

Review

# Machine Learning for Human-Autonomy Teaming in Surgical Skill Assessment: Scoping Review

Kamal Barati<sup>1</sup>; Michael S Ramirez Campos<sup>2</sup>, BEng; Thomas E Doyle<sup>1,2</sup>, PhD

<sup>1</sup>Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada

<sup>2</sup>School of Biomedical Engineering, McMaster University, Hamilton, ON, Canada

**Corresponding Author:**

Kamal Barati

Department of Electrical and Computer Engineering

McMaster University

1280 Main Street West

ITB A202

Hamilton, ON, L8S 4L7

Canada

Phone: 1 (905) 525 9140 ext 22588

Email: [baratik@mcmaster.ca](mailto:baratik@mcmaster.ca)

## Abstract

**Background:** Human-autonomy teaming (HAT) has the potential to reshape surgical practice by fostering true partnership between surgeons and intelligent systems. To achieve this, AI must move beyond static scoring to provide adaptive, real-time guidance based on reliable skill assessment.

**Objective:** This scoping review aims to map the current landscape of machine learning methods for surgical skill assessment and evaluate their technical readiness for integration into functional HAT systems.

**Methods:** Following PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines, we conducted a systematic search across 3 major scientific databases (PubMed, IEEE Xplore, and Web of Science). We identified and analyzed 92 peer-reviewed studies published between 2019 and 2025. The review focused on data modalities (kinematics, video, and biosignals), model architectures, and validation environments. To assess translational maturity, each study was evaluated using the NASA technology readiness level (TRL) scale and mapped to the Yang et al levels of autonomy for surgical robotics, alongside 6 additional dimensions: real-time capability, interpretability, adaptivity, validation environment, external validation, and clinical integration.

**Results:** Our analysis of the 92 included studies reveals a dominant shift toward multimodal data integration and deep learning architectures. While high performance is frequently reported on benchmark datasets, significant barriers to HAT integration persist. The TRL analysis shows that 77.2% (71/92) of studies remain at TRL 3 (proof-of-concept), with only 2.2% (2/92) reaching TRL 5 or above. Furthermore, 96.7% (89/92) of studies operate at Yang autonomy level 0, providing no autonomous functionality beyond classification. We identified that 91.3% (84/92) of models are static, with no user-specific adaptation, 38% (35/92) lack any external validation, and validation practices remain inconsistent across the field.

**Conclusions:** Current AI techniques provide a robust foundation for objective skill assessment, but they are not yet ready for autonomous teaming. Future development must prioritize model robustness, interpretability, and seamless integration into clinical environments to transition from stand-alone assessment tools to effective surgical teammates.

**Trial Registration:** OSF Registries 10.17605/OSF.IO/PQWS5; <https://osf.io/pqws5>

(JMIR AI 2026;5:e94109) doi: [10.2196/94109](https://doi.org/10.2196/94109)

**KEYWORDS**

deep learning; human-autonomy teaming; machine learning; multimodal data; surgical skill assessment; technology readiness level

## Introduction

Assessing surgical skill is essential for both training and ensuring safe and effective procedures. Traditionally, this has been done through expert observation and standardized rating tools such as the Objective Structured Assessment of Technical Skills (OSATS) [1] or the Global Rating Scale (GRS). Although widely used, these methods are time-consuming, subjective, and not always consistent between raters. Advances in surgical simulation platforms, robotic systems, and sensing technologies—such as motion tracking, video, and biosignal acquisition—offer new opportunities for scalable, objective, and data-driven assessment. In this context, machine learning (ML) has emerged as a transformative tool [2,3] capable of analyzing multimodal data [4,5] to evaluate surgical performance with high accuracy and consistency. Surgical competence encompasses both technical skills (eg, instrument handling and tissue manipulation) and nontechnical skills (eg, communication, decision-making, and situational awareness) [6-8]. This review focuses specifically on the assessment of technical skills using ML methods.

Researchers have explored various ML models to analyze surgical performance using data sources such as tool trajectories, surgical video, force signals, and physiological measurements. These models are often trained to classify levels of expertise, recognize gestures or procedural phases, or evaluate metrics such as efficiency, motion, and safety. For instance, convolutional neural networks (CNNs) and recurrent models such as long short-term memory (LSTM) networks have been applied to video and motion data, while more recent work incorporates transformers for sequence modeling [4,5,9,10]. This computational approach enables consistent and scalable evaluation of surgical skill across settings.

Much of the foundational work in ML for surgical skill assessment has relied on benchmark datasets such as JIGSAWS (Johns Hopkins University and Intuitive Surgical Inc Gesture and Skill Assessment Working Set), which includes synchronized kinematic and video data of simulated robotic tasks performed by participants with varying levels of expertise [11-13]. While this dataset has enabled reproducible research and model benchmarking, it is limited by its small sample size, controlled simulation environment, and narrow scope of tasks. As a result, models trained on such data may struggle to generalize to real-world surgical settings, which are inherently more complex and variable.

To address these limitations, recent studies have shifted toward using clinical data, including videos of real procedures, cadaver models, and live intraoperative recordings [3,5,14]. These sources offer greater realism and diversity, enabling ML models to capture variability in anatomy, surgeon technique, and surgical context. In addition, advances in sensor technologies have made it possible to integrate physiological signals (eg, electroencephalography and electrocardiography), instrument motion, and force data into multimodal learning frameworks [2,10]. This shift opens the door to not only more accurate post hoc evaluations but also the development of real-time, adaptive

systems that can assist, guide, or learn alongside human operators [2,3].

Human-autonomy teaming (HAT) represents a paradigm shift from traditional human-machine interaction toward collaborative partnerships between humans and intelligent systems. Unlike conventional automation, which operates in isolation or under strict human control, HAT involves dynamic cooperation—where decision-making, situational awareness, and task execution are shared across the human and autonomous agent [15,16]. In surgical settings, this could involve AI systems that analyze procedural context in real time, adapt their behavior based on a surgeon's skill level, or provide tailored feedback to enhance performance and learning. As such, HAT is not just a technological evolution but a transformation in how expertise, responsibility, and adaptability are distributed in high-stakes clinical environments.

This review uses an established conceptual framework, adopting the levels of autonomy for surgical robotics proposed by Yang et al [6], which defines 6 stages ranging from level 0 (no autonomy; pure teleoperation) through level 1 (robot assistance, such as virtual fixtures or tremor filtering), level 2 (task autonomy, where the system executes specific subtasks under continuous supervision), level 3 (conditional autonomy, with the robot performing extended sequences while the surgeon monitors and intervenes as needed), to levels 4 and 5 (high and full autonomy, requiring minimal or no human involvement). This taxonomy provides a structured lens through which the technical maturity of current ML-based skill assessment methods can be evaluated. Most studies reviewed in this work operate at or below level 1—providing post hoc classification or passive feedback—while the transition toward HAT demands capabilities aligned with levels 2 and 3, where the system must interpret surgical context in real time, adapt to individual performance, and actively support decision-making within the procedural workflow. By mapping the reviewed literature against this framework, we aim to identify not only what current models achieve but also how far the field remains from enabling genuine human-autonomy collaboration in surgery.

While HAT holds great promise for enhancing surgical decision-making and skill development, its implementation presents several open challenges. Effective teaming requires AI systems that can interpret the surgical context in real time and provide feedback tailored to the surgeon's level of expertise—whether novice, intermediate, or expert. However, developing such adaptive systems is complex. Many current models are trained on limited datasets that fail to capture the variability of real-world surgical environments, reducing their generalizability and robustness [12,14]. Additionally, the lack of transparency in deep learning systems can erode clinical trust, particularly when decisions affect patient outcomes. Real-time integration into the operating room also raises concerns around system latency, interoperability, and regulatory approval. Finally, ethical considerations such as accountability, autonomy, and the transparency of shared decision-making remain underexplored. These limitations highlight the need for a comprehensive synthesis of current research, particularly regarding how HAT systems can adapt to different users and safely support surgical workflows in real time.

To guide this review, we posed the following research questions (RQs):

1. How can HAT models be adapted to assist in real-time surgical decision-making and skill development, focusing on personalized guidance and feedback?
2. What ML methods exist to classify surgical movements and assess surgeon proficiency, and how can multimodal data (eg, video, kinematic, and force sensors) enhance this classification for real-time feedback?
3. What are the benchmarks for evaluating the accuracy and effectiveness of AI-driven movement-based skill assessments, and how do they compare with expert evaluations?

This review presents a systematic account of the current landscape of ML techniques for surgical skill assessment, with particular attention to (1) model design, (2) data modality integration, and (3) readiness for HAT. We synthesize findings across 92 peer-reviewed studies and identify gaps in validation practices, generalizability, and clinical integration. These insights are critical as the field moves toward deploying AI systems that not only evaluate performance but also collaborate with human surgeons in real-time, ultimately shaping the future of surgical education and patient care.

## Methods

### Overview

This scoping review follows the methodological framework proposed by Arksey and O'Malley [17], as refined by Levac et al [18], and is reported following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines (Multimedia Appendix 1) [19,20]. As outlined, this review presents a systematic mapping of existing AI applications in surgical skill assessment, with a particular emphasis on HAT, real-time guidance, and multimodal learning approaches. The protocol for this review was registered on the Open Science Framework prior to data extraction [21]. Two deviations from the registered protocol occurred: (1) 3 additional records were identified through manual searching of reference lists, supplementing the database search described in the protocol; and (2) a technology readiness level (TRL) assessment framework and supplementary translational readiness dimensions (Yang autonomy level, real-time capability, interpretability, adaptivity, validation environment, external validation, and clinical integration) were added during revision in response to peer review feedback to strengthen the analytical evaluation of included studies. Neither deviation affected the screening criteria or the final corpus of included studies.

To ensure consistent application, the following operational definitions guided the coding. Real-time capability was

classified as “real-time” when inference was demonstrated during live task execution, “near-real-time” when the authors reported fast inference speeds but did not demonstrate integration into a live workflow, and “offline” when analysis was performed retrospectively. Interpretability was coded as “built-in” for architectures with inherently transparent components (eg, decision trees and rule-based systems), “post hoc” when dedicated explanation methods such as Shapley additive explanations (SHAP) or gradient-weighted class activation mapping (Grad-CAM) were applied, and “none” otherwise; attention mechanisms and feature importance rankings were coded as “built-in” where the study explicitly used them to explain model decisions, but this classification is acknowledged as a simplification. External validation was subcategorized as cross-task (different tasks within the same dataset), cross-dataset (a different dataset entirely), and cross-site (data from a different institution), reflecting increasing levels of generalizability rigor. Clinical integration was coded as “detailed” when the study included substantive discussion of deployment considerations, workflow integration, or regulatory aspects, and “brief” when clinical relevance was mentioned without elaboration.

### Search Strategy

A structured search was conducted in Web of Science, IEEE Xplore, and PubMed to identify peer-reviewed articles published between January 2019 and July 2025. The search targeted studies on surgical skill assessment and assisted decision-making using ML or related AI approaches. Search terms included combinations of the following: *human-autonomy teaming*, *surgical skill assessment*, *surgical gesture recognition*, *robotic surgery*, *machine learning*, *deep learning*, *multimodal data*, and *real-time feedback*. Boolean operators (AND, OR) were used to structure the queries and adapt them to each database's syntax (Multimedia Appendix 2).

An example search string used for Web of Science was:

ALL=((“surgical movement classification” OR “surgical skill assessment” OR “surgical training AI” OR “gesture recognition surgery” OR “kinematic analysis surgery” OR “surgical education”) OR (“AI-assisted decision making” OR “human-autonomy teaming in surgery” OR “real-time surgical feedback” OR “personalized surgical guidance”)) AND (ALL=(“Artificial Intelligence” OR “Machine Learning” OR “Deep Learning”))

Search results were imported into Covidence, and duplicate entries were removed prior to screening.

### Inclusion and Exclusion Criteria

Textbox 1 defines the inclusion and exclusion criteria to select studies for full-text review.

**Textbox 1.** Inclusion and exclusion criteria for studies.**Inclusion criteria**

- Studies focused on surgical skill assessment, AI-based classification of surgical movements, human-autonomy teaming, or decision-making for surgical tasks.
- Use of machine learning techniques (eg, convolutional neural networks, recurrent neural networks, transformers, or graph convolutional networks) or statistical approaches such as linear or logistic regression.
- Studies addressing real-time or near-real-time feedback, movement classification, or benchmarks related to AI-driven skill assessment.
- Populations involving surgical trainees (medical students, residents, and fellows), experienced surgeons, or robotic systems in surgical skill development contexts.
- Studies using kinematic, video, electroencephalography, or electrocardiogram data for analysis.
- Original research articles (excluding reviews or summaries).
- Published within the past 7 years (2019-2025).

**Exclusion criteria**

- Studies unrelated to surgical skill assessment, AI methods, or decision-making.
- Articles that do not focus on AI-driven movement classification, decision support, or benchmarking of skills.
- Studies that do not use any type of recorded or measured data.
- Preprints, review articles, or non-peer-reviewed content.
- Studies lacking experimental validation or empirical results.
- Articles focusing solely on ethics, philosophy, or conceptual discussions without technical or methodological contributions.

**Screening and Selection**

A total of 1139 records were identified, comprising 391 from PubMed, 193 from IEEE Xplore, 552 from Web of Science, and 3 additional records identified through manual searching of reference lists and related literature. After duplicate removal, a total of 506 papers were screened based on title and abstract to remove clearly irrelevant studies. This left 130 articles for full-text review. After applying the inclusion/exclusion criteria, 92 papers were selected for final synthesis. Screening and selection were independently performed by 2 reviewers (the first author and a collaborator), with disagreements resolved through discussion.

**Data Extraction and Analysis**

For each included study, information was extracted regarding the publication, ML methods used, type of surgical task, data modalities (eg, kinematics, video, and electroencephalography), population (eg, novices, experts, and robotic systems), model outputs, and how performance was evaluated.

Studies were then grouped into thematic clusters based on learning method (eg, deep learning vs rule-based), data modality (eg, video-only vs multimodal), and clinical focus (eg, simulation tasks vs real surgeries). This allowed us to compare trends across time, methods, and application domains.

**Analytical Framework for Technology Readiness**

To systematically evaluate the translational maturity of the reviewed studies, we adopted the NASA TRL scale [22], a widely used framework for assessing the maturity of emerging technologies from basic research (TRL 1) through proven operational deployment (TRL 9). We adapted the 9-level scale to the surgical ML context as follows: TRL 1-2 corresponds to

basic algorithmic research and concept formulation; TRL 3 to proof-of-concept models tested on limited or benchmark data (eg, JIGSAWS); TRL 4 to models validated in laboratory or simulation environments with standard metrics; TRL 5 to models tested on clinical or cross-institutional data; TRL 6-7 to prototype systems demonstrated in simulated or real operating room environments; and TRL 8-9 to clinically validated, regulatory-ready, or routinely deployed systems.

In addition to TRL, each study was assessed along 6 translational dimensions: real-time capability (offline, near-real-time, or real-time inference), interpretability (none, post hoc, or built-in), adaptivity to individual users (none, skill-stratified, or adaptive), validation environment (simulation, laboratory/cadaver, clinical, or mixed), external validation (none, cross-task, cross-dataset, or cross-site), and clinical integration discussed (none, brief, or detailed). These dimensions, together with the TRL and the Yang et al [6] autonomy levels introduced in the Introduction, provide a structured lens for evaluating how far current approaches have progressed toward functional HAT integration. It should be noted that both the TRL and Yang frameworks were applied post hoc as analytical lenses rather than as instruments originally designed for evaluating ML classification studies. In particular, Yang level 0 was used in this review to indicate the absence of any autonomous system functionality, rather than its original meaning of strict teleoperation, since the majority of reviewed studies evaluate algorithms in isolation rather than within an integrated robotic system. These adaptations represent a pragmatic approximation, and their limitations should be considered when interpreting the results.

The TRL and translational dimension assessments were conducted in 2 stages. First, TRLs were assigned independently by 2 reviewers (KB and MSRC) for all 92 studies. Initial percent

agreement on TRL assignment was 67.4% (62/92 studies). Disagreements occurred predominantly at the TRL 3-4 boundary, reflecting differing interpretations of whether laboratory-based validation with human participants constituted component validation (TRL 4) or remained at proof-of-concept level (TRL 3). All 30 disagreements were resolved through discussion, during which explicit decision rules were refined: studies validated exclusively on benchmark or preexisting datasets were assigned TRL 3, while those integrating multiple components (hardware, software, and participants) in a laboratory setting were assigned TRL 4. Second, the remaining translational dimensions (Yang autonomy level, real-time capability, interpretability, adaptivity, validation environment, external validation, and clinical integration) were coded by the first author (KB) using the operational definitions described above and verified by the second reviewer (MSRC), with discrepancies resolved through discussion. The final consensus ratings are reported throughout this review.

## Ethics Approval

This study is a scoping review of previously published research and did not involve the collection of new data from human participants. Therefore, it did not require ethical approval or informed consent according to institutional and national guidelines. All studies included in this review were assumed to have obtained appropriate ethical approval and to have adhered to relevant regulations.

## Results

### Overview

The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram outlining the study selection process is presented in [Figure 1](#). This review includes 92 studies published between 2019 and early 2025. Most studies were published after 2020, reflecting a growing interest in applying ML to surgical skill assessment, especially in simulation and robotic-assisted surgery contexts.

**Figure 1.** PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram outlining the selection process for included studies.

