole-procedure analysis provides a stable global assessment, phase-based analysis—particularly in Phase 1—can capture strong and interpretable indicators of certification-level skill.

4.3.3 Confusion matrix analysis

To further characterize model behavior beyond overall accuracy, confusion matrices were generated for the best-performing classifier of each phase: Random Forest for Phase 2 and Phase 3, and SVM for Phase 1. Figures 4.6a–4.6c summarize the class-wise prediction performance.

In Phase 1 (Fig. 4.6a), all three skill groups were well distinguished. Novice trials were correctly classified at 76%, and experts at 67.3%, with most misclassifications occurring between adjacent levels (e.g., novice → intermediate). Only a very small fraction of experts were misclassified as novices (0.3%), indicating a clear separation in kinematic patterns.

For Phase 2 (Fig. 4.6b), the diagonal elements remained dominant (69–72%), suggesting that the motion characteristics during hilar dissection retain moderate discriminability. Novice trials were occasionally predicted as intermediate (24.0%), and experts were sometimes confused with intermediate (25.1%), reflecting the increased complexity and variability of this phase.

In Phase 3 (Fig. 4.6c), class separability declined further, especially for novice and expert surgeons, consistent with the reduced discriminative power seen in overall accuracy. Intermediate surgeons remained the most distinguishable group (68.3%), whereas novices and experts exhibited greater mutual overlap.

Overall, the confusion matrix analysis reveals a consistent trend: earlier phases of the nephrectomy task (Phase 1) provide clearer group differences, while later phases (Phase 3) introduce higher variability and reduce classification robustness. These results support the importance of structured, phase-specific evaluation in surgical skill assessment.

4.3.4 Summary of classification performance

Overall, the classification results provide the following insights:

Table 4.2: Classification accuracies for ESSQ qualification prediction (two-class). Values represent median accuracy with interquartile range (Q1–Q3).

Participant	Model	Accuracy (%) (Q1–Q3)
Whole Phase	Random Forest	89.11 (86.89–89.33)
	SVM	84.89 (82.67–87.11)
	PCA-SVM	86.78 (82.78–89.11)
P1	Random Forest	91.11 (89.11–93.33)
	SVM	87.11 (84.89–89.33)
	PCA-SVM	88.89 (84.89–91.11)
P2	Random Forest	85.22 (84.44–87.11)
	SVM	80.44 (78.22–82.67)
	PCA-SVM	80.44 (78.22–82.61)
P3	Random Forest	78.44 (76.28–81.89)
	SVM	82.67 (80.44–84.89)
	PCA-SVM	82.67 (80.44–84.67)

- Random Forest was the most reliable model across all tasks and phases, showing strong robustness and generalization.
- ESSQ classification achieved substantially higher accuracy than the three-class experience classification, indicating that ESSQ certification is more strongly reflected in motion-kinematic patterns.
- Phase-wise analysis revealed that Phase 1 carries more discriminative information than Phases 2 and 3.
- PCA-SVM did not outperform standard SVM, suggesting that PCA-based dimensionality reduction may suppress informative variance in this dataset.

These findings validate the feasibility of motion-capture–based classification of surgical proficiency and support the use of Random Forest as a practical and effective model for objective surgical skill assessment.

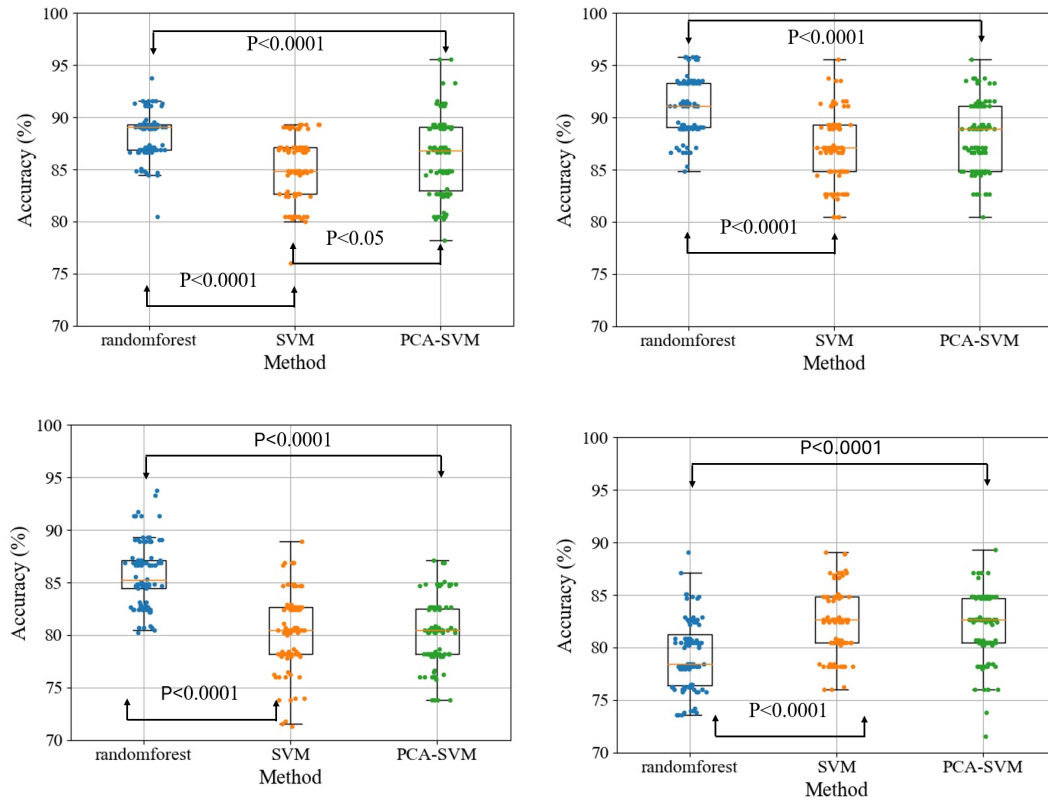


Figure 4.5: ESSQ classification accuracy for the whole phase and phases 1–3. Overall accuracy was substantially higher than in the three-class task. Random Forest consistently achieved the highest accuracy.

Table 4.3: Classification accuracies for ESSQ qualification prediction (two-class). Values represent median accuracy with interquartile range (Q1–Q3).

Participant	Model	Accuracy (%) (Q1–Q3)
Whole Phase	Random Forest	89.11 (86.89–89.33)
	SVM	84.89 (82.67–87.11)
	PCA-SVM	86.78 (82.78–89.11)
P1	Random Forest	91.11 (89.11–93.33)
	SVM	87.11 (84.89–89.33)
	PCA-SVM	88.89 (84.89–91.11)
P2	Random Forest	85.22 (84.44–87.11)
	SVM	80.44 (78.22–82.67)
	PCA-SVM	80.44 (78.22–82.61)
P3	Random Forest	78.44 (76.28–81.89)
	SVM	82.67 (80.44–84.89)
	PCA-SVM	82.67 (80.44–84.67)

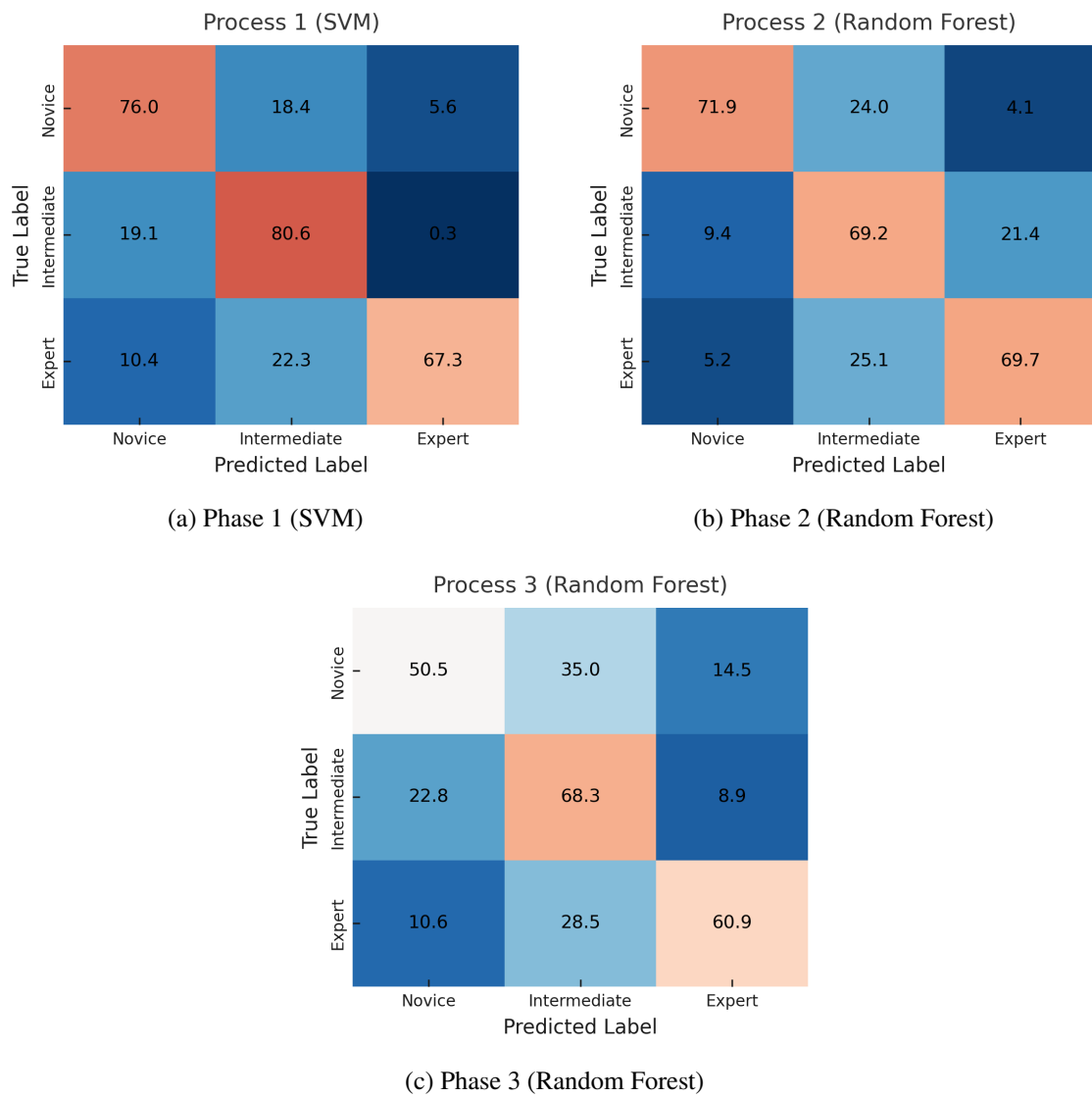


Figure 4.6: Confusion matrices for phases 1–3.

4.4 SHAP-based Interpretation of Classification Models

To interpret how individual kinematic features contributed to the classification performance, Shapley Additive exPlanations (SHAP) analysis was applied to the PCA-SVM models trained for each phase (P1–P3). SHAP provides feature-wise contribution scores for each prediction based on cooperative game theory, enabling model-agnostic and quantitative interpretation of feature importance.

4.4.1 SHAP analysis for Phase 1

As shown in Figure 4.7, the most influential features in Phase 1 were dominated by scissor forceps (SF) usage time (SF_UT), relative path length (RPL), and orientation-related features of the grasping forceps (GF_AL-Roll and GF_AL-PitchYaw).

SF_UT exhibited the highest SHAP magnitude, indicating that the temporal usage pattern of the scissor forceps strongly contributed to skill classification. In addition, larger RPL and smaller angular variability (GF_AL-Roll and GF_AL-PitchYaw) were associated with increased model output toward the novice class, reflecting less stable and less efficient instrument control.

Conversely, experts tended to show more compact trajectories and reduced orientation variability, which contributed to predictions toward higher skill levels.

These findings suggest that, in the early phase, skill differentiation is primarily characterized by instrument usage patterns and motion stability rather than solely by velocity-based smoothness indices.

4.4.2 SHAP analysis for Phase 2

In Phase 2 (renal pedicle dissection and vessel isolation), the dominant features shifted toward distribution-based and efficiency-related metrics (Figure 4.8).

The proportion of low-velocity motion of the scissor forceps (SF_PLow) and overall operation time (OT) showed the highest SHAP contributions, followed by ED_PLow and SF_PMiddle. These results indicate that motion distribution characteristics, rather than single-point kinematic values, play a central role in this phase.

Higher proportions of low-velocity motion and longer operation time were associated with increased likelihood of novice classification, suggesting that inexperienced surgeons tend to perform slower and less decisive manipulations.

In contrast, experts demonstrated more balanced velocity distributions and shorter operation times, reflecting efficient and well-paced dissection during vessel handling.

Additionally, path length of the grasping forceps (GF_PL) and working area of the energy device (ED_WA) contributed moderately, indicating the importance of coordinated bimanual control and spatial precision during this technically demanding phase.

4.4.3 SHAP analysis for Phase 3

Phase 3 (remaining tissue removal and final detachment) demonstrated a more heterogeneous feature pattern (Figure 4.9).

Operation time (OT) showed the largest SHAP contribution, followed by scissor forceps jerk (SF_j) and energy device path-related features (ED_DPL and RPL). These results indicate that both temporal efficiency and motion smoothness are critical for distinguishing skill levels in the final phase.

