



Surgical computer vision for intraoperative decision-support: a scoping review on performance metrics and readiness for real-time deployment

 Jayvee Buchanan^{1,2} , Saxon Connor² , John Pearson³ , Bruce Carey-Smith¹, Tim Eglinton^{1,2}

Keywords:

Artificial intelligence, computer vision, deep learning, surgical video analysis, intraoperative decision support, minimally invasive surgery, laparoscopic surgery

Citation:

Buchanan J, Connor S, Pearson J, Carey-Smith B, Eglinton T. Surgical computer vision for intraoperative decision-support: a scoping review on performance metrics and readiness for real-time deployment. *Art Int Surg.* 2026;6:150-70. <https://dx.doi.org/10.20517/ais.2025.76>

Received: 11 Aug 2025

First Decision: 28 Nov 2025

Revised: 14 Jan 2026

Accepted: 4 Feb 2026

Published: 28 Feb 2026

Academic Editor:

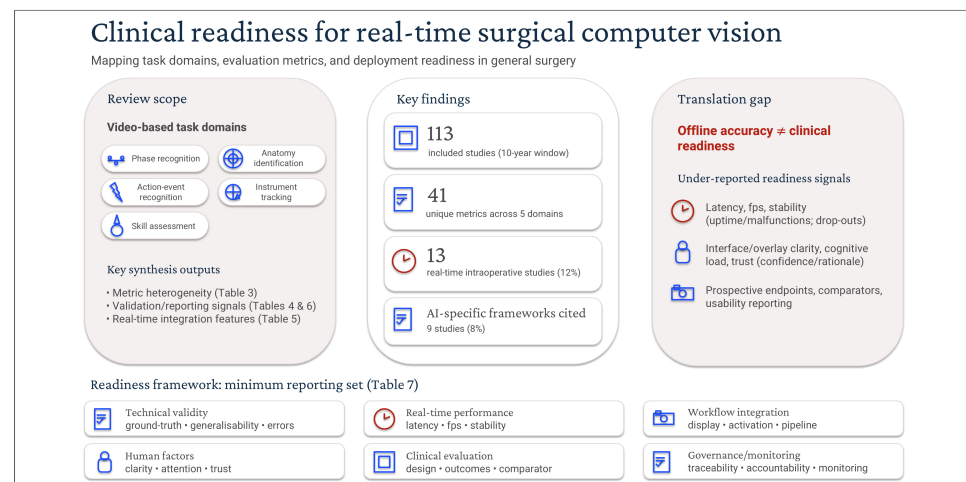
Eyad Elyan

Copy Editor:

Xing-Yue Zhang

Production Editor:

Xing-Yue Zhang



Abstract

Background: Real-time computer vision-based artificial intelligence (CV-AI) systems for surgical video analysis are rapidly advancing. Current evaluation strategies and clinical-readiness reporting, however, remain inconsistent. This scoping review mapped contemporary CV-AI task domains, performance metrics, and evidence of readiness for real-time intraoperative deployment within general surgery.

Methods: This study followed Joanna Briggs Institute methodology for scoping reviews, and was reported in accordance with PRISMA-ScR. Eligible studies were identified by systematic literature search of the MEDLINE, Embase, PubMed, and Scopus databases. All studies published between 1 June 2015 and 1 June 2025 were eligible.

Results: A total of 490 articles were screened, with 113 studies meeting the inclusion criteria after full-text review. Retrospective feasibility analyses predominated, with only 13 studies (12%) evaluating real-time intraoperative integration. Five task domains were identified (phase recognition, anatomy identification, action-event recognition, instrument

¹Department of Surgery and Critical Care, University of Otago Christchurch, Christchurch 8140, New Zealand.

²Department of Surgery, Christchurch Hospital, Te Whatu Ora Waitaha Canterbury, Christchurch 8140, New Zealand.

³Biostatistics and Computational Biology Unit, University of Otago Christchurch, Christchurch 8140, New Zealand.

Correspondence to: Dr. Jayvee Buchanan, Department of Surgery and Critical Care, University of Otago Christchurch, Christchurch 8140, New Zealand. E-mail: jayvee.buchanan3@cdhb.health.nz

tracking, and skill-assessment). Forty-one unique performance metrics were reported, with predominant use of discrimination-style summary measures (e.g., accuracy, recall, F1 score), and comparatively sparse reporting of class imbalance, boundary-aware (e.g., Hausdorff distance) or real-time workflow factors (e.g., latency/stability, interface design, surgeon feedback). External validation was described in 13 (12%) studies. Nine studies (8%) referenced artificial intelligence-specific reporting frameworks.

Conclusion: Surgical CV-AI is advancing technically, but remains predominantly at an early feasibility stage. Variability in current metric application and limited real-time clinical evaluation limit potential for comparability, applicability and widespread adoption. Standardised metrics, evaluation frameworks, prospective clinical trials, and collaborative end-user engagement are critical to translate conceptual promise to reliable real-time decision-support tools that support surgeon judgement and integrate seamlessly into routine operative workflows.

INTRODUCTION

Computer vision (CV)-based artificial intelligence (AI) (CV-AI) systems in general surgery are rapidly evolving, driven by substantial advances in image-based computational analysis^[1,2]. Video footage captured during minimally invasive surgical procedures can be leveraged and analysed in real-time using CV machine-learning techniques, with potential to automate surgical phase recognition, anatomical landmark identification, instrument tracking, and objective skill-assessment tasks^[1-6]. Through integration into routine operative workflows, CV-AI is poised to significantly enhance surgical performance and decision-making^[1,7,8]. These efforts are broadly classified under three quality-improvement domains: superior patient safety or outcomes, enhanced surgeon training, and optimisation of theatre processes.

Robust quantitative measures are essential for evaluating surgical CV-AI systems, with metric selection often specific to the task addressed^[9,10]. Classification models, such as for phase recognition and operative-difficulty grading, typically report accuracy, precision, recall/sensitivity, F1 score, or area under receiver-operating-curve (AUROC), reflecting prediction and class-balance^[5,9,11]. Object-detection tasks such as instrument tracking employ spatial metrics such as intersection-over-union (IoU) and mean average precision (mAP), evaluating positional accuracy^[9]. Segmentation models, for identifying anatomical or instrument boundaries, frequently apply Dice coefficient and IoU, quantifying overlap with ground-truth segmentations^[9,12-14]. Collectively, these metrics facilitate performance benchmarking and comparative analyses between algorithms^[10]. Although current CV-AI literature in general surgery largely comprises retrospective feasibility studies, a diverse set of approaches and metrics have been reported^[9,13]. This variability reinforces the need to assess how CV-AI is evaluated as systems advance toward clinical integration, establishing the basis for evaluation of surgeon-centred measures such as impact on cognitive load, end-user experience, and patient outcomes^[6,15,16]. Defining the contemporary CV-AI landscape also serves as a necessary step toward developing standardised evaluation frameworks supporting clinically applicable real-time deployment of surgical intelligence in the operating theatre^[9,13,15,17].

The primary aim of this scoping review was to map contemporary CV-AI task domains and performance metrics used to evaluate them within general surgery, clarifying current CV-AI translational gaps toward safe, clinician-facing decision-support within the operating theatre. Secondary aims were to characterise clinical-readiness signals, including extent of real-time intraoperative integration, workflow integration features, and uptake of recognised reporting guidelines.

METHODS

This study was conducted in accordance with Joanna Briggs Institute (JBI) methodology for scoping reviews^[18]. The PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analysis

extension for Scoping Reviews) checklist was completed [Supplementary Figure 1]^[19]. The protocol was registered *a priori* on Open Science Framework (<https://osf.io/xgyv>).

Eligibility criteria

Peer-reviewed studies of any design meeting all the following criteria were included: (1) specific to the discipline of general surgery or its subspecialties; (2) evaluated laparoscopic or robot-assisted surgical procedures; (3) described a CV-AI system; and (4) performed retrospective or prospective analyses. Open procedures were excluded given the requirement for continuous visual data input, such as that provided by laparoscopic cameras. Studies that described minimally invasive transanal surgical techniques, endoluminal endoscopic or radiological CV-AI, or those lacking practical intraoperative relevance for surgeons (including commercial platforms primarily providing automated retrospective case-feedback without directly assessing algorithm performance), were excluded. Studies that reported a subset of general surgical operations were eligible if the findings were stratified accordingly and metrics were reported separately. Conference abstracts and study protocols were also excluded.

Search strategy

Eligible studies were identified by systematic literature search of the MEDLINE, Embase, PubMed, and Scopus databases. Search terms included a combination of: (laparoscopic OR robotic) AND (surgical video OR computer vision) AND (artificial intelligence). Boolean operators AND and OR were used to connect the search terms. The search was restricted to studies published in peer-reviewed journals in English-text only. All studies published between 1 June 2015 and 1 June 2025 were eligible. Reference lists of included articles were examined, with citation tracking and hand-searching also performed. Search records were imported into EndNote (EndNote 20, Clarivate, Philadelphia, PA, USA), with duplications removed. Covidence review software was used for screening and full-text review (Covidence, Veritas Health Innovation, Melbourne, Australia).

Data extraction and synthesis

All titles and abstracts retrieved were screened by a single reviewer (JB) due to feasibility constraints. To mitigate selection bias, eligibility criteria were pre-specified per the registered protocol, with uncertainties regarding application of eligibility criteria resolved through consensus discussion with senior authors (SC/TE). Studies satisfying the eligibility criteria proceeded to full-text review. Results from each study were tabulated in a spreadsheet with data charted in accordance with JBI guidelines^[20]. Detailed quantitative syntheses combining individual study results and formal critical appraisals/risk-of-bias assessments were not performed, as this lay outside the exploratory objectives of this scoping review.

RESULTS

Search results and study characteristics

Four hundred and ninety articles were screened by title and abstract following removal of duplicate entries. Forty-nine full-text articles underwent eligibility assessment, with a further 11 excluded. Reference and citation tracking yielded an additional 75 publications, manually checked to confirm structural similarity, resulting in 113 studies^[21-133] in the final dataset [Figure 1]. An increasing trend in annual CV-AI publications was observed [Figure 2]. Fifty-five studies (47%) evaluated general procedures, mainly in laparoscopic cholecystectomy (LC) [Table 1]. Colorectal and upper gastrointestinal procedures were the most frequently studied subspecialty domains, accounting for 25% and 19% of studies. The predominant focus was utility of CV-AI in laparoscopic procedures (89%), with robot-assisted applications comprising the remainder (11%). Training datasets ranged between five and 700 surgical videos^[29,73].

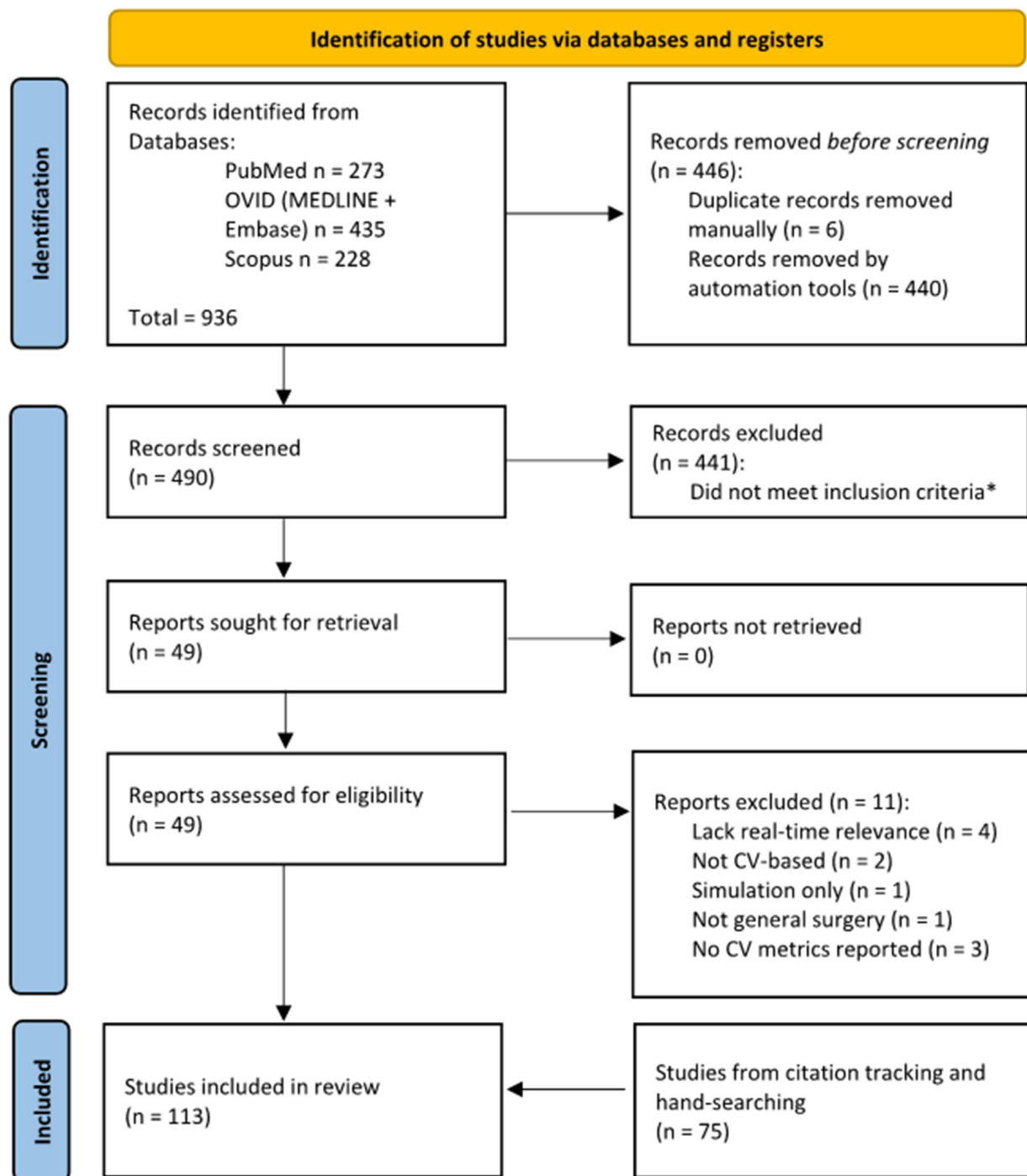


Figure 1. PRISMA flowchart. *Records excluded at title/abstract screening stage failed to meet all of the following inclusion criteria: (1) specific to general surgery or its subspecialties; (2) evaluated laparoscopic or robot-assisted procedures; (3) described a computer vision (CV)-based artificial intelligence (AI) system; (4) performed retrospective or prospective analyses. Conference abstracts and study protocols were also excluded.

Task domains and performance metrics

Five principal CV-AI task domains were identified, with performance metric selection dependent on the task evaluated [Table 2]. Anatomy detection and phase recognition were prioritised, evaluated in 88 studies (78%). Anatomical detection models were used for identifying important surgical landmarks such as the cystic duct/artery^[37,75,86,92], common bile duct (CBD)^[37,86,119,128], pancreas^[87], ileocolic vessels^[97], inferior

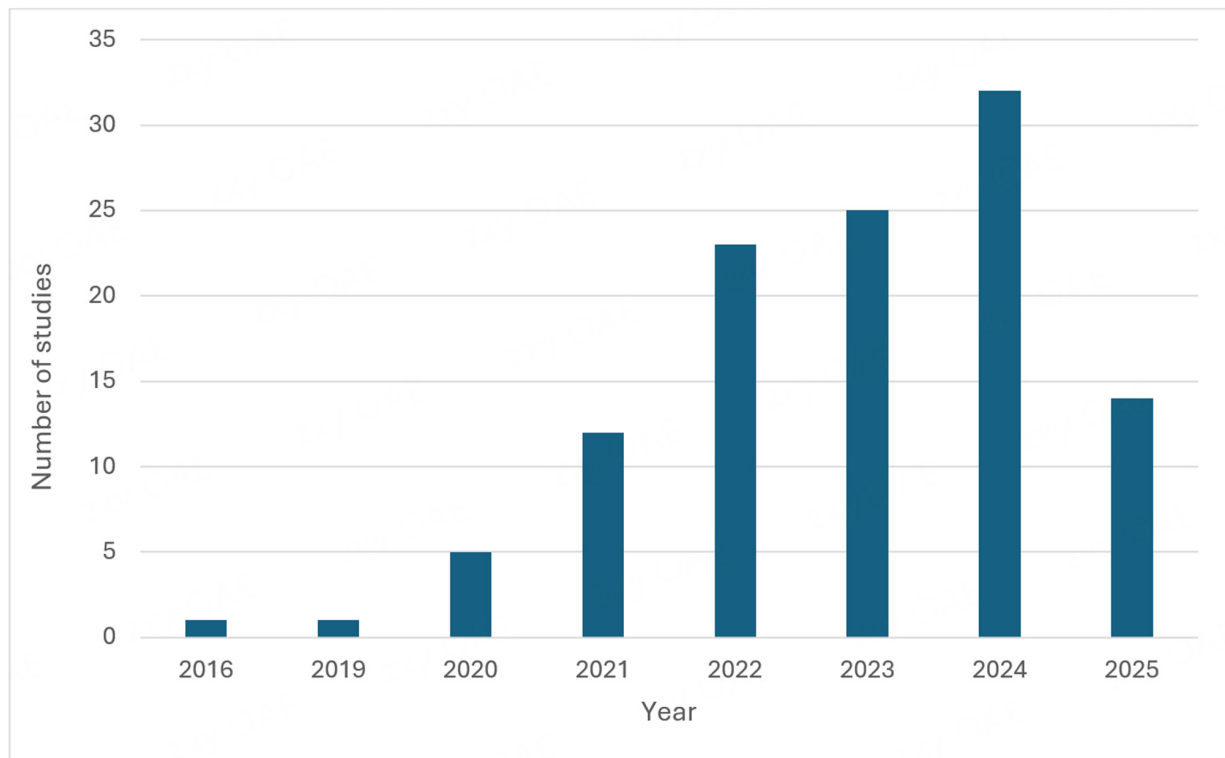


Figure 2. Computer vision-based artificial intelligence (CV-AI) publication trend over time. Data are current as of 1 June 2025.

Table 1. Included studies by surgical subspecialty and procedures

Subspecialty	Studies (n, %)*	Procedures (n)
General surgery	55 (47)	Laparoscopic cholecystectomy (47), laparoscopic inguinal hernia repair [†] (7), robot-assisted inguinal hernia repair (1)
Colorectal surgery	29 (25)	Laparoscopic colon resection, right (2); laparoscopic colon resection, left (including high anterior resection) (15), laparoscopic rectal resection (including low anterior resection) (1), other laparoscopic colorectal procedures [‡] (8), robot-assisted rectal procedures [§] (3)
Endocrine	2 (2)	Laparoscopic adrenalectomy (2)
Hepatopancreatobiliary	7 (6)	Laparoscopic hepatectomy (4), laparoscopic pancreatectomy (Whipple or distal) (2), robotic pancreatojejunostomy (1)
Transplant	1 (1)	Laparoscopic donor hepatectomy (1)
Upper gastrointestinal/bariatric surgery	22 (19)	Laparoscopic/thoracoscopic oesophagectomy (1), laparoscopic gastrectomy (total or distal) [¶] (7), laparoscopic sleeve gastrectomy (2), laparoscopic gastric bypass (3), other laparoscopic procedures [•] (1), robotic gastrectomy (5), robotic oesophagectomy (3)

Summary counts of included studies by subspecialty and representative procedures. *Studies involving multiple procedures counted separately within each relevant subspecialty discipline; [†]Includes TEP (totally extraperitoneal) and TAPP (transabdominal preperitoneal) approaches; [‡]Includes combinations of listed procedures if analysed collectively, and ileocaecal resection, Hartmann's procedure, intersphincteric dissection, abdominoperineal resection, and pelvic exenteration; [§]Includes robot-assisted anterior resection, intersphincteric resection, abdominoperineal resection, and rectal extirpations; [¶]Includes radical gastrectomy with D2 lymphadenectomy; [•]Includes exploratory laparoscopy for gastric malignancy.

mesenteric artery^[61], ureter^[54,62,88], and recurrent laryngeal nerve (RLN)^[40]; as well as providing important contextual features, such as distinguishing 'safe' from 'unsafe' zones of dissection^[53,78,109], confirming satisfactory exposure prior to mesh placement in hernia repair^[115], or automatically confirming the critical view of safety (CVS) in LC^[23,52,67,73,79-82,93]. Conversely, phase recognition models segmented operations into defined steps. Phase definitions were heterogeneous; for example, LC was segmented into between four and

Table 2. Task domains and performance metrics

Task domain	Description	Typical evaluation metrics	Example use cases	Number of studies (%) [*]
Phase recognition	Procedure-step classification	Accuracy, F1 score	- Segmentation of LC into defined operative phases^[34,42,46,51,79] - Step-level recognition in pancreaticojejunostomy during robotic pancreatoduodenectomy^[22] - Phase-mapping in laparoscopic sleeve gastrectomy^[45]	38 (34)
Anatomy identification	Landmark detection and segmentation	Dice coefficient, IoU	- Automatic verification of the CVS in LC^[52,67,73] - Ureter segmentation during left-sided colorectal resections^[54,62,88,98] - Vessel segmentation of the SMV/ICA/ICV in right hemicolectomy^[97]	56 (50)
Action-event recognition	Event or manoeuvre detection	Accuracy, F1 score	- Real-time bleeding detection in colorectal resections^[29,48] - Automated operative-difficulty grading in LC, e.g., using scarring/inflammation scoring^[21,90,105] - Perfusion-adequacy of colorectal anastomosis using ICG fluorescence^[25]	17 (15)
Instrument tracking	Tool presence and trajectory tracking	Accuracy, mAP, IoU	- Presence/absence detection of instruments in the operative field during LC^[26,31,76] - Continuous tool-trajectory tracking integrated with phase recognition in LC^[82] - Instrument segmentation for laparoscopic colorectal procedures^[60]	18 (16)
Skill-assessment	Automated performance scoring	Accuracy, correlation statistics	- Surgical-field-based skill scoring correlated with ESSQS performance in sigmoid colectomy^[50,59] - Objective dissection-efficiency metric in colorectal surgery^[85] - Composite competency scoring combining action detection, instrument metrics and phase recognition in LC^[127]	6 (5)

^{*}Studies evaluating multiple tasks counted separately under each relevant domain. LC: Laparoscopic cholecystectomy; mAP: mean average precision; IoU: intersection-over-union; CVS: critical view of safety; SMV: superior mesenteric vein; ICA: ileocolic artery; ICV: ileocolic vein; ICG: indocyanine-green; ESSQS: Endoscopic Surgical Skill Qualification System.

12 constituent phases, with few papers accommodating common real-world variances such as intraoperative cholangiogram or CBD exploration^[30,46,56]. Several authors denoted additional phases such as ‘out of body’, ‘idle’, ‘bleeding’ or ‘Po’ (redundant phase), to improve overall precision and accuracy^[39,42,43,108].

Action or event recognition tasks identified discrete intraoperative events, capturing key details indicative of increased procedural complexity or risk, such as scarring, inflammation^[67,77,90,123], specific surgical actions (dissecting, clipping, coagulating, and suctioning), dissection exposure quality, and operative-difficulty scoring^[21,26,29,36,48,112,127]. Instrument tracking tasks monitored instrument movement, presence, or field entry/exit^[27,60,76,110]. Skill-assessment was the least-represented task domain, inferring surgeon performance by combining tissue handling, dissection efficiency, and blood loss; composite measures included Global Operative Assessment of Laparoscopic Skills (GOALS), Endoscopic Surgical Skill Qualification System (ESSQS), and efficient-dissection time ratio (tissue-dissection time per monopolar device appearance time)^[51,59,85,94,122]. Annotation methods for training CV-AI models were poorly described, typically completed by expert surgeons from the host institution. Agreement between expert surgeons was suboptimal, even within studies^[119]. Reference standards for annotating training protocols were seldom reported; among those cited, the Society of American Gastrointestinal and Endoscopic Surgeons (SAGES) consensus statement was most commonly referenced^[46,134].

Table 3. Metric heterogeneity by task domain

Task domain	Unique metrics used	Metrics reported once within domain	Most frequently reported metrics (<i>n</i> studies)
Phase recognition	20	12	Accuracy (29); F1 (19); recall/sensitivity (14); precision (14); IoU (4)
Anatomy identification	29	16	Recall/sensitivity (25); Dice coefficient (25); IoU (22); precision (17); F1 (15)
Action-event recognition	20	13	Accuracy (10); F1 (5); recall/sensitivity (4); precision (3); Dice coefficient (2)
Instrument tracking	15	7	Accuracy (9); recall/sensitivity (6); precision (6); mAP (6); IoU (5)
Skill-assessment	7	2	Accuracy (4); recall/sensitivity (4); AUROC (2); correlation (2); specificity (2)

Total counts represent number of studies within each task domain reporting a given metric (unique metrics, reported once within domain). IoU: Intersection-over-union; F1: F1 score (harmonic mean of precision and recall); mAP: mean average precision; AUROC: area under receiver-operating-curve.

Table 4. Study stage and evaluation setting of included studies

Study stage/evaluation setting	Studies (<i>n</i> , %)*
Feasibility	
Retrospective	93 (82.3)
Retrospective with simulation component	1 (0.9)
Retrospective with real-time experimental verification	7 (6.2)
Prospective real-time	5 (4.4)
Validation	
Retrospective	7 (6.2)

*Study stage categories are descriptive, with each included study assigned to a single primary study stage for summary reporting.

Forty-one unique performance metrics were identified across the five task domains (Table 3; full distribution in Supplementary Table 1). Metric selection was highly heterogeneous, with predominant use of discrimination-style summary measures (e.g., accuracy, recall, F1 score) and comparatively sparse reporting of class imbalance, boundary-aware (e.g., Hausdorff distance), and real-time workflow factors (e.g., latency/stability, interface design, and surgeon feedback).

Deployment stage and validation

All studies were observational, with 93 (82%) retrospective feasibility analyses [Table 4]. Internal model validation was usually completed using open-access or publicly-listed datasets (such as Cholec80 in LC^[121] or LapSig300 in laparoscopic sigmoidectomy^[57,61]) in combination with institutional datasets. External or multicentric validation was reported in 12% of included studies, typically involving testing on non-local datasets rather than independent model evaluation. The largest comparative analysis assessed 12 algorithms on a single LC video dataset, developing the HeiChole benchmark for phase recognition and instrument detection models^[122]. Zang *et al.* reported a similar comparative analysis evaluating seven phase recognition models in robot-assisted inguinal hernia repair^[131]. Benchmarking efforts were not reported across the other task domains. Real-time components comprised both experimental theatre-based evaluations and live prospective analyses. Several authors advocated for independent evaluation panels to increase objectivity in model assessment^[24,39,86,93].

Real-time applications in theatre

Thirteen studies (12%)^[24,25,39,62,63,73,82,86,93,96,99,119,120] described real-time CV-AI implementation [Table 5]. Study designs were predominantly small, single-centre feasibility or pilot evaluations; including between one and 51 procedures^[99,119,120]. Participant groups comprised select cohorts, for instance those without prior history of

abdominal surgery^[86], or included elective-only procedures^[39,73,93]. Five studies blinded operating surgeons to CV-AI outputs, citing ethical considerations^[25,39,73,82,120]. Task domains assessed in live analyses included anatomical landmark detection^[24,39,62,63,82,86,96,99,119,120], perfusion assessment of colorectal anastomoses^[25], CVS validation^[73,82,93], and phase recognition^[39,82]. Two real-time studies^[39,82] evaluated multi-task CV-AI.

Reported endpoints broadly clustered into three groups: (i) technical feasibility/stability (e.g., malfunction rates, system uptime and intraoperative performance measures such as landmark/nerve detection accuracy), (ii) workflow/process measures (e.g., timing appropriateness of prompts, latency/frame-rate, and integration characteristics), and (iii) behavioural or early clinical signals, where a minority assessed surgeon-reported usability and/or short-term perioperative outcomes. Few real-time studies incorporated comparators or prespecified clinical endpoints, providing only preliminary signals of clinical efficacy; for example, Mascagni *et al.*^[82] reported ‘rate of malfunctions’ as their primary endpoint, while Fujinaga *et al.* described ‘appropriateness of landmark detection-timing’^[39]. In two separate studies, Ryu *et al.* reported that surgical intelligence systems improved pelvic nerve detection during left-sided colorectal resections, demonstrating statistically significant improvements in nerve detection rates independent of surgeon experience^[96,99]. In one study, short-term surgical outcomes such as blood loss, operative duration, length of postoperative stay, and postoperative complication rates were measured, without associated increase in complications^[99]. Leifman *et al.* reported increased rate of CVS attainment despite the algorithm outputs not being visible intraoperatively to surgeons, presumably due to a Hawthorne effect^[73].

Clinical integration and workflow readiness

Latency, display modality, and activation requirement reporting were inconsistent, constraining cross-study comparability and confidence in claims of intraoperative readiness. Latency was documented in 28% of included studies, with inference times ranging between sub-30ms to multi-second delays, illustrating wide variability in technical optimisation^[35,39,48,76]. In real-time analyses, CV-AI outputs were most commonly displayed as overlays on auxiliary monitors, frequently requiring manual activation rather than operating fully autonomously^[25,39,73,86,93]. Compatibility across several different endoscopic systems was demonstrated in only one study^[120]. Additionally, surgeon input was elicited in only a subset of studies involving real-time use [Table 5]^[24,39,62,82,86,120]. Studies predominantly collated end-user feedback using Likert scales; however, formal usability endpoints were uncommon with reporting poorly standardised^[24,39,62,86,120]. Median usefulness-scores were generally positive (e.g., 3.6/4, 4.2/5)^[39,120], alongside similarly high scores for perceived accuracy (e.g., 3.3-4.6/5 for postoperative pancreatic fistula in laparoscopic gastrectomy^[24], 4-4.2/5 for CBD detection during LC^[86]). User interfaces were typically considered under development, with several authors highlighting a need for optimised visual layout, simplified overlays, or feedback loops; with limited value in their current form^[82,86,119]. Nakanuma *et al.* reported improved surgeon intraoperative CBD detection with ‘tile-shaped’ compared to ‘bounding box’ display overlays, citing reduced screen flicker as an observed benefit^[86].

Alignment with implementation and reporting guidelines

Adherence to formal implementation and reporting frameworks was limited [Table 6]. Nine studies^[34,48,50,59,60,82,103,104,122] (8%) cited standardised reporting guidelines, typically referencing general-imaging or clinical reporting standards such as BIAS (Biomedical Image Analysis ChallengeS)^[135], STARD (Standards for Reporting Diagnostic Accuracy)^[136] (excluding the STARD-AI extension^[137]), STROBE (Strengthening the Reporting of Observational Studies in Epidemiology)^[138], or SQUIRE (Standards for Quality Improvement Reporting Excellence)^[139]. Reference to AI-specific frameworks was less frequent, described in 4% of studies: CLAIM (Checklist for Artificial Intelligence in Medical Imaging)^[140,141] or MI-CLAIM (Minimum Information about Clinical Artificial Intelligence Modeling)^[142] appeared in three^[48,103,104], DECIDE-AI (Developmental and Exploratory Clinical Investigations of DEcision support systems driven by Artificial Intelligence)^[143] in one^[82], with SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional

Table 5. Real-time intraoperative computer vision study characteristics

Authors (year)	Procedure (number of real-time cases)	Study design*	CV-AI function	Training dataset	Real-time readiness†	Display and surgeon feedback	Endpoints (group‡; primary; comparator)
Aoyama <i>et al.</i> ^[24] (2024)	Laparoscopic gastrectomy (10)	RTE	“HyperSeg”: identification of “dimpling lines” (landmarks to reduce POPF occurrence)	2,771 frames from 50 videos with 2,493 frames from 45 videos in training dataset	fps NR · latency 210 ms	AI-overlay on endoscopic camera image; overall EEC accuracy scores 3.3-4.6/5	B; surgeon perceived dimpling line accuracy (five-point Likert scale); NR
Arpaia <i>et al.</i> ^[25] (2022)	Laparoscopic colorectal resection (NR)	RTE	Maps ICG perfusion at anastomosis	470 frames from 11 videos	30 fps · latency NR	ROI screen outside theatre; no formal feedback	W; ensure theatre-compatible integration; NR; surgeons blinded
Fujinaga <i>et al.</i> ^[39] (2023)	Laparoscopic cholecystectomy (20)	PRT	Landmark identification (CBD, cystic duct, S4, Rouviere’s sulcus) and automatic phase cue	1,826 frames from 92 videos (landmark detection); 106 videos with 10,000 data per phase	30 fps · 33 ms (≥ 5 s phase display)	Sub-monitor; IEC/EEC usefulness 3.6-3.7/4	W; appropriateness of landmark-detection timing; EEC vs. AI-landmark detected phases; surgeons blinded
Kitaguchi <i>et al.</i> ^[62] (2023)	Laparoscopic colorectal resection (20)	RTE	Ureter and pelvic nerve detection	10,711 frames from 252 videos (ureter); 14,577 frames from 194 videos (nerve)	8-10 fps · latency NR	Display NR; surgeons scored 3.7-4.1/5 for ureter/nerve-injury prevention	T; model output speed (frame-rate) and accuracy; NR
Kojima <i>et al.</i> ^[63] (2023)	Laparoscopic colorectal resection (12)	RTE	Hypogastric nerve and plexus detection	12,978 frames from 245 videos (hypogastric nerve), 5,198 frames from 44 videos (superior hypogastric plexus)	12.5 fps · latency NR	Second video feed beside main monitor; no formal feedback	W; timing of target nerve detection; surgeon intraoperative assessment
Leifman <i>et al.</i> ^[73] (2024)	Laparoscopic cholecystectomy (40)	PRT	CVS verification	Frames NR; 700 videos with 80:20 training testing split	fps NR · latency 49 ms (peak 55 ms)	Touch-screen logger only (no live overlay); no formal feedback	T; CVS detection accuracy; surgeon intraoperative assessment; surgeons blinded
Mascagni <i>et al.</i> ^[82] (2024)	Laparoscopic cholecystectomy (3)	RTE	“SurgFlow”: phase and instrument recognition, anatomical segmentation, (gallbladder, cystic duct/artery, cystic plate, hepatocystic triangle), CVS validation	Frames NR; 80 videos (instrument tracking), 120 videos (phase), 201 videos (hepatocystic anatomy and CVS assessment)	fps NR · latency NR	Live overlay on laparoscopic feed; current UI offers limited direct value to surgeons, more for staff/administrators (unstructured interviews)	T; malfunction rate (any technical/non-technical problem); NR (single-arm feasibility); surgeons blinded
Nakanuma <i>et al.</i> ^[86] (2023)	Laparoscopic cholecystectomy (10)	PRT	Landmark identification (CBD, cystic duct, S4, Rouviere’s sulcus)	NR	fps NR · latency 90 ms	Sub-monitor; EEC 4/5 for CBD accuracy (bounding box), 4.2/5 for tile display	B; surgeon perceived landmark detection accuracy (five-point Likert scale); bounding box vs. tile shaped displays
Petracchi <i>et al.</i> ^[93] (2024)	Laparoscopic cholecystectomy (40)	PRT	Voice-activated CVS validation	131 frames augmented to 402; videos NR	fps NR · latency NR	Accessory monitor; no formal feedback	T; in-theatre feasibility of CVS detection; expert surgeon consensus

Ryu <i>et al.</i> ^[96] (2023)	Laparoscopic left-sided colorectal resections (10)	RTE	"Eureka": autonomic pelvic nerve identification	NR	fps NR · latency NR	AI navigation monitor; no formal feedback	B; rate of anatomic recognition assistance provided by AI navigation for each nerve; NR
Ryu <i>et al.</i> ^[99] (2025)	Laparoscopic left-sided colorectal resections (51)	PRT	"Eureka": autonomic pelvic nerve identification	NR	fps NR · latency NR	AI navigation monitor; no formal feedback	B; rate of anatomic recognition assistance provided by AI navigation for each nerve; NR
Tokuyasu <i>et al.</i> ^[119] (2021)	Laparoscopic cholecystectomy (1)	RTE	Landmark identification (CBD, cystic duct, S4, Rouviere's sulcus)	2,339 frames from 99 videos (augmented 26 times) with 76 videos in training dataset	37 fps · latency NR	Secondary monitor; no formal feedback	T; in-theatre feasibility of landmark detection; NR
Tomioaka <i>et al.</i> ^[120] (2023)	Laparoscopic hepatectomy (1)	RTE	Colour-coding of hepatic veins and Glissonean pedicle	350 frames from 10 videos	> 30 fps · latency 118 ms	Screen outside theatre; HPB surgeons 4.2/5 usefulness, trainees 1.8/2	T; in-theatre feasibility of landmark detection; NR; surgeons blinded

[†]PRT: Prospective real-time feasibility study; [‡]Frame-rate (fps) and/or latency; [§]Endpoint group (applicable only to real-time component for experimental pilots following retrospective training): T = technical feasibility/performance (e.g., capacity to run intraoperatively, fps/latency, malfunction/uptime, object detection performance), W = workflow/process (e.g., timing appropriateness, prompt timing, activation burden, display comparisons), B = behavioural/utility/human factors [e.g., clinical (e.g., surgeon-rated usefulness/accuracy, assistance rates, changes in intraoperative behaviour, early clinical/process outcomes)]. CV-AI: Computer vision-based artificial intelligence; RTE: real-time experimental pilot (live theatre test) following retrospective training; NR: not reported; POPF: postoperative pancreatic fistula; ICG: indocyanine green; ROI: region of interest; CBD: common bile duct; S4: lower edge of the left medial hepatic segment; IEC: internal evaluation committee; EEC: external evaluation committee; CVS: critical view of safety; HPB: hepatopancreatobiliary.

Trials-Artificial Intelligence)^[144], CONSORT-AI (Consolidated Standards of Reporting Trials-Artificial Intelligence)^[145], TRIPOD+AI (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis-Artificial Intelligence)^[146], and FUTURE-AI (Fairness Universality Traceability Usability Robustness Explainability - Artificial Intelligence)^[147] not appearing.

DISCUSSION

Surgical CV-AI is advancing technically but remains predominantly feasibility-stage, with most models confined to retrospective analyses rather than real-time use. Only 12% of studies evaluated real-time integration, comprising small, single-centre series with feasibility-oriented endpoints, variable blinding, and limited comparators; highlighting a major translational gap between offline validation and clinical deployment. This review confirmed substantial heterogeneity in metric selection and clinical integration of CV-AI. Across five principal task domains over 40 unique performance metrics were identified, though were variably applied, consistent with two recent reviews evaluating anatomical segmentation CV-AI in laparoscopic procedures, including those outside general surgery^[9,14]. Surgeon feedback was collected post-hoc after model development and validation, suggesting that end-user engagement typically follows, rather than precedes, system readiness. Additionally, agreement between expert annotators within studies was uneven, with adherence to recognised frameworks seldom reported. While anatomical delineation and phase recognition signal areas of emerging promise, studies remain yet to determine whether CV-AI meaningfully improves surgeon performance, intraoperative judgement, or operative outcomes. The study findings are synthesised into a concise evaluation framework in [Table 7](#), highlighting the concentration of current evidence around retrospective feasibility, and relative scarcity of reporting on real-time system performance (i.e., latency, usability, and prospective outcomes). This framework especially underscores that CV-AI translation hinges on workflow integration and human factors alongside conventional discrimination metrics.

Table 6. Reporting and implementation guidance in included studies

Authors (year)	Procedure/setting	Study stage	General guideline(s) cited	AI-specific guideline(s) cited	Items reported (translation-relevant)
Cheng <i>et al.</i> ^[34] (2022)	Laparoscopic cholecystectomy videos from four centres; surgical phase recognition and analysis	Retrospective feasibility	STROBE ^[138]	None stated	Ground-truth/annotation protocol; comparator (surgeon ground-truth)
Horita <i>et al.</i> ^[48] (2024)	Laparoscopic colectomy videos from a nationwide surgical video dataset (Japan); detection of active bleeding	Retrospective feasibility	None stated	CLAIM ^[140,141]	Ground-truth/annotation protocol; inference speed; usability/surgeon feedback
Igaki <i>et al.</i> ^[50] (2023)	Laparoscopic sigmoid colectomy videos submitted to the Japan Society for Endoscopy Surgery; automated surgical skill assessment	Retrospective feasibility	STARD ^[136]	None stated	Dataset selection; comparator (ESSQS score groups); ground-truth; AI confidence score
Kitaguchi <i>et al.</i> ^[59] (2021)	Laparoscopic sigmoid colectomy videos submitted to the Japan Society for Endoscopy Surgery; automated surgical skill assessment	Retrospective feasibility	SQUIRE ^[139] ; STARD	None stated	Dataset selection; comparator (ESSQS score groups); ground-truth
Kitaguchi <i>et al.</i> ^[60] (2022)	Laparoscopic colorectal surgery videos from a multi-institutional video dataset; instrument recognition	Retrospective feasibility	SQUIRE; STROBE	None stated	Dataset selection; ground-truth/annotation protocol
Mascagni <i>et al.</i> ^[82] (2024)	Operating-room deployment of “SurgFlow” for real-time AI assistance during three live laparoscopic cholecystectomies; phase and instrument recognition, anatomical segmentation	Prospective real-time feasibility (blinded)	None stated	DECIDE-AI ^[143]	Stability/uptime; stopping rules/safety safeguards; usability/surgeon feedback
Sengun <i>et al.</i> ^[103] (2023)	Laparoscopic transabdominal left adrenalectomy videos (single-centre); semantic segmentation of the left adrenal vein	Retrospective feasibility	STROBE	MI-CLAIM ^[142]	Dataset selection; ground-truth/annotation protocol
Sengun <i>et al.</i> ^[104] (2024)	Laparoscopic right adrenalectomy intraoperative videos (single-centre); semantic segmentation of the liver, inferior vena cava, and right adrenal gland	Retrospective feasibility	STROBE	MI-CLAIM	Dataset selection; ground-truth annotation protocol; comparator (SwinUNETR vs. MedNeXt segmentation models)
Wagner <i>et al.</i> ^[122] (2023)	HeiChole dataset incorporating 33 laparoscopic cholecystectomy videos from three centres; used in 2019 EndoVis challenge for phase, action, instrument recognition and/or skill assessment	Retrospective validation (external)	None stated	BIAS ^[135]	Dataset selection; ground-truth annotation protocol

Guidelines/frameworks explicitly cited by authors in included studies. AI-specific guidance includes reporting checklists (e.g., CLAIM^[140,141]/MI-CLAIM^[142], CONSORT-AI^[145], SPIRIT-AI^[144], TRIPOD+AI^[146]) and staged evaluation/implementation frameworks (e.g., DECIDE-AI^[143]), where reported. AI: Artificial intelligence; STROBE: Strengthening the Reporting of Observational Studies in Epidemiology; CLAIM: Checklist for Artificial Intelligence in Medical Imaging; STARD: Standards for Reporting Diagnostic Accuracy; ESSQS: Endoscopic Surgical Skill Qualification System; SQUIRE: Standards for QUality Improvement Reporting Excellence; DECIDE-AI: Developmental and Exploratory Clinical Investigations of DEcision support systems driven by Artificial Intelligence; MI-CLAIM: Minimum Information about Clinical Artificial Intelligence Modeling; BIAS: Biomedical Image Analysis ChallengeS; CONSORT-AI: Consolidated Standards of Reporting Trials-Artificial Intelligence; SPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials-Artificial Intelligence; TRIPOD+AI: Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis-Artificial Intelligence.

Metric selection

Robust, understandable, and task-specific metrics are fundamental to meaningful evaluation of CV-AI, a principle well-established in medical image segmentation research^[10,12,13,16,17]. Widely-used metrics offer familiarity and easy computation, although they are limited by a lack of clinical context; intraoperatively, should an error occur, surgeons are concerned more by their nature than error rate^[9,10,16]. For segmentation models, overlap-based metrics such as Dice coefficient are necessary, but do not penalise small-yet-critical boundary-errors^[10,13,16]. Anatomical delineation tasks guiding surgical decision-making, such as identifying the ureter or RLN, or recognising the base of segment four of the liver/Rouviere’s sulcus, inherently demand

Table 7. Evaluation framework for clinical translation of surgical computer vision

Evaluation domain	Minimum reporting set	Example outputs/measures
Technical validity (offline)	• Task definition and ground-truth protocol • Dataset representativeness and robust video quality • Generalisability (external validation/benchmarking) • Error patterns and failure modes	Accuracy/recall/F1; Dice coefficient/IoU; calibration; benchmark algorithm performance
Real-time system performance	• Latency and frame-rate on theatre hardware • Robustness to camera artefacts (e.g., smoke/blood/glare) • Failure behaviour and recovery (drop-outs/flicker)	Latency; frame-rate; uptime/malfunction rate; runtime stability; failure logs
Workflow integration	• Output display (direct overlay vs. auxiliary monitor) • Activation burden (manual vs. autonomous) • Stack compatibility and capture/processing pipeline • Data pathway and privacy controls	Integration manual; compatible in-theatre systems; activation steps; data flow description
Human factors and usability	• Output clarity and interpretability • Distraction/attention impact • Trust safeguards (dismissible outputs; model confidence display) • Surgeon feedback	User ratings; qualitative themes; surgeon task-load measures; override/dismiss rates
Clinical evaluation and outcomes	• Study stage (feasibility → prospective multicentre trials) • Safeguards (blinding/stopping rules where appropriate) • Outcomes beyond accuracy (process, patient outcomes) • Comparator (standard care/expert clinical judgement)	Procedure time; complications; landmark detection; predefined clinical endpoints
Reporting, governance and monitoring	• Reporting AI-specific guidance used • Traceability • Accountability • Post-deployment monitoring	DECIDE-AI ^[143] /SPIRIT-AI ^[144] /CONSORT-AI ^[145] alignment; audit plan

Evaluation framework derived from patterns observed in the present scoping synthesis for clinical translation of surgical computer vision to real-time use, presented as a pragmatic study minimum reporting set rather than a fixed prescriptive guideline. F1: F1 score (harmonic mean of precision and recall); IoU: intersection over union; AI: artificial intelligence; DECIDE-AI: Developmental and Exploratory Clinical Investigations of DEcision support systems driven by Artificial Intelligence; SPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials-Artificial Intelligence; CONSORT-AI: Consolidated Standards of Reporting Trials-Artificial Intelligence.

a high degree of precision^[9,10,16]. In these contexts, inclusion of boundary-aware parameters such as Hausdorff distance is essential^[9,10,12,13]. Comparatively, whilst plain accuracy is intuitive and provides a straightforward snapshot of algorithm performance, it performs poorly on class imbalanced datasets^[9,10,13]. Operative-difficulty classification models predominantly trained on patients without significant inflammation may achieve high accuracy rates by consistently predicting lower scores, although they may dangerously fail to recognise instances of severe pathology, including gangrenous or perforated disease^[9,10]. Such limitations may mask failure-modes and potentiate injury if not proactively recognised by surgeons, especially in complex procedures where enhanced situational awareness and safe decision-making are critical^[9,10]. Balanced accuracy, precision-recall curves, and F1 scores offer a more informative assessment in such contexts^[9,13]. Metrics should also evaluate all relevant aspects of model performance, including both discrimination metrics (e.g., AUROC) and calibration measures (e.g., Brier score, calibration plots) to confirm algorithm predictions are both accurate and clinically meaningful^[16,148]. This need for clinically relevant, standardised metric selection in surgical CV-AI is increasingly recognised, with guidance in this area in development by the Computer Vision in Surgical International Collaborative^[9,17]. Although thresholds that constitute ‘acceptable’ accuracy and precision rates are presently undefined, benchmarks above 90%

compared with expert surgeon consensus are likely required to establish surgeon trust and utility, comparable to that recommended for resect-and-discard strategies for diminutive polyps in colonoscopy^[26,149].

Clinical readiness

Achieving reliable real-time performance in the operating theatre introduces unique technical and workflow demands^[150]. Real-world environments introduce complexities often not captured in development datasets, such as anatomical variability, patient-specific factors such as high body mass index (BMI)/adhesions, and unpredictable conditions, including haemorrhage, inflammation, and camera artefacts (e.g., shake, lens blur or smoke)^[62,108,151,152]. Algorithms trained on retrospective datasets must demonstrate accuracy and reliability across diverse, dynamic surgical environments, accommodating variations in port placement, equipment (e.g., electrocautery devices, meshes), and non-sequential phase order (e.g., right hemicolectomy)^[122]. Specific situations also present unique technological challenges. Kojima *et al.* noted that their autonomic nerve CV-AI system struggled segmenting indistinct nerve boundaries, only worsened by poor pelvic visibility, underscoring the difficulty of replicating expert-level recognition in situations difficult in practice for surgeons themselves^[14,63]. Similarly, major adverse events such as bile duct injuries (BDIs) are rarely represented in idealised training datasets, despite critical importance for decision-support models. Several authors have proposed large multicentre open-access video repositories with equitable representation of diverse patient populations as a solution to support robust model training, although ultimately, intraoperative utility will also require consistent video quality, standardised annotations, and overall clinical relevance^[14,17,151,153,154]. Adherence to standard safe surgical practice protocols should be apparent in such datasets as some publicly-available surgical videos lack technical quality^[155]. Despite the current focus on algorithm performance, few studies emphasised usability-focused outcomes such as latency, overlay clarity, interpretability, and surgeon trust; factors essential to support widespread adoption and integration into surgical workflows^[17,152,154,156]. Outputs among real-time evaluations were commonly delivered via auxiliary displays and required manual activation, adding cognitive load and workflow burden^[25,39,73,86,93]. Additionally, reports of distracting artefacts such as screen flicker highlight that interface characteristics will shape surgeon trust and adoption^[86]. Simulation-based practice with CV-AI support and clear communication of confidence levels or rationale may help surgeons understand when to rely on or override algorithm outputs^[156]. Equally important is evaluation of human factors, such as mitigating over-reliance or distraction, and organisational considerations including cost-effectiveness, regulatory burden, and potential workflow-disruptions^[152-154,156].

Translation roadmap

Translating retrospective CV-AI research to routine surgical practice demands a clearly-defined roadmap for future clinical development and evaluation^[156]. Although surgical CV-AI reporting standards are still in development, existing AI-implementation guidelines should be applied to prevent incomplete/‘black-box’ reporting^[17]. Early-stage prospective studies should follow DECIDE-AI, though these recommendations lack specific guidance on CV-AI-metric reporting^[143]. Fundamental principles outlined in FUTURE-AI such as traceability, fairness, and explainability, should be incorporated into study design and model development^[147]. Real-time CV-AI should follow staged clinical innovation pathways, such as the Idea, Development, Exploration, Assessment, Long-term study (IDEAL) framework^[143,157,158], used previously during development cycles for surgical robotics and transanal total mesorectal excision (TaTME)^[159,160]. Under the IDEAL framework, current CV-AI literature aligns with the early-phase feasibility stages (1/2a)^[157,158], with progression to multicentre randomised trials (Stage 2b/3)^[161] for evaluation against surgical care standards, whilst adhering to reporting standards such as SPIRIT-AI^[144] and CONSORT-AI^[145]. Designing such trials poses unique practical challenges, such as blinding surgeon awareness of AI involvement^[73]. Beyond methodological rigour, ethical considerations must be addressed; for instance, legal

clarity is essential in situations where CV-AI contributes to errors in clinical judgement, or where AI-guidance is disregarded and patient harm eventuates^[154,156]. In parallel, standardised CV-AI benchmarking studies across all task domains are needed, akin to the HeiChole benchmark^[122]. Ultimately, translation to routine intraoperative use will not only require better algorithms, but an integrated ecosystem of clinical testing, surgeon training, regulatory oversight, and post-deployment monitoring^[17,156]. Execution will depend on a coordinated, multidisciplinary approach involving surgeons, data scientists, hospitals, and regulatory bodies^[17,154,156].

Study limitations

This review has several limitations. Surgical CV-AI research is evolving at substantial pace, and as such more recently published studies, tasks, metrics, and guidelines may not be captured. Only English-language, peer-reviewed publications were included. Title/abstract screening was performed by a single reviewer, introducing a residual risk of selection bias. Limiting the scope to minimally invasive CV-AI applications within general surgery may constrain generalisability beyond this discipline, such as to open procedures. Transanal surgical techniques, endoluminal, and radiological domains were excluded. Studies were excluded if they lacked intraoperative surgeon utility, although audit CV-AI platforms may play a role in optimising perioperative workflows and quality assurance processes. As few included studies evaluated real-time intraoperative deployment, inferences regarding workflow integration, usability, and clinical impact are constrained. No formal risk-of-bias appraisal or quantitative syntheses were performed, consistent with the scoping methodology and exploratory study objectives.

Future directions

Future research should prioritise evaluation designs that address CV-AI translational gaps, particularly external/multicentre validation, standardised reporting of real-time performance (latency, stability, and failure modes), and assessment of interface design and surgeon interaction. Although evidence for robust real-time performance and clinical impact within general surgery remains early, emerging multimodal vision-language approaches may offer an additional pathway to higher-level scene understanding, such as generating structured intraoperative summaries or context-aware explanations from surgical video analysis^[162,163].

CONCLUSION

In summary, CV-AI applications in general surgery remain at a nascent stage, characterised by considerable heterogeneity in metric selection and limited clinical integration. The present review highlights several crucial clinical translational gaps, including limited external validation, real-time intraoperative evaluation, and inconsistent usability reporting. Prospective, real-time CV-AI evaluation using standardised metric selection, annotation protocols, and transparent reporting aligned with AI-specific reporting frameworks is a necessary step moving forward. Strengthened governance, reporting standards, and collaborative end-user engagement are critical factors required to successfully translate conceptual promise to reliable real-time decision-support tools that support surgeon judgement and integrate seamlessly into routine operative workflows.

DECLARATIONS

Authors' contributions

Conceptualisation: Buchanan J, Eglinton T

Project administration: Buchanan J, Connor S, Eglinton T

Writing - original draft: Buchanan J

Writing - review and editing: Buchanan J, Connor S, Pearson J, Carey-Smith B, Eglinton T

Methodology: Buchanan J, Connor S, Pearson J, Eglinton T

Preparation of figures/images: Buchanan J

Manuscript revision: Buchanan J, Connor S, Pearson J, Carey-Smith B, Eglinton T

Supervision: Connor S, Pearson J, Carey-Smith B, Eglinton T

Availability of data and materials

Not applicable.

AI and AI-assisted tools statement

Not applicable.

Financial support and sponsorship

None.

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2026.

REFERENCES

1. Ward TM, Mascagni P, Ban Y, et al. Computer vision in surgery. *Surgery*. 2021;169:1253-6. [DOI PubMed](#)
2. Kourounis G, Elmahmudi AA, Thomson B, Hunter J, Ugail H, Wilson C. Computer image analysis with artificial intelligence: a practical introduction to convolutional neural networks for medical professionals. *Postgrad Med J*. 2023;99:1287-94. [DOI PubMed PMC](#)
3. Fernicola A, Palomba G, Capuano M, De Palma GD, Aprea G. Artificial intelligence applied to laparoscopic cholecystectomy: what is the next step? *Updates Surg*. 2024;76:1655-67. [DOI PubMed PMC](#)
4. Kankanamge D, Wijeweera C, Ong Z, et al. Artificial intelligence based assessment of minimally invasive surgical skills using standardised objective metrics - a narrative review. *Am J Surg*. 2025;241:116074. [DOI PubMed](#)
5. Garrow CR, Kowalewski KF, Li L, et al. Machine learning for surgical phase recognition: a systematic review. *Ann Surg*. 2021;273:684-93. [DOI PubMed](#)
6. Kehagias D, Lampropoulos C, Bellou A, Kehagias I. Detection of anatomic landmarks during laparoscopic cholecystectomy with the use of artificial intelligence - a systematic review of the literature. *Updates Surg*. 2026;78:229-40. [DOI PubMed](#)
7. Mascagni P, Alapatt D, Sestini L, et al. Computer vision in surgery: from potential to clinical value. *NPJ Digit Med*. 2022;5:163. [DOI PubMed PMC](#)
8. Varghese C, Harrison EM, O'Grady G, Topol EJ. Artificial intelligence in surgery. *Nat Med*. 2024;30:1257-68. [DOI PubMed](#)
9. Zhou R, Wang D, Zhang H, et al. Vision techniques for anatomical structures in laparoscopic surgery: a comprehensive review. *Front Surg*. 2025;12:1557153. [DOI PubMed PMC](#)
10. Reinke A, Tizabi MD, Baumgartner M, et al. Understanding metric-related pitfalls in image analysis validation. *Nat Methods*. 2024;21:182-94. [DOI PubMed PMC](#)
11. Miller C, Portlock T, Nyaga DM, O'Sullivan JM. A review of model evaluation metrics for machine learning in genetics and genomics. *Front Bioinform*. 2024;4:1457619. [DOI PubMed PMC](#)
12. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging*. 2015;15:29. [DOI PubMed PMC](#)
13. Müller D, Soto-Rey I, Kramer F. Towards a guideline for evaluation metrics in medical image segmentation. *BMC Res Notes*. 2022;15:210. [DOI PubMed PMC](#)
14. Khojah B, Enani G, Malibari A, Alhalabi W, Sabbagh A. Deep learning-based intraoperative guidance system for anatomical identification in laparoscopic surgery: a review. *J Infrastruct Policy Dev*. 2024;8:8939. [DOI](#)
15. Ibrahim AM, Dimick JB. What metrics accurately reflect surgical quality? *Annu Rev Med*. 2018;69:481-91. [DOI PubMed](#)
16. Maier-Hein L, Reinke A, Godau P, et al. Metrics reloaded: recommendations for image analysis validation. *Nat Methods*. 2024;21:195-212. [DOI PubMed PMC](#)

17. Kitaguchi D, Watanabe Y, Madani A, et al.; Computer Vision in Surgery International Collaborative. Artificial intelligence for computer vision in surgery: a call for developing reporting guidelines. *Ann Surg.* 2022;275:e609-11. [DOI PubMed](#)
18. Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z. *JBIManual for Evidence Synthesis*. JBI; 2024. Available from <https://synthesismanual.jbi.global> [accessed 25 February 2026].
19. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med.* 2018;169:467-73. [DOI PubMed](#)
20. Pollock D, Peters MDJ, Khalil H, et al. Recommendations for the extraction, analysis, and presentation of results in scoping reviews. *JBIM Evid Synth.* 2023;21:520-32. [DOI PubMed](#)
21. Abbing JR, Voskens FJ, Gerats BGA, Egging RM, Milletari F, Broeders IAMJ. Towards an AI-based assessment model of surgical difficulty during early phase laparoscopic cholecystectomy. *Comput Methods Biomech Biomed Eng Imaging Vis.* 2023;11:1299-306. [DOI](#)
22. Al Abbas AI, Namazi B, Radi I, et al. The development of a deep learning model for automated segmentation of the robotic pancreaticojejunostomy. *Surg Endosc.* 2024;38:2553-61. [DOI PubMed](#)
23. Alkhamaiseh KN, Grantner JL, Shebrain S, Abdel-Qader I. Towards reliable hepatocytic anatomy segmentation in laparoscopic cholecystectomy using U-Net with Auto-Encoder. *Surg Endosc.* 2023;37:7358-69. [DOI PubMed](#)
24. Aoyama Y, Matsunobu Y, Etoh T, et al. Artificial intelligence for surgical safety during laparoscopic gastrectomy for gastric cancer: indication of anatomical landmarks related to postoperative pancreatic fistula using deep learning. *Surg Endosc.* 2024;38:5601-12. [DOI PubMed](#)
25. Arpaia P, Bracale U, Corcione F, et al. Assessment of blood perfusion quality in laparoscopic colorectal surgery by means of Machine Learning. *Sci Rep.* 2022;12:14682. [DOI PubMed PMC](#)
26. Aspart F, Bolmgren JL, Lavanchy JL, et al. ClipAssistNet: bringing real-time safety feedback to operating rooms. *Int J Comput Assist Radiol Surg.* 2022;17:5-13. [DOI PubMed PMC](#)
27. Bamba Y, Ogawa S, Itabashi M, Kameoka S, Okamoto T, Yamamoto M. Automated recognition of objects and types of forceps in surgical images using deep learning. *Sci Rep.* 2021;11:22571. [DOI PubMed PMC](#)
28. Bamba Y, Ogawa S, Itabashi M, et al. Object and anatomical feature recognition in surgical video images based on a convolutional neural network. *Int J Comput Assist Radiol Surg.* 2021;16:2045-54. [DOI PubMed PMC](#)
29. Bamba Y, Ogawa S, Itabