Original Article

Journal of Advances in Medical Education & Professionalism

Artificial Intelligence–Based Computer Vision Feedback to Improve Clinical Performance in Surgical Technology Education

KIARASH KAMBOOZIA[1*], MSc Student; SEDIGHEH HANNANI[1], PhD; NAZANIN SARRAF SHAHRI[1,3], PhD; ZEYNAB BARZEGAR[2], PhD; SAEID ZARIF GHASEMIAN[2], PhD Candidate

[1]Department of Operating Room, Faculty of Allied Health Medicine, Iran University of Medical Sciences, Tehran, Iran; [2]Department of Artificial Intelligence in Medicine, Faculty of Advanced Technologies in Medicine, Iran University of Medical Sciences, Tehran, Iran; [3]Center for Educational Research in Medical Science (CERMS), Department of Medical Education, School of Medicine, Iran University of Medical Sciences, Tehran, Iran

Abstract

Introduction: Advancements in artificial intelligence (AI) and computer vision offer new opportunities to improve clinical skills training. This study developed a real-time, AI-based computer vision system for detecting surgical instruments and evaluated its association with clinical performance among surgical technology students.

Methods: A quasi-experimental pre-test/post-test study with a control group was conducted among 48 eligible surgical technology students at Iran University of Medical Sciences during the 2025–2026 academic year. Participants who met the predefined inclusion criteria were enrolled using census sampling. Following ethics approval (IR.IUMS.REC.1403.928), the participants were allocated to an intervention group (n=31), which received immediate AI-generated visual feedback after each instrument arrangement task, and a control group (n=17), which received routine instructor-led training without AI feedback. A YOLOv8-based deep learning model was trained using transfer learning on annotated images of surgical table setups. Clinical performance was assessed using a validated observational checklist (CVI=0.90; Cronbach's α=0.87). Data were analyzed using analysis of covariance (ANCOVA), the Wilcoxon signed-rank test, Spearman's correlation, and the intraclass correlation coefficient (ICC) in SPSS version 27. All 48 participants completed the study (attrition rate=0%).

Results: The AI model achieved excellent object detection performance (precision=0.974, recall=0.979, mAP@0.5=0.981, mAP@0.5:0.95=0.808), with an average inference time of 1.6 seconds per image. Clinical performance was significantly higher following AI-assisted feedback (Wilcoxon signed-rank test, Z=−4.879, p<0.001, r=0.88). After adjustment for baseline performance using ANCOVA, AI-assisted feedback was associated with higher post-intervention performance scores than conventional training (F (1,45)=19.51, p<0.001, partial η^2=0.302). Excellent agreement was observed between AI-generated assessments and expert human ratings (ICC=0.984).

Conclusions: Real-time, AI-based computer vision feedback was associated with improved surgical instrument arrangement performance among surgical technology students while providing an objective and reliable assessment of practical skills. The proposed system represents a promising proof-of-concept educational tool for simulation-based training and competency assessment and may complement conventional instructional approaches in surgical technology education.

Keywords: Artificial intelligence, Surgical instrument, Medical education

**Corresponding author:*
Kiarash Kamboozia, MSc Student;
Department of Operating Room, Faculty of Allied Health Medicine,
Iran University of Medical Sciences, Tehran, Iran
Tel: +98-9129272609
Email: K.Kamboozia@gmail.com

Please cite this paper as:
Kamboozia K, Hannani S, Sarraf Shahri N, Barzegar Z, Zarif Ghasemian S. Artificial Intelligence–Based Computer Vision Feedback to Improve Clinical Performance in Surgical Technology Education. J Adv Med Educ Prof. 2026;14(4):364-377. DOI: 10.30476/jamp.2026.111362.2415.

Received: 15 May 2026
Accepted: 22 August 2026

Introduction

In the rapidly evolving field of healthcare education, the integration of artificial intelligence (AI) and computer vision technologies has emerged as a promising approach to enhancing surgical training (1-3). Computer vision, a subfield of artificial intelligence, enables the automated interpretation of images and videos using deep learning techniques and has been increasingly applied across healthcare for diagnosis, clinical decision support, workflow optimization, and medical education (4-8).

In recent years, computer vision has become an integral component of surgical video analysis. Deep learning–based algorithms have demonstrated excellent performance in surgical phase recognition, workflow analysis, and gesture classification, with reported accuracies frequently exceeding 85–95% in benchmark datasets such as Cholec80 and JIGSAWS, highlighting the feasibility of automated surgical activity recognition. However, these approaches have primarily focused on intraoperative video analysis rather than the assessment of surgical instrument arrangement during preoperative preparation, leaving an important educational gap in surgical technology training (9-13). One prominent application is automated skill assessment, which typically occurs in two forms: objective evaluation based on intraoperative performance (11, 12), and analysis of surgical instrument position and motion (12-15).

The operating room is a specialized clinical environment for surgical technology students. This setting is characterized by high levels of stress and anxiety, a diversity of surgical procedures involving multidisciplinary teams (16) (surgery and anesthesiology), and the simultaneous conduct of emergency and elective cases, all of which pose serious challenges to delivering effective education while preserving patient safety. Students must acquire professional performance skills, critical thinking, and decision-making abilities in this high-pressure environment, prioritizing patient safety while operating under routine and critical conditions. Additional constraints, such as the limited number of clinical instructors, further reduce learning opportunities (17). Consequently, inadequate clinical exposure often leads to fear and anxiety regarding clinical tasks, increased professional errors, and suboptimal performance (18, 19). Therefore, allowing students to learn and rehearse specific performance scenarios before clinical exposure may help them develop appropriate skills and perform effectively in actual operating room settings (7, 20).

One of the most significant challenges in clinical education is the objective evaluation of student performance (21). For instance, a study by Farrokhi in Mashhad found that 62% of students believed their assigned grades did not accurately reflect their true abilities, and 77% requested re-evaluation or reconsideration by instructors (22). Traditional learning of clinical skills relies on an apprenticeship model, in which an assessor observes the individual and provides a qualitative ranking. Overall, this approach may be less effective for objective clinical skill assessment because of limited standardization. The current common methods for evaluating surgical skills include questionnaires, manual scoring systems, and virtual reality devices (23-27). However, these methods can be inconsistent and are susceptible to human bias.

To address this issue, several tools have been developed for the objective assessment of surgical performance (25, 28, 29). One widely recognized instrument is the Objective Structured Assessment of Technical Skill (OSATS), which is considered the gold standard for the structured evaluation of technical competencies. In OSATS, surgeons and instructors observe trainees performing predefined surgical tasks within a limited timeframe and assign global ratings using a standardized checklist (23, 30). Nevertheless, OSATS is time-consuming and exhibits inter-rater variability.

Another approach involves sensor-based systems that precisely track hand and instrument movements, which are important indicators of surgical performance (31). Given the delicate and highly precise nature of surgical procedures, wearable sensors may have important limitations, including potential interference with task performance. As demonstrated by Tee et al., the use of wearable sensors can disrupt an operator's skill performance (32). Furthermore, sterilizing these wearable devices can be challenging, and their relatively high cost may limit their practical and economic feasibility (32, 33).