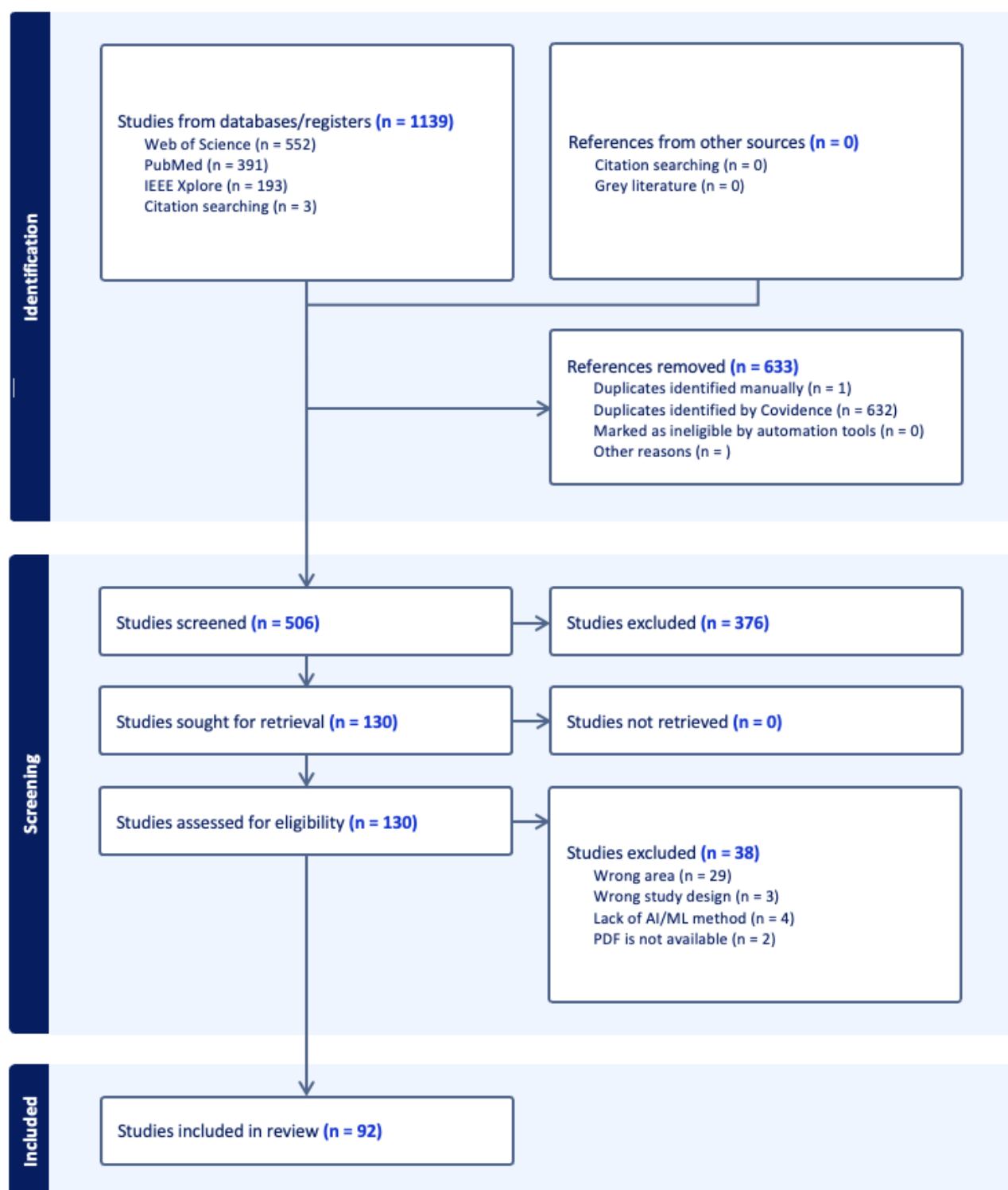


Table 1 summarizes the distribution of translational readiness characteristics across the 92 included studies. The full per-study assessment is provided in Table S1 in [Multimedia Appendix 3](#).

**Table 1.** Distribution of translational readiness characteristics across included studies (n=92).

| Dimension and category        | Studies, n (%) |
|-------------------------------|----------------|
| <b>TRL<sup>a</sup> level</b>  |                |
| Level 3                       | 71 (77.2)      |
| Level 4                       | 19 (20.7)      |
| Level 5                       | 1 (1.1)        |
| Level 7                       | 1 (1.1)        |
| <b>Yang autonomy level</b>    |                |
| Level 0                       | 89 (96.7)      |
| Level 1                       | 3 (3.3)        |
| <b>Real-time capability</b>   |                |
| Offline                       | 53 (57.6)      |
| Near-real-time                | 16 (17.4)      |
| Real-time                     | 23 (25.0)      |
| <b>Interpretability</b>       |                |
| None                          | 21 (22.8)      |
| Post hoc                      | 5 (5.4)        |
| Built-in                      | 66 (71.7)      |
| <b>Adaptivity</b>             |                |
| None                          | 84 (91.3)      |
| Skill-stratified              | 2 (2.2)        |
| Adaptive                      | 6 (6.5)        |
| <b>Validation environment</b> |                |
| Simulation only               | 14 (15.2)      |
| Laboratory/cadaver            | 39 (42.4)      |
| Clinical                      | 26 (28.3)      |
| Mixed                         | 13 (14.1)      |
| <b>External validation</b>    |                |
| No                            | 35 (38)        |
| Cross-task                    | 24 (26.1)      |
| Cross-dataset                 | 19 (20.7)      |
| Cross-site                    | 14 (15.2)      |
| <b>Clinical integration</b>   |                |
| Brief                         | 78 (84.8)      |
| Detailed                      | 14 (15.2)      |

<sup>a</sup>TRL: technology readiness level.

The studies covered a diverse range of surgical tasks and environments (Table 2). This review found an even split between simulation-based and clinical-focused research. Specifically, 46 of the papers used simulation-based setups, such as benchmark datasets (eg, JIGSAWS or Cholecystectomy 80 [Cholec80]), virtual reality (VR) simulators (eg, NeuroVR, ANGIO Mentor, or Fundamentals of Robotic Surgery Dome), bench-top models, or simple box trainers. The remaining 46 papers focused on clinical or semiclinical data, including real

patient procedures, wet laboratory training on animal models, cadaver laboratories, or in vivo videos recorded in operating rooms. This balanced approach demonstrates the field's maturity in transitioning from controlled laboratory validation to real-world clinical evaluation. A few studies notably focused on low-cost custom-built platforms such as the Advanced Robotics and Automated System-Farabi platform for capsulorhexis training.

**Table 2.** Distribution of included studies by application focus and data modality.

|                                                                  | Studies, n |
|------------------------------------------------------------------|------------|
| <b>Application focus</b>                                         |            |
| Gesture recognition or segmentation                              | 17         |
| Skill classification                                             | 56         |
| Tool tracking and motion analysis                                | 29         |
| Real-time feedback or assessment systems                         | 18         |
| <b>Data modality</b>                                             |            |
| Kinematic data                                                   | 18         |
| Video data (RGB, optical flow, etc)                              | 58         |
| Physiological data (EEG <sup>a</sup> , fNIRS <sup>b</sup> , etc) | 3          |
| Force data                                                       | 1          |
| Multimodal approaches (2+modalities fused)                       | 12         |

<sup>a</sup>EEG: electroencephalography.

<sup>b</sup>fNIRS: functional near-infrared spectroscopy.

The data modalities varied significantly (Table 2): video was the primary input for 58 (63%) papers; 18 (19.6%) used kinematic data, often from the JIGSAWS [12] or da Vinci Research Kit platforms; a smaller group of 12 (13%) used multimodal data combining inputs such as video, kinematics, pose estimation, or tool trajectory features [23]; 3 looked at physiological signals such as electroencephalography or functional near-infrared spectroscopy (n=3, 3.3%); and 1 focused on force signals (n=1, 1.1%).

The study populations also varied across the literature. The vast majority of studies, 68 (73.9%) papers, involved mixed-skill groups that included a combination of trainees, residents, fellows, and expert surgeons [2,24]. Another 14 (15.2%) papers did not focus on a specific clinical cohort, instead using preexisting datasets (eg, JIGSAWS) or concentrating on the development of the AI system itself [25,26]. A smaller number of studies focused exclusively on a single group: 6 (6.5%) papers studied only trainees such as medical students [27], while 4 (4.3%) papers analyzed data from only expert surgeons [28].

The data sources used for analysis were highly diverse. The most common approach, found in 29 (31.5%) studies, involved

simulation platforms other than the JIGSAWS benchmark; these included high-fidelity VR trainers for neurosurgery and robotic tasks, as well as physical box trainers for fundamental surgical skills [4,29]. Data from real clinical or in vivo procedures on patients or animal models formed the basis for 28 (30.4%) studies, which analyzed a wide range of surgeries such as live robotic prostatectomies, cataract surgery, and cholecystectomies [5,30]. Another 18 (19.6%) studies developed or used custom or other public datasets, including newly annotated video collections (eg, Cholec80) and unique experimental setups [14,31]. Finally, the standardized JIGSAWS benchmark dataset was the primary data source for 17 (18.5%) studies, which leveraged its kinematic and video data for tasks such as suturing and knot tying [9,32].

Overall, these 92 studies cover a wide cross-section of methods, datasets, and goals, which are presented in Tables 3-6. There has been a clear shift toward integrating different data types, moving beyond offline analysis, and thinking more about how AI systems might work alongside people—rather than evaluate them. These directions are especially important when we start to think about AI as an active teammate, not just a passive tool.

**Table 3.** Summary of included studies on skill classification and regression.

| Study | Year | Data modality                     | Data source                                | Sample size                                 | Assessment target/outcome                                                    | ML <sup>a</sup> method                                                                                                                                                    | Data availability |
|-------|------|-----------------------------------|--------------------------------------------|---------------------------------------------|------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------|
| [33]  | 2025 | Video                             | Simulator                                  | N/A <sup>b</sup>                            | Laparoscopic skill validation                                                | YOLOv4 <sup>c</sup>                                                                                                                                                       | Private           |
| [34]  | 2025 | Video                             | Simulator                                  | N/A                                         | Technical skill classification                                               | Vision Transformer                                                                                                                                                        | Private           |
| [35]  | 2025 | Text (free-text comments)         | Simulator (Vertex Pursuit)                 | 600 data points (24 participants)           | Skill level classification (novice, intermediate, and expert)                | LLM <sup>d</sup> (Claude 3 Opus, GPT-4, Gemini Pro); BERT <sup>e</sup> ; SVM <sup>f</sup> ; KNN <sup>g</sup> ; RF <sup>h</sup> ; XG-Boost <sup>i</sup> ; MLP <sup>j</sup> | Private           |
| [36]  | 2025 | Video                             | Clinical (colorectal)                      | N/A                                         | Skill assessment                                                             | CNN <sup>k</sup> (Efficient-NetB7)                                                                                                                                        | Private           |
| [37]  | 2025 | Motion (rotation)                 | Simulator                                  | N/A                                         | Suturing skill assessment (hand rotation)                                    | Functional data analysis                                                                                                                                                  | Private           |
| [38]  | 2025 | Video (tool motion)               | Simulator (EA OJT <sup>l</sup> Model)      | 45 videos                                   | Skill level classification (good vs poor)                                    | YOLOv3; CNN (4-layer)                                                                                                                                                     | Private           |
| [39]  | 2025 | Video                             | Clinical (cataract)                        | >1000 videos                                | OSACSS <sup>m</sup> performance ratings                                      | YOLACT <sup>n</sup>                                                                                                                                                       | Private           |
| [10]  | 2024 | Kinematics; force                 | VR <sup>o</sup> /AR <sup>p</sup> simulator | 27 participants                             | Expertise classification                                                     | ANN <sup>q</sup> ; transfer learning                                                                                                                                      | Private           |
| [40]  | 2024 | Video (surgeries)                 | Clinical (laparoscopic cholecystectomy)    | 75 videos (33 surgeons)                     | mGOALS <sup>r</sup> skill classification (competent vs incompetent)          | SVM; LR <sup>s</sup> ; RF; GB-DT <sup>t</sup> ; AdaBoost                                                                                                                  | Private           |
| [41]  | 2024 | Video                             | EndoVis SkillNet                           | N/A                                         | Skill classification                                                         | Video Transformer (multitask)                                                                                                                                             | Public            |
| [42]  | 2024 | Video                             | Suture simulator                           | 314 videos                                  | GRS <sup>u</sup> score classification (novice, intermediate, and proficient) | Video Swin Transformer; I3D <sup>v</sup>                                                                                                                                  | Public            |
| [43]  | 2024 | Video                             | Clinical (laparoscopic distal gastrectomy) | 256 videos (training); 180 videos (testing) | ESSQS <sup>w</sup> skill score (phase duration and AICS <sup>x</sup> )       | CNN (Efficient-NetB7)                                                                                                                                                     | Private           |
| [44]  | 2024 | Video                             | Clinical (laparoscopic colorectal)         | 1572 videos                                 | ESSQS skill score regression                                                 | EfficientNetB7; regression                                                                                                                                                | Private           |
| [24]  | 2023 | Video (motion)                    | Laparoscopic trainer (LMICs <sup>y</sup> ) | 74 videos (47 trainees)                     | OSATS <sup>z</sup> score prediction                                          | KNN; YOLOv5                                                                                                                                                               | Private           |
| [45]  | 2024 | EEG <sup>aa</sup> ; eye-tracking  | Clinical (porcine model)                   | 43 surgeries (11 surgeons)                  | GEARS <sup>ab</sup> skill level classification                               | RF; XGBoost; LR; gradient boosting                                                                                                                                        | Private           |
| [46]  | 2024 | Force                             | Simulator                                  | N/A                                         | Skill assessment (Smoothness of Force)                                       | Time/frequency analysis                                                                                                                                                   | Private           |
| [47]  | 2024 | Motion (EM <sup>ac</sup> sensors) | SutureCoach Simulator                      | 97 participants                             | Expertise level classification (novice, intermediate, and expert)            | Statistical metrics                                                                                                                                                       | Private           |
| [48]  | 2024 | Kinematics                        | dVRK <sup>ad</sup>                         | N/A                                         | Bimanual coordination assessment                                             | Information theory; DTW <sup>ae</sup>                                                                                                                                     | Private           |
| [30]  | 2024 | Video                             | Clinical (cataract surgery)                | 197 videos (12 surgeons)                    | Skill level classification (expert vs novice)                                | 2D CNN-LSTM <sup>af</sup> ; 3D CNN                                                                                                                                        | Private           |
| [49]  | 2024 | Kinematics                        | JIGSAWS <sup>ag</sup> ; Wolf dataset       | 8 (JIGSAWS); 32 (Wolf)                      | GRS item score prediction (GOALS <sup>ah</sup> /OSATS)                       | Multiheaded MLP                                                                                                                                                           | Public            |
| [9]   | 2024 | Kinematics (as images)            | JIGSAWS                                    | 8 participants                              | Skill level classification                                                   | Vision Transformer; CWT <sup>ai</sup>                                                                                                                                     | Public            |

| Study | Year | Data modality        | Data source                                  | Sample size                                | Assessment target/outcome                                     | ML <sup>a</sup> method                                                                   | Data availability |
|-------|------|----------------------|----------------------------------------------|--------------------------------------------|---------------------------------------------------------------|------------------------------------------------------------------------------------------|-------------------|
| [50]  | 2023 | Video                | JIGSAWS                                      | 8 participants                             | GRS score prediction (relative to expert)                     | Transformer; TCN <sup>aj</sup> ; ResNet <sup>ak</sup> -18                                | Public            |
| [51]  | 2023 | Video                | Clinical (laparoscopic sigmoidectomy)        | 120 videos                                 | ESSQS score correlation (AICS)                                | CNN (Xception)                                                                           | Private           |
| [52]  | 2023 | Kinematics           | JIGSAWS                                      | 8 participants                             | Skill level classification (novice, intermediate, and expert) | Dempster-Shafer theory; SVM; XGBoost                                                     | Public            |
| [2]   | 2023 | EEG                  | VR simulator (NeuroVR)                       | 21 participants                            | Expertise level classification (skilled vs less-skilled)      | ANN; SVM; LDA <sup>al</sup> ; RF; KNN; LR; naive Bayes                                   | Private           |
| [53]  | 2023 | Video (tool motion)  | JIGSAWS                                      | 24 videos (8 participants)                 | Skill level classification (2-class and 3-class)              | ResNet; KCF <sup>am</sup> tracking                                                       | Public            |
| [54]  | 2023 | Simulator metrics    | ANGIO Mentor Simulator                       | 22 participants                            | Modified Reznick scale score prediction                       | Feedforward neural network                                                               | Private           |
| [16]  | 2023 | Kinematics           | JIGSAWS; VR simulator                        | 8 (JIGSAWS); 16 (VR)                       | Skill level classification (novice vs expert)                 | CNN; Bi-LSTM <sup>an</sup> ; MC <sup>ao</sup> Dropout                                    | Public            |
| [55]  | 2023 | Video (tool motion)  | JIGSAWS; FLS <sup>ap</sup> trainer           | 8 (JIGSAWS); 33 (FLS)                      | Pass/fail classification                                      | Res-CNN <sup>aq</sup> ; DAE <sup>ar</sup> ; Mask R-CNN <sup>as</sup> ; CAM <sup>at</sup> | Public            |
| [56]  | 2022 | Motion capture       | Wet Lab box trainer                          | 89 sessions (70 participants)              | GOALS score prediction                                        | SVR <sup>au</sup> ; PCA <sup>av</sup> ; Ridge; PLS <sup>aw</sup> Regression              | Private           |
| [57]  | 2022 | Video (hand motion)  | Laparoscopic simulator                       | 9 participants                             | Hand movement performance assessment                          | Fuzzy logic; SSD <sup>ax</sup> ResNet50                                                  | Public            |
| [58]  | 2022 | Video (tool motion)  | Cholec80                                     | 80 videos (13 surgeons)                    | GOALS efficiency score classification                         | Transformer; random forest; YOLOv5                                                       | Public            |
| [59]  | 2022 | Kinematics; image    | Knot-tying simulator                         | 360 trials (72 trainees)                   | OSATS score prediction (4 domains)                            | ResNet; Bi-LSTM                                                                          | Public            |
| [60]  | 2022 | VR simulators        | VR simulators                                | 254 sessions                               | GEARS score classification (novice, intermediate, and expert) | DNN <sup>ay</sup> (Google AutoML)                                                        | Private           |
| [27]  | 2022 | fNIRS <sup>az</sup>  | FLS Trainer                                  | 474 trials (7 students)                    | Pass/fail classification (FLS task)                           | Dilated causal CNN (WaveNet-style)                                                       | Private           |
| [61]  | 2022 | Motion; force        | Phantom; clinical                            | 12 surgeons                                | Skill level classification (expert vs novice)                 | RF; XGBoost; SHAP <sup>ba</sup>                                                          | Private           |
| [62]  | 2022 | Kinematics (1D)      | Clinical (porcine model)                     | 100 trials (10 surgeons)                   | Intra- and intersubject skill similarity                      | Dynamic time warping                                                                     | Private           |
| [4]   | 2021 | Kinematics; force    | VR Simulator (Sim-Ortho)                     | 23 participants                            | Expertise classification                                      | ANN (multilayer perceptron)                                                              | Private           |
| [63]  | 2021 | Kinematics           | JIGSAWS                                      | 33 trials (7 surgeons)                     | Skill level classification (novice, intermediate, and expert) | KNN; PCA; Relief                                                                         | Public            |
| [64]  | 2021 | Video (optical flow) | JIGSAWS                                      | 8 participants                             | Skill level classification (novice vs expert)                 | ResNet; CNN; LSTM; ConvAuto                                                              | Public            |
| [65]  | 2021 | Video (tool motion)  | Clinical (cholecystectomy)                   | 949 clips                                  | Likert scale skill score prediction (1-5)                     | Linear regression; ResNet50-FPN <sup>bb</sup>                                            | Private           |
| [66]  | 2021 | Video; kinematics    | JIGSAWS; Clinical (laparoscopic gastrectomy) | 206 videos (JIGSAWS); 20 videos (clinical) | GRS score prediction (modified OSATS)                         | TCN; ResNet-101; MS-TCN <sup>bc</sup> ; MLP; contrastive learning                        | Public, Private   |

| Study | Year | Data modality                     | Data source                                              | Sample size                | Assessment target/outcome                                      | ML <sup>a</sup> method                                                   | Data availability |
|-------|------|-----------------------------------|----------------------------------------------------------|----------------------------|----------------------------------------------------------------|--------------------------------------------------------------------------|-------------------|
| [67]  | 2021 | Video                             | JIGSAWS                                                  | 8 participants             | GRS score classification (expert vs novice)                    | SVM; BoF <sup>bd</sup> (STIP <sup>be</sup> , iDT <sup>bf</sup> )         | Public            |
| [12]  | 2021 | Video                             | JIGSAWS                                                  | 8 participants             | Skill level classification (novice, intermediate, and expert)  | ResNet50; FFT <sup>bg</sup> ; 1D CNN                                     | Public            |
| [68]  | 2021 | Sensor (force, time)              | FRS <sup>bh</sup> Dome Simulator                         | 27 performances            | FRS performance score (1-3)                                    | ANFIS <sup>bi</sup> ; fuzzy logic                                        | Private           |
| [32]  | 2020 | Kinematics                        | JIGSAWS                                                  | 8 participants             | Skill level classification                                     | CNN; LSTM; PCA; DFT <sup>bj</sup> ; DCT <sup>bk</sup> ; Autoencoder; SVM | Public            |
| [69]  | 2020 | Video                             | JIGSAWS                                                  | 8 participants             | OSATS score prediction                                         | I3D; TSN <sup>bl</sup> (multi-task)                                      | Public            |
| [70]  | 2020 | Video (tool motion)               | Clinical (robotic thyroidectomy); Simulator (BABA Model) | 54 video segments          | OSATS/GEARS skill classification (novice, skilled, and expert) | Mask R-CNN; Deep SORT <sup>bm</sup> ; SVM; random forest                 | Private           |
| [71]  | 2020 | Kinematics                        | JIGSAWS                                                  | 8 participants             | GRS score classification (expert vs novice)                    | Kernel SVM                                                               | Public            |
| [72]  | 2019 | Kinematics                        | JIGSAWS                                                  | 8 participants             | OSATS score prediction                                         | FCN <sup>bn</sup> ; CAM                                                  | Public            |
| [73]  | 2019 | Video; optical flow               | JIGSAWS                                                  | 8 participants             | Skill level classification (novice, intermediate, and expert)  | 3D CNN (I3D); TSN                                                        | Public            |
| [74]  | 2019 | Video (tool motion)               | UVA <sup>bo</sup> Dataset                                | 12 surgeons                | Skill level classification (novice vs expert)                  | HMM <sup>bp</sup> ; SVM; LDA                                             | Public            |
| [75]  | 2019 | Video (tool motion, optical flow) | Clinical (cataract surgery)                              | 99 videos                  | ICO-OSCAR <sup>bq</sup> score prediction (expert vs nonexpert) | TCN                                                                      | Private           |
| [76]  | 2019 | EMG <sup>br</sup> ; motion        | Myo Armband (box trainer)                                | 99 knots (28 participants) | OSATS score prediction                                         | Neural network; decision trees                                           | Private           |

| Study | Year | Data modality           | Data source                         | Sample size           | Assessment target/outcome                | ML <sup>a</sup> method            | Data availability |
|-------|------|-------------------------|-------------------------------------|-----------------------|------------------------------------------|-----------------------------------|-------------------|
| [77]  | 2019 | Video (color, semantic) | Clinical (laparoscopic gastrectomy) | 57 videos (1 surgeon) | OSATS/COF <sup>bs</sup> score prediction | ResNet-101; regression; rank loss | Private           |

<sup>a</sup>ML: machine learning.