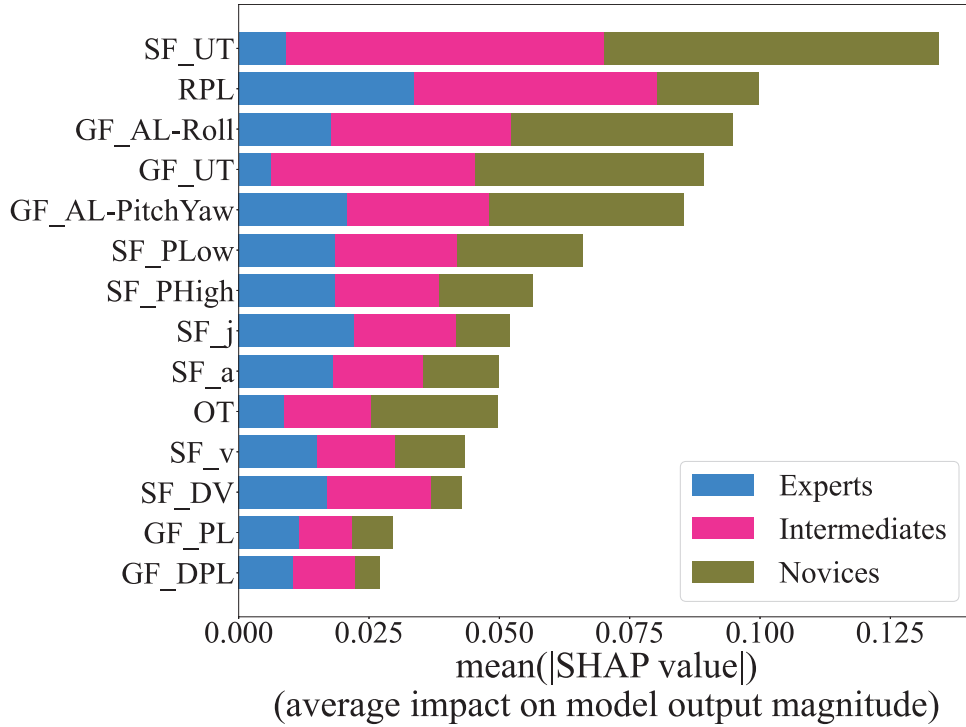


Figure 4.7: SHAP summary plot for Phase 1 using the PCA-SVM classifier. Positive SHAP values contribute to predicting higher skill classes.

Higher jerk values and longer path-related measures were associated with novice predictions, reflecting unstable and redundant movements during tissue detachment.

Orientation-related features of the energy device (ED_AL-PitchYaw and ED_AL-Roll) and scissor forceps velocity (SF_v) also contributed to the model, suggesting that fine control and precise instrument alignment are important for safe and efficient completion of the procedure.

Overall, Phase 3 emphasized integrated multi-instrument coordination, where both temporal efficiency and spatial control jointly determine surgical proficiency.

4.4.4 Summary of SHAP-based interpretation

Across all three phases, several consistent patterns were observed:

- Efficiency-related metrics (e.g., operation time and path length) and motion stability indicators (e.g., jerk and angular variability) were consistently important for skill classification.
- Distribution-based features (e.g., low-velocity and mid-velocity proportions) played a prominent role in Phase 2, highlighting the importance of motion regulation during technically demanding phases.
- The dominant contributing features varied by phase: instrument usage and orientation control in Phase 1, motion distribution and efficiency in Phase 2, and integrated temporal and spatial control in Phase 3.
- The SHAP-derived patterns were consistent with known surgical workflow characteristics, supporting the validity of the engineered kinematic feature set.

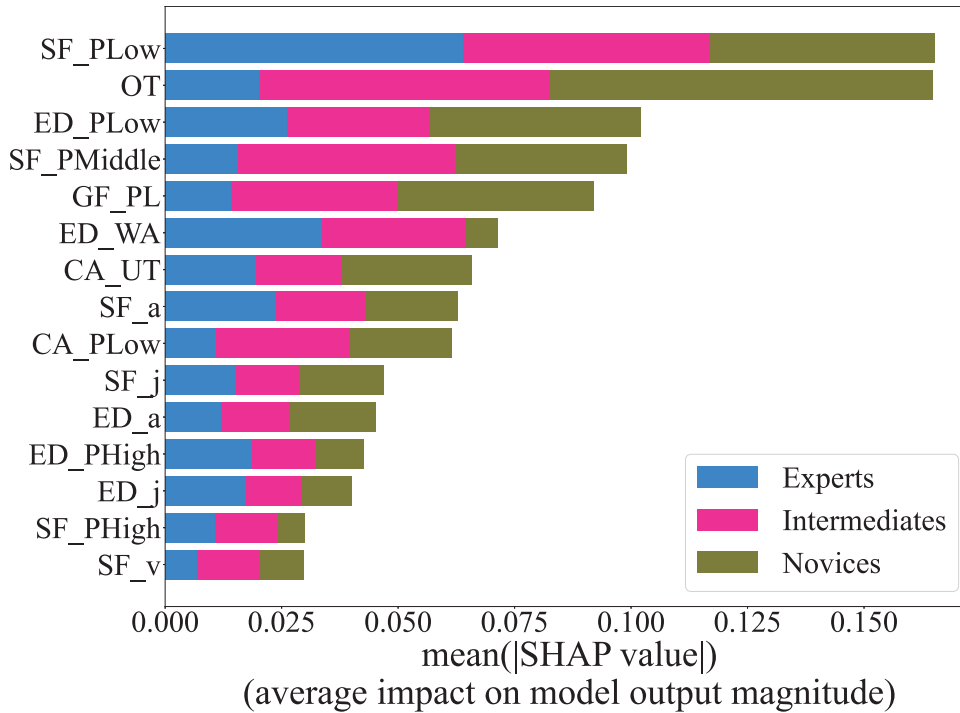


Figure 4.8: SHAP summary plot for Phase 2 using the PCA-SVM classifier.

These findings demonstrate that SHAP analysis provides interpretable, phase-specific insights into surgical skill, bridging the gap between machine-learning predictions and clinically meaningful behavior patterns. This interpretability is essential for future applications such as feedback-based training, surgical coaching, and automated assessment systems.

4.5 Summary

This chapter presented a comprehensive machine-learning framework for objective surgical skill classification based on kinematic features extracted from cadaveric laparoscopic nephrectomy procedures. Three classification models—Random Forest, SVM, and PCA-SVM—were evaluated under a rigorous nested and repeated 5-fold cross-validation design, ensuring unbiased estimation of model performance.

The results demonstrated that all three models were capable of distinguishing novice, intermediate, and expert surgeons, with Random Forest achieving the highest accuracy for the three-class task. When the surgical workflow was divided into three manually annotated phases (P1–P3), the classification accuracy increased in P1 and P2, indicating that skill-relevant information is not uniformly distributed throughout the procedure and is concentrated in technically demanding phases such as vascular dissection.

ESSQ-based binary classification yielded even higher performance, with Random Forest consistently outperforming SVM and PCA-SVM, achieving median accuracies above 89% across all phases. These results suggest that motion-capture-derived kinematic indices contain robust information correlating with formally certified surgical proficiency.

Furthermore, SHAP analysis provided detailed, interpretable insights into the feature sets underlying each model’s predictions.

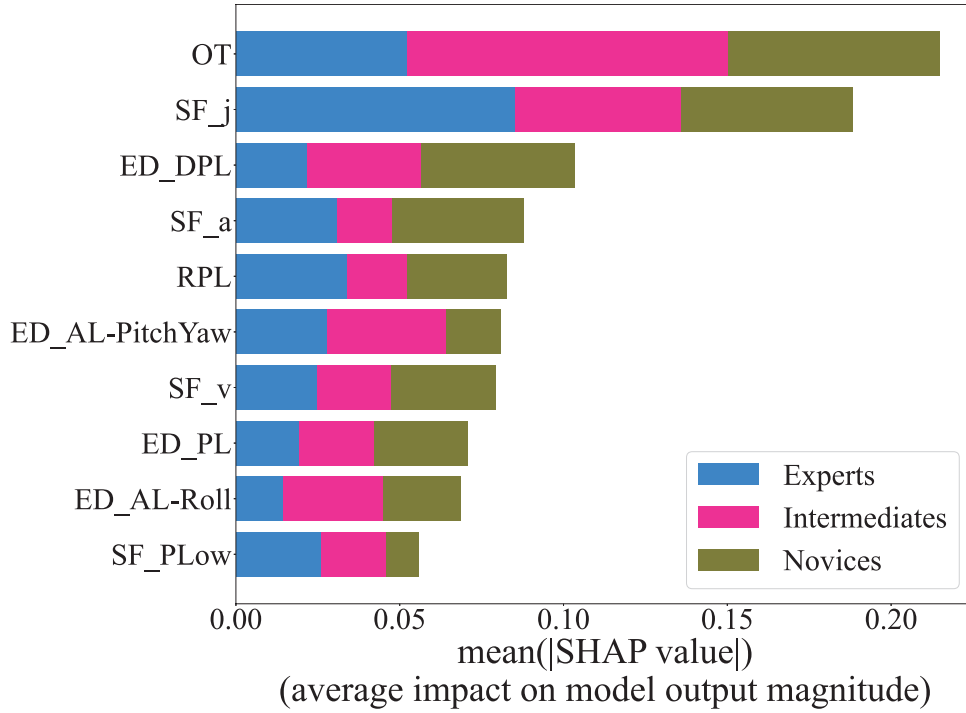


Figure 4.9: SHAP summary plot for Phase 3 using the PCA-SVM classifier.

Efficiency-related metrics (e.g., operation time and path length) and motion stability indicators (e.g., jerk and angular variability) were consistently important, although their relative contributions varied across phases.

In particular, instrument usage and orientation control were more prominent in Phase 1, distribution-based features (e.g., low-velocity proportion) dominated in Phase 2, and combined temporal and spatial control played a key role in Phase 3.

These interpretable patterns aligned closely with surgical workflow characteristics and supported the validity of the engineered kinematic features.

Overall, the findings in this chapter demonstrate that:

- Kinematic data alone enable reliable, high-performing surgical skill classification.
- Phase-specific analysis improves performance and reveals phase-specific skill characteristics.
- SHAP interpretability bridges the gap between machine-learning output and surgical reasoning, supporting educational and clinical applicability.

The next chapter extends this framework toward continuous skill assessment by constructing regression models that predict surgeon performance using GOALS (Global Operative Assessment of Laparoscopic Skills) scores.

Chapter 5

Regression Modeling for Continuous Surgical Skill Assessment

Building on the classification framework established in Chapter 4, this chapter advances the objective assessment of laparoscopic surgical skills from discrete skill categorization toward continuous, quantitatively interpretable scoring. While classification models distinguish broad skill levels (e.g., novice, intermediate, expert), they are inherently limited in capturing subtle, within-level variations or providing granular feedback that can guide individualized training.

To address this limitation, we develop regression models that estimate the Global Operative Assessment of Laparoscopic Skills (GOALS) score [27] directly from motion-capture-derived kinematic features. GOALS is a standardized and widely accepted metric for evaluating laparoscopic technical proficiency across five domains: depth perception, bimanual dexterity, efficiency, tissue handling, and autonomy. In this study, two expert surgeons independently rated each cadaveric nephrectomy procedure, and their averaged scores served as the reference labels for regression.

This chapter describes the construction, validation, and interpretation of the regression models, including Support Vector Regression (SVR), Ridge Regression, and Partial Least Squares (PLS).

Through these analyses, this chapter aims to demonstrate that continuous skill assessment—grounded in objectively measured motion data—can not only approximate expert ratings but also provide interpretable, fine-grained feedback suitable for surgical training and competency evaluation.

5.1 GOALS rating and inter-rater reliability

5.1.1 GOALS rating procedure

To obtain criterion-based measures of technical performance, all cadaveric nephrectomy trials were evaluated using the *Global Operative Assessment of Laparoscopic Skills* (GOALS) score. Two experienced laparoscopic surgeons (Reviewer A and Reviewer B), who were not involved in the machine-learning analysis, independently rated each trial.

For every participant, GOALS was scored separately for the three manually segmented phases (Phase 1–3) according to the definitions described in Section 3.7. For each phase, both raters evaluated five domains—depth perception (D), bimanual dexterity (B), efficiency (E), tissue handling (T), and autonomy (A)—on a 5-point Likert scale (1 = poor, 5 = excellent). A total GOALS score

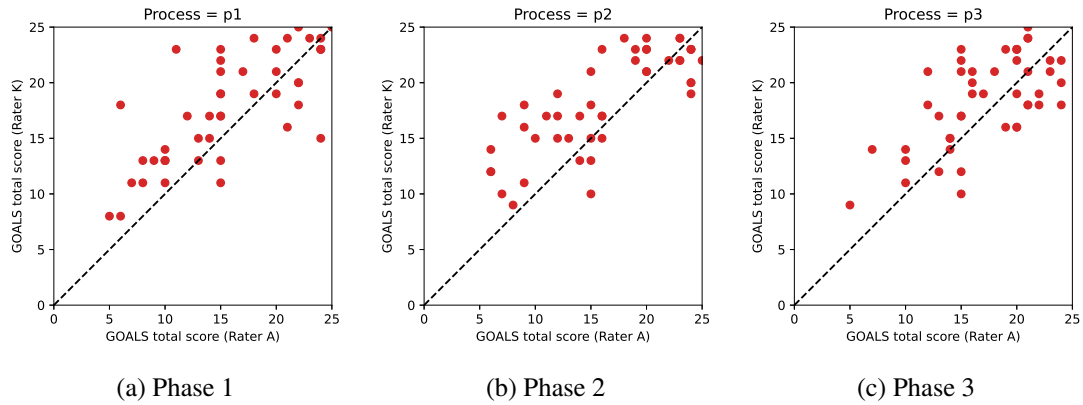


Figure 5.1: Agreement between two expert raters for GOALS total scores in each phase. Each point represents one cadaveric nephrectomy trial, and the dashed line indicates perfect agreement ($y = x$).

(range 5–25) was obtained by summing the five domain scores. Raters were blinded to the motion-capture features, to the experience group of each participant, and to each other’s ratings. All videos were anonymized and presented in random order, and raters were allowed to replay sequences as needed.

