Original Article | **Open Access**

Artificial Intelligence Surgery

Kewalramani *et al. Art Int Surg.* 2026;6:697-709 **DOI:**10.20517/ais.2026.72

Automated machine learning assessment and feedback for the Fundamentals of Laparoscopic Surgery peg transfer task

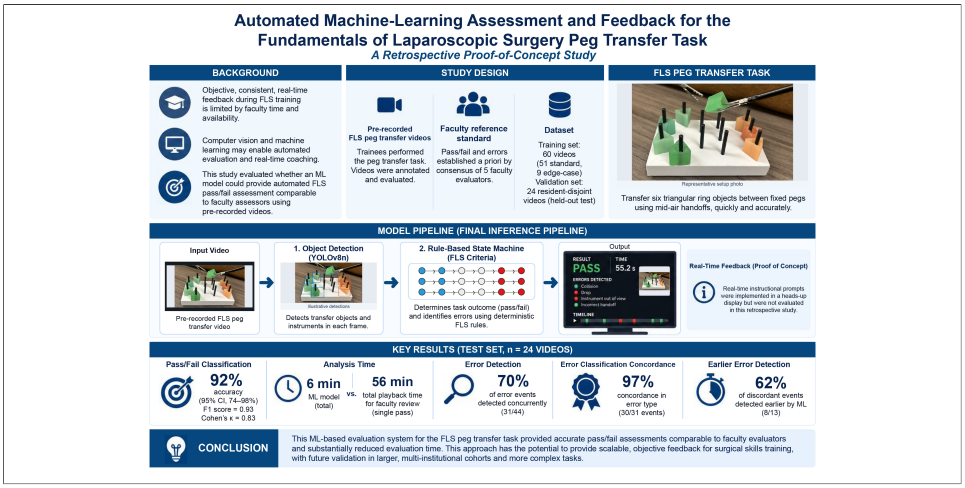
Divya Kewalramani[1], Andrew Hu[1], Yunzhe Xue[2], Ricardo Nunez[1], Finn Kliewer[1], Jenny Cai[1], Joelle Getrajdman[1], Nell Maloney Patel[1], Issam Koleilat[1], Philip S. Barie[3], Amin Madani[4], Usman Roshan[5], Mayur Narayan[1], Advaith Bongu[1]

Keywords:
Computer vision, feasibility study, Fundamentals of Laparoscopic Surgery, laparoscopy, machine learning, real-time feedback, surgical education

Citation: Kewalramani D, Hu A, Xue Y, Nunez R, Kliewer F, Cai J, Getrajdman J, Patel NM, Koleilat I, Barie PS, Madani A, Roshan U, Narayan M, Bongu A. Automated machine learning assessment and feedback for the Fundamentals of Laparoscopic Surgery peg transfer task. *Art Int Surg.* 2026;6:697-709. https://dx.doi.org/10.20517/ais.2026.72

Received: 19 Jul 2026
First Decision: 4 Sep 2026
Revised: 22 Oct 2026
Accepted: 30 Oct 2026
Published: 19 Nov 2026

Academic Editor: Andrew Gumbs
Copy Editor: Tong Wang
Production Editor: Tong Wang



Abstract

Aim: Whereas computer vision has advanced the capabilities of Fundamentals of Laparoscopic Surgery (FLS) trainers, real-time coaching using machine learning (ML) remains limited. This retrospective proof-of-concept study evaluated whether an ML model could provide automated FLS pass/fail assessment comparable to faculty, using pre-recorded videos.

Methods: Residents performed video-recorded FLS peg transfers scored pass/fail. A training set of 60 videos (51 standard cases, 9 edge-case videos) was used to develop a multi-stage pipeline combining a "you only look once" (YOLOv8n) object detector, and a deterministic rule engine encoding FLS pass/fail criteria. The model was evaluated on 24 resident-disjoint, held-out videos. A faculty majority-vote reference standard was set by five evaluators per video. Model predictions were compared against this reference standard.

[1]Department of Surgery, Rutgers Robert Wood Johnson Medical School, New Brunswick, NJ 08901, USA.
[2]Robust AI, LLC, Westfield, NJ 07090, USA.
[3]Division of Trauma, Burns, Acute and Critical Care, Weill Cornell Medicine, New York, NY 10065, USA.
[4]Department of Surgery, University of Toronto, Toronto M5T 1P5, Canada.
[5]Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102, USA.

Correspondence to: Dr. Advaith Bongu, Department of Surgery, Rutgers Robert Wood Johnson Medical School, New Brunswick, NJ 08901,

USA. E-mail: ab497@rwjms.rutgers.edu

Results: The ML model classified outcomes accurately in 92% of cases [95%CI: 74%-98%; F1 score = 0.93 (pass positive); Cohen's κ = 0.83]. Total analysis time was 6 min *vs.* approximately 56 min of video playback. The model concurrently identified errors in 70% (31/44) of error events, with 97% (30/31) concordance in error classification. Among the 13 remaining events, the model preceded the faculty-annotated timestamp in 8/13 (62%). Real-time instructional prompts were implemented in a heads-up display but not evaluated here.

Conclusion: The system provided accurate FLS peg-transfer assessment comparable to faculty while reducing evaluation time. As a retrospective proof-of-concept, it supports automated assessment and, pending prospective validation, could deliver scalable real-time feedback in skills training. External validation in larger, multi-institutional cohorts should precede adoption.

INTRODUCTION

Over the past two decades, surgical education has shifted progressively from apprenticeship-based training toward competency-based assessment while preserving the central importance of graduated operative responsibility[1]. What has not changed is that trainees acquire technical skills at markedly different rates[2]. Coaching, defined broadly as structured faculty guidance aimed at improving performance, remains central to this process[3]. However, duty-hour regulations have constrained opportunities for hands-on mentorship, elevating simulation-based training as a necessary adjunct for technical skill acquisition[2].