<sup>b</sup>N/A: not available.

<sup>c</sup>YOLOv4: You Only Look Once version 4.

<sup>d</sup>LLM: large language model.

<sup>e</sup>BERT: Bidirectional Encoder Representations from Transformers.

<sup>f</sup>SVM: support vector machine.

<sup>g</sup>KNN: k-nearest neighbors.

<sup>h</sup>RF: random forest.

<sup>i</sup>XGBoost: extreme gradient boosting.

<sup>j</sup>MPL: multilayer perceptron.

<sup>k</sup>CNN: convolutional neural network.

<sup>l</sup>EA OJT: esophageal atresia off-the-job training.

<sup>m</sup>OSACSS: Objective Structured Assessment of Cataract Surgical Skill.

<sup>n</sup>YOLACT: You Only Look at Coefficients.

<sup>o</sup>VR: virtual reality.

<sup>p</sup>AR: augmented reality.

<sup>q</sup>ANN: artificial neural network.

<sup>r</sup>mGOALS: Modified Global Objective Assessment of Laparoscopic Skills.

<sup>s</sup>LR: logistic regression.

<sup>t</sup>GBDT: gradient boosted decision trees.

<sup>u</sup>GRS: Global Rating Scale.

<sup>v</sup>I3D: Inflated 3D ConvNet.

<sup>w</sup>ESSQS: Endoscopic Surgical Skill Qualification System.

<sup>x</sup>AICS: Artificial Intelligence Confidence Score.

<sup>y</sup>LMIC: low-and-middle-income country.

<sup>z</sup>OSATS: Objective Structured Assessment of Technical Skills.

<sup>aa</sup>EEG: electroencephalography.

<sup>ab</sup>GEARS: Global Evaluative Assessment of Robotic Skills.

<sup>ac</sup>EM: electromagnetic.

<sup>ad</sup>dVRK: da Vinci Research Kit.

<sup>ae</sup>DTW: dynamic time warping.

<sup>af</sup>LSTM: long short-term memory.

<sup>ag</sup>JIGSAWS: Johns Hopkins University and Intuitive Surgical Inc Gesture and Skill Assessment Working Set.

<sup>ah</sup>GOALS: Global Objective Assessment of Laparoscopic Skills.

<sup>ai</sup>CWT: continuous wavelet transform.

<sup>aj</sup>TCN: temporal convolutional network.

<sup>ak</sup>ResNet: residual network.

<sup>al</sup>LDA: linear discriminant analysis.

<sup>am</sup>ANGIO: angiotherapy.

<sup>an</sup>Bi-LSTM: bidirectional long short-term memory.

<sup>ao</sup>MC: Monte Carlo.

<sup>ap</sup>FLS: Fundamentals of Laparoscopic Surgery.

<sup>aq</sup>Res-CNN: residual convolutional neural network.

<sup>ar</sup>DAE: denoising autoencoder.

<sup>as</sup>Mask R-CNN: mask region-based convolutional neural network.

<sup>at</sup>CAM: class activation mapping.

<sup>au</sup>SVR: support vector regression.

<sup>av</sup>PCA: principal component analysis.

<sup>aw</sup>PLS: partial least squares.

<sup>ax</sup>SSD: single shot detector.

<sup>ay</sup>DNN: deep neural network.

<sup>az</sup>fNIRS: functional near-infrared spectroscopy.

<sup>ba</sup>SHAP: Shapley additive explanations.

<sup>bb</sup>FPN: feature pyramid network.

<sup>bc</sup>MS-TCN: multistage temporal convolutional networks.

<sup>bd</sup>BoF: bag of features.

<sup>be</sup>STIP: spatio-temporal interest points.

<sup>bf</sup>iDT: improved dense trajectories.

<sup>bg</sup>FFT: fast Fourier transform.

<sup>bh</sup>FRS: Fundamentals of Robotic Surgery.

<sup>bi</sup>ANFIS: adaptive neuro-fuzzy inference system.

<sup>bj</sup>DFT: discrete Fourier transform.

<sup>bk</sup>DCT: discrete cosine transform.

<sup>bl</sup>TSN: temporal segment network.

<sup>bm</sup>SORT: simple online and realtime tracking.

<sup>bn</sup>FCN: fully convolutional network.

<sup>bo</sup>UVA: urethro-vesicle anastomosis.

<sup>bp</sup>HMM: hidden Markov model.

<sup>bq</sup>ICO-OSCAR: International Council of Ophthalmology's Ophthalmology Surgical Competency Assessment Rubric.

<sup>br</sup>EMG: electromyography.

<sup>bs</sup>COF: clearness of operating field.

**Table 4.** Summary of included studies on gesture and phase recognition.

| Study | Year | Data modality       | Data source                                         | Sample size                              | ML <sup>a</sup> method                                                                   | Data availability |
|-------|------|---------------------|-----------------------------------------------------|------------------------------------------|------------------------------------------------------------------------------------------|-------------------|
| [78]  | 2025 | Video               | Microsurgery videos                                 | 132 sutures (11 surgeons)                | T-LRCN <sup>b</sup> ; ConvLSTM <sup>c</sup>                                              | Private           |
| [79]  | 2025 | Video               | Clinical (robotic prostatectomy)                    | N/A <sup>d</sup>                         | CNN <sup>e</sup>                                                                         | Private           |
| [80]  | 2025 | Video               | Clinical (cholecystectomy)                          | N/A                                      | Action recognition model                                                                 | Private           |
| [81]  | 2024 | Video               | Clinical (laparoscopic pancreaticoduodenectomy)     | 69 videos                                | CNN (TSM <sup>f</sup> -based)                                                            | Private           |
| [82]  | 2023 | Video               | Simulator (low-cost appendectomy trainer)           | 43 videos (12 experts, 31 residents)     | CNN (ResNet50 <sup>g</sup> ); Bi-LSTM <sup>h</sup>                                       | Private           |
| [83]  | 2023 | Video; hand pose    | Open surgery simulator                              | 100 videos (25 clinicians)               | MS <sup>i</sup> -TCN <sup>++j</sup> ; I3D <sup>k</sup> ; YOLOX <sup>l</sup>              | Public            |
| [13]  | 2022 | Video               | JIGSAWS <sup>m</sup>                                | 39 videos (8 surgeons)                   | Calibrated MS-TCN; I3D; SlowFast                                                         | Public            |
| [84]  | 2022 | Video; kinematics   | JIGSAWS; RARP-45                                    | 39 (JIGSAWS); 45 (RARP)                  | MA <sup>n</sup> -TCN                                                                     | Public, private   |
| [85]  | 2022 | Video               | Clinical (laparoscopic TAPP hernia repair)          | 119 videos (11 surgeons)                 | TeCNO <sup>o</sup> (MS-TCN); HMM <sup>p</sup> ; SV-RCNet <sup>q</sup> ; KNN <sup>r</sup> | Private           |
| [86]  | 2021 | Video; kinematics   | JIGSAWS; dVRK <sup>s</sup> simulator (peg transfer) | 75 videos (JIGSAWS); 24 sequences (dVRK) | Relational GCN <sup>t</sup> (MRG-Net <sup>u</sup> ); ResNet-18; TCN; LSTM                | Public, private   |
| [5]   | 2021 | Video; optical flow | Clinical (robotic prostatectomy)                    | 511 clips                                | AlexNet; LSTM; ConvLSTM                                                                  | Private           |
| [28]  | 2021 | EEG <sup>v</sup>    | Clinical (robotic prostatectomy)                    | 34 surgeries (5 surgeons)                | Extra Trees; RF <sup>w</sup> ; bagging; KNN                                              | Private           |
| [87]  | 2020 | Kinematics          | JIGSAWS                                             | 39 trials (8 surgeons)                   | Multitask Bi-LSTM                                                                        | Public            |
| [14]  | 2020 | Video               | Clinical (laparoscopic colorectal)                  | 300 videos                               | CNN (Xception); U-Net <sup>x</sup>                                                       | Private           |
| [88]  | 2019 | Kinematics          | JIGSAWS; V-RASTED <sup>y</sup>                      | 8 (JIGSAWS); 18 (V-RASTED)               | Time-delay neural network (TDNN <sup>z</sup> )                                           | Public            |

<sup>a</sup>ML: machine learning.<sup>b</sup>T-LRCN: temporal long-term recurrent convolutional network.<sup>c</sup>ConvLSTM: convolutional long short-term memory.<sup>d</sup>N/A: not available.<sup>e</sup>CNN: convolutional neural network.<sup>f</sup>TSM: temporal shift module.<sup>g</sup>ResNet50: Residual Network With 50 Layers.<sup>h</sup>Bi-LSTM: bidirectional long short-term memory.<sup>i</sup>MS: multistage.<sup>j</sup>TCN: temporal convolutional network.<sup>k</sup>I3D: inflated 3D convolutional network.<sup>l</sup>YOLOX: You Only Look Once X.<sup>m</sup>JIGSAWS: Johns Hopkins University and Intuitive Surgical Inc Gesture and Skill Assessment Working Set.<sup>n</sup>MA-TCN: multiaxis temporal convolutional network.<sup>o</sup>TeCNO: temporal convolutional network for operating room phase recognition.<sup>p</sup>HMM: hidden Markov model.<sup>q</sup>SV-RCNet: surgical video recurrent convolutional network.

<sup>r</sup>KNN: k-nearest neighbors.

<sup>s</sup>dVRK: da Vinci Research Kit.

<sup>t</sup>GCN: graph convolutional network.

<sup>u</sup>MRG-Net: multirelational graph network.

<sup>v</sup>EEG: electroencephalography.

<sup>w</sup>RF: random forest.

<sup>x</sup>U-Net: U-shaped convolutional network.

<sup>y</sup>V-RASTED: Virtual Reality Assessment of Surgical Tool-Handling and Ergonomic Dexterity.

<sup>z</sup>TDNN: time-delay neural network.

**Table 5.** Summary of included studies on real-time assistance and safety analysis.

| Study                                | Year | Data modality            | Data source                                                                  | Sample size                              | ML <sup>a</sup> method                                                                          | Data availability |
|--------------------------------------|------|--------------------------|------------------------------------------------------------------------------|------------------------------------------|-------------------------------------------------------------------------------------------------|-------------------|
| <b>Real-time feedback/guidance</b>   |      |                          |                                                                              |                                          |                                                                                                 |                   |
| [89]                                 | 2025 | Video                    | Clinical (robotic mastectomy)                                                | 10 surgeries                             | Dissection plane segmentation model                                                             | Private           |
| [90]                                 | 2025 | Video                    | Clinical (colorectal)                                                        | N/A <sup>b</sup>                         | Nerve recognition model                                                                         | Private           |
| [91]                                 | 2024 | Video                    | Clinical (pituitary surgery)                                                 | >200 videos                              | Mask R-CNN <sup>c</sup>                                                                         | Private           |
| [3]                                  | 2024 | Video                    | Clinical (robotic urology)                                                   | 2 surgeries                              | Video transformer network                                                                       | Private           |
| [92]                                 | 2023 | Video (multicamera)      | Laparoscopic box-trainer (Fundamentals of Laparoscopic Surgery peg transfer) | 55 videos; 5000+ images (9 participants) | SSD <sup>d</sup> (ResNet50 <sup>e</sup> V1 FPN <sup>f</sup> ); rule-based sequential assessment | Public            |
| [57]                                 | 2022 | Video (top-down cameras) | Laparoscopic box-trainer                                                     | 55 videos (9 residents)                  | SSD; ResNet50; rule-based                                                                       | Public            |
| [29]                                 | 2022 | Video; kinematics        | dVRK <sup>g</sup> simulator (SurRoL)                                         | >1000 trials                             | Reinforcement learning (RL <sup>h</sup> )                                                       | Private           |
| [93]                                 | 2022 | Video                    | Clinical (cholecystectomy)                                                   | 290 surgeries                            | Semantic segmentation (ResNet50+PSPNet <sup>i</sup> )                                           | Private           |
| <b>Tool tracking/pose estimation</b> |      |                          |                                                                              |                                          |                                                                                                 |                   |
| [94]                                 | 2025 | Video                    | Laparoscopic simulator                                                       | 10 surgeons                              | DeepLabCut (Pose Estimation)                                                                    | Private           |
| [15]                                 | 2025 | Video                    | 3D printed model                                                             | 15 procedures                            | YOLOv8 <sup>j</sup> ; BoT <sup>k</sup> -SORT <sup>l</sup>                                       | Private           |
| [95]                                 | 2024 | Video; IMU               | SutureCoach Simulator                                                        | 81 participants                          | CNN <sup>m</sup> (ResNet-50)                                                                    | Private           |
| [23]                                 | 2023 | Video; kinematics        | Custom Multiprocedure Dataset                                                | 950 frames (5 surgeries)                 | GCN <sup>n</sup> ; ResNet; Transformer                                                          | Private           |
| [96]                                 | 2022 | Video (simulated)        | da Vinci simulator                                                           | 9 videos                                 | 6-DoF <sup>o</sup> Motion Planning                                                              | Private           |
| [97]                                 | 2022 | Video (synthetic)        | Synthetic MICCAI <sup>p</sup> ; JIGSAWS <sup>q</sup>                         | 8 videos                                 | PnP <sup>r</sup> ; optical flow; time series forest                                             | Public            |
| [98]                                 | 2022 | Video                    | ATLAS Dione <sup>s</sup> ; Endovis <sup>t</sup> ; Cholec80 <sup>u</sup>      | 99 clips + 1083 frames                   | DSCNet-CLSTM <sup>v</sup> ; CNN regression                                                      | Public            |
| [31]                                 | 2021 | Video                    | Clinical (cataract surgery); CaDIS <sup>w</sup> ; Cataract-101               | 629 images                               | Faster R-CNN; SSD; RetinaNet; Scaled-YOLOv4                                                     | Public, Private   |
| <b>Error detection/safety</b>        |      |                          |                                                                              |                                          |                                                                                                 |                   |
| [99]                                 | 2024 | Video                    | JIGSAWS                                                                      | N/A                                      | Transformer                                                                                     | Public            |
| [100]                                | 2024 | Video                    | Microsurgery simulator                                                       | 24 trials (12 surgeons)                  | Semantic segmentation (ResNet-50)                                                               | Private           |

<sup>a</sup>ML: machine learning.<sup>b</sup>N/A: not available.<sup>c</sup>R-CNN: region-based convolutional neural network.<sup>d</sup>SSD: single shot detector.<sup>e</sup>ResNet: residual network.<sup>f</sup>FPN: feature pyramid network.<sup>g</sup>dVRK: da Vinci research kit.<sup>h</sup>RL: reinforcement learning.<sup>i</sup>PSPNet: pyramid scene parsing network.<sup>j</sup>YOLO: You Only Look Once.<sup>k</sup>BoT: bag of tricks.<sup>l</sup>SORT: simple online and realtime tracking.<sup>m</sup>CNN: convolutional neural network.<sup>n</sup>GCN: graph convolutional network.<sup>o</sup>6-DoF: 6 degrees of freedom.

<sup>p</sup>MICCAI: Medical Image Computing and Computer Assisted Intervention.  
<sup>q</sup>JIGSAWS: Johns Hopkins University and Intuitive Surgical Inc Gesture and Skill Assessment Working Set.  
<sup>r</sup>PnP: perspective-n-point.  
<sup>s</sup>ATLAS Dione: Applied Technology Laboratory for Advanced Surgery Dione dataset.  
<sup>t</sup>EndoVis: Endoscopic Vision Challenge.  
<sup>u</sup>Cholec80: a dataset of 80 cholecystectomy procedure videos.  
<sup>v</sup>DSCNet-CLSTM: depth-wise separable convolutional network with convolutional long short-term memory.  
<sup>w</sup>CaDIS: cataract dataset for image segmentation.