Other studies have explored automated surgical skill assessment using video- or sensor-based analysis of instrument movements (34, 35). These approaches evaluate standardized simulated surgical performance based on surgeon positioning or instrument kinematics as proxies for skill level. Despite these advances, specialized tools for assessing the performance of surgical technology students remain limited, particularly for the critical tasks of surgical instrument preparation and table setup (36).

Thus, there remains a pressing need for an affordable, time-efficient, objective, and reliable assessment tool. Although significant progress has been made in AI and computer vision, empirical evidence regarding the impact of these algorithms on the clinical performance of surgical technology students remains limited. In particular, it remains unclear whether intelligent algorithm-based assessments can serve as a valid and reliable substitute for human evaluation in clinical educational settings.

This gap highlights the need for innovative, technology-enhanced interventions to optimize surgical education. The present study introduces a real-time, AI-based computer vision system specifically designed to recognize surgical instruments, with the aim of supporting learning and improving clinical performance in simulated surgical setups. Despite the growing use of AI in surgical video analysis, empirical evidence regarding the educational effectiveness of AI-driven computer vision systems as formative feedback tools for clinical skills training remains limited, particularly in undergraduate surgical technology education.

Methods

This study employed a quasi-experimental pre-test/post-test design with a nonrandomized control group. Due to logistical constraints within the educational setting and the structure of practical classes, random allocation of participants was not feasible; instead, participants were assigned to the intervention and control groups according to pre-existing practical classes. This quasi-experimental study was reported in accordance with the TREND (Transparent Reporting of Evaluations with Nonrandomized Designs) statement to enhance transparency and reproducibility.

Participants and Setting

The target population comprised all operating room technology students (n=48) enrolled in semesters 5 and 7 at Iran University of Medical Sciences during the 2025–2026 academic year. Because the number of eligible students enrolled during the study period was limited, all students meeting the inclusion criteria were invited to participate. Therefore, a census sampling approach was used, and no priori sample size calculation was performed. The study was conducted in the clinical skills training laboratory of the School of Allied Medical Sciences at the same institution. Convenience sampling with nonrandom assignment resulted in an intervention group (n=31) receiving AI-based feedback and a control group (n=17) receiving conventional training.

Inclusion criteria were completion of at least one operating room internship semester; theoretical and practical familiarity with operating room principles, equipment, and medical terminology; and provision of written informed consent. All participants had completed the same standardized operating room internship curriculum prescribed by the university. The duration and core educational content of the internship were identical for all students, ensuring comparable prior clinical training. Exclusion criteria included failure to participate in any stage of the study (initial setup, feedback session, or repeat setup) or voluntary withdrawal at any point.

Ethical Considerations

The study protocol was approved by the Research Ethics Committee of Iran University of Medical Sciences (approval code: IR.IUMS.REC.1403.928). Participants received a detailed explanation of the study objectives, procedures, and voluntary nature of participation during an orientation session. Written informed consent was obtained from each student before participation. Data confidentiality was ensured by using unique identification codes (S01–S48), and all personal information was anonymized.

Development of the AI-Based Computer Vision System and Dataset Construction

Following ethical approval and institutional research council endorsement, a deep learning-based computer vision algorithm was developed using the YOLOv8 architecture to enable intelligent evaluation of students' surgical table setups. The system detects surgical instruments in images, analyzes their spatial arrangement, compares it against a predefined standard layout for abdominal surgery, and generates quantitative performance scores and educational feedback.

A custom dataset of 140 high-resolution digital images of simulated surgical table setups was created under faculty supervision. Images were acquired using a fixed digital camera positioned directly above the operating table at a height of 100 cm, with an approximately 90° overhead angle relative to the table surface. Image acquisition was performed under standardized indoor laboratory lighting conditions to minimize shadows and illumination variability. The same camera setup and environmental conditions were maintained throughout data collection. The model was implemented in Python using the Ultralytics YOLOv8 framework with PyTorch version 2.6.0.

Images included both correct and incorrect arrangements and were captured in a controlled educational environment. Annotations (class labels and normalized bounding box coordinates) followed the standard YOLO format. The dataset was randomly split into training (80%, n=112), validation (10%, n=14), and test (10%, n=14) sets. Five-fold cross-validation was applied to the training set to minimize bias and enhance generalizability. To reduce the risk of overfitting associated with the relatively limited dataset, we employed transfer learning, data augmentation, and five-fold cross-validation during model development. Nevertheless, external validation on an independent dataset was not performed which is acknowledged as a limitation.

The YOLOv8x model, pretrained on a large-scale dataset, was fine-tuned using transfer learning to optimize performance with limited training data and accelerate model convergence. Figure 1 illustrates representative qualitative detection results produced by the proposed YOLOv8-based computer vision model.

Post-training evaluation on the independent test set included computing Precision, Recall, and mean Average Precision (mAP@0.5 and mAP@0.5:0.95), as well as confusion matrices, to assess detection accuracy and error patterns. The final system integrated a scoring module evaluating two primary criteria: correct instrument selection and accurate spatial placement, enabling both quantitative output for statistical analysis and real-time visual/verbal feedback to learners.

Instruments and Data Collection Tools

Performance was assessed using a researcher-developed observational checklist comprising two sections: Section 1 (observational performance evaluation) and Section 2 (instrument checklist).

Content validity was established through a review by 10 faculty experts in surgical technology, guided by the Association of Perioperative Registered Nurses (AORN) standards and the WHO Safe Surgery Checklists. The Content Validity Ratio (CVR) exceeded Lawshe's critical values for all items, and the overall Content Validity Index (CVI) was 0.91. Reliability was assessed via test–retest (two administrations one week apart on 10 volunteers) and inter-rater agreement using the Intraclass Correlation Coefficient (ICC=0.85), indicating acceptable stability and consistency.

The AI algorithm served dual roles: as the educational intervention for the feedback group and as an objective assessment tool for both groups. Following ethics approval and informed consent, baseline demographic data (gender, semester, prior scrub experience) were collected via a brief questionnaire.

Pre-test

Participants independently set up a simulated general abdominal surgery table (e.g., laparotomy) under sterile conditions without AI access. Performance was assessed immediately after task completion using the validated observational checklist. High-quality digital images of each setup were captured for AI input. Following completion of the baseline assessment, both groups participated in the same number of educational sessions and received an equivalent duration of practical training. The intervention group received AI-generated visual and textual feedback based on their pre-test performance, whereas the control group received conventional instructor feedback covering the same educational content without access to the AI-based computer vision system.