The Fundamentals of Laparoscopic Surgery (FLS) program, introduced in 1997, became a prerequisite for American Board of Surgery certification in the 2009-2010 academic year. Its task trainer assesses core laparoscopic competencies across five standardized exercises, including peg transfer (PT), pattern cutting, ligating-loop placement, intracorporeal suturing, and extracorporeal suturing, and performance on these tasks correlates with intraoperative proficiency, although competence is attained only through sustained, deliberate practice[4,5]. Skill acquisition is optimized further by distributed practice, with shorter inter-session intervals conferring greater benefit. Delivering such spaced training remains logistically challenging within conventional residency structures[6-8].

This study focused on the PT task, in which six objects are transferred as rapidly as possible from pegs on one side of a board to the other by way of a mid-air handoff. Although the mechanics are simple, the task trains complementary use of the dominant and non-dominant hands and is scored against explicit pass/fail criteria[9], features that make it well-suited to objective, automated analysis. Consequently, PT videos have become a prime target for computer vision (CV)-based machine learning (ML)[10,11]. Real-time "you only look once" (YOLO) detectors have since been applied to streaming video in other domains, underscoring the maturity of this architecture for on-the-fly analysis[12]. ML is well suited to simulation-based assessment because expert evaluation of laparoscopic skill likewise depends on recognizing recurring visual and temporal patterns of movement[13-17].

Prior CV and ML systems for laparoscopic skill assessment have served a largely summative role, generating a score or performance metric only after a task is completed[18,19]. Such retrospective feedback is valuable for benchmarking but arrives too late to shape a trainee's actions during the attempt itself and does little to offload the moment-to-moment burden of supervision on faculty. Providing guidance in real time is a distinct and more demanding task, both computationally and pedagogically, and remains largely unaddressed. Whether ML can meet this challenge and augment faculty coaching during the early phase of technical skill acquisition is unresolved. Therefore, we developed and retrospectively evaluated an ML system

Table 1. Held-out test-set composition by PGY and outcome (faculty reference standard)

PGY	*n*	Pass	Fail	Pass %
PGY1	8	5	3	63%
PGY2	3	3	0	100%
PGY3	6	3	3	50%
PGY4	3	2	1	67%
PGY5	4	2	2	50%
Total	24	15	9	63%

PGY: Postgraduate year.

for automated assessment of the FLS PT. We hypothesized that its real-time evaluations would achieve substantial agreement with expert faculty consensus, comparable to the agreement observed among individual faculty evaluators.

METHODS

Study design and ethics

Study approval was obtained from the Rutgers University Institutional Review Board (IRB) (Pro2022000164). The IRB determined the study to be non-human-subjects research; recordings contained no identifiable participant features, so individual informed consent was not required. De-identified videos were stored on a password-encrypted workstation and shared only via password-protected cloud storage. The training and held-out test sets were resident-disjoint: videos were split by resident so that no resident (and therefore no recording session) contributed videos to both sets, ensuring that the model was evaluated on residents it had not been exposed to during training. Each resident was unique and contributed a single video; the held-out test set comprised 24 residents [postgraduate year (PGY)1, n = 8; PGY2, n = 3; PGY3, n = 6; PGY4, n = 3; PGY5, n = 4], as detailed in Table 1.

Video acquisition and dataset

Videos of general surgery residents performing the FLS PT were recorded using a standard FLS box trainer camera (Limbs & Things, Savannah, GA). The initial training dataset consisted of 51 videos, a size selected to target a preliminary detection accuracy of at least 80%[20]. To enhance model performance in detecting edge cases, particularly drops off the pegboard, an additional nine videos showcasing different configurations of these errors were incorporated. Each of the 60 development videos was contributed by a distinct resident (one video per resident), comprising 51 standard cases and 9 additional residents whose recordings captured naturally occurring edge-case errors; development and held-out test residents did not overlap (resident-level split). Video frames were extracted at 1 frame/s for training-set construction; for evaluation, the pipeline processed video at 10 frames/s, and the live system supports 5, 10, or 15 frames/s (default 15 frames/s). Objects of interest were annotated by manually-drawn bounding boxes using in-house customized Python software (Python Software Foundation, Beaverton, OR), which recorded the box coordinates as training labels for the YOLO model. Annotations were made frame-by-frame at one-second granularity, primarily by the first author with assistance from trained laboratory staff for a subset of videos (particularly outlier cases); a random 10% of non-primary annotations were quality-checked against a fixed action key. Formal inter-annotator agreement was not quantified. The custom software is available from the authors upon reasonable request.

Model development

Several CV approaches have been applied to laparoscopic skill assessment, including region-based (two-stage) convolutional neural network (CNN) detectors such as Faster R-CNNs, single-stage detectors such as the YOLO family, and pose- or kinematics-based tracking methods. Two-stage detectors may offer marginal

