Surgical Endoscopy
https://doi.org/10.1007/s00464-023-10335-z

REVIEW ARTICLE

Technical skill assessment in minimally invasive surgery using artificial intelligence: a systematic review

Romina Pedrett[1] · Pietro Mascagni[2,3] · Guido Beldi[1] · Nicolas Padoy[2,4] · Joël L. Lavanchy[1,2,5]

Received: 15 May 2023 / Accepted: 20 July 2023
© The Author(s) 2023

Abstract

Background Technical skill assessment in surgery relies on expert opinion. Therefore, it is time-consuming, costly, and often lacks objectivity. Analysis of intraoperative data by artificial intelligence (AI) has the potential for automated technical skill assessment. The aim of this systematic review was to analyze the performance, external validity, and generalizability of AI models for 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 the 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, 1958 articles were identified, 50 articles met eligibility criteria and were analyzed. Motion data extracted from surgical videos ($n=25$) or kinematic data from robotic systems or sensors ($n=22$) were the most frequent input data for AI. Most studies used deep learning ($n=34$) and predicted technical skills using an ordinal assessment scale ($n=36$) with good accuracies in simulated settings. However, all proposed models were in development stage, only 4 studies were externally validated and 8 showed a low RoB.

Conclusion AI showed good performance in technical skill assessment in minimally invasive surgery. However, models often lacked external validity and generalizability. Therefore, models should be benchmarked using predefined performance metrics and tested in clinical implementation studies.

Keywords Technical skill assessment · Surgical skill assessment · Artificial intelligence · Minimally invasive surgery · Surgical data science

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 among others in bariatric [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.

In surgical education, technical skills of trainees are often assessed by staff surgeons through direct observations in the operating room. These instantaneous assessments by supervisors are frequently unstructured and

A previous version of this manuscript was published on medRxiv: https://www.medrxiv.org/content/10.1101/2022.11.08.22282058v1.

✉ Joël L. Lavanchy
joel.lavanchy@clarunis.ch

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

2 IHU Strasbourg, Strasbourg, France

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

4 ICube, CNRS, University of Strasbourg, Strasbourg, France

5 University Digestive Health Care Center Basel – Clarunis, PO Box, 4002 Basel, Switzerland

Published online: 16 August 2023

might only be snapshots of the actual technical performance of a trainee. Furthermore, they often lack objectivity due to peer review bias [5]. Aiming to improve the objectivity and construct validity of technical skill assessment, video-based assessment has been introduced [6]. Video-based assessment allows for retrospective review of full-length procedures or critical phases of an intervention by one or multiple experts. Despite the improvement of technical skill assessment by video-based assessment, it is still limited by the need for manual review of procedures by experts. Therefore, technical skill assessment is time-consuming, costly, and not scalable.

Automation of video-based assessment using artificial intelligence (AI) could lead to affordable, objective, and consistent technical skill assessment in real-time.

Despite the great potential of AI in technical skill assessment, it remains uncertain how accurate, valid, and generalizable AI models are to date. Therefore, the aim of this systematic review was to analyze the performance, external validity, and generalizability of AI models 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) [7] guidelines and was prospectively registered at PROSPERO (2021 CRD42021267714). The PRISMA checklist can be found in the Supplementary (Table S1).

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 at the University Library, University of Bern performed the literature search combining the following terms using Boolean operators: (1) Minimally invasive surgery including laparoscopic, or robotic surgery, and box model trainer. (2) AI including machine learning (ML), 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 are shown in the Supplementary (Table S2). The literature search was re-run prior to final analysis on February 25th, 2022 and May 31st, 2023.

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 (08/2016-08/2021, updated 02/2022 & 05/2023) in English language were included. Review articles, conference abstracts, comments, and letters to the editor were excluded.

Any form of quantitative or qualitative evaluation of manual surgical performance was considered a technical skill assessment.

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) [8]. 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 (a dataset is a defined collection of data either especially collected for the aim of the study or reused from previous studies), the assessment scale (ordinal scale vs. interval scale), the AI models used [ML or deep learning (DL)], the performance and the maturity level (development, validation, implementation) of AI models were extracted from the included studies. Missing or incomplete data was not imputed.

Performance metrics

The performance of AI models in technical skill assessment can be measured as accuracy, precision, recall, F1-score, and Area Under the Curve of Receiver Operator Characteristic (AUC-ROC). This paragraph gives a short definition of the used performance metrics. 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 (RoB) of the included studies was assessed using the modified version of Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) criteria [9]. This tool is commonly used for RoB evaluation in quality assessment studies. The quality of studies was evaluated using the modified Joanna Briggs Institute critical appraisal checklist for cross-sectional research in ML as used in [10, 11].

Results

The literature search retrieved a total of 1958 studies. After removing all duplicates, the remaining 1714 studies were screened by title and abstract. Thereafter, 120 studies remained, of which 70 were excluded after full-text screening. In summary, 50 studies [12–61] met the eligibility criteria and thus were included into this systematic review (Fig. 1). Two of the 50 studies [34, 61] included in this review were found to match the inclusion criteria during the process of full-text screening and were thus included through cross-referencing. Six studies [21, 29, 37, 45, 55, 58] were obtained during the re-run prior to final analysis six months after the initial literature search and 13 [13, 17, 34, 36, 38, 42, 48, 50, 52–54, 56, 61] during the second update on May 31st, 2023. Table 1 gives an overview of the 50 studies included in this systematic review (for full information extracted see Supplementary Table S3).

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 ($n=27$, 54%) [13, 14, 16–20, 22, 23, 25, 29, 32, 37, 43, 45–48, 50, 53–56, 58–61]. Laparoscopic simulators were the second most common setting for data collection ($n=15$, 30%) [12, 21, 24, 26, 27, 30, 31, 35, 36, 40, 42, 49, 51, 54, 57].

The most common tasks assessed were suturing ($n=31$, 62%) [13, 14, 16, 17, 19, 20, 22–25, 27, 29, 31, 32, 34, 35, 37, 45–51, 55–61], knot-tying ($n=21$, 42%) [13, 14, 16, 17, 19, 20, 22, 23, 29, 32, 35, 37, 45, 47, 48, 54–56, 59–61], and needle passing ($n=18$, 36%) [13, 14, 16, 17, 19, 22, 23, 29, 32, 37, 45, 47, 48, 55, 56, 59–61]. Other tasks assessed were peg transfers ($n=10$, 20%) [18, 24, 27, 30, 31, 36, 40, 42, 51, 54] and pattern cutting ($n=7$) [12, 21, 24, 26, 27, 31, 51]. All these tasks are part of the Fundamentals of Laparoscopic Surgery program, a well-established training curriculum for laparoscopic surgery with proven construct validity [62, 63].

Eleven studies (22%) [15, 28, 33, 34, 38, 39, 41, 43–45, 52] used data of real surgical procedures. Eight [15, 28, 33, 38, 39, 44, 45, 52] of them using videos of laparoscopic surgeries as for example laparoscopic cholecystectomies [28, 39] or laparoscopic pelvic lymph node dissections [15]. Three studies [34, 41, 43] used video data obtained from robotic surgeries such as robotic prostatectomy [41] or robotic thyroid surgery [43]. The tasks assessed in surgical procedures ranged from entire interventions to specific steps (e.g., lymph node dissection [15], clip application [39]).