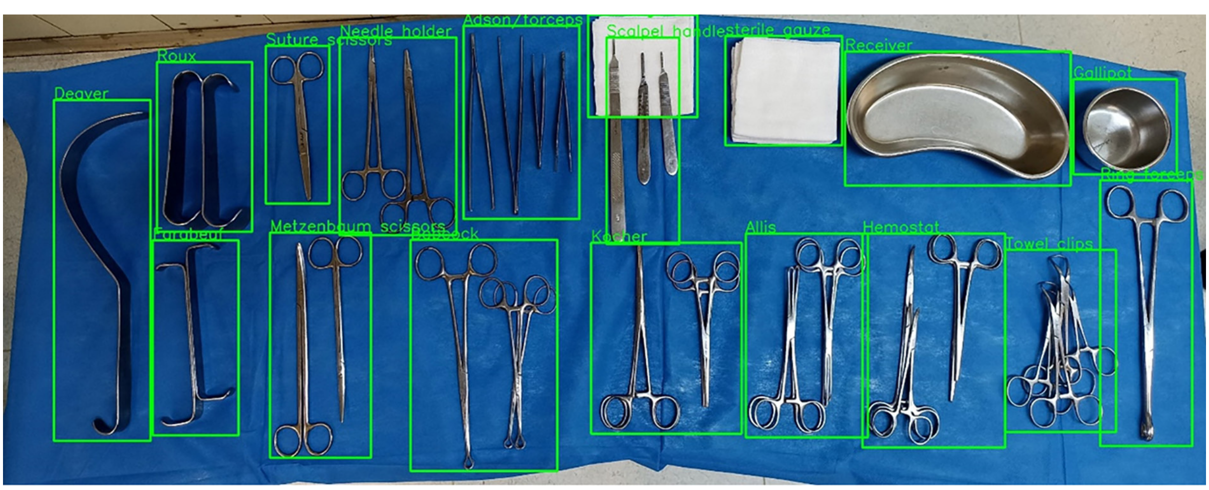

Figure 1. Qualitative results of surgical instrument detection using the proposed YOLO model (along with confidence scores)

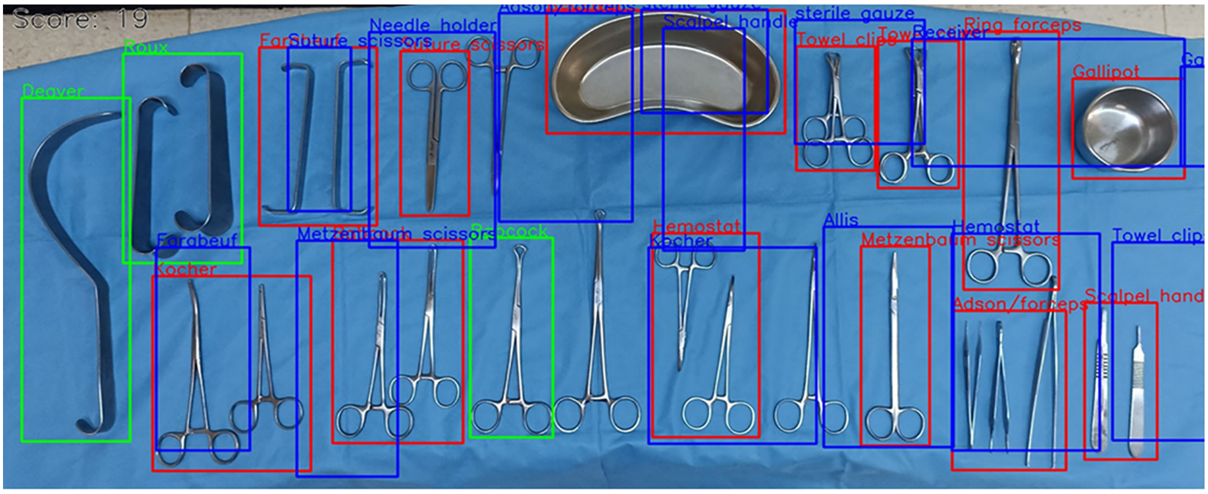

Figure 2. Example of YOLOv8 detection output on a surgical table setup image. Bounding boxes were color-coded as follows: green for the correct instrument and position, red for the correct instrument but incorrect position, and blue for the correct designated position. The total performance score was 19 (based on the presence and placement accuracy of essential instruments).

Intervention

Following the baseline assessment, participants assigned to the intervention group underwent an AI-assisted educational program designed to improve surgical instrument arrangement skills. The intervention consisted of automated performance assessment, immediate AI-generated feedback, a structured educational session, and repeated hands-on practice. Immediately after each instrument arrangement task, a digital image of the operating table was captured using a fixed overhead camera. The image was automatically processed by the trained YOLOv8-based computer vision model, which detected each surgical instrument and compared its position with a predefined reference arrangement. The reference arrangement was based on the standard surgical instrument layout approved by experienced surgical technology faculty members.

Within a few seconds, the software displayed the analyzed image with bounding boxes around all detected instruments, together with the predicted instrument labels and corresponding confidence scores (Figure 2). The system automatically classified performance by identifying correctly positioned, misplaced, omitted, and unnecessary instruments. Incorrectly arranged instruments were visually highlighted, enabling students to immediately recognize their errors and compare their performance with the standard reference arrangement.

Following the automated assessment, participants attended one structured face-to-face educational session delivered by the principal investigator. The educational intervention consisted of a single 90-minute session. During the session, the principles of standard surgical instrument arrangement, common placement errors, and interpretation of AI-generated feedback were explained using representative examples generated by the computer vision system. Students were encouraged to discuss their errors and appropriate corrective strategies.

After completion of the educational session, each participant independently repeated the surgical instrument arrangement task twice under identical simulation conditions. Immediately after each practice attempt, the AI system generated individualized visual feedback using the same automated assessment procedure. The same finalized pretrained YOLOv8 model was used to evaluate both the pre-test and post-test images. No recalibration, retraining, fine-tuning, or modification of the model was performed during the study, ensuring identical evaluation conditions for all assessments. This iterative feedback process enabled students to identify, understand, and correct their errors before the subsequent practice attempt. Throughout the intervention, the researcher supervised the training sessions, answered the participants' questions, and ensured adherence to the educational protocol without modifying the AI-generated assessment or feedback.

Participants allocated to the control group received the routine educational program provided during the surgical technology internship, consisting of conventional instructor-led demonstrations and standard practical training without access to the AI-based computer vision feedback system. Following completion of the intervention period, the participants in both groups underwent the post-intervention assessment using the same standardized clinical performance checklist, assessment procedure, camera position, operating table setup, and evaluator as those used during the baseline assessment.

A single experienced evaluator performed human assessments using a predefined standardized checklist. The evaluator was not blinded to group allocation or assessment time point.

Post-test

Both groups repeated the table setup task. The intervention group repeated the task after receiving AI-generated feedback, whereas the control group followed routine instructor-led training without AI feedback. Final performance was scored using the same checklist, and new images were obtained. AI evaluation was applied to all post-test images (for the control group solely as an assessor, without feedback). Data were analyzed using SPSS version 27. Before statistical analyses, the assumptions underlying each statistical test were evaluated. The normality of continuous variables was assessed using the Shapiro–Wilk test. Because some variables did not follow a normal distribution, the Wilcoxon signed-rank test and Spearman's rank correlation coefficient were used where appropriate. Before performing ANCOVA, the assumptions of homogeneity of variances (Levene's test) were evaluated and confirmed. Inter-rater agreement between the AI system and the human evaluator was assessed using the intraclass correlation coefficient (ICC). All analyses were two-tailed, with statistical significance set at $p<0.05$.