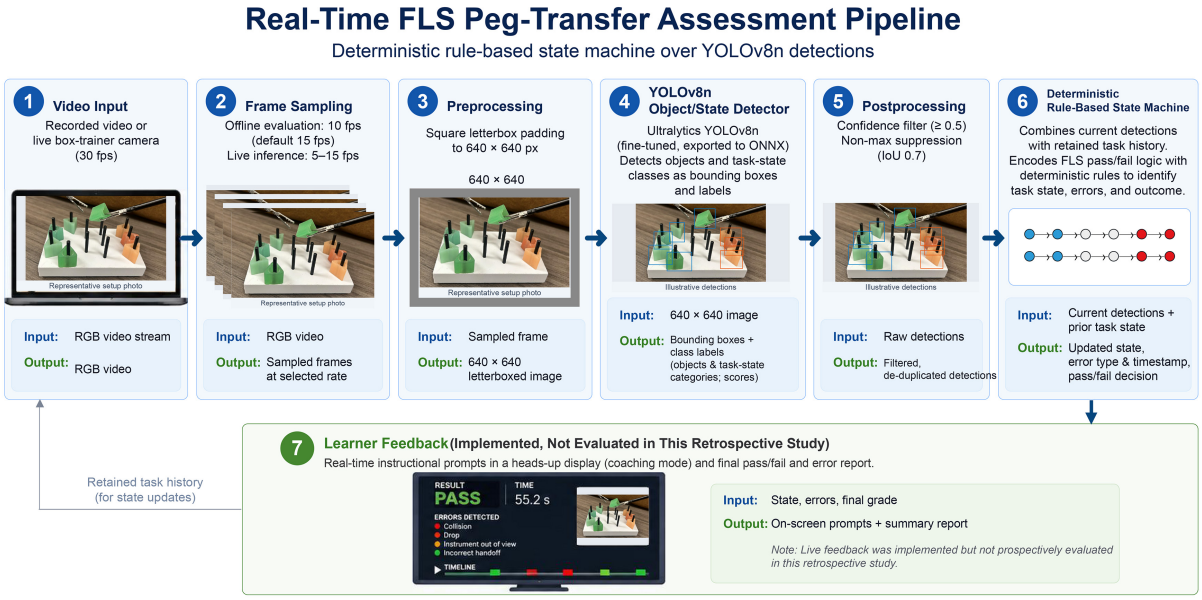


Figure 1. Processing pipeline of the automated assessment system. Recorded-video or live-camera input is sampled at the selected frame rate, letterbox-padded to 640 × 640 pixels, and passed to the YOLOv8n detector; confidence filtering and non-maximum suppression yield object and task-state detections, which a deterministic rule-based state machine combines with retained task history to update task state, detect errors, and render a pass/fail decision with on-screen learner feedback. RGB: Red, green, blue; px: pixels; YOLO: You Only Look Once; ONNX: Open Neural Network Exchange; IoU: intersection of union score; FLS: Fundamentals of Laparoscopic Surgery.

gains in accuracy but are computationally slower, a crucial limitation for real-time feedback. For this model, the YOLOv8n object detection model (Ultralytics) was utilized due to its speed and accuracy in detecting and classifying multiple objects simultaneously[11] through a deep CNN[21]. The detector was initialized from pretrained YOLOv8n weights and fine-tuned on the annotated frames, then exported to Open Neural Network Exchange (ONNX) (opset 12) at 640 × 640 input resolution; inference used a confidence threshold of 0.5 and non-maximum suppression at an intersection of union score (IoU) of 0.7. Training used the Ultralytics default schedule (automatic optimizer selection, batch size 16, initial and final learning-rate factors of 0.01, momentum 0.937, weight decay 0.0005, and a three-epoch warm-up) for 80 epochs, with mosaic augmentation disabled for the final 10 epochs; offline augmentation included horizontal mirroring with coordinate reflection and left/right class relabeling. Hyperparameters were selected on an 80/20 resident-grouped split, after which the final model was retrained on all 60 development videos (60 unique residents). The complete processing pipeline is shown in Figure 1.

Task-specific actions were identified and labeled (e.g., Left grasper pick, Transfer, Right grasper pick, Left peg drop, Right peg drop, Drop on board, Drop off board). A temporal convolutional network[22] was developed for temporal activity recognition, using a sequence of preceding frames to predict the activity in the final frame. In subsequent development, the YOLOv8n detector - trained on manually annotated example frames - recognized the same activity classes directly from individual frames, and these detections became the primary activity-recognition signal in the scoring pipeline; the temporal network was retained for incorporating temporal information but was not the primary source of activity classification. The activity labels themselves were learned by the YOLO detector, which was trained on manually-labeled example frames (e.g., frames showing a left-grasper pick were annotated as such). The predicted activity for each frame was then compared with the preceding activity using a deterministic rule set. For example, if the preceding activity was “pick with left grasper” and the current prediction was “place with left grasper”, the sequence was flagged as an error and reported to the learner; if the current prediction was “transfer”, the sequence was scored as correct, and the task was allowed to proceed. These logical rules were implemented

from established FLS scoring guidelines[23], faithfully translating the existing pass/fail criteria into software. For example, task completion required sequential actions: initiation by left grasper, mid-air transfer, proper drop on peg. Drops on board could be resumed if the object was picked up by the same grasper that initiated the drop. Errors such as off-board drops and completion time exceeding 300 s were encoded as failures.

Ground truth (GT) and outcomes

Held-out test videos underwent independent evaluation by a panel of five faculty members; the reference-standard outcome (pass/fail) for each video was the majority vote of the five. Because the panel comprised five evaluators, the consensus label was determined by simple majority vote, which precluded ties. The primary outcome was task classification (pass/fail); secondary analyses addressed error detection and the type and timing of errors. For each failed task, the type and timing of errors were compared between faculty assessments and ML predictions. A two-second tolerance window was imposed to account for delays in frame detection and reporting for the human component. Error classifications included: "No transfer detected", "Wrong grasper error", "Improper transfer", "Improper pickup", "Time exceeded", and "Drop off board".

Statistical analysis

As a feasibility study, the validation-set size was pragmatic rather than formally powered for efficacy. Using inter-rater reliability hypothesis testing (Cohen's κ), with a minimum acceptable reliability of 0.70, an expected κ of 0.90, two-tailed significance of 0.05, power of 80%, and two raters (model *vs.* reference standard), the required sample size was $n = 24$; the calculation was performed with the Rutgers Biostatistics and Epidemiology Consulting Service. Model-consensus agreement on task outcome was quantified by Cohen's κ, with $\kappa \geq 0.70$ pre-specified as the threshold for feasibility. Accuracy (with Wilson score 95% confidence intervals) and the F1 score were reported alongside κ, and McNemar's test assessed directional asymmetry in model-consensus disagreements. Analyses were performed in Statistical Package for the Social Sciences (SPSS) v29 (IBM Corp., Armonk, NY).

