# Exploring the Limitations and Implications of the JIGSAWS Dataset for Robot-Assisted Surgery

Antonio Hendricks, Max Panoff, Kaiwen Xiao, Zhaoqi Wang, Shuo Wang, Christophe Bobda

***Abstract*—The JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) dataset has proven to be a foundational component of modern work on the skill analysis of robotic surgeons. In particular, methods using either the system's kinematics or video data have shown to be able to classify operators into distinct experience levels, and recent approaches have even ventured to recover numeric skill ratings assigned to assessment sessions. Although prior works have achieved positive results in these directions, challenges still remain with classification across all three levels of operator training amounts and objective skill rating regressions. To this end, we perform the first statistical analysis of the dataset itself and compile the results here. We find limited relationships between the amount of experience or training of an operator and their performance in JIGSAWS. Moreover, as operator-side kinematics have well-known relationships with their skill, previous works have used both robot and operator-side kinematics to classify operator skill; we find the first explicit relationships between pure robot-side kinematics and surgical performance. Finally, we analyze the robotic kinematic trends associated with high performance in JIGSAWS tasks and present how they may be used as indicators in human and automated surgeon training.**

***Index Terms*—Surgical Robotics: Laparoscopy; Data Sets for Robot Learning; Performance Evaluation and Benchmarking; Computer Vision for Medical Robotics; Deep Learning Methods**

## I. INTRODUCTION

ROBOT- Assisted Minimally Invasive Surgery (RAMIS) is a rapidly growing approach in healthcare, with over 640,000 such surgeries performed in the US alone, some procedures seeing a nearly 45× increase in the use of such robotic systems [1]. These approaches are well posed for even broader adoption due to their minimally invasive nature, greatly improved patient outcomes, and lighter burden on surgeons [2]–[4]. However, robotic surgery greatly differs from traditional surgical approaches in several ways, one of which is the absence of haptic and inertial feedback that a robotic surgeon operator experiences while performing their duties [5]. As a result, surgeons must undergo retraining, which, although costly and inefficient [6], is essential for maintaining high standards of quality care. Current methods for determining proficiency in RAMIS are either highly subjective (e.g., human ratings) or based on poor indicators (e.g., number of hours logged) [7]–[9]. As the robotics and medical communities come together, the RAMIS approach will evolve with ever-better patient outcomes, learning and utility curves, and new surgical techniques and insights [10].

Several solutions have been proposed and evaluated on the JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) dataset, which analyzes direct kinematics [11]–[13] or visual data [14]–[16] to identify the skill level of RAMIS system operators. Solutions using kinematics tend to rely heavily on readings from the operator side rather than solely robot-side, limiting their extension into rating the performance of autonomous robotic surgeons. While there are known to be strong correlations between operator hand-steadiness [17] and surgical experience, these metrics may not be well suited to adaption into robotics due to electro-mechanical dampening control filters. Additionally, while these previous works confirm that high-order derivatives of patient-side kinematics are indicative of skill, they do not examine the overall movement paths that we use in this study. While there have been a few attempts at extracting kinematics from RAMIS operations using image data, these seem to be limited to recovering the position of tools in an image (i.e., 2D position) rather than in the real world (i.e., 3D position) [18], or they rely heavily on foreknowledge of the specific tools being used [19]. Vision-based techniques [14], [20] have demonstrated the ability to match or exceed the performance of kinematics-based approaches.

Most works evaluated on the JIGSAWS dataset focus on classifying the approximate skill of surgeons into 3 classes: Novice, Intermediate, and Expert (NIE) [12], [14], [21], where these class distinctions denote experience in hours, with a RAMIS system, not a ranking based on imperially quantifiable skill level [22]. The Objective Structured Assessment of Technical Skills (OSATS) sought to provide a better assessment method than task-specific checklists and assumptions based on seniority [23], [24]. There has been difficulty developing a state-of-the-art system that targets JIGSAWS' Global Rating Scale (GRS) [25] to recover more specific skill ratings, with the highest performing method achieving only a 72% Spearman's rank correlation without reports of subcategory score recovery accuracy [20].

The results of the methods evaluated on this dataset may be easily misinterpreted due to the limitations of the chosen metrics and the dataset itself. This warrants a closer

Manuscript received 30 Jan. 2024; accepted 6 June 2024. Date of publication DD MM YYYY; date of current revision 21 June 2024. This paper was recommended for publication by Editor Jessica Burgner-Kahrs upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by the National Science Foundation under Grant Nos 2007210 and 2106610. Corresponding author: Antonio Hendricks

Great thanks to M.D. Ali Zarrinpar and Dr. Sergio Duarte, for our discussion on the implications of our findings as hepatobiliary surgeons and scientists, we are mathematicians and engineers appreciative of your help and time.

The authors are with the Electrical & Computer Engineering Department, University of Florida, 32603 USA (email: a.hendricks1, m.panoff, kaiwen.xiao, wangzhaoqi, shuo.wang, c.bobda @ufl.edu)