Results

Model Performance Metrics

To assess the performance of the YOLOv8-based computer vision algorithm, the model's outputs were evaluated on the test dataset. The evaluation results, based on standard metrics including Precision, Recall, and mean Average Precision (mAP@0.5 and mAP@0.5:0.95), are presented in Table 1. Overall, the model achieved a global Precision of 0.974 and a Recall of 0.979. The mean Average Precision at an IoU threshold of 0.5 (mAP@0.5) was 0.981, while the more comprehensive mAP@0.5:0.95 reached 0.808.

As shown in Table 1, the YOLOv8-based model demonstrated excellent object detection performance, with Precision, Recall, and mAP@0.5 values exceeding 0.97. The model also achieved an mAP@0.5:0.95 of 0.808, indicating robust detection performance across different IoU thresholds.

Processing Time: The average inference time of the algorithm was reported as 1.6 seconds per image.

An example of the AI-generated detection output is presented in Figure 2, illustrating the classification of correctly positioned, misplaced, and missing surgical instruments, together with the corresponding performance score.

Figure 3 presents the confusion matrix of the proposed computer vision model, illustrating the distribution of correct classifications and misclassification patterns across all instrument classes.

The confusion matrix results indicate that the designed computer vision algorithm, in addition to achieving high accuracy, exhibited a controlled and acceptable error pattern. Errors were concentrated in a limited number of classes with no evidence of systematic bias, reflecting adequate network training and high-quality training data.

Table 1. Evaluation results of the surgical instrument detection model based on precision and recall metrics

Class of Instrument	Image	Samples	Precision	Recall	mAP@0.5	mAP@0.5:0.95
All Images	30	508	0.974	0.979	0.981	0.808
Adson	30	30	1.000	1.000	0.995	0.878
Allis	29	31	0.996	0.903	0.936	0.792
Babcock	27	27	1.000	1.000	0.995	0.809
Deaver	28	28	0.996	0.966	0.971	0.796
Farabeuf	27	27	1.000	1.000	0.995	0.803
Gallipot	29	29	1.000	1.000	0.995	0.779
Hemostat	28	30	0.967	0.967	0.981	0.758
Kocher	28	31	0.829	0.935	0.933	0.774
Metzenbaum Scissors	29	29	0.967	1.000	0.991	0.806
Needle Holder	29	30	1.000	0.967	0.983	0.850
Receiver	30	30	1.000	1.000	0.995	0.879
Ring Forceps	30	30	1.000	1.000	0.995	0.787
Roux Retractor	9	10	0.900	0.900	0.944	0.767
needle Handle	29	29	1.000	1.000	0.995	0.785
Towel clips	30	30	1.000	1.000	0.995	0.794
Suture Scissors	28	28	0.996	1.000	0.989	0.811
Sterile Gauze	29	58	1.000	1.000	0.995	0.864

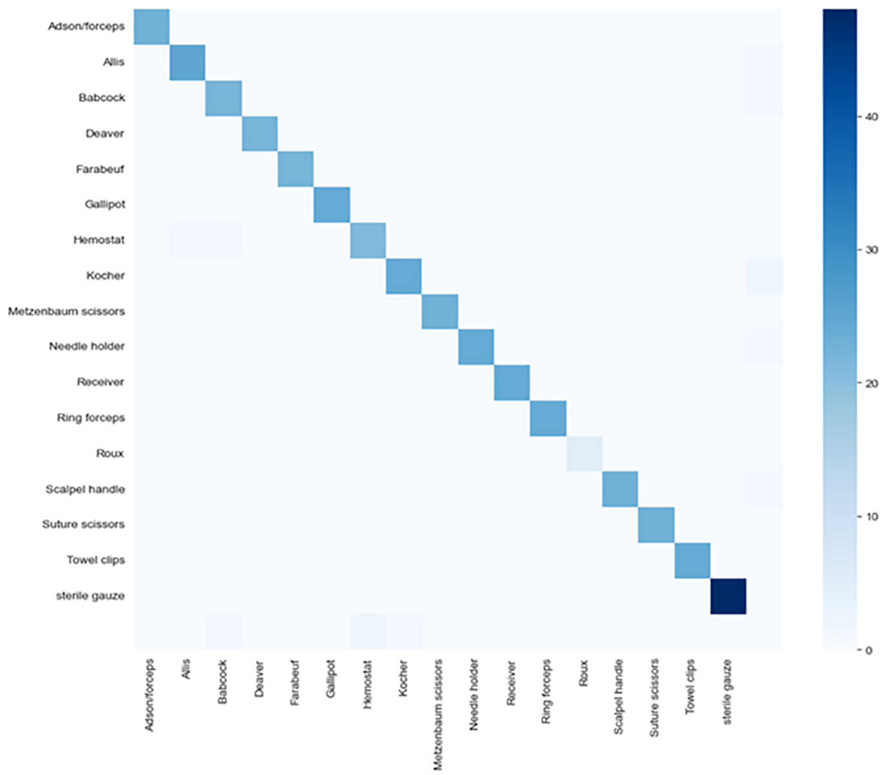

Figure 3. Confusion matrix of the deep learning–based computer vision model for surgical instrument recognition

Table 2. Frequency distribution of study participants by study group, academic semester, gender, and prior experience in surgical operations. Results are presented as numbers and percentages.

Variable	Category	Number (n)	Percentage (%)
Study Group	Intervention (received AI feedback)	31	64.6
	Control	17	35.4
Academic Semester	Fifth semester	19	39.6
	Seventh semester	29	60.4
Gender	Male	21	43.8
	Female	27	56.2
Prior Experience in Operating Room Procedures	No	21	43.8
	Yes	27	56.2

Analysis of the Model Training Trend and Performance Metrics

The following graphs (Figure 4) illustrate the trends in the loss function components, namely Box Loss, Class Loss, and DFL Loss, on both the training and validation sets. The results indicate that all three loss components exhibited a consistent downward and converging trend across the training and validation datasets throughout the training process. Furthermore, the trends of the performance metrics Precision, Recall, mAP@0.5, and mAP@0.5:0.95 are also shown. A gradual and steady increase in these metrics was observed across the training epochs.

Participant Flow and Baseline Characteristics

Of the 48 eligible surgical technology students (semesters 5 and 7), all consented and were allocated non-randomly to the intervention (n=31) or control group (n=17) based on class scheduling constraints. No participant was excluded from the study. All 48 participants completed the pre-test, intervention/control phase, and post-test without dropout.

The baseline demographic characteristics are displayed in Table 2. No statistically significant differences were observed between the groups at baseline (all $p>0.05$).

Effect of AI-Assisted Feedback on Clinical Performance

Assessment of Normality

The normality of continuous variables was assessed using the Shapiro–Wilk test, which was considered the primary normality test considering the study sample size (n=48), as shown in Table 3.

