Multi-modal Deep Learning for Assessing Surgeon Technical Skill on a Surgical Knot-tying Task

Kevin Kasa [1], David Burns [2], Mitchell Goldenberg [2], Omar Selim [2], Cari Whyne [2], and Michael Hardisty [2]

[1]Sunnybrook Research Institute
[2]Affiliation not available

June 21, 2022

Abstract

This paper introduces a new dataset of a surgical knot-tying task, and a multi-modal deep learning model that achieves comparable performance to expert human raters on this skill assessment task. Seventy-two surgical trainees and faculty were recruited for the knot-tying task, and were recorded using video, kinematic, and image data. Three expert human raters conducted the skills assessment using the Objective Structured Assessment of Technical Skill (OSATS) Global Rating Scale (GRS). We also designed and developed three deep learning models: a ResNet-based image model, a ResNet-LSTM kinematic model, and a multi-modal model leveraging the image and timeseries kinematic data. Results: All three models demonstrate performance comparable to the expert human raters on most GRS domains. The multi-modal model demonstrates the best overall performance, as measured using the mean squared error (MSE) and intraclass correlation coefficient (ICC). We found that multi-modal deep learning has the potential to replicate human raters on a challenging human-performed knot-tying task. As objective assessment of technical skill continues to be a growing, but resource-heavy, element of surgical education, this study is an important step towards automated surgical skill assessment, ultimately leading to reduced burden on training faculty and institutes.

Multi-modal Deep Learning for Assessing Surgeon Technical Skill on a Surgical Knot-tying Task

Kevin Kasa, David Burns, Mitchell G. Goldenberg, Omar Selim, Cari Whyne, Michael Hardisty

***Abstract*—*Goal:* This paper introduces a new dataset of a surgical knot-tying task, and a multi-modal deep learning model that achieves comparable performance to expert human raters on this skill assessment task. *Methods:* Seventy-two surgical trainees and faculty were recruited for the knot-tying task, and were recorded using video, kinematic, and image data. Three expert human raters conducted the skills assessment using the Objective Structured Assessment of Technical Skill (OSATS) Global Rating Scale (GRS). We also designed and developed three deep learning models: a ResNet-based image model, a ResNet-LSTM kinematic model, and a multi-modal model leveraging the image and time-series kinematic data. *Results:* All three models demonstrate performance comparable to the expert human raters on most GRS domains. The multi-modal model demonstrates the best overall performance, as measured using the mean squared error (MSE) and intraclass correlation coefficient (ICC). *Conclusion:* Multi-modal deep learning has the potential to replicate human raters on a challenging human-performed knot-tying task. *Significance:* As objective assessment of technical skill continues to be a growing, but resource-heavy, element of surgical education, this study is an important step towards automated surgical skill assessment, ultimately leading to reduced burden on training faculty and institutes.**

I. Introduction

THERE has been a gradual evolution in surgical education towards objective assessment of competence as a requirement for trainee advancement and an increased reliance on simulation-based training [1]. This paradigm responds to mounting pressures to shorten the surgical trainee work-week, and improve operating room efficiency and safety at teaching institutions. However, competency-based medical education (CBME) can icrease the burden on supervising surgical faculty and increase program reliance on the objectivity and validity of their CBME assessments [2].

Deep learning offers the ability to tackle these challenges by automating some surgical skills assessments, potentially improving their objectivity and reducing the burden of CBME on training faculty and institutes. Deep learning is well suited for tackling technical skills assessment due to its robustness to noise and flexibility to learn an optimal feature set representative of task performance from large, unstructured, and multi-modal data sources. This is especially powerful when combined with low-cost data-collection devices and accessible high-performance computing resources.

In contrast to deep learning, "classical" machine learning algorithms (e.g., random forest, K-Nearest Neighbour, Linear Regression) rely on feature sets that are hand-crafted or heuristically extracted from the data to use in the classification or regression task of interest. Deep learning algorithms simultaneously extract the most pertinent features from the data while performing the downstream classification or regression task. Further, deep learning benefits from the large amounts of data that can be collected, and its performance improves with greater quantity and variety of available data. Deep learning has been show to outperform other methods on problems relying on unstructured data such as images or time series, including object detection, image classification, and speech recognition.

Deep residual models (ResNets) are particularly powerful in training deeper neural networks with increased capacity to learn and model complicated relationships, achieving state-of-the-art performance on many image recognition tasks [3]. These improvements largely stem from the use of "skip connections", or residual blocks, between layers which allow for deeper networks without suffering from vanishing gradient problems. This ability to effectively train very deep networks is the major advantage of the ResNet architecture. Although ResNet's are often employed in image related tasks, they can also be implemented using one-dimensional convolutions for time-series data. Another innovation that lends itself to timeseries analysis is the LSTM, or Long Short-Term Memory, network. These are a specific kind of recurrent neural networks (RNN) designed to learn long-term dependencies in time-series data, and have been used in language-modeling tasks, robotic control, and human activity recognition.

Although these advances in deep learning algorithms present the opportunity to automate some surgical technical skill assessments, previous research in this area has largely relied on classical machine learning algorithms leveraging engineered features in the data to classify performance, and thus far has largely been used to assess, global but not domain-specific, performance [4], [5], [6], [7], [8]. Deep learning has also been used successfully for object detection in skill assessment tasks [9], [10].

