Technical Skill Assessment in Minimally Invasive Surgery Using Artificial Intelligence:

A Systematic Review

Romina Pedrett, MD[1], Pietro Mascagni, MD, PhD[2,3], Guido Beldi, MD[1], Nicolas Padoy, PhD[2,4], Joël L. Lavanchy, MD[1,2]

1 Department of Visceral Surgery and Medicine, Inselspital, Bern University Hospital, University of Bern, Switzerland

2 IHU Strasbourg, France

3 Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, Italy

4 ICube, University of Strasbourg, CNRS, France

Correspondence and requests for reprints to:
Joël L. Lavanchy, MD
Department of Visceral Surgery and Medicine
Inselspital, Bern University Hospital
Freiburgstrasse
3010 Bern, Switzerland
joel.lavanchy@insel.ch

Joël Lavanchy was funded by the Swiss National Science Foundation (grant No P500PM_206724). This work was partially supported by French state funds managed by the ANR within the Investments for the future program under Grant ANR-10-IAHU-02 (IHU Strasbourg).

Running head: Technical Skill Assessment Using AI

Mini Abstract

Technical skill assessment in minimally invasive surgery is time consuming and costly. Artificial intelligence is a promising technology to facilitate and automate technical skill assessment. Therefore, this article systematically reviews artificial intelligence applications for the assessment of technical skills in minimally invasive surgery.

Abstract

Objective: To review artificial intelligence (AI) based applications for the assessment of technical skills in minimally invasive surgery.

Background: As technical skill assessment in surgery relies on expert opinion, it is time-consuming, costly, and often lacks objectivity. Analysis of routinely generated data by AI methods has the potential for automatic technical skill assessment in minimally invasive surgery.

Methods: A systematic search of Medline, Embase, Web of Science and IEEE Xplore was performed to identify original articles reporting the use of AI in the assessment of technical skill in minimally invasive surgery. Risk of bias (RoB) and quality of included studies were analyzed according to Quality Assessment of Diagnostic Accuracy Studies criteria and the modified Joanna Briggs Institute checklists, respectively. Findings were reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement.

Results: In total, 1467 articles were identified, 37 articles met eligibility criteria and were analyzed. Motion data extracted from surgical videos (49%) or kinematic data from robotic systems or sensors (46%) were the most frequent input data for AI. Most studies used deep learning (73%) and predicted technical skills using an ordinal assessment scale (73%) with good accuracies in simulated settings. However, all proposed models were in development stage, only 8% were externally validated and 16% showed a low RoB.

Conclusion: AI is promising to automate technical skill assessment in minimally invasive surgery. However, models should be benchmarked on representative datasets using predefined performance metrics and tested in clinical implementation studies.

INTRODUCTION

The assessment of technical skill is of major importance in surgical education and quality improvement programs given the association of technical skills with clinical outcomes[1–4]. This correlation has been demonstrated amongst others in bariactric[1], upper gastrointestinal[2] and colorectal surgery[3,4]. In addition, data from the American Colleges of Surgeons National Surgical Quality Improvement Program revealed that surgeon's technical skills as assessed by peers during right hemicolectomy are correlated with outcomes in colorectal as well as in non-colorectal surgeries performed by the same surgeon[3], showing the overarching impact of technical skills on surgical outcomes.

To date, technical skills are assessed through direct observations of surgeons' performance or retrospectively by reviewing surgical video recordings. Generally, this process involves either classifying skill levels in ordinal scales (e.g., novice, intermediate and expert) through unstructured observations or assessing performance intervals through the use of structured, validated checklists (e.g., Objective Structured Assessment of Technical Skills (OSATS)[5], Global Evaluative Assessment of Robotic Skills (GEARS)[6] (Figure 1). Therefore, technical skill assessment is complex and time consuming, hence costly. Moreover, technical skill assessment is limited by inter-observer variability and reviewer bias[7].

The growing adoption of minimally invasive surgery and recent developments in artificial intelligence (AI) could lead to automatic, objective, and consistent technical skill assessment in surgery.

As in minimally invasive surgery the surgical field is visualized by cameras, surgical videos are easily recorded and readily available at a large scale. Surgical videos can be used to extract information about technical skills. In robotic surgery the movements of the surgeon are translated to robotic arms holding the endoscope and the instruments. This allows the extraction of kinematic data such as moving trajectories directly form the robotic system. Based on kinematic data performance metrics of technical skills were developed[8].

AI is a very promising technology that is widely adopted in medicine[9,10]. For example, AI is able to detect diabetic retinopathy[11,12] and to screen for lung cancer[13] and malignant skin cancer[14] with an accuracy comparable to expert clinician screening.

Two subfields of AI are particularly used to extract and analyze motion data from surgical videos or robotic systems: machine learning (ML) and deep learning (DL). ML can be defined as computer algorithms that learn distinct features iterating over data without explicit programming. DL designates computer algorithms that analyze unstructured data using neural networks (NN). NN are computer algorithms designed in analogy to the synaptic network of the human brain. The input data is processed through multiple interconnected layers of artificial neurons, each performing mathematical operations on the input data to predict an output. The predicted output is compared to the human labeled output to optimize the operations of the NN, which makes it a self-learning system. From an AI perspective technical skill assessment is a classification (prediction of expert levels) or a regression task (prediction of a score). Figure 2 illustrates how different input data types are processed by AI models to predict technical skills.