Digital Object Identifier (DOI): see top of this page

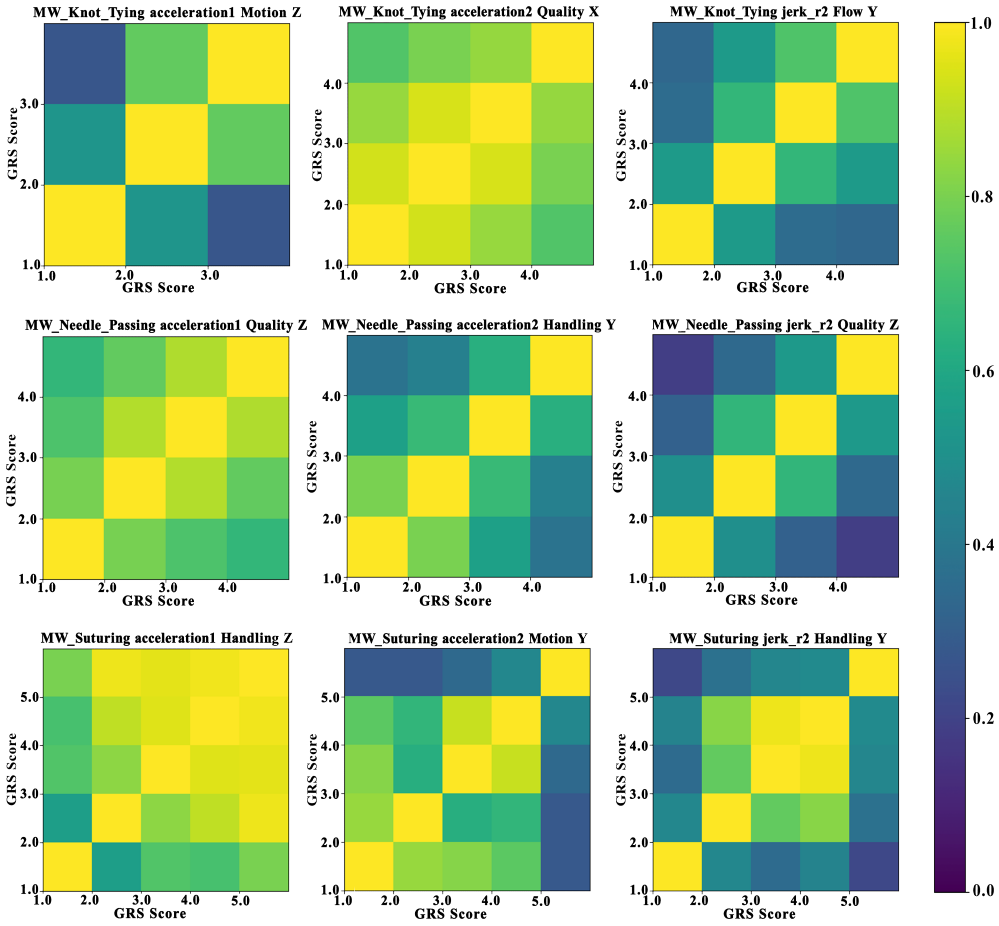


Fig. 1: Mann-Whitney p-values between various cross-task kinematic profiles and performance ratings, the likelihood that two samples were selected from different populations. Darker colors indicate lower p-values or more significant differences[1].

examination of the JIGSAWS dataset, prior works using the dataset, its overall usefulness, and potentially current surgical robotics training procedures. For example, studies that do not recover OSATS, only report Spearman's correlation metrics, rely solely on MTM data, or use PSM data without correlating it with operator experience or skill are all types of potential studies that may require further scrutiny.

To address these questions, we present a detailed analysis of the JIGSAWS dataset. We find a minimal correlation between the hours logged and GRS scores, and correlations that are present do not remain consistent across tasks. However, there are relationships between certain robot-side kinematic behaviors and GRS scores, as shown in Figure 1, that carry over between tasks. Simply stated, the contributions of this work are as follows:

- We perform, to our knowledge, the first in-depth analysis of the JIGSAWS dataset itself.
- We find that RAMIS surgeon performance in JIGSAWS is unrelated to the number of hours of operator training with a high statistical significance ($p < 0.005$).
- We are the first to analyze the relationship between (exclusive) patient-side kinematic paths and surgical task ratings in JIGSAWS.
- We analyze and compare current methods of evaluation of robotic surgeon performance to novel metrics.

This paper continues as follows: Section II provides relevant background on the composition of the JIGSAWS dataset and a meta-analysis of prior work using it. Section III covers our analytic methods and rationale. We then discuss findings, potential implications and explanations, and new metrics and methods to extend JIGSAWS in Section IV before concluding our work in Section V.

## II. The JIGSAWS Dataset

### A. RAMIS

With an array of deployable sensors, machine learning software packages, and algorithms, RAMIS systems can analyze surgical data to identify patterns and make predictions that may help guide the surgeon's actions during the procedure. Common forms of intelligent sensing include tissue elasticity, blood flow, and the precise location of surgical instruments. Robot-assisted surgery is a necessary step toward the end goal of fully autonomous robotic surgery. Until then, RAMIS systems continue to rely on human operators interacting with master tool manipulators within an operator-side console, remotely controlling the electro-mechanical robotic surgical

[1]JIGSAWS explicitly aligns only the coordinate system, not the operational area per task per trial. Thus, $X$, $Y$, and $Z$ may not reflect movement directions accurately, so cross-direction analysis is used.

| Approach | Method | ST NIE | NP NIE | KT NIE | ST OSATS | NP OSATS | KT OSATS | Notes |
|---|---|---|---|---|---|---|---|---|
| [15] | Visual | †100% | †96.4% | †95.8% | N/A | N/A | N/A | 3D CNN |
| [21] | Visual | †79.29% | †87.01% | †72.57% | N/A | N/A | N/A | STIP |
| [21] | Visual | †76.69% | †83.81% | †82.82% | N/A | N/A | N/A | iDT |
| [26] | Visual | †80.72% | †79.66% | †80.41% | N/A | N/A | N/A | CNN |
| [26] | Visual | †81.58% | †83.19% | †82.82% | N/A | N/A | N/A | CNN + LSTM |
| [26] | Visual | †81.89% | †84.23% | †83.54% | N/A | N/A | N/A | ResNet |
| [27] | Visual | *97.27% | *97.27% | *97.27% | N/A | N/A | N/A | CNN |
| [11] | Kinematic | 100% | 100% | 100% | $0.60\sigma$ | $0.57\sigma$ | $0.65\sigma$ | FCN |
| [12] | Kinematic | 100% | 100% | 99.9% | $0.31\sigma$ | $0.16\sigma$ | $0.26\sigma$ | ApEn |
| [12] | Kinematic | N/A | N/A | N/A | $0.59\sigma$ | $0.37\sigma$ | $0.57\sigma$ | DCT + DFT + ApEn |
| [20] | Visual | 100% | 97.2% | 100% | $0.68\sigma$ | $0.62\sigma$ | $0.74\sigma$ | MT-TSN |
| [20] | Visual | 100% | 97.2% | 100% | $0.72\sigma$ | $0.68\sigma$ | $0.75\sigma$ | MT-TSN + Attention |
| [28] | Both | N/A | N/A | N/A | $\diamond$ $0.45\sigma$ | $\diamond$ $0.62\sigma$ | $\diamond$ $0.58\sigma$ | VTP |
| [29] | Both | N/A | N/A | N/A | $\diamond$ $0.45\sigma$ | $\diamond$ $0.34\sigma$ | $\diamond$ $0.61\sigma$ | AIM |

