ANNALS OF SURGERY OPEN

OPEN

Developing Surgical Skill Level Classification Model Using Visual Metrics and a Gradient Boosting Algorithm

Somayeh B. Shafiei*, Saeed Shadpour†, James L. Mohler*, Kristopher Attwood‡, Qian Liu‡, Camille Gutierrez§, and Mehdi Seilanian Toussi*

Objective: Assessment of surgical skills is crucial for improving training standards and ensuring the quality of primary care. This study aimed to develop a gradient-boosting classification model to classify surgical expertise into inexperienced, competent, and experienced levels in robot-assisted surgery (RAS) using visual metrics.

Methods: Eye gaze data were recorded from 11 participants performing 4 subtasks; blunt dissection, retraction, cold dissection, and hot dissection using live pigs and the da Vinci robot. Eye gaze data were used to extract the visual metrics. One expert RAS surgeon evaluated each participant's performance and expertise level using the modified Global Evaluative Assessment of Robotic Skills (GEARS) assessment tool. The extracted visual metrics were used to classify surgical skill levels and to evaluate individual GEARS metrics. Analysis of Variance (ANOVA) was used to test the differences for each feature across skill levels.

Results: Classification accuracies for blunt dissection, retraction, cold dissection, and burn dissection were 95%, 96%, 96%, and 96%, respectively. The time to complete only the retraction was significantly different among the 3 skill levels (P value = 0.04). Performance was significantly different for 3 categories of surgical skill level for all subtasks (P values < 0.01). The extracted visual metrics were strongly associated with GEARS metrics ($R^2 > 0.7$ for GEARS metrics evaluation models).

Conclusions: Machine learning algorithms trained by visual metrics of RAS surgeons can classify surgical skill levels and evaluate GEARS measures. The time to complete a surgical subtask may not be considered a stand-alone factor for skill level assessment.

Keywords: expertise level, robot-assisted surgery, visual metrics

INTRODUCTION

RAS Surgical Skill Levels Evaluation Challenge in Surgical Training

A large proportion of the cases of surgical malpractice are due to a lack of technical competence.[1] Robotic-assisted surgery (RAS), specifically the da Vinci Surgical System, has become popular in a variety of specialties, especially surgical oncology, urology, and gynecology, due to its benefits such as smaller incisions, less pain, lower risk of infection, and shorter hospital stay.[2] However, surgeons must acquire the required proficiency level to ensure both the safety of patients and high-quality surgical outcomes. RAS procedures require high dexterity, motor planning and control, and hand-eye coordination. Acquisition of these skills requires effective surgical training and assessment methods.[3] Evaluation of the level of surgical expertise is important for feedback during training and programmatic changes.[4] Procedural-based assessment and surgical logbooks are examples of traditional (ie, nonautomated) surgical skill assessment methods. In all these procedures, an expert surgeon monitors the trainee's activity and evaluates surgical skills. These methods of assessing surgical skills are inconsistent and change by the rater.[5] The challenge of universal surgical skill assessment has yet to be addressed.

Available Objective Skill Evaluation Methods in RAS

An "objective" surgical skill assessment means to assess expertise level that eliminates inconsistencies. Objective assessment methods have not been developed for existing surgical practice protocols and volume-based skills assessment. Several studies in the literature have proposed objective and automated surgical skill assessment methods using physiological data including brain activity,[6] eye movement,[4] kinematics,[7] or surgical videos.[8]

From the *Department of Urology, Roswell Park Comprehensive Cancer Center in Buffalo, NY; †Department of Animal Biosciences, University of Guelph, Guelph, Ontario, Canada; ‡Department of Biostatistics and Bioinformatics, Roswell Park Comprehensive Cancer Center, Buffalo, NY; and §Obstetrics and Gynecology Residency Program, Sisters of Charity Health System, Buffalo, NY.

SDC Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.annalsofsurgery.com).

Research reported in this publication was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under award number R01EB029398 and the Alliance Foundation of Roswell Park Comprehensive Cancer Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work was supported by National Cancer Institute (NCI) grant P30CA016056 involving the use of Roswell Park Comprehensive Cancer Center's Comparative Oncology Shared Resource and Applied Technology Laboratory for Advanced Surgery (ATLAS).

The data that support the findings of this study are available from the corresponding author (SBS) upon reasonable request.

Disclosure: The authors declare that they have nothing to disclose.

Reprints: Somayeh B. Shafiei, PhD, Department of Urology, Roswell Park Comprehensive Cancer Center, Elm & Carlton St Buffalo, NY 14263. Email: Somayeh.besharatshafiei@roswellpark.org.

Annals of Surgery Open (2023) 2:e292

Received: 25 October 2022; Accepted 24 April 2023

Published online 24 May 2023

DOI: 10.1097/AS9.0000000000000292

The proposed methods for objective RAS skill assessment have reported promising results and opened the door for a new era of automated surgical skill evaluation. However, those methods have shortcomings limited to very basic tasks performed using models in the dry lab, using small groups of participants, introducing biases, or developing computationally costly models that cannot be integrated into surgical robot systems.

Importance of Objectifying RAS Surgical Skill Evaluation in Urology

Hysterectomy, cystectomy, and nephrectomy are 3 common urology and gynecological surgery procedures. Between 2000 and 2013, 3194 adverse events (death, injury, malfunction, etc.) related to robotic surgery were reported to the US Food and Drug Administration Manufacturer and User Facility Device Experience (MAUDE) database in gynecology, 2331 of which were related to hysterectomy. In addition, 138 of the 1565 adverse events reported in urology were related to nephrectomy and 48 to cystectomy.[9,10] Surgical trainees would best acquire experience in performing these operations before operating on humans in an actual operating room (OR). However, there is currently no comprehensive objective surgical skill assessment paradigm applicable in clinical settings.