Input data

Four different types of input data were used throughout the 50 studies: video data ($n=25$, 50%) [12, 13, 15, 21, 24, 25, 27, 28, 31–34, 36, 38, 39, 41–45, 50–52, 55, 58], kinematic data ($n=22$, 44%) [14, 16–20, 22, 23, 29, 35, 37, 40, 46–49, 54, 56, 57, 59–61], eye tracking data ($n=2$) [36, 53], and functional near-infrared spectroscopy (fNIRS) data ($n=2$) [26, 30]. Video recordings either from laparoscopic/robotic cameras or external cameras are used in 25 studies (50%). Kinematic data was obtained from DaVinci robotic systems (Intuitive Surgical Inc., CA, USA) in 17 studies (34%) [14, 16–20, 22, 23, 29, 37, 46–48, 56, 59–61] and from external sensors in five studies [35, 40, 49, 54, 57]. For example, electromyography sensors (Myo armband, Thalmic Labs, Ontario, CA) [35], optical sensors (Apple Watch, Apple, CA, USA) [40] or magnetic sensors attached to the instruments [49, 57] were used as external sensors to collect kinematic data. Two studies [26, 30] recorded fNIRS data from participants while they performed laparoscopic tasks. For example, Keles et al. [30] 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. Another approach was the tracking of eye gaze data. For example, Kuo et al. [36] used the Pro

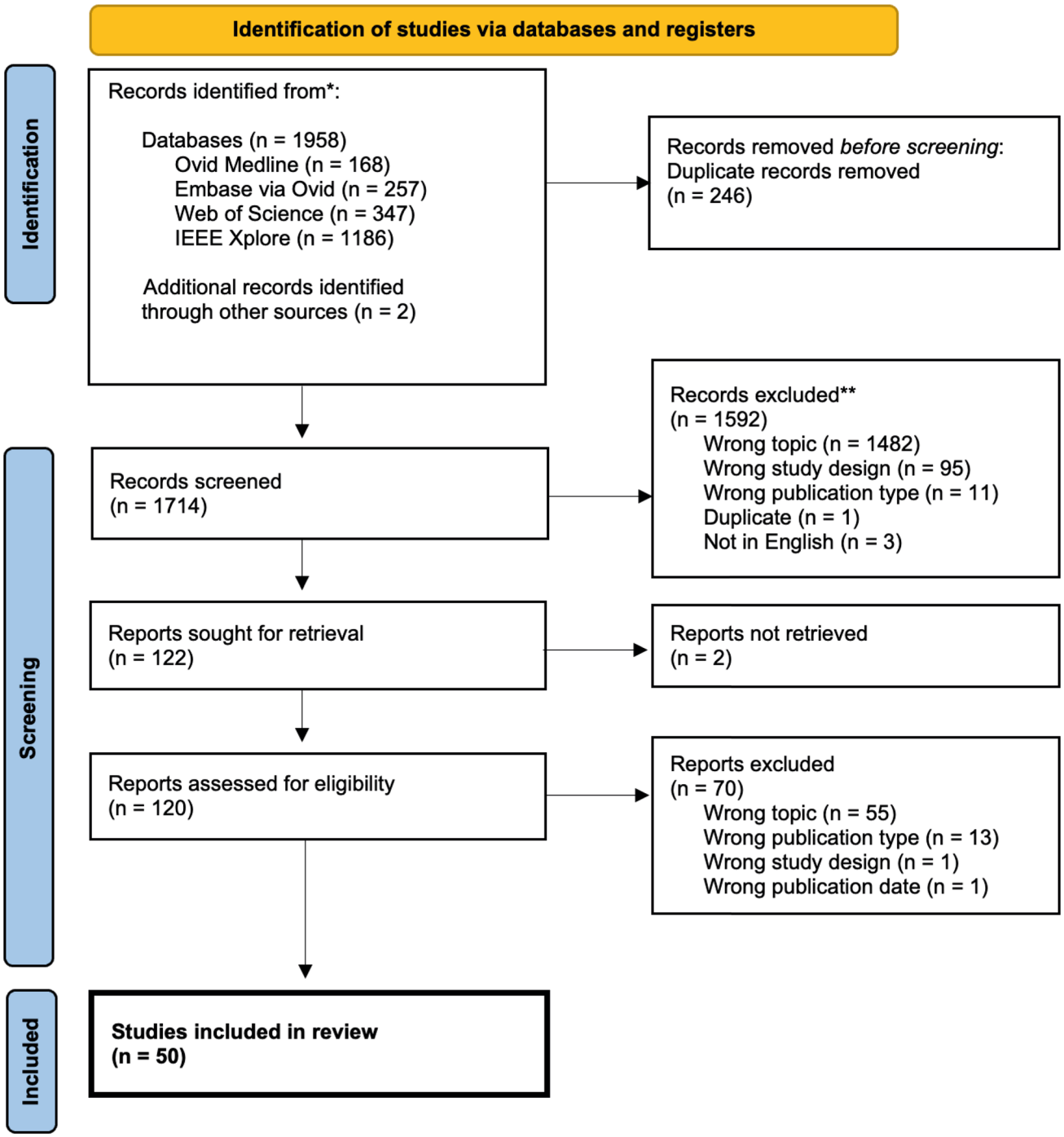

Fig. 1 PRISMA flow diagram of the study selection process (from PRISMA Statement 2020) [7]

Nano (Tobii Technology, Stockholm, Sweden) remote eye tracker to record gaze points during the tasks.

Datasets and external validation

Publicly available datasets were used in 22 studies (44%) [13, 14, 16, 17, 19, 20, 22, 23, 25, 28, 29, 32, 37, 45, 47, 48, 50, 55, 56, 59–61]. Of those, the JIGSAWS (Johns Hopkins University and Intuitive Surgical, Inc. Gesture and Skill Assessment Working Set) [64] dataset was most frequently used ($n = 21$, 42%) [13, 14, 16, 17, 19, 20, 22, 23, 25, 29, 32, 37, 45, 47, 48, 50, 55, 56, 59–61]. It contains video and kinematic data together with human annotated skill ratings of eight surgeons performing three surgical tasks in