Sensitivity, specificity, predictive values, and F1 were additionally computed for a failure-positive framing with Wilson intervals, and inter-rater reliability among the five faculty was summarized with Fleiss' κ. Error-detection and timing analyses were exploratory and descriptive given clustering of events within videos. A two-sided α of 0.05 was applied throughout.

RESULTS

Classification performance

General surgery residents from postgraduate years (PGY) 1-5 contributed videos; the distribution of the 60 training videos by resident PGY level is shown in Table 2, and the composition of the 24-video held-out test set by PGY and outcome is shown in Table 1. Model validity metrics for the ML model and each of the five faculty evaluators are summarized in Table 3. Against the faculty-consensus GT, the ML model achieved an accuracy of 92% (22/24 videos; Wilson 95%CI: 74.2%-97.7%), an F1 score of 0.93 (with pass as the positive class), and a Cohen's κ of 0.83, indicating substantial agreement with the faculty-derived GT. McNemar's test showed no significant difference in the direction of disagreements between the ML model and the faculty consensus ($P = 0.50$). With only a small number of discordant pairs, however, McNemar's test is underpowered and indicates absence of evidence of a difference rather than evidence of equivalence; a formally powered non-inferiority study would be required to establish that the model performs as well as faculty.

Individual faculty agreement with the consensus GT varied widely (accuracy 67%-100%; Cohen's κ 0.40-1.0) [Table 3], and the model's performance fell within, and toward the upper end of, this range. This spread among expert raters is itself notable. To make the classification interpretable given the high pass rate, the GT

Table 2. Training video breakdown by resident PGY level (*N* = 60 videos: 51 standard cases plus 9 edge-case videos)

Level	*N* = 60
PGY1	16
PGY2	18
PGY3	6
PGY4	6
PGY5	5
EDGE*	9

*EDGE: edge-case videos curated deliberately to capture uncommon error configurations (e.g., drops off the pegboard) to improve model robustness; these are included in the 60-video training total. PGY: Postgraduate year.

Table 3. Validity metrics for the ML model

		% Correct	Cohen's κ	McNemar's test	Total time (min)*	F1 score
Model		92% (22/24)	0.830	0.500	6	0.93
	1.	96% (23/24)	0.913	1.000		
	2.	67% (16/24)	0.396	0.008		
Faculty	3.	100% (24/24)	1.000	1.000	56	-
	4.	83% (20/24)	0.673	0.125		
	5.	96% (23/24)	0.909	1.000		

Each of the five faculty evaluators shown against the faculty-consensus ground truth with percent correct, Cohen's κ, McNemar's test *P*-value and F1 score. Wide variation across individual faculty shows inter-rater variability among expert evaluators. *Model "Total time" (6 min) is the end-to-end time to process all 24 test videos sequentially. Faculty "Total time" (≈ 56 min) is the cumulative playback duration of the 24 videos, used only as a lower-bound proxy for a single-pass review, not a measured review time; the two values are not directly comparable as evaluation efficiency. ML: Machine learning.

Table 4. Confusion matrix: model *vs.* faculty reference standard (pass = positive; *n* = 24)

	Model: Pass	Model: Fail
Reference: Pass	13 (true pass)	2 (false fail)
Reference: Fail	0 (false pass)	9 (true fail)

class distribution and confusion matrix are as follows: both misclassifications were false failures (subjects 4 and 19), indicating the model's errors were directionally conservative, it was over-strict rather than over-lenient, a potentially desirable property for a learner-facing tool.

Treating failure as the positive (clinically important) class, the model detected every true failure (sensitivity 100%, 9/9; 95%CI: 70%-100%) with a specificity of 86.7% (13/15), a positive predictive value of 81.8%, a negative predictive value of 100%, an F1 of 0.90, and a balanced accuracy of 93.3%. The full confusion matrix is shown in Table 4.

Inter-rater reliability among the five faculty was moderate (Fleiss' κ = 0.60); the model's agreement with the reference standard (κ = 0.83) therefore exceeded the faculty's agreement with one another. When performance was stratified by strength of faculty agreement, the model classified all 14 unanimously graded tasks correctly and all four tasks with a 3/5 split correctly; both misclassifications occurred among the six 4/5-split tasks, i.e., cases in which the faculty themselves were divided.

As pass outcomes predominated in this validation cohort, overall classification accuracy was influenced by the high base rate of task success, analogous to the way predictive values are shaped by disease prevalence. To

characterize performance independent of this base rate, we report metrics that are less sensitive to class imbalance: Cohen's κ (0.83) and the F1 score (0.93), both of which account for chance agreement and the balance of precision and recall, respectively. These agreement-based measures support the conclusion that model performance was not an artifact of the high pass rate alone.

Evaluation efficiency

Analysis time was markedly reduced with the ML model, which processed all 24 validation videos in a total of 6 min, representing total wall-clock processing time, inclusive of both computer processing unit (CPU) and graphics processing unit (GPU) computation. By comparison, the 24 videos have a combined playback duration of approximately 56 min (mean 2.3 min per video); this total is used as a lower-bound proxy for a single-pass faculty review rather than a measured review time.