TABLE I: Comparison of NIE classification and OSATS regression results taken from cited methods evaluated on the JIGSAW dataset. $\sigma$ denotes Spearman's correlation, ST denotes Suturing, NP denotes Needle Passing, and KT denotes Knot Tying. * The per-task performance was not released for this binary (IE) classification method. †The method evaluates NI Classes instead of NIE. $\diamond$ The work evaluates their method using LOUO folds instead of LOSO.

instruments (robotic arms, endoscopes, and end-effectors) at the patient's bedside.

### *B. JIGSAWS*

We evaluated the JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) dataset [22], one of the largest and most widely used surgical skill assessment datasets in the world [30]. The dataset contains synchronous stereo-video and kinematic recordings of three standard surgical training tasks (*Knot-Tying*, *Needle-Passing*, and *Suturing*) conducted by human operators of varying skill levels using a robotic da Vinci Surgical System. The dataset compiles annotated assessments for each of the trials completed by the operators in the dataset. Annotations include surgical gesture labels, operator Novice, Intermediate, and Expert (NIE) ordinal classification labels, and efficacy scores using a modified OSATS approach that excludes any evaluation categories not applicable to the training sessions in the dataset (e.g. use of assistants). The *Global Rating Score* (GRS) represents a measure of technical skill over the entire trial in the categories of "Respect for tissue", "Suture/needle handling", "Time and motion", "Flow of operation", "Overall performance", and "Quality of final product"; each is rated on an interval scale of 1 to 5, then given a final cumulative score with a possible range of 6 to 30 points.

*1) Tasks:* Knot-Tying, Needle-Passing, and Suturing are essential skills for all practicing RAMIS operators. JIGSAWS uses these three most basic tasks to benchmark the skills of the participating operators, as is typical in most surgical skills training curricula. The Knot-Tying task consists of two sutures (needle and thread) to be separately tied around a flexible tube attached to the workbench at both ends. Knots should be fully taut and secure around the elastic conduit. The Needle-Passing task requires the subjects to pick up a threaded needle (which may not be recorded in the video or kinematic data) and pass it through a small maze of small metal hoops that are fixed to rubber mounts at a small variable height above the surface of the bench-top model. While suturing, the surgeon must pick up a suture and pass the needle through the "tissue", entering at the dot marked on one side of the incision (marked by a line on the fabric) and exiting at the corresponding dot marked on the other side of the incision. After the first needle pass, the subject extracts the needle out of the tissue, passes it to the right hand and repeats the needle pass three more times [22]. These *tasks* demonstrate the surgeon's ability to handle needles and suturing equipment and operate smoothly, efficiently, and carefully, with quality in the final product of the operation. The elastic material represents the malleable soft internal tissue encountered during surgery. As a limitation, the operators were not allowed to move the camera (even by activating the clutch) to adjust the alignment for better vantage and cardinal manipulator control. Each operator made five attempts at each task, *trials*. We denote a particular trial for a particular operator as a *session*.

*2) Components:* The JIGSAWS dataset comprises three components: synchronized kinematic and video data, manual annotations of associated gestures, and GRS scores and NIE classification. The kinematics of both *patient-side* (i.e. *robot-side*) and *operator-side* manipulators were sampled with a shared coordinate system by the da Vinci Surgical System API at a rate of 30 Hz, along with both left and right laparoscopic camera views of the surgical trial at the same frame rate. Video of each assessment trial was recorded at a resolution of 640 x 480. Calibration parameters for the two endoscopic cameras were not provided in the dataset. For this work, we did not statistically evaluate JIGSAWS regarding gesture recognition or operator-side kinematics and GRS for NIE classification as GRS scores are provided based exclusively on the recorded patient-side video.

*3) Analysis Methods:* JIGSAWS provides two different formats for analyzing performance: *Leave One User Out* (LOUO) and *Leave One Super-trial Out* (LOSO). In LOUO, a fold is created per task per operator, while in LOSO, a fold is created for each super-trial (e.g. all the 4th trials across all users). Cross-fold validation can then be used to analyze performance, training on all non-fold data, and testing on all folds pairwise. For example, in LOSO, all operators' $n$-th attempt at a task will be withheld from machine learning training and may be

used for validation or testing instead.

### C. Skill Estimation

There are a few common methods for rating robotic surgeon proficiency. JIGSAW (and thus many prior works) uses two of them. Firstly, a simple controller classification based on the number of hours of experience into one of three ratings: Novice, Intermediate, and Expert (NIE). Secondly, a multi-axis rating (OSATS and GRS [25]) was manually given by a rating surgeon who viewed video recordings of the surgeon controlling the robot. A brief overview of proposed autonomous RAMIS skill assessment methods and their results using LOSO can be found in Table I.

**Kinematics Based Skill Estimation** Several methods have been explored using extracted kinematics to directly predict an operating surgeon's skill. Nagy et al. proposed using jerk (3rd order derivative of movement) of operator side end-effectors as a metric for motor skill and found that it is a significantly poorer indicator than the jerk of other professions [17], [31]. In [13], Fard et al. use machine learning kinematics analysis to classify operators in JIGSAW into either Novice or Expert with ∼80% accuracy. Deep Learning approaches have seen 100% success in classification. Zia et al. [12] and Ismail et al. [11] also used machine learning kinematics analysis to recover OSATS values and obtained 60% success.