Importance of Time to Complete a Surgical Task in Surgical Skill Acquisition

In several surgical fields, residents or fellows attend the OR, and an expert surgeon teaches them, as a part of the training program, by considering patient safety and surgical outcome as the 2 main priorities. This teaching style lengthens operative time,[11] while operative time is a major factor affecting the risk of surgical complications.[12] Various surgical specialties, including urology,[13] obstetrics, and gynecology[14] have shown a strong association between prolonged operative time and increased risk of complications. A meta-analysis of 33 studies showed that the risk of complications approximately doubled with operative time greater than 2 hours.[15]

Several studies have found a link between longer operative times and surgical site infections.[16] Logical reasons have been suggested to explain the association between surgical site infection and operative time, which include longer exposure of tissues to environmental bacteria, increased risk of tissue desiccation and ischemia, diminished concentration of prophylactic antibiotics, increased probability of sterile rule violation, and enhanced risk of venous thromboembolism through longer blood stasis and an increased risk of endothelial damage.[17] Lengthy operative times may cause fatigue in the surgical team, which consequently adversely affects their decision-making and concentration.[18]

Operative time is an important metric in surgical performance and skill level evaluation tools, but it is not a sufficient quantitative metric for performance and skill level assessment.[19] Operative time has been considered in RAS surgical skill evaluation tools; Global Assessment of Robotic Skills (GEARS) considers time as part of the ‘efficiency’ metric.[20]

Connection Between Eye Metrics and Surgical Skill Assessment

Research in a variety of disciplines, including the evaluation of surgical skill and surgical training has suggested using eye gaze metrics as an assessment tool.[21–23] Eyeglasses record the corneal reflection of infrared lighting to track pupil location, mapping the subject’s focus of attention.[24] These recordings have enabled the measurement of various eye metrics including gaze entropy and gaze velocity, time to first fixation, total fixation duration, and saccade rate.[21,25,26] Differences in these metrics between subjects of varying skill levels suggest their use as markers of skill level.[23,27–29] Recording the eye gaze of experts and junior resident surgeons performing laparoscopic cholecystectomy, experts who watched the video had much more overlap (55% of the time) with the reference (expert) surgeons than junior residents (43.8%).[28]

Eye gaze measurements have provided insight into learning and skill level improvement in several applications. However, the utility of measurements retrieved from eye movement behavior during RAS skill evaluation have yet to be documented.

Potential Use of Machine Learning Approaches for Surgical Skill Assessment

The application of machine learning (ML) to objectively assess surgical skills and offer timely, helpful surgical feedback is growing quickly. Several studies suggested using ML methods to objectify RAS surgical skill assessment and performance evaluation using various types of physiological data collected from surgeons and surgical trainees.[30,31] Development of ML models that use physiological data (eg, eye gaze data) for RAS surgical skill assessment and performance evaluation in the operating room needs to be investigated. Gradient Boosting and Generalized Linear Mixed Model using penalized Lasso method (known as GLMMLASSO) are appropriate ML techniques for this purpose.

Gradient Boosting

Ensemble learning is a ML method that uses multiple predictors, instead of using a single predictor, trains them on the data, and combines their results, usually giving a better score than using a single model. Boosting is a special type of ensemble learning method that combines several weak learning algorithms (ie, decision trees; predictors with poor accuracy) into a strong learner (a model with strong accuracy).[32] It first fits an initial model (such as a tree or a linear regression) to the data. Next, a second model is created by accurately predicting the data that the first model could not. The combined performance of these 2 models should outperform their individual results. Then this process is repeated several times. Each successive model (ie, tree) corrects the shortcomings of the combined boosted ensemble of all previous models by learning from its predecessor’s mistakes.[32] Boosted trees are the trees that have been modified by boosting process.[32] Gradient Boosting is an ensemble learning and boosting technique that improves predictions by having each predictor try to reduce its predecessors’ errors. Instead of fitting a prediction to the data at each iteration, gradient boosting fits a new predictor to the residual errors produced by the prior predictor. Gradient Boosting methods have shown significant success in several applications.[33]

Generalized Linear Mixed Models by L1- penalized estimation

The Generalized Linear Mixed Model (GLMM) is an ML method that extends the generalized linear model by incorporating random effects in linear predictors to account for clustering. The penalized Lasso method selects variables and estimates coefficients simultaneously in GLMM (ie, GLMMLASSO).[34]

The Focus of the Study on Addressing the Surgical Skill Classification Challenge

In this study, (1) the retrieved information from visual metrics was used in an ML model to develop a RAS surgical skill classification model for application to a clinical setting; and (2) the time to complete each subtask, performance scores, as well as the visual metrics, were compared between surgical skill levels to find the change of these metrics across different levels. And (3) the retrieved information from visual metrics was used to develop ML models to evaluate individual GEARS metrics.

METHODS

This study was conducted in accordance with relevant guidelines and regulations approved by the Institutional Review Board (IRB) of Roswell Park Comprehensive Cancer Center (IRB: I-241913) and Institutional animal care and use committee approval (IACUC 1179S). Participants completed questionnaires that captured their age, gender, RAS surgical cases, and the number of laparoscopic surgical cases as the primary surgeon. The IRB issued a waiver of documentation of consent. Participants were given a research study information sheet and provided verbal consent.

Data Recording

The eye gaze data was recorded from participants using TobiiPro2 eyeglasses with a frequency of 50 Hz. Videos were recorded in the OR during task performances.