5.1.2 Statistical analysis of inter-rater reliability

Inter-rater reliability of the GOALS ratings was examined using both correlation- and agreement-based indices. Because GOALS scores are ordinal and subsequent analyses in this dissertation rely primarily on rank-based statistics, the main association measure between the two raters was Spearman’s rank correlation coefficient (ρ). Spearman’s ρ was computed for each GOALS domain as well as for the total score, separately for Phases 1–3.

To visually assess rater agreement, Figure 5.1 presents a comparison of the GOALS total scores assigned by the two expert raters for each cadaveric nephrectomy trial across Phases 1–3. Each point corresponds to a single trial, with the two raters’ scores plotted on the horizontal and vertical axes. Most points lie close to the identity line ($y = x$), indicating a high degree of agreement and suggesting that both raters applied the GOALS criteria in a consistent manner.

5.1.3 Results of inter-rater reliability

Across all three phases, the two raters showed good to excellent agreement for the GOALS total scores. Spearman’s ρ for the total score ranged from approximately 0.63 to 0.78 across Phase 1–3, indicating a strong monotonic relationship between the two raters. Domain-level coefficients were generally in the moderate-to-good range ($\rho \approx 0.36$ – 0.76). Agreement was highest for efficiency and tissue handling (E and T), while depth perception and bimanual dexterity (D and B) showed relatively lower, but still moderate, correlations, particularly in Phase 3.

Given this level of reliability, the mean of the two raters’ scores was used as the final GOALS domain and total scores for all subsequent analyses.

5.2 Regression model construction for GOALS score prediction

To develop regression models capable of predicting GOALS scores from motion-derived features, three types of regression algorithms were employed: Ridge regression, partial least squares regression (PLS), and PCA-based support vector regression (PCA-SVR). The target variable was defined as the mean GOALS total score assigned independently by the two expert reviewers, as described in Section 5.1.

5.2.1 Regression algorithms

Ridge regression. Ridge regression introduces an L2 regularization term to the loss function and provides stable coefficient estimation in the presence of multicollinearity among kinematic features. This property is particularly advantageous given the highly correlated nature of surgical motion metrics.

Partial least squares (PLS). PLS regression extracts latent components that maximize the covariance between predictors and the target variable. It is suitable for datasets with strong collinearity and provides interpretable low-dimensional representations of surgeons' motion characteristics.

PCA-SVR. This model combines robust normalization, PCA-based dimensionality reduction, and support vector regression (SVR with an RBF kernel). PCA transformation was fitted using only the training data and then applied to the validation/test data to avoid information leakage. Hyperparameters (number of components, C , γ , and ϵ) were determined through an inner-loop grid search.

5.2.2 Data preprocessing

All regression models underwent identical preprocessing steps: robust normalization using the median and normalized IQR (NIQR), selection of significant features based on Spearman correlation ($|r| > 0.4$), and strict separation of training and test sets. PCA was performed exclusively on the training folds to avoid information leakage.

5.2.3 Overall Regression-Model Framework

Figure 5.2 provides an overview of the analytical pipeline used to construct and evaluate regression models for GOALS score prediction.

The upper panel illustrates the preprocessing pathway, including: (i) motion-feature extraction from cadaveric trials, (ii) robust normalization based on median and NIQR scaling, (iii) removal of outliers and non-informative features, and (iv) filtering of statistically significant variables via Spearman tests. This process ensured that only stable and meaningful kinematic indices were retained for subsequent modeling.

The lower panel summarizes the regression architecture. Three algorithms were implemented—PCA-SVR, Ridge regression, and Partial Least Squares (PLS). Nested cross-validation was employed to avoid optimistic bias in model selection and to provide an unbiased estimate of generalization performance. Model accuracy was quantified using mean absolute error (MAE), and visual assessment was performed using scatter plots of predicted versus actual GOALS scores for each phase.

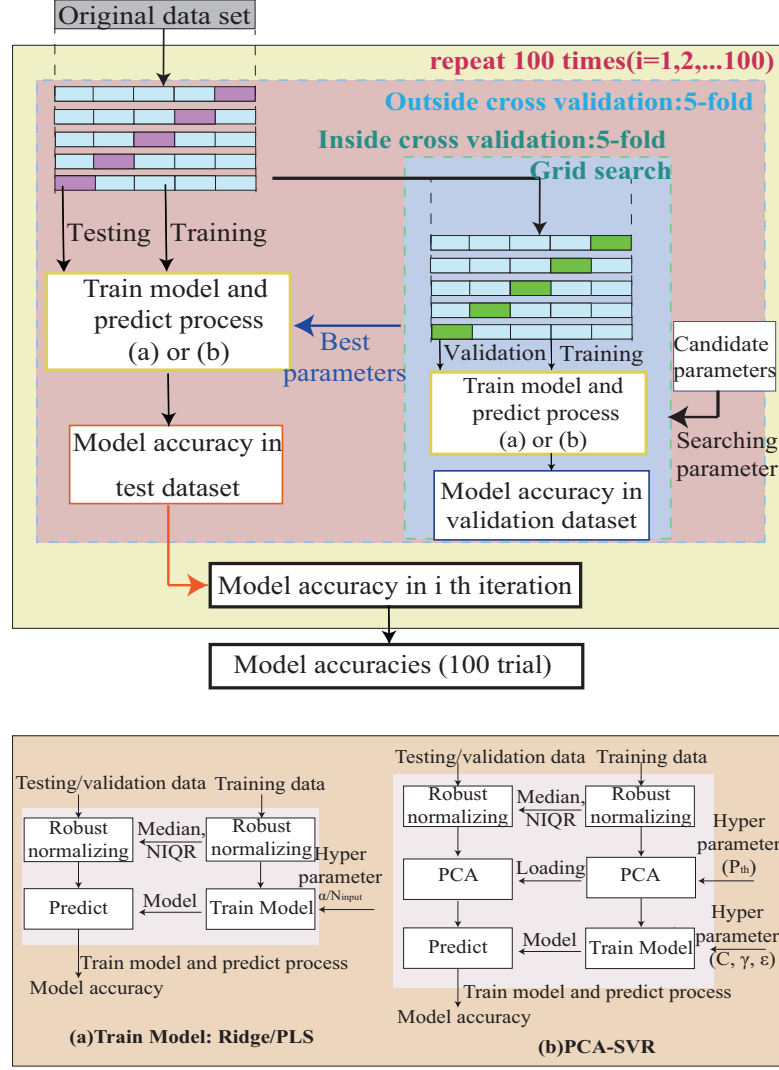


Figure 5.2: Overall regression-model pipeline. Feature preprocessing, robust normalization, and selection of significant kinematic indices.

This integrated framework enabled both interpretable and quantitatively stable prediction of expert-derived GOALS ratings, forming the foundation for the results presented in Section 5.3.

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (5.1)$$

5.3 Regression analysis for GOALS scores in Phase 1

This section presents the regression results for GOALS score prediction using the framework described in Sections 5.1 and 5.2.

5.3.1 Model accuracy (MAE)

Figure 5.4 summarizes the distribution of mean absolute errors (MAE) for the three models across the 100 repeated validation cycles. Ridge regression consistently achieved the lowest MAE for most of the GOALS items, followed by PLS. PCA-SVR showed the highest variability and generally larger errors.

The statistical comparison using the Kruskal–Wallis test and post-hoc pairwise tests is summarized in Table 5.1. Ridge regression significantly outperformed PCA-SVR for all five GOALS items and the total score. Ridge also outperformed PLS in depth perception, bimanual dexterity, and total score.

Table 5.1: Comparison of MAE among PCA-SVR, PLS, and Ridge regression models in Phase 1. Values indicate median MAE. Statistical comparisons were performed using the Kruskal–Wallis test with post-hoc pairwise testing.

Item	Median MAE				p Value			
	PCA-SVR	PLS	Ridge	Overall	SVR-PLS	SVR-Ridge	Ridge-PLS	
(a) Autonomy	0.66	0.60	0.59	< 0.0001	< 0.0001	< 0.0001	0.0406	
(b) Depth Perception	0.76	0.72	0.74	< 0.0001	< 0.0001	< 0.0001	< 0.0001	
(c) Bimanual Dexterity	0.62	0.61	0.57	< 0.0001	< 0.0001	0.0026	< 0.0001	
(d) Efficiency	0.57	0.57	0.54	< 0.0001	< 0.0001	0.8815	< 0.0001	
(e) Tissue Handling	0.63	0.61	0.63	< 0.0001	0.3582	< 0.0001	< 0.0001	
(f) Total Score	3.28	2.71	2.66	< 0.0001	< 0.0001	< 0.0001	< 0.0001	

5.3.2 Comparison of predicted and actual scores

Figure 5.3 illustrates the relationship between actual GOALS scores and Ridge regression predictions for Phase 1. Each point represents the median across 100 predictions for a single trial, with error bars showing the interquartile range. The predictions follow the identity line closely, especially in the mid-range scores (10–20), indicating stable calibration of the model.

5.3.3 Item-wise prediction performance

To examine model behavior for each GOALS domain, boxplots of predicted scores for autonomy, depth perception, bimanual dexterity, efficiency, tissue handling, and total score are provided in Figures 5.4. Across all items, Ridge regression demonstrates narrower interquartile ranges and fewer extreme deviations compared to PCA-SVR.

5.3.4 Summary of Phase 1

- Ridge regression achieved the highest accuracy across all GOALS items.
- PCA-SVR exhibited the greatest error and instability.
- PLS offered intermediate performance, with accuracy close to Ridge for some items.
- Prediction errors were smallest in the efficiency and bimanual dexterity domains.
- The strong correspondence between predicted and actual scores suggests that motion-capture–derived kinematic features are highly informative in the early operative phase.

Overall, Phase 1 results demonstrate that GOALS scores can be estimated with substantial reliability using motion-based features alone, highlighting the feasibility of automated skill assessment in laparoscopic training.

5.3.5 Regression performance in Phase 2

Table 5.2 summarizes the MAE values of the three regression models for each GOALS item and the total score in Phase 2. The median MAE values of PCA-SVR, PLS, and Ridge were 0.60, 0.62, and 0.61 for Autonomy; 0.59, 0.63, and 0.61 for Depth Perception; 0.72, 0.63, and 0.63 for Bimanual Dexterity; 0.57, 0.60, and 0.54 for Efficiency; and 0.57, 0.58, and 0.59 for Tissue Handling, respectively. The corresponding median MAE values for the GOALS total score were 3.05, 3.07, and 2.99. Kruskal–Wallis tests showed significant differences among the three models for all items and the total score, and post-hoc comparisons confirmed that PCA-SVR yielded significantly larger errors than Ridge, whereas PLS showed intermediate performance.

Figure 5.5 illustrates the relationship between predicted and actual GOALS total scores obtained with the Ridge model. Most boxplots are located close to the identity line, indicating good agreement over the entire range of scores. Similar to Phase 1, a slight underestimation was observed for the highest-scoring trials, whereas no systematic bias was evident for low- and medium-scoring procedures.

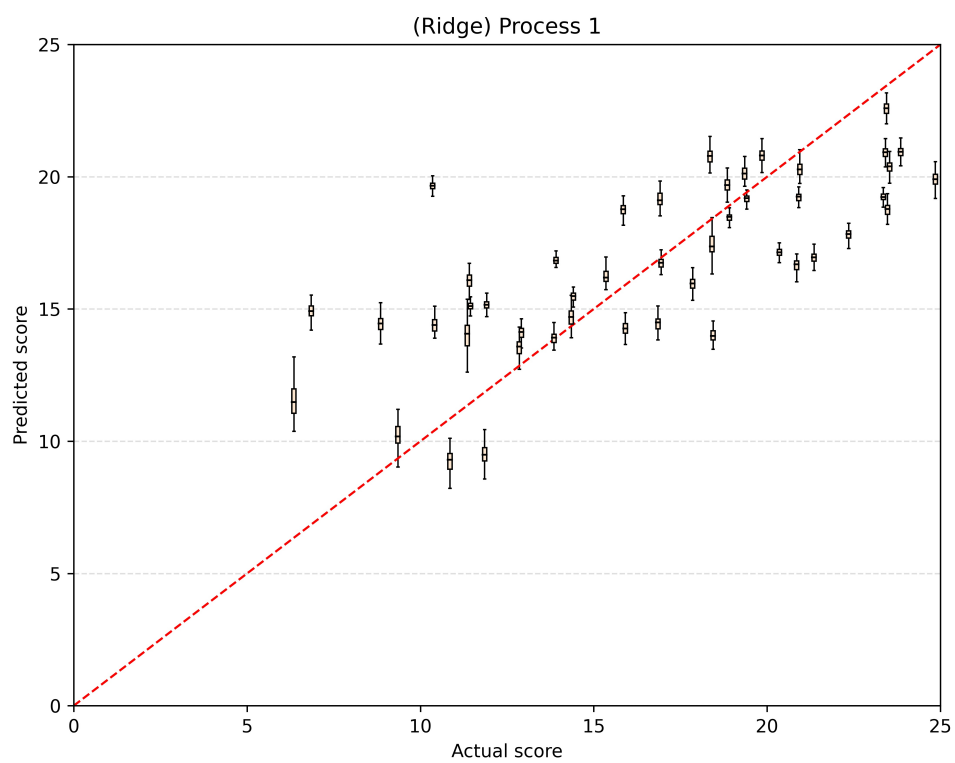


Figure 5.3: Predicted versus actual GOALS total scores for Ridge regression (Phase 1). The dashed red line represents the identity line.