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. [12]	2018	20	LS	PC	VR	Private	Binary (experienced, non-experienced)	DL	0.94	Dev
Anastasiou et al. [13]	2023	8	RS	SU, NP, KT	VR	JIGSAWS	Modified OSATS score	DL	na	Dev
Anh et al. [14]	2020	8	RS	SU, NP, KT	KD (dV)	JIGSAWS	N, I, E	DL	0.97	Dev
Baghdadi et al. [15]	2018	na	Lap	Pelvic lymph node dissection	VR	Private	PLACE score	ML	0.83	Dev
Benmansour et al. [16]	2018	6	RS	SU, NP, KT	KD (dV)	JIGSAWS	Custom score	DL	na	Dev
Benmansour et al. [17]	2023	8	RS	SU, NP, KT	KD (dV)	JIGSAWS	OSATS score	DL	na	Dev
Brown et al. [18]	2017	38	RS	PT	KD (dV)	Private	GEARS score (1–5: exact rating)	ML	0.75	Dev
Castro et al. [19]	2019	8	RS	SU, NP, KT	KD (dV)	JIGSAWS	N, I, E	DL	0.98	Dev
Fard et al. [20]	2017	8	RS	SU, KT	KD (dV)	JIGSAWS	Binary (N, E)	ML	0.9	Dev
Fathabadi et al. [21]	2021	na	LS	PC	VR	Private	Level A (excellent)—E (very bad)	DL	na	Dev
Fawaz et al. [22]	2019	8	RS	SU, NP, KT	KD (dV)	JIGSAWS	N, I, E	DL	1	Dev
Forestier et al. [23]	2018	8	RS	SU, NP, KT	KD (dV)	JIGSAWS	N, I, E	ML	0.96	Dev
French et al. [24]	2017	98	LS	PT, SU, PC	VR	Private	Binary (N, E)	ML	0.9	Dev
Funke et al. [25]	2019	8	RS	SU	VR	JIGSAWS	N, I, E	DL	1	Dev
Gao et al. [26]	2020	13	LS	PC	fNIRS data	Private	FLS score: pass/fail	DL	0.91	Dev
Islam et al. [27]	2016	52	LS	PT, SU, PC	VR	Private	Custom score	DL	na	Dev
Jin et al. [28]	2018	na	Lap	Lap cholecystectomy	VR	m2cai16-tools-location	Qualitative description	DL	na	Dev
Juarez-Villalobos et al. [29]	2021	8	RS	SU, NP, KT	KD (dV)	JIGSAWS	Binary (N, E)	ML	1	Dev
Keles et al. [30]	2021	33	LS	PT, threading	fNIRS data	Private	Binary (student vs. attending)	ML	~0.9	Dev
Kelly et al. [31]	2020	na	LS	PT, SU, PC, Clipping	VR	Private	Binary (N, E)	DL	0.97	Dev
Khalid et al. [32]	2020	8	RS	SU, NP, KT	VR	JIGSAWS	N, I, E	DL	0.77	Dev
Kitaguchi et al. [33]	2021	na	Lap	Lap colorectal surgery	VR	Private	ESSQS score	DL	0.75	Dev
Kiyasseh et al. [34]	2023	42	Rob	SU	VR	Private	Binary (low vs. high skill level)	ML	na	Dev
Kowalewski et al. [35]	2019	28	LS	SU, KT	KD (s)	Private	N, I, E	DL	0.7	Dev
Kuo et al. [36]	2022	10	LS	PT	VR, ETD	Private	N, I, E	ML, DL	0.83	Dev
Lajkó et al. [37]	2021	8	RS	SU, NP, KT	KD (dV)	JIGSAWS	Binary (N, E)	DL	0.84	Dev
Lam et al. [38]	2022	na	Lap	Lap gastric band insertion	VR	Private	Binary (trainee vs. expert)	DL	na	Dev
Lavanchy et al. [39]	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. [40]	2018	7	LS	PT	KD (s)	Private	N, I, E	DL	na	Dev
Law et al. [41]	2017	12	Rob	Robotic prostatectomy	VR	Private	Binary (good vs. poor)	ML, DL	0.92	Dev
Lazar et al. [42]	2023	27	LS	PT	VR	Private	N, I, E	DL	na	Dev

Table 1 (continued)

Author	Year	Population	Setting	Tasks	Input data	Dataset	Assessment	AI model	Accuracy	Maturity level
Lee et al. [43]	2020	1/ na	RS, Rob	Robotic thyroid surgery/ simulation	VR	Private	N, I, E	DL	0.83	Dev
Liu et al. [44]	2020	na	Lap	Lap gastrectomy	VR	Private	Modified OSATS score	DL	na	Dev
Liu et al. [45]	2021	8 / na	RS, Lap	SU, NP, KT / Lap surgery gastric cancer	VR	JIGSAWS	Modified OSATS score	ML	na	Dev
Lyman et al. [46]	2021	2	RS	SU of hepaticojejunostomy	KD (dV)	Private	Binary (N, I)	ML	0.89	Dev
Nguyen et al. [47]	2019	8	RS	SU, NP, KT	KD (dV)	JIGSAWS	N, I, E	DL	0.98	Dev
Oğul et al. [48]	2022	8 / 12	RS	SU, NP, KT / post and sleeve, pea on a peg and wire chaser	KD (dV)	JIGSAWS / ROSMA	Binary (JIGSAWS)	CNN	0.79	Dev
							Pairwise ranking (ROSMA)		0.75	
Oquendo et al. [49]	2018	32	LS	SU	KD (s)	Private	OSATS score (±4 points)	ML	0.89	Dev
Pan et al. [50]	2023	8	RS	SU	VR	JIGSAWS	Binary	DL	0.92	Dev
							N, I, E		0.85	
Pérez-Escamirosa et al. [51]	2019	43	LS	PT, SU, PC	VR	Private	Binary (experienced vs. non-experienced)	ML	0.98	Dev
Sasaki et al. [52]	2022	na	Lap	Lap sigmoidectomy	VR	Private	N, I, E (based on blood pixels)	ML	na	Dev
Shafiei et al. [53]	2023	11	RS	bluntD, retraction, coldD, burnD	ETD	Private	N, I, E	ML	0.96	Dev
Soangra et al. [54]	2022	26	LS, RS	PT, KT	KD (s)	Private	N, I, E	ML	0.58	Dev
Soleymani et al. [55]	2021	8	RS	SU, NP, KT	VR	JIGSAWS	N, I, E	DL	0.97	Dev
Soleymani et al. [56]	2022	8	RS	SU, NP, KT	KD (dV)	JIGSAWS	Qualitative description	DL	na	Dev
Uemura et al. [57]	2018	67	LS	SU	KD (s)	Private	Binary (N, E)	DL	0.79	Dev
Wang Y. et al. [58]	2021	18	RS	SU	VR	Private	N, I, E	DL	0.83	Dev
							GEARS score (± 1 point)		0.86	
Wang Z. et al. [59]	2018	8	RS	SU, NP, KT	KD (dV)	JIGSAWS	N, I, E	DL	0.95	Dev
Wang Z. et al. [60]	2018	8	RS	SU, NP, KT	KD (dV)	JIGSAWS	N, I, E	DL	0.96	Dev
Zia et al. [61]	2018	8	RS	SU, NP, KT	KD (dV)	JIGSAWS	N, I, E	ML	na	Dev
							Modified OSATS score		na	

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 S3

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, *bluntD* blunt dissection, *coldD* cold dissection (incl. cutting with scissors), *burnD* burn dissection, *VR* video recordings, *KD (dv)* kinematic data collected by Da Vinci systems, *fNIRS* functional near-infrared spectroscopy, *KD (s)* kinematic data collected by external sensors, *ETD* eye tracker data, *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, *AICS* AI confidence score, *ESSQS* Endoscopic Surgical Skill Qualification System, *OSATS* Objective Structured Assessment of Technical Skills, *dev* development

five-fold repetition in a robotic box model trainer. Oğul et al. [48] used another, newly released publicly available dataset called Robotic Surgical Maneuvers (ROSMA) dataset [65]. This dataset recorded using the Da Vinci Research kit provides dynamic and kinematic data as well as a performance score calculated from time to completion and penalty points. One study [28] extended the publicly available m2cai16-tool dataset [66] with locations of surgical tools and published it as m2cai16-tools-localisation dataset. Though, most studies ($n=28$, 56%) [12, 15, 18, 21, 24, 26, 27, 30, 31, 33–36, 38–44, 46, 49, 51–54, 57, 58] created private datasets, that were not publicly released. Most datasets ($n=46$, 92%) [12–23, 25–32, 35–40, 42–61] were monocentric. However, four studies used a multicentric dataset: French, et al. [24] used a multi-institutional dataset from three centers, Kitagutchi, et al. [33] drew a sample form a national Japan Society of Endoscopic Surgeons database, Kiyasseh et al. [34] trained on data from one center and deployed the model to two other centers, and Law, et al. [41] used a part of a statewide national quality improvement database collected by the Michigan Urological Surgical Improvement Collaborative. Four of the 50 studies included [23, 34, 45, 47], 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) (Fig. 2). In 36 of the studies (72%) an ordinal scale was applied [12, 14, 19–25, 29–32, 34, 35, 37–43, 46–48, 50–55, 57–61]. In 16 studies (32%) participants were categorized in two different skill levels [12, 20, 24, 29–31, 34, 37–39, 41, 46, 48, 50, 51, 57] and in 20 studies (40%) into three different expert levels (novice, intermediate, expert) [14, 19, 22, 23, 25, 32, 35, 36, 40, 43, 47, 50, 52–55, 58–61]. Twelve studies (24%) applied different proficiency scores: Pelvic Lymphadenectomy Assessment and Completion Evaluation (PLACE [67]), Fundamentals of Laparoscopic Surgery (FLS [68]), Endoscopic Surgical Skill Qualification System (ESSQS [69]), Objective Structured Assessment of Technical Skills (OSATS [70]), and Global Evaluative Assessment of Robotic Skills (GEARS [71]) [15–18, 26, 27, 33, 44, 45, 49, 58, 61].