Further, available machine learning datasets for surgical skill assessment have thus far had limited sample sizes to train and thoroughly evaluate deep learning models — for example the widely-used JIGSAWS dataset consists of eight total subjects and only two "experts" [11]. The JIGSAWS

Kevin Kasa is with the Holland Bone and Joint Program at the Sunnybrook Research Institute

David Burns and Cari Whyne are with the Holland Bone and Joint Program at Sunnybrook Reserach Institute, the Institute of Biomedical Engineering at the University of Toronto, and the Department of Surgery at the University of Toronto

Mitchell G. Goldenberg is with the Department of Surgery at the University of Toronto

Omar Selim is with the Department of Surgery at the Royal Victoria Regional Health Center

dataset consists of video and kinematic data captured using a DaVinci Robotic system [12]. Indeed, many previous studies focus solely on data acquired using robotic systems or virtual simulators [13], [14], [15], and not on human-performed surgical tasks. Additionally, existing research generally relies on classifying performance into broad categories (beginner, intermediate, advanced), detecting general events such as incisions, or categorizing surgical tasks (suturing, knot tying, etc.) [15]. Some studies [13], [16] do predict OSATS scores in a regression framework. However, previous studies often only report metrics comparing the model performance to ground truth data, and do not present in-depth comparisons between the model predictions and expert human evaluators [14], [5].

In this paper, we introduce a novel dataset of seventy-two surgical trainees and surgeons performing a one-handed knot-tying task during a University of Toronto Department of Surgery Prep Camp and Orthopaedics Bootcamp. This consists of image, video, and kinematic data of the simulated surgical task, as well as skills assessment evaluations performed by three expert raters. We also develop, and evaluate three deep learning models to automate the surgical skills assessment: an image-based model, a kinematic-based model, and a multi-modal model leveraging both image and kinematic data. All three models are evaluated using the mean squared error and intraclass correlation coefficient metrics, with the multi-modal exhibiting the best performance. We demonstrate that this model rates the surgical trainees with comparable performance to the expert human raters on three of the four OSATS domains. A preliminary version of this work has been reported in [17].

II. Methods

This project seeks to develop and validate deep learning models for automated surgical skill assessment, specifically for the assessment of technical skill for a simulated knot-tying task. To facilitate this, 72 participants performed a knot-tying task, which were subsequently rated by human experts. Video and kinematic data of the task was recorded, as well as a photograph of the final product. In this study, the anonymized video recording was used for assessment by the human raters; the machine learning models used only image and kinematic data. This study was approved by the Sunnybrook Health Sciences Center Research Ethics Board (REB) on August 1, 2018 (REB protocol # 248-2018), and the Mount Sinai Hospital REB on June 22, 2018 (REB protocol # 18-0149-E).

A. Surgical Task

Seventy-two surgical trainees and surgeons were recruited for participation in this study during the 2018 University of Toronto Department of Surgery Prep Camp and Orthopaedics Bootcamp suturing modules. Participants performed a simulated vessel ligature task using one-handed knot-tying with 0-silk ties on polypropylene tubing. Each participant performed the task five times consecutively, with each performance as a separate task. No feedback was provided to participants between executions of the task. The overall goal of this task is to determine if the trainee's can correctly tie off, or occlude, the simulated blood vessel using the silk suture.

B. Data Collection

The vessel ligature tasks were recorded using three modalities, which are visualized in Figure 1:

1) High resolution digital photograph of the final product
2) Anonymized video recording of the operative field
3) 3D kinematic motion tracking of the hands using a Leap Sensor

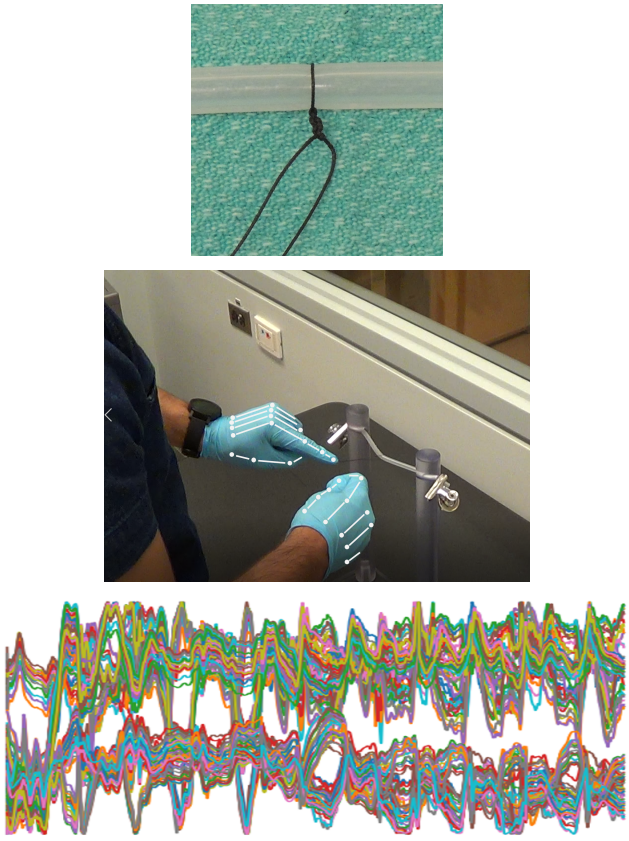

Fig. 1: The trials were recorded using three modalities. The top is an image of the final product, the middle is a screen capture of the video data with a visualization of the joints tracked by the Leap sensor. The bottom is an example of the kinematic time series data, representing the temporal 3-dimensional movement of the hand joints during the knot tying task.

C. Task Ratings

Three blinded independent raters conducted the technical skills assessment from the recorded video and photograph of the final product. The raters were senior surgical residents (PGY4 and above) with expertise in the assessed skill. Performance at the simulated surgical task was assessed by each rater using the Objective Structured Assessment of Technical Skill (OSATS) Global Rating Scale (GRS) [18] on the following four domains:

1) Respect for Tissue

2) Time and Motion
3) Quality of Final Product
4) Overall Performance

Each domain was scored on a 5-point scale (1-5). All raters were oriented to the OSATS GRS and domain specific anchors using example performances and suggested ratings. An example of the rating scale used by the human raters can be seen in Table I.