The aim of this systematic review was to analyze studies using AI for technical skill assessment in minimally invasive surgery.

METHODS

This systematic review is reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)[15] guidelines and was prospectively registered at PROSPERO (2021 CRD42021267714).

Literature search

A systematic literature search of the databases Medline/Ovid, Embase/Ovid, Web of Science and IEEE Explore was conducted on August 25th, 2021. The first three databases account for biomedical literature and IEEE Explore for technical literature. A librarian of the University Library, University of Bern performed the literature search combining the following terms using Boolean operators: 1) Minimally invasive surgery including endoscopic, laparoscopic, or robotic surgery, and box model trainer. 2) AI including machine learning, supervised learning, unsupervised learning, computer vision and convolutional neural networks. 3) Technical skill assessment including surgical skill assessment, surgical performance assessment, and task performance analysis. The full-text search terms can be found in the Supplementary. The literature search was re-run prior to final analysis on February 25th, 2022.

Eligibility criteria

Studies presenting original research on AI applications for technical skills assessment in minimally invasive surgery including box model trainers published within the last 5 years (08/2016-02/2022) in English language were included. Review articles, conference abstracts, comments, and letters to the editor were excluded.

Study selection

Before screening, the identified records were automatically deduplicated using the reference manager program Endnote™ (Clarivate Analytics). After removal of the duplicates, two authors (R.P. & J.L.L.) independently screened the titles and abstracts of the identified records for inclusion using the web-tool Rayyan (https://www.rayyan.ai)[16]. Disagreement of the two authors regarding study selection was settled

in joint discussion. Of all included records the full-text articles were acquired. Articles not fulfilling the inclusion criteria after full-text screening were excluded.

Data extraction

Besides bibliographic data (title, author, publication year, journal name), the study population, the setting (laparoscopic/robotic simulation or surgery), the task assessed (e.g., peg transfer, cutting, knot-tying), the data input (motion data from video recordings, kinematic data from robotic systems or sensors), the dataset used, the assessment scale (ordinal scale vs. interval scale), the AI models used (ML or DL), the performance and the maturity level (development, validation, implementation) of AI models were extracted from the included studies.

Performance metrics

Performance metrics included accuracy, precision, recall, F1-score, and Area Under the Curve of Receiver Operator Characteristic (AUC-ROC). Accuracy is the proportion of correct predictions among the total number of observations. Precision is the proportion of true positive predictions among all (true and false) positive predictions and referred to as the positive predictive value. Recall is the proportion of true positive predictions among all relevant observations (true positives and false negatives) and referred to as sensitivity. F1-score is the harmonic mean of precision and recall and is a measure of model performance. A ROC curve plots the true positive against the false positive predictions at various thresholds and the AUC describes performance of the model to distinguish true positive from false positive predictions.

Risk of bias and quality assessment

The risk of bias of the included studies was assessed using the modified version of Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) criteria[17]. The quality of studies was evaluated using the modified Joanna Briggs Institute critical appraisal checklist for cross-sectional research in ML as used in[18,19].

RESULTS

The literature search retrieved a total of 1467 studies. After removing all duplicates, the remaining 1236 studies were screened by title and abstract. Thereafter, 88 studies remained, of which 51 were excluded after full-text screening. In summary, 37 studies[20–56] met eligibility criteria and thus were included into this systematic review (Figure 3). Out of these 37 studies, six (16%)[27,35,41,47,52,54] were obtained during the re-run prior to final analysis six months after the initial literature search was conducted. Table 1 gives an overview on the 37 studies included in this systematic review (for full information extracted see Supplementary Table S1).

Settings and tasks:

Most often, motion data from surgical videos or kinematic data from robotic systems or sensors were collected from simulators rather than during actual surgical procedures. The most common simulators used were robotic box models (51%, n=19) [21,23–26,28,29,31,35,38,41,45,47–49,52,54–56]. Laparoscopic simulators were the second most common setting for data collection (32%, n=12)[20,27,30,32,33,36,37,40,43,50,51,53].

The most common tasks assessed were suturing (62%, n=23)[21,23,25,26,28–31,33,35,37,38,40,41,47,49–56], knot-tying (41%, n=15)[21,23,25,26,28,29,35,38,40,41,47,49,52,55,56] and needle passing (35%, n=13)[21,23,25,28,29,35,38,41,47,49,52,55,56]. Other tasks assessed were peg transfers (19%, n=7)[24,30,33,36,37,43,51] and pattern cutting (19%, n=7)[20,27,30,32,33,37,51]. All of these tasks are part of the Fundamentals of Laparoscopic Surgery program, a well-established and validated training curriculum for laparoscopic surgery[57,58].

Eight studies (22%)[22,34,39,42,44–47] used data of real surgical procedures. Six (16%)[22,34,39,42,46,47] of them using videos of laparoscopic surgeries as for example laparoscopic cholecystectomies[34,42] or laparoscopic pelvic lymph node dissections[22]. Two studies (5%)[44,45] used video data obtained from robotic surgeries such as robotic prostatectomy[44] or robotic thyroid surgery[45]. The tasks assessed in surgical procedures ranged from entire interventions to specific steps (e.g., lymph node dissection[22], clip application[42]).