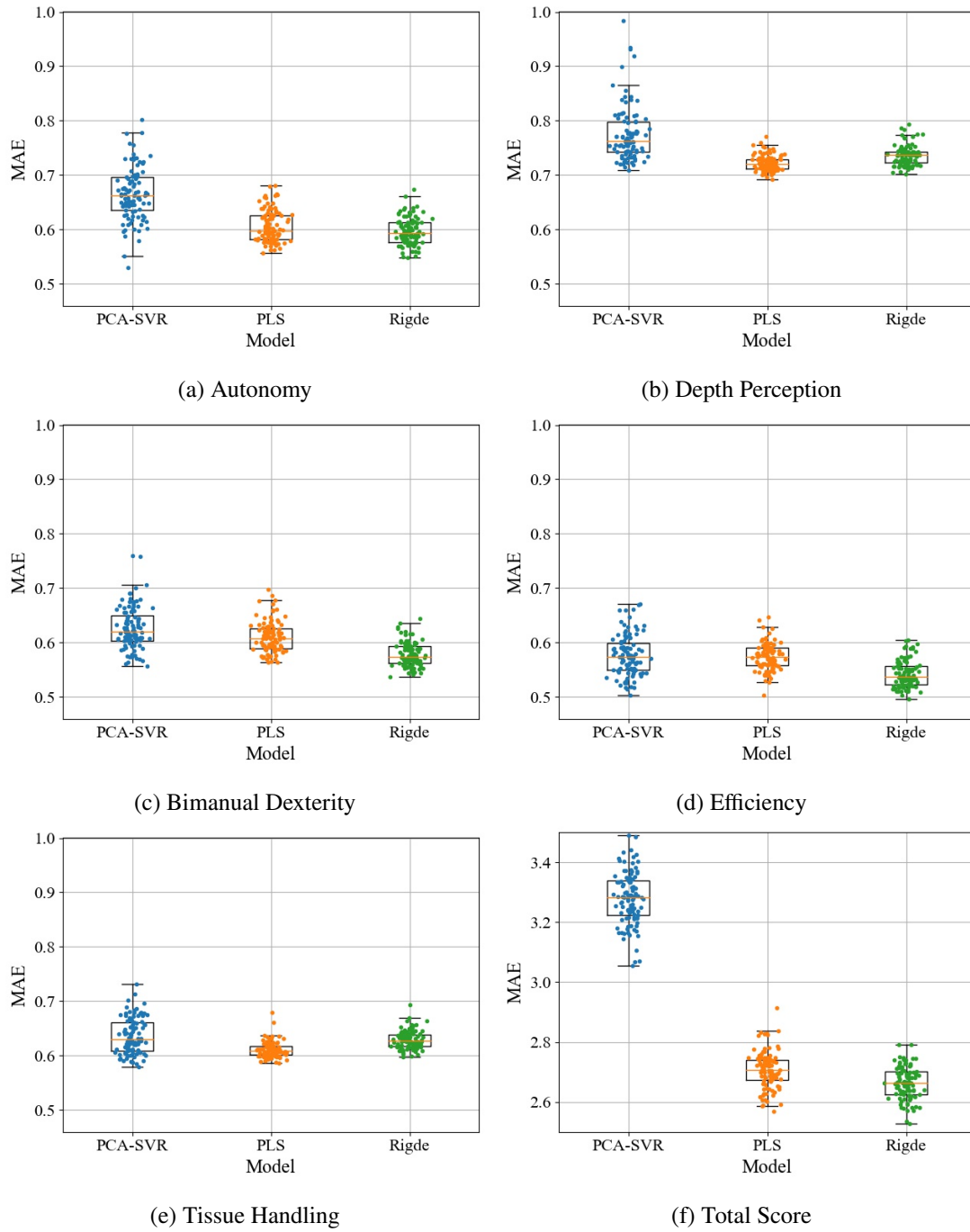


Figure 5.4: Prediction performance of six GOALS items using Ridge, PLS, and PCA-SVR models (Phase 1).

Table 5.2: Comparison of MAE among PCA-SVR, PLS, and Ridge regression models in Phase 2. Values indicate median MAE across all cases. Statistical comparisons were performed using the Kruskal–Wallis test with post-hoc pairwise testing.

Item	Median MAE				p Value			
	PCA-SVR	PLS	Ridge	Overall	SVR-PLS	SVR-Ridge	Ridge-PLS	
(a) Autonomy	0.60	0.62	0.61	< 0.0001	0.0007	0.0130	< 0.0001	
(b) Depth Perception	0.59	0.63	0.61	< 0.0001	< 0.0001	< 0.0001	< 0.0001	
(c) Bimanual Dexterity	0.72	0.63	0.63	< 0.0001	< 0.0001	< 0.0001	0.1979	
(d) Efficiency	0.57	0.60	0.54	< 0.0001	< 0.0001	< 0.0001	0.2154	
(e) Tissue Handling	0.57	0.58	0.59	< 0.0001	0.0207	0.0003	0.0183	
(f) Total Score	3.05	3.07	2.99	< 0.0001	0.0173	< 0.0001	< 0.0001	

5.3.6 Regression performance in Phase 3

In Phase 3, the prediction errors further decreased compared with Phases 1 and 2. The median MAE values of PCA-SVR, PLS, and Ridge were 0.63, 0.62, and 0.59 for Autonomy; 0.53, 0.52, and 0.51 for Depth Perception; 0.56, 0.52, and 0.51 for Bimanual Dexterity; 0.63, 0.56, and 0.55 for Efficiency; and 0.53, 0.51, and 0.49 for Tissue Handling, respectively. The corresponding median MAE values for the GOALS total score were 2.90, 2.72, and 2.66 (Table 5.3), which were the lowest among the three phases.

Kruskal–Wallis tests again revealed significant differences among the three models for all items and the total score. Post-hoc analyses showed that Ridge and PLS consistently outperformed PCA-SVR. Although the difference between Ridge and PLS was smaller than that between Ridge and PCA-SVR, Ridge still achieved the smallest MAE for the total score.

Figure 5.7 displays the relationship between predicted and actual GOALS total scores in Phase 3 for the Ridge model. Compared with Phases 1 and 2, the boxplots are more tightly clustered around the identity line, particularly in the mid-to-high score range, indicating improved precision and reduced variability of the predictions. Only a few high-scoring trials showed slight underestimation.

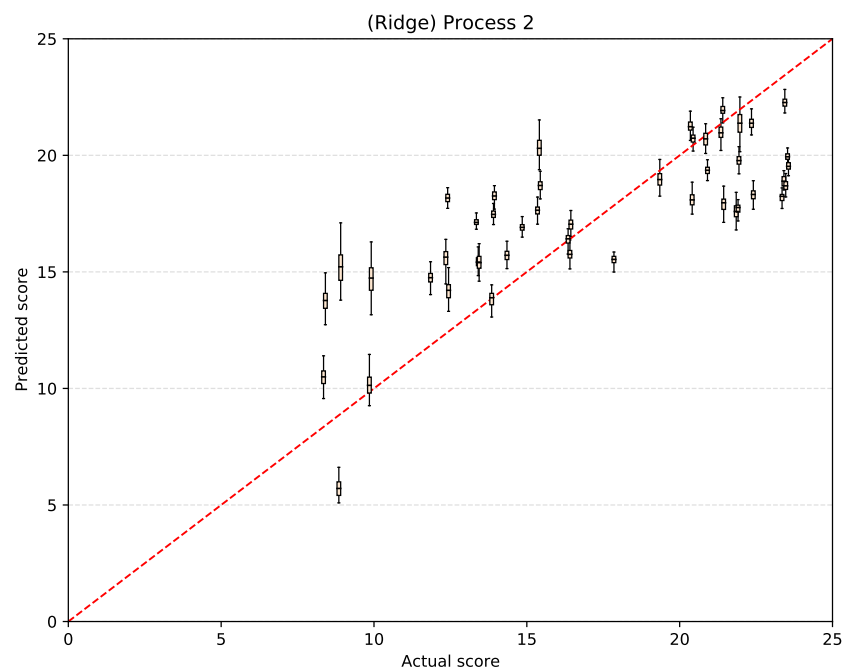


Figure 5.5: Relationship between predicted and actual GOALS total scores for the Ridge regression model in Phase 2. Each boxplot shows the distribution of predicted scores over the repeated cross-validation runs for a given trial; the red dashed line indicates the identity line.

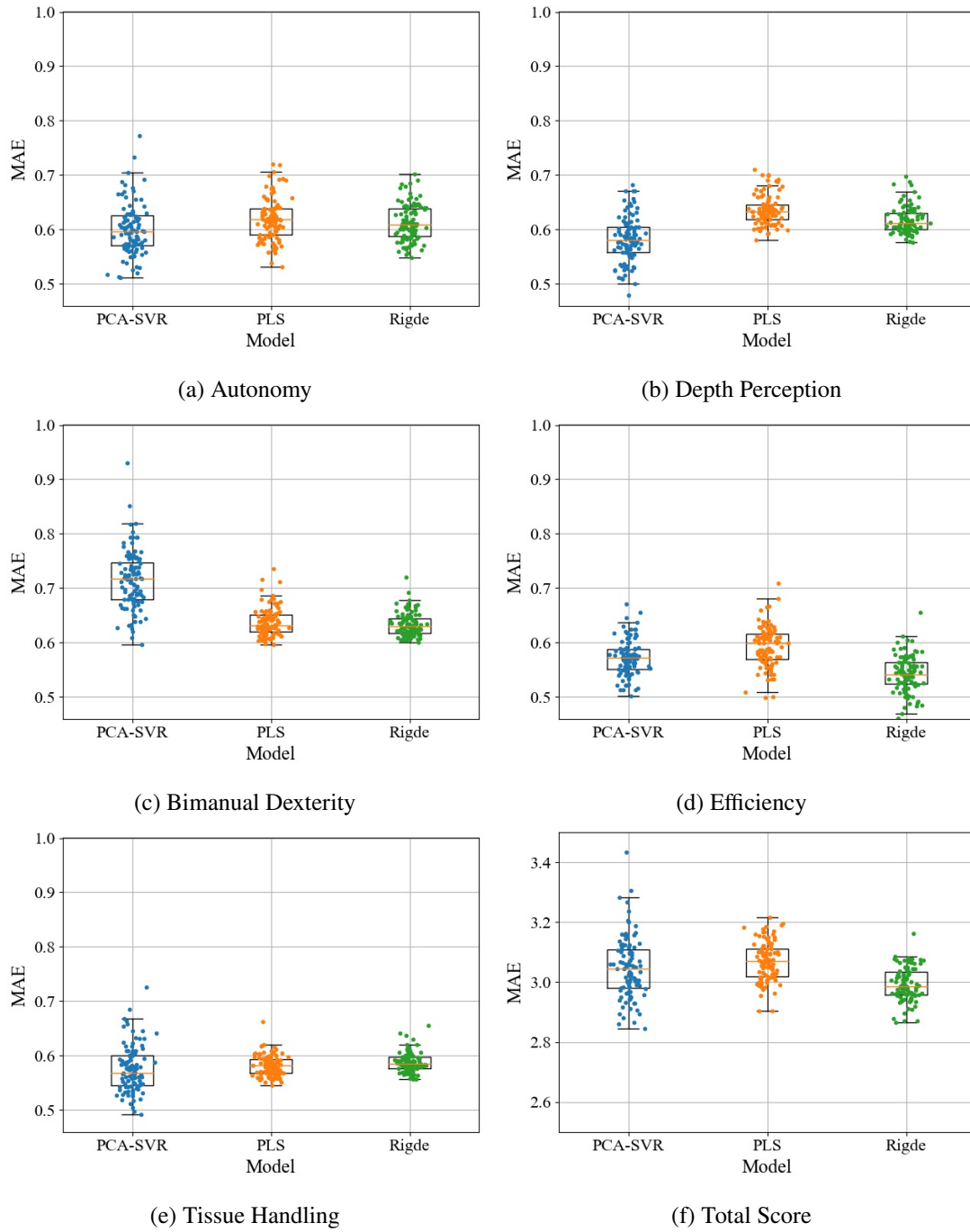


Figure 5.6: Prediction performance of GOALS items using Ridge, PLS, and PCA-SVR models (Phase 2).

Table 5.3: Comparison of MAE among PCA-SVR, PLS, and Ridge regression models in Phase 3. Values indicate median MAE across all cases. Statistical comparisons were performed using the Kruskal–Wallis test with post-hoc pairwise testing.

Item	Median MAE				p Value			
	PCA-SVR	PLS	Ridge	Overall	SVR-PLS	SVR-Ridge	Ridge-PLS	
(a) Autonomy	0.63	0.62	0.59	< 0.0001	0.0446	< 0.0001	< 0.0001	
(b) Depth Perception	0.53	0.52	0.51	< 0.0001	0.0063	< 0.0001	0.0181	
(c) Bimanual Dexterity	0.56	0.52	0.51	< 0.0001	< 0.0001	< 0.0001	< 0.0001	
(d) Efficiency	0.63	0.56	0.55	< 0.0001	< 0.0001	< 0.0001	0.2331	
(e) Tissue Handling	0.53	0.51	0.49	< 0.0001	< 0.0001	< 0.0001	0.0016	
(f) Total Score	2.90	2.72	2.66	< 0.0001	< 0.0001	< 0.0001	< 0.0001	

Table 5.4: Spearman correlation (ρ) between expert scores and Ridge regression predictions.

Phase	Mean-prediction ρ	Trial range	Trial mean ρ
Phase 1	0.756	0.727–0.775	0.750
Phase 2	0.780	0.734–0.804	0.768
Phase 3	0.682	0.624–0.729	0.675

5.3.7 Correlation between Expert Ratings and Model Predictions

To further assess how well the regression model reproduced expert-assigned GOALS scores, we computed Spearman’s rank correlation coefficient (ρ) between the Ridge regression predictions and the expert labels. Because the model was trained 100 times with different random seeds, we computed the correlation for each trial separately as well as for the mean-prediction (the average of 100 predictions for each subject).

The regression performance across the three phases showed consistent but phase-dependent differences. For Phase 1, the Ridge model demonstrated a strong and stable correlation with expert ratings, with a mean-prediction correlation of $\rho = 0.756$ ($p = 1.21 \times 10^{-9}$). Across 100 trials, Spearman’s ρ ranged from 0.727 to 0.775 (mean = 0.750), indicating stable performance.