It was also important to ensure that the dataset was collected from a diverse and representative set of participants, including diversity in aspects such as surgical division, and prior experience level. The plurality of participants were from the division of orthopaedics, with participants from nine other surgical divisions included. Most participants were also Post-Graduate Year 1 (PGY1) trainees, with experience levels ranging up to Staff and Fellow surgeons.

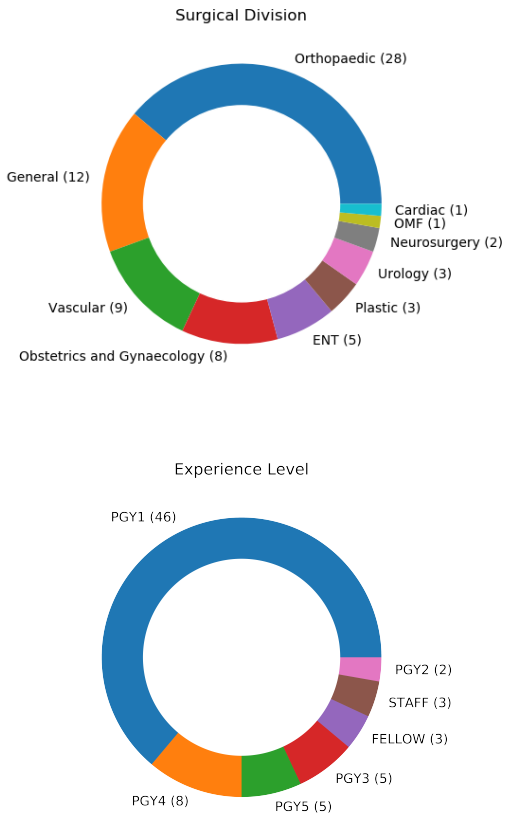

Fig. 2: Participants came from 10 surgical divisions, with experiences ranging from PGY1 to Fellow.

The sequence of tasks was randomized so that the raters were not consecutively exposed to tasks performed by the same individual. Further, the randomization was seeded separately for each rater, providing each rater with a different random order of tasks to assess. Forty random samples were also selected to be rated a second time by each rater for test-retest reliability assessment.

Domain	Rating Scale
Respect for Tissue	1 - Very poor: Frequent or excessive pulling or sawing of tissue 3 - Competent: Careful handling of tissue with occasional sawing or pulling 5 - Clearly superior: Consistent atraumatic handling of tissue
Time and Motion	1 - Very poor: Many unnecessary movements 3 - Competent: Efficient time/motion but some unnecessary moves 5 - Clearly superior: Clear economy of movement and maximum efficiency
Quality of Final Product	1 - Very poor 3 - Competent 5 - Clearly superior
Overall Performance	1 - Very poor 3 - Competent 5 - Clearly superior

TABLE I: Rating scale used when evaluating surgical skill on the GRS Domains.

D. Data Pre-processing

The three-dimensional position data of each joint in the phalanges from both hands was extracted from the Leap Motion Sensor's kinematic data capture. This 120 channel timeseries data was used as input into the deep learning models. The kinematic models require a fixed-length input, and the trials were not uniform in length. The Seglearn library [19] was used to truncate or zero-pad each data sample to a length of 4223 samples, which represents the 90th percentile of the sample lengths. This means that most samples were padded instead of truncated, so that as much information as possible was preserved. With a sampling rate of 110 Hz, this 4223 timestamp sequence is approximately 36 seconds long.

The Python implementation of OpenCV was used to pre-process the image data. The images were first temporarily masked to a binary image, isolating the black suture from the background. A dilating operation was applied to this image to enlarge the knot center. The OpenCV blob detector was then used to detect the suture knot, and a 512x512 bounding box was drawn around the center. The cropped image was then unmasked back to full RGB color. The kinematic and image data were also normalized between [0,1]. This is a standard deep learning procedure to speed computation time and avoid local minima in model optimization.

E. Data Augmentation

Although our dataset is not small relative to other relevant datasets, deep learning almost always benefits from larger quantities of data. Thus, the entire dataset was randomly oversampled to increase the number of training examples. Additionally, the trials with ratings that were greater or less than one standard deviation from the mean were further oversampled by a factor of three. By more evenly balancing the score distribution, the network can better learn to predict these minority classes.

However, increasing the size of the dataset without introducing any variation may lead to degraded performance, as the network may rely on memorizing specific features

of the training data and fail to generalize to unseen data. Data augmentation may be used to alter the input instances, thus artificially increasing the variety of training data and the networks ability to generalize. To minimize the model overfitting to the training data, the oversampled data was also augmented prior to input into the networks. The images were augmented with random 90-degree rotations and reflections about the x- or y-axis, largely to help mirror the varying knot orientation in the real data. The kinematic data was augmented based on recommendations in previous literature [20]: random rotations, reflections, and injection of Gaussian noise.

F. Machine Learning Models

We developed and analyzed three deep learning models. The first uses the RGB image data of the simulated vessel and ligature as input and the Quality rating as output. The second model uses the hand kinematic data as input and predicted the three other domains (Respect for Tissue, Time and Motion, and Overall Performance). The final is a composite model containing both RGB and kinematic modalities and output all four GRS rating domains. The video data was not used by the model.