**Table 6.** Summary of included studies on surgical workflow and data analysis.

| Study                                        | Year | Data modality                   | Data source                                                  | Sample size                       | ML <sup>a</sup> method                                           | Data availability |
|----------------------------------------------|------|---------------------------------|--------------------------------------------------------------|-----------------------------------|------------------------------------------------------------------|-------------------|
| <b>Learning curve analysis</b>               |      |                                 |                                                              |                                   |                                                                  |                   |
| [101]                                        | 2024 | Video                           | Simulator                                                    | N/A <sup>b</sup>                  | 3D CNN <sup>c</sup> ; TSN <sup>d</sup>                           | Private           |
| [102]                                        | 2022 | Kinematics; force               | VR <sup>e</sup> simulator (NeuroVR)                          | 50 participants                   | KNN <sup>f</sup> (for metric selection); ANOVA                   | Private           |
| <b>Video retrieval/content understanding</b> |      |                                 |                                                              |                                   |                                                                  |                   |
| [26]                                         | 2024 | Image; text (VQA <sup>g</sup> ) | EndoVis17/18; DAISI <sup>h</sup> -VQA                        | 545 QA <sup>i</sup> pairs (DAISI) | Multimodal LLM <sup>j</sup> (Instruct-BLIP); CL <sup>k</sup>     | Public            |
| [103]                                        | 2024 | Video                           | Cholec80 <sup>l</sup> ; Cataract-101; CholecT45 <sup>m</sup> | 80, 101, 45 videos                | Hashing; Transformer                                             | Public            |
| [25]                                         | 2022 | Video                           | Surgical Actions 160; Cataract-101                           | 160 clips; 101 surgeries          | CNN; LSTM <sup>n</sup> ; TCN <sup>o</sup> ; contrastive learning | Public            |
| [104]                                        | 2019 | Kinematics                      | JIGSAWS <sup>p</sup>                                         | 8 participants                    | DTW <sup>q</sup> ; DBA <sup>r</sup> ; NLTS <sup>s</sup>          | Public            |

<sup>a</sup>ML: machine learning.  
<sup>b</sup>N/A: not available.  
<sup>c</sup>CNN: convolutional neural network.  
<sup>d</sup>TSN: temporal segment network.  
<sup>e</sup>VR: virtual reality.  
<sup>f</sup>KNN: k-nearest neighbors.  
<sup>g</sup>VQA: visual question answering.  
<sup>h</sup>DAISI: Database for AI Surgical Instruction.  
<sup>i</sup>QA: question-answer.  
<sup>j</sup>LLM: large language model.  
<sup>k</sup>CL: contrastive learning.  
<sup>l</sup>Cholec80: a dataset of 80 cholecystectomy procedure videos.  
<sup>m</sup>CholecT45: a dataset of 45 cholecystectomy procedure videos with triplet annotations.  
<sup>n</sup>LSTM: long short-term memory.  
<sup>o</sup>TCN: temporal convolutional network.  
<sup>p</sup>JIGSAWS: Johns Hopkins University and Intuitive Surgical Inc Gesture and Skill Assessment Working Set.  
<sup>q</sup>DTW: dynamic time warping.  
<sup>r</sup>DBA: DTW barycenter averaging.  
<sup>s</sup>NLTS: nonlinear time series.

AI-Based Methods for Surgical Skill Assessment

Overview

Across the 92 papers, there’s a wide range of ML techniques used to classify surgical movements, recognize gestures, or assess technical skill (Table 7). The dominant approach involved CNNs, appearing in 55 (59.8%) studies. This broad category includes 2D networks such as ResNet for feature extraction,

spatiotemporal models such as Inflated 3D ConvNet (I3D) for video analysis, and temporal convolutional networks (TCNs) for sequence modeling [72,73]. Classical ML methods also remained prevalent, with 29 (31.5%) studies using techniques such as support vector machine (SVM), random forests, and regression models, particularly for skill classification from handcrafted features [24,56]. Object detection and segmentation models (eg, You Only Look Once[YOLO] and U-shaped

convolutional network [U-Net]) were specifically used in 16 (17.4%) studies to track instruments or identify anatomy [14,57]. Recurrent neural networks (RNNs), primarily LSTMs, were featured in 14 (15.2%) studies to capture temporal dependencies [5]. More recent architectures based on transformers and

attention were used in 11 (12%) studies [3,9], while other approaches included time series and probabilistic methods (eg, dynamic time warping or hidden Markov models [HMM]) in 9 (9.8%) studies [104] and unsupervised or self-supervised learning in 7 (7.6%) studies [32,66].

**Table 7.** Frequency of different machine learning (ML) techniques across the 92 reviewed papers. Convolutional neural network (CNN)-based architectures were the most common. The total count exceeds 92 because many studies used multiple techniques (eg, a CNN for feature extraction and a long short-term memory (LSTM) for temporal analysis).

| Frequency of machine learning techniques used                                            | Studies (n=92), n |
|------------------------------------------------------------------------------------------|-------------------|
| Unsupervised & self-supervised learning                                                  | 7                 |
| Time series & probabilistic methods (DTW <sup>a</sup> , HMM <sup>b</sup> , etc)          | 9                 |
| Transformers & attention mechanisms                                                      | 11                |
| RNN <sup>c</sup> -based architectures (LSTM, etc)                                        | 14                |
| Object detection and segmentation (YOLO <sup>d</sup> , U-Net <sup>e</sup> , etc)         | 16                |
| Classical ML (SVM <sup>f</sup> , random forest, KNN <sup>g</sup> , etc)                  | 29                |
| CNN-based architecture (ResNet <sup>h</sup> , I3D <sup>i</sup> , TCN <sup>j</sup> , etc) | 55                |

<sup>a</sup>DTW: dynamic time warping.  
<sup>b</sup>HMM: hidden Markov model.  
<sup>c</sup>RNN: recurrent neural network.  
<sup>d</sup>YOLO: You Only Look Once.  
<sup>e</sup>U-Net: U-shaped convolutional network.  
<sup>f</sup>SVM: support vector machine.  
<sup>g</sup>KNN: k-nearest neighbors.  
<sup>h</sup>ResNet: residual network.  
<sup>i</sup>I3D: inflated 3D convolutional network.  
<sup>j</sup>TCN: temporal convolutional network.

**Classical ML**

Across the reviewed literature, 29 (31.5%) studies used classical ML methods such as SVMs, random forests, and k-nearest neighbors. These models were typically applied to features extracted from kinematics data or motion trajectories such as path length, speed, and jerk. For instance, Ebina et al [56] applied support vector regression to predict Global Objective Assessment of Laparoscopic Skills (GOALS) scores from motion capture data, while Hasani et al [63] used a k-nearest neighbors classifier on kinematic features to assess suturing skill.

Separately, 9 (9.8%) studies used time-series, probabilistic, or rule-based models to address temporal dependencies and uncertainty. Gorantla and Esfahani [74], for example, used motor control descriptors as input to HMMs, demonstrating that structured, time-aware models could infer meaningful surgical patterns even with limited data. To tackle label ambiguity directly, Iranfar et al [52] incorporated Dempster-Shafer theory to quantify classification uncertainty. Their method fused soft evidence and handled ambiguous gesture transitions more robustly than standard decision boundaries, highlighting the value of belief-based reasoning when surgical labels are imprecise or noisy.

While these methods provided valuable early insights, their reliance on handcrafted features and limited scalability led to a broader adoption of deep learning models capable of learning directly from raw, high-dimensional inputs.

**Deep Learning Models**

Deep learning has become the dominant approach in recent years, with various architectures used to learn from high-dimensional data. The most common technique, involving CNNs, appeared in 55 (59.8%) studies. This was often complemented by RNNs in 14 (15.2%) studies, and more recently, transformers and attention-based models in 11 (12%) studies.

CNNs were commonly applied to video data or spatially structured kinematic sequences. A primary application was in object detection and segmentation, which was central to 16 (17.4%) studies. For instance, Nakajima et al [36] used a CNN-based semantic segmentation model to estimate the dissected tissue area from laparoscopic colorectal videos, allowing a phase-specific efficiency metric that correlated with surgical expertise. Beyond segmentation, CNNs were used for end-to-end skill assessment. For example, Jian et al [69] implemented a multitask CNN framework for video-based surgical skill assessment, jointly learning to predict gesture classes and global skill labels. Similarly, Tanin et al [30] used an ensemble of 2D and 3D CNNs to assess skill across cataract

surgery phases using only endoscopic video, demonstrating the feasibility of sensor-free performance evaluation. Some work focused on temporal convolutions, such as Menegozzo et al [88], who applied a time-delay neural network to robotic kinematic data for real-time gesture segmentation during knot tying, showcasing the early use of temporal convolutional models for skill-related sequence modeling. Yanik et al [101] further advanced this by developing a deep learning pipeline to track surgical skill acquisition over time using only laparoscopic training videos. Their model combined 3D CNNs with temporal segment networks to classify skill levels across repeated training sessions, capturing performance trajectories without relying on external sensors. The approach demonstrated strong accuracy in distinguishing novice from experienced participants and offered a scalable solution for monitoring skill progression in simulation-based curricula.

For sequential modeling, 14 (15.2%) studies used RNN variants, such as LSTM networks and gated recurrent units. These models processed frame-level or kinematic data over time to detect transitions between surgical gestures or estimate proficiency [5,32]. However, they often required fixed-length inputs and struggled with long-range dependencies.

To address these limitations, hybrid architectures such as long-term recurrent convolutional networks have gained traction by combining spatial and temporal representations more effectively. Recent implementations leverage variations such as convolutional long short-term memory (ConvLSTM), temporal pooling, and 2-stream fusion to improve fine-grained subphase classification and enable skill assessment. These models not only improve accuracy but also provide task-aware metrics such as phase-wise confidence or timing, offering interpretable insights into surgical performance. While such architectures perform well even in challenging video-only setups, they often remain constrained by unimodal inputs and require careful handling of visually similar transitions [78].

Reflecting a recent trend in sequence modeling, 11 (12%) studies introduced transformer-based models to leverage their strength in capturing long-range dependencies. For example, Zuluaga et al [3] applied a video transformer network for real-time procedural annotation during robotic-assisted prostatectomy, demonstrating high alignment with expert labels and strong temporal generalization. Zhang et al [9] used a continuous wavelet transform-vision transformer (ViT) architecture combining continuous wavelet transforms and vision transformers to assess suturing skill, achieving over 90% classification accuracy on the JIGSAWS dataset. Compared to RNNs, transformers offered better parallelization and interpretability through attention weights, making them attractive for real-time surgical guidance applications.

### ***Self-Supervised and Domain Adaptation Methods***

As labeling surgical data is time-consuming and often subjective, 7 (7.6%) studies explored self-supervised, semisupervised, or transfer learning strategies to reduce dependence on annotated datasets. These methods include contrastive learning to differentiate between samples [50], reinforcement learning for generating optimal trajectories [29], and pseudolabeling to leverage unlabeled data [16].

Several studies focused on domain adaptation to improve generalizability across different surgical tasks. For example, Wang et al [16] developed a domain-adaptive model that used uncertainty-aware self-supervised learning to generate pseudolabels for an unlabeled surgical task. This allowed the model to generalize from the JIGSAWS dataset to a novel VR simulation without requiring new annotations. Similarly, Alkadri et al [10] improved skill classification accuracy by applying transfer learning from a model trained on a similar surgical task, demonstrating its effectiveness on small datasets.

In a different approach, Anastasiou et al [50] applied contrastive regression within a transformer architecture, where a surgical video was scored based on its deviation from an expert reference performance, providing a fine-grained and interpretable assessment.

### ***Multimodal and Graph-Based Approaches***

To capture the complexity of surgical actions, 12 (13%) studies integrated multiple data streams—such as video, kinematics, and biosignals—within unified deep learning architectures [45,59,84]. These multimodal systems aim to improve robustness by fusing complementary performance perspectives, with techniques ranging from simple feature concatenation to more sophisticated attention and graph-based methods. A common strategy involves using attention mechanisms to dynamically weigh information; for instance, Amsterdam et al [84] developed a multimodal attention TCN that dynamically fuses video and kinematic features, allowing the model to rely more on one modality when the other is compromised (eg, by visual occlusion).

A prominent direction in this space is the use of graph-based learning for structural alignment across modalities [23,86]. Long et al [86] designed a relational graph network that jointly processes visual and kinematic embeddings to enhance gesture recognition. Their model treats video, left-hand kinematics, and right-hand kinematics as distinct nodes in a graph, refining their representations through relation-aware message passing that models concepts such as “vision-to-motion” and “in-between-motions” [86]. In a different application, Liu et al [23] introduced a visual-kinematics graph learning framework focused specifically on instrument tip segmentation. Their model constructs directed graphs from both image features and rendered kinematic silhouettes, using graph convolutional networks (GCNs) and a node-wise contrastive loss to enable procedure-agnostic segmentation [23].

These architectural strategies reflect a broader shift toward models that internalize surgical scene structure. Graph-based representations capture spatial relationships between instrument components or modalities [23,86], while attention mechanisms dynamically filter relevant signals across time [84]. Compared to early fusion techniques that rigidly concatenate inputs, these flexible approaches are better suited to cope with occlusions, variable motion dynamics, and visual noise in clinical environments [84,86].

## Data Modalities and Multimodal Fusion

### Overview

An analysis of the 92 included studies reveals 3 primary data modalities used for surgical skill assessment. The most common approach, found in 58 (63%) studies, was the use of video-based methods that analyze endoscopic or external camera footage. This was followed by 18 (19.6%) studies that relied on kinematic-only data from robotic systems or sensors. Finally, 12 (13%) studies used multimodal approaches, fusing 2 or more data streams to create more robust models.

### Kinematic-Only Studies

Kinematic data—comprising tool tip positions, velocities, forces, and trajectory-derived metrics—was a foundational modality, particularly in studies using robotic systems or standardized datasets such as JIGSAWS [32,72]. Early studies in this area focused on handcrafted features such as motion smoothness and jerk [63]. More recent work has applied deep learning directly to kinematic sequences; for example, Anh et al [32] compared 9 different feature extraction techniques, including CNN and LSTM, for evaluating skills in near real time from robotic motion data.

Building on the pursuit of interpretable motion features, Shayan et al [37] applied functional data analysis to hand rotation signals captured during open suturing. Using functional principal component analysis and functional ANOVA, they extracted temporally structured features that significantly distinguished expert from novice performance. This approach highlights the potential of statistical modeling for capturing subtle skill differences while preserving interpretability.

Soleymani et al [48] introduced an information-theoretical framework to quantify bimanual coordination through mutual information and dynamic time warping between hand trajectories. By analyzing peg transfer tasks, they demonstrated that expert surgeons exhibited higher levels of interhand coordination, reinforcing the value of motion coupling as a skill indicator. Their method adds a relational perspective to kinematic analysis and aligns with the broader trend toward interpretable and quantifiable skill metrics.

### Video-Based Methods

Video-based methods use endoscopic or external surgical footage to extract visual features for gesture recognition, procedural phase segmentation, or skill evaluation. These approaches typically rely on CNNs [5,54] to extract spatial patterns from frames, or on temporal models such as 3D-CNNs and transformers [3,73] to handle motion and sequence information. In particular, transformer-based models have gained traction due to their ability to capture long-range dependencies and provide interpretable attention maps, which are useful for real-time applications and explainability. Recent work by Guo et al [41] introduced a multitask transformer (multitask vision transformer [MT-ViT]) capable of both phase recognition and skill assessment from streaming long-form surgical videos. Their model handled real-time inference over continuous video using temporally local attention blocks, achieving competitive accuracy across both classification and regression tasks.

One important axis of variation is the application setting. In simulation-based environments, models often benefit from standardized camera views and task repetition. Liu et al [23] integrated video features with motion data using a spatiotemporal attention model, enhancing gesture recognition in a bench-top robotic setup. Their study also noted that performance gains were partially attributed to clean and well-labeled synthetic datasets, raising concerns about generalizability to more variable clinical data. In a similar direction, Erlich-Feingold et al [34] conducted a pilot study using deep learning to classify surgeon skill levels from video recordings of laparoscopic simulations. Their model, trained on simulated tasks involving novice and expert participants, achieved high agreement with human raters, particularly in identifying novice-level performance. While limited to a simulation setting and video-only input, this work reinforces the role of AI in formative assessment and early-stage skill differentiation, highlighting its practical relevance for surgical training programs. Another approach to simulation-based feedback involves extracting interpretable kinematic proxies from video alone. For instance, recent work estimated hand roll angles in suturing tasks using CNN-based vision models trained on hand-segmented frames [95]. The predicted “number of rolls”—a well-established metric for suturing fluency—successfully distinguished expertise levels in simulator environments, offering a low-cost and contactless method of skill assessment without requiring wearable sensors. This direction highlights how proxy motion cues derived from video can support scalable assessment pipelines when direct kinematics are unavailable.

In contrast, clinical studies often work with noisier, lower-frame-rate data and fewer labeled samples. Saricilar et al [54] trained CNNs on laparoscopic gastrectomy recordings, showing that even small datasets can be useful when phase boundaries are well annotated. They emphasized that transfer learning from simulation-trained models significantly boosted performance in data-scarce clinical environments. Zuluaga et al [3] demonstrated real-time procedural annotation using transformer models on robotic-assisted urological procedures, bridging the gap between offline learning and intraoperative support. Recent work has also leveraged deep action recognition models to directly link surgical step classification with surgeon proficiency. Yen et al [80] trained a model to recognize key actions—such as dissecting, exposing, and cutting—in laparoscopic cholecystectomy, showing that the predicted action patterns correlated significantly with GOALS-based competency scores. Their annotated video dataset, created from real clinical procedures, adds to the growing pool of surgical video benchmarks designed for both action understanding and skill assessment.

Following this trend of direct video-to-score pipelines, Yeh et al [39] developed PhacoTrainer, an AI tool trained on annotated cataract surgeries that achieved high agreement with expert Objective Structured Assessment of Cataract Surgical Skill ratings. Its design focused on efficiency and skill differentiation in ophthalmic microsurgery, reinforcing the feasibility of end-to-end assessment in standardized clinical workflows. Tanin et al [30] proposed an ensemble of 2D and 3D convolutional

networks to classify surgical skill levels in cataract surgery using raw video data. Their architecture combined time-distributed 2D CNNs with LSTM layers and a parallel 3D CNN stream, enabling skill prediction across multiple phases without requiring external sensors or tool annotations. This study highlights the feasibility of pure vision-based assessment in ophthalmic microsurgery, although its generalizability to other surgical contexts remains to be validated.

Similar efforts in microsurgical contexts have explored long-term recurrent convolutional network-based architectures for both phase recognition and skill assessment. By analyzing short video segments using hybrid recurrent-convolutional models, these methods can estimate both the likelihood of a subphase and the confidence or duration associated with its execution. This dual-metric evaluation enables more nuanced feedback—distinguishing expert control from novice hesitation—even in the absence of kinematic or tool-tracking data. Such techniques show particular promise for vision-only feedback systems in training environments, though their robustness under varying visual conditions remains an open question [78].

Beyond standard frame-level classification, several studies investigated specialized visual features to improve robustness in surgical settings. For example, sparse optical flow has been used to derive minimal yet informative motion descriptors from video sequences [64]. Others have proposed evaluating surgeon performance based on field-of-view clarity as a surrogate for procedural control and efficiency [77]. Pose estimation techniques—derived from articulated tool tracking or skeleton inference—were also applied to augment gesture classification in video-only settings [97], particularly where direct kinematic data were unavailable. Building on this, Fukuta et al [94] used DeepLabCut-based pose estimation in a structured simulator environment, tracking laparoscopic forceps trajectories using top-down video. Unlike Elek et al [97], who focused on skeleton-based gesture inference, Fukuta et al's [94] approach emphasizes fine-grained tool motion tracking for skill validation and feedback. Their system demonstrates the promise of simulation-calibrated vision-only methods, especially when domain constraints (eg, lighting or tool occlusion) can be controlled to improve tracking stability.

Moreover, deep segmentation models have been explored to extract fine-grained surgical signals from video, such as vessel displacement during microvascular suturing [100]. Tang et al [100] also noted that segmentation-driven features enabled interpretable feedback, allowing for localized performance assessment rather than coarse global scores. These approaches suggest that video alone, when processed with task-specific priors, can provide insight into subtle aspects of surgical technique.

While video offers richer contextual cues—such as tool-tissue interaction and field visibility—it introduces challenges including variable lighting, occlusion, and camera motion. As a result, several studies combine video with additional modalities such as kinematics or pose estimation to stabilize predictions and reduce uncertainty, particularly in real-world surgical environments.

## Multimodal Fusion

Multimodal approaches combine distinct data types—such as video, kinematics, force, and electroencephalography—to provide a more comprehensive representation of surgical performance. These methods aim to leverage the complementary nature of different modalities to improve classification accuracy, interpretability, and resilience in real-world scenarios.

Fusion strategies typically fall into 3 categories: early fusion (input-level concatenation), late fusion (decision-level), and intermediate fusion using attention or graph-based mechanisms. Luongo et al [5] demonstrated that integrating video and kinematics within an LSTM architecture yielded superior gesture segmentation compared to unimodal models. Similarly, Liu et al [23] implemented a 3D CNN with temporal attention fusion to combine spatiotemporal cues from video and motion signals, leading to more stable predictions in robotic tasks. However, they also reported latency-performance trade-offs, posing challenges for deployment in time-critical HAT applications. Time-synchronized multimodal systems have also been explored for procedural understanding. For example, Fawaz et al [104] used aligned video and kinematic streams to improve surgical phase segmentation accuracy on benchmark datasets.

Force sensing is another valuable modality in surgical simulation. Alkadri et al [4] used position and force metrics from a VR simulator to train a multilayer perceptron classifier, identifying key skill-related features through feature selection. Their work emphasized interpretable modeling and efficiency, though performance was demonstrated primarily on a single simulator task. In a follow-up study, the same group explored transfer learning with force and motion metrics in a hybrid VR/augmented reality (AR) simulator [10]. While transfer learning improved test accuracy, the study acknowledged the need to address modality shifts and environmental variability for broader generalization. Takács and Haidegger [68] developed an adaptive neuro-fuzzy inference system using force-based metrics across multiple training tasks, producing task-specific subscores for skill assessment. Multiview and multi-instrument tracking systems have been developed to capture fine-grained tool behavior across training and clinical environments, expanding the coverage and fidelity of motion-based assessment [70].

Some studies explored cognitive and biometric signals to expand beyond physical motion. Natheir et al [2] used electroencephalography signals during neurosurgical simulation to train a semisupervised artificial neural network (ANN), achieving high accuracy in differentiating skill levels. However, their approach exhibited limited generalizability outside neurosurgical contexts, indicating a need for task-agnostic models. Shafiei et al [45] combined electroencephalography and eye-tracking data from live animal surgeries and applied ensemble models to classify subtasks and skill levels based on both physiological and behavioral metrics. Their study emphasized cognitive load inference as a critical component for adaptive HAT feedback. In a related effort, Shafiei et al [28] demonstrated that electroencephalography signals alone could be used to classify surgical gestures, using features derived from functional brain networks and power spectral densities. This

work provided early evidence linking cognitive state and motor intention in surgical performance modeling. Gesture tracking using temporal alignment techniques—such as dynamic time warping—has also been proposed to accommodate variations in execution pace and movement timing [62].

Bkheet et al [83] used pose estimation combined with I3D features to inform gesture segmentation models in open surgery training videos, enabling interpretable skill proxy-based feedback on hand-tool interaction. Singh et al [47] captured multisensor motion data using electromagnetic tracking and inertial measurement units to extract fine motion metrics such as angular change and path length during open suturing. Despite providing high-fidelity motion data, Singh et al [47] reported calibration drift and alignment issues, posing challenges for long-duration tasks.

## Benchmarks and Validation Practices

### Overview

Benchmarking plays a central role in evaluating and comparing ML models for surgical skill assessment. However, the landscape is fragmented, with diverse datasets, inconsistent evaluation criteria, and limited external validation. This section examines the most commonly used datasets, assessment standards, evaluation metrics, and the methodological challenges that influence reproducibility and generalizability in the literature.

### Datasets Used

The JIGSAWS dataset was the most frequently used benchmark, providing synchronized video and kinematic data for 3 basic tasks (suturing, knot tying, and needle passing) [9,12,63,66,67,69,72,84]. However, JIGSAWS is limited by its low task diversity and controlled recording environment, making it less representative of real-world surgical variability. Despite these limitations, it enabled classification and regression models to evaluate motion-based surgical skill in controlled environments. Other datasets included Cholec80 [58,80,103], Cataract-101 [25,60], EndoVis [26,98], LapSig300 [14], and various simulated VR platforms [4,10,47,59,68]. These datasets vary widely in data types, incorporating video-only streams, kinematics, tool usage, or even cognitive signals such as electroencephalography and eye tracking [27,45]. The EndoVis challenge series provided a standardized benchmark for segmentation and pose estimation tasks but lacks longitudinal datasets necessary for tracking surgeon progression over time. LapSig300, in contrast, offers specialized tool vibration signals for laparoscopic task analysis, representing a more niche application. Simulator-based systems incorporating AI-driven validation and feedback loops have also been explored, contributing to benchmark creation and structured training platforms [94]. However, studies such as Alkadri et al [4,10] highlighted the domain gap challenges when transferring models trained on simulator data to clinical environments, underscoring the need for robust domain adaptation techniques. Additionally, electroencephalography and eye-tracking datasets remain limited by small sample sizes and are often restricted to animal or phantom studies, raising concerns about generalizability to human surgical performance. [27,45]

### Skill Assessment Standards

Multiple studies used structured scoring systems such as OSATS, GRS, or Endoscopic Surgical Skill Qualification System (ESSQS) [42,43,51,66,69,72]. However, studies such as Komatsu et al [43] reported significant interrater variability in ESSQS scoring, raising concerns about reproducibility in subjective assessments. Some adopted task-specific rubrics such as International Council of Ophthalmology's Ophthalmology Surgical Competency Assessment Rubric for ophthalmic microsurgery [75], or GOALS for laparoscopic tasks [49]. Labels were often derived from expert review, standardized rubrics, or thresholds applied to continuous scores (eg, binarizing OSATS/GRS for classification [71]). Hoffmann et al [42] introduced a hybrid approach by combining rubric-based scores with automated kinematic metrics, aiming to balance interpretability with objectivity. Several works used self-reported skill level [53] or pass/fail outcomes [27]. Pan et al [53] observed discrepancies between self-reported skill levels and objective performance metrics, highlighting limitations in self-assessment reliability. Others generated derived metrics from kinematic analysis [47,61,63]. Classification frameworks based on explicit gesture features have also been evaluated, providing a more interpretable bridge between visual motion and standardized rubrics [40]. Chen et al [40] further enhanced interpretability by incorporating attention-based visualization maps, supporting explainable AI in surgical skill assessment.

### Evaluation Metrics

Model performance was evaluated using accuracy,  $F_1$ -score, and area under the curve (AUC) for classification tasks [43,58,66,73,75]. Komatsu et al [43] further used confusion matrices to diagnose class-wise prediction errors, aiding in understanding model misclassifications. Fathollahi et al [58] highlighted that AUC and  $F_1$ -scores provided more reliable performance indicators in the presence of imbalanced class distributions, a common challenge in surgical datasets. Regression-based systems predicting continuous skill scores used metrics such as mean absolute error (MAE), root-mean-square error, and Spearman correlation [10,23,72]. Liu et al [23] emphasized that Spearman correlation better captured subjective skill ranking alignments compared to absolute error metrics such as MAE, making it particularly suitable for rubric-based assessment. Interrater agreement and expert consensus were also reported for label consistency [65,85], with Cohen  $\kappa$  or intraclass correlation coefficient values supporting annotation reliability [81]. Lavanchy et al [65] observed reduced interrater agreement in complex procedural annotations, underscoring the influence of task complexity on labeling consistency. You et al [81] advocated for standardized annotation protocols to enhance interrater reliability, particularly when aggregating multi-institutional datasets. In addition, motion analysis techniques applied to wet-laboratory laparoscopic training environments have shown measurable skill discrimination, supporting their inclusion as quantitative ground truths [56]. Recently, Singh et al [46] introduced the Smoothness of Force metric, derived from time- and frequency-domain analysis of force signals during simulated tissue handling. Their results showed strong skill-level separation, suggesting that signal-based smoothness measures

may serve as low-cost, interpretable indicators for assessing manipulation quality. However, Ebina et al [56] cautioned that such metrics are context-sensitive, limiting their generalizability beyond specific task types.

### **Challenges in Benchmarking**

Despite growing datasets, benchmarking in surgical AI remains inconsistent. The size and diversity of available datasets are still limited, particularly in real-world procedures. Alkadri et al [10] noted that leave-one-user-out (LOUO) validation becomes challenging when user-specific data is scarce, exacerbating class imbalance issues. Labels may be subjective or derived from inconsistent evaluation standards, leading to variability in reported outcomes. Validation strategies vary, with LOUO and cross-validation applied unevenly across studies [10,16,24,73]. Wang et al [16] emphasized that inconsistent reporting of model inputs and preprocessing pipelines further hinders reproducibility across studies. Only a few works have tested models across institutions, tasks, or input modalities [3,66,85], which limits generalization. Zuluaga et al [3] identified mismatches in input modality characteristics as a major barrier to generalizing models trained on homogeneous datasets. Takeuchi et al [85] further highlighted interinstitutional annotation drift, which complicates benchmarking consistency. Early studies also underscored the role of sensor-rich laparoscopic systems in capturing workflow states and expertise levels, laying the groundwork for modern validation strategies [76]. Improved benchmarking will require harmonized rubrics, open multicenter datasets, and transparent reporting of model inputs and validation logic. Alternative benchmarking tools, such as fuzzy-logic scoring engines for simulation-based peg transfer tasks, have been proposed to model uncertainty and provide flexible scoring interpretations [57,92]. Beyond uncertainty modeling, Fathabadi et al [92] demonstrated that fuzzy scoring systems enhanced interpretability, particularly in assessing novice-level performance with more nuanced feedback.

## **From Assessment to Decision-Making: Toward HAT**

### **Overview**

While traditional research in surgical AI has focused on post hoc skill assessment, a growing number of studies are moving toward systems that support real-time interaction, adapt to users, and contribute to intelligent surgical guidance. This transition forms the foundation for HAT, in which AI systems not only evaluate but also actively collaborate with surgeons.

### **Real-Time and Feedback-Capable Models**

Several studies developed models optimized for real-time inference and feedback. Anh et al [32] evaluated lightweight deep learning pipelines that prioritize low-latency inference, making them suitable for looped skill-feedback systems, though not without trade-offs in accuracy. Recent transformer-based methods have expanded this space by enabling on-the-fly labeling and procedural awareness. For instance, Zuluaga et al [3] deployed a transformer for robotic-assisted prostatectomies that delivered real-time procedural annotations, but reported difficulties in generalizing to surgical variations not seen during training. In a similar direction, Guo et al [41] introduced a

multitask transformer that performs both phase segmentation and skill prediction over continuous video streams, supporting feedback that evolves with procedural progression. While these systems focus on task structure and performance cues, others have emphasized anatomy-aware guidance. Ryu et al [90] trained a deep learning model to identify nerves in laparoscopic colorectal video, enabling real-time overlays that support both intraoperative navigation and surgical education—especially critical in high-risk regions where visual precision informs safety. Similarly, Madani et al [93] used semantic segmentation to distinguish anatomical zones and hazard boundaries in cholecystectomy, offering interpretable heatmaps aligned with expert visual reasoning.

Khan et al [91] extended anatomy-aware AI to the neurosurgical domain by developing a mask region-based convolutional neural network-based model for recognizing critical structures during endoscopic pituitary surgery. Trained on a large dataset of over 2 million annotated video frames from 200 clinical procedures, their model achieved a mean average precision exceeding 0.8 across multiple anatomical labels, including the optic nerve and carotid artery. The study sets a high benchmark for anatomical segmentation quality and reinforces the role of AI in intraoperative safety and education. Although not yet deployed in real time, the scale, annotation rigor, and surgical relevance position it as a promising system for operative decision support.

Building on these anatomy-aware segmentation systems, Lee et al [89] developed a model to identify dissection planes during robotic mastectomy. Using annotated video frames extracted at 1-second intervals from 10 procedures, their deep learning pipeline produced accurate and interpretable visual overlays, achieving a mean Dice similarity coefficient of 0.78. Although limited by dataset size and the lack of real-time deployment, the system emphasizes clinically meaningful labels and frame-level precision, presenting a viable direction for embedding AI-based visual guidance into surgical education platforms. This aligns with the broader shift toward using intraoperative video not only for assessment, but also for training-aware assistive decision support. These approaches illustrate how AI can contribute to decision support without requiring gesture classification. Outside of vision-based overlays, motion-tracking systems have also advanced, such as Nwosu et al [15], who built a drill-tracking pipeline for otologic surgery that outputs motion smoothness and trajectory metrics with minimal latency, demonstrating practical utility in feedback-driven training scenarios.

These systems focus not only on accuracy, but also on speed and interpretability, critical components for intraoperative utility. Some models provided low-latency outputs by reducing input complexity or model depth [32]. Agarwal et al [13] used fine-tuned spatiotemporal backbones and a calibrated multistage TCN model to perform frame-level gesture segmentation on untrimmed surgical videos, achieving high segmental accuracy and edit distance scores. While not real-time, their approach demonstrates the importance of temporally precise modeling for gesture-aware feedback systems. Others aimed to balance latency and robustness by limiting model complexity or reducing input dimensionality [63]. Hasani et al [63] demonstrated that

input dimensionality reduction was effective in enhancing model robustness, especially in noisy surgical environments.

### ***Adaptive Systems and Sequential Decision-Making***

Beyond inference speed, a second group of studies emphasized adaptability to different users, tasks, or conditions. These models incorporated temporal memory or personal context to adjust feedback and scoring dynamically.

Alkadri et al [10] trained ANN models on force and motion data in VR/AR simulators and applied transfer learning across procedures, indicating cross-task adaptability. However, they also acknowledged limitations due to domain shifts between simulated and clinical environments, which constrained full adaptation. Natheir et al [2] used semisupervised learning on electroencephalography and performance metrics to personalize skill estimation without full supervision. Their approach demonstrated the effectiveness of low-labeled data strategies in adapting models to individual users. Shafiei et al [45] integrated electroencephalography and eye-tracking in live robotic setups to detect workload and subtask transitions in real time. This integration improved system responsiveness to user cognitive states, enhancing its suitability for HAT scenarios.

Sequential decision-making models—such as TCNs or attention-based systems—were explored for gesture segmentation and phase tracking [9,23]. Liu et al [23] showed that temporal attention mechanisms significantly improved sensitivity to gesture transitions, improving the accuracy of real-time feedback. In a similar multitask framework, Guo et al [41] designed MT-ViT to output both phase segmentation and continuous skill scoring from video streams. The model's sequential attention over long video segments allows it to track procedural progression while simultaneously inferring performance, supporting time-aware feedback across tasks. Sato et al [79] proposed a rule-based pipeline that extracts temporal dissection and exposure metrics by parsing instrument activation from the da Vinci surgical interface. Although their method does not involve learning-based adaptation, the extracted phase durations showed strong correlations with surgeon expertise, reinforcing the role of coarse temporal structure in modeling surgical performance. RNN-based architectures have also been applied in multitask settings, where systems simultaneously predict gesture class and procedural progress [87]. Amsterdam et al [87] further highlighted that such sequential outputs from these models enhance interpretability, offering users clearer insights into procedural state and system reasoning. These architectures allow systems to evolve predictions over time, a necessary feature for interactive and supportive autonomy.

Sequential reasoning has gained renewed interest with the application of transformer-based architectures tailored for surgical workflows. Building on the idea that decision-making in surgery often involves causally dependent gestures, recent models have introduced reasoning frameworks that emulate step-by-step thought processes. One such approach leverages gesture context to improve downstream error detection, highlighting the value of temporally grounded prompting within robotic video streams [99]. This design enables the system to anticipate and interpret surgical deviations by chaining likely action sequences—an advance that improves robustness and

interpretability compared to traditional frame-based classifiers. Such methods suggest a move toward proactive AI systems that not only assess skill retrospectively but also reason through procedural intent in near real time.

### ***Integration With Surgical Training and Guidance***

Some works directly positioned their models within existing surgical education or robotic platforms. For example, Singh et al [47] used wearable sensors to provide motion quality metrics in open suturing, enabling dynamic coaching in physical training setups. Tonetti et al [49] designed a feedback system that maps model errors to OSATS subdomains, mirroring the structure of expert verbal guidance. Their rubric-aligned feedback approach notably enhanced trainee engagement by providing familiar, structured performance insights. Amsterdam et al [84] aligned transformer attention maps with human-defined gesture cues, facilitating interpretable handover of perceptual roles. This interpretability was shown to reduce cognitive load for trainees, making AI-assisted feedback more accessible. End-to-end video-based feedback systems have also been proposed for formative and summative evaluation within surgical curricula [55]. Yanik et al emphasized the scalability of such systems, supporting broader curricular integration.

In simulation environments, force-enabled feedback [4,68], visual annotations [3], or neural attention markers [23] were incorporated into intelligent tutoring systems or robotic task interfaces. Beyond technical feedback, some systems have prioritized scalability and accessibility. Bhatia et al [82] demonstrated the deployment of an AI-driven training system for low-resource environments, aiming to address instructor shortages in low- and middle-income countries and support foundational surgical education. Cruz et al [33] developed a scalable video-based AI platform that used YOLOv4 for laparoscopic instrument detection and feedback generation. Their model aligned performance scores with standardized surgical training frameworks and successfully replicated expert ratings across simulation tasks, showing the practical integration of real-time object detection in education pipelines. Liu et al reported that attention markers enhanced the granularity of localized feedback, improving the precision of corrective cues. Neurosurgical simulation has also benefited from AI-driven metric selection to track learning curves, helping to personalize instructional content and improve skill acquisition [102]. Ledwos et al [102] found that personalized metrics not only tailored instruction but also improved long-term skill retention. These developments mark early advances in shared control and multimodal feedback integration.

### ***Gaps in Current HAT Readiness***

Despite progress, few current systems achieve the robustness, adaptivity, and trust required for surgical HAT. Most studies rely on simulated tasks or preannotated data, limiting their ecological validity [12,66]. Generalization across users, procedures, and clinical setups remains underexplored, with only a handful testing multisite or cross-task robustness [10,16]. Wang et al [16] also reported inconsistencies in data pipelines and annotation protocols across clinical sites, further complicating generalization efforts. Recent clinical-phase models—for instance, in colorectal surgery—have shown that

deep learning can accurately recognize procedural stages in real time. However, such tools remain scarce and largely confined to specialized domains [44]. Nakajima et al [44] emphasized that even when technically successful, these models face significant challenges in integrating with existing surgical workflows.

Interpretability is still treated as an afterthought in many deep models, although recent studies using SHAP [61] or attention visualization [84] have made promising strides. Yibulayimu et al [61] demonstrated improvements in local interpretability through SHAP summary and dependence plots and further implemented sample-by-sample force plots to provide real-time feedback during liposuction training. Similarly, Amsterdam et al's [84] attention maps facilitated gesture-level insights but did not address higher-order procedural reasoning. True teaming will require models to communicate confidence, defer to the human when uncertain, and offer justifications in a format that fits surgical workflows.

In addition, almost no system integrates patient-specific variables or anticipates clinical decisions, key elements for future decision augmentation. Bridging the gap from skill scoring to surgical collaboration will require not only better models but better integration with the social, cognitive, and procedural context of the operating room.

## Discussion

### Principal Findings

This scoping review mapped recent advances in AI-based surgical skill assessment, focusing on methods that support HAT. Our synthesis addresses the 3 RQs posed in the Introduction:

#### **RQ1: Adapting HAT Models for Real-Time Guidance and Personalized Feedback**

The literature reveals a growing emphasis on real-time inference and feedback (one of our primary trends). This transition is critical for HAT adoption. Lightweight models and transformer-based architectures have enabled systems capable of intraoperative gesture recognition and phase annotation. Furthermore, emerging efforts toward adaptive and interpretable models align directly with the need for personalized guidance. Studies using SHAP explanations, attention visualizations, and personalized feedback pipelines represent initial steps toward systems that can communicate their reasoning and adjust to individual surgeon profiles. However, true real-time HAT remains limited: these systems are largely task-specific and constrained by generalization challenges across users, procedures, and clinical settings.

#### **RQ2: ML Methods and the Enhancement of Multimodal Data**

Our analysis confirms an evolution from unimodal to multimodal data integration, which serves to enhance classification and proficiency assessment. Multimodal approaches, combining kinematics, video, force, and biosignals, have been shown to enhance classification accuracy and robustness by leveraging complementary information sources.

Techniques such as attention-based fusion and GCNs reflect this trend toward richer data representations, allowing models to process diverse data streams effectively. This validates the use of multimodal data as a key pathway to enhancing surgical movement classification for real-time applications.

#### **RQ3: Benchmarks and Comparison With Expert Evaluations**

The studies reviewed show that AI-driven skill assessment relies on a fragmented ecosystem of benchmarks. While standardized tools such as OSATS and GRS are widely used as ground truth for comparison, inconsistencies in evaluation criteria and validation practices remain a critical barrier. Validation strategies vary (eg, LOUO vs cross-validation), leading to variability in reported outcomes. Benchmarking is often constrained by simulation-based datasets (eg, JIGSAWS), which limits their ecological validity when comparing results to real-world expert evaluations. We conclude that standardized, harmonized protocols are essential to ensure the reproducibility and clinical relevance of AI assessments.

Together, these developments lay the foundation for transitioning from isolated assessment tools to intelligent surgical teammates. However, critical gaps remain before true HAT can be realized in surgical practice.

### Limitations and Research Gaps in Current Literature

Despite notable advancements, current AI-based surgical skill assessment methods face several limitations that hinder their readiness for HAT in real-world surgical environments. Our translational readiness analysis (Table 1) reveals that these limitations are not isolated shortcomings but stem from interconnected structural and methodological root causes.

First, ecological validity remains a significant concern. Most studies rely on simulation-based datasets or controlled laboratory environments, which do not fully capture the variability, complexity, and unpredictability of actual surgical procedures. Our TRL analysis confirms this pattern: 77.2% (71/92) of studies remain at TRL 3, indicating proof-of-concept validation on limited or benchmark data, while only 2.2% (2/92) have reached TRL 5 or above with clinical or cross-institutional testing. This concentration at low TRL levels reflects a fundamental bottleneck in the field: the scarcity of large-scale, annotated clinical datasets. Collecting and labeling intraoperative data requires expert time, institutional ethics approval, and standardized annotation protocols—resources that remain unevenly distributed across research groups [12,66]. This data scarcity also increases the risk of overfitting, particularly for deep learning architectures with high parameter counts trained on small, homogeneous datasets such as JIGSAWS (39 trials from 8 participants). High reported accuracies on such benchmarks may therefore reflect memorization of dataset-specific patterns rather than genuine learning of transferable surgical skill features.

Second, generalization across users, tasks, and clinical settings is underexplored. Our analysis shows that 38% (35/92) of studies lack any form of external validation, and only 15.2% (14/92) have conducted cross-site testing. The root causes are twofold: (1) pipeline inconsistencies and annotation discrepancies across

institutions complicate robust evaluation [10,16,85], and (2) domain shift between simulated and clinical environments degrades model performance when training and deployment conditions differ. These factors create a cycle in which models achieve high accuracy on narrow benchmarks but fail to transfer, reinforcing reliance on familiar datasets.

Third, interpretability continues to be treated as an ancillary feature rather than a core design principle. Although 71.7% (66/92) of studies report some form of built-in interpretability, this high proportion is largely driven by inherently interpretable architectures (eg, attention mechanisms or feature importance rankings) rather than deliberate explainability design for clinical end users. Only 5.4% (5/92) used dedicated post hoc explanation methods such as SHAP [61] or Grad-CAM [84]. The underlying cause is a disconnect between the ML research community's notion of interpretability (model-centric) and what surgeons need (decision-centric): actionable, real-time justifications embedded in clinical workflows rather than offline feature attribution maps.

Fourth, adaptivity to individual users is largely absent. Our analysis reveals that 91.3% (84/92) of studies use static models with no user-specific adjustment, and only 6.5% (6/92) demonstrate any form of adaptive behavior. This gap stems from a data scarcity problem at the individual level: adapting to a specific surgeon requires sufficient per-user data, which is rarely available in current datasets. Furthermore, techniques such as online learning and meta-learning that could enable few-shot personalization remain largely unexplored in the surgical domain, partly due to concerns about catastrophic forgetting and the safety implications of models that update during procedures.

Several technical limitations also arise from the nature of the data itself. Vision-based datasets, while attractive due to their ease of collection and minimal interference with the surgical workflow, present known challenges. Fukuta et al [94] reported issues such as key point occlusion, weak contrast on metallic tools, background variation, and motion blur—all of which undermine pose estimation performance, especially when moving from simulators to real-world settings. These issues are often masked in standardized setups but remain critical for deployment.

Finally, the near-total absence of system-level autonomy further underscores the field's distance from HAT. Our mapping to the Yang et al [6] autonomy framework reveals that 96.7% (89/92) of studies operate at level 0—providing no autonomous functionality beyond classification output—while only 3 studies reach level 1 with basic assistive capabilities. No study in our corpus demonstrates level 2 or higher autonomy. This indicates that current research remains firmly in the assessment paradigm, with the transition to collaborative teaming requiring not only better models but fundamentally different system architectures that integrate perception, reasoning, and action within the surgical workflow.

These limitations point not only to the current bottlenecks but also to concrete directions where further development is needed. The following section outlines key research priorities that can support the transition toward HAT-readiness in surgery.

## Research Opportunities for HAT Readiness

The limitations identified in the preceding section point to 6 concrete research priorities that can accelerate the transition from skill assessment toward functional HAT in surgery. These priorities represent the authors' forward-looking recommendations, informed by the gaps identified in this review, rather than direct findings from the 92 included studies.

First, addressing the data bottleneck that confines 77.2% (71/92) of studies to TRL 3 (“Limitations and Research Gaps in Current Literature” section) will require coordinated, multicenter dataset initiatives. Federated learning offers a promising pathway, enabling institutions to collaboratively train models on distributed clinical data without sharing sensitive recordings, thereby sidestepping many privacy and ethics barriers. In parallel, domain adaptation and transfer learning techniques—such as adversarial domain alignment and curriculum-based fine-tuning—should be systematically investigated to bridge the simulation-to-clinical gap. Synthetic data augmentation through procedurally generated surgical scenes or physics-based instrument simulations could further expand training distributions without additional clinical data collection.

Second, the near-absence of user adaptivity (84/92, 91.3% of studies use static models; “Limitations and Research Gaps in Current Literature” section) demands dedicated investigation into personalization mechanisms. Concrete approaches include (1) meta-learning frameworks (eg, model-agnostic meta-learning) that learn transferable initialization parameters, enabling rapid few-shot adaptation to a new surgeon from minimal calibration trials; (2) online continual learning with replay buffers to update models incrementally as per-user data accumulates, while mitigating catastrophic forgetting through elastic weight consolidation or similar regularization; (3) hierarchical Bayesian models that maintain population-level priors while estimating surgeon-specific parameters; and (4) user-calibration protocols embedded in preoperative warm-up routines, where a short standardized task establishes the individual baseline against which intraoperative performance is measured. The associated challenges—data scarcity at the individual level, safety validation of adaptive models, and the computational overhead of real-time updates—represent important open problems that the field has yet to systematically address.

Third, closing the interpretability gap between model-centric and decision-centric explainability (“Limitations and Research Gaps in Current Literature” section) requires a design-level shift. Rather than appending post hoc explanation tools to existing architectures, future systems should embed interpretability through inherently transparent modules—such as concept bottleneck layers that map internal representations to clinically meaningful features (eg, tissue tension, instrument angle, or proximity to critical structures). Uncertainty quantification via calibrated confidence estimates or conformal prediction can enable systems to communicate when they are unsure and defer to the surgeon, a prerequisite for trustworthy teaming. Co-design with surgical teams through participatory design studies will be essential to ensure that explanations are

not only technically faithful but also operationally useful within the time constraints and cognitive demands of the operating room.

Fourth, incorporating patient-specific context represents a key frontier for moving from generic assessment to individualized decision augmentation. Future models should integrate preoperative imaging, anatomical variation maps, and pathology-specific risk profiles as conditioning inputs, enabling the system to anticipate procedure-specific challenges and adjust its guidance accordingly. Predictive modeling of procedural trajectories—forecasting likely next steps and potential complications based on the current surgical state—would represent a qualitative leap from retrospective assessment toward anticipatory decision support aligned with HAT requirements.

Fifth, the fragmented benchmarking ecosystem identified in this review (“Limitations and Research Gaps in Current Literature” section; RQ3) calls for community-level standardization. Specific actions include developing open, multicenter benchmark suites that span diverse procedures, skill levels, and clinical environments; adopting harmonized evaluation protocols that specify validation strategy (eg, LOUO), reporting metrics, and statistical testing; and establishing shared annotation guidelines to reduce interinstitutional labeling drift. Organizations such as the Medical Image Computing and Computer Assisted Intervention community and surgical societies are well positioned to coordinate these efforts, following precedents set by challenges such as EndoVis and Cholec80.

Finally, advancing from state recognition to procedural policy learning is essential for decision support. While current AI can identify what gesture is occurring, the critical gap is the inability to recommend what should happen next. This requires learning the underlying policy or grammar of a surgical task by abstracting low-level gestures into structured, step-based workflow representations. Sequential decision-making frameworks—including model-based reinforcement learning, inverse reinforcement learning from expert demonstrations, and probabilistic graphical models over gesture sequences—provide the architectural foundation for systems capable of modeling optimal procedural paths and providing guidance tailored to a surgeon’s proficiency. This direction represents perhaps the most ambitious yet transformative step toward genuine human-autonomy collaboration in surgery.

Addressing these priorities in a coordinated manner offers the opportunity to advance from isolated AI tools toward intelligent, adaptive surgical teammates capable of supporting decision-making, enhancing training, and improving patient outcomes.

### Translating AI Skill Assessment to HAT

The transition from AI-driven surgical skill assessment to true HAT represents not only a technical challenge but also a sociotechnical one. Effective deployment of HAT systems in clinical practice demands a holistic approach that integrates technological capabilities with human factors, workflow dynamics, regulatory requirements, and organizational considerations. Our finding that 84.8% (78/92) of studies discuss

clinical integration only briefly (Table 1) underscores how underexplored these translational dimensions remain.

A primary barrier is the requirement for real-time performance under strict safety constraints. Intraoperative AI systems must deliver inference within milliseconds to avoid disrupting surgical flow, yet must simultaneously maintain high reliability—false or delayed feedback during a critical procedural step could compromise patient safety. This creates an engineering tension between model complexity and latency that few current studies address: only 25% (23/92) of reviewed studies demonstrated real-time capability, and most of these were validated in controlled rather than live operating room conditions. Hardware constraints further complicate deployment, as operating rooms may lack the GPU infrastructure required for computationally intensive deep learning inference, necessitating either model compression techniques (eg, pruning, quantization, or knowledge distillation) or edge computing architectures optimized for surgical environments.

Seamless integration into existing surgical workflows represents another critical challenge. AI systems must be designed to complement, rather than disrupt, established practices. This requires careful consideration of how feedback is delivered (eg, visual overlays, auditory cues, or haptic signals), when it is delivered (avoiding information overload during high-cognitive-demand phases), and how autonomy levels are modulated based on context and user preferences. Human-centered design principles and codevelopment with surgical teams are essential to ensure usability, trust, and adoption. Importantly, the mode of feedback must be tailored to the surgical context: a real-time warning during a high-risk dissection step requires a fundamentally different interface design than a postoperative performance summary for training purposes.

Surgeon acceptance and trust represent perhaps the most underappreciated translational barrier. Surgeons are trained to exercise autonomous clinical judgment, and introducing an AI system that offers guidance or correction during procedures raises concerns about professional autonomy, overreliance on technology, and the potential for skill degradation over time. Building trust requires not only technical reliability but also transparency in system behavior: AI must communicate its confidence levels, provide clear justifications for its recommendations, and defer to human expertise when faced with uncertainty. Participatory design—involving surgeons as codevelopers rather than end users—and graduated exposure through simulation-based familiarization can help bridge the acceptance gap. Evidence from adjacent domains such as aviation crew resource management suggests that teaming is most effective when both human and autonomous agents have clearly defined, mutually understood roles.

Regulatory and medico-legal considerations present additional hurdles that the reviewed literature largely overlooks. AI systems that influence intraoperative decisions will require regulatory clearance through pathways such as the US Food and Drug Administration’s Software as a Medical Device framework or the EU Medical Device Regulation, both of which demand rigorous clinical validation, risk classification, and postmarket

surveillance. The question of liability is equally unresolved: if an AI-guided intervention leads to an adverse outcome, the allocation of responsibility among the surgeon, the institution, and the system developer remains legally ambiguous in most jurisdictions. These uncertainties create institutional risk aversion that may slow adoption even when technical readiness is achieved. Proactive engagement with regulatory bodies and the development of domain-specific guidelines for surgical AI will be necessary to establish clear pathways from prototype to clinical deployment.

Finally, iterative evaluation in real-world settings is necessary to refine HAT systems continuously. Pilot deployments, multi-institutional collaborations, and longitudinal studies will help identify practical barriers, optimize system performance, and ensure that AI teammates contribute meaningfully to surgical outcomes and team dynamics. Such evaluations should assess not only technical performance metrics but also human factors outcomes—including cognitive workload, situation awareness, team communication patterns, and long-term skill development trajectories—to ensure that AI integration enhances rather than undermines surgical team performance.

By addressing these translational challenges, AI systems can evolve from passive assessment tools into active, adaptive teammates that enhance surgical decision-making, training, and patient care.

## Conclusion

This scoping review successfully examined the landscape of AI-based methods for surgical skill assessment, focusing on their potential to enable HAT. Our synthesis provides clear answers to the 3 RQs that guided this review:

Regarding real-time HAT adaptation and guidance (RQ1), we conclude that current AI systems require significant adaptation

to transition from post hoc assessment to real-time teaming. This demands that future HAT models incorporate adaptive, personalized feedback mechanisms and address critical challenges in ecological validity and interpretability. Necessary adaptations include embedding uncertainty communication and ensuring seamless alignment of AI outputs with existing surgical workflows.

Regarding ML methods and multimodal data enhancement (RQ2), significant technical progress has been made, with deep learning becoming the dominant methodology ( $n=55$ , 59.8% of studies) for movement classification. Multimodal data fusion (combining video, kinematics, and biosignals), enhanced by architectural strategies such as attention mechanisms and GCNs, is essential for integrating these diverse data streams to improve the robustness and reliability of surgical movement classification for real-time feedback.

Regarding benchmarks and expert evaluations (RQ3), while standardized rubrics such as OSATS and GRS serve as the ground truth for expert evaluation, the AI landscape is characterized by inconsistent benchmarking protocols. The prevalent reliance on small, simulated datasets (eg, JIGSAWS) limits the ability to generalize results and accurately compare AI performance with established expert skill rankings in diverse clinical environments.

Looking ahead, advancing HAT in surgery requires not only technical innovation but also a foundation built on human-centered design, strict adherence to regulatory considerations, and iterative validation in real-world clinical settings. By addressing these multifaceted challenges, AI systems can evolve into intelligent surgical teammates that enhance decision-making, support skill development, and ultimately contribute to improved patient care.

## Acknowledgments

During the revision of this manuscript, a generative AI tool (Claude; Anthropic) was used to assist with grammar checking, organizing data, and entering data into LaTeX tables. The AI tool was not used during the original study design, literature search, screening, data extraction, or analysis. All AI-assisted content was critically reviewed, verified, and edited by the authors, who take full responsibility for the accuracy and integrity of the final manuscript.

## Funding

This work was supported by the National Sciences and Engineering Research Council of Canada (NSERC) through NSERC-RGPIN-2022-05438.

## Authors' Contributions

All authors contributed to the conception and design of the study. KB contributed to conceptualization, methodology, investigation, formal analysis, and writing—original draft. MSRC contributed to methodology, investigation, formal analysis, and writing—review and editing. TED contributed to conceptualization, methodology, formal analysis, validation, writing—review and editing, supervision, and project administration. All authors read and approved the final manuscript.

All authors read and approved the final manuscript.

## Conflicts of Interest

None declared.

## Multimedia Appendix 1

PRISMA-ScR checklist.

[\[DOCX File , 108 KB-Multimedia Appendix 1\]](#)

## Multimedia Appendix 2

Database search strategies.

[\[DOCX File , 22 KB-Multimedia Appendix 2\]](#)

## Multimedia Appendix 3

Full per-study assessment.

[\[DOCX File , 34 KB-Multimedia Appendix 3\]](#)

## References

1. Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg*. 1997;84(2):273-278. [doi: [10.1046/j.1365-2168.1997.02502.x](https://doi.org/10.1046/j.1365-2168.1997.02502.x)] [Medline: [9052454](#)]
2. Natheir S, Christie S, Yilmaz R, Winkler-Schwartz A, Bajunaid K, Sabbagh A. Utilizing artificial intelligence and electroencephalography to assess expertise on a simulated neurosurgical task. *Comput Biol Med*. 2023;152:106286. [FREE Full text] [Medline: [36502696](#)]
3. Zuluaga L, Rich JM, Gupta R, Pedraza A, Ucpinar B, Okhawere K, et al. AI-powered real-time annotations during urologic surgery: the future of training and quality metrics. *Urol Oncol*. 2024;42(3):57-66. [doi: [10.1016/j.urolonc.2023.11.002](https://doi.org/10.1016/j.urolonc.2023.11.002)] [Medline: [38142209](#)]
4. Alkadri S, Ledwos N, Mirchi N, Reich A, Yilmaz R, Driscoll M, et al. Utilizing a multilayer perceptron artificial neural network to assess a virtual reality surgical procedure. *Comput Biol Med*. 2021;136:104770. [FREE Full text] [doi: [10.1016/j.compbimed.2021.104770](https://doi.org/10.1016/j.compbimed.2021.104770)] [Medline: [34426170](#)]
5. Luongo F, Hakim R, Nguyen JH, Anandkumar A, Hung AJ. Deep learning-based computer vision to recognize and classify suturing gestures in robot-assisted surgery. *Surgery*. 2021;169(5):1240-1244. [FREE Full text] [doi: [10.1016/j.surg.2020.08.016](https://doi.org/10.1016/j.surg.2020.08.016)] [Medline: [32988620](#)]
6. Yang GZ, Cambias J, Cleary K, Daimler E, Drake J, Dupont PE, et al. Medical robotics-regulatory, ethical, and legal considerations for increasing levels of autonomy. *Sci Robot*. 2017;2(4):eaam8638. [doi: [10.1126/scirobotics.aam8638](https://doi.org/10.1126/scirobotics.aam8638)] [Medline: [33157870](#)]
7. Takács K, Lukács E, Levendovics R, Pekli D, Szijártó A, Haidegger T. Assessment of surgeons' stress levels with digital sensors during robot-assisted surgery: an experimental study. *Sensors (Basel)*. 2024;24(9):2915. [FREE Full text] [doi: [10.3390/s24092915](https://doi.org/10.3390/s24092915)] [Medline: [38733021](#)]
8. Yule S, Flin R, Paterson-Brown S, Maran N. Non-technical skills for surgeons in the operating room: a review of the literature. *Surgery*. 2006;139(2):140-149. [doi: [10.1016/j.surg.2005.06.017](https://doi.org/10.1016/j.surg.2005.06.017)]
9. Zhang Y, Weng Y, Wang B. CWT-ViT: a time-frequency representation and vision transformer-based framework for automated robotic surgical skill assessment. *Expert Syst Appl*. 2024;258:125064. [FREE Full text]
10. Alkadri S, Del Maestro RF, Driscoll M. Unveiling surgical expertise through machine learning in a novel VR/AR spinal simulator: a multilayered approach using transfer learning and connection weights analysis. *Comput Biol Med*. 2024;179:108809. [doi: [10.1016/j.compbimed.2024.108809](https://doi.org/10.1016/j.compbimed.2024.108809)] [Medline: [38944904](#)]
11. Gao Y, Vedula S, Reiley C, Ahmidi N, Varadarajan B, Lin HC, et al. The JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS): a surgical activity dataset for human motion modeling. In: *Modeling and Monitoring of Computer Assisted Interventions (M2CAI)*. Strasbourg, France. MICCAI workshop; 2014.
12. Soleymani A, Sadat AA, Yeganejou M, Dick S, Tavakoli M, Li X. Surgical skill evaluation from robot-assisted surgery recordings. 2021. Presented at: 2021 International Symposium on Medical Robotics (ISMR); 2021 November 17-19:1-6; Atlanta, GA, USA. [doi: [10.1109/ismr48346.2021.9661527](https://doi.org/10.1109/ismr48346.2021.9661527)]
13. Agarwal S, Pradeep C, Sinha N. Temporal surgical gesture segmentation and classification in multi-gesture robotic surgery using fine-tuned features and calibrated MS-TCN. 2022. Presented at: 2022 IEEE International Conference on Signal Processing and Communications (SPCOM); 2022 July 11-15:1-5; Bangalore, India. [doi: [10.1109/spcom55316.2022.9840779](https://doi.org/10.1109/spcom55316.2022.9840779)]
14. Kitaguchi D, Takeshita N, Matsuzaki H, Oda T, Watanabe M, Mori K, et al. Automated laparoscopic colorectal surgery workflow recognition using artificial intelligence: experimental research. *Int J Surg*. 2020;79:88-94. [FREE Full text] [doi: [10.1016/j.ijsu.2020.05.015](https://doi.org/10.1016/j.ijsu.2020.05.015)] [Medline: [32413503](#)]
15. Nwosu OI, Ota M, Xu LJ, Crowson MG. Automated real-time otologic drill motion analysis. *Laryngoscope*. 2025;135(2):836-839. [FREE Full text] [doi: [10.1002/lary.31795](https://doi.org/10.1002/lary.31795)] [Medline: [39319613](#)]
16. Wang Z, Mariani A, Mencias A, De ME, Fey AM. Uncertainty-aware self-supervised learning for cross-domain technical skill assessment in robot-assisted surgery. *IEEE Trans Med Robot Bionics*. 2023;5(2):301-311. [FREE Full text]
17. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol*. 2005;8(1):19-32. [FREE Full text]

18. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. *Implement Sci.* 2010;5:69. [FREE Full text] [doi: [10.1186/1748-5908-5-69](https://doi.org/10.1186/1748-5908-5-69)] [Medline: [20854677](https://pubmed.ncbi.nlm.nih.gov/20854677/)]
19. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med.* 2018;169(7):467-473. [FREE Full text] [doi: [10.7326/M18-0850](https://doi.org/10.7326/M18-0850)] [Medline: [30178033](https://pubmed.ncbi.nlm.nih.gov/30178033/)]
20. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ.* 2021;372:n71. [FREE Full text] [doi: [10.1136/bmj.n71](https://doi.org/10.1136/bmj.n71)] [Medline: [33782057](https://pubmed.ncbi.nlm.nih.gov/33782057/)]
21. Barati K. Protocol for: from skill assessment to surgical teammates: a scoping review of machine learning for human-autonomy teaming. OSF. 2026. URL: <https://doi.org/10.17605/OSF.IO/PQWS5> [accessed 2026-09-25]
22. Technology readiness levels. National Aeronautics and Space Administration. 2023. URL: <https://www.nasa.gov/directorates/somd/space-communications-navigation-program/technology-readiness-levels/>; [accessed 2026-06-25]
23. Liu J, Long Y, Chen K, Leung CH, Wang Z, Dou Q. Visual-kinematics graph learning for procedure-agnostic instrument tip segmentation in robotic surgeries. 2023. Presented at: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2023 September 2; Detroit, MI.
24. Ryder CY, Mott NM, Gross CL, Anidi C, Shigut L, Bidwell SS, et al. Using artificial intelligence to gauge competency on a novel laparoscopic training system. *J Surg Educ.* 2024;81(2):267-274. [doi: [10.1016/j.jsurg.2023.10.007](https://doi.org/10.1016/j.jsurg.2023.10.007)] [Medline: [38160118](https://pubmed.ncbi.nlm.nih.gov/38160118/)]
25. Kumar V, Tripathi V, Pant B, Alshamrani S, Dumka A, Gehlot A, et al. Hybrid spatiotemporal contrastive representation learning for content-based surgical video retrieval. *Electronics.* 2022;11(9):1353. [FREE Full text] [doi: [10.3390/electronics11091353](https://doi.org/10.3390/electronics11091353)]
26. Chen K, Du Y, You T, Islam M, Guo Z, Jin Y. LLM-assisted multi-teacher continual learning for visual question answering in robotic surgery. 2024. Presented at: 2024 IEEE International Conference on Robotics and Automation (ICRA); 2024 May 13-17; Yokohama, Japan. [doi: [10.1109/icra57147.2024.10610603](https://doi.org/10.1109/icra57147.2024.10610603)]
27. Subedi A, Rahul F, Dutta A, Norfleet J, Makled B, Intes XR, et al. A dilated causal convolutional model for surgical skill assessment using optical neuroimaging. 2022. Presented at: Proceedings of SPIE: Clinical and Translational Neurophotonics 2022; 2022 March 4; Cergy, France. URL: <https://doi.org/10.1117/12.2610190>
28. Shafiei SB, Durrani M, Jing Z, Mostowy M, Doherty P, Hussein AA, et al. Surgical hand gesture recognition utilizing electroencephalogram as input to the machine learning and network neuroscience algorithms. *Sensors (Basel).* 2021;21(5):1733. [FREE Full text] [doi: [10.3390/s21051733](https://doi.org/10.3390/s21051733)] [Medline: [33802372](https://pubmed.ncbi.nlm.nih.gov/33802372/)]
29. Long Y, Cao J, Deguet A, Taylor RH, Dou Q. Integrating artificial intelligence and augmented reality in robotic surgery: an initial dVRK study using a surgical education scenario. 2022. Presented at: 2022 International Symposium on Medical Robotics (ISMR); 2022 April 13-15:1-8; Atlanta, GA. [doi: [10.1109/ismr48347.2022.9807505](https://doi.org/10.1109/ismr48347.2022.9807505)]
30. Tanin U, Duimering A, Law C, Ruzicki J, Luna G, Holden M. Performance evaluation in cataract surgery with an ensemble of 2D-3D convolutional neural networks. *Healthc Technol Lett.* 2024;11(2-3):189-195. [FREE Full text] [doi: [10.1049/htl2.12078](https://doi.org/10.1049/htl2.12078)] [Medline: [38638495](https://pubmed.ncbi.nlm.nih.gov/38638495/)]
31. Ahmadi MJ, Allahkaram MS, Rashvand A, Lotfi F, Abdi P, Motaharifar M, et al. ARAS-Farabi experimental framework for skill assessment in capsulorhexis surgery. 2021. Presented at: 2021 9th RSI International Conference on Robotics and Mechatronics (ICRoM); 2021 November 17-19; Tehran, Iran. [doi: [10.1109/icrom54204.2021.9663494](https://doi.org/10.1109/icrom54204.2021.9663494)]
32. Anh NX, Nataraja RM, Chauhan S. Towards near real-time assessment of surgical skills: a comparison of feature extraction techniques. *Comput Methods Programs Biomed.* 2020;187:105234. [doi: [10.1016/j.cmpb.2019.105234](https://doi.org/10.1016/j.cmpb.2019.105234)] [Medline: [31794913](https://pubmed.ncbi.nlm.nih.gov/31794913/)]
33. Cruz E, Selman R, Figueroa U, Belmar F, Jarry C, Sanhueza D, et al. A scalable solution: effective AI implementation in laparoscopic simulation training assessments. *Global Surg Educ.* 2025;4(1):46. [FREE Full text]
34. Erlich-Feingold O, Anteby R, Klang E, Soffer S, Cordoba M, Nachmany I, et al. Artificial intelligence classifies surgical technical skills in simulated laparoscopy: a pilot study. *Surg Endosc.* 2025;39(6):3592-3599. [doi: [10.1007/s00464-025-11715-3](https://doi.org/10.1007/s00464-025-11715-3)] [Medline: [40259088](https://pubmed.ncbi.nlm.nih.gov/40259088/)]
35. Iranfar A, Soleymannejad M, Moshiri B, Taghirad HD. Natural language processing and soft data for motor skill assessment: a case study in surgical training simulations. *Comput Methods Programs Biomed.* 2025;264:108686. [doi: [10.1016/j.cmpb.2025.108686](https://doi.org/10.1016/j.cmpb.2025.108686)] [Medline: [40081199](https://pubmed.ncbi.nlm.nih.gov/40081199/)]
36. Nakajima K, Takenaka S, Kitaguchi D, Tanaka A, Ryu K, Takeshita N, et al. Artificial intelligence assessment of tissue-dissection efficiency in laparoscopic colorectal surgery. *Langenbecks Arch Surg.* 2025;410(1):80. [doi: [10.1007/s00423-025-03641-8](https://doi.org/10.1007/s00423-025-03641-8)] [Medline: [39984705](https://pubmed.ncbi.nlm.nih.gov/39984705/)]
37. Shayan AM, Hitchcock DB, Singh SP, Gao J, Groff RE, Singapogu RB. Functional data analysis of hand rotation for open surgical suturing skill assessment. *IEEE J Biomed Health Inform.* 2025;29(4):2981-2992. [doi: [10.1109/JBHI.2024.3496122](https://doi.org/10.1109/JBHI.2024.3496122)] [Medline: [39527415](https://pubmed.ncbi.nlm.nih.gov/39527415/)]
38. Yasui A, Hayashi Y, Hinoki A, Amano H, Shirota C, Tainaka T, et al. Developing an effective off-the-job training model and an automated evaluation system for thoracoscopic esophageal atresia surgery. *J Pediatr Surg.* 2025;60(2):161615. [doi: [10.1016/j.jpedsurg.2024.06.023](https://doi.org/10.1016/j.jpedsurg.2024.06.023)] [Medline: [39054116](https://pubmed.ncbi.nlm.nih.gov/39054116/)]

39. Yeh HH, Sen S, Chou JC, Christopher KL, Wang SY. PhacoTrainer: automatic artificial intelligence-generated performance ratings for cataract surgery. *Transl Vis Sci Technol.* 2025;14(5):2. [FREE Full text] [doi: [10.1167/tvst.14.5.2](https://doi.org/10.1167/tvst.14.5.2)] [Medline: [40310637](https://pubmed.ncbi.nlm.nih.gov/40310637/)]
40. Chen Z, Yang D, Li A, Sun L, Zhao J, Liu J, et al. Decoding surgical skill: an objective and efficient algorithm for surgical skill classification based on surgical gesture features-experimental studies. *Int J Surg.* 2024;110(3):1441-1449. [FREE Full text] [doi: [10.1097/JS9.0000000000000975](https://doi.org/10.1097/JS9.0000000000000975)] [Medline: [38079605](https://pubmed.ncbi.nlm.nih.gov/38079605/)]
41. Guo J, Han S, Liu YH. MT-ViT: multi-task video transformer for surgical skill assessment from streaming long-term vide. 2024. Presented at: 2024 IEEE International Conference on Robotics and Biomimetics (ROBIO); 2024 December 10-14:1752-1757; Bangkok, Thailand. [doi: [10.1109/robio64047.2024.10907318](https://doi.org/10.1109/robio64047.2024.10907318)]
42. Hoffmann H, Funke I, Peters P, Venkatesh DK, Egger J, Rivoir D, et al. AIxSuture: vision-based assessment of open suturing skills. *Int J Comput Assist Radiol Surg.* 2024;19(6):1045-1052. [FREE Full text] [doi: [10.1007/s11548-024-03093-3](https://doi.org/10.1007/s11548-024-03093-3)] [Medline: [38526613](https://pubmed.ncbi.nlm.nih.gov/38526613/)]
43. Komatsu M, Kitaguchi D, Yura M, Takeshita N, Yoshida M, Yamaguchi M, et al. Automatic surgical phase recognition-based skill assessment in laparoscopic distal gastrectomy using multicenter videos. *Gastric Cancer.* 2024;27(1):187-196. [doi: [10.1007/s10120-023-01450-w](https://doi.org/10.1007/s10120-023-01450-w)] [Medline: [38038811](https://pubmed.ncbi.nlm.nih.gov/38038811/)]
44. Nakajima K, Kitaguchi D, Takenaka S, Tanaka A, Ryu K, Takeshita N, et al. Automated surgical skill assessment in colorectal surgery using a deep learning-based surgical phase recognition model. *Surg Endosc.* 2024;38(11):6347-6355. [doi: [10.1007/s00464-024-11208-9](https://doi.org/10.1007/s00464-024-11208-9)] [Medline: [39214877](https://pubmed.ncbi.nlm.nih.gov/39214877/)]
45. Shafiei SB, Shadpour S, Mohler JL, Kauffman EC, Holden M, Gutierrez C. Classification of subtask types and skill levels in robot-assisted surgery using EEG, eye-tracking, and machine learning. *Surg Endosc.* 2024;38(9):5137-5147. [doi: [10.1007/s00464-024-11049-6](https://doi.org/10.1007/s00464-024-11049-6)] [Medline: [39039296](https://pubmed.ncbi.nlm.nih.gov/39039296/)]
46. Singh SP, Shayan AM, Gao J, Biblev J, Groff RE, Singapogu R. Evaluating smoothness of force for surgical skill assessment. 2024. Presented at: 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2024 July 15-19:1-4; Orlando, FL.
47. Singh SP, Shayan AM, Gao J, Bible J, Groff RE, Singapogu R. Objective and automated quantification of instrument handling for open surgical suturing skill assessment: a simulation-based study. *IEEE Open J Eng Med Biol.* 2024;5:485-493. [doi: [10.1109/OJEMB.2024.3402393](https://doi.org/10.1109/OJEMB.2024.3402393)] [Medline: [39050974](https://pubmed.ncbi.nlm.nih.gov/39050974/)]
48. Soleymani A, Tavakoli M, Aghazadeh F, Ou Y, Rouhani H, Zheng B, et al. Hands collaboration evaluation for surgical skills assessment: an information theoretical approach. *IEEE Trans Med Robot Bionics.* 2024;6(4):1490-1501. [FREE Full text]
49. Tonetti G, Våpenstad C, Zemiti N, Voros S. Providing automatic formative feedback along surgical skill assessment. 2024. Presented at: Proceedings of SPIE: Medical Imaging 2024: Image-Guided Procedures, Robotic Interventions, and Modeling; 2024 March 29:1292821; Cergy, France. URL: <https://doi.org/10.1117/12.3006508>
50. Anastasiou D, Jin Y, Stoyanov D, Mazomenos E, Anastasiou D. Keep your eye on the best: contrastive regression transformer for skill assessment in robotic surgery. *IEEE Robot Autom Lett.* 2023;8(3):1755-1762. [doi: [10.1109/lra.2023.3242466](https://doi.org/10.1109/lra.2023.3242466)]
51. Igaki T, Kitaguchi D, Matsuzaki H, Nakajima K, Kojima S, Hasegawa H, et al. Automatic surgical skill assessment system based on concordance of standardized surgical field development using artificial intelligence. *JAMA Surg.* 2023;158(8):e231131. [FREE Full text] [doi: [10.1001/jamasurg.2023.1131](https://doi.org/10.1001/jamasurg.2023.1131)] [Medline: [37285142](https://pubmed.ncbi.nlm.nih.gov/37285142/)]
52. Iranfar A, Soleymannejad M, Moshiri B, Taghirad HD. A modified Dempster Shafer approach to classification in surgical skill assessment. 2023. Presented at: 2023 31st International Conference on Electrical Engineering (ICEE); 2023 May 9-11:820-825; Tehran, Iran. [doi: [10.1109/icee59167.2023.10334703](https://doi.org/10.1109/icee59167.2023.10334703)]
53. 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. [FREE Full text] [doi: [10.3390/s23094496](https://doi.org/10.3390/s23094496)] [Medline: [37177699](https://pubmed.ncbi.nlm.nih.gov/37177699/)]
54. Saricilar EC, Burgess A, Freeman A. A pilot study of the use of artificial intelligence with high-fidelity simulations in assessing endovascular procedural competence independent of a human examiner. *ANZ J Surg.* 2023;93(6):1525-1531. [doi: [10.1111/ans.18484](https://doi.org/10.1111/ans.18484)] [Medline: [37088922](https://pubmed.ncbi.nlm.nih.gov/37088922/)]
55. Yanik E, Kruger U, Intes X, Rahul R, De S. Video-based formative and summative assessment of surgical tasks using deep learning. *Sci Rep.* 2023;13(1):1038. [FREE Full text] [doi: [10.1038/s41598-022-26367-9](https://doi.org/10.1038/s41598-022-26367-9)] [Medline: [36658186](https://pubmed.ncbi.nlm.nih.gov/36658186/)]
56. Ebina K, Abe T, Hotta K, Higuchi M, Furumido J, Iwahara N, et al. Objective evaluation of laparoscopic surgical skills in wet lab training based on motion analysis and machine learning. *Langenbecks Arch Surg.* 2022;407(5):2123-2132. [FREE Full text] [doi: [10.1007/s00423-022-02505-9](https://doi.org/10.1007/s00423-022-02505-9)] [Medline: [35394212](https://pubmed.ncbi.nlm.nih.gov/35394212/)]
57. Fathabadi FR, Grantner JL, Shebrain SA, Abdel-Qader I. Two-level fuzzy logic evaluation system for surgeon's hand movement using object detection. 2022. Presented at: 2022 IEEE Symposium Series on Computational Intelligence (SSCI); 2023 December 4-7; Singapore, Singapore.
58. Fathollahi M, Sarhan MH, Pena R, DiMonte L, Gupta A, Ataliwala A, et al. Video-based surgical skills assessment using long term tool tracking. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022.* Cham. Springer; 2022:546-556.

59. Kasa K, Burns D, Goldenberg MG, Selim O, Whyne C, Hardisty M. Multi-modal deep learning for assessing surgeon technical skill. *Sensors (Basel)*. 2022;22(19):7328. [[FREE Full text](#)] [doi: [10.3390/s22197328](https://doi.org/10.3390/s22197328)] [Medline: [36236424](#)]
60. Smith R, Julian D, Dubin A. Deep neural networks are effective tools for assessing performance during surgical training. *J Robot Surg*. 2022;16(3):559-562. [doi: [10.1007/s11701-021-01284-7](https://doi.org/10.1007/s11701-021-01284-7)] [Medline: [34268699](#)]
61. Yibulayimu S, Wang Y, Liu Y, Sun Z, Wang Y, Jiang H, et al. An explainable machine learning method for assessing surgical skill in liposuction surgery. *Int J Comput Assist Radiol Surg*. 2022;17(12):2325-2336. [doi: [10.1007/s11548-022-02739-4](https://doi.org/10.1007/s11548-022-02739-4)] [Medline: [36167953](#)]
62. Zhou XH, Xie XL, Liu SQ, Feng ZQ, Gui MJ, Wang JL, et al. Surgical skill assessment based on dynamic warping manipulations. *IEEE Trans Med Robot Bionics*. 2022;4(1):50-61. [[FREE Full text](#)]
63. Hasani P, Lotfi F, Taghirad HD. Towards an efficient computational framework for surgical skill assessment: suturing task by kinematic data. 2021. Presented at: 2021 9th RSI International Conference on Robotics and Mechatronics (ICRoM); November 17-19, 2021; Tehran, Iran. [doi: [10.1109/icrom54204.2021.9663473](https://doi.org/10.1109/icrom54204.2021.9663473)]
64. Lajkó G, Elek RN, Haidegger T. Surgical skill assessment automation based on sparse optical flow data. 2021. Presented at: 2021 IEEE 25th International Conference on Intelligent Engineering Systems (INES); 2021 July 7-9; Budapest, Hungary. [doi: [10.1109/ines52918.2021.9512917](https://doi.org/10.1109/ines52918.2021.9512917)]
65. Lavanchy JL, Zindel J, Kirtac K, Twick I, Hosgor E, Candinas D, et al. Automation of surgical skill assessment using a three-stage machine learning algorithm. *Sci Rep*. 2021;11(1):5197. [[FREE Full text](#)] [doi: [10.1038/s41598-021-84295-6](https://doi.org/10.1038/s41598-021-84295-6)] [Medline: [33664317](#)]
66. Liu D, Li Q, Jiang T, Wang Y, Miao R, Shan F, et al. Towards unified surgical skill assessment. 2021. Presented at: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 June 20-25:9517-9526; Nashville, TN.
67. Ming Y, Cheng Y, Jing Y, Liangzhe L, Pengcheng Y, Guang Z, et al. Surgical skills assessment from robot assisted surgery video data. 2021. Presented at: 2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA); 2021 January 22-24; Shenyang, China.
68. Takács K, Haidegger T. Adaptive neuro-fuzzy inference system for automated skill assessment in robot-assisted minimally invasive surgery. 2021. Presented at: 2021 IEEE 25th International Conference on Intelligent Engineering Systems (INES); 2021 July 7-9; Budapest, Hungary. [doi: [10.1109/ines52918.2021.9512924](https://doi.org/10.1109/ines52918.2021.9512924)]
69. Jian Z, Yue W, Wu Q, Li W, Wang Z, Lam V. Multitask learning for video-based surgical skill assessment. 2020. Presented at: 2020 Digital Image Computing: Techniques and Applications (DICTA); 2020 December 2:1-8; Melbourne, Australia. [doi: [10.1109/dicta51227.2020.9363408](https://doi.org/10.1109/dicta51227.2020.9363408)]
70. Lee D, Yu HW, Kwon H, Kong HJ, Lee KE, Kim HC. Evaluation of surgical skills during robotic surgery by deep learning-based multiple surgical instrument tracking in training and actual operations. *J Clin Med*. 2020;9(6):1964. [[FREE Full text](#)] [doi: [10.3390/jcm9061964](https://doi.org/10.3390/jcm9061964)] [Medline: [32585953](#)]
71. Ming Y, Cheng Y, Chunchen W, Meng L, Guang Z, Feng C. Automated objective basic surgical skills assessment: overall kinematic performance assessment method. 2020. Presented at: 2020 3rd International Conference on Mechatronics, Robotics and Automation (ICMRA); 2020 October 16-18:74-78; Shanghai, China. [doi: [10.1109/icmra51221.2020.9398349](https://doi.org/10.1109/icmra51221.2020.9398349)]
72. Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller PA. Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks. *Int J Comput Assist Radiol Surg*. 2019;14(9):1611-1617. [doi: [10.1007/s11548-019-02039-4](https://doi.org/10.1007/s11548-019-02039-4)] [Medline: [31363983](#)]
73. Funke I, Mees ST, Weitz J, Speidel S. Video-based surgical skill assessment using 3D convolutional neural networks. *Int J Comput Assist Radiol Surg*. 2019;14(7):1217-1225. [doi: [10.1007/s11548-019-01995-1](https://doi.org/10.1007/s11548-019-01995-1)] [Medline: [31104257](#)]
74. Gorantla KR, Esfahani ET. Surgical skill assessment using motor control features and hidden Markov model. 2019. Presented at: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2019 July 23-27:5842-5845; Berlin, Germany.
75. Kim TS, O'Brien M, Zafar S, Hager GD, Sikder S, Vedula SS. Objective assessment of intraoperative technical skill in capsulorhexis using videos of cataract surgery. *Int J Comput Assist Radiol Surg*. 2019;14(6):1097-1105. [doi: [10.1007/s11548-019-01956-8](https://doi.org/10.1007/s11548-019-01956-8)] [Medline: [30977091](#)]
76. Kowalewski KF, Garrow CR, Schmidt MW, Benner L, Müller-Stich BP, Nickel F. Sensor-based machine learning for workflow detection and as key to detect expert level in laparoscopic suturing and knot-tying. *Surg Endosc*. 2019;33(11):3732-3740. [doi: [10.1007/s00464-019-06667-4](https://doi.org/10.1007/s00464-019-06667-4)] [Medline: [30790048](#)]
77. Liu D, Jiang T, Wang Y, Miao R, Shan F, Li Z. Surgical skill assessment on in-vivo clinical data via the clearness of operating field. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Cham. Springer; 2019:476-484.
78. Khalil S, Shah HA, Bednarik R. Improving microsurgical suture training with automated phase recognition and skill assessment via deep learning. *Comput Biol Med*. 2025;192(Pt B):110238. [[FREE Full text](#)] [doi: [10.1016/j.compbiomed.2025.110238](https://doi.org/10.1016/j.compbiomed.2025.110238)] [Medline: [40339525](#)]
79. Sato K, Takenaka S, Kitaguchi D, Zhao X, Yamada A, Ishikawa Y, et al. Objective surgical skill assessment based on automatic recognition of dissection and exposure times in robot-assisted radical prostatectomy. *Langenbecks Arch Surg*. 2025;410(1):39. [doi: [10.1007/s00423-024-03598-0](https://doi.org/10.1007/s00423-024-03598-0)] [Medline: [39812861](#)]

80. Yen HH, Hsiao YH, Yang MH, Huang JY, Lin HT, Huang CC, et al. Automated surgical action recognition and competency assessment in laparoscopic cholecystectomy: a proof-of-concept study. *Surg Endosc.* 2025;39(5):3006-3016. [doi: [10.1007/s00464-025-11663-y](https://doi.org/10.1007/s00464-025-11663-y)] [Medline: [40116897](#)]
81. You J, Cai H, Wang Y, Bian A, Cheng K, Meng L, et al. Artificial intelligence automated surgical phases recognition in intraoperative videos of laparoscopic pancreatoduodenectomy. *Surg Endosc.* 2024;38(9):4894-4905. [doi: [10.1007/s00464-024-10916-6](https://doi.org/10.1007/s00464-024-10916-6)] [Medline: [38958719](#)]
82. Bhatia MB, Namazi B, Matthews J, Thomas C, Doster D, Martinez C, et al. Use of artificial intelligence to support surgical education personnel shortages in low- and middle-income countries: Developing a safer surgeon. *Global Surg Educ.* 2023;2(1):64. [FREE Full text]
83. Bkheet E, D'Angelo AL, Goldbraikh A, Laufer S. Using hand pose estimation to automate open surgery training feedback. *Int J Comput Assist Radiol Surg.* 2023;18(7):1279-1285. [doi: [10.1007/s11548-023-02947-6](https://doi.org/10.1007/s11548-023-02947-6)] [Medline: [37253925](#)]
84. 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):1677-1687. [doi: [10.1109/TMI.2022.3147640](https://doi.org/10.1109/TMI.2022.3147640)] [Medline: [35108200](#)]
85. Takeuchi M, Collins T, Ndagijimana A, Kawakubo H, Kitagawa Y, Marescaux J, et al. Automatic surgical phase recognition in laparoscopic inguinal hernia repair with artificial intelligence. *Hernia.* 2022;26(6):1669-1678. [doi: [10.1007/s10029-022-02621-x](https://doi.org/10.1007/s10029-022-02621-x)] [Medline: [35536371](#)]
86. Long Y, Wang JY, Reiter A, Hager GD. Relational graph learning on visual and kinematics embeddings for accurate gesture recognition in robotic surgery. 2021. Presented at: 2021 IEEE International Conference on Robotics and Automation (ICRA); 2021 June 5:13346-13353; Xi'an, China.
87. Amsterdam BV, Clarkson MJ, Stoyanov D. Multi-task recurrent neural network for surgical gesture recognition and progress prediction. 2020. Presented at: 2020 IEEE International Conference on Robotics and Automation (ICRA); 2020 May 31:1380-1386; Paris, France. [doi: [10.1109/icra40945.2020.9197301](https://doi.org/10.1109/icra40945.2020.9197301)]
88. Menegozzo G, Dall'Alba D, Zandonà C, Fiorini P. Surgical gesture recognition with time delay neural network based on kinematic data. 2019. Presented at: 2019 International Symposium on Medical Robotics (ISMR); 2019 April 3-5:1-7; Atlanta, GA. [doi: [10.1109/ismr.2019.8710178](https://doi.org/10.1109/ismr.2019.8710178)]
89. Lee J, Ham S, Kim N, Park HS. Development of a deep learning-based model for guiding a dissection during robotic breast surgery. *Breast Cancer Res.* 2025;27(1):34. [FREE Full text] [doi: [10.1186/s13058-025-01981-3](https://doi.org/10.1186/s13058-025-01981-3)] [Medline: [40065440](#)]
90. Ryu S, Imaizumi Y, Goto K, Iwauchi S, Kobayashi T, Ito R, et al. Artificial intelligence-enhanced navigation for nerve recognition and surgical education in laparoscopic colorectal surgery. *Surg Endosc.* 2025;39(2):1388-1396. [doi: [10.1007/s00464-024-11489-0](https://doi.org/10.1007/s00464-024-11489-0)] [Medline: [39762611](#)]
91. Khan DZ, Valetopoulou A, Das A, Hanrahan JG, Williams SC, Bano S, et al. Artificial intelligence assisted operative anatomy recognition in endoscopic pituitary surgery. *NPJ Digit Med.* 2024;7(1):314. [FREE Full text] [doi: [10.1038/s41746-024-01273-8](https://doi.org/10.1038/s41746-024-01273-8)] [Medline: [39521895](#)]
92. Rashidi FF, Grantner JL, Shebrain SA, Abdel-Qader I. Autonomous sequential surgical skills assessment for the peg transfer task in a laparoscopic box-trainer system with three cameras. *Robotica.* 2023;41(6):1837-1855. [FREE Full text] [doi: [10.1017/s0263574723000218](https://doi.org/10.1017/s0263574723000218)]
93. Madani A, Namazi B, Altieri MS, Hashimoto DA, Rivera AM, Pucher PH, et al. Artificial intelligence for intraoperative guidance: using semantic segmentation to identify surgical anatomy during laparoscopic cholecystectomy. *Ann Surg.* 2022;276(2):363-369. [FREE Full text] [doi: [10.1097/SLA.0000000000004594](https://doi.org/10.1097/SLA.0000000000004594)] [Medline: [33196488](#)]
94. Fukuta A, Yamashita S, 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. [doi: [10.1007/s11548-024-03253-5](https://doi.org/10.1007/s11548-024-03253-5)] [Medline: [39160382](#)]
95. Gao J, Shayan AM, Singh SP, Bible J, Singapogu R, Groff RE. Surgical suturing skill assessment using estimated hand roll angle from a deep-learning computer vision algorithm. 2024. Presented at: 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2024 July 15-19; Orlando, FL.
96. Abdelaal AE, Hong N, Avinash A, Budihal D, Sakr M, Hager GD, et al. Orientation matters: 6-DoF autonomous camera movement for video-based skill assessment in robot-assisted surgery. 2022. Presented at: 2022 9th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechanics (BioRob); 2022 August 21-24:1-7; Seoul, Korea. [doi: [10.1109/biorob52689.2022.9925374](https://doi.org/10.1109/biorob52689.2022.9925374)]
97. Elek RN, Haidegger T. Towards autonomous endoscopic image-based surgical skill assessment: articulated tool pose estimation. 2022. Presented at: 2022 IEEE 10th Jubilee International Conference on Computational Cybernetics and Cyber-Medical Systems; 2022 July 6-9; Reykjavík, Iceland. [doi: [10.1109/iccc20225925.2022.9922725](https://doi.org/10.1109/iccc20225925.2022.9922725)]
98. Liu Y, Zhao Z, Shi P, Li F. Towards surgical tools detection and operative skill assessment based on deep learning. *IEEE Trans Med Robot Bionics.* 2022;4(1):62-71. [FREE Full text]
99. Shao Z, Xu J, Stoyanov D, Mazomenos EB, Jin Y. Think step by step: chain-of-gesture prompting for error detection in robotic surgical videos. *IEEE Robot Autom Lett.* 2024;9(12):11513-11520. [doi: [10.1109/LRA.2024.3495452](https://doi.org/10.1109/LRA.2024.3495452)]

100. Tang M, Sugiyama T, Takahari R, Sugimori H, Yoshimura T, Ogasawara K, et al. Assessment of changes in vessel area during needle manipulation in microvascular anastomosis using a deep learning-based semantic segmentation algorithm: a pilot study. *Neurosurg Rev.* 2024;47(1):200. [FREE Full text] [doi: [10.1007/s10143-024-02437-6](https://doi.org/10.1007/s10143-024-02437-6)] [Medline: [38722409](https://pubmed.ncbi.nlm.nih.gov/38722409/)]
101. Yanik E, Ainam JP, Fu Y, Schwaitzberg S, Cavuoto L, De S. Video-based skill acquisition assessment in laparoscopic surgery using deep learning. *Global Surg Educ.* 2024;3(1):26. [doi: [10.1007/s44186-023-00223-4](https://doi.org/10.1007/s44186-023-00223-4)]
102. Ledwos N, Mirchi N, Yilmaz R, Winkler-Schwartz A, Sawni A, Fazlollahi AM, et al. Assessment of learning curves on a simulated neurosurgical task using metrics selected by artificial intelligence. *J Neurosurg.* 2022;137(4):1160-1171. [doi: [10.3171/2021.12.JNS211563](https://doi.org/10.3171/2021.12.JNS211563)] [Medline: [35120309](https://pubmed.ncbi.nlm.nih.gov/35120309/)]
103. Yang Y, Wang H, Wang J, Dong K, Ding S. Semantic-preserving surgical video retrieval with phase and behavior coordinated hashing. *IEEE Trans Med Imaging.* 2024;43(2):807-819. [doi: [10.1109/TMI.2023.3321382](https://doi.org/10.1109/TMI.2023.3321382)] [Medline: [37788194](https://pubmed.ncbi.nlm.nih.gov/37788194/)]
104. Fawaz HI, Forestier G, Weber J, Petitjean F, Idoumghar L, Muller PA. Automatic alignment of surgical videos using kinematic data. In: Riaño D, Wilk S, Teije A, editors. *Artificial Intelligence in Medicine AIME 2019*. Cham, Switzerland. Springer; 2019:139-148.

## Abbreviations

**ANN:** artificial neural network  
**AR:** augmented reality  
**AUC:** area under the curve  
**Bi-LSTM:** bidirectional long short-term memory  
**BoF:** bag of features  
**Cholec80:** Cholecystectomy 80  
**CNN:** convolutional neural network  
**ConvLSTM:** convolutional long short-term memory  
**ESSQS:** Endoscopic Surgical Skill Qualification System  
**GCN:** graph convolutional network  
**GOALS:** Global Objective Assessment of Laparoscopic Skills  
**Grad-CAM:** gradient-weighted class activation mapping  
**GRS:** Global Rating Scale  
**HAT:** human-autonomy teaming  
**HMM:** hidden Markov model  
**I3D:** Inflated 3D ConvNet  
**JIGSAWS:** Johns Hopkins University and Intuitive Surgical Inc Gesture and Skill Assessment Working Set  
**LOUO:** leave-one-user-out  
**LSTM:** long short-term memory  
**MAE:** mean absolute error  
**ML:** machine learning  
**MT-ViT:** multitask vision transformer  
**OSATS:** Objective Structured Assessment of Technical Skills  
**PRISMA:** Preferred Reporting Items for Systematic Reviews and Meta-Analyses  
**PRISMA-ScR:** Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews  
**RNN:** recurrent neural network  
**RQ:** research question  
**SHAP:** Shapley additive explanations  
**SVM:** support vector machine  
**TCN:** temporal convolutional network  
**TRL:** technology readiness level  
**U-Net:** U-shaped convolutional network  
**ViT:** vision transformer  
**VR:** virtual reality  
**YOLO:** You Only Look Once

*Edited by I Steenstra; submitted 24.Feb.2026; peer-reviewed by M Chakit, CS Biyani, E Lukács ; comments to author 23.Jul.2026; revised version received 28.Aug.2026; accepted 31.Aug.2026; published 02.Oct.2026*

*Please cite as:*

*Barati K, Ramirez Campos MS, Doyle TE*

*Machine Learning for Human-Autonomy Teaming in Surgical Skill Assessment: Scoping Review*

*JMIR AI 2026;5:e94109*

URL: <https://ai.jmir.org/2026/1/e94109>

doi: [10.2196/94109](https://doi.org/10.2196/94109)

PMID:

©Kamal Barati, Michael S Ramirez Campos, Thomas E Doyle. Originally published in JMIR AI (<https://ai.jmir.org>), 02.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.