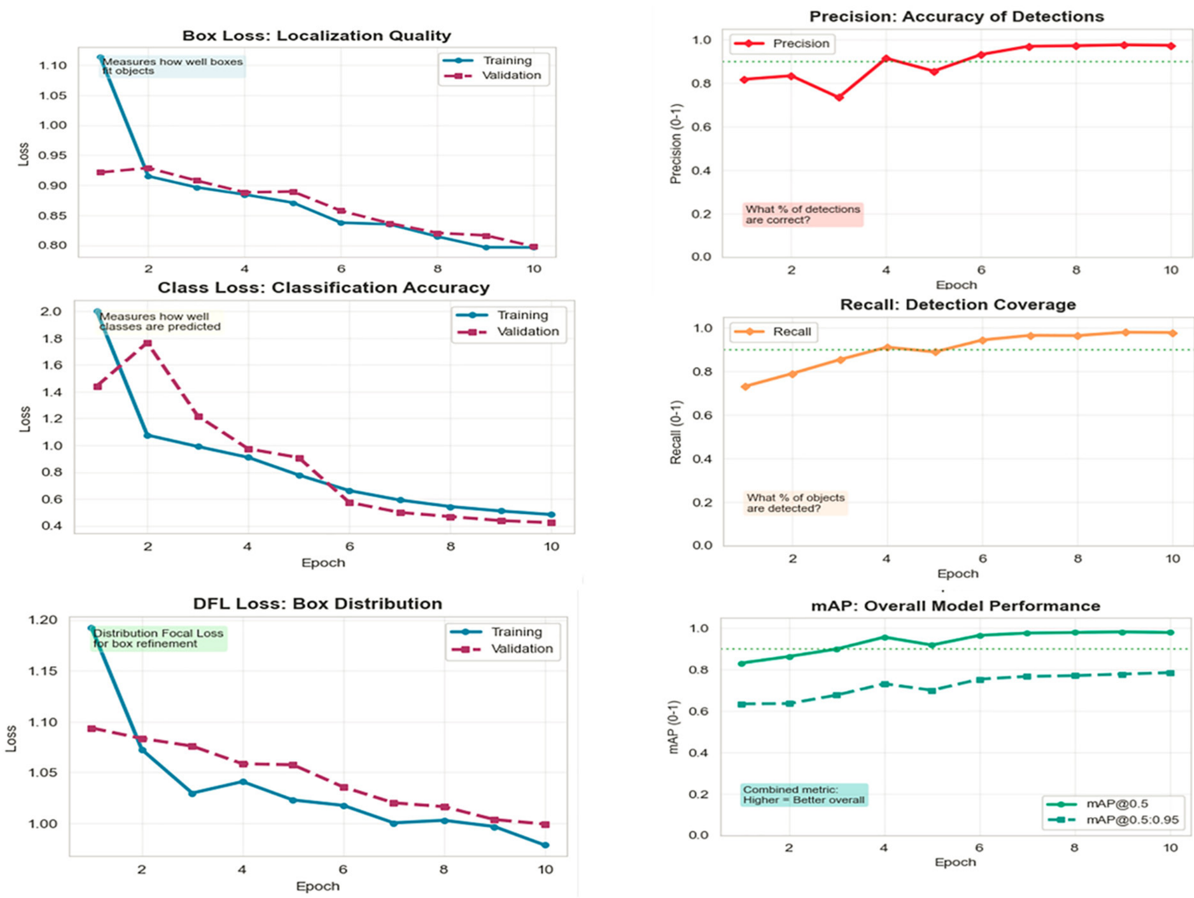

Figure 4. Convergence trends of the YOLOv8 model's training and validation losses alongside the progressive changes in key performance metrics throughout the training process

Table 3. Results of the Shapiro–Wilk Normality Test for Study Variables

Variable	Shapiro–Wilk Statistic	df	p	Interpretation
Human evaluator score	0.950	48	0.009	Non-normal
AI score (Pre-test)	0.949	48	0.005	Non-normal
AI score (Post-test)	0.954	48	0.059	Normal
Clinical performance checklist	0.987	48	0.200	Normal
Overall performance score	0.985	48	0.200	Normal

The results indicated that the distributions of the human evaluator score and the baseline AI assessment score deviated significantly from normality ($p<0.05$). In contrast, the post-intervention AI assessment score, clinical performance checklist score, and overall performance score did not significantly deviate from a normal distribution ($p>0.05$). Accordingly, the distribution of each variable was considered when selecting the appropriate statistical analyses. The primary study outcome was subsequently analyzed using analysis of covariance (ANCOVA), with baseline performance included as a covariate.

The algorithm-derived performance scores for the students in the intervention group before receiving AI feedback averaged 17.23±2.91, indicating a moderate level of performance in surgical tray arrangement, as assessed by the model. Following the provision of AI-generated feedback, the mean score increased to 22.35±2.35.

Within the intervention group, AI-assisted feedback was associated with a significant increase in algorithm-derived performance scores (Table 4), from 17.23±2.91 before the intervention to 22.35±2.35 after the intervention (Wilcoxon signed-rank test: $Z=-4.879$, $p<0.001$). Positive ranks were observed for all participants, with no negative ranks or ties, indicating that algorithm-derived performance scores increased for all students following the intervention. The calculated effect size was very large ($r=0.88$), suggesting a substantial educational impact of the intervention. The primary between-group analysis was performed using ANCOVA, with post-intervention performance as the dependent variable and baseline performance as the covariate.

Table 4. Within-group comparison of algorithm-derived performance scores before and after AI-assisted feedback

Variable	n	Mean±SD	Med (Q1–Q3)	Min– Max	Z	p
Algorithm performance score (pre-AI feedback)	31	17.23±2.91	17.00 (15–19)	13–25	—	—
Algorithm performance score (post-AI feedback)	31	22.35±2.35	23.00 (21–25)	19–28	Z = −4.879	p < 0.001

Table 5. ANCOVA Results for Post-test Performance Adjusted for Baseline Performance

Source	F	df	p	Partial η²
Baseline performance	142.43	1,45	<0.001	0.760
Study group	19.51	1,45	<0.001	0.302

Effect of AI-Assisted Feedback on Clinical Performance

Before conducting the ANCOVA, we assessed the assumptions of homogeneity of variances and homogeneity of regression slopes. The homogeneity of variances was supported by Levene's test ($F(1,46)=2.00$, $p=0.164$). The homogeneity of regression slopes was evaluated by testing the interaction between the study group and baseline performance, which was not statistically significant ($F(1,44)=3.677$, $p=0.062$); this indicates that this assumption was met.

An analysis of covariance (ANCOVA) was performed with the post-intervention algorithm-derived performance score as the dependent variable, the study group as the fixed factor, and the baseline algorithm-derived performance score as the covariate (Table 5). After adjustment for baseline performance, the study group was significantly associated with post-intervention performance ($F(1,45)=19.51$, $p<0.001$, partial $\eta^2=0.302$), with students receiving AI-assisted feedback showing higher post-intervention performance scores than those receiving conventional instruction (Table 5). Baseline performance was also significantly associated with post-intervention performance ($F(1,45)=142.43$, $p<0.001$, partial $\eta^2=0.760$). The adjusted mean post-intervention performance score was 23.00 (95% CI: 22.46–23.53) in the intervention group and 21.01 (95% CI: 20.28–21.73) in the control group (Figure 5).

Box plot showing the distribution of post-intervention performance scores in the intervention and control groups. The horizontal line within each box represents the median, the box indicates the interquartile range (IQR), and the whiskers represent the minimum and maximum values.