The models were trained in a supervised regression learning framework, with the mean scores of the three expert raters as the ground truth. We trained the models to minimize a mean-squared error loss, however the number of output targets varied between the models, since some predicted only one OSATS domain and others multiple.

$$L = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}(y_i^k - \hat{y_j^k}) \tag{1}$$

Here, N is the number of samples in the training batch, and K is the number of output targets. For example, the image-only model has a K=1 since it predicts only the Quality score, whereas the multi-modal has K=4 since all four domains are predicted.

The image model is depicted depicted in the bottom branch of Figure 3, and consists of a ResNet-50 backbone with pre-trained weights from the ImageNet dataset. Prior to input, the images were resampled to 1024x1024, further cropped 30% tighter, and normalized based on the ImageNet metrics. For the first 200 epochs, the ResNet layers were frozen and only the final dense layer was trained. Subsequently, the learning rate was reduced and the top layers of the ResNet model were fine-tuned for another 200 epochs.

Previous works demonstrate that convolutional-recurrent neural networks can been used to successfully perform human activity recognition from kinematic data [21], [22]. In our work, the network is tasked with scoring surgical skill across multiple domains from a relatively high-dimensional dataset (120 channels). To ensure the network has the capacity to perform these tasks, a one-dimensional ResNet-18 model is used as a feature extractor on the kinematic data. The extracted features are then inputted into two bi-directional LSTM layers to model the temporal nature of the data. Finally, three dense layers are used to score the 'Overall Performance', 'Respect for Tissue', and 'Time and Motion' from the learned features. This model was trained for 200 epochs, and the architecture can be seen in the top branch of Figure 3.

The previous two models are combined so that all four GRS domains can be scored. The time series and image networks are trained concurrently, and the extracted feature sets are concatenated. These are then inputted into fully-connected layers to perform the final task scoring for each domain, as seen in Figure 3. The 2D ResNet network also leveraged pre-trained ImageNet weights and followed a fine-tuning scheme similar to that described above, where the ResNet layers were initially frozen for 50 epochs and used solely as a feature extractor, followed by fine-tuning the top layers of the ResNet for another 50 epochs.

G. Statistical Analysis

The collected dataset was analyzed to ensure its reliability and validity prior to being used for training and evaluating the deep learning models. The analysis of the expert human raters also serves as a baseline for understanding the model's best achievable performance. The Intraclass Correlation Coefficient (ICC) and Standard Error of Measure (SEM) were used to analyze the human and AI ratings for agreement and consistency. To assess the interrater reliability on the entire collected dataset, the ICC (2,3), ICC(2,1), and SEM scores were calculated for each of the GRS domains [23]. The ICC (2, 3) model is selected since our raters are chosen as representative of a larger population, and the mean of the three raters is used as the ground-truth. The ICC (2,1) was also used to assess the human raters on their test-retest consistency, using the randomly repeated trials that were rated twice. Our hypothesis was that the human raters show moderate to good agreement on the GRS domains and good consistency in their ratings.

In addition to measuring the average human rater reliability on the entire dataset, we also looked at the ICC score of the raters on the held-out testing subset of the data. Since the AI models were evaluated on this test set, finding the human rater's reliability on this subset alone can allow for a more direct comparison with the network performance.

The experience levels of the participants and their ratings were also investigated to help establish construct validity. A one-way ANOVA was performed between the beginner (PGY1 & PGY2, n=48), intermediate (PGY3, PGY4, & PGY5, n=18), and expert (Staff & Fellow, n=6) level participants. A Tukey–Kramer post-hoc test was then done to determine which groups were different from each other. These tests were all done using the participants performance on the "Overall Performance" GRS domain.

Several tests were done to evaluate the model's performance. The point difference between the model's predictions and the human ratings with the ground truth was evaluated using the mean squared error (MSE). The goodness of fit of the model was evaluated using R^2. Finally, the agreement amongst each (human or AI) rater and the ground truth was determined using the ICC (2,1) score. This means that the ICC between the AI ratings and the ground truth was determined,

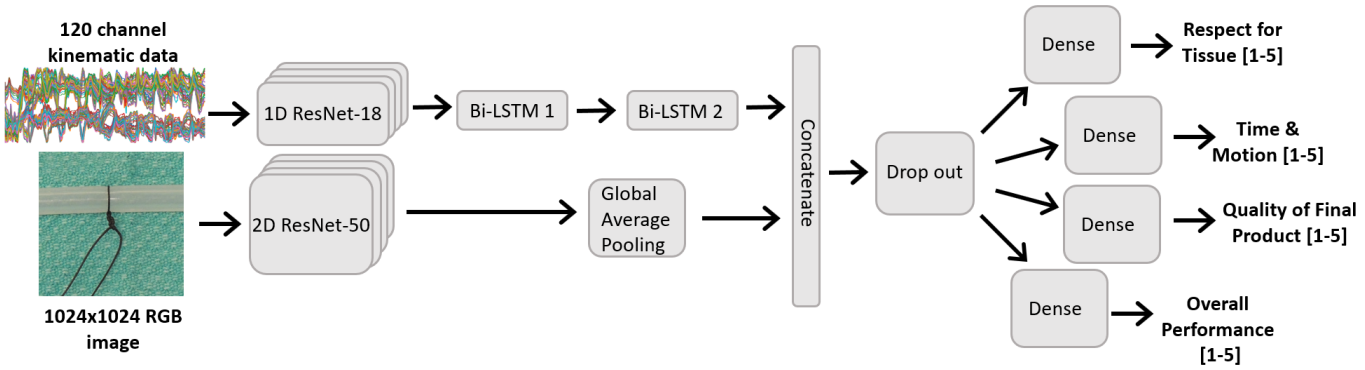

Fig. 3: Images were analyzed using a ResNet-based network, and the kinematic data was analyzed using a 1D ResNet-18 as a 'feature extractor', followed by 2 bidirectional LSTM layers. The combined multi-modal network is concurrently trained on both the image and kinematic data as input, and predicts all four GRS domains.

as well as the ICC between each human rater and the ground truth. This allows us to consider how our model performs as a single generalized rater [23] in terms of its agreement with the ground-truth data, as well as compare the AI agreement with that exhibited by the humans. Our hypothesis was that the AI would demonstrate comparable point errors (MSE) and agreement (ICC) with the ground truth data as the human raters.