In Phase 2, the mean-prediction correlation slightly increased to $\rho = 0.780$ ($p = 1.74 \times 10^{-10}$), with trial-wise correlations ranging from 0.734 to 0.804 (mean = 0.768). However, this phase exhibited greater variability across trials compared with Phase 1.

In Phase 3, the correlation decreased to $\rho = 0.682$ ($p = 1.80 \times 10^{-7}$), with a wider spread of trial-wise correlations (0.624–0.729, mean = 0.675). This reduction likely reflects increased variability and task complexity in the later procedural stage.

Overall, these results indicate that while correlation remains relatively high across all phases, later stages of the procedure tend to exhibit increased variability and reduced predictive stability.

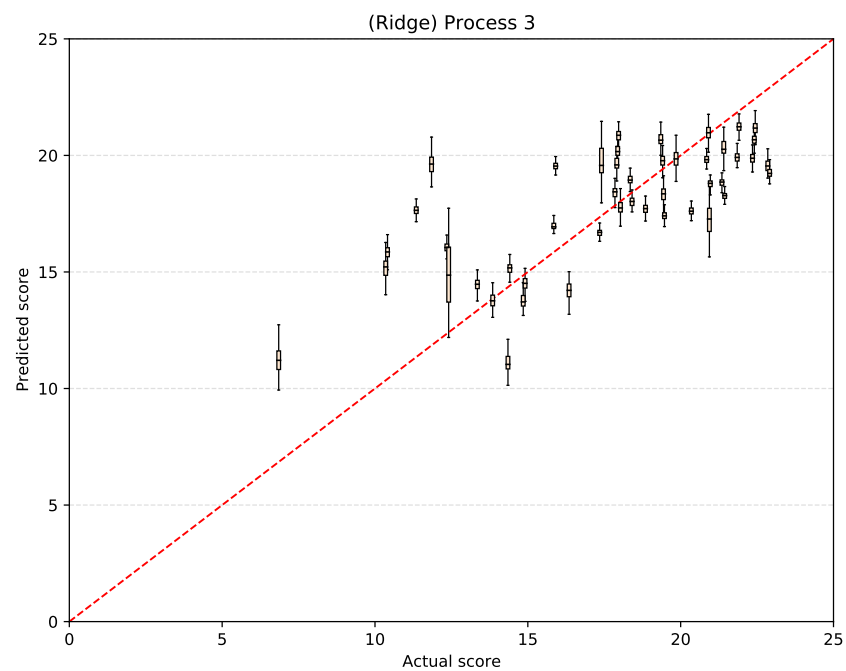


Figure 5.7: Relationship between predicted and actual GOALS total scores for the Ridge regression model in Phase 3. Compared with Phases 1 and 2, the boxplots are more tightly clustered around the identity line, indicating improved prediction accuracy.

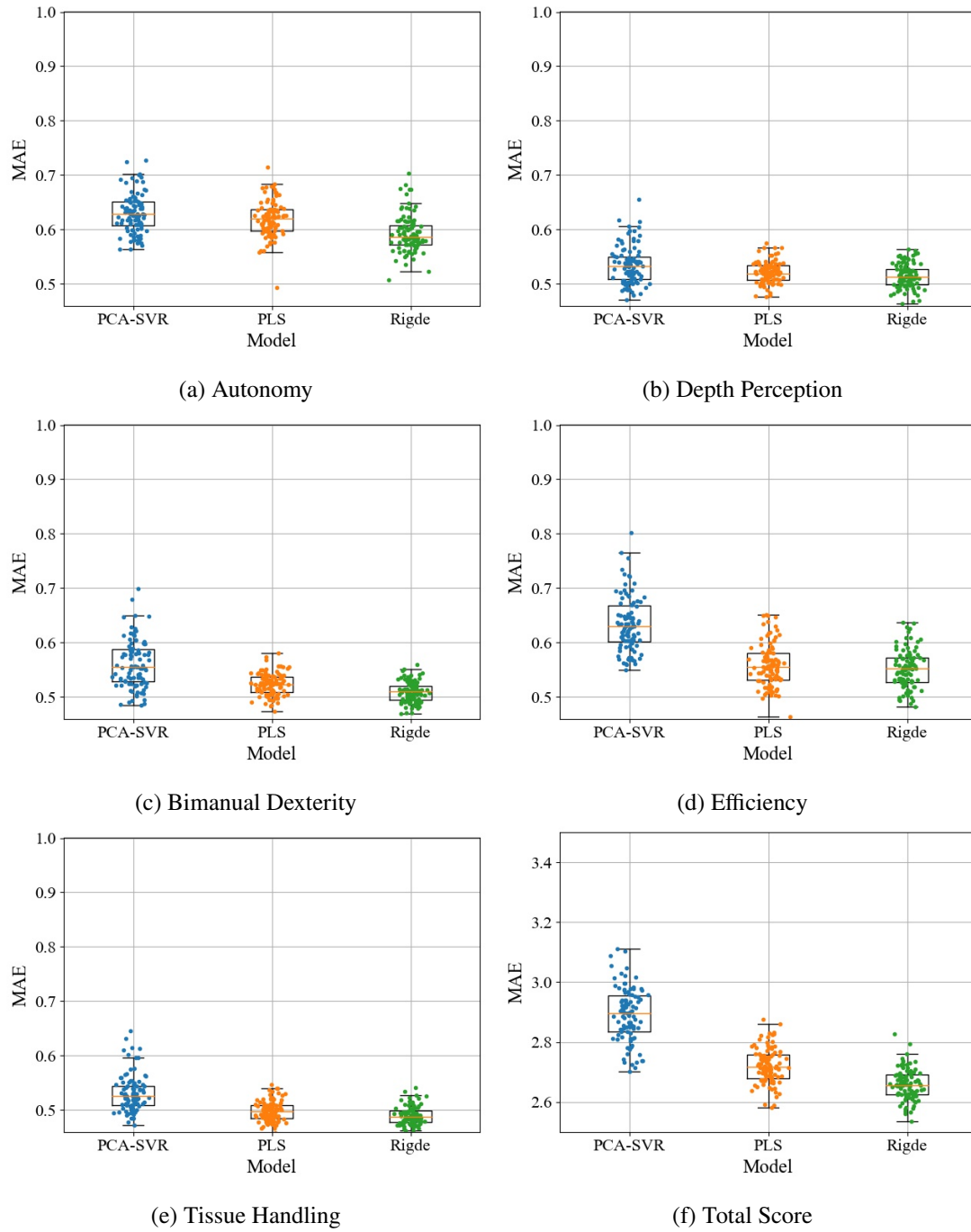


Figure 5.8: Prediction performance of GOALS items using Ridge, PLS, and PCA-SVR models (Phase 3).

5.4 Summary

This chapter developed regression models to estimate expert-assigned GOALS scores from motion-derived quantitative features. Ridge regression, PCA-SVR, and PLS models were evaluated across individual GOALS domains as well as total scores. Ridge regression consistently achieved the lowest MAE among the three models across all phases.

Correlation analysis further demonstrated that the Ridge model reproduced the expert ranking with strong monotonic consistency. Phase 1 showed the most stable performance (mean $\rho = 0.756$), followed by Phase 2 (mean $\rho = 0.780$ but with larger variability). Phase 3 exhibited the lowest consistency (mean $\rho = 0.682$), likely due to its higher task complexity and greater inter-subject variability.

Taken together, these findings demonstrate that (i) motion-derived quantitative metrics can reconstruct expert GOALS evaluation with high fidelity, (ii) model performance varies across procedural phases, and (iii) early procedural segments (Phase 1) provide the most reliable basis for automated skill assessment.

Chapter 6

Educational Applications and Integration Into Tele-mentoring Systems

This chapter examines educational applications of the motion-analysis framework developed in this dissertation. Three complementary environments were explored: (1) real laparoscopic surgery using a sterilizable motion-capture configuration; (2) porcine wet-lab training with on-site expert scoring; and (3) early-stage tele-mentoring experiments integrating network-latency evaluation. Together, these efforts illustrate the feasibility and current limitations of applying objective motion analysis in realistic operative contexts and provide insights into future directions for surgical education and remote guidance systems.

6.1 Real Surgery Experiments

6.1.1 Measurement System and Rigid-Body Marker Attachment

To evaluate the feasibility of instrument motion tracking in real clinical settings, we conducted pilot measurements during multiple clinical laparoscopic procedures, including donor nephrectomy. All real-surgery measurements were performed under a separate ethics approval from the Hokkaido University Hospital Institutional Review Board (IRB no. 024-0363).

A sterilizable rigid-body marker set was designed using autoclavable guide resin and attached to the surgical instruments with sterile rubber bands. Each instrument had a unique marker geometry to ensure correct identification and stable tracking. The motion-capture system (OptiTrack V120:Trio) recorded 3D trajectories at a sampling rate of 120 Hz. In addition to the instrument markers, anatomical reference points—including trocar locations and patient landmarks (shoulder, hip, and knee)—were recorded to define the region of interest (ROI).

6.1.2 Trajectory Recording and Preprocessing

Six surgical cases with complete tracking data were analyzed as a pilot feasibility series. Representative 3D trajectories of three instruments—grasping forceps (Karl-Storz), LigaSure, and scissor forceps (Olympus)—are shown in Fig. 6.2. The raw trajectory data were first filtered to remove points located outside the predefined ROI. Missing segments caused by temporary occlusion were interpolated to construct continuous trajectories suitable for subsequent quantitative analysis.

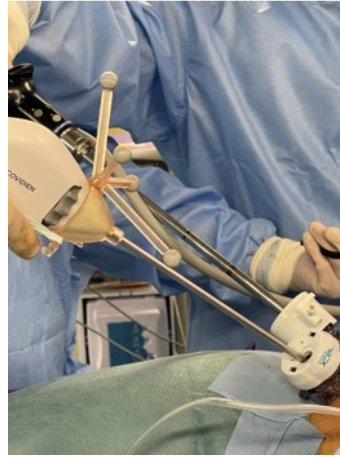


Figure 6.1: Sterilized rigid-body marker attached to a laparoscopic instrument during a real donor nephrectomy. The rigid body was fabricated using autoclavable guide resin and attached using sterile bands.

Table 6.1: Tracking success rates across six completed surgical cases.

Case	Grasping Forceps (Karl-Storz)	LigaSure	Scissor Forceps (Olympus)
1	92.30%	88.17%	—
2	93.19%	88.32%	90.96%
3	89.18%	86.10%	—
4	90.75%	94.10%	84.48%
5	90.55%	90.33%	90.28%
6	95.78%	98.99%	92.27%

To characterize the temporal dynamics of the surgical workflow, active usage periods were extracted for each instrument. Figure 6.3 shows a representative timeline of instrument activation over the duration of a full nephrectomy, illustrating clear task segmentation and switching patterns.

6.1.3 Tracking Success Rate Across the Six Surgical Cases

Table 6.1 shows the successful tracking rate for each instrument across the six fully analyzed cases. The success rate was defined as the proportion of validly reconstructed frames among all captured frames within the ROI. In most cases, the tracking success rate exceeded 90%; reductions were mainly observed during deep pelvic phases or when instruments overlapped.

6.1.4 Surgeon Questionnaire on Marker Interference

To assess the usability of the rigid-body markers during clinical surgery, participating surgeons evaluated whether the attached markers interfered with instrument manipulation. A five-point Likert scale was used (1 = strongly disagree, 5 = strongly agree), with 0 indicating that the instrument was not used. The distribution of responses is shown in Fig. 6.4, and the average score for each instrument is summarized in Table 6.2.

Overall, the majority of responses clustered around scores of 2–3, indicating that surgeons

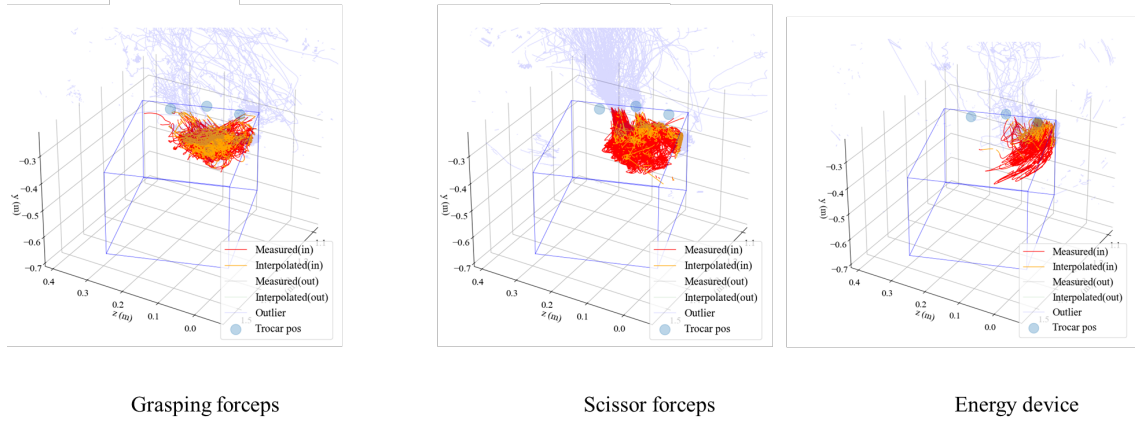


Figure 6.2: Example 3D trajectories recorded during laparoscopic donor nephrectomy (Case 4). Left: grasping forceps; center: energy device; right: scissor forceps. Red indicates measured points inside the ROI, orange indicates interpolated points, and gray denotes out-of-ROI samples.

Table 6.2: Average perceived-interference score for each instrument.

Instrument	Mean Score
Scissor Forceps (Olympus)	2.59
Grasping Forceps (Karl-Storz)	2.63
LigaSure	2.42

generally did not perceive the markers as obstructive in routine manipulation. Only a small number of higher scores were reported, typically during instrument crossings or deep pelvic manipulation. These pilot results suggest that sterilizable rigid-body markers can be incorporated into selected clinical laparoscopic procedures with acceptable workflow impact, although further refinement is needed before widespread adoption.