AI models

All AI models in this review are either ML- or DL-based (Fig. 3). ML was applied in 19 studies (38%) [18, 20, 23, 24, 29, 30, 34–36, 39, 41, 45, 46, 49, 51–54, 61] and DL in 34 studies (68%) [12–14, 16, 17, 19, 21, 22, 25–28, 31–33, 35–44, 47, 48, 50, 55–60]. Three studies used a combination of ML and DL models [36, 39, 41].

Performance

The most common performance metric reported in the studies included in this systematic review is accuracy ($n=35$, 70%) [12, 14, 15, 18–20, 22–26, 29–33, 35–37, 39, 41, 43, 46–51, 53–55, 57–60]. Accuracies of the best performing models range between 0.58 and 1. Other performance metrics reported include F1-score ($n=9$) [25, 29, 32, 40, 50, 51, 53, 54, 60], recall also known as sensitivity ($n=11$, 22%) [12, 18, 25, 26, 32, 35, 50, 51, 53, 54, 60], specificity ($n=4$) [14, 26, 44, 45], and AUC-ROC ($n=4$) [14, 29, 44, 45]. Six studies [16, 21, 27, 28, 42, 56] did not report a performance metric at all.

Risk of bias and quality assessment

Eight of the included studies [18, 31, 33–35, 39, 49, 53] had an overall low probability of bias in the RoB assessment. The other studies had one ($n=15$, 30%) [25, 26, 30, 32, 36, 40, 41, 44, 45, 50–52, 56, 59, 61], two ($n=13$, 26%) [12, 13, 19, 22–24, 29, 42, 47, 48, 54, 55, 58], three

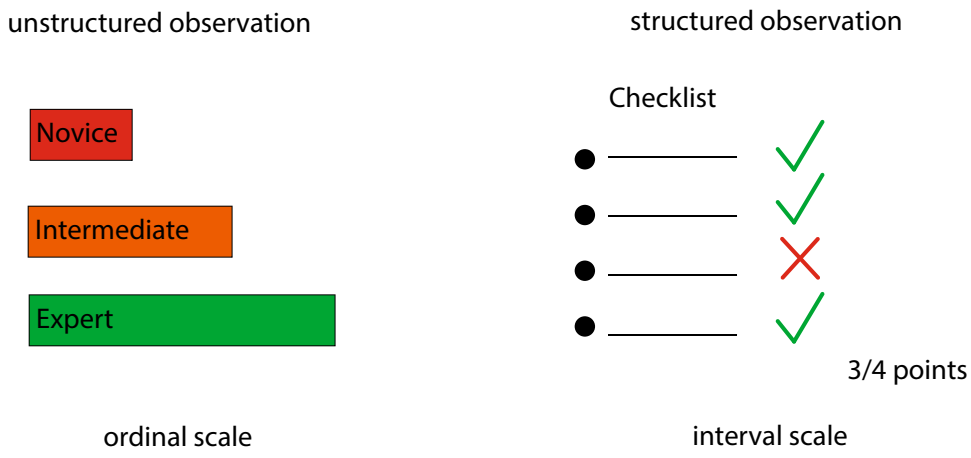

Fig. 2 Human technical skill assessment in minimally invasive surgery

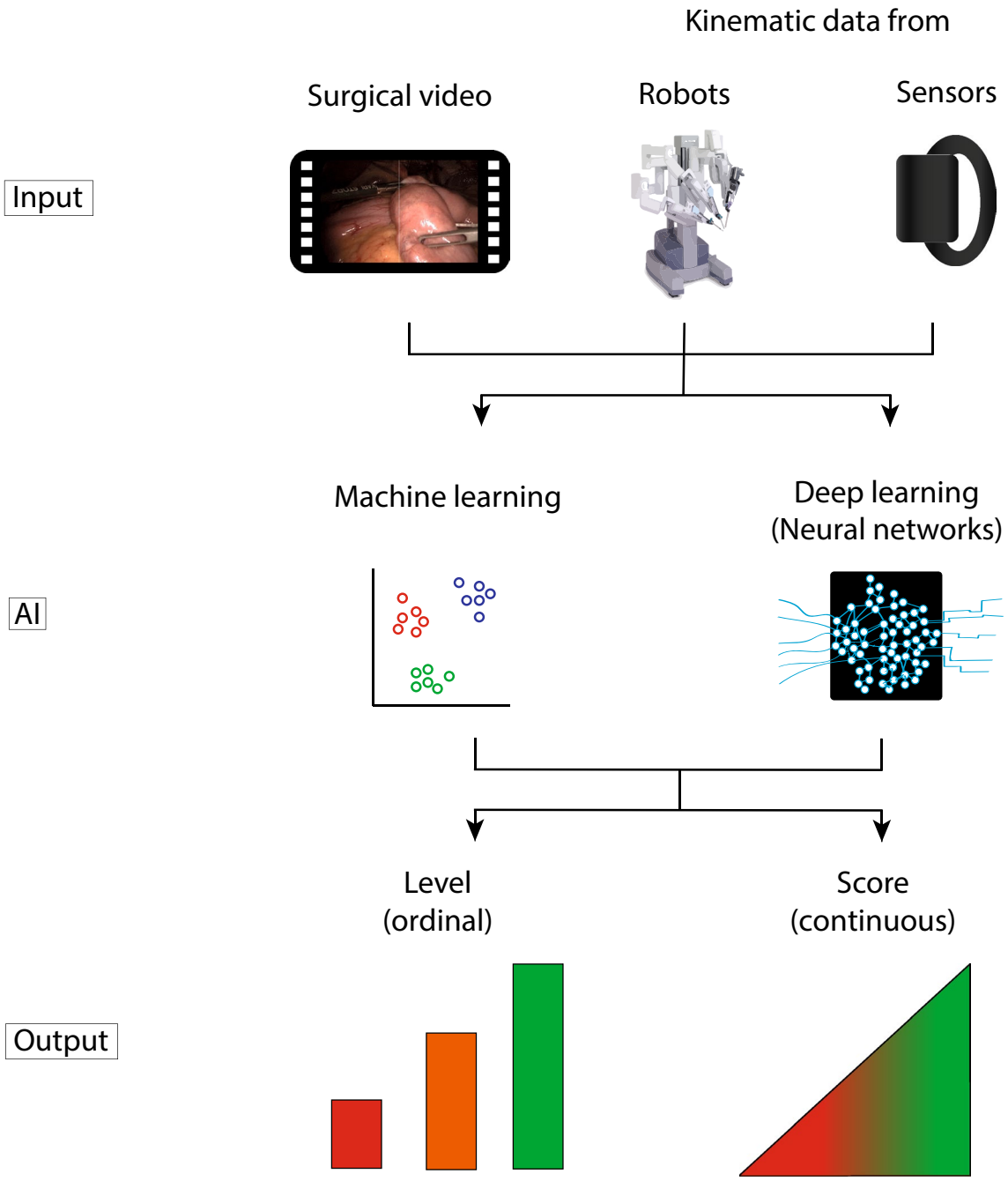

Fig. 3 Automated technical skill assessment in minimally invasive surgery by artificial intelligence