Input data and datasets:

Three different types of input data were used throughout the 37 studies: video data (49%, n=18)[20,22,27,30,31,33,34,37–39,42,44–47,51,52,54], kinematic data (46%, n=17)[21,23–26,28,29,35,40,41,43,48–50,53,55,56] and functional near-infrared spectroscopy (fNIRS) data (5%, n=2)[32,36]. Video recordings either from endoscopic laparoscopic/robotic camera or external cameras are used in 18 studies (49%). Kinematic data was obtained from DaVinci robotic systems (Intuitive Surgical Inc., CA, USA) in 13 studies (35%)[21,23–26,28,29,35,41,48,49,55,56] and from external sensors in four studies (11%)[40,43,50,53]. For example, electromyography sensors (Myo armband, Thalmic Labs, Ontario, CA)[40], optical sensors (Apple Watch, Apple, CA, USA)[43] or magnetic sensors attached to the instruments[50,53] were used as external sensors to collect kinematic data. Two studies[32,36] recorded fNIRS data from participants while they performed laparoscopic tasks. For example, Keles et al.[36] collected fNIRS data using a wireless, high density NIRS device, measuring functional brain activation of the prefrontal cortex. The NIRS device was adjacent to the surgeons' foreheads while they performed different laparoscopic tasks.

Publicly available datasets were used in 16 studies (43%)[21,23,25,26,28,29,31,34,35,38,41,47,49,52,55,56]. Of those, the JIGSAWS (Johns Hopkins University and Intuitive Surgical, Inc. Gesture and Skill Assessment Working Set)[59] dataset was most frequently used (n=15, 41%)[21,23,25,26,28,29,31,35,38,41,47,49,52,55,56]. It contains video and kinematic data together with human annotated skill ratings of eight surgeons performing three surgical tasks in five-fold repetition in a robotic box model trainer. One study[34] extended the publicly available m2cai16-tool dataset[60] with locations of surgical tools and published it as m2cai16-tools-localisation dataset. Though, most studies (n=21, 57%) created private datasets, that were not publicly released. Most datasets (n=34, 92%) were monocentric. However, three studies (8%) used a multicentric dataset: French, et al.[30] used a multi-institutional dataset from three centers, Kitagutchi, et al.[39] draw a sample form a national Japan Society of Endoscopic Surgeons database, and Law, et al.[44] used a part of a statewide national quality improvement database collected by the Michigan Urological Surgical Improvement Collaborative. Three of the 37 studies included (8%)[29,47,49], reported external validation on a second independent dataset.

Assessment:

Technical surgical skills can be assessed using expert levels (ordinal scale) or proficiency scores (interval scale) (Figure 1). In 27 of the studies (73%) an ordinal scale was applied[20,21,25–31,35–38,40–45,48,49,51–56]. In twelve studies (32%) participants were categorized in two different skill levels[20,26,30,35–37,41,42,44,48,51,53] and in 14 studies (38%) into three different expert levels (novice, intermediate, expert)[21,25,28,29,31,38,40,43,45,49,52,54–56]. Ten studies (27%) applied different proficiency scores: Pelvic Lymphadenectomy Assessment and Completion Evaluation (PLACE[61]), Fundamentals of Laparoscopic Surgery (FLS[62]), Endoscopic Surgical Skill Qualification System (ESSQS[63]), Objective Structured Assessment of Technical Skills (OSATS[64]), and Global Evaluative Assessment of Robotic Skills (GEARS[6])[22–24,32,33,39,46,47,50,54].

AI models:

All AI models in this review are either ML- or DL-based. ML was applied in 12 studies (32%)[24,26,29,30,35,36,40,42,47,48,50,51] and DL in 27 studies (73%)[20–23,25,27,28,31–34,37–46,49,52–56]. Two studies (5%) used a combination of ML and DL models[42,44].

Performance:

The most common performance metrics reported in the studies included in this systematic review is accuracy (n=30, 81%)[20–22,24–26,28–32,35–42,44,45,48–56]. Accuracies of the best performing models range between 0.7 – 1. Other performance metrics reported include F1-score (n=6, 16%)[31,35,38,43,50,51], recall (n=4, 11%)[24,31,38,40], sensitivity (n=4, 11%)[20,32,50,51], specificity (n=4, 11%)[20,32,50,51], and AUC-ROC (n=4, 11%)[20,35,50,51]. Four studies (11%)[23,27,33,34] did not report a performance metric at all.

Risk of bias and quality assessment:

Six of the included studies (16%)[24,37,39,40,42,50] had an overall low probability of bias in the risk of bias assessment. The other studies had one (n=10, 27%), two (n=9, 24%), three (n=8, 22%), four (n=3, 8%) or five criteria (n=1, 3%) at risk of bias. The full risk of bias assessment table is presented in the

Supplementary (Table S2). The quality assessment of the included studies is displayed in Figure 4. All proposed AI models were in a developmental preclinical stage of maturity, none was implemented in routine clinical use.