Participants

Eleven participants with varying RAS experience levels completed hysterectomy, cystectomy, and/or nephrectomy (Table 1).

Operations

Participants performed operations, using live pigs and the da Vinci robot, during 1 session that lasted for 4–6 hours. An expert surgeon attended the session as the mentor if a participant did not have operative experience.

Subtask Extraction

Four subtasks were extracted from operations (Table 2) that included blunt dissection (subtask 1), retraction (subtask 2), cold dissection and cold cutting using scissors (subtask 3), and burn dissection (subtask 4). The start and end times for each subtask were derived from recorded videos. Eye gaze information for each subtask was extracted.

Eye Gaze Data Pre-processing

Tobii Pro Lab was used to preprocess eye gaze data before extracting visual metrics. The Tobii Pro Lab framework applies a moving average filter with a window size of 3 points to reduce noise in eye gaze data. A velocity-threshold identification fixation filter, with a threshold of 30 degrees per second, was applied to the data to identify fixation and saccadic points. Extracted visual metrics for both eyes were defined in Table 3.[35]

True Surgical Expertise Levels

The surgical performance and level of expertise of each participant were evaluated using the modified GEARS assessment method by an expert RAS surgeon who watched the recorded videos. GEARS is an assessment tool for the evaluation of overall technical proficiency for robotic surgery. GEARS metrics are depth perception, bimanual dexterity, efficiency, force sensitivity, robotic control, and autonomy. Each metric is scored on a Likert scale from 1 to 5, and the overall performance score ranges from 6 to 30. GEARS also has 3 levels for assessing expertise: (1) inexperienced: requires extensive practice/improvement, (2) competent: adequate, and (3) experienced: excellent/established. The rater (expert RAS surgeon) provided an explanation of each score and indicated whether certain subtasks were performed with a higher or lower level of competence, and performance than others. These skill level variations were considered during analysis (assigning a true surgical expertise level to each subtask).

Machine Learning Classification Model

The extracted visual metrics for each subtask and the true surgical expertise levels were used as input to a gradient-boosting classification model (GBM) to classify the 3 classes of expertise. Twenty percent of the samples were chosen at random and used as a test set. The remaining 80% were used to train and validate the model. This process was repeated 10 times, and the average measurements were reported. The number of trees to construct (ie, the number of boosting stages to perform), the maximum depth of trees (the maximum depth limits the number of nodes in the tree), and the learning rate (this parameter scales the contribution of each tree), were randomly tuned using the grid search method during GBM model development. The process for tuning each parameter was explained in Table 4.[36]. The developed model's performance in classifying the surgical skill levels of participants was evaluated using statistical measurements that were explained in Table 4.

Statistical Analysis to Find the Change in Each Eye Gaze Metric, Time, and Performance Across Surgical Skill Levels

A linear mixed model was fitted for each eye gaze feature, the time to complete, and performance where the skill levels were treated as 3 factors (levels 1: inexperienced; 2: competent; and 3: experienced), and participant Identifier (ID) was treated as a random effect to accommodate for repeated measurement.

TABLE 1.

Demographics of participants and the number of operations performed by each participant

Participant	1	2	3	4	5	6	7	8	9	10	11
Gender	M	M	M	M	M	M	M	M	M	M	F
Age (years)	28	67	36	44	47	44	61	32	33	39	32
Specialty*	U	U	U	T	U	U	T	U	G	U	G
Position	Resident	Faculty	Fellow	Faculty	Faculty	Faculty	Faculty	Fellow	Fellow	Fellow	Resident
RAS practice (hours)	<100	>1000	>100	>500	>1000	>1000	<100	500	>100	<100	<100
RAS clinical experience (cases)	0	>500	<150	>150	>500	>500	0	<150	<150	0	0
Laparoscopic surgeries as the primary surgeon (cases)	0	0	<50	>250	Between 50 and 100	>250	>250	0	Between 50 and 100	0	0
Hysterectomy	1	1	1	1	1	1	0	1	1	2	1
Cystectomy	1	1	1	1	1	1	0	1	1	2	1
Left Nephrectomy	1	1	1	1	1	1	1	1	1	2	1
Right Nephrectomy	1	1	1	0	1	1	1	0	1	2	0

*Specialty: Gynecology (G); Urology (U); Thoracic (T).
RAS indicates robot-assisted surgery.

TABLE 2.

Total number of samples for each subtask

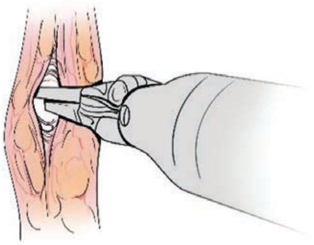
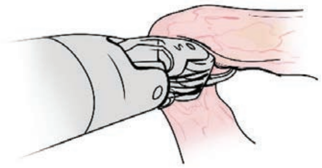
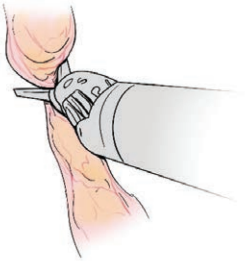
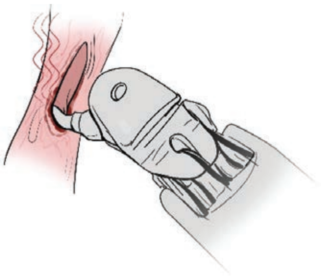
Subtasks	Total Samples	Inexperienced	Competent	Experienced
Subtask 1. Blunt dissection: separating tissue planes by pushing rather than cutting or cautery 	219	50	134	35
Subtask 2. Retraction: hold structures aside to improve the visibility of the operative field 	1082	182	611	289
Subtask 3. Cold dissection and cold cutting: using scissors: use scissors to cut tissue 	376	40	272	64
Subtask 4. Burn dissection: use a hook tool to cut tissue with heat 	374	68	179	127

TABLE 3.