**Vision Based Skill Estimation** has been the championed evaluation medium for many presented works. Funke et al. achieved high *binary* classification accuracy of 95.1% to 100% on each of JIGSAWS 3 tasks by serving stack video snippets through an inflated 3D ConvNet and Temporal Segment Network during training [15]. Jian et al. presented a method for simultaneous skill level classification and OSATS score regression [20] by processing video snippets through an Attention-enhanced Two-Stream Inflated 3D CNN (I3D) shared for all tasks. Ming et al. approached binary skill classification by modeling motion dynamics fed through a non-linear support vector machine (SVM) and histogram test, achieving 72.6% to 83.8% accuracy depending on the modeled dynamics and training task [21]. A series of works have also explored the utility of segmentation for classification from generated sparse optical flow data on various other RNN and DNN learning methods [14], [26]. Soleymani et al. continued work with sparse optical flow data and deployed 10 × 10 cross-fold validation to improve accuracy on binary classification [27].

A few recent works have combined these approaches, [28], [29]. Many works have similarly identified the challenge of recovering NIE based on kinematics and vision, and have started evaluating methods for recovering GRS [32], [33]. In this work, we support that the challenge of recovering NIE on JIGSAWS is not due to the inherent complexity of the challenge requiring more advanced analysis techniques, but that the underlying data does not have meaningful statistical differences between these categories.

## III. Analysis Methodology

While the JIGSAWS dataset is commonly used and consists of a large amount of data, of types useful to machine learning approaches, no statistical evaluation of meaningful trends in this data has ever been completed to our knowledge. We thus statistically analyze the relationship between three key components: the forward kinematics as continuous trajectories, the Novice, Intermediate, and Expert (NIE) rating, and the Global Rating Scale (GRS) scores for each session. These were chosen for a variety of reasons. Firstly, the system's kinematics, particularly the patient-side kinematics, convey information about the continuous paths taken by the robotic manipulators through physical space over time. This effectively distills pertinent information in the captured videos into a more compressed, temporally coherent format, while GRS reflects the true skill level of the operator. Finally, the NIE rating is a popular target for classification and should be related to the GRS score, as NIE tracks hours of training and the GRS tracks performance directly. As the NIE classifications are definitive classifications with rank order, we perform statistical tests valid for ordinal data. As a last note, we only present methods and findings in this section, saving discussion of those findings for Section IV.

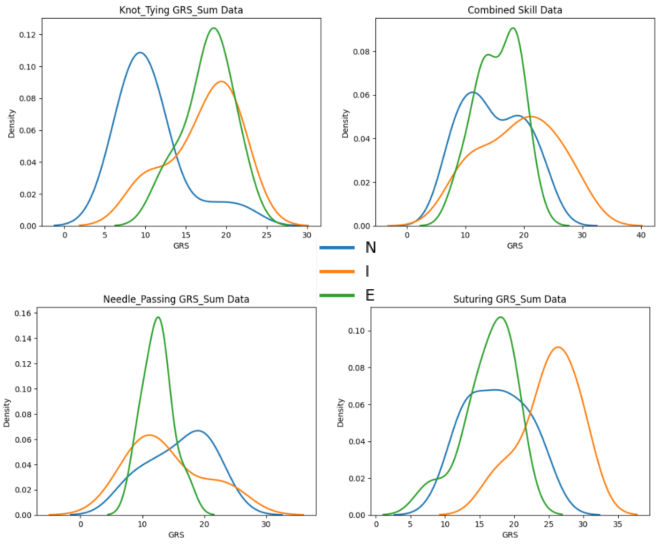


Fig. 2: The approximately Gaussian distributions of the raw NIE data across the three tasks and combined.

### A. Novice, Intermediate, and Expert vs GRS

We begin by analyzing any relationships between GRS and NIE. Firstly, we confirm that the total GRS score and NIE rating follow a roughly Gaussian distribution as shown in Figure 2. As they do, we continue calculating the mean GRS across each task and the whole dataset, which is recorded in Table II.

| Task | Mean N Score | Mean I Score | Mean E Score |
|---|---|---|---|
| Knot Tying | 10.68750 | 17.10000 | **17.70000** |
| Needle Passing | **16.0000** | 14.00000 | 12.44445 |
| Suturing | 17.47368 | **25.10000** | 16.30000 |
| Combined | 14.76087 | **19.07143** | 15.58621 |

TABLE II: Mean GRS scores for Novice, Intermediate, and Expert operators. The score for the highest-performing class in each task is **bolded**.

From there, we performed a Mann-Whitney U-test, also known as the Wilcoxon rank-sum test, to test against the null hypothesis that the distributions of any two ordinal populations are identical [34]. We do this for the total summed GRS score and each equally spaced rank order category in GRS. These tests are also done across all tasks and on the combined dataset. These results can be seen in Table III, which records the p-value for each test.

| Task | GRS Element | N to I | I to E | N to E |
|---|---|---|---|---|
| Knot Tying | Total | **0.00389** | 1.00000 | **0.00070** |
| Knot Tying | Respect | **0.00103** | 0.31268 | **0.00403** |
| Knot Tying | Handling | **0.03999** | 0.73895 | **0.02095** |
| Knot Tying | Time | **0.00888** | 0.76120 | **0.00081** |
| Knot Tying | Flow | **0.00450** | 0.16725 | **0.00092** |
| Knot Tying | Performance | **0.02311** | 0.59344 | **0.00126** |
| Knot Tying | Quality | **0.00068** | 0.69338 | **0.00104** |
| Needle Passing | Total | 0.48141 | 1.00000 | 0.08607 |
| Needle Passing | Respect | 0.15732 | 0.31735 | 0.40179 |
| Needle Passing | Handling | 0.83127 | 0.71358 | 0.17122 |
| Needle Passing | Time | 0.40900 | 0.33857 | **0.00687** |
| Needle Passing | Flow | 0.68554 | 0.18696 | 0.09041 |
| Needle Passing | Performance | 0.57599 | 0.95492 | 0.18646 |
| Needle Passing | Quality | 0.28216 | 0.80070 | 0.07405 |
| Suturing | Total | **0.00047** | **0.00095** | 0.72915 |
| Suturing | Respect | **0.03534** | **0.00279** | **0.04770** |
| Suturing | Handling | **0.00125** | **0.00043** | 0.09894 |
| Suturing | Time | **0.00131** | **0.00573** | 0.33899 |
| Suturing | Flow | **0.00103** | **0.00282** | 0.920885 |
| Suturing | Performance | **0.00105** | **0.00059** | 0.42741 |
| Suturing | Quality | **0.00009** | **0.00053** | 0.80005 |
| Combined | Total | **0.00625** | **0.02462** | 0.42247 |
| Combined | Respect | **0.04214** | **0.03751** | 0.91382 |
| Combined | Handling | **0.00973** | **0.00932** | 0.95828 |
| Combined | Time | **0.00753** | 0.05510 | 0.29860 |
| Combined | Flow | **0.01053** | 0.13588 | 0.19690 |
| Combined | Performance | **0.01298** | **0.04388** | 0.41763 |
| Combined | Quality | **0.00319** | **0.01253** | 0.55831 |