Agreement between AI and Human Assessments

Spearman's rank correlation demonstrated a strong positive association between AI-derived scores and expert evaluator scores ($r_s=0.957$, $p<0.001$). Furthermore, the intraclass correlation coefficient (ICC) demonstrated excellent agreement between the two assessment methods (ICC=0.984, 95% CI: 0.969–0.992, $p<0.001$), confirming the high consistency of the proposed AI-based assessment system with expert human evaluation (37). Spearman's correlation analysis revealed a strong, positive, and statistically significant relationship between the scores generated by the computer vision algorithm and those recorded by the human evaluator ($r_s=0.957$, $p<0.001$).

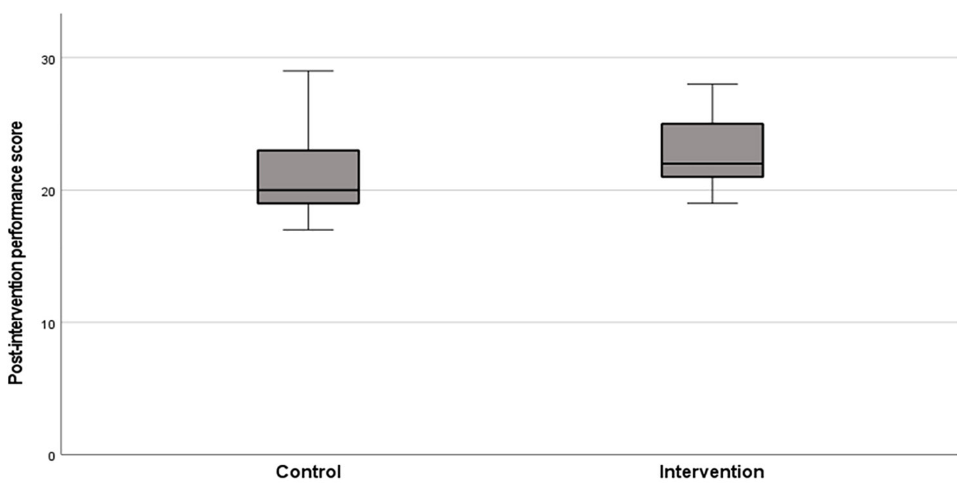

Figure 5. Box plot of post-intervention clinical performance scores in the intervention and control group

Table 6. Reliability indices and agreement between the two assessment methods

Measure	Value	95% CI	p
Spearman correlation	0.957	—	<0.001
ICC	0.984	0.969–0.992	<0.001

Additionally, reliability and inter-rater agreement were assessed using the intraclass correlation coefficient (ICC). The reliability indices and agreement between AI-generated and expert assessments are presented in Table 6.

The intraclass correlation coefficient (ICC) for average measures reached 0.984 (95% CI: 0.969–0.992, $p<0.001$), which falls within the excellent agreement category according to the standard interpretive guidelines ($ICC \geq 0.90$=excellent agreement). The excellent agreement between the AI algorithm and the human evaluator, as indicated by the high ICC, supports the reliability of the proposed computer vision system for surgical instrument assessment. The strong Spearman correlation further indicates consistency in ranking the participants between the two assessment methods.

Discussion

This quasi-experimental study suggests that AI-assisted computer vision feedback is associated with improved surgical instrument arrangement performance among students in the intervention group. After adjustment for baseline performance using analysis of covariance (ANCOVA), students receiving AI-generated feedback achieved significantly higher post-intervention performance scores than those receiving conventional training, with a moderate-to-large effect size (partial $\eta^2=0.302$). Furthermore, the AI-based assessment system demonstrated excellent agreement with expert human ratings (ICC=0.984), supporting its reliability as an objective educational assessment tool. This level of agreement suggests that the proposed system could support instructors by providing consistent formative assessment while reducing evaluator variability (38-40).

The YOLOv8 model demonstrated excellent object detection performance, achieving high precision (0.974), recall (0.979), and mAP@0.5 (0.981), together with a rapid inference time of approximately 1.6 seconds per image. Although this processing speed may be lower than that of fully real-time systems designed for intraoperative applications or continuous surgical monitoring, it is well-suited to simulation-based educational environments (41). In this context, the primary objective is to provide learners with accurate, reliable, and timely formative feedback rather than continuous high-frame-rate video processing. Consequently, the inference time obtained represents an appropriate balance between detection accuracy and response time while supporting effective learning through prompt feedback. These findings indicate that the proposed AI-based computer vision system is suitable for providing timely and objective feedback during simulation-based surgical skills training. These findings also support previous evidence, indicating that simulation-based education provides an effective environment for acquiring procedural skills before clinical practice (38). Our findings are consistent with previous studies reporting the educational benefits of technology-enhanced feedback, including video-based and motion-tracking systems used in minimally invasive and robotic surgery (39-46). Our findings are also consistent with previous educational technology studies demonstrating that digital learning platforms and interactive educational software can improve learner motivation and educational outcomes in medical education (47). However, unlike previous approaches, the present study focused on surgical instrument arrangement, an essential but relatively underexplored component of perioperative training that may contribute to operating room efficiency, maintenance of sterility, and reduction of procedural errors. The consistency between our findings and previous studies may be explained by the ability of AI-based feedback to provide immediate, objective, and standardized performance evaluation, thereby facilitating more effective learning than conventional subjective feedback alone.

The strengths of this study include the use of validated assessment checklists developed in accordance with AORN and WHO recommendations, the dual use of the AI system as an educational intervention and an objective assessment tool, and the excellent agreement between AI-generated and expert evaluator scores. Collectively, these findings suggest that AI-assisted assessment may help standardize performance evaluation and complement conventional instructor-based feedback in surgical technology education.

The observed improvement may be explained by the educational advantages of immediate, objective, and individualized feedback. Unlike conventional instructor-based feedback, which may vary depending on instructor availability and individual judgment, the proposed AI system provided consistent and immediate feedback after each practice attempt. Such timely feedback enables learners to recognize and correct their mistakes during repeated practice, facilitating deliberate practice and reinforcing skill acquisition. This mechanism may explain the substantial improvement observed in the intervention group.

Overall, these findings indicate that AI-based computer vision feedback may improve clinical performance and represent a promising educational tool for competency assessment in surgical technology education. Similarly, previous research among operating room students has shown that technology-assisted educational interventions can improve perceived competence and the quality of surgical training (48). The proposed system demonstrated high agreement with expert human evaluation and may serve as a valuable educational tool to support technology-enhanced learning and competency assessment in surgical technology education. Further multicenter studies with larger sample sizes are warranted to confirm the generalizability of these findings.

Despite the promising findings, several implementation challenges should be considered before integrating AI-based computer vision systems into routine surgical technology education. Successful implementation requires adequate hardware resources, standardized image acquisition conditions, and technical infrastructure to ensure consistent model performance. Faculty members may also require training to effectively interpret AI-generated feedback and integrate it into existing educational workflows. Furthermore, developing and maintaining high-quality annotated datasets for different surgical procedures may require substantial time and institutional resources. Finally, although the proposed system demonstrated high performance under standardized simulation conditions, further validation in diverse educational settings and across a broader range of surgical procedures is necessary before widespread implementation.