Definition of visual metrics

Visual Metric	Definition
Rate of fixation	Number of eye-tracking time points that fell below the threshold of 30 degrees per second divided by the number of total time points of the recording.
Rate of saccade	Number of eye-tracking time points with an angular velocity higher than the threshold of 30 degrees per second, divided by the number of total time points of the recording.
Average pupil diameter	Average pupil diameter of each eye throughout a recording. This feature was calculated for both eyes.
Shannon entropy of pupil diameter	Shannon entropy: average rate at which information is produced by a stochastic source of data.[35] For a signal X(t), Shannon entropy is calculated as $S(X) = -\sum_{i=1}^{N} p(x_i)\, log_2\,(p(x_i))$ where p (x_i) is the probability of obtaining the value x_i. This feature was calculated for both eyes.
Rate of change of eye gaze direction	Total number of time points at which the direction of the eye changes, divided by the total number of time points. This feature was calculated for both eyes and both directions (horizontal and vertical).
Total length of eye-tracking trajectory	The length of the eye pupil trajectory was extracted using the geometry of the eye pupil trajectory. This feature was calculated for both eyes.

TABLE 4.

Tuning process for optimizing hyperparameters of the gradient boosting classification model, and the model's evaluation metrics

Gradient Boosting Model's Hyperparameters Tuning	
Parameter	Tuning process
number of trees	was determined in increments of 25 in the range of 25–500.
maximum depth of trees	was determined in increments of 2 in the range of 1–21.
learning rate	was determined in increments of 0.1 in the range of 0.1–2.
The hyperparameters of the model were optimized using a 5-fold cross-validation that was repeated 5 times, and samples were distributed in folds in a stratified fashion. Each fold consisted of samples from inexperienced, competent, and experienced participants, and the distribution of samples from each expertise level was the same as the distribution of the original population. The Adaptive Synthetic algorithm was applied to the training sets to address the imbalance of samples from 3 categories in the dataset during training.[36]	
Evaluation Metrics for Classification Model	
Metric	Definition
Average accuracy	the ratio of the sum of correct predictions to the total number of predictions
Precision	the ratio of correct positive predictions (T_p) and the total positive results predicted by the classifier (T_p+F_p)
Recall	the ratio of positive predictions (T_p) and the total positive results predicted by the classifier (T_p+F_p)

T_p and F_p were the numbers of true positives and false positives, while T_n and F_n were the numbers of true negatives and false negatives.

ANOVA was fitted to test whether there was any difference in measurements between different skill levels. A *P* value less than 0.05 was considered a statistically significant difference between skill levels.

Machine Learning Models for Evaluation of GEARS Metrics

The extracted visual metrics for each subtask were used to develop GLMM by L1- penalized estimation (ie, Least Absolute Shrinkage and Selection Operator), known as GLMMLASSO, to evaluate individual GEARS metrics. Five-fold cross-validation was used to select the optimum lambda value based on the Bayesian information criterion. Each GEARS metric was considered as the dependent variable, visual metrics were considered as possible independent variables, and participant ID was treated as a random effect. The R^2 metric was calculated to measure the proportion of variance in the dependent variable that can be explained by the independent variables.

RESULTS

Classification results and confusion matrices were represented in Tables 5 and 6, respectively.

Change of Eye Gaze Features, Performance, and Time to Complete Surgical Subtasks Across Skill Levels

Eye gaze metrics that were significantly different for 3 categories of surgical skill levels, and differences between those identified eye gaze metrics across skill levels, were identified (Table 7).

The rate of the left eye's gaze direction changes in horizontal and vertical directions increased by skill level when performing blunt dissection, retraction, and burn dissection. Also, the average pupil diameter of left and right eyes decreased as skill level changed from competent to experienced during the retraction and burn dissection subtask.

Time to complete only subtask 2 (retraction by nondominant hand) was significantly different for the 3 categories of skill level (*P* value = 0.04; Table 8). The time required to complete the

TABLE 5.

Evaluation metrics representing the performance of the GBM model classifying surgical expertise of participants performing blunt dissection, retraction, cold dissection, and burn dissection subtasks into 3 levels: inexperienced, competent, or experienced

Average % (Standard Deviation%)	Subtask 1: Blunt Dissection	Subtask 2: Retraction	Subtask 3: Cold Dissection	Subtask 4: Burn Dissection
Precision	92 (4)	96 (1)	94 (4)	96 (1)
Recall	95 (4)	96 (2)	96 (3)	96 (2)
F1-score	93 (4)	96 (1)	94 (3)	96 (2)
Accuracy	95 (3)	96 (1)	96 (2)	96 (2)

TABLE 6.

Confusion matrices for classification of surgical expertise levels for blunt dissection (A), retraction (B), cold dissection (C), and burn dissection (D) subtasks into 3 levels of inexperienced, competent, and experienced

True skill	Inexperienced	93	5	3
	competent	2	98	0
	experienced	1	13	86
		Inexperienced	competent	experienced
			Predicted skill	

A)

True skill	Inexperienced	94	6	0
	competent	1	98	1
	experienced	0	4	96
		Inexperienced	Competent	experienced
			Predicted skill	

B)

True skill	Inexperienced	88	12	0
	competent	1	98	1
	experienced	0	5	95
		Inexperienced	competent	experienced
			Predicted skill	

C)

True skill	Inexperienced	98	2	0
	competent	1	95	4
	experienced	1	3	96
		Inexperienced	competent	Experienced
			Predicted skill	

D)

TABLE 7.