DISCUSSION

This systematic review of AI applications for technical skill assessment in minimally invasive surgery included 37 studies. Technical surgical skills were assessed in 51% of studies in robotic simulators, in 32% of studies in laparoscopic simulators, and in 22% of studies in actual surgical procedures. The input data to AI models were video data (49%), kinematic data from robotic systems or sensors (46%), and fNIRS data (2%). Technical skills were classified in 73% of studies using skill levels and in 27% of studies using proficiency scores. In total, 32% of AI models were ML-based and 73% of AI models were DL-based. Most studies (81%) reported accuracy as performance metric. Overall, 84% of studies were at risk of bias and only 3 studies tested their AI model for external validity using a second independent dataset. None of the proposed models was implemented in routine clinical use.

The comparability of studies included in this systematic review is limited due to several fundamental differences between them. Most studies (57%) use private datasets of different settings, tasks, and sizes. However, 15 studies included in this systematic review used JIGSAWS, a robotic simulator dataset and the most frequently used dataset in technical skill assessment. The use of simulators for technical skill assessment has advantages and disadvantages. On the one hand, simulators allow to control the experimental setting and enable reproducibility of studies. On the other hand, box model trainers simulate surgical tasks and have only a restricted degree of realism. In addition, simulators are well established in surgical training but have limited significance in the assessment of fully trained surgeons. The use of video recordings and motion data of actual surgeries as input data improves the validity of technical skill assessment models. However, in actual surgeries the experimental setting cannot be standardized and therefore, lacks reproducibility.

Moreover, the comparison of studies is impaired by the different scales and scores used to measure technical skill. Some studies use ordinal scales with different numbers of skill levels (good vs. bad[20,26,35–37,41,42,44,48,51,53], novice vs. intermediate vs. expert[21,25,28–31,38,40,43,45,49,52,54–56]) others use different interval scales (OSATS scores[28,40,50], GEARS scores[24,54], or Likert scales[42]). This finding represents the general difficulty to define and measure technical surgical skills.

Most of the studies included in this systematic review have methodologic limitations. Overall, 84% of studies included in this review are at risk of bias. The quality assessment of the included studies revealed that only 32% of studies discussed the findings and implications in detail. Furthermore, only three studies included in this review have a multicentric dataset. Only three of the AI models studied are validated on an independent external dataset. Therefore, it is questionable whether the AI models included in this review would generalize to other settings, tasks, and institutions. Out of 37 included studies, 30 report on accuracy. However, there is a large variation of reported performance metrics among studies included in this systematic review. Due to the novelty of AI application in the healthcare domain and in surgery in particular, the literature lacks standards in the evaluation of AI methods and their performance. There is an urgent need for the application of guidelines to assess AI models and for studies comparing them head-to-head. Guidelines for early stage clinical evaluation of AI[65] and clinical trials involving AI[66] have been published recently. However, the studies included in this review are all at a preclinical stage where these guidelines do not apply. A multi-stakeholder initiative recently introduced guidelines and flowcharts on the choice of AI evaluation metrics in the medical image domain[67]. For surgical video analysis this effort still needs to be taken[68]. To overcome the limitations of the proposed AI models for technical skill assessment, valid and representative datasets using predefined performance metrics, and external validation in clinical implementation studies will be essential.

Looking at the educational benefits of AI algorithms, the current models allow an estimation of individual skill levels in comparison with the population the algorithm was trained on. However, no direct or concrete feedback on how to improve technical skills is provided. Potentially training AI models on

assessment scores divided in different domains of technical skills (e.g. bimanual dexterity, tissue handling,) could help to give automated but actionable feedback.

In conclusion, AI has great potential to automate technical skill assessment in minimally invasive surgery. Various AI models, that analyze surgical video or movement data from simulators or actual surgical procedures and correlate them with technical surgical skills, have been studied. However, the studies included in this review lack standardization of datasets, performance metrics and external validation. Therefore, we advocate for benchmarking of AI models on valid and representative datasets using predefined performance metrics and testing in clinical implementation studies.

ACKNOLEDGEMENTS

We would like to acknowledge the help of Tanya Karrer, Information Specialist Medicine, University Library, University of Bern with the literature search.