Limitations

This study has several limitations that should be considered when interpreting the findings. First, the quasi-experimental design with non-random group allocation may have introduced selection bias although baseline differences in performance were statistically controlled using ANCOVA. Second, the study was conducted at a single academic institution with a relatively small sample of surgical technology students. Moreover, a priori sample size calculation was not performed because all eligible students were included using a census sampling approach. These factors may limit the generalizability of the findings to other educational settings. Third, participants were aware that their performance was being observed, which may have introduced a Hawthorne effect and influenced their performance during the simulation sessions. In addition, the outcome assessor was not blinded to group allocation or assessment time point, potentially introducing detection bias.

Nevertheless, assessments were conducted using a predefined standardized checklist to minimize subjective variation. Furthermore, the AI model was developed and evaluated using a relatively limited image dataset acquired under standardized and controlled laboratory conditions, and external validation using independent datasets from other institutions was not performed. Therefore, the current system should be considered a proof-of-concept prototype rather than a fully validated system for broad clinical or educational deployment. Although the model demonstrated high performance on the available test data, its robustness and generalizability under more diverse conditions—including variations in imaging devices, lighting, instrument configurations, and educational or clinical environments—remain to be established. Future studies should validate the system using larger, more diverse, and independently collected datasets across multiple institutions before broader implementation.

The proposed AI-based feedback system was also developed and evaluated exclusively for abdominal surgical instrument arrangement; therefore, its applicability to other surgical specialties requires further investigation. Additionally, the educational intervention and performance assessment were conducted in a simulated educational environment rather than during actual surgical procedures. Although the study evaluated clinical performance and self-efficacy, it did not examine other important educational outcomes, such as students' confidence, anxiety, or acceptance of AI-assisted learning; as a result, the findings may not fully reflect the students' performance in real operating room settings. Finally, the cost-effectiveness and practical feasibility of implementing AI-based computer vision systems in routine educational settings were not assessed and should be investigated in future studies. No long-term follow-up was conducted to determine whether the observed improvements in surgical instrument arrangement performance were sustained over time.

Conclusion

This study demonstrated that a real-time AI-based computer vision system using YOLOv8 could effectively support surgical technology education by providing objective, immediate, and automated feedback on the arrangement of surgical instruments. Students who received AI-assisted feedback achieved significantly higher post-intervention performance than those receiving conventional training, even after adjustment for baseline performance. In addition, the proposed system showed excellent agreement with expert human evaluation, supporting its potential as a reliable assessment tool for simulation-based learning. These findings suggest that AI-assisted computer vision can complement conventional educational approaches by enhancing both formative assessment and practical skill development in surgical technology training. Future multicenter studies using larger and more diverse datasets and populations are recommended to evaluate the long-term educational impact, external validity, and broader applicability of this approach across different surgical specialties.

Acknowledgment

The authors acknowledge the institutional support provided by Iran University of Medical Sciences for conducting this study. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' Contribution

K.K contributed to the study conception and design. Material preparation, data collection, and analysis were performed by K.K and S.ZGh. The first draft of the manuscript was written by K.K and all authors commented on previous versions of the manuscript. S.H and N.SSh supervised the project. All authors read and approved the final manuscript.

Conflicts of Interest

The authors declare that there is no conflict of interest.

Declaration of AI

During the preparation of this manuscript, the authors used Grammarly for language editing, grammar improvement, and enhancement of clarity and readability. Grammarly was not used to generate scientific content, analyze data, or draw conclusions. The authors reviewed and edited the output as appropriate and take full responsibility for the content of the manuscript.

References

1. Thacharodi A, Singh P, Meenatchi R, Ahmed ZHT, Kumar RSR, Neha V, et al. Revolutionizing healthcare and medicine: The impact of modern technologies for a healthier future-A comprehensive review. Health Care Sci. 2024;3(5):329-49.
2. Jafari F, Keykha A, Taheriankalati A, Taghavi Monfared A. The Role of AI in Shaping Medical Education: Insights from an Umbrella Review of Review Studies. J Adv Med Educ Prof. 2025;13(4):270-93.
3. Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. Artificial Intelligence in Healthcare. 2020;26:25-60.
4. Bhattacharya S, Reddy Maddikunta KP, Pham QV, Reddy Gadekallu T, Krishnan SRS, Chowdhary CL, et al. Deep learning and medical image processing for coronavirus (COVID-19) pandemic: A survey. Sustainable Cities and Society. 2021;65:102589.
5. Ward TM, Mascagni P, Ban Y, Rosman G, Padoy N, Meireles O, et al. Computer vision in surgery. Surgery. 2021;169(5):1253-6.
6. Kabir MM, Rahman A, Hasan MdN, Mridha MF. Computer vision algorithms in healthcare: Recent advancements and future challenges. Computers in Biology and Medicine. 2025;185:109531.
7. Li J, Guan Z, Wang J, Cheung CY, Zheng Y, Lim LL, et al. Integrated image-based deep learning and language models for primary diabetes care. Nat Med. 2024;30:2886–96.
8. Castillo-Segura P, Panadero CF, Alario-Hoyos C, Muñoz-Merino PJ, Kloos CD. Objective and automated assessment of surgical technical skills with IoT systems: A systematic literature review. Artificial Intelligence in Medicine. 2021;112:102007.
9. Kostiuchik G, Lalith Sh, Benedikt M, Ivo W, Bernhard P, Sandy E. Surgical phase and instrument recognition: how to identify appropriate dataset splits. International Journal of Computer Assisted Radiology and Surgery. 2024;19:699-711.
10. Graëff C, Padoy N, Liverneaux P, Lampert T. Introducing surgical workflow recognition in orthopaedic surgery with timestamp supervision. Comput Biol Med. 2025;197(Pt A):110995.
11. Ramesh S, Dall'Alba D, Gonzalez C, Yu T, Mascagni P, Mutter D, et al. Weakly Supervised Temporal Convolutional Networks for Fine-Grained Surgical Activity Recognition. IEEE Transactions on Medical Imaging. 2023;10:1109.
12. Yunlong L, Zhao Z, Li R, Li F. Deep learning for surgical workflow analysis: a survey of progresses, limitations, and trends. Artificial Intelligence Review. 2024;57:10.
13. Oza RS, Mehta MA, Kotecha K, Lin JCW. Analytics of deep model-based spatiotemporal and spatial feature learning methods for surgical action classification. Multimedia Tools and Applications. 2024;83(17):52275-303.
14. van Amsterdam B, Funke I, Edwards E, Speidel S, Collins J, Sridhar A, et al. Gesture Recognition in Robotic Surgery With Multimodal Attention. IEEE Trans Med Imaging. 2022;41(7):1677-87.