Eye gaze metrics that were significantly different (*P* value <0.05) for 3 categories of surgical skill levels

	Subtask Name	Feature Name	*P* value
A	Blunt dissection	Rate of gaze direction change, left eye, the horizontal direction	**0.004**
B	Blunt dissection	Rate of gaze direction change, left eye, the vertical direction	**0.004**
C	Retraction	Average pupil diameter, left eye	**2×10^{-9}**
D	Retraction	Average pupil diameter, right eye	**9×10^{-6}**
E	Retraction	Rate of gaze direction change, left eye, the horizontal direction	**0.002**
F	Retraction	Rate of gaze direction change, left eye, the vertical direction	**0.002**
G	Burn dissection	Average pupil diameter, left eye	**0.037**
H	Burn dissection	Average Pupil diameter, right eye	**0.013**
I	Burn dissection	Rate of gaze direction change, left eye, the horizontal direction	**3×10^{-5}**
J	Burn dissection	Rate of gaze direction change, left eye, the vertical direction	**0.007**

The difference in eye gaze metrics across skill levels. Significant differences (*P* value <0.05).

		A	B	C	D	E	F	G	H	I	J
Inexperienced to competent.	coefficient	0.007	0.002	0.11	0.20	0.004	0.003	–0.18	0.03	0.001	0.003
	P value	**0.04**	0.28	0.87	0.74	0.47	0.44	0.81	0.96	0.85	0.37
Inexperienced to experienced.	coefficient	0.018	0.01	–0.46	–0.22	0.015	0.012	–0.38	–0.17	0.017	0.01
	P value	**1×10^{-4}**	**2×10^{-4}**	0.51	0.73	**9×10^{-3}**	**5×10^{-3}**	0.64	0.81	**0.02**	**5×10^{-3}**
Competent to experienced.	coefficient	0.011	0.008	–0.57	–0.42	0.011	0.008	–0.19	–0.21	0.016	0.007
	P value	**5×10^{-3}**	**5×10^{-4}**	**2×10^{-11}**	**2×10^{-7}**	**4×10^{-4}**	**6×10^{-4}**	**4×10^{-3}**	**7×10^{-4}**	**7×10^{-7}**	**1×10^{-3}**

TABLE 8.

Significance level for the difference in time to complete surgical subtasks among 3 categories of surgical skill levels, and across pairs of skill levels

The Significance Level for 3 Categories of Skill Level		Blunt Dissection	Retraction	Cold Cut	Burn Dissection
P value		0.06	**0.04**	0.30	0.48
The significance level for change of time to complete each subtask across skill levels					
Inexperienced to component	Coefficient	1.34	–1.19	4.77	0.72
	P value	0.61	0.39	0.13	0.75
Inexperienced to experienced	Coefficient	–6.21	**–3.94**	3.36	–2.00
	P value	0.08	**0.02**	0.28	0.42
Competent to experienced	Coefficient	**–7.56**	–2.74	–1.41	–2.72
	P value	**0.02**	0.07	0.66	0.24

retraction subtask with the nondominant hand differed significantly across 3 skill levels, and it decreased significantly from inexperienced to experienced samples.

Performance scores for inexperienced, competent, and experienced samples of each subtask were represented in Table 9. Performance was significantly different for 3 categories of surgical skill level and increased significantly from inexperienced to experienced for all subtasks, from inexperienced to competent for blunt dissection and burn dissection. It increased significantly from competent to experienced for blunt dissection, retraction, and burn dissection.

GEARS Metrics Evaluation Models

The GLMMLASSO models for evaluation of GEARS metrics were represented in Supplemental Data File, http://links.lww.com/AOSO/A224 for different surgical subtasks. Gears metrics were significantly associated with visual metrics for all subtasks and the R^2 of all developed models for GEARS metrics were strong (>0.7).

DISCUSSION

Better methods for surgical skill assessment are necessary to improve training while ensuring patient safety. Methods for

TABLE 9.

Distribution of performance across skill levels, and the significance level for the difference in performance in completing surgical subtasks among three categories of surgical skill levels, and across pairs of skill levels

		Performance (Standard Deviation)			
Skill Level		Subtask: Blunt Dissection	Subtask: Retraction	Subtask: Cold Dissection	Subtask: Burn Dissection
Inexperienced		13.2 (2.96)	12.7 (2.82)	13.6 (3.03)	11.1 (3.22)
Competent		21.7 (1.42)	19.7 (1.35)	19.6 (1.75)	20.6 (1.52)
Experienced		27.3 (1.52)	27.7 (1.38)	28.5 (2.59)	26.7 (1.53)
The significance level for the difference in performance among three categories of surgical skill levels					
		Blunt dissection	Retraction	Cold cut	Burn dissection
P value		**1.32×10^{-6}**	**8×10^{-10}**	**0.01**	**2.33×10^{-11}**
The significance level for change of performance in completing each subtask across skill levels					
Inexperienced to component	Coefficient	8.49	7.04	5.99	9.46
	P value	**0.03**	0.05	0.25	**0.02**
Inexperienced to experienced	Coefficient	14.03	15.08	14.89	15.56
	P value	**2×10^{-4}**	**$<1 \times 10^{-5}$**	**0.01**	**$<1 \times 10^{-4}$**
Competent to experienced	Coefficient	5.54	8.04	8.90	6.03
	P value	**$<1 \times 10^{-4}$**	**$<1 \times 10^{-5}$**	0.05	**$<1 \times 10^{-4}$**

evaluating RAS technical proficiency can be divided into manual and automated. Several proven manual evaluation techniques are simple to use but require an expert panel that is prone to bias. The best existing skill assessment approach is direct observation of operative performance using global rating scales,[37] which is subject to bias and requires experts to be present throughout the session and spend a significant amount of time. Objective skill assessment methods enable individualized skill development, which ultimately improves surgical outcomes.