($n = 9$) [12, 13, 19, 22–24, 29, 42, 47, 48, 54, 55, 58], four ($n = 4$) [15, 17, 21, 46] or five criteria ($n = 1$) [16] at RoB. The full RoB assessment table is presented in the Supplementary (Table S4). The quality assessment of the included studies is displayed in Fig. 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 assessed the performance, external validity, and generalizability

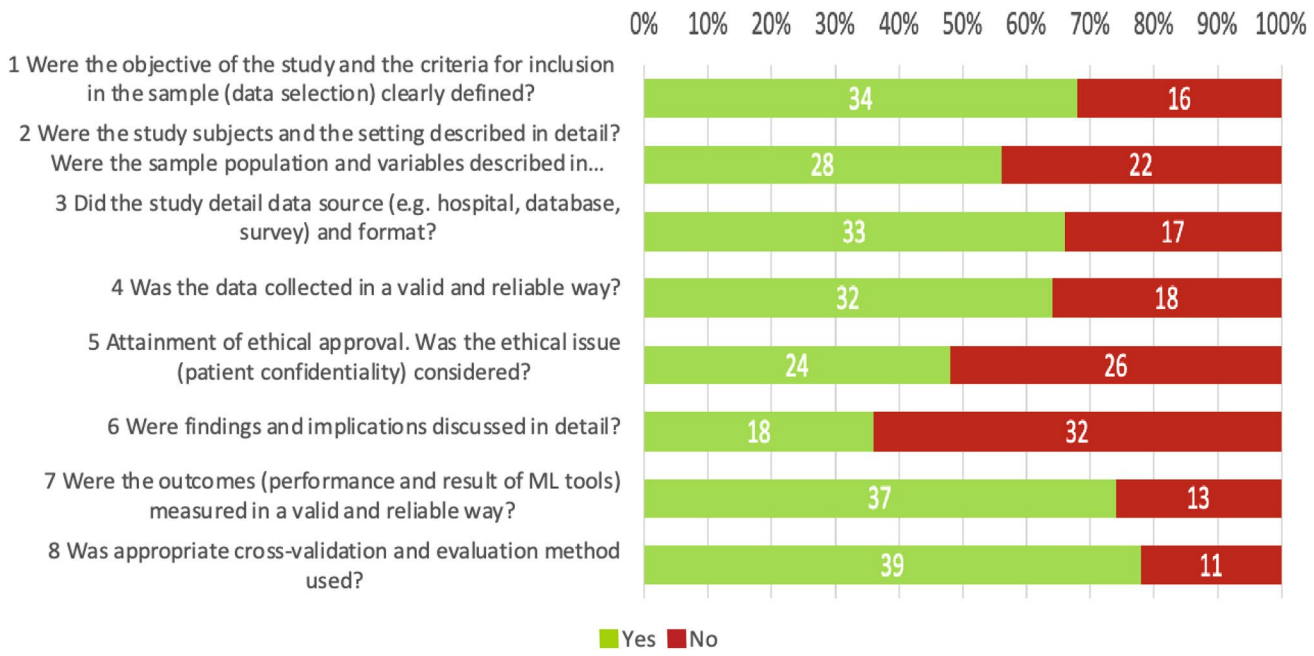

Fig. 4 Quality assessment of the included studies. The numbers within the bars represent the respective number of studies

of 50 included studies. A large variety of task, settings, datasets, and AI models have been studied.

In general, technical skill assessment involves either classifying skill levels in ordinal scales (e.g., novice, intermediate and expert) through unstructured observations or assessing performance intervals using structured checklists (e.g., Objective Structured Assessment of Technical Skills (OSATS) [70], Global Evaluative Assessment of Robotic Skills (GEARS) [71]) (Fig. 2). OSATS for example evaluates technical skills in seven dimensions (respect for tissue, time and motion, instrument handling, knowledge of instruments, use of assistants, flow of operation and forward planning, and knowledge of specific procedure) assigning a 5-point Likert scale from 1 (low skill) to 5 (high skill) to every dimension. Thus, 35 points is the maximum OSATS score reflecting highest technical skills. The ideal automated skill assessment model would not just output a skill level or overall score, but rather multiple dimensions of skill to provide actionable feedback to trainees.

Two subfields of AI are particularly used to extract and analyze motion data from surgical videos or robotic systems to assess technical skill: ML and 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 3 illustrates how different input data types are processed by AI models to predict technical skills.

The generalizability of the studies included in this systematic review is limited due to several fundamental differences between them. Most studies (56%) used private datasets of different settings, tasks, and sizes. However, 21 studies (42%) 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 construct validity of technical skill assessment models. However, in actual surgeries the experimental setting cannot be standardized and therefore, lacks reproducibility. This brings up the potential of virtual reality (VR) simulation in technical skill assessment [72]. VR enables simulation and assessment of complex tasks, as faced in actual surgery, without exposing patients to any harm. Furthermore, the management of rare

but far-reaching intraoperative adverse events like hemorrhage or vascular injury can be trained to proficiency in VR simulation.

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, novice vs. intermediate vs. expert). Dichotomous classification of technical skill in good or bad performance seems obvious, however, remains highly subjective. Skill levels distinguishing novice, intermediate, and expert surgeons are often based on quantitative measures like operative volume or years in training but fail to reflect individual technical skill levels. Other studies used different interval scales (OSATS scores, GEARS scores, or Likert scales). In contrast to expert annotated or quantitatively derived skill levels, OSATS and GEARS are scores, that have proven reliability and construct validity for direct observation or video-based assessment [70, 71]. However, for the purpose of AI model training there is no standardization of skill annotation. Which part of the task, using which ontology, and in which interval technical skill should be annotated by experts to reflect the overall skill level of study participants remains to be defined.

Most of the studies included in this systematic review have methodologic limitations. Overall, 84% of studies included in this review are at RoB. The quality assessment of the included studies revealed that only 36% of the studies discussed the findings and implications in detail. Furthermore, only four studies included in this review have a multicentric dataset. Only four 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 50 included studies, 35 (70%) report on accuracy. However, there is a large variation of reported performance metrics among the 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 [73] and clinical trials involving AI [74] 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 [75]. For surgical video analysis this effort still needs to be taken [76].

This systematic review is limited by the lack of generalizability and methodologic limitations of the included studies. Therefore, the direct comparison of AI models and a meta-analysis summarizing the evidence of included studies is not meaningful. To overcome these limitations valid and representative datasets, the use of predefined performance metrics, and external validation in clinical implementation studies will be essential to develop robust and generalizable AI models for technical skill assessment. In conclusion, AI has great potential to automate technical skill assessment in minimally invasive surgery. AI models showed moderate to high accuracy in technical skill assessment. 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.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00464-023-10335-z.

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

Funding Open access funding provided by University of Basel. Nicolas Padoy was funded by ANR grants ANR-10-IAHU-02 and ANR-20-CHIA-0029-01 and BPI grant CONDOR. Joël Lavanchy was funded by the Swiss National Science Foundation (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).

Data availability All data produced in the present work are contained in the manuscript.