Although previous research seeking to directly predict GRS scores is sparse, existing studies report performance using the mean Spearman Correlation Coefficient ρ across the predicted vs true GRS scores [16]. For consistency in the reported metrics, we also evaluate the Spearman Coefficient on the multi-modal model.

Finally, some studies that directly predict the GRS domain scores report their performance in terms of accuracy [13]. For a comparable metric, we also find the accuracy of our multi-modal model. Since our predictions are continuous and accuracy deals with discrete data, we first round the ground-truth and model predictions; for example a score of 2.7 will get rounded to 3.0, which is necessary to compute the accuracy metric. Our model is designed to predict continuous scores so this is not a perfect metric, but serves to gain a general comparison with previous studies.

III. Results

A. Dataset Analysis

The human raters showed ICC scores corresponding to moderate agreement on the four GRS domains, when measured on the entire collected dataset, as summarized in Table II.

GRS Domain	ICC (2,3)	SEM (2,3)	ICC (2,1)	SEM (2,1)
Respect for Tissue	0.71	0.45	0.47	0.62
Time and Motion	0.70	0.47	0.44	0.64
Quality of Final Product	0.83	0.40	0.63	0.61
Overall Performance	0.73	0.39	0.47	0.55

TABLE II: The expert human raters demonstrate moderate to good agreement on their evaluations when as measured using the mean. The AI model was trained & evaluated on the mean value of the ratings.

GRS Domains	Rater 1		Rater 2		Rater 3	
	ICC	SEM	ICC	SEM	ICC	SEM
Respect for Tissue	0.84	0.43	0.49	0.55	0.55	0.54
Time and Motion	0.83	0.46	0.57	0.58	0.62	0.48
Quality of Final Product	0.88	0.40	0.79	0.47	0.69	0.43
Overall Performance	0.85	0.37	0.60	0.49	0.58	0.48

TABLE III: Test-retest performance of the human raters on the forty repeated trials. Although the raters performance varies, they all show moderate to good consistency.

GRS Domain	ICC (2,3)	SEM (2,3)	ICC (2,1)	SEM (2,1)
Respect for Tissue	0.78	0.44	0.54	0.63
Time and Motion	0.81	0.41	0.58	0.61
Quality of Final Product	0.93	0.30	0.82	0.49
Overall Performance	0.86	0.30	0.68	0.30

TABLE IV: Human raters show good to excellent agreement on the held-out test set. Determining agreement on the same test set the AI model is evaluated on can help provide a better baseline for expected performance.

There is some variance in the test-retest performance of the human raters, with ICC scores ranging from 0.49 to 0.88, and SEM ranging from 0.37 to 0.58. Overall Rater 1 demonstrated better consistency amongst their ratings than Rater 2 or 3. Although some raters performed better than others, overall they all show moderate to good consistency, and the results are summarized in Table III.

On the held-out test set, the human raters show good to excellent agreement, as seen in Table IV. Greater agreement is seen on this smaller subset of the overall data likely because there are fewer samples for the human raters to disagree on.

The one-way ANOVA returned a p-value of 0.0038, suggesting there was a significant performance difference amongst the surgeon experience groups. The Tukey analysis resulted in a significant difference between the Beginner (n = 48, mean = 2.31) and Intermediate (n = 36, mean = 2.79) groups (p = 0.003), and no significance between the Expert group (n = 6, mean = 2.50) and either of the two groups. The lack of significance in the Expert group may be due to the relatively small sample size compared to the other two.

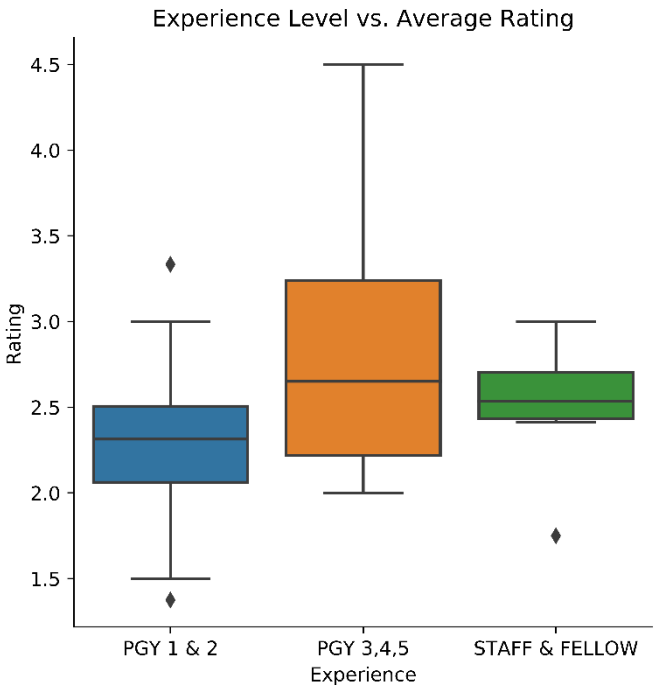

Fig. 4: Participant experience and rating on the 'Overall Performance' domain. A significant difference was found between the Beginner and Intermediate groups.