TABLE III: Wilcoxon–Mann–Whitney U-Test metrics of GRS scores between Intermediate, Novice, and Expert users in JIGSAWS. Statistically significant (p-value $< 0.05$) differences are **bolded**. Values showing highly significant differences (p-value $< 0.005$) are also highlighted in **purple**.

### *B. Kinematic Analysis*

We repeat this process for the kinematic analysis by calculating the time-dependent derivatives of PSM trajectories (position, velocity, acceleration, then jerk, and jounce) for each robotic manipulator's translational and rotational movement, rather than treating individual 3D/6D positions as independent data points. Data is trimmed to exclude periods before task commencement or after completion using Algorithm 1; total PSM traversal distance is also continuously calculated at each point in time.

While prior works have found meaning in the translation jerk of operator controls [17], this may not carry over as strongly to the patient-side manipulators due to movement dampening [4], [5]. Moreover, as GRS scores are determined by watching the robot-side and not the operator-side performance, it should be possible to similarly replicate these findings based solely on robot-side kinematics. However, we acknowledge that in uncontrolled scenarios with varying environmental interactions, the interpretation of kinematic data might require additional nuance. Nonetheless, the JIGSAWS tasks are highly standardized, presenting minimal variation in environment or interactions. This consistency allows a direct comparison of kinematic profiles across different GRS ratings. As these kinematic values (*e.g.* translational position, rotational velocity, etc.) do not follow largely Gaussian distributions, we use the Mann-Whitney non-parametric test [34], [35] to check against the null hypothesis that kinematic profiles across different GRS ratings are from the same distribution. These results are generally less indicative than those mentioned in Section III-A. As there are $3 \times 2 \times 6 = 36$ comparisons for each task, with each having between 3 and 16 unique GRS ratings to compare between, we instead focus on a few examples in which actual trends can be seen to show in Figures 1 and 4. In these figures, brighter colors indicate more similarity and darker colors indicate more dissimilarity.

**Algorithm 1:** Identify Frames of Interest

**Parameter:** $m$ is the number of kinematic metrics
**Parameter:** B, a scalar, buffer size
**Parameter:** T, a scalar, minimum $\Delta$ threshold
**Input :** $K_i$, a $n_i \times m$ matrix
**Output :** $K_o$, a $n_o \times m$ matrix
$s_1 \leftarrow 1$;
$s_2, s_0 \leftarrow K_i$.dim[0]; `/* Get length of dimension 0 */`
**for** $s_1$ *in range*$(s_0)$ **do**
  diffs $\leftarrow K_i[0, :]$ - $K_i[s_1, :]$;
  diffs $\leftarrow$ abs(diffs);
  **if** *any(diffs* $\geq$ *T)* **then**
    break;
  **end**
**end**
$s_1 \leftarrow s_1 - B$;
**if** $s_1 < 0$ **then**
  $s_1 = 0$
**end**
**for** $s_2$ *in range*$(s_0)$ **do**
  $s_2 \leftarrow s_0 - s_2$;
  diffs $\leftarrow K_i[s_2, :]$ - $K_i[s_0 - 1, :]$;
  diffs $\leftarrow$ abs(diffs);
  **if** *any(diffs* $\geq$ *T)* **then**
    $s_2 \leftarrow s_2 + B$;
    break;
  **end**
**end**
**if** $s_2 > s_0$ **then**
  $s_2 = s_0$
**end**
$K_o = K_i[s_1 : s_2, :]$; `/* Remove the still frames, keeping the middle */`
return $K_o$

## IV. Discussion

### A. NIE and GRS

*1) Analysis:* As many prior works were able to recover at least a subset of Novice, Intermediate, and Expert (NIE) ratings but not OSATS/GRS scores as seen in the meta-analysis of Table I, we decided to examine possible reasoning for this gap in performance in more detail. Surprisingly we found very little statistical significance between the NIE labels, which are based on time trained on the machine, and the GRS score given by experienced surgeons. As shown in Table II, each NIE class scored better on average than the others in at least one task. This is counter-intuitive to expectations, where Experts should consistently outperform Intermediates, who in turn should outperform Novices. As the NIE classes do not strongly correlate with actual performance, this in turn raises potential concerns about the ability of prior work to truly interpret the skill of a robotic surgeon rather than recover other patterns in operator movement.

This discrepancy in the data can be seen through a series of statistical tests, the results of which can be found in Table III. Here, performance along every axis of GRS rating exhibits very little relationship with the amount of robotic training an operator has, indicating that hours of training are a very poor metric of proficiency in JIGSAWS. The Needle Passing task shows no statistical significance between operators based on hours trained, while the other two seem to prove opposite conclusions. In Knot Tying, statistical differences exist between Novice and Expert users, as well as between Novice and Intermediate, but not between Experts and Intermediates. Conversely, the Suturing Task finds no difference between Novices and Experts, with Intermediate operators greatly outperforming the two. As each task has a different distribution between GRS and NIE rating, this indicates that even the statistically significant difference in each task is not maintained in general skill rating. This similarity among classes is evident in Figure 3, where we've applied Kernel Density Estimation [35] to estimate the probability distributions mapped on a normal Gaussian scale with minimal smoothing.