Declarations

Disclosures Pietro Mascagni is the scientific director of the Global Surgical AI Collaborative. Nicolas Padoy received consulting fees from Caresyntax outside of the submitted work. Romina Pedrett, Guido Beldi, and Joël Lavanchy have no conflict of interest or financial ties to disclose.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Birkmeyer JD, Finks JF, O'Reilly A, Oerline M, Carlin AM, Nunn AR, Dimick J, Banerjee M, Birkmeyer NJO (2013) Surgical skill

and complication rates after bariatric surgery. N Engl J Med 369:1434–1442. https://doi.org/10.1056/NEJMsa1300625
2. Fecso AB, Bhatti JA, Stotland PK, Quereshy FA, Grantcharov TP (2019) Technical performance as a predictor of clinical outcomes in laparoscopic gastric cancer surgery. Ann Surg 270:115–120. https://doi.org/10.1097/SLA.0000000000002741
3. Stulberg JJ, Huang R, Kreutzer L, Ban K, Champagne BJ, Steele SR, Johnson JK, Holl JL, Greenberg CC, Bilimoria KY (2020) Association between surgeon technical skills and patient outcomes. JAMA Surg 155:960–968. https://doi.org/10.1001/jamasurg.2020.3007
4. Curtis NJ, Foster JD, Miskovic D, Brown CSB, Hewett PJ, Abbott S, Hanna GB, Stevenson ARL, Francis NK (2020) Association of surgical skill assessment with clinical outcomes in cancer surgery. JAMA Surg 155:590–598. https://doi.org/10.1001/jamasurg.2020.1004
5. Lendvay TS, White L, Kowalewski T (2015) Crowdsourcing to assess surgical skill. JAMA Surg 150:1086–1087. https://doi.org/10.1001/jamasurg.2015.2405
6. Aggarwal R, Grantcharov T, Moorthy K, Milland T, Darzi A (2008) Toward feasible, valid, and reliable video-based assessments of technical surgical skills in the operating room. Ann Surg 247:372–379. https://doi.org/10.1097/SLA.0b013e318160b371
7. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, Moher D (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372:n71. https://doi.org/10.1136/bmj.n71
8. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A (2016) Rayyan—a web and mobile app for systematic reviews. Syst Rev 5:210. https://doi.org/10.1186/s13643-016-0384-4
9. Whiting PF (2011) QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 155:529. https://doi.org/10.7326/0003-4819-155-8-201110180-00009
10. Anteby R, Horesh N, Soffer S, Zager Y, Barash Y, Amiel I, Rosin D, Gutman M, Klang E (2021) Deep learning visual analysis in laparoscopic surgery: a systematic review and diagnostic test accuracy meta-analysis. Surg Endosc 35:1521–1533. https://doi.org/10.1007/s00464-020-08168-1
11. Kwong MT, Colopy GW, Weber AM, Ercole A, Bergmann JHM (2019) The efficacy and effectiveness of machine learning for weaning in mechanically ventilated patients at the intensive care unit: a systematic review. Bio-Des Manuf 2:31–40. https://doi.org/10.1007/s42242-018-0030-1
12. Alonso-Silverio GA, Perez-Escamirosa F, Bruno-Sanchez R, Ortiz-Simon JL, Munoz-Guerrero R, Minor-Martinez A, Alarcon-Paredes A (2018) Development of a laparoscopic box trainer based on open source hardware and artificial intelligence for objective assessment of surgical psychomotor skills. Surg Innov 25:380–388. https://doi.org/10.1177/1553350618777045
13. Anastasiou D, Jin Y, Stoyanov D, Mazomenos E (2023) Keep your eye on the best: contrastive regression transformer for skill assessment in robotic surgery. IEEE Robot Autom Lett 8:1755–1762. https://doi.org/10.1109/LRA.2023.3242466
14. Anh NX, Chauhan S, Nataraja RM (2020) Towards near real-time assessment of surgical skills: A comparison of feature extraction techniques. Comput Methods Programs Biomed 187:105234. https://doi.org/10.1016/j.cmpb.2019.105234
15. Baghdadi A, Hussein AA, Ahmed Y, Guru KA, Cavuoto LA (2019) A computer vision technique for automated assessment of surgical performance using surgeons' console-feed videos. Int J Comput Assist Radiol Surg 14:697–707. https://doi.org/10.1007/s11548-018-1881-9
16. Benmansour M, Handouzi W, Malti A (2018) A neural network architecture for automatic and objective surgical skill assessment. In: Proceedings of 2018 3rd international conference on electrical sciences and technologies in Maghreb (CISTEM), pp 1–5. https://doi.org/10.1109/CISTEM.2018.8613550
17. Benmansour M, Malti A, Jannin P (2023) Deep neural network architecture for automated soft surgical skills evaluation using objective structured assessment of technical skills criteria. Int J Comput Assist Radiol Surg 18:929–937. https://doi.org/10.1007/s11548-022-02827-5
18. Brown JD, Brien CEO, Leung SC, Dumon KR, Lee DI, Kuchenbecker KJ (2017) Using contact forces and robot arm accelerations to automatically rate surgeon skill at peg transfer. IEEE Trans Biomed Eng 64:2263–2275. https://doi.org/10.1109/TBME.2016.2634861
19. Castro D, Pereira D, Zanchettin C, Macêdo D, Bezerra BLD (2019) Towards optimizing convolutional neural networks for robotic surgery skill evaluation. In: Proceedings of 2019 international joint conference on neural networks (IJCNN), pp 1–8. https://doi.org/10.1109/IJCNN.2019.8852341
20. Fard MJ, Ameri S, Darin Ellis R, Chinnam RB, Pandya AK, Klein MD (2018) Automated robot-assisted surgical skill evaluation: predictive analytics approach. Int J Med Robot 14:e1850. https://doi.org/10.1002/rcs.1850
21. Fathabadi FR, Grantner JL, Shebrain SA, Abdel-Qader I (2021) Surgical skill assessment system using fuzzy logic in a multi-class detection of laparoscopic box-trainer instruments. In: Proceedings of 2021 IEEE international conference on systems, man, and cybernetics (SMC), pp 1248–1253. https://doi.org/10.1109/SMC52423.2021.9658766
22. Fawaz HI, Forestier G, Weber J, Idoumghar L, Muller PA (2019) Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks. Int J Comput Assist Radiol Surg 14:1611–1617. https://doi.org/10.1007/s11548-019-02039-4
23. Forestier G, Petitjean F, Senin P, Despinoy F, Huaulme A, Fawaz HI, Weber J, Idoumghar L, Muller PA, Jannin P (2018) Surgical motion analysis using discriminative interpretable patterns. Artif Intell Med 91:3–11. https://doi.org/10.1016/j.artmed.2018.08.002
24. French A, Kowalewski TM, Lendvay TS, Sweet RM (2017) 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 12:1161–1170. https://doi.org/10.1007/s11548-017-1606-5
25. Funke I, Speidel S, Mees ST, Weitz J (2019) Video-based surgical skill assessment using 3D convolutional neural networks. Int J Comput Assist Radiol Surg 14:1217–1225. https://doi.org/10.1007/s11548-019-01995-1
26. Gao Y, Yan P, Kruger U, Cavuoto L, Schwaitzberg S, De S, Intes X (2021) Functional brain imaging reliably predicts bimanual motor skill performance in a standardized surgical task. IEEE Trans Biomed Eng 68:2058–2066. https://doi.org/10.1109/TBME.2020.3014299
27. Islam G, Kahol K, Li BX, Smith M, Patel VL (2016) Affordable, web-based surgical skill training and evaluation tool. J Biomed Inform 59:102–114. https://doi.org/10.1016/j.jbi.2015.11.002
28. Jin A, Yeung S, Jopling J, Krause J, Azagury D, Milstein A, Fei-Fei L (2018) Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In: Proceedings of 2018 IEEE winter conference on applications of computer vision (WACV), pp 691–699. https://doi.org/10.1109/WACV.2018.00081
29. Juarez-Villalobos L, Hevia-Montiel N, Perez-Gonzalez J (2021) Machine learning based classification of local robotic surgical skills in a training tasks set. In: Annual international conference