Error detection and timing

Of the 44 error events identified by faculty during failed tasks, which arose within nine failed tasks (a single failed video could contain more than one error event), the ML model independently detected an error concurrently with the faculty assessment in 70% (31/44) of events. Among these 31 concurrently detected errors, the model assigned the same error category as the faculty in 97% (30/31), indicating near-perfect agreement on error type once an error had been jointly identified. For the remaining 30% (13/44) of error events, the model and faculty did not flag the error at the same moment; in these discordant cases, the model's timestamp preceded the faculty-annotated timestamp in 8/13 (62%) and followed it in 5/13. Because faculty annotated recorded video without real-time constraints, an earlier flag is not necessarily superior: it could reflect genuine early detection of the same error or a premature trigger on a precursor movement. We therefore report these as discordant in timing rather than as evidence of model superiority. Concurrent detection was robust to the choice of matching window (27/44, 31/44, 33/44 and 34/44 within ±1, ±2, ±3, and ±5 s, respectively); matching was driven by wrong-grasper/no-transfer events (31 of 41 matched within ±2 s), whereas none of the three improper-transfer events matched (0 of 3); these two categories together account for all 44 faculty error events.

Real-time feedback interface

Real-time instructional prompts and task-specific feedback were integrated successfully into a heads-up display interface [Figure 2]. We distinguish two operating modes: an assessment mode (retrospective pass/fail and error scoring, evaluated here) and a coaching mode (live on-screen prompts), the latter implemented but not evaluated in this study. The architecture of the real-time coaching system is summarized in Table 5. The ML model provided real-time identification of the movements performed [Figure 2A], offered instructional guidance for subsequent steps through an integrated "notes" section on the heads-up display [Figure 2B], and generated comprehensive final reports detailing task performance and grading outcomes [Figure 2C].

DISCUSSION

This study demonstrated that an ML model employing YOLOv8n can provide accurate automated evaluation and feedback for the FLS PT, demonstrated here retrospectively on recorded video, achieving substantial concordance with expert faculty assessments. The high accuracy (92%), strong inter-rater agreement (Cohen's κ = 0.83), and efficiency gains, marked by a significant reduction in analysis time compared to manual faculty review, illustrate the potential of ML to enhance surgical skills training.

Compared with existing artificial intelligence (AI) tools in laparoscopic skills assessment, which typically provide summative evaluations after task completion, our model uniquely offers real-time, stepwise instructional guidance[24,25]. Lazar *et al.* employed YOLOv4-based systems to provide detailed metrics such as

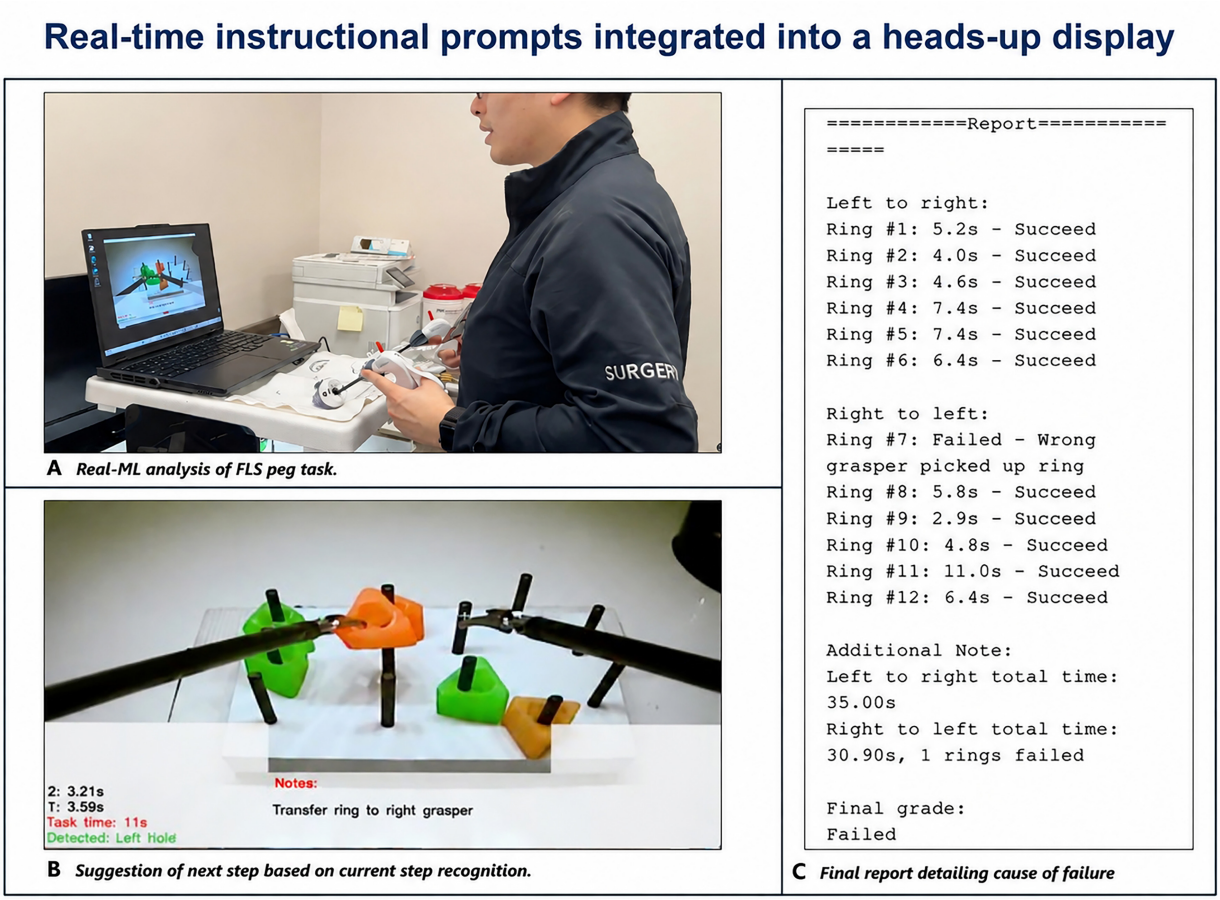


Figure 2. Demonstration of real-time peg transfer assessment model. (A) real-time identification of the movements performed; (B) step-by-step instructional guidance presented in the on-screen "notes" section of the heads-up display; and (C) the comprehensive final report detailing task performance and grading outcomes. Original figure; interface screenshots were captured by the authors from the study software, and the photographic image was taken by the authors with an iPhone (Apple Inc., Cupertino, CA). ML: Machine learning; FLS: Fundamentals of Laparoscopic Surgery.