*2) Explanations:* There are several potential explanations for these findings. Firstly, the JIGSAWS dataset may have either a random flaw due to operators behaving outside their typical performance or the experimental setup was not well representative of actual surgical tasks. Support for the former can be found in the limitation on camera movement, which experienced operators may be more accustomed to [36] while concerning the latter; there is always the potential that the simulated tasks and materials within them differ greatly from the true surgical experience. However, in this case, we would still expect higher ratings for Experts in GRS elements such as "Suture/Needle Handling," "Flow of Operation," and "Time and Motion" which were not present. In many domains, including robotic surgery, it is generally believed that as individuals gain more experience, they acquire a deeper understanding of the task, develop refined techniques, and become more efficient and effective in their execution. Expertise comes with extensive experience and practice, which should translate into superior performance compared to less experienced individuals.

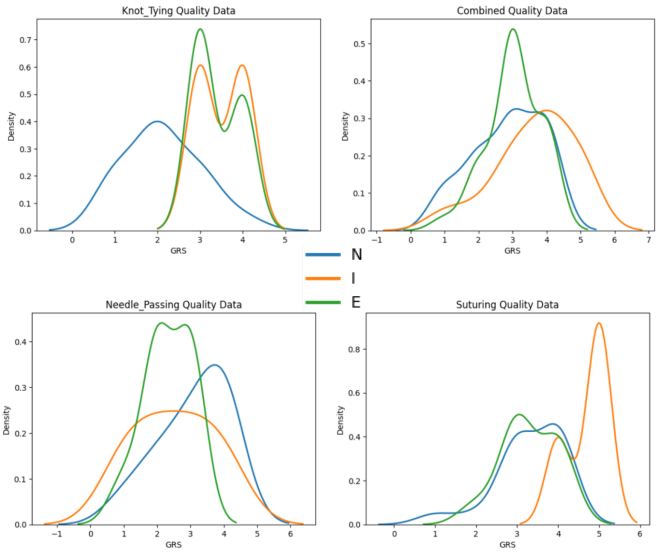


Fig. 3: Estimated Probability Density Functions showing the high but varying overlaps between different NIE classes.

An alternative explanation might be that the manual GRS scoring process is prone to significant variations, since the sole rating surgeon, being human, may not be accustomed to assessing performance in these simulated tasks. Additionally, as the GRS ratings were stated to be done blind, this removes doubt that data presents bias towards certain operators, influencing the rating outcomes. We can confirm this by finding meaningful relationships between the kinematics of the dataset, as presented in Section IV-B.

Further work is necessary to investigate the implications of these findings. One final possible explanation is that existing operator training methods may not translate well to teleoperation in practice, as suggested by the kinematic analysis. If further substantiated, this would indicate that the training and experience operators received did not substantially improve their performance in this dataset and, in some cases, impeded their results. This highlights the need for ongoing evaluation and potential revision of robotic surgery training programs, emphasizing the importance of continuous training.

### B. GRS and Kinematics

From Figure 1, it should be apparent that some robot-side kinematics have relationships with performance across all three tasks. It should be noted that as the operational area and direction do not seem to remain constant between tasks, the $X$, $Y$, and $Z$ directions may not be particularly indicative of movement within the task's operational area, and as such we allow for cross-direction analysis. In particular, acceleration along both the Patient Side Manipulators, as well as the rotational jerk (3rd order derivative of movement) of the second manipulator seems to be closely tied to scores given for handling, motion and quality in all three tasks[2]. This is

[2]Revelations exclusive to a single PSM may simply indicated operator handedness, as all 8 JIGSAWS participants were reportedly right-handed.

especially interesting as it indicates that the final outcome has direct ties to movement paths during the session, which while intuitive, is good to confirm. This also potentially contradicts the above explanation of poor labeling by the human grader, as actual kinematic trends are present. Others have found kinematic differences and skill levels to correlate in radical prostatectomy [37]. As the experience levels in JIGSAWS (the most common evaluation metric) do not align with the evaluated skill level; this does support the hypothesis that there are potential issues with the simpleness of the phantom tasks in JIGSAWS. To use this information, we propose Confidence Through Acceleration in Section IV-C2 when training human surgeons and evaluating both human and automated robotic surgeon performance.

There are also some interesting per-task relationships, most notably in the Suturing Task. In Figure 4, we present kinematic profiles that exhibit high correlations with skill rating variability in Suturing that are not found as prominently in the other tasks. This suggests that certain strong movement patterns are task-specific, which aligns with expectations. Notably, the deep purple (low correlation) regions follow a pattern that suggests a greater change in performance as GRS scores decrease; high-performing operators become gradually less dissimilar in certain motion-based profiles in certain tasks. JIGSAWS provides segmented gesture classification annotations for each trial, which may be analyzed to provide additional evidence and insights.

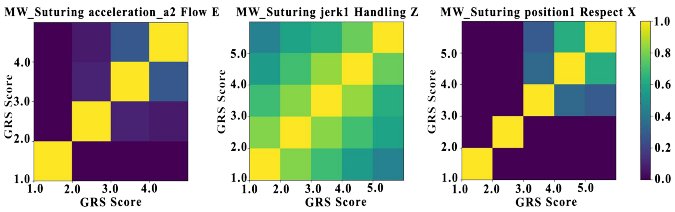


Fig. 4: Heat maps of Kinematic Profiles in Suturing for a few motion metrics across both arms.

### C. Learning Folds and Metrics

*1) Leave One Task Out:* JIGSAWS provides two suggested types of folds for cross-fold validation and ensuring generalization of results: Leave One Super-trial Out (LOSO) and Leave One User Out (LOUO). These breaks ensure intra-task generalization but do not effectively scale for cross-task analysis due to limitations in the dataset's representation of experts and intermediates. For instance, when applying LOUO, the training data often lacks the necessary intra-class variability, as demonstrated in [15]. To this end, we propose an additional fold type, Leave One Task Out (LOTO) in which deep/machine learning solutions are trained on two tasks and evaluated on a third. For example, a solution may be trained on Suturing and Knot-Tying, but then evaluated on Needle-Passing. In this way, the performance of the models is ensured to generalize to surgical performance as a whole rather than being limited to a single task.

*2) Metrics:* We propose two new metrics for JIGSAWS or JIGSAWS-like datasets: Correlation Weighted Mean Square Error (CWMSE) to evaluate the performance of skill rating systems; and Confidence Through Acceleration (CTA) to evaluate surgeon general performance across tasks using only robot-side kinematics.

**CTA**: There seems to be a clear relationship between the absolute value of an operator's mean acceleration and the range of accelerations they work at with the overall score. While not directly linear, this relationship persists across tasks in JIGSAWS and may have a reasonable intuitive explanation. Namely, operators who are confident of their next move switch into it more quickly than others while also operating at a wider range of accelerations as they adjust manipulators to ideal positions. This relationship between acceleration variance and higher constant speed denoting higher scores holds for both PSMs across all three tasks, as shown in Table IV.

| Motion | Task | Score | Mean | Std | CTA |
|---|---|---|---|---|---|
| PSM1 Accel | KT | 1 | -1.8376e-05 | 0.1351 | 2.4820e-6 |
| PSM1 Accel | KT | 2 | 7.1750e-05 | 0.2076 | 1.4897e-5 |
| PSM1 Accel | KT | 3 | 8.2151e-05 | 0.2495 | 2.0495e-5 |
| PSM1 Accel | NP | 1 | -4.7709e-05 | 0.1750 | 8.3502e-6 |
| PSM1 Accel | NP | 2 | 6.1302e-05 | 0.1436 | 8.80031e-6 |
| PSM1 Accel | NP | 3 | **4.2674e-05** | 0.1392 | **5.93931e-6** |
| PSM1 Accel | NP | 4 | 0.00014 | 0.1809 | 2.5366e-5 |
| PSM1 Accel | SU | 1 | 3.1511e-05 | 0.1724 | 5.4328e-6 |
| PSM1 Accel | SU | 2 | 0.00010 | 0.1731 | 1.7712e-5 |
| PSM1 Accel | SU | 3 | 0.00015 | 0.1568 | 2.3691e-5 |
| PSM1 Accel | SU | 4 | **0.00012** | 0.2467 | 3.1633e-5 |
| PSM1 Accel | SU | 5 | 0.00015 | 0.2969 | 4.5736e-5 |

TABLE IV: Alignment of the proposed Confidence Through Acceleration metric with the scores in JIGSAWS. Note that CTA **fails** to follow GRS Score for one trial in NP.

To use this relationship, we propose Equation 1 as a metric. In Equation 1, $X$ is the set of all accelerations throughout a session, $\bar{X}$ is the mean of $X$, $n$ is the number of elements in $X$, and $x_i$ denotes the $i$-th element of $X$. To be clear, we are not proposing optimization based on this, as greedy maximization will likely result in poor outcomes, but rather as an insight into surgeon performance.

$$CTA(X) = |\bar{X}|\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n}} \tag{1}$$

**CWMSE**: It should be noted that the correlation in Table I is Spearman's correlation rather than Pearson's. Spearman's correlation conveys the similarity in the ranked order of two sets rather than their similarity in a linear space. In Table V, we demonstrate how this may lead to confusion about the similarity of two sets with a simple example. Two sets, X and Y, consisting of 5 elements, are compared to their Spearman's and Pearson's correlations, along with the mean square error (MSE). This means that despite a high Spearman correlation, the two sets are quite distant in space which may not be ideal for GRS as it is interval data.

| Set | Elem 1 | Elem 2 | Elem 3 | Elem 4 | Elem 5 |
|---|---|---|---|---|---|
| X | 1 | 2 | 3 | 2 | 1 |
| Y | -300 | 1 | 1.00001 | 0.9999 | -20 |
| Spearman's | **0.94** | Pearson's | **0.58** | MSE | **18209** |

TABLE V: Difference between similarity measures.

$$CWMSE(X,Y) = \frac{\sum_{i=0}^{n}(y_i - x_i)^2}{n} \times \left(1 + \frac{\sum_{i=0}^{n}(\bar{X} - x_i)(\bar{Y} - y_i)}{\sqrt{\sum_{i=0}^{n}(\bar{X} - x_i)^2(\bar{Y} - y_i)^2}}\right) \quad (2)$$

For this reason, we propose Equation 2 be used instead, where $x_i$ and $y_i$ are the $i$-th element of $X$ and $Y$, $\bar{X}$ and $\bar{Y}$ are the means, and $n$ denotes the number of elements in each set. This way, the distance in space between the estimated and true values of the sets is reflected while also rewarding alignment between the particular categories of GRS.

## V. Conclusion

In this work, we identify several potential issues in the JIGSAWS dataset, most notably a lack of relationship between the commonly used NIE classes with actual operator performance, which indicates that results using it may have different impacts than previously understood. This gap may imply an issue with current training procedures for Robot-Assisted Minimally Invasive Surgery to be further investigated. Regardless of the true cause of this difference, we find strong support for the need for a new, robust dataset. Finally, we identify kinematic behaviors associated with strong performance using robot-side kinematics and present a non-task-specific metric to evaluate human and robot surgeon performance going forward.

## References


[1] R. H. Grogan, "Current status of robotic adrenalectomy in the united states," *Gland Surgery*, vol. 9, no. 3, p. 840, 2020.

[2] R. M. Flores and N. Alam, "Video-assisted thoracic surgery lobectomy (vats), open thoracotomy, and the robot for lung cancer," *The Annals of Thoracic Surgery*, vol. 85, no. 2, pp. S710–S715, 2008.

[3] J. H. Palep, "Robotic assisted minimally invasive surgery," *Journal of minimal access surgery*, vol. 5, no. 1, p. 1, 2009.