of the IEEE Engineering in Medicine & Biology Society 2021, pp 4596–4599. https://doi.org/10.1109/EMBC46164.2021.9629579
30. Keles HO, Cengiz C, Demiral I, Ozmen MM, Omurtag A (2021) High density optical neuroimaging predicts surgeons's subjective experience and skill levels. PLoS ONE 16:e0247117. https://doi.org/10.1371/journal.pone.0247117
31. Kelly JD, Kowalewski TM, Petersen A, Lendvay TS (2020) Bidirectional long short-term memory for surgical skill classification of temporally segmented tasks. Int J Comput Assist Radiol Surg 15:2079–2088. https://doi.org/10.1007/s11548-020-02269-x
32. Khalid S, Goldenberg M, Grantcharov T, Taati B, Rudzicz F (2020) Evaluation of deep learning models for identifying surgical actions and measuring performance. JAMA Netw OPEN. https://doi.org/10.1001/jamanetworkopen.2020.1664
33. Kitaguchi D, Takeshita N, Matsuzaki H, Igaki T, Hasegawa H, Ito M (2021) Development and validation of a 3-dimensional convolutional neural network for automatic surgical skill assessment based on spatiotemporal video analysis. JAMA Netw Open 4:e2120786. https://doi.org/10.1001/jamanetworkopen.2021.20786
34. Kiyasseh D, Ma R, Haque TF, Miles BJ, Wagner C, Donoho DA, Anandkumar A, Hung AJ (2023) A vision transformer for decoding surgeon activity from surgical videos. Nat Biomed Eng. https://doi.org/10.1038/s41551-023-01010-8
35. Kowalewski KF, Garrow CR, Schmidt MW, Benner L, Muller-Stich BP, Nickel F (2019) Sensor-based machine learning for workflow detection and as key to detect expert level in laparoscopic suturing and knot-tying. Surg Endosc Interv Tech 33:3732–3740. https://doi.org/10.1007/s00464-019-06667-4
36. Kuo RJ, Chen H-J, Kuo Y-H (2022) The development of an eye movement-based deep learning system for laparoscopic surgical skills assessment. Sci Rep 12:11036. https://doi.org/10.1038/s41598-022-15053-5
37. Lajko G, NagyneElek R, Haidegger T (2021) Endoscopic image-based skill assessment in robot-assisted minimally invasive surgery. Sensors. https://doi.org/10.3390/s21165412
38. Lam K, Lo FP-W, An Y, Darzi A, Kinross JM, Purkayastha S, Lo B (2022) Deep learning for instrument detection and assessment of operative skill in surgical videos. IEEE Trans Med Robot Bionics 4:1068–1071. https://doi.org/10.1109/TMRB.2022.3214377
39. Lavanchy JL, Zindel J, Kirtac K, Twick I, Hosgor E, Candinas D, Beldi G (2021) Automation of surgical skill assessment using a three-stage machine learning algorithm. Sci Rep 11:5197. https://doi.org/10.1038/s41598-021-84295-6
40. Laverde R, Rueda C, Amado L, Rojas D, Altuve M (2018) Artificial neural network for laparoscopic skills classification using motion signals from apple watch. In: Proceedings of 2018 40th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp 5434–5437. https://doi.org/10.1109/EMBC.2018.8513561
41. Law H, Zhang Y, Kim T-K, Miller D, Montie J, Deng J, Ghani K (2018) Surgeon technical skill assessment using computer vision-based analysis. J Urol 199:e1138
42. Lazar A, Sroka G, Laufer S (2023) Automatic assessment of performance in the FLS trainer using computer vision. Surg Endosc. https://doi.org/10.1007/s00464-023-10132-8
43. Lee D, Yu HW, Kwon H, Kong H-J, Lee KE, Kim HC (2020) Evaluation of surgical skills during robotic surgery by deep learning-based multiple surgical instrument tracking in training and actual operations. J Clin Med 9:1–15. https://doi.org/10.3390/jcm9061964
44. Liu D, Jiang T, Wang Y, Miao R, Shan F, Li Z (2020) Clearness of operating field: a surrogate for surgical skills on in vivo clinical data. Int J Comput Assist Radiol Surg 15:1817–1824. https://doi.org/10.1007/s11548-020-02267-z
45. Liu D, Li Q, Jiang T, Wang Y, Miao R, Shan F, Li Z (2021) Towards unified surgical skill assessment. In: Proceedings of 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 9517–9526. https://doi.org/10.1109/CVPR46437.2021.00940
46. Lyman WB, Passeri MJ, Murphy K, Iannitti DA, Martinie JB, Baker EH, Vrochides D, Siddiqui IA, Khan AS (2021) An objective approach to evaluate novice robotic surgeons using a combination of kinematics and stepwise cumulative sum (CUSUM) analyses. Surg Endosc 35:2765–2772. https://doi.org/10.1007/s00464-020-07708-z
47. Nguyen XA, Ljuhar D, Pacilli M, Nataraja RM, Chauhan S (2019) Surgical skill levels: classification and analysis using deep neural network model and motion signals. Comput Methods Programs Biomed 177:1–8. https://doi.org/10.1016/j.cmpb.2019.05.008
48. Oğul BB, Gilgien M, Özdemir S (2022) Ranking surgical skills using an attention-enhanced Siamese network with piecewise aggregated kinematic data. Int J Comput Assist Radiol Surg 17:1039–1048. https://doi.org/10.1007/s11548-022-02581-8
49. Oquendo YA, Riddle EW, Hiller D, Blinman TA, Kuchenbecker KJ (2018) Automatically rating trainee skill at a pediatric laparoscopic suturing task. Surg Endosc Interv Tech 32:1840–1857. https://doi.org/10.1007/s00464-017-5873-6
50. Pan M, Wang S, Li J, Li J, Yang X, Liang K (2023) An automated skill assessment framework based on visual motion signals and a deep neural network in robot-assisted minimally invasive surgery. Sensors 23:4496. https://doi.org/10.3390/s23094496
51. Perez-Escamirosa F, Alarcon-Paredes A, Alonso-Silverio GA, Oropesa I, Camacho-Nieto O, Lorias-Espinoza D, Minor-Martinez A (2020) Objective classification of psychomotor laparoscopic skills of surgeons based on three different approaches. Int J Comput Assist Radiol Surg 15:27–40. https://doi.org/10.1007/s11548-019-02073-2
52. Sasaki S, Kitaguchi D, Takenaka S, Nakajima K, Sasaki K, Ogane T, Takeshita N, Gotohda N, Ito M (2022) Machine learning-based automatic evaluation of tissue handling skills in laparoscopic colorectal surgery: a retrospective experimental study. Ann Surg. https://doi.org/10.1097/SLA.0000000000005731
53. Shafiei SB, Shadpour S, Mohler JL, Attwood K, Liu Q, Gutierrez C, Toussi MS (2023) Developing surgical skill level classification model using visual metrics and a gradient boosting algorithm. Ann Surg Open 4:e292. https://doi.org/10.1097/AS9.0000000000000292
54. Soangra R, Sivakumar R, Anirudh ER, Sai Viswanth Reddy Y, John EB (2022) Evaluation of surgical skill using machine learning with optimal wearable sensor locations. PLOS ONE 17:e0267936. https://doi.org/10.1371/journal.pone.0267936
55. Soleymani A, Asl AAS, Yeganejou M, Dick S, Tavakoli M, Li X (2021) Surgical skill evaluation from robot-assisted surgery recordings. In: Proceedings of 2021 international symposium on medical robotics (ISMR), pp 1–6. https://doi.org/10.1109/ISMR48346.2021.9661527
56. Soleymani A, Li X, Tavakoli M (2022) A domain-adapted machine learning approach for visual evaluation and interpretation of robot-assisted surgery skills. IEEE Robot Autom Lett 7:8202–8208. https://doi.org/10.1109/LRA.2022.3186769
57. Uemura M, Tomikawa M, Akahoshi T, Lefor AK, Hashizume M, Miao T, Souzaki R, Ieiri S (2018) Feasibility of an AI-based measure of the hand motions of expert and novice surgeons. Comput Math Methods Med 2018:9873273. https://doi.org/10.1155/2018/9873273
58. Wang Y, Dai J, Morgan TN, Elsaied M, Garbens A, Qu X, Steinberg R, Gahan J, Larson EC (2021) Evaluating robotic-assisted surgery training videos with multi-task convolutional neural networks. J Robot Surg. https://doi.org/10.1007/s11701-021-01316-2

59. Wang ZH, Fey AM (2018) Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. Int J Comput Assist Radiol Surg 13:1959–1970. https://doi.org/10.1007/s11548-018-1860-1
60. Wang Z, Fey AM (2018) SATR-DL: improving surgical skill assessment and task recognition in robot-assisted surgery with deep neural networks. In: Annual international conference of the IEEE Engineering in Medicine and Biology Society 2018, pp 1793–1796. https://doi.org/10.1109/EMBC.2018.8512575
61. Zia A, Essa I (2018) Automated surgical skill assessment in RMIS training. Int J Comput Assist Radiol Surg 13:731–739. https://doi.org/10.1007/s11548-018-1735-5
62. Zendejas B, Ruparel RK, Cook DA (2016) Validity evidence for the Fundamentals of Laparoscopic Surgery (FLS) program as an assessment tool: a systematic review. Surg Endosc 30:512–520. https://doi.org/10.1007/s00464-015-4233-7
63. Fried GM, Feldman LS, Vassiliou MC, Fraser SA, Stanbridge D, Ghitulescu G, Andrew CG (2004) Proving the value of simulation in laparoscopic surgery. Ann Surg 240:518–528. https://doi.org/10.1097/01.sla.0000136941.46529.56
64. Gao Y, Vedula SS, Reiley CE, Ahmidi N, Varadarajan B, Lin HC, Tao L, Zappella L, Béjar B, Yuh DD, Chen CCG, Vidal R, Khudanpur S, Hager GD (2014) 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
65. Rivas-Blanco I, Pérez-del-Pulgar C, Mariani A, Quaglia C, Tortora G, Reina AJ V (2021) A surgical dataset from the da Vinci Research Kit for task automation and recognition. arXiv preprint. https://doi.org/10.48550/arXiv.2102.03643
66. Twinanda AP, Shehata S, Mutter D, Marescaux J, Marescaux J, de Mathelin M, Padoy N (2017) EndoNet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans Med Imaging 36:86–97. https://doi.org/10.1109/tmi.2016.2593957
67. Hussein AA, Ghani KR, Peabody J, Sarle R, Abaza R, Eun D, Hu J, Fumo M, Lane B, Montgomery JS, Hinata N, Rooney D, Comstock B, Chan HK, Mane SS, Mohler JL, Wilding G, Miller D, Guru KA, Michigan Urological Surgery Improvement Collaborative and Applied Technology Laboratory for Advanced Surgery Program (2017) Development and validation of an objective scoring tool for robot-assisted radical prostatectomy: prostatectomy assessment and competency evaluation. J Urol 197:1237–1244. https://doi.org/10.1016/j.juro.2016.11.100
68. Peters JH, Fried GM, Swanstrom LL, Soper NJ, Sillin LF, Schirmer B, Hoffman K, the SAGES FLS Committee (2004) Development and validation of a comprehensive program of education and assessment of the basic fundamentals of laparoscopic surgery. Surgery 135:21–27. https://doi.org/10.1016/S0039-6060(03)00156-9
69. Mori T, Kimura T, Kitajima M (2010) Skill accreditation system for laparoscopic gastroenterologic surgeons in Japan. Minim Invasive Ther Allied Technol 19:18–23. https://doi.org/10.3109/13645700903492969
70. Martin JA, Regehr G, Reznick R, Macrae H, Murnaghan J, Hutchison C, Brown M (1997) Objective structured assessment of technical skill (OSATS) for surgical residents: objective structured assessment of technical skill. Br J Surg 84:273–278. https://doi.org/10.1046/j.1365-2168.1997.02502.x
71. Goh AC, Goldfarb DW, Sander JC, Miles BJ, Dunkin BJ (2012) Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. J Urol 187:247–252. https://doi.org/10.1016/j.juro.2011.09.032
72. Winkler-Schwartz A, Mirchi N, Ponnudurai N, Yilmaz R, Ledwos N, Karlik B, Del Maestro RF, Bissonnette V, Siyar S, Azarnoush H (2019) Artificial intelligence in medical education: best practices using machine learning to assess surgical expertise in virtual reality simulation. J Surg Educ 76:1681–1690. https://doi.org/10.1016/j.jsurg.2019.05.015
73. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, Denniston AK, Faes L, Geerts B, Ibrahim M, Liu X, Mateen BA, Mathur P, McCradden MD, Morgan L, Ordish J, Rogers C, Saria S, Ting DSW, Watkinson P, Weber W, Wheatstone P, McCulloch P (2022) Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ 377:e070904. https://doi.org/10.1136/bmj-2022-070904
74. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group (2020) Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 26:1364–1374. https://doi.org/10.1038/s41591-020-1034-x
75. Maier-Hein L, Reinke A, Godau P, Tizabi MD, Büttner F, Christodoulou E, Glocker B, Isensee F, Kleesiek J, Kozubek M, Reyes M, Riegler MA, Wiesenfarth M, Kavur AE, Sudre CH, Baumgartner M, Eisenmann M, Heckmann-Nötzel D, Rädsch AT, Acion L, Antonelli M, Arbel T, Bakas S, Benis A, Blaschko M, Cardoso MJ, Cheplygina V, Cimini BA, Collins GS, Farahani K, Ferrer L, Galdran A, van Ginneken B, Haase R, Hashimoto DA, Hoffman MM, Huisman M, Jannin P, Kahn CE, Kainmueller D, Kainz B, Karargyris A, Karthikesalingam A, Kenngott H, Kofler F, Kopp-Schneider A, Kreshuk A, Kurc T, Landman BA, Litjens G, Madani A, Maier-Hein K, Martel AL, Mattson P, Meijering E, Menze B, Moons KGM, Müller H, Nichyporuk B, Nickel F, Petersen J, Rajpoot N, Rieke N, Saez-Rodriguez J, Sánchez CI, Shetty S, van Smeden M, Summers RM, Taha AA, Tiulpin A, Tsaftaris SA, Van Calster B, Varoquaux G, Jäger PF (2022) Metrics reloaded: pitfalls and recommendations for image analysis validation. arXiv preprint. https://doi.org/10.48550/arXiv.2206.01653
76. Kitaguchi D, Watanabe Y, Madani A, Hashimoto DA, Meireles OR, Takeshita N, Mori K, Ito M, on behalf of the Computer Vision in Surgery International Collaborative (2022) Artificial intelligence for computer vision in surgery: a call for developing reporting guidelines. Ann Surg 275:e609–e611. https://doi.org/10.1097/SLA.0000000000005319

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.