Recent developments in RAS have increased the demand for effective methods for objective skill evaluation.[38] Establishing surgical skills and competency evaluation method is crucial to decreasing the high rate of medical errors. Several methods have been proposed to address this need. However, objective skill evaluation for RAS in clinical settings remains challenging. Some have suggested that experience alone can serve as a stand-in for skill level, by correlating surgeon operating volume to outcome.[39] However, this approach has its shortcomings; a surgeon who lacks expertise in 1 task may be an expert in another or a surgeon who performs many operations can continuously perform certain activities with poor results.

The primary goal of this research was to introduce a GBM classification model developed using visual metrics as an objective RAS skill levels classification model. Visual metrics can provide information about expertise level in RAS. Visual metrics were used to demonstrate that experts have focused attention that differentiates them from novices.[40] The use of modalities to record eye gaze data is straightforward and inexpensive. These advantages make using eye gaze data for surgical skill assessment in the clinic practical.

Eye Gaze Metrics for Skill Level Classification

The results of this study showed that the use of eye gaze metrics and a GBM classification model is a promising approach to the efficient and objective assessment of surgical skill and can distinguish inexperienced, competent, and experienced participants from each other, with high accuracy while performing RAS subtasks using live pigs.

Patient safety is the most important priority in an OR and expert surgeons should provide enhanced safety. Therefore, a surgical skill assessment model should minimize the misclassification of inexperienced surgeons as experienced. The developed skill classification model performed well in differentiating experienced and inexperienced samples. The developed RAS skill classification model misclassified only 3% of inexperienced samples as experienced when performing blunt dissection, while it did not misclassify any inexperienced as experienced when performing retraction, cold dissection, or burn dissection.

Classification accuracy for blunt dissection was lower than that for retraction, cold dissection, or burn dissection subtasks, which could have resulted from fewer samples for blunt dissection than for other subtasks (219 compared to 1082, 376, and 374, respectively). However, there were fewer samples for burn dissection (374) than for retraction (1082) and the classification accuracies were similar. Perhaps, evaluating expertise in performing blunt dissection is more difficult than for other subtasks. Finally, blunt dissection is a complicated subtask so information beyond visual metrics may be required to distinguish among expertise levels.

A major portion of misclassified samples for the blunt dissection subtask was related to experts misclassified as competent (13%). This finding may show that even when the outcome of a surgical subtask is very good, the eyes of the surgeon could still behave like a competent surgeon. Hence, visual metrics may not be sufficient for differentiating experienced from competent RAS surgeons.

Existing studies about surgical skill assessment have reported poor classification of the competent category since competent surgeons have developed some skills but not all the skills required to become experienced.[41] The results of this study showed the developed classification model outperformed the existing models since 98%, 98%, 98%, and 95% of the blunt dissection, retraction, cold cut, and burn dissection subtasks performed by competent surgeons were correctly classified.

Numerous studies have been performed to evaluate a surgeon's skill using pupil size, fixation time, and saccade time.[42,43] To the best of our knowledge, this is the first study to suggest 12 eye gaze features to differentiate inexperienced, competent, and experienced RAS surgeons while they perform surgical procedures on pigs.

The findings of this study demonstrated that visual metrics have the potential for RAS surgical skill classification to be used in clinical settings.

Relationship Between Eye Gaze Metrics and Skill Level

The increase in the rate of the gaze direction changes of the left eye in horizontal and vertical directions by skill level during blunt dissection, retraction, and burn dissection can be interpreted based on improvement in the level of engagement and shift in attention by surgical skill level. This finding may point to the fact that more skilled surgeons switch their attention more frequently to concentrate on the target at hand, gather information from the surgical scene, and use it to inform their decisions rather than maintaining a static focus on a single target for an extended period.

Furthermore, pupil diameter has been proposed as a component for measuring cognitive load,[44] and a greater change in pupil dilation was associated with a higher working memory load.[45] Hence, the smaller average pupil diameter of the left and right eyes in more experienced surgeons during retraction and burn dissection may indicate a lower cognitive load. This result is consistent with what research on learning has found[46]; an increase in pupil diameter is related to response latency and target selection uncertainty.[47] The results of this study also support the idea that more experienced surgeons are less uncertain about choosing the correct target because they have more knowledge and experience.

Relationship Between Operative Time and Skill Level

Although operative time influences patient safety and should be improved during skill acquisition, the amount of time it takes to perform a surgical subtask is not a single criterion for determining surgical skill level. Results showed that time to complete only the retraction subtask performed by the nondominant hand differed among the surgical skill level groups. The time to completion for surgeons of 3 different skill levels performing blunt dissection, cold dissection, and burn dissection subtasks, all performed by the dominant hand was not significantly different.

GEARS Metrics Evaluation Models Using Eye Gaze Metrics

Results showed that eye gaze metrics are informative in evaluating GEARS metrics in performing subtasks. Promising results may suggest using the extracted visual metrics for evaluating GEARS metrics with strong R^2.

Practical Implications of Results in RAS Training

The developed GBM can be used to identify whether a RAS trainee still needs practice (inexperienced or competent) in performing a specific subtask. As a result, the learning process is accelerated, and the learning expense is reduced because the trainee can focus on practicing those subtasks rather than repeating the entire operation.