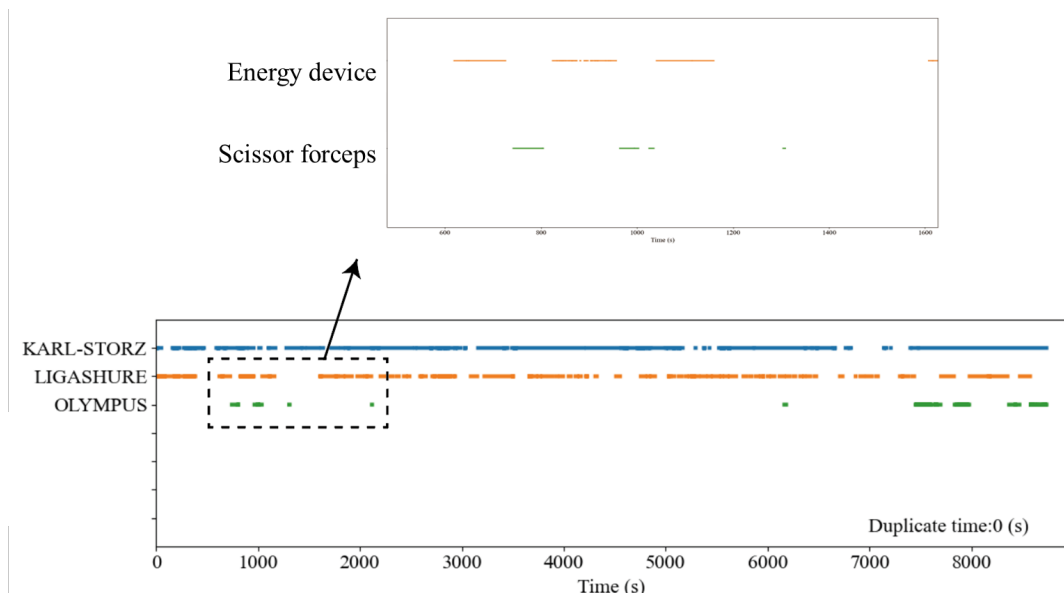


Figure 6.3: Representative instrument usage timeline for a full nephrectomy. Usage intervals are shown for grasping forceps (Karl-Storz), LigaSure, and scissor forceps (Olympus). Duplicate usage time was zero, indicating consistent identification of each instrument.

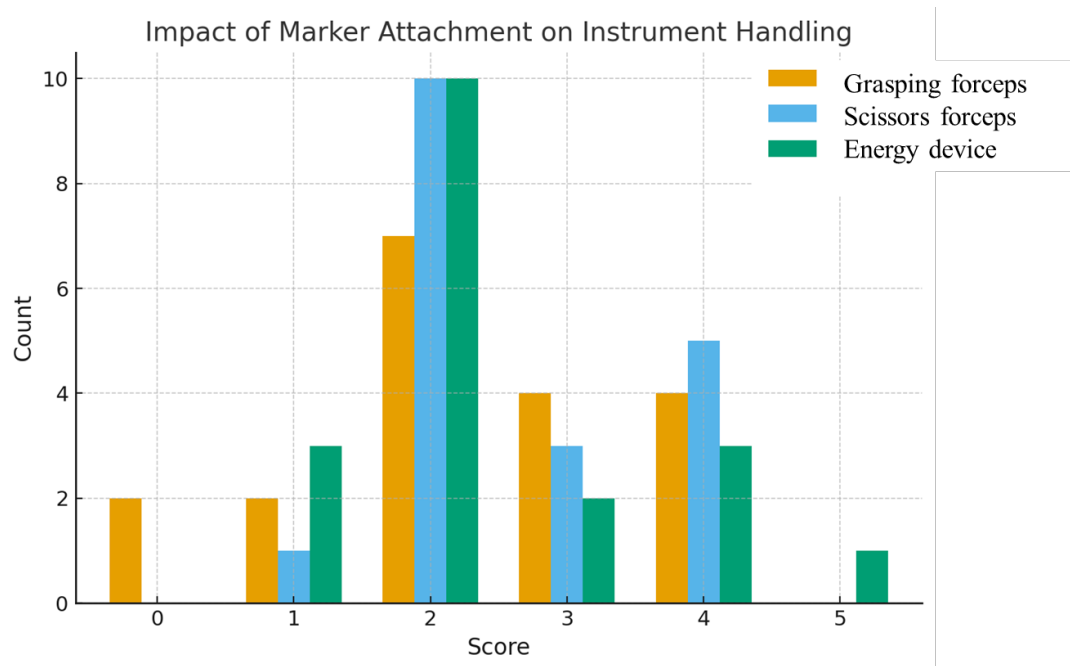


Figure 6.4: Survey results assessing perceived interference caused by the instrument-mounted rigid-body markers. Scores were evaluated on a five-point Likert scale (1 = no interference; 5 = strong interference).

6.2 Wet-Lab (Porcine) Experiments

6.2.1 Overview of the Wet-Lab Environment

To evaluate the applicability of the proposed quantitative metrics in a realistic tissue environment, wet-lab experiments were conducted using porcine wet-lab models. Such models are frequently used in laparoscopic education because they provide anatomically realistic tissue-handling properties similar to those encountered in clinical surgery [16].

These experiments reproduced essential components of laparoscopic procedures, including trocar placement, manipulation of soft tissue, and vessel exposure. The same motion-tracking system and sterilizable rigid-body markers introduced in Chapters 2 and 6 were used. Since the full system configuration has already been presented in Chapter 2, it is not repeated here. Tracking stability in the wet-lab environment was generally high due to controlled lighting conditions and fewer visual occlusions compared with real clinical settings.

6.2.2 On-Site Skill Assessment by Expert Surgeons

During each wet-lab session, experienced laparoscopic surgeons provided immediate performance assessments based on the core domains of the Global Operative Assessment of Laparoscopic Skills (GOALS). Similar wet-lab scoring approaches have been successfully used to validate motion-based skill metrics in previous work [16].

The direct on-site scoring minimized recall bias and enabled reliable comparison between subjective expert evaluation and objective quantitative metrics extracted from instrument trajectories. The tissue-handling domain was included in Task 1 (mucosal dissection) but was not evaluated in Task 2 (renal suturing), as the suturing task does not involve tissue-handling assessment.

6.2.3 Integrated Feedback Interface

To support structured feedback for trainees, an integrated feedback interface developed by a previous member of our group was used. This interface combines: (1) GOALS domain scores; (2) PCA-based visualization comparing the participant with the cohort median; (3) hand-symmetry metrics such as velocity differences between dominant and nondominant hands; (4) expert textual feedback; and (5) quantitative indices extracted from motion data for both right-hand and left-hand needle holders.

An example of this interface is shown in Fig. 6.5. This tool allowed participants to review their performance immediately after completing the wet-lab task, providing a standardized and data-driven training feedback mechanism.

6.2.4 Summary

The wet-lab experiments demonstrated that: (1) the proposed measurement and evaluation system is technically feasible under realistic porcine tissue conditions; (2) expert on-site scoring can be consistently collected and aligned with quantitative metrics; and (3) the integrated feedback interface provides interpretable, structured, and immediate performance feedback.

These findings support the use of wet-lab experiments as an intermediate validation step between controlled dry-lab tasks and real laparoscopic surgery, consistent with prior work showing the value of high-fidelity tissue models in skill quantification [16].

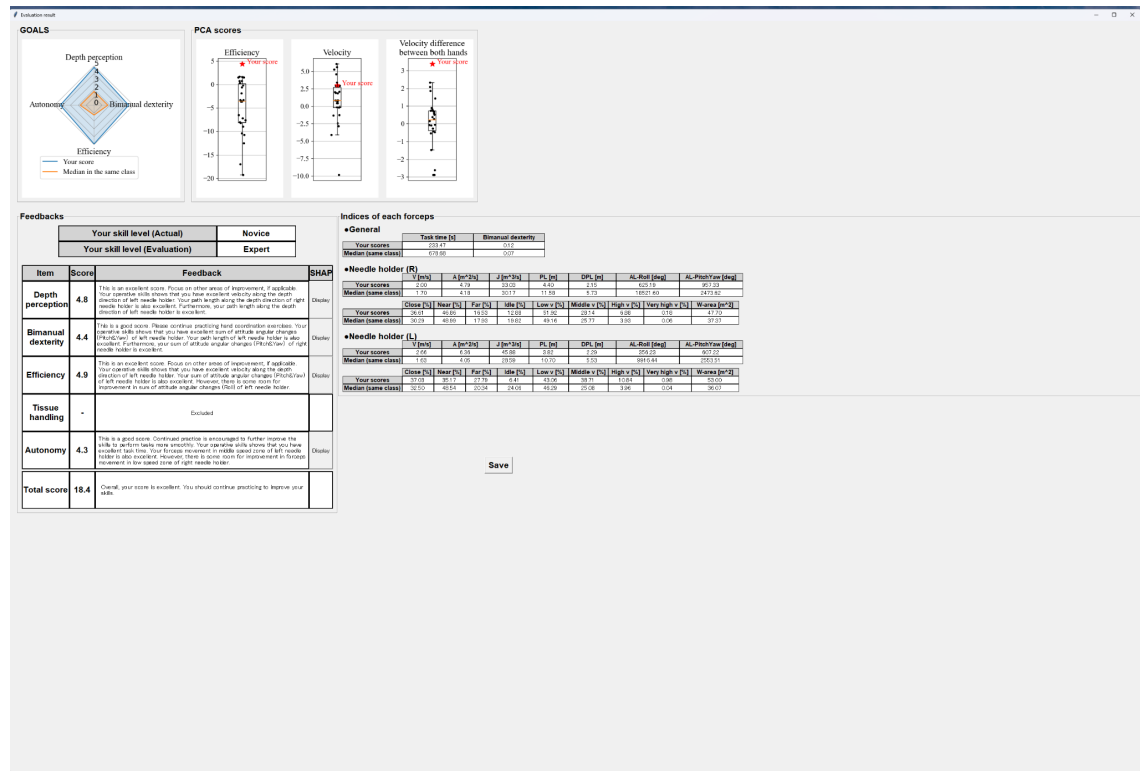


Figure 6.5: Example of the integrated feedback interface used in the wet-lab porcine experiments. The interface displays GOALS assessment results, PCA score comparison, velocity symmetry metrics, expert feedback, and quantitative indices extracted for each instrument.

6.3 Remote Surgery Experiments

6.3.1 Motivation and Background

Remote surgical assistance has become increasingly relevant due to the demand for expert guidance in regions with limited access to experienced surgeons. Modern telecommunication systems enable real-time transmission of surgical video, allowing experts to provide instructions even when physically distant. To explore basic feasibility in laparoscopic environments, we conducted pilot experiments evaluating communication delay, video quality, and system usability during remote surgical observation.

These remote experiments were conducted as a separate exploratory study focusing on telecommunication performance. No motion-capture data were recorded during these sessions, and the system was not used for intraoperative clinical decision-making.

6.3.2 Remote Multi-View Surgical Streaming System

A multi-channel remote-streaming system (Zao Cloud View) was used to transmit real-time laparoscopic and room-view video from the operating site to mentors located off-site. The system simultaneously streamed (1) laparoscopic view, (2) overhead robotic camera view, and (3) room camera view. An example of the interface used during the experiment is shown in Fig. 6.6.

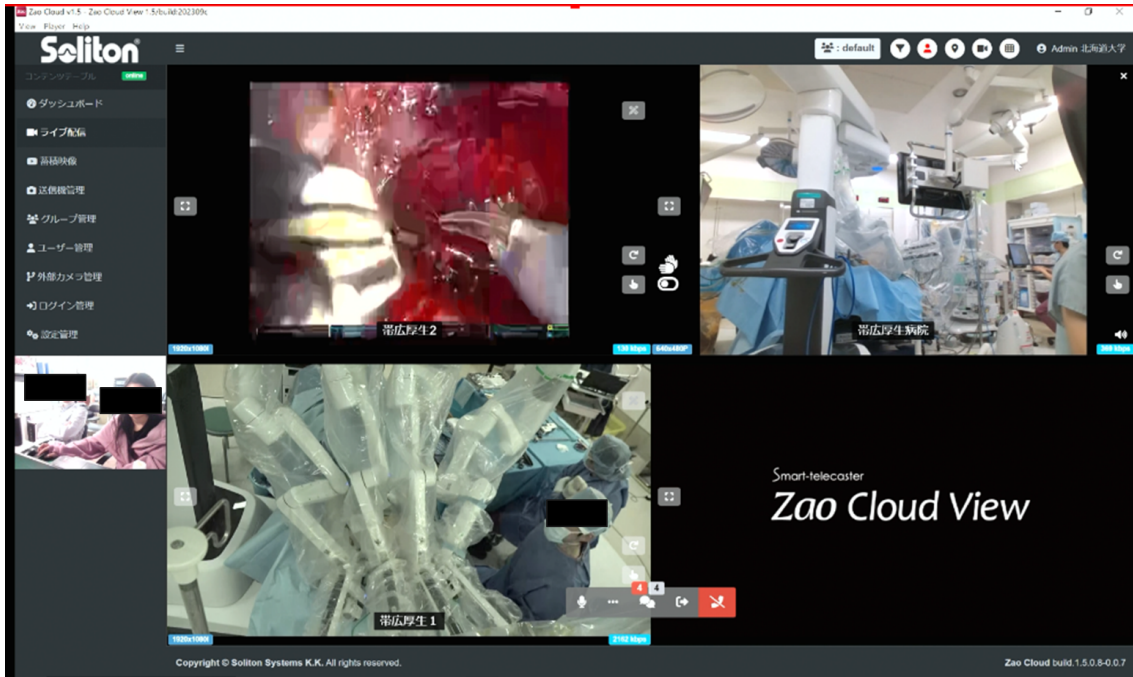


Figure 6.6: Example of the remote surgical streaming interface used in the pilot experiment. Multiple video sources—including laparoscopic view, robotic surgical view, and room-view camera—were transmitted simultaneously using the Zao Cloud View system®. Video compression artifacts and temporary pixelation were observed during network congestion, highlighting the impact of bandwidth and latency on remote surgical observation.