B. Deep Learning Model Performance

The kinematic, image, and multi-modal models were all trained and evaluated independently of each other on the same reserved testing set. The model performance was evaluated by how well it can predict the mean OSATS GRS ratings provided by the raters, as well as the intrarater reliability between the model predictions and the expert raters.

Table V highlights the performance relative to the ground-truth. For a direct comparison with the human performance, the same metrics are presented for each individual rater's score compared to their mean scores, for the test-set trials. These metrics serve as an understanding for how close the model predictions are to the dataset's ground truth. The model's predictions do appear close to the ground-truth, with lower point errors than two of three human raters, and with the multi-modal model exhibiting the lowest point error overall.

The error between the ground-truth and the model predictions, as well as human ratings, is also seen in Figure 5. The improvements of the multi-modal model is particularly noted on the Overall Performance domain.

The AI model demonstrates ICC scores ranging from 0.3 to 0.90, with the human raters ranging from 0.60 to 0.92. The multi-modal model demonstrates better agreement based on the ICC and SEM than the kinematic or image-only models on all domains except for Respect for Tissue. The multi-modal model also demonstrates better agreement with the ground truth than 2 of the 3 human raters on the Overall Performance and Quality of Final Product domains, however its performance is poorer on the remaining two domains.

The Spearman Correlation Coefficient, ρ, of our multi-modal model is reported in Table VII. This represents the correlation between the model's predictions and the ground truth.

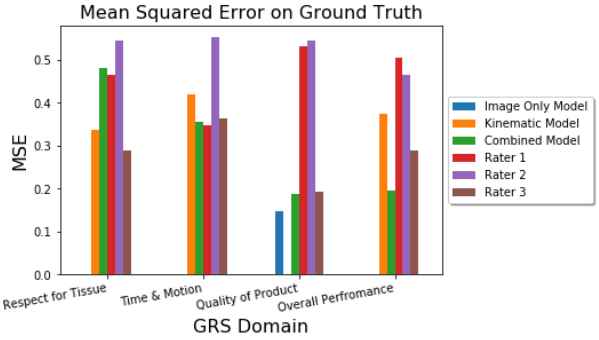

Fig. 5: Graphical comparison of the MSE on the GRS Domains — lower MSE is better.

The discretized scores are used to evaluate the model's accuracy, and are summarized in Table VIII. As mentioned, accuracy is not a perfect metric for our continuous data predictions, however it is indicative of the difference between the predictions and ground-truth on the datasets.

Overall, the multi-modal model demonstrates comparable results to the humans on most of the GRS domains. The AI has a lower point error on the ground truth scores than the human raters on three of the four GRS domains, as exhibited by the lower MSE. The ICC metrics suggest that in general, the human raters are in better agreement with the ground-truth scores. The multi-modal model demonstrates the best performance, with higher ICC on some domains (e.g. Quality of Final Product) than two of the three raters.

IV. Discussion

This paper presented a new dataset consisting of multi-modal recordings (image, video, & kinematic) of a simulated surgical knot-tying task, with skill assessment conducted by expert human raters based on the OSATS GRS framework. A thorough statistical analysis was conducted to ensure the validity of the dataset. Three deep-learning models were trained and evaluated on this dataset: a ResNet-50 image model, a unique "ResLSTM" kinematic model, and a combined multi-modal model.

All three models were able to successfully perform the skills assessment, with the multi-modal model performing the best overall. In comparison to previous studies conducted on the JIGSAWS dataset [11], which contains video and kinematic data from eight surgeons performing three surgical actions (knot tying, needle passing, and suturing) using the DaVinci Robotic System [12], out multi-modal model achieves better performance. For comparison, previous literature report a mean Spearman Correlation of $\rho = 0.65$ on the knot-tying task in the JIGSAWS dataset [16], as seen in Table VII. This means that on average, our multi-modal model demonstrates better correlation between its predictions and the ground-truth on our dataset, than reported on similar datasets in previous literature. Further, Khalid et. al. [13] present a study that

Model	Metric	Respect for Tissue	Time and Motion	Quality of Final Product	Overall Performance
Image Model	MSE	-	-	***0.146***	-
	R2	-	-	0.778	-
Kinematic Model	MSE	***0.336***	0.420	-	0.373
	R2	***0.337***	0.244	-	0.453
Multi-modal Model	MSE	0.480	***0.356***	0.186	***0.194***
	R2	0.136	***0.476***	***0.838***	***0.618***
Rater 1	MSE	0.464	**0.348**	0.531	0.505
Rater 2	MSE	0.546	0.553	0.545	0.466
Rater 3	MSE	**0.288**	0.363	0.193	0.290

TABLE V: Mean Squared Error (MSE) of the AI predictions and human ratings, compared to the ground truth (mean of human scores).

Model	Metric	Respect for Tissue	Time and Motion	Quality of Final Product	Overall Performance
Image Model	ICC(2,1)	-	-	0.888	-
	SEM(2,1)	-	-	***0.257***	-
Kinematic Model	ICC(2,1)	***0.477***	***0.621***	-	0.534
	SEM(2,1)	***0.464***	0.441	-	0.416
Multi-modal Model	ICC(2,1)	0.301	0.591	***0.904***	***0.746***
	SEM(2,1)	0.499	***0.428***	0.309	***0.305***
Rater 1	ICC(2,1)	0.717	0.779	0.823	0.616
	SEM(2,1)	0.476	0.414	0.512	0.502
Rater 2	ICC(2,1)	0.606	0.627	0.758	0.508
	SEM(2,1)	0.516	0.524	0.521	0.689
Rater 3	ICC(2,1)	**0.797**	**0.797**	**0.924**	**0.789**
	SEM(2,1)	**0.377**	**0.423**	**0.308**	0.379

TABLE VI: Intraclass Correlation Coefficient (ICC) and Standard Error of Measurement (SEM) scores between the ground truth and the AI models & human raters.