Moreover, results showed that extracted eye gaze features can evaluate individual GEARS metrics. The developed models might be employed in RAS training to give trainees specific feedback regarding their skills without the presence of an expert panel. This feedback will speed up the RAS learning process.

When the training process is shortened, more RAS trainees are accepted into training programs and residents complete the program more rapidly. Also, each year more RAS surgeons will be trained, and more patients will be able to benefit from RAS technology. RAS has a shorter hospital stay and fewer surgical complications than conventional methods of surgery,[48] thus hospitals will also benefit from this.

The developed models for skill classification and GEARS metrics evaluation lay down the foundation for objective evaluation of RAS skill and performance, and they might be used to provide trainees objective feedback. As a result, RAS training in various surgical training programs would become standard for all trainees rather than relying on the opinion of an expert panel.

Limitations of This Study and Future Research

Limitations exist despite the novelty of this study. Only 11 participants were included, and 1 expert RAS surgeon assessed GEARS metrics. It is required to validate the developed models by including data from more participants with various specialties from different training programs and including assessments from more expert RAS surgeons. It is required raters watch the video associated with each subtask and assess GEARS metrics.

Using the findings of this study, the next research steps could be (1) including assessments from more expert RAS surgeons from different institutes to develop RAS surgical skill classification and performance evaluation models; (2) developing automatic subtask extraction models; (3) expanding the models developed in this study using data from more participants with a variety of specialties and RAS experiences, from different institutes; and (4) predicting the number of attempts that would move an inexperienced surgeon in the certain subtask to each of the subsequent skill levels. Developing a fully automatic model that receives eye gaze data, extracts subtasks, and detects the skill level and the score of GEARS metrics in performing each subtask could significantly improve the RAS training process. Only by using inexpensive eyeglasses, such a model can be used to provide trainees with feedback regarding their skill and performance that is superior to what a panel of experts currently does.

CONCLUSION

The developed GBM classification model appears promising because it assigned true skill labels using easily and quickly measurable outcomes rather than using operating volume and hours of experience and considered surgical subtasks in skill level classification rather than assigning one skill level label for each participant performing all the tasks that comprise an operation. Even more significant, RAS surgeons' visual metrics can be used in ML models to classify surgical skill levels and to evaluate GEARS metrics objectively. The time to complete a surgical subtask may not necessarily be significantly different across skill levels, and this measurement may not be considered a stand-alone factor for skill-level assessment.

Acknowledgements

The authors would like to thank all participants for participating in the study.

REFERENCES

1. Rogers SO Jr, Gawande AA, Kwaan M, et al. Analysis of surgical errors in closed malpractice claims at 4 liability insurers. Surgery. 2006;140:25–33.
2. Lanfranco AR, Castellanos AE, Desai JP, et al. Robotic surgery: a current perspective. Ann Surg. 2004;239:14–21.
3. Chen J, Cheng N, Cacciamani G, et al. Objective assessment of robotic surgical technical skill: a systematic review. J Urol. 2019;201:461–469.
4. Menekse Dalveren GG, Cagiltay NE. Distinguishing intermediate and novice surgeons by eye movements. Front Psychol. 2020;11:542752.
5. Shah J, Darzi A. Surgical skills assessment: an ongoing debate. BJU Int. 2001;88:655–660.
6. Shafiei SB, Hussein AA, Guru KA. Cognitive learning and its future in urology: surgical skills teaching and assessment. Curr Opin Urol. 2017;27:342–347.
7. Oğul BB, Gilgien M.F, Şahin PD. Ranking robot-assisted surgery skills using kinematic sensors. In: *Ambient Intelligence: 15th European Conference, AmI 2019, Rome, Italy, November 13–15, 2019, Proceedings 15*. Springer, 2019.
8. Funke I, Mees ST, Weitz J, et al. Video-based surgical skill assessment using 3D convolutional neural networks. Int J Comput Assist Radiol Surg. 2019;14:1217–1225.
9. Usluoğulları FH, Tıplamaz S, Yaycı N. Robotic surgery and malpractice. Turk J Urol. 2017;43:425–428.
10. Alemzadeh H, Raman J, Leveson N, et al. Adverse events in robotic surgery: a retrospective study of 14 years of FDA data. PLoS One. 2016;11:e0151470.
11. Babineau TJ, Becker J, Gibbons G, et al. The cost of operative training for surgical residents. Arch Surg. 2004;139:366–370.
12. Triantafyllopoulos G, Stundner O, Memtsoudis S, et al. Patient, surgery, and hospital related risk factors for surgical site infections following total hip arthroplasty. Sci World J. 2015;2015:11979560–11979569.
13. Freilich DA, Cilento BG, Graham D, et al. Perioperative risk factors for surgical complications in pediatric urology: a pilot study in preoperative risk assessment in children. Urology. 2010;76:3–8.
14. Dowdy SC, Borah BJ, Bakkum-Gamez JN, et al. Factors predictive of postoperative morbidity and cost in patients with endometrial cancer. Obstet Gynecol. 2012;120:1419–1427.
15. Cheng H, Clymer JW, Po-Han Chen B, et al. Prolonged operative duration is associated with complications: a systematic review and meta-analysis. J Surg Res. 2018;229:134–144.
16. Campbell DA Jr, Henderson WG, Englesbe MJ, et al. Surgical site infection prevention: the importance of operative duration and blood transfusion—results of the first American College of Surgeons–National Surgical Quality Improvement Program Best Practices Initiative. J Am Coll Surg. 2008;207:810–820.
17. Piper K, Algattas H, DeAndrea-Lazarus IA, et al. Risk factors associated with venous thromboembolism in patients undergoing spine surgery. J Neurosurg Spine. 2017;26:90–96.
18. Kurmann A, Tschan F, Semmer NK, et al. Human factors in the operating room–the surgeon's view. Trends Anaesth Crit Care. 2012;2:224–227.
19. Smith CD, Farrell TM, McNatt SS, et al. Assessing laparoscopic manipulative skills. Am J Surg. 2001;181:547–550.
20. Sánchez R, Rodríguez O, Rosciano J, et al. Robotic surgery training: construct validity of Global Evaluative Assessment of Robotic Skills (GEARS). J Robot Surg. 2016;10:227–231.
21. Diaz-Piedra C, Sanchez-Carrion JM, Rieiro H, et al. Gaze-based technology as a tool for surgical skills assessment and training in urology. Urology. 2017;107:26–30.
22. Tien T, Pucher PH, Sodergren MH, et al. Eye tracking for skills assessment and training: a systematic review. J Surg Res. 2014;191:169–178.
23. Richstone L, Schwartz MJ, Seideman C, et al. Eye metrics as an objective assessment of surgical skill. Ann Surg. 2010;252:177–182.
24. Duchowski AT, Duchowski AT. *Eye Tracking Methodology: Theory and Practice*. Springer; 2017.
25. Van der Gijp A, Ravesloot CJ, Jarodzka H, et al. How visual search relates to visual diagnostic performance: a narrative systematic review of eye-tracking research in radiology. *Adv Health Sci Educ Theory Pract*. 2017;22:765–787.
26. Steichen B, Carenini G, Conati C. User-adaptive information visualization: using eye gaze data to infer visualization tasks and user cognitive abilities. in Proceedings of the 2013 international conference on Intelligent user interfaces. ACM; 2013.
27. Koh RY, Park T, Wickens CD, et al. Differences in attentional strategies by novice and experienced operating theatre scrub nurses. J Exp Psychol Appl. 2011;17:233–246.