The interface allowed the remote expert to observe multiple video streams in real time and switch between different camera views. Furthermore, the system provided a real-time annotation function, enabling the expert to highlight critical regions and visually support intraoperative guidance.

6.3.3 Measurement of Communication Latency

To evaluate communication performance, end-to-end latency was measured by embedding synchronized timestamps into the transmitted video stream. Under stable network conditions, round-trip latency typically ranged from 30–80 ms. However, occasional increases above 100 ms were observed during peak network load, which caused temporary pixelation and reduced visual clarity (as visible in Fig. 6.6).

These pilot results underscore the importance of stable bandwidth and optimized video encoding for effective remote surgical observation and mentoring.

6.3.4 Remote Surgical Guidance Trial

A pilot remote-mentoring session was performed in which an off-site expert provided guidance to an on-site surgeon. The remote expert used the streamed video to advise on trocar positioning, robotic arm configuration, and procedural steps. Despite occasional latency fluctuations, the expert reported that essential visual information could be followed and that communication delay

remained within a tolerable range for non-critical guidance tasks. This preliminary experience suggests that real-time remote mentoring is technically achievable under favorable network conditions, although further studies are required before routine clinical implementation.

6.3.5 Technical Considerations and Limitations

Several limitations were observed:

- **Latency variability:** Even modest increases in delay affected the perceived fluidity of dialog between on-site and off-site teams.
- **Compression artifacts:** As shown in the laparoscopic feed (Fig. 6.6), unstable bandwidth introduced noticeable pixelation, reducing the clarity of fine tissue structures.
- **Multichannel synchronization:** Slight time offsets occurred between different camera streams, which may influence future multi-view analysis and could impact user perception.

6.3.6 Future Directions

These pilot experiments demonstrated the technical feasibility of remote surgical observation and basic mentoring while identifying important technical challenges. Future development will focus on:

1. Implementing dedicated low-latency communication networks and quality-of-service control.
2. Exploring the integration of quantitative motion-based skill evaluation with remote mentoring platforms in a stepwise and carefully validated manner.
3. Coupling with robotic teleoperation systems for research on remote surgical interventions under strictly controlled conditions.
4. Applying the system to disaster-response or resource-limited scenarios where specialist access is particularly constrained.

In the context of this dissertation, remote experiments are considered a complementary feasibility study, pointing toward potential future integration with the motion-analysis framework rather than representing a completed, unified system.

Furthermore, future work will investigate the causal relationships between extracted motion features and surgical skill, with the aim of establishing a framework that can provide effective and actionable feedback to support skill improvement.

6.4 Summary

The additional experiments presented in this chapter evaluated the feasibility, robustness, and applicability of the proposed motion-analysis and skill-assessment framework in increasingly realistic and complex environments.

First, the real laparoscopic surgery measurements suggested that sterilizable rigid-body markers can be integrated into selected clinical workflows with acceptable interference. High tracking

success rates and stable motion-capture performance in the analyzed cases indicate that quantitative trajectory analysis is technically feasible in the operating room, although further optimization and larger-scale studies are required.

Second, the porcine wet-lab experiments served as an intermediate validation environment between controlled dry-lab tasks and real clinical surgery. Expert GOALS scores were consistently obtained, and the integrated feedback interface enabled structured and immediate performance review. These results highlight the value of combining subjective expert evaluation with objective motion-derived metrics for surgical education and skill development.

Finally, the remote surgery experiments explored the applicability of telecommunication platforms for real-time remote surgical support in an observational setting. Multi-view video streaming allowed off-site experts to observe the operative field, while latency assessments suggested that real-time guidance can be feasible under stable network conditions. Technical limitations such as video compression artifacts and latency fluctuations remain important barriers that must be addressed for future clinical deployment.

Overall, the findings from this chapter indicate that the proposed motion-tracking and quantitative evaluation system can be extended beyond controlled cadaveric environments to more realistic settings, and that tele-mentoring technologies provide a promising avenue for future integration. These results provide groundwork for subsequent development of comprehensive, data-driven surgical skill assessment and tele-mentoring platforms, which will require additional technical refinement and large-scale validation.

Chapter 7

Conclusions

7.1 Summary

This dissertation presented an integrated framework for quantitative evaluation of laparoscopic surgical skills using motion-capture data, machine-learning techniques, and multi-environment validation. The primary goal of this study was to establish objective, reproducible, and interpretable metrics for surgical skill assessment, thereby supporting more structured training and complementing conventional subjective scoring.

The proposed framework first introduced a complete preprocessing pipeline for reconstructing 3D instrument trajectories, validating measurement rates, and extracting motion features robustly across multiple environments. These features included spatial-temporal metrics such as velocity, acceleration, jerk, path length, working volume, idle and high-activity distributions, and angular motion characteristics. Dimensionality reduction using principal component analysis enabled visualization of skill-related structure, while machine-learning classifiers demonstrated reliable discrimination between novice, intermediate, and expert surgeons as well as between ESSQ-certified and non-certified surgeons.

Regression analyses based on GOALS scores further showed that objective motion features can approximate expert evaluations with reasonably high accuracy. The incorporation of SHAP-based model interpretation provided transparent linkage between motion patterns and scoring criteria, supporting pedagogical feedback and explainability in surgical education.

Beyond controlled cadaveric tasks, this study implemented instrument-motion tracking in high-fidelity environments, including porcine wet-lab procedures, Thiel-embalmed cadaver surgery, and real clinical laparoscopic procedures (including donor nephrectomy). These experiments evaluated the technical feasibility of sterilizable rigid-body markers, demonstrated high tracking success rates in representative cases, and suggested minimal interference with surgical manipulation as judged by surgeon questionnaires. An integrated feedback interface was also developed, enabling participants to receive immediate, structured, and data-driven performance summaries in the wet-lab setting.

Additionally, remote surgery experiments explored the applicability of modern telecommunication systems for surgical mentoring in an observational context. Multi-view streaming allowed off-site experts to observe operative scenes in real time, and latency measurements indicated that real-time communication is technically achievable under stable network conditions. These findings highlight the potential for combining objective motion analysis with tele-mentoring platforms in future work, while emphasizing that further research is necessary before clinical integration.

Taken together, the results of this dissertation indicate that motion-capture-based quantitative assessment is technically feasible and educationally meaningful across a spectrum of surgical train-

ing and clinical environments. By combining objective metrics with expert-derived scoring systems and remote communication technologies, this work contributes to the development of next-generation surgical education tools. Future applications may include integration with intelligent tutoring systems, robotic platforms, and large-scale training databases, further advancing the field of surgical data science.

7.2 Future Work

The following directions are expected as future extensions of this research.

Automatic segmentation of surgical workflows using motion data and multi-modal inputs

In this study, the analysis was conducted on manually segmented task units. However, real laparoscopic procedures consist of multiple phases and sub-procedures, and automatic workflow segmentation is essential for practical deployment. Future work will explore automatic segmentation algorithms using accumulated trajectory data as training data. Recurrent neural networks such as Long Short-Term Memory (LSTM), Temporal Convolutional Networks (TCN), and Transformer-based sequence models may be investigated to classify motion patterns into surgical phases. Furthermore, integrating endoscopic video, instrument usage timelines, and forceps kinematics in a multi-modal framework may enable more robust segmentation without reliance on labor-intensive manual annotation.

Integration of instrument- and tissue-recognition using image-based AI

Although endoscopic video was recorded in all training sessions, it was used only for GOALS evaluation in this study. With recent advances in computer vision models such as YOLO, DETR, and SAM, future work will apply image-recognition AI to automatically detect surgical instruments and anatomical landmarks from endoscopic video. By combining instrument detection with tracking algorithms, it may become possible to estimate motion trajectories directly from video without external markers, enabling markerless assessment systems. Additionally, tissue recognition and semantic segmentation could provide objective evaluation of procedure-specific performance, such as clip application to vessels, depth of dissection, or unintended tissue damage. These techniques may also contribute to automatic workflow segmentation mentioned above.

Large-scale validation of educational impact

This study proposed a quantitative evaluation system that provides on-site feedback combining motion features, machine-learning predictions, and GOALS-based assessment. While preliminary results support its potential usefulness, future work will focus on verifying the educational effect of the system through longitudinal studies. Repeated training sessions with novice and intermediate surgeons will be conducted to evaluate whether the system accelerates skill acquisition compared with conventional training methods. In addition, user questionnaires and expert reviews will be used to identify usability issues and refine the feedback interface. Ultimately, large-scale multi-center studies will be necessary to confirm the generalizability, reliability, and pedagogical impact of the proposed assessment framework.

Deployment in real surgery and remote surgical support environments

Real laparoscopic surgery experiments demonstrated the technical feasibility of motion tracking with sterilizable rigid bodies, and remote surgery experiments suggested that low-latency transmission allows real-time mentoring under favorable conditions. Future work includes extending the quantitative assessment framework to real clinical cases by integrating motion data from multiple instruments and examining associations with operative outcomes. For remote surgical support, combining motion-analysis feedback with tele-mentoring platforms may enable more objective assessment even when the expert is off-site. Improved latency compensation, video-quality stabilization, and multi-view synchronization will be necessary for practical deployment.

Toward a comprehensive data-driven surgical training platform

The long-term goal of this line of research is to construct a unified, data-driven surgical training platform that integrates motion capture, video analysis, machine learning, phase recognition, and real-time feedback. Future development will explore the incorporation of robotics, digital twins, and cloud-based computation for large-scale data storage and personalized skill assessment. By combining objective motion metrics with multi-modal observations and expert knowledge, the proposed framework has the potential to significantly advance the field of surgical education and contribute to safer and more efficient operative training.

Acknowledgements

This study was conducted at the Intelligent Robots and Systems Laboratory, Graduate School of Information Science and Technology, in collaboration with the Department of Renal and Urological Surgery, Graduate School of Medicine, and Hokkaido University Hospital.

First and foremost, I would like to express my deepest gratitude to my Regent Professor, Atsushi Konno, for his continuous guidance and unwavering support throughout my doctoral studies. From him, I learned not only research methodologies and scientific thinking but also the qualities, mindset, and responsibilities that define a true researcher. His mentorship has been invaluable to both my academic development and personal growth.

I would also like to extend my sincere appreciation to Professor Takashige Abe of the Department of Renal and Genitourinary Surgery, Hokkaido University. He provided extensive support for this research, including assistance with measurement experiments, evaluation of GOALS scores, and financial support through my employment as a research assistant (RA). His advice on research direction and conceptual development has been essential to the progress of this work.

My heartfelt thanks also go to Professor Takayuki Tanaka and Professor Natsuki Miyata for their constructive comments, insightful guidance, and patient reading of my dissertation. Their feedback greatly strengthened the quality of this work.

I am sincerely grateful to Lecturer Masafumi Kon, who provided tremendous support by evaluating the GOALS scores for the surgical videos. My appreciation further extends to Associate Professor Shunsuke Komizunai, who has supported me from my master's studies through my Ph.D. journey. From him, I learned many practical skills related to microcomputers, 3D modeling, and motion capture.

I would also like to thank Associate Professor Teppei Tsujita, Associate Professor Kazuya Sase, and Associate Professor Chen Xiaoshuai for their valuable advice and constructive comments through numerous online meetings. In addition, I sincerely appreciate the support from M.D. Chihiro Kamijo, Lecturer Kiyohiko Hotta, Assistant Professor Madoka Higuchi, Assistant Professor Naoya Iwahara, and Dr. Jun Furumido, who assisted me during the measurement experiments.

My appreciation further goes to Mr. Koki Ebina, whose contributions to data collection, coding, and system development were essential for this research. I am also deeply grateful to Mr. Kaito Yamada for his continuous support in the experimental work, his thoughtful feedback during our group meetings, and the many academic discussions that helped refine and deepen this study.

I would also like to thank the members of the surgical simulator group— Mr. Hiroki Usui, Mr. Daichi Aoki, Ms. Chisaho Miura, and Mr. Daichi Miura—for their diligent discussions, cooperation, and teamwork. I truly enjoyed the productive and memorable time shared with all members of our laboratory.

I am also deeply grateful to our laboratory secretaries, Ms. Mikie Watanabe and Ms. Keiko Makino, as well as the administrative staff of the Department of Renal and Genitourinary Surgery, for their indispensable assistance with procedures related to research purchases, RA employment,

and travel arrangements.

Lastly, I wish to express my heartfelt thanks to my husband, who has been my greatest support throughout this long doctoral journey. His constant encouragement, quiet devotion, and warm companionship sustained me through every challenge. I am deeply grateful for his help in taking care of our daily life, preparing meals, and ensuring that I could work with peace of mind. This dissertation is as much his achievement as it is mine.

Finally, I would like to express my heartfelt thanks to my parents and friends for their unwavering encouragement and support. Without their strength and kindness, this dissertation would not have been possible.