GRS Domain	ρ	
	Multi-modal model (ours)	FCN [16]
Respect for Tissue	0.18	-
Time and Motion	0.73	-
Quality of Final Product	0.95	-
Overall Performance	0.82	-
Mean	**0.67**	0.65

TABLE VII: Spearman Correlation Coefficient between the multi-modal AI predictions and the ground truth. Best performing model on the JIGSAWS dataset included as reference [16].

GRS Domain	Accuracy	
	Multi-modal model (ours)	Embedding Analysis [13]
Time and Motion	**0.54**	0.32
Quality of Final Product	**0.76**	0.51
Overall Performance	**0.76**	0.41

TABLE VIII: Accuracy of the multi-modal model, determined by first rounding the continuous ground-truth and predicted scores. Best performing model on the JIGSAWS dataset included as reference [13].

directly predicts the GRS scores in a regression fashion, using the video data of the JIGSAWS dataset. As seen in Table VIII, they report a mean accuracy of 0.32 for Time and Motion, 0.51 for Quality of Final Product, and 0.41 for Overall Performance.

This is particulalry encouraging as assessing surgical skill from human performed knot-tying is seemingly more challenging than evaluating a robotically operated dataset. This result means that our model can be used in a wider range of environments and facilities, where robotic surgery systems are not available for surgical trainees or faculty. Further, while some studies attempt to indirectly compute performance metrics for surgical skill [9], [10], our model directly predicts performance on the GRS domains and provides the most pertinent assessment of surgical skill to trainees.

The AI performance was comparable to the human rater on three out of the four GRS domains. Further experiments are required to determine why the model consistently struggles on the Respect for Tissue domain. A hypothesis is that since the Leap Sensor is only tracking the subjects hands, important information on the handling of the "tissue" (or polypropylene tubing) is not captured using this modality. Respect for Tissue was better assessed on video which was available to raters but not used by the model. Future analysis will investigate leveraging the video modality within the multimodal model to improve performance on this domain.

The image-only model was trained solely on the Quality of Final Product domain, since it is not likely that the images alone contain enough relevant information for this model to perform well on the other categories (e.g., Time and Motion). Smaller models were investigated for this task, such as 5- and 7-layer convolutional neural networks, however these all exhibited poor performance in the rating task and were abandoned. This suggests that the ResNet's increased capacity to extract important and meaningful features from the image data is important in assessing surgical skill. Similarly, shallow recurrent neural networks exhibited poor performance on the kinematic data and were also discarded. Learning to score various categories of surgical skill is a complex task and these models likely did not have the capacity to extract the necessary features from the kinematic data. This justifies the

development of a deeper, more powerful "ResLSTM" model; the one-dimensional ResNet-18 backbone and bi-directional LSTM layers exhibited far better performance on our dataset than shallower networks. This outperforms a LSTM-only network for two likely reasons: 1) the ResNet extracts meaningful features from the raw sensor data, and 2) the convolution operators reduce the length of the time-series sequences, which are easier for the LSTM layers to learn than longer sequences.

Leveraging transfer learning was also important to increasing the image model's performance. Training a ResNet-50 model without weights pre-trained on ImageNet leads to an RMSE of 0.523 (0.274) for the quality of final product score, compared to the RMSE score of 0.392 (0.146) exhibited with pre-training. Although the ImageNet dataset does not contain examples of surgical sutures, the low-level features learned on the large-scale generic dataset are helpful starting points when transitioning to a domain-specific task.

Combining both the kinematic and image modalities allows for a single model to rate all four surgical skill assessment categories. Further, training a single model on both modalities led to an increase in performance across all the categories, except for Respect for Tissue. It is unclear why this modal sees a degradation in performance in this category compared to the kinematic-only model; further experiments are required to discern this. Notably, the Overall Performance category saw a large increase in MSE and R^2 scores. Training on both kinematic and image data allows for the combined model to learn a more optimal feature set that is better representative of the task performances.

This study is limited in that the AI was trained and evaluated on data collected from a single training center. It remains to be studied how the model performance is affected by increased participant diversity, e.g., trainees from different institutes or countries. Future studies can investigate how the the model generalizes to new participants. Further, while the OSATS was used in this study to evaluate the knot tying performance, improved assessment tools, such as a modified OSATS score which incorporates additional domains [24], may be more suitable in future studies as more complex tasks are considered in more physiologically challenging environments.

V. Conclusion

We presented and analyzed a novel dataset of a surgical knot-tying task. We also designed and developed three deep learning models that were trained and evaluated on this dataset, with the goal of automating the surgical skills assessment. This includes a multi-modal model that takes image and time-series kinematic data as input which demonstrated the best performance. Its results were comparable to the expert human raters on most of the GRS domains. The AI model generally had a lower point error, and the humans show slightly better agreement.

Thus, the main contributions of this work are 1) introducing a new surgical skills dataset, and 2) developing a multi-modal deep learning model that replicates human raters on most of the GRS domains. These represent important steps towards automated surgical skill assessment, with the potential for reduced surgical faculty burden while allowing instant & objective feedback for surgical trainees.

Acknowledgment

The authors would like to acknowledge Lisa Satterthwaite, Dr. Oleg Safir, and the staff at the Mount Sinai Hospital Surgical Skills Centre for their support of this work, as well as the Wyss Medical Foundation for their funding.