15. Wang T, Jin M, Li M. Towards accurate and interpretable surgical skill assessment: a video-based method for skill score prediction and guiding feedback generation. International Journal of Computer Assisted Radiology and Surgery. 2021;16(9):1595-1605.
16. Amirafzali Nasab M, Nouhi E. The impact of simulated operating rooms in a clinical skill center on operating room students' skills and satisfaction: A semi experimental study. J Med Edu Dev. 2024;17(55):31-39.
17. Hannani S, Sadati L, Sarraf shahri N, Arab Khazaei A, Arab Khazaei A, Kamboozia K, et al . Investigating the Effect of the Task-Based Educational Method on Operating Room Technologist Students Speed of Action in Iran University of Medical Sciences. Paramedical Sciences and Military Health. 2022;17(2):37-44.
18. Seshabela HD, Shakwane S. Impact of anxiety-related conditions on nursing students' academic excellence. Curationis. 2024;47(1):e1-e10.
19. Sadati L, Askarkhah A, Hannani S, Moazamfard M, Abedinzade M, Mohammad Alinejad P, et al. Assessment of Staff Performance in CSSD Unit by 360 Degree Evaluation Method. Asia Pacific Journal of Health Management. 2020;15:71-7.
20. Elendu C, Amaechi DC, Okatta AU, Amaechi EC, Elendu TC, Ezeh CP, et al. The impact of simulation-based training in medical education: A review. Medicine (Baltimore). 2024;103(27):e38813.
21. Aziz S, Hameed M, Azeez S, Omer Sh, Ameen K, Abdullah T, et al. Challenges in student assessment in medical and health sciences education in Northern Iraq. BMC Med Educ. 2025;25(1):1052.
22. Khadivzadeh T, Farrokhi F. A study of the strengths and weaknesses of clinical education from the perspective of students of day and evening courses at the Faculty of Nursing and Midwifery of Mashhad in 2003. Iranian Journal of Education in Medical Sciences. 2003;3(Special Issue 10):67-68.
23. Berg M, Dahlin LB, Kjellman M. Overview of surgical training and assessment of surgical skills: a narrative review. Front Surg. 2025;12:1605495.
24. Guerrero JG, Alqarni AS, Estadilla LT, Benjamin LS, Rani VI. Raters and examinees training for objective structured clinical examination: comparing the effectiveness of three instructional methodologies. BMC Nurs. 2024;23(1):500.
25. Hussein AA, Dibaj Sh, Hinata N, Field E, O'leary d K, Kuvshinoff B, et al. Development and Validation of a Quality Assurance Score for Robot-assisted Radical Cystectomy: A 10-year Analysis. Urology. 2016;97:124-9.
26. Morineau T, Riffaud L, Morandi X, Villain J, Jannin P. Work domain constraints for modelling surgical performance. Int J Comput Assist Radiol Surg. 2015;10(10):1589-97.
27. Zia A, Essa I. Automated surgical skill assessment in RMIS training. Int J Comput Assist Radiol Surg. 2018;13(5):731-9.
28. Goh AC, Goldfarb DW, Sander JC, Miles BJ, Dunkin BJ. Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. J Urol. 2012;187(1):247-52.
29. Khan H, Kozlowski JD, Hussein AA, Sharif M, Ahmed Y, May P, et al. Use of Robotic Anastomosis Competency Evaluation (RACE) for assessment of surgical competency during urethrovesical anastomosis. Canadian Urological Association Journal. 2018;13(1):E10-E16.
30. O'Regan DJ, Caddy C, Rajandram R, Barach P. A Scoping Review of the Objective Structured Assessment of Technical Surgical Skills (OSATS). J Surg Educ. 2026;83(6):103946.
31. Stan C, Ujvary PL, Blebea C, Tănase MI, Tănase M, Pop SS, et al. Hand Motion Analysis Using Accelerometer-Based Sensors and Sheep's Head Model for Basic Training in Functional Endoscopic Sinus Surgery. Cureus. 2024;16(5):e59725.
32. Roos LG, Slavich GM. Wearable technologies for health research: Opportunities, limitations, and practical and conceptual considerations. Brain Behav Immun. 2023;113:444-52.
33. Canali S, Schiaffonati V, Aliverti A. Challenges and recommendations for wearable devices in digital health: Data quality, interoperability, health equity, fairness. PLOS Digit Health. 2022;1(10):e0000104.
34. Pan M, Wang S, Li J, Li J, Yang X, Liang K. An Automated Skill Assessment Framework Based on Visual Motion Signals and a Deep Neural Network in Robot-Assisted Minimally Invasive Surgery. Sensors (Basel). 2023;23(9):4496.
35. Zia A, Sharma Y, Bettadapura V, Sarin EL, Essa I. Video and accelerometer-based motion analysis for automated surgical skills assessment. Int J Comput Assist Radiol Surg. 2018;13(3):443-55.
36. Hamza H, & Shabir D, Aboumarzouk O, Al-Ansari A, Shaban Kh, Navkar N. Automated skills assessment in open surgery: A scoping review. Engineering Applications of Artificial Intelligence. 2025;153:110893.
37. Mukaka MM. Statistics Corner: A guide to appropriate use of Correlation coefficient in medical research. Malawi medical journal: the journal of Medical Association of Malawi. 2012;24(3):69-71.
38. Barzegar M, Boushehri E, Amini M, Faghihi SA, Zarifsanaiey N, Daneshvar M. What Happened to Simulation-Based Education in Outpatient Setting in The 21st Century: A Scoping Review. J Adv Med Educ Prof. 2024;12(2):69-78.
39. Oh J. Artificial intelligence-driven real-time assistance in minimally invasive surgery: a technology-oriented narrative review. J Minim Invasive Surg. 2026;29(2):70-81.
40. Choudhry O, Ali Sh, Biyani CS, Dominic J. Real-Time Tool Detection in Laparoscopic Datasets for Surgical Training in Low-Resource Settings. Healthcare Technology Letters. 2025;12(1):e70045.
41. Wang Y, He G, Ahmed MF. A Real-Time Laparoscopic Surgical Instrument Detection System Based on YOLOv5. 2024 IEEE International Conference on Robotics and Biomimetics (ROBIO), China; 2024. pp. 1-6.
42. Bajraktari F, Fleissner K, Pott P. A comparison of two CNN-based instrument detection approaches for automated surgical assistance systems. Current Directions in Biomedical Engineering. 2023;9:599-602.
43. Gautier B, Tugal H, Tang B, Nabi Gh, Erden MS. Real-Time 3D Tracking of Laparoscopy Training Instruments for Assessment and Feedback. Front Robot

AI. 2021;8:751741.
44. Wiese L, Hinz L, Reithmeier E, Korn Ph, Neuhaus M. Detection of Surgical Instruments Based on Synthetic Training Data. Computers. 2025;14(2):69.
45. Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial Intelligence in Surgery: Promises and Perils. Ann Surg. 2018;268(1):70-6.
46. Fukuta A, Yamashita Sh, Maniwa J, Tamaki A, Kondo T, Kawakubo N, et al. Artificial intelligence facilitates the potential of simulator training: An innovative laparoscopic surgical skill validation system using artificial intelligence technology. Int J Comput Assist Radiol Surg. 2025;20(3):597-603.
47. Hannani S, Salehi M, Amiri F, Ali Azadi N. Design and Effect of Neurosurgical Educational Software Using Gamification on Students' Learning and Motivation. J Adv Med Educ Prof. 2024;12(3):189-98.
48. Ghasemi S, Imani B, Jafarkhani A, Hosseinefard H. Roles of Two Learning Methods in the Perceived Competence of Surgery and Quality of Teaching: A Quasi-experimental Study among Operating Room Nursing Students. J Adv Med Educ Prof. 2024;12(3):180-188.