Table 5. Components of the real-time machine learning coaching system

Component	Function	Output to Learner
Object detection (YOLOv8n)	Identifies graspers, pegs, and objects	Continuous object localization
Activity recognition	Determines current task step	Current procedural state
Rule engine	Applies FLS scoring logic	Correct/incorrect action determination
Error detection	Identifies rule violations	Immediate error notification
Feedback generator	Produces instructional prompts	Real-time coaching message
Final scoring module	Summarizes performance	Pass/fail and error report

FLS: Fundamentals of Laparoscopic Surgery; YOLO: you only look once.

motion efficiency and bimanual coordination, distinguishing skill levels; however, their system offered post-task analysis[11]. In exploring ML innovations in surgical training, Varas *et al.* highlights the opportunity for personalized real-time feedback[13]. Our approach extends these foundational studies by integrating instantaneous error detection and immediate corrective guidance, thereby facilitating in-the-moment skill refinement. Additionally, initial research with PT has focused on the use of sophisticated biosensors, virtual reality, three-dimensional movement, or multiple cameras to track progress[26]. Our model's compatibility with standard FLS box trainers enhances accessibility without requiring specialized hardware or advanced virtual reality systems.

The integration of automated real-time feedback into surgical training carries substantial educational advantages[27-31]. Instantaneous error identification and guidance enable learners to rectify mistakes promptly, potentially accelerating skill acquisition and reducing dependence on continuous faculty supervision. Additionally, the marked reduction in evaluation time offers efficiency benefits, freeing educators to focus on advanced teaching and personalized mentorship. Analysis of error timing and error-type concordance with faculty provided insight into the features the model used to identify errors.

A central question raised by this study is not merely whether a machine can match faculty on a binary pass/fail decision, but what kind of feedback supports learning most effectively. Pass/fail scoring alone conveys little about how a trainee should improve, and effective coaching encompasses far more than error detection[32]. Coaching also includes diagnosing the cause of an error, modeling correct technique, and adapting guidance to the individual learner. Our current system provides stepwise procedural feedback (identifying which action was performed and whether it violated a rule) but does not yet provide the qualitative, explanatory feedback a faculty coach offers in person. We did not compare the instructional content of ML-generated *vs.* in-person faculty feedback; doing so is an important direction for future work. We therefore consider this system an adjunct that can offload high-frequency, objective error detection, freeing faculty time for the higher-order teaching that remains beyond machine capabilities, rather than a replacement for human coaching.

The PT is the most structured FLS exercise, with discrete objects and largely binary, readily observable failure criteria, which makes it particularly amenable to computer-vision analysis. Performance on this task should therefore be interpreted as a favorable initial case rather than as evidence that the approach will generalize directly to more complex exercises. Tasks such as intracorporeal suturing involve continuous assessments of technical quality (for example, tissue handling, knot security, and economy of motion) that are not captured well by binary rules and on which even experts may reasonably disagree. Extending ML to these tasks will likely require larger and more diverse training sets, richer annotation schemes, and models that output graded or narrative assessments rather than pass/fail labels; inter-rater agreement is also likely to be lower for such non-binary outcomes. Establishing whether the efficiency and agreement observed here can be replicated under different conditions is an essential next step.

Our model's performance was generally comparable to, and in some cases exceeded, that of individual faculty evaluators. However, it graded two cases incorrectly as failures due to its rigid interpretation of predefined rules: one instance (subject 4) involved premature advancement without fully securing a triangle on the peg, and another (subject 19) occurred when the resident grasped the next triangle before the last one was released. A larger and more diverse training dataset may reduce these errors. Furthermore, only 58% (14/24 videos) of cases reached faculty unanimity on a grade, highlighting inherent inter-rater variability and the potential for grading fatigue among human evaluators. That expert evaluators disagreed on the correct grade in nearly one-half of cases has a direct practical implication: when human raters themselves lack consensus, treating any single human assessment, or an automated system trained to reproduce it, as definitive GT in high-stakes evaluation (e.g., certification or promotion decisions) risks encoding that variability as if it were objective. This supports using automated tools to augment rather than replace human judgment, and for careful calibration before such systems inform consequential decisions. A further caveat is that each faculty member contributed to the consensus against which they were then scored, so individual faculty-GT agreement [Table 3] is inflated by this circularity.

Although the current ML model provides simple, instructional feedback, its strength lies in its ability to evolve for continued development. The initial reliance on edge cases demonstrated the feasibility of expanding feedback modules to address additional performance factors such as "improper grasper position" or "hand weakness". As the model becomes more adept at recognizing these nuanced features, richer and more sophisticated real-time feedback can be integrated seamlessly into the training interface.

Nevertheless, this study has several limitations. First, the relatively small dataset and single-institution design may limit the generalizability of results. Future research involving larger, diverse datasets from multiple institutions would enhance the robustness and applicability of the model. Second, the current evaluation focuses solely on PT; thus, validation across other FLS exercises and more complex procedural tasks is necessary. Finally, our study did not assess long-term educational outcomes, such as skill acquisition, retention, proficiency progression, or performance on certification examinations. Further work will be required to determine the true educational impact of such tools. An additional limitation concerns the GT itself: faculty reached unanimous agreement on only 58% of cases, so the consensus standard against which the model was judged carries inherent uncertainty, and agreement metrics should be interpreted in that light. The high task-pass rate in this cohort also means overall accuracy is partly a function of the base rate; performance on cohorts or tasks with higher failure rates may differ. Finally, the rule-based scoring that makes the system transparent also makes it rigid, as illustrated by the two cases the model failed incorrectly for narrow rule violations; more flexible models may be needed as task complexity grows. Importantly, although the system is designed for real-time use, validation in this study was retrospective: the model was evaluated on recorded performances and no trainee received live, in-session feedback. Prospective evaluation in live training sessions is needed to confirm real-time feasibility and educational value. In addition, although offline evaluation processed video at 10 frames/s, sampling density may still need to increase to capture faster movements in more complex tasks or in demanding live use.