References

[1] H. M. Richard K Reznick, "Teaching surgical skills–changes in the wind." *The New England Journal of Medicine*, vol. 355, no. 25, pp. 2664–2669, 2006.

[2] R. R. Sonnadara, C. Mui, S. McQueen, P. Mironova, M. Nousiainen, O. Safir, W. Kraemer, P. Ferguson, B. Alman, and R. Reznick., "Reflections on competency-based education and training for surgical residents." *Journal of Surgical Education*, vol. 71, no. 1, pp. 151–158, 2014.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 06 2016, pp. 770–778.

[4] M. J. Fard, S. Ameri, R. D. Ellis, R. B. Chinnam, A. K. Pandya, and M. D. Klein, "Automated robot-assisted surgical skill evaluation: Predictive analytics approach," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 14, 2018.

[5] B. Poursartip, M.-E. LeBel, L. C. McCracken, A. Escoto, R. V. Patel, M. D. Naish, and A. L. Trejos, "Energy-based metrics for arthroscopic skills assessment," *Sensors*, vol. 17, 2017.

[6] H. Law, "Skill assessment using computer vision based analysis," 2017.

[7] R. A. Watson, "Use of a machine learning algorithm to classify expertise: analysis of hand motion patterns during a simulated surgical task," 2014.

[8] Z. Aneeq, S. Yachna., and B. Vinay, "Video and accelerometer-based motion analysis for automated surgical skills assessment," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, 2018.

[9] O. O'Driscoll, R. Hisey, D. Camire, J. Erb, D. Howes, G. Fichtinger, and T. Ungi, "Object detection to compute performance metrics for skill assessment in central venous catheterization," in *Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling*, C. A. Linte and J. H. Siewerdsen, Eds., vol. 11598, International Society for Optics and Photonics. SPIE, 2021, pp. 315 – 322. [Online]. Available: https://doi.org/10.1117/12.2581889

[10] O. O'Driscoll, R. Hisey, M. Holden, D. Camire, J. Erb, D. Howes, T. Ungi, and G. Fichtinger, "Feasibility of object detection for skill assessment in central venous catheterization," in *Medical Imaging 2022: Image-Guided Procedures, Robotic Interventions, and Modeling*, C. A. Linte and J. H. Siewerdsen, Eds., vol. 12034, International Society for Optics and Photonics. SPIE, 2022, pp. 358 – 365. [Online]. Available: https://doi.org/10.1117/12.2607294

[11] Y. Gao, S. S. Vedula, C. E. Reiley, N. Ahmidi, B. Varadarajan, H. C. Lin, L. Tao, L. Zappella, B. Béjar, D. D. Yuh, C. C. G. Chen, R. Vidal, S. Khudanpur, and G. Hager, "Jhu-isi gesture and skill assessment working set (jigsaws) : A surgical activity dataset for human motion modeling," 2014.

[12] I. Surgical, "Davinci by intuitive." [Online]. Available: https://www.intuitive.com/en-us/products-and-services/da-vinci

[13] S. Khalid, M. G. Goldenberg, T. P. Grantcharov, B. Taati, and F. Rudzicz, "Evaluation of deep learning models for identifying surgical actions and measuring performance." *JAMA network open*, vol. 3 3, p. e201664, 2020.

[14] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Evaluating surgical skills from kinematic data using convolutional neural networks," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham: Springer International Publishing, 2018, pp. 214–221.

[15] E. Yanik, X. Intes, U. Kruger, P. Yan, D. Diller, B. Voorst, B. Makled, J. Norfleet, and S. De, "Deep neural networks for the assessment of surgical skills: A systematic review," *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, p. 154851292110345, 08 2021.

[16] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "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, 07 2019.

[17] K. Kasa, D. Burns, M. G. Goldenber, O. Selim, C. Whyne, O. Safir, and M. Hardisty, "Machine learning the assessment of surgeon technical skill for one handed surgical knot tying," in *Imaging Network of Ontario 2022 Symposium*, 2022.

[18] J. A. Martin, G. Regehr, R. Reznick, H. Macrae, J. Murnaghan, C. Hutchison, and M. Brown, "Objective structured assessment of technical skill (osats) for surgical residents," *British Journal of Surgery*, vol. 84, pp. 273–278, 1997.

[19] D. M. Burns and C. M. Whyne, "Seglearn: A python package for learning sequences and time series," *Journal of Machine Learning Research*, vol. 19, no. 83, pp. 1–7, 2018. [Online]. Available: http://jmlr.org/papers/v19/18-160.html

[20] D. Itzkovich, Y. Sharon, A. Jarc, Y. Refaely, and I. Nisky, "Using augmentation to improve the robustness to rotation of deep learning segmentation in robotic-assisted surgical data," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 5068–5075.

[21] D. M. Burns, N. Leung, M. Hardisty, C. M. Whyne, P. Henry, and S. McLachlin, "Shoulder physiotherapy exercise recognition: Machine learning the inertial signals from a smartwatch," *Physiological measurement*, vol. 39 7, p. 075007, 2018.

[22] F. J. Ordonez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors (Basel, Switzerland)*, vol. 16, 2016.

[23] T. Koo and M. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of Chiropractic Medicine*, vol. 15, 03 2016.

[24] C. J. Hopmans, P. T. den Hoed, L. van der Laan, E. van der Harst, M. van der Elst, G. H. H. Mannaerts, I. Dawson, R. Timman, B. P. Wijnhoven, and J. N. M. Ijzermans, "Assessment of surgery residents' operative skills in the operating theater using a modified objective structured assessment of technical skills (osats): a prospective multicenter study." *Surgery*, vol. 156 5, pp. 1078–88, 2014.