[4] A. Ghasem *et al.*, "The arrival of robotics in spine surgery: a review of the literature," *Spine*, vol. 43, no. 23, pp. 1670–1677, 2018.

[5] S. P. Díez *et al.*, "Evaluation of haptic feedback on bimanually teleoperated laparoscopy for endometriosis surgery," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 5, pp. 1207–1221, 2019.

[6] R. J. Trute *et al.*, "Development of a robotic surgery training system," *Frontiers in Robotics and AI*, vol. 8, p. 434, 2022.

[7] K. R. Wanzel, M. Ward, and R. K. Reznick, "Teaching the surgical craft: from selection to certification," *Current problems in surgery*, vol. 39, no. 6, pp. 583–659, 2002.

[8] J. D. Brown *et al.*, "Using contact forces and robot arm accelerations to automatically rate surgeon skill at peg transfer," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 9, pp. 2263–2275, 2017.

[9] K. Ahmed, D. Miskovic, A. Darzi, *et al.*, "Observational tools for assessment of procedural skills: a systematic review," *The American Journal of Surgery*, vol. 202, no. 4, pp. 469–480, 2011.

[10] Y. Ren, T. J. Loftus, S. Datta, M. M. Ruppert, *et al.*, "Performance of a machine learning algorithm using electronic health record data to predict postoperative complications and report on a mobile platform," *JAMA Network Open*, vol. 5, no. 5, pp. e2 211 973–e2 211 973, 2022.

[11] H. Ismail Fawaz *et al.*, "Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks," *International journal of computer assisted radiology and surgery*, vol. 14, pp. 1611–1617, 2019.

[12] A. Zia and I. Essa, "Automated surgical skill assessment in rmis training," *International journal of computer assisted radiology and surgery*, vol. 13, pp. 731–739, 2018.

[13] M. J. Fard, S. Ameri, R. Darin Ellis, *et al.*, "Automated robot-assisted surgical skill evaluation: Predictive analytics approach," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 14, no. 1, p. e1850, 2018.

[14] G. Lajkó, R. Nagyné Elek, and T. Haidegger, "Endoscopic image-based skill assessment in robot-assisted minimally invasive surgery," *Sensors*, vol. 21, no. 16, p. 5412, 2021.

[15] I. Funke *et al.*, "Video-based surgical skill assessment using 3d convolutional neural networks," *International journal of computer assisted radiology and surgery*, vol. 14, pp. 1217–1225, 2019.

[16] J. L. Lavanchy *et al.*, "Automation of surgical skill assessment using a three-stage machine learning algorithm," *Scientific Reports*, vol. 11, no. 1, p. 5197, 2021.

[17] A. L. Trejos, R. V. Patel, R. A. Malthaner, and C. M. Schlachta, "Development of force-based metrics for skills assessment in minimally invasive surgery," *Surgical endoscopy*, vol. 28, pp. 2106–2119, 2014.

[18] E. Colleoni *et al.*, "Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2714–2721, 2019.

[19] L. Zhang *et al.*, "Real-time surgical tool tracking and pose estimation using a hybrid cylindrical marker," *International journal of computer assisted radiology and surgery*, vol. 12, pp. 921–930, 2017.

[20] Z. Jian, W. Yue, Q. Wu, W. Li, *et al.*, "Multitask learning for video-based surgical skill assessment," in *2020 Digital Image Computing: Techniques and Applications (DICTA)*, 2020, pp. 1–8.

[21] Y. Ming *et al.*, "Surgical skills assessment from robot assisted surgery video data," in *2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA)*. IEEE, 2021, pp. 392–396.

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

[23] J. Martin, G. Regehr, *et al.*, "Objective structured assessment of technical skill (osats) for surgical residents," *British journal of surgery*, vol. 84, no. 2, pp. 273–278, 1997.

[24] V. Datta, S. Bann, M. Mandalia, and A. Darzi, "The surgical efficiency score: a feasible, reliable, and valid method of skills assessment," *The American journal of surgery*, vol. 192, no. 3, pp. 372–378, 2006.

[25] H. Niitsu, N. Hirabayashi, M. Yoshimitsu, T. Mimura, J. Taomoto, *et al.*, "Using the objective structured assessment of technical skills (osats) global rating scale to evaluate the skills of surgical trainees in the operating room," *Surgery today*, vol. 43, pp. 271–275, 2013.

[26] G. Lajkó *et al.*, "Surgical skill assessment automation based on sparse optical flow data," in *2021 IEEE 25th International Conference on Intelligent Engineering Systems (INES)*, 2021, pp. 000 201–000 208.

[27] A. Soleymani *et al.*, "Surgical skill evaluation from robot-assisted surgery recordings," in *2021 International Symposium on Medical Robotics (ISMR)*, 2021, pp. 1–6.

[28] D. Liu *et al.*, "Towards unified surgical skill assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9522–9531.

[29] J. Gao *et al.*, "An asymmetric modeling for action assessment," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*. Springer, 2020, pp. 222–238.

[30] E. Yanik and *et al.*, "Deep neural networks for the assessment of surgical skills: A systematic review," *The Journal of Defense Modeling and Simulation*, vol. 19, no. 2, pp. 159–171, 2022.

[31] T. D. Nagy and T. Haidegger, "Performance and capability assessment in surgical subtask automation," *Sensors*, vol. 22, no. 7, p. 2501, 2022.

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

[33] Z. Li *et al.*, "Surgical skill assessment via video semantic aggregation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 410–420.

[34] H. B. Mann and D. R. Whitney, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," *Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50 – 60, 1947.

[35] M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," *Annals of Mathematical Statistics*, pp. 832–837, 1956.

[36] A. J. Hung *et al.*, "Utilizing machine learning and automated performance metrics to evaluate robot-assisted radical prostatectomy performance and predict outcomes," *Journal of Endourology*, vol. 32, no. 5, pp. 438–444, 2018.

[37] A. J. Hung and Others, "Development and validation of objective performance metrics for robot-assisted radical prostatectomy: a pilot study," *The Journal of urology*, vol. 199, no. 1, pp. 296–304, 2018.