Extension of this approach to additional FLS tasks, individualized feedback, and longitudinal assessment would provide a more comprehensive evaluation of its educational utility. The most important remaining question is whether real-time ML feedback improves learning rather than simply reproducing faculty assessment. Prospective, multi-institutional comparisons of ML-assisted and conventional feedback would help determine whether the observed efficiency translates into measurable educational benefit.

Conclusion

Our results indicate that an ML-based evaluation system can accurately and efficiently provide objective assessments aligning closely with expert faculty evaluations. Ultimately development and validation of comprehensive scalable AI-based feedback systems for real-time use could advance surgical education meaningfully[33,34], providing trainees with consistent, effective skill-enhancement tools.

DECLARATIONS

Authors' contributions

Made substantial contributions to formal analysis, original draft preparation, review & editing: Kewalramani D, Narayan M, Barie PS, Madani A, Koleilat I

Provided support for data curation, formal analysis, investigation, methodology, project administration, validation, visualization, review & editing: Hu A, Kliewer F, Nunez R

Made substantial contributions to conceptualization, data curation, formal analysis, supervision, review & editing: Patel NM, Cai J, Getrajdman J

Made substantial contributions to conceptualization, data curation, formal analysis, resources, software, validation: Xue Y, Roshan U

Contributed to conceptualization, data curation, formal analysis, investigation, methodology, resources, project administration, funding acquisition, validation, visualization, supervision: Bongu A

Availability of data and materials

The data that support our findings are available from the corresponding author upon reasonable request. The videos themselves are not publicly available due to IRB restrictions.

AI and AI-assisted tools statement

Not applicable.

Financial support and sponsorship

This project was supported through an internal university grant (Rutgers University Innovator Investigator Award, August 2023; Project No. 303998; Task No. 300; award amount: $10,000; principal investigator: Bongu A). Grant funding was used for project resources, including hardware and software services.

Conflicts of interest

Kewalramani D, Barie PS, and Narayan M are Guest Editors of *Artificial Intelligence Surgery*. Narayan M is an Editorial Board Member of *Artificial Intelligence Surgery*. They were not involved in any steps of editorial processing, notably including reviewers' selection, manuscript handling, and decision-making. The other authors declare that there are no conflicts of interest.

Ethical approval and consent to participate

Study approval was obtained from the Rutgers University Institutional Review Board (Pro2022000164). No specific consent was required since this was deemed non-human subject research.

Consent for publication

The author, Hu A, has provided written consent for the publication of his identifiable image shown in Figure 2A and in the accompanying video abstract.

Copyright

© The Author(s) 2026.

REFERENCES

1. Petrucci AM, Chand M, Wexner SD. Social media: changing the paradigm for surgical education. *Clin Colon Rectal Surg.* 2017;30:244-51. DOI PubMed PMC
2. Grantcharov TP, Funch-Jensen P. Can everyone achieve proficiency with the laparoscopic technique? Learning curve patterns in technical skills acquisition. *Am J Surg.* 2009;197:447-9. DOI PubMed
3. Louridas M, Sachdeva AK, Yuen A, Blair P, Macrae H. Coaching in surgical education: a systematic review. *Ann Surg.* 2022;275:80-4. DOI PubMed
4. Higgins RM, Turbati MS, Goldblatt MI. Preparing for and passing the fundamentals of laparoscopic surgery (FLS) exam improves general surgery resident operative performance and autonomy. *Surg Endosc.* 2023;37:6438-44. DOI PubMed
5. Sroka G, Feldman LS, Vassiliou MC, Kaneva PA, Fayez R, Fried GM. Fundamentals of Laparoscopic Surgery simulator training to proficiency improves laparoscopic performance in the operating room - a randomized controlled trial. *Am J Surg.* 2010;199:115-20. DOI PubMed
6. Gallagher AG, Jordan-Black JA, O'Sullivan GC. Prospective, randomized assessment of the acquisition, maintenance, and loss of laparoscopic skills. *Ann Surg.* 2012;256:387-93. DOI PubMed
7. Stefanidis D, Walters KC, Mostafavi A, Heniford BT. What is the ideal interval between training sessions during proficiency-based laparoscopic simulator training? *Am J Surg.* 2009;197:126-9. DOI PubMed
8. Jørgensen HK, Vamadevan A, Konge L, Hertz P, Bjerrum F. Distributed training vs. massed practice for surgical skills training - a systematic review. *Surg Endosc.* 2025;39:39-63. DOI PubMed
9. Fraser SA, Klassen DR, Feldman LS, Ghitulescu GA, Stanbridge D, Fried GM. Evaluating laparoscopic skills: setting the pass/fail score for the MISTELS system. *Surg Endosc.* 2003;17:964-7. DOI PubMed
10. Belmar F, Gaete MI, Escalona G, et al. Artificial intelligence in laparoscopic simulation: a promising future for large-scale automated evaluations. *Surg Endosc.* 2023;37:4942-6. DOI PubMed PMC
11. Lazar A, Sroka G, Laufer S. Automatic assessment of performance in the FLS trainer using computer vision. *Surg Endosc.* 2023;37:6476-82. DOI PubMed