REFERENCES

1. Birkmeyer JD, Finks JF, O'Reilly A, et al. Surgical Skill and Complication Rates after Bariatric Surgery. *N Engl J Med*. 2013;369(15):1434-1442. doi:10.1056/NEJMsa1300625
2. Fecso AB, Bhatti JA, Stotland PK, Quereshy FA, Grantcharov TP. Technical Performance as a Predictor of Clinical Outcomes in Laparoscopic Gastric Cancer Surgery. *Ann Surg*. 2019;270(1):115-120. doi:10.1097/SLA.0000000000002741
3. Stulberg JJ, Huang R, Kreutzer L, et al. Association Between Surgeon Technical Skills and Patient Outcomes. *JAMA Surg*. 2020;155(10):960-968. doi:10.1001/jamasurg.2020.3007
4. Curtis NJ, Foster JD, Miskovic D, et al. Association of Surgical Skill Assessment With Clinical Outcomes in Cancer Surgery. *JAMA Surg*. 2020;155(7):590-598. doi:10.1001/jamasurg.2020.1004
5. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents: OBJECTIVE STRUCTURED ASSESSMENT OF TECHNICAL SKILL. *Br J Surg*. 1997;84(2):273-278. doi:10.1046/j.1365-2168.1997.02502.x
6. Goh Alvin C., Goldfarb David W., Sander James C., Miles Brian J., Dunkin Brian J. Global Evaluative Assessment of Robotic Skills: Validation of a Clinical Assessment Tool to Measure Robotic Surgical Skills. *J Urol*. 2012;187(1):247-252. doi:10.1016/j.juro.2011.09.032
7. Lendvay TS, White L, Kowalewski T. Crowdsourcing to Assess Surgical Skill. *JAMA Surg*. 2015;150(11):1086-1087. doi:10.1001/jamasurg.2015.2405
8. Hung Andrew J., Chen Jian, Jarc Anthony, Hatcher David, Djaladat Hooman, Gill Inderbir S. Development and Validation of Objective Performance Metrics for Robot-Assisted Radical Prostatectomy: A Pilot Study. *J Urol*. 2018;199(1):296-304. doi:10.1016/j.juro.2017.07.081
9. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56. doi:10.1038/s41591-018-0300-7
10. Plana D, Shung DL, Grimshaw AA, Saraf A, Sung JJY, Kann BH. Randomized Clinical Trials of Machine Learning Interventions in Health Care: A Systematic Review. *JAMA Netw Open*. 2022;5(9):e2233946. doi:10.1001/jamanetworkopen.2022.33946
11. Gulshan V, Peng L, Coram M, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*. 2016;316(22):2402-2410. doi:10.1001/jama.2016.17216
12. Ting DSW, Cheung CYL, Lim G, et al. Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. *JAMA*. 2017;318(22):2211-2223. doi:10.1001/jama.2017.18152
13. Ardila D, Kiraly AP, Bharadwaj S, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med*. 2019;25(6):954-961. doi:10.1038/s41591-019-0447-x
14. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118. doi:10.1038/nature21056
15. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71. doi:10.1136/bmj.n71

16. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. *Syst Rev*. 2016;5(1):210. doi:10.1186/s13643-016-0384-4

17. Whiting PF. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann Intern Med*. 2011;155(8):529. doi:10.7326/0003-4819-155-8-201110180-00009

18. Anteby R, Horesh N, Soffer S, et al. Deep learning visual analysis in laparoscopic surgery: a systematic review and diagnostic test accuracy meta-analysis. *Surg Endosc*. 2021;35(4):1521-1533. doi:10.1007/s00464-020-08168-1

19. Kwong MT, Colopy GW, Weber AM, Ercole A, Bergmann JHM. The efficacy and effectiveness of machine learning for weaning in mechanically ventilated patients at the intensive care unit: a systematic review. *Bio-Des Manuf*. 2019;2(1):31-40. doi:10.1007/s42242-018-0030-1

20. Alonso-Silverio GA, Perez-Escamirosa F, Bruno-Sanchez R, et al. Development of a Laparoscopic Box Trainer Based on Open Source Hardware and Artificial Intelligence for Objective Assessment of Surgical Psychomotor Skills. *Surg Innov*. 2018;25(4):380-388. doi:10.1177/1553350618777045

21. Anh NX, Chauhan S, Nataraja RM. 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

22. Baghdadi A, Hussein AA, Ahmed Y, Guru KA, Cavuoto LA. A computer vision technique for automated assessment of surgical performance using surgeons' console-feed videos. *Int J Comput Assist Radiol Surg*. 2019;14(4):697-707. doi:10.1007/s11548-018-1881-9

23. Benmansour M, Handouzi W, Malti A. A Neural Network Architecture for Automatic and Objective Surgical Skill Assessment. *In: Proceedings 2018 3rd International Conference on Electrical Sciences and Technologies in Maghreb (CISTEM); 2018.* doi:10.1109/CISTEM.2018.8613550

24. Brown JD, O Brien CE, Leung SC, Dumon KR, Lee DI, Kuchenbecker KJ. Using Contact Forces and Robot Arm Accelerations to Automatically Rate Surgeon Skill at Peg Transfer. *IEEE Trans Biomed Eng*. 2017;64(9):2263-2275. doi:10.1109/TBME.2016.2634861

25. Castro D, Pereira D, Zanchettin C, Macêdo D, Bezerra BLD. Towards Optimizing Convolutional Neural Networks for Robotic Surgery Skill Evaluation. In: ; 2019:1-8. doi:10.1109/IJCNN.2019.8852341

26. Fard MJ, Ameri S, Darin Ellis R, Chinnam RB, Pandya AK, Klein MD. Automated robot-assisted surgical skill evaluation: Predictive analytics approach. *Int J Med Robot*. 2018;14(1):e1850. doi:10.1002/rcs.1850

27. Fathabadi FR, Grantner JL, Shebrain SA, Abdel-Qader I. Surgical Skill Assessment System Using Fuzzy Logic in a Multi-Class Detection of Laparoscopic Box-Trainer Instruments. In: ; 2021. doi:10.1109/SMC52423.2021.9658766

28. Fawaz HI, 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

29. Forestier G, Petitjean F, Senin P, et al. Surgical motion analysis using discriminative interpretable patterns. *Artif Intell Med*. 2018;91:3-11. doi:10.1016/j.artmed.2018.08.002

30. French A, Kowalewski TM, Lendvay TS, Sweet RM. Predicting surgical skill from the first N

seconds of a task: value over task time using the isogony principle. *Int J Comput Assist Radiol Surg*. 2017;12(7):1161-1170. doi:10.1007/s11548-017-1606-5