Bibliography

- [1] T. Inoue, H. Kinoshita, M. Satou, N. Oguchi, G. Kawa, K. Muguruma, T. Murota, and T. Matsuda, “Complications of Urologic Laparoscopic Surgery: A Single Institute Experience of 1017 Procedures,” *Journal of Endourology*, **24**(2), pp. 253–260, 2010. doi:10.1089/end.2009.0322
- [2] G. M. Fried, “Surgical training and simulation: a new paradigm,” *Current Problems in Surgery*, **41**(9), pp. 757–808, 2004. doi:10.1067/j.cpsurg.2004.06.002
- [3] B. Wentink, “Eye-hand coordination in laparoscopy an overview of experiments and supporting aids,” *Minimally Invasive Therapy & Allied Technologies*, **10**(3), pp. 155–162, 2001. doi:10.1080/136457001753192277
- [4] N. Hogan and D. Sternad, “Sensitivity of smoothness measures to movement duration, amplitude, and arrests,” *Journal of Motor Behavior*, **36**(1), pp. 3–14, 1984.
- [5] T. Flash and N. Hogan, “The coordination of arm movements: an experimentally confirmed mathematical model,” *Journal of Neuroscience*, **5**(7), pp. 1688–1703, 1985.
- [6] T. Matsuda, Y. Ono, T. Terachi, S. Naito, S. Baba, T. Miki, Y. Hirao, and A. Okuyama, “The Endoscopic Surgical Skill Qualification System in Urological Laparoscopy: A Novel System in Japan,” *Journal of Urology*, **176**(5), pp. 2168–2172, 2006. doi:10.1016/j.juro.2006.07.034
- [7] X. Liu, A. Gupta, and K. S. Madhavan, “Artificial intelligence in surgical skill assessment: current status and future directions,” *Surgical Endoscopy*, **36**(1), pp. 524–537, 2022. doi:10.1007/s00464-021-08377-1
- [8] K. Tanoue, S. Ieiri, K. Konishi, T. Yasunaga, K. Okazaki, S. Yamaguchi, D. Yoshida, Y. Kakeji, and M. Hashizume, “Effectiveness of endoscopic surgery training for medical students using a virtual reality simulator versus a box trainer: a randomized controlled trial,” *Surgical Endoscopy*, **22**(4), pp. 985–990, 2008. doi:10.1007/s00464-007-9554-8
- [9] K. W. van Dongen, E. Tournoij, D. C. van der Zee, M. P. Schijven, and I. A. M. J. Broeders, “Construct validity of the Lap-Sim: Can the LapSim virtual reality simulator distinguish between novices and experts?” *Surgical Endoscopy*, **21**, pp. 1413–1417, 2007. doi:10.1007/s00464-006-9188-2
- [10] L. Beyer, J.D. Troyer, J. Mancini, F. Bladou, S.V. Berdah, and G. Karsenty, “Impact of laparoscopy simulator training on the technical skills of future surgeons in the operating room: a prospective study,” *The American Journal of Surgery*, **202**(3), pp. 265–272, 2011. doi:10.1016/j.amjsurg.2010.11.008

- [11] “Lap mentor (Surgical Science Inc.),” <https://simbionix.com/simulators/lap-mentor/platforms>, [Online; accessed 05-October-2023].
- [12] A. Chellali, H. Mentis, A. Miller, W. Ahn, V. S. Arikatla, G. Sankaranarayanan, S. De, S. D. Schwaitzberg, and C. G. L. Cao, “Achieving interface and environment fidelity in the Virtual Basic Laparoscopic Surgical Trainer,” *International Journal of Human-Computer Studies*, **96**, pp. 22–37, 2016. doi:10.1016/j.ijhcs.2016.07.005
- [13] Z. Lin, M. Uemura, M. Zecca, S. Sessa, H. Ishii, M. Tomikawa, M. Hashizume, and A. Takanishi, “Objective Skill Evaluation for Laparoscopic Training Based on Motion Analysis,” *IEEE Transactions on Biomedical Engineering* **60**(4), pp. 977–985, 2013. doi:10.1109/TBME.2012.2230260
- [14] E. F. Hofstad, C. Våpenstad, L. E. Bø, T. Langø, E. Kuhry, and R. Mårvik, “Psychomotor skills assessment by motion analysis in minimally invasive surgery on an animal organ,” *Minimally Invasive Therapy & Allied Technologies*, **26**(4), pp. 240–248, 2017. doi:10.1080/13645706.2017.1284131
- [15] K. Ebina, T. Abe, M. Higuchi, J. Furumido, N. Iwahara, M. Kon, K. Hotta, S. Komizunai, Y. Kurashima, H. Kikuchi, R. Matsumoto, T. Osawa, S. Murai, T. Tsujita, K. Sase, X. Chen, A. Konno and N. Shinohara, “Motion analysis for better understanding of psychomotor skills in laparoscopy: objective assessment-based simulation training using animal organs,” *Surgical Endoscopy*, **35**(8), pp. 4399–4416, 2021. doi:10.1007/s00464-020-07940-7
- [16] K. Ebina, T. Abe, K. Hotta, M. Higuchi, J. Furumido, N. Iwahara, M. Kon, K. Miyaji, S. Shibuya, Y. Lingbo, S. Komizunai, Y. Kurashima, H. Kikuchi, R. Matsumoto, T. Osawa, S. Murai, T. Tsujita, K. Sase, X. Chen, A. Konno and N. Shinohara, “Objective evaluation of laparoscopic surgical skills in wet lab training based on motion analysis and machine learning,” *Langenbeck’s Archives of Surgery*, **407**(5), pp. 2123–2132, 2022. doi:10.1007/s00423-022-02505-9
- [17] K. Ebina, T. Abe, K. Hotta, M. Higuchi, J. Furumido, N. Iwahara, M. Kon, K. Miyaji, S. Shibuya, Y. Lingbo, S. Komizunai, Y. Kurashima, H. Kikuchi, R. Matsumoto, T. Osawa, S. Murai, T. Tsujita, K. Sase, X. Chen, A. Konno and N. Shinohara, “Automatic assessment of laparoscopic surgical skill competence based on motion metrics,” *Plos one*, **17**(11), pp. 1–13, 2022. doi:10.1371/journal.pone.0277105
- [18] S. G. T. Smith, J. Torkington, T. J. Brown, N. J. Taffinder, and A. Darzi, “Motion analysis,” *Surgical Endoscopy*, **16**(4), pp. 640–645, 2002. doi:10.1007/s004640080081
- [19] J. A. Sánchez-Margallo, F. M. Sánchez-Margallo, I. Oropesa, S. Enciso, and E. J. Gómez, “Objective assessment based on motion-related metrics and technical performance in laparoscopic suturing,” *International Journal of Computer Assisted Radiology and Surgery*, **12**(2), pp. 307–314, 2017. doi:10.1007/s11548-016-1459-3
- [20] I. Oropesa, P. Sánchez-González, P. Lamata, M. K. Chmarra, J. B. Pagador, J. A. Sánchez-Margallo, F. M. Sánchez-Margallo, and E. J. Gómez, “Methods and tools for objective assessment of psychomotor skills in laparoscopic surgery,” *Journal of Surgical Research*, **171**(1), pp. e81–e95, 2011. doi:10.1016/j.jss.2011.06.034

- [21] K. Moorthy, Y. Munz, A. Dosis, F. Bello, A. Chang, and A. Darzi, “Bimodal assessment of laparoscopic suturing skills,” *Surgical Endoscopy*, **18**(11), pp. 1608–1612, 2004. doi:10.1007/s00464-003-9312-5
- [22] S. Yamaguchi, D. Yoshida, H. Kenmotsu, T. Yasunaga, K. Konishi, S. Ieiri, H. Nakashima, K. Tanoue, and M. Hashizume, “Objective assessment of laparoscopic suturing skills using a motion-tracking system,” *Surgical Endoscopy*, **25**(3) pp. 771–775, 2011. doi:10.1007/s00464-010-1251-3
- [23] M. K. Chmarra, S. Klein, J. C. F. De Winter, F.-W. Jansen, and J. Dankelman, “Objective classification of residents based on their psychomotor laparoscopic skills,” *Surgical Endoscopy*, **24**(5), pp. 1031–1039, 2010. doi:10.1007/s00464-009-0721-y
- [24] J. C. Rosser, S. P. Prosst, V. N. Murayama, and L. N. Malangoni, “Telementoring: an application whose time has come,” *Surgical Endoscopy*, **11**(11), pp. 1081–1085, 1997. doi:10.1007/s004649900564
- [25] E. Hiemstra, M. K. Chmarra, J. Dankelman, and F. W. Jansen, “Intracorporeal Suturing: Economy of Instrument Movements Using a Box Trainer Model,” *Journal of Minimally Invasive Gynecology*, **18**(4), pp. 494–499, 2011. doi:10.1016/j.jmig.2011.04.003
- [26] 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*, **84**(2), pp. 273–278, 1997. doi:10.1046/j.1365-2168.1997.02502.x
- [27] M. C. Vassiliou, L. S. Feldman, C. G. Andrew, S. Bergman, K. Leffondré, D. Stanbridge, and G. M. Fried, “A global assessment tool for evaluation of intraoperative laparoscopic skills,” *The American Journal of Surgery*, **190**(1), pp. 107–113, 2005. doi:10.1016/j.amjsurg.2005.04.004
- [28] A. C. Goh, D. W. Goldfarb, J. C. Sander, B. J. Miles, and B. J. Dunkin, “Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills,” *The Journal of Urology*, **187**(1), pp. 247–252, 2012. doi:10.1016/j.juro.2011.09.032
- [29] M. A. Pereira-Sampaio, L. A. Favorito, and F. J. Sampaio, “Pig kidney: anatomical relationships between the intrarenal arteries and the kidney collecting system,” *Journal of Urology*, **172**(5), pp. 2077–2081, 2004. doi:10.1097/01.ju.0000138085.19352.b5
- [30] E. Gazel, A. Ozgok, A. M. Bayraktar, O. Ergun, and Y. Ozgok, “Animal ethics in laparoscopic urology training,” *Turkish Journal of Urology*, **43**(4), pp. 429–433, 2017. doi:10.5152/tud.2017.74501
- [31] E. Kobayashi and J. Nudeshima, “Current state of surgical training using cadavers in Japan compared with Western countries,” *Surgery Today*, **48**(9), pp. 891–893, 2018. doi:10.1007/s00595-018-1673-4
- [32] M. Higuchi, T. Abe, K. Hotta, K. Morita, H. Miyata, J. Furumido, N. Iwahara, M. Kon, T. Osawa, R. Matsumoto, H. Kikuchi, Y. Kurashima, S. Murai, A. Aydin, N. Raison, K. Ahmed, M. S. Khan, P. Dasgupta, N. Shinohara, “Development and validation of a porcine organ model for training in essential laparoscopic surgical skills,” *International Journal of Urology*, **27**, pp. 929–938, 2020. doi:10.1111/iju.14315

- [33] W. Thiel, “An arterial substance for subsequent injection during the preservation of the whole corpse,” *Annals of Anatomy*, **174**(3), pp. 197–200, 1992. doi:[10.1016/S0940-9602\(11\)80347-X](https://doi.org/10.1016/S0940-9602(11)80347-X)
- [34] Japan Surgical Society and Japanese Association of Anatomists, “Guidelines for cadaver dissection in education and research of clinical medicine,” *Acta Anatomica Nipponica*, **87**, pp. 21–23, 2012. (In Japanese)
- [35] I. Oropesa, P. Sánchez-González, M. K. Chmarra, P. Lamata, Á. Fernández, J. A. Sánchez-Margallo, F. W. Jansen, J. Dankelman, F. M. Sánchez-Margallo, and E. J. Gómez, “EVA: Laparoscopic Instrument Tracking Based on Endoscopic Video Analysis,” *Surgical Endoscopy*, **27**(3), pp. 1029–1039, 2013. doi:[10.1007/s00464-012-2513-z](https://doi.org/10.1007/s00464-012-2513-z)
- [36] S. Ganni, S. M. B. I. Botden, M. Chmarra, R. H. M. Goossens, and J. J. Jakimowicz, “A software-based tool for video motion tracking in the surgical skills assessment landscape,” *Surgical Endoscopy*, **32**(6), pp. 2994–2999, 2018. doi:[10.1007/s00464-018-6023-5](https://doi.org/10.1007/s00464-018-6023-5)
- [37] S. Ganni, S. M. B. I. Botden, M. Chmarra, M. Li, R. H. M. Goossens, and J. J. Jakimowicz, “Validation of Motion Tracking Software for Evaluation of Surgical Performance in Laparoscopic Cholecystectomy,” *Journal of Medical Systems*, **44**(3), 2020. doi:[10.1007/s10916-020-1525-9](https://doi.org/10.1007/s10916-020-1525-9)
- [38] C. Heiliger, D. Andrade, C. Geister, A. Winkler, K. Ahmed, A. Deodati, V. H. E. V. Treuenstätt, J. Werner, A. Eursch, K. Karcz, and A. Frank, “Tracking and evaluating motion skills in laparoscopy with inertial sensors,” *Surgical Endoscopy*, **37**(7), pp. 5274–5284, 2023. doi:[10.1007/s00464-023-09983-y](https://doi.org/10.1007/s00464-023-09983-y)
- [39] “V120 Duo & Trio Features,”
<https://d111srqycjesc9.cloudfront.net/V120-Duo-and-Trio-Data-Sheet.pdf>
- [40] S. G. Hart and L. E. Staveland, “Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research,” in *Advances in Psychology*, vol. 52, pp. 139–183, North-Holland, 1988. doi:[10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- [41] A. Zia, Y. Sharma, V. Bettadapura, E. L. Sarin, M. A. Clements, and I. Essa, “Automated assessment of surgical skills using frequency analysis,” in *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015)*, Lecture Notes in Computer Science, vol. 9349, pp. 430–438, Springer, 2015. doi:[10.1007/978-3-319-24553-9_53](https://doi.org/10.1007/978-3-319-24553-9_53)
- [42] A. Zia and I. Essa, “Automated surgical skill assessment in RMIS training,” *International Journal of Computer Assisted Radiology and Surgery*, **13**(5), pp. 731–739, 2018. doi:[10.1007/s11548-018-1735-5](https://doi.org/10.1007/s11548-018-1735-5)
- [43] K. Lam, J. Chen, Z. Wang, F. M. Iqbal, A. Darzi, B. Lo, S. Purkayastha, and J. M. Kinross, “Machine learning for technical skill assessment in surgery: a systematic review,” *npj Digital Medicine*, **5**(1), Article 24, 2022. doi:[10.1038/s41746-022-00566-0](https://doi.org/10.1038/s41746-022-00566-0)