12. Moshayedi AJ, Wang Z, Sharifdoust M, et al. Smart farming solutions: a user-friendly GUI for maize tassel estimation using YOLO with dynamic and fixed labelling, featuring video support. *IEEE Access*. 2025;13:57809-33. DOI
13. Varas J, Coronel BV, Villagrán I, et al. Innovations in surgical training: exploring the role of artificial intelligence and large language models (LLM). *Rev Col Bras Cir*. 2023;50:e20233605. DOI PubMed PMC
14. Ogbonnaya CN, Li S, Tang C, et al. Exploring the role of artificial intelligence (AI)-driven training in laparoscopic suturing: a systematic review of skills mastery, retention, and clinical performance in surgical education. *Healthcare*. 2025;13:571. DOI PubMed PMC
15. Winkler-Schwartz A, Bissonnette V, Mirchi N, et al. Artificial intelligence in medical education: best practices using machine learning to assess surgical expertise in virtual reality simulation. *J Surg Educ*. 2019;76:1681-90. DOI PubMed
16. Maier-Hein L, Vedula SS, Speidel S, et al. Surgical data science for next-generation interventions. *Nat Biomed Eng*. 2017;1:691-6. DOI PubMed
17. Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial intelligence in surgery: promises and perils. *Ann Surg*. 2018;268:70-6. DOI PubMed PMC
18. Riddle EW, Kewalramani D, Narayan M, Jones DB. Surgical simulation: virtual reality to artificial intelligence. *Curr Probl Surg*. 2024;61:101625. DOI PubMed
19. Kewalramani D, Jawa RS, Martin CA, et al. Position statement from the society of University surgeons, surgical education committee: artificial intelligence in surgical training for medical students, residents, and fellows. *Surgery*. 2026;190:109849. DOI PubMed
20. Bar O, Neimark D, Zohar M, et al. Impact of data on generalization of AI for surgical intelligence applications. *Sci Rep*. 2020;10:22208. DOI PubMed PMC
21. Jocher G, Qiu J, Liu M, Lyu S, Akyon FC, Kalfaoglu ME. Ultralytics YOLO26: Unified Real-Time End-to-End Vision Models. Available from: https://github.com/ultralytics/ultralytics. [Last accessed on 4 Nov 2026].
22. Bai S, Kolter JZ, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv* 2018;arXiv:1803.01271. Available from https://doi.org/10.48550/arXiv.1803.01271 [accessed 4 November 2026].
23. Fundamentals of Laparoscopic Surgery. FLS manual skills: written instructions and performance guidelines (revised February 2014). Los Angeles (CA): FLS Program; 2014. Available from: https://www.flsprogram.org/wp-content/uploads/2014/03/Revised-Manual-Skills-Guidelines-February-2014.pdf. [Last accessed on 4 Nov 2026].
24. Power D, Burke C, Madden MG, Ullah I. Automated assessment of simulated laparoscopic surgical skill performance using deep learning. *Sci Rep*. 2025;15:13591. DOI PubMed PMC
25. Hung AJ, Chen J, Gill IS. Automated performance metrics and machine learning algorithms to measure surgeon performance and anticipate clinical outcomes in robotic surgery. *JAMA Surg*. 2018;153:770-1. DOI PubMed PMC
26. Castillo-Segura P, Fernández-Panadero C, Alario-Hoyos C, Muñoz-Merino PJ, Delgado Kloos C. A cost-effective IoT learning environment for the training and assessment of surgical technical skills with visual learning analytics. *J Biomed Inform*. 2021;124:103952. DOI PubMed
27. Kewalramani D, Roman DS, Lagos SA, et al. Quality matters: artificial intelligence-based assessment of feedback quality predicts technical skill improvement. *Surgery*. 2025;187:109625. DOI PubMed
28. Cheng K, Wu S, Peng B, Wang X. An artificial intelligence-enhanced coaching mode. *Int J Surg*. 2025;111:6469-72. DOI PubMed PMC
29. Malpani A, Vedula SS, Lin HC, Hager GD, Taylor RH. Real-time teaching cues for automated surgical coaching. *arXiv* 2017;arXiv:1704.07436. Available from https://doi.org/10.48550/arXiv.1704.07436 [accessed 4 November 2026].
30. Ma X, Wijewickrema SNR, Zhou Y, Zhou S, O'Leary S, Bailey J. Providing effective real-time feedback in simulation-based surgical training. In: Descoteaux M, Maier-Hein L, Franz A, Jannin P, Collins D, Duchesne S, Editors. Medical Image Computing and Computer-Assisted Intervention - MICCAI 2017. Lecture Notes in Computer Science, Vol 10434. Springer, Cham; 2017. pp. 566-74. DOI
31. Lin J, Cella L, Goldenberg MG, Hung AJ. Artificial intelligence-based approaches to augmenting and automating surgical training. *BJU Int*. 2026;137:941-8. DOI PubMed PMC
32. Kewalramani D, Grunhut J, Jopling J, Narayan M. Distinguishing signal from perception: what video-based assessment reveals about surgical decision-making and technical performance. *J Am Coll Surg*. 2026;242:1478-80. DOI PubMed
33. Benton J, Kewalramani D, Nieman DR, et al. Familiar but unprepared: artificial intelligence training needs in graduate medical education. *Surgery*. 2026;197:110389. DOI PubMed
34. Castillo-Medina A, Calleja-Zardain R, Kewalramani D, Narayan M, Julio Mayol J. Inteligencia artificial como herramienta de la cirugía global en América Latina. *Rev Colomb Cir*. 2025;40:25-32. (in Spanish). DOI

Disclaimer/Publisher's Note: All statements, opinions, and data contained in this publication are solely those of the individual author(s) and contributor(s) and do not necessarily reflect those of OAE and/or the editor(s). OAE and/or the editor(s) disclaim any responsibility for harm to persons or property resulting from the use of any ideas, methods, instructions, or products mentioned in the content.

© The Author(s) 2026. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.