31. Funke I, Speidel S, Mees ST, Weitz J. 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

32. Gao Y, Yan P, Kruger U, et al. Functional Brain Imaging Reliably Predicts Bimanual Motor Skill Performance in a Standardized Surgical Task. *IEEE Trans Biomed Eng*. 2021;68(7):2058-2066. doi:10.1109/TBME.2020.3014299

33. Islam G, Kahol K, Li BX, Smith M, Patel VL. Affordable, web-based surgical skill training and evaluation tool. *J Biomed Inform*. 2016;59:102-114. doi:10.1016/j.jbi.2015.11.002

34. Jin A, Yeung S, Jopling J, et al. Tool Detection and Operative Skill Assessment in Surgical Videos Using Region-Based Convolutional Neural Networks. In: ; 2018:691-699. doi:10.1109/WACV.2018.00081

35. Juarez-Villalobos L, Hevia-Montiel N, Perez-Gonzalez J. Machine Learning based Classification of Local Robotic Surgical Skills in a Training Tasks Set. *Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Int Conf*. 2021;2021. doi:10.1109/EMBC46164.2021.9629579

36. Keles HO, Cengiz C, Demiral I, Ozmen MM, Omurtag A. High density optical neuroimaging predicts surgeons's subjective experience and skill levels. *PLoS ONE*. 2021;16(2 February):e0247117. doi:10.1371/journal.pone.0247117

37. Kelly JD, Kowalewski TM, Petersen A, Lendvay TS. Bidirectional long short-term memory for surgical skill classification of temporally segmented tasks. *Int J Comput Assist Radiol Surg*. 2020;15(12):2079-2088. doi:10.1007/s11548-020-02269-x

38. Khalid S, Goldenberg M, Grantcharov T, Taati B, Rudzicz F. Evaluation of Deep Learning Models for Identifying Surgical Actions and Measuring Performance. *JAMA Netw OPEN*. 2020;3(3). doi:10.1001/jamanetworkopen.2020.1664

39. Kitaguchi D, Takeshita N, Matsuzaki H, Igaki T, Hasegawa H, Ito M. Development and Validation of a 3-Dimensional Convolutional Neural Network for Automatic Surgical Skill Assessment Based on Spatiotemporal Video Analysis. *JAMA Netw Open*. 2021;4(8):e2120786. doi:10.1001/jamanetworkopen.2021.20786

40. Kowalewski KF, Garrow CR, Schmidt MW, Benner L, Muller-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 Interv Tech*. 2019;33(11):3732-3740. doi:10.1007/s00464-019-06667-4

41. Lajko G, Nagyne Elek R, Haidegger T. Endoscopic Image-Based Skill Assessment in Robot-Assisted Minimally Invasive Surgery. *Sensors*. 2021;21(16). doi:10.3390/s21165412

42. Lavanchy JL, Zindel J, Kirtac K, et al. Automation of surgical skill assessment using a three-stage machine learning algorithm. *Sci Rep*. 2021;11(1):5197. doi:10.1038/s41598-021-84295-6

43. Laverde R, Rueda C, Amado L, Rojas D, Altuve M. Artificial Neural Network for Laparoscopic Skills Classification Using Motion Signals from Apple Watch. In: ; 2018:5434-5437. doi:10.1109/EMBC.2018.8513561

44. Law H, Zhang Y, Kim TK, et al. Surgeon technical skill assessment using computer vision-based

analysis. *In: Proceedings of the 2nd Machine Learning for Healthcare Conference 2018*. https://proceedings.mlr.press/v68/law17a.html

45. 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):1-15. doi:10.3390/jcm9061964

46. Liu D, Jiang T, Wang Y, Miao R, Shan F, Li Z. Clearness of operating field: a surrogate for surgical skills on in vivo clinical data. *Int J Comput Assist Radiol Surg*. 2020;15(11):1817-1824. doi:10.1007/s11548-020-02267-z

47. Liu D, Li Q, Jiang T, et al. Towards Unified Surgical Skill Assessment. In: ; 2021. doi:10.1109/CVPR46437.2021.00940

48. Lyman WB, Passeri MJ, Murphy K, et al. An objective approach to evaluate novice robotic surgeons using a combination of kinematics and stepwise cumulative sum (CUSUM) analyses. *Surg Endosc*. 2021;35(6):2765-2772. doi:10.1007/s00464-020-07708-z

49. Nguyen XA, Ljuhar D, Pacilli M, Nataraja RM, Chauhan S. Surgical skill levels: Classification and analysis using deep neural network model and motion signals. *Comput METHODS PROGRAMS Biomed*. 2019;177:1-8. doi:10.1016/j.cmpb.2019.05.008

50. Oquendo YA, Riddle EW, Hiller D, Blinman TA, Kuchenbecker KJ. Automatically rating trainee skill at a pediatric laparoscopic suturing task. *Surg Endosc Interv Tech*. 2018;32(4):1840-1857. doi:10.1007/s00464-017-5873-6

51. Perez-Escamirosa F, Alarcon-Paredes A, Alonso-Silverio GA, et al. Objective classification of psychomotor laparoscopic skills of surgeons based on three different approaches. *Int J Comput Assist Radiol Surg*. 2020;15(1):27-40. doi:10.1007/s11548-019-02073-2

52. Soleymani A, Asl AAS, Yeganejou M, Dick S, Tavakoli M, Li X. Surgical Skill Evaluation From Robot-Assisted Surgery Recordings. In: ; 2021. doi:10.1109/ISMR48346.2021.9661527

53. Uemura M, Tomikawa M, Akahoshi T, et al. Feasibility of an AI-Based Measure of the Hand Motions of Expert and Novice Surgeons. *Comput Math Methods Med*. 2018;2018:9873273. doi:10.1155/2018/9873273

54. Wang Y, Dai J, Morgan TN, et al. Evaluating robotic-assisted surgery training videos with multi-task convolutional neural networks. *J Robot Surg*. Published online 2021. doi:10.1007/s11701-021-01316-2

55. Wang ZH, Fey AM. Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. *Int J Comput Assist Radiol Surg*. 2018;13(12):1959-1970. doi:10.1007/s11548-018-1860-1

56. Wang Z, Fey AM. SATR-DL: Improving Surgical Skill Assessment And Task Recognition In Robot-Assisted Surgery With Deep Neural Networks. *Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Int Conf*. 2018;2018:1793-1796. doi:10.1109/EMBC.2018.8512575

57. Zendejas B, Ruparel RK, Cook DA. Validity evidence for the Fundamentals of Laparoscopic Surgery (FLS) program as an assessment tool: a systematic review. *Surg Endosc*. 2016;30(2):512-520. doi:10.1007/s00464-015-4233-7

58. Fried GM, Feldman LS, Vassiliou MC, et al. Proving the Value of Simulation in Laparoscopic

Surgery. *Ann Surg*. 2004;240(3):518-528. doi:10.1097/01.sla.0000136941.46529.56

59. Gao Y, Vedula SS, Reiley CE, et al. 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) – MICCAI Workshop*.

60. Twinanda AP, Shehata S, Mutter D, et al. EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos. *IEEE Trans Med Imaging*. 2017;36(1):86-97. doi:10.1109/tmi.2016.2593957

61. Hussein AA, Ghani KR, Peabody J, et al. Development and Validation of an Objective Scoring Tool for Robot-Assisted Radical Prostatectomy: Prostatectomy Assessment and Competency Evaluation. *J Urol*. 2017;197(5):1237-1244. doi:10.1016/j.juro.2016.11.100

62. Peters JH, Fried GM, Swanstrom LL, et al. Development and validation of a comprehensive program of education and assessment of the basic fundamentals of laparoscopic surgery. *Surgery*. 2004;135(1):21-27. doi:10.1016/S0039-6060(03)00156-9

63. Mori T, Kimura T, Kitajima M. Skill accreditation system for laparoscopic gastroenterologic surgeons in Japan. *Minim Invasive Ther Allied Technol*. 2010;19(1):18-23. doi:10.3109/13645700903492969

64. Vassiliou MC, Feldman LS, Andrew CG, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg*. 2005;190(1):107-113. doi:10.1016/j.amjsurg.2005.04.004

65. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ*. 2022;377:e070904. doi:10.1136/bmj-2022-070904

66. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26(9):1364-1374. doi:10.1038/s41591-020-1034-x

67. Maier-Hein L, Reinke A, Christodoulou E, et al. Metrics reloaded: Pitfalls and recommendations for image analysis validation. Published online July 7, 2022. Accessed August 31, 2022. http://arxiv.org/abs/2206.01653

68. Kitaguchi D, Watanabe Y, Madani A, et al. Artificial Intelligence for Computer Vision in Surgery: A Call for Developing Reporting Guidelines. *Ann Surg*. 2022;275(4):e609-e611. doi:10.1097/SLA.0000000000005319

TABLES AND FIGURES

Figure 1: Human technical skill assessment in minimally invasive surgery.

Figure 2: Automated technical skill assessment in minimally invasive surgery by artificial intelligence.

Figure 3: PRISMA flow diagram of the study selection process (from PRISMA Statement 2020[15]).

Figure 4: Quality assessment of the included studies. The numbers within the bars represent the respective number of studies.

Table 1: Information summary of all studies included in this review

Author	*Year*	*Population*	*Setting*	*Tasks*	*Input Data*	*Dataset*	*Assessment*	*AI model*	*Accuracy*	*Maturity level*
Alonso-Silverio et al.[20]	2018	20	LS	PC	VR	private	binary (experienced, non-experienced)	DL	0.94	dev
Anh et al.[21]	2020	8	RS	SU, NP, KT	KD (dV)	JIGSAWS	N, I, E	DL	0.97	dev
Baghdadi et al.[22]	2018	na	Lap	Pelvic lymph node dissection	VR	private	PLACE score	DL	0.83	dev
Benmansour et al.[23]	2018	6	RS	SU, NP, KT	KD (dV)	JIGSAWS	Custom score	DL	na	dev
Brown et al.[24]	2017	38	RS	PT	KD (dV)	private	GEARS score (1-5: exact rating)	ML	0.75	dev
Castro et al.[25]	2019	8	RS	SU, NP, KT	KD (dV)	JIGSAWS	N, I, E	DL	0.98	dev
Fard et al.[26]	2017	8	RS	SU, KT	KD (dV)	JIGSAWS	binary (N, E)	ML	0.9	dev
Fathabadi et al.[27]	2021	na	LS	PC	VR	private	Level A (excellent) - E (very bad)	DL	na	dev
Fawaz et al.[28]	2019	8	RS	SU, NP, KT	KD (dV)	JIGSAWS	N, I, E	DL	1	dev
Forestier et al.[29]	2018	8	RS	SU, NP, KT	KD (dV)	JIGSAWS	N, I, E	ML	0.96	dev
French et al.[30]	2017	98	LS	PT, SU, PC	VR	private	binary (N, E)	ML	0.9	dev
Funke et al.[31]	2019	8	RS	SU	VR	JIGSAWS	N, I, E	DL	1	dev
Gao et al.[32]	2020	13	LS	PC	fNIRS data	private	FLS score: pass/fail	DL	0.91	dev
Islam et al.[33]	2016	52	LS	PT, SU, PC	VR	private	Custom score	DL	na	dev
Jin et al.[34]	2018	na	Lap	Lap cholecystectomy	VR	m2cai16-tools-location	unknown	DL	na	dev
Juarez-Villalobos et al.[35]	2021	8	RS	SU, NP, KT	KD (dV)	JIGSAWS	binary (N, E)	ML	1	dev
Keles et al.[36]	2021	33	LS	PT, threading	fNIRS data	private	binary (student vs. attending)	ML	~ 0.9	dev
Kelly et al.[37]	2020	na	LS	PT, SU, PC, Clipping	VR	private	binary (N, E)	DL	0.97	dev
Khalid et al.[38]	2020	8	RS	SU, NP, KT	VR	JIGSAWS	N, I, E	DL	0.77	dev
Kitaguchi et al.[39]	2021	na	Lap	Lap colorectal surgery	VR	private	ESSQS score	DL	0.75	dev
Kowalewski et al.[40]	2019	28	LS	SU, KT	KD (s)	private	N, I, E	DL	0.7	dev
Lajkó et al.[41]	2021	8	RS	SU, NP, KT	KD (dV)	JIGSAWS	binary (N, E)	DL	0.84	dev
Lavanchy et al.[42]	2021	40	Lap	Lap cholecystectomy	VR	private	binary (good vs. poor)	ML, DL	0.87	dev
							5-point Likert scale (+/- 1 point)		0.7	
Laverde et al.[43]	2018	7	LS	PT	KD (s)	private	N, I, E	DL	na	dev
Law et al.[44]	2017	12	Rob	Robotic prostatectomy	VR	private	binary (good vs. poor)	ML, DL	0.92	dev
Lee et al.[45]	2020	1/ na	RS, Rob	Robotic thyroid surgery/simulation	VR	private	N, I, E	DL	0.83	dev
Liu et al.[46]	2020	na	Lap	Lap gastrectomy	VR	private	modified OSATS score	DL	na	dev
Liu et al.[47]	2021	8 / na	RS, Lap	SU, NP, KT / Lap surgery gastric cancer	VR	JIGSAWS	modified OSATS score	ML	na	dev
Lyman et al.[48]	2021	2	RS	SU of hepaticojejunostomy	KD (dV)	private	binary (N, I)	ML	0.89	dev
Nguyen et al.[49]	2019	8	RS	SU, NP, KT	KD (dV)	JIGSAWS	N, I, E	DL	0.98	dev
Oquendo et al.[50]	2018	32	LS	SU	KD (s)	private	OSATS scores (+/- 4 points)	ML	0.89	dev
Pérez-Escamirosa et al.[51]	2019	43	LS	PT, SU, PC	VR	private	binary (experienced vs. non-experienced)	ML	0.98	dev
Soleymani et al.[52]	2021	8	RS	SU, NP, KT	VR	JIGSAWS	N, I, E	DL	0.97	dev
Uemura et al.[53]	2018	67	LS	SU	KD (s)	private	binary (N, E)	DL	0.79	dev
Wang Y. et al.[54]	2021	18	RS	SU	VR	private	N, I, E	DL	0.83	dev
							GEARS score (+/- 1 point)		0.86	
Wang Z. et al.[55]	2018	8	RS	SU, NP, KT	KD (dV)	JIGSAWS	N, I, E	DL	0.95	dev
Wang Z. et al.[56]	2018	8	RS	SU, NP, KT	KD (dV)	JIGSAWS	N, I, E	DL	0.96	dev

Of note, to ensure legibility the data provided in Table 1 is limited to accuracy metrics of the best performing model presented in each study. For full information extracted see Supplementary Material Table S2. Abbreviations: na = not available, LS = laparoscopic simulator, Lap = laparoscopic surgery, RS = robotic simulator, Rob = robotic surgery, PC = pattern cutting, SU = suturing, NP = needle-passing, KT = knot-tying, PT = peg transfer, VR = video recordings, KD (dv) = kinematic data collected by daVinci systems, fNIRS = functional near-infrared spectroscopy, KD (s) = kinematic data collected by external sensors, DL = deep learning, ML = machine learning, N = novice, I = intermediate, E = expert, PLACE = Pelvic Lymphadenectomy Assessment and Completion Evaluation, GEARS = Global Evaluative Assessments of Robotic Skills, FLS = Fundamentals of Laparoscopic Surgery, ESSQS = Endoscopic Surgical Skill Qualification System, OSATS = Objective Structured Assessment of Technical Skills, dev = development.