28. Khan RS, Tien G, Atkins MS, et al. Analysis of eye gaze: do novice surgeons look at the same location as expert surgeons during a laparoscopic operation?. Surg Endosc. 2012;26:3536–3540.
29. Wilson MR, McGrath JS, Vine SJ, et al. Perceptual impairment and psychomotor control in virtual laparoscopic surgery. Surg Endosc. 2011;25:2268–2274.
30. Lam K, Chen J, Wang Z, et al. Machine learning for technical skill assessment in surgery: a systematic review. npj Digital Med. 2022;5:24.
31. Fard MJ, Ameri S, Darin Ellis R, et al. Automated robot-assisted surgical skill evaluation: predictive analytics approach. Int J Med Robot. 2018;14:e1850.
32. Takimoto E, Maruoka A. Top-down decision tree learning as information based boosting. Theor Comput Sci. 2003;292:447–464.
33. Hutchinson R, Liu L-P, Dietterich T. Incorporating boosted regression trees into ecological latent variable models. in Proceedings of the AAAI Conference on Artificial Intelligence. 2011.
34. Santi V, Notodiputro KA, Sartono B, et al. Generalized Linear Mixed Models by penalized Lasso in modelling the scores of Indonesian students. J Phys Conf Ser. 2021;1869:012140. IOP Publishing
35. Garcia AAT, Garcia CAR, Villasenor-Pineda L, et al. Biosignal Processing and Classification Using Computational Learning and Intelligence: Principles, Algorithms, and Applications. 2021: Academic Press.
36. He H, Bai Y, Garcia EA, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. in 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). 2008. IEEE.
37. Vassiliou MC, Feldman LS, Andrew CG, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. Am J Surg. 2005;190:107–113.
38. Vedula SS, Ishii M, Hager GD. Objective assessment of surgical technical skill and competency in the operating room. Annu Rev Biomed Eng. 2017;19:301–325.
39. Eppsteiner RW, Csikesz NG, McPhee JT, et al. Surgeon volume impacts hospital mortality for pancreatic resection. Ann Surg. 2009;249:635–640.
40. Ericsson KA, Ward P. Capturing the naturally occurring superior performance of experts in the laboratory: toward a science of expert and exceptional performance. Curr Dir Psychol Sci. 2007;16:346–350.
41. Wang Z, Majewicz Fey A. Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. Int J Comput Assist Radiol Surg. 2018;13:1959–1970.
42. Menekse Dalveren GG, Cagiltay NE. Insights from surgeons' eye-movement data in a virtual simulation surgical training environment: effect of experience level and hand conditions. BIT. 2018;37:517–537.
43. Zheng B, Jiang X, Bednarik R, et al. Action-related eye measures to assess surgical expertise. BJS open. 2021;5:zrab068.
44. Appel T, Scharinger C, Gerjets, P, et al. Cross-subject workload classification using pupil-related measures. in Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications. 2018.
45. Beatty J. Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychol Bull. 1982;91:276–292.
46. Sibley C, Coyne J, Baldwin C. Pupil dilation as an index of learning. in Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2011; SAGE Publications Sage CA: Los Angeles, CA.
47. Geng JJ, Blumenfeld Z, Tyson TL, et al. Pupil diameter reflects uncertainty in attentional selection during visual search. Front Hum Neurosci. 2015;9:435.
48. Khorgami Z, Li WT, Jackson TN, et al. The cost of robotics: an analysis of the added costs of robotic-assisted versus laparoscopic surgery using the National Inpatient Sample. Surg Endosc. 2019;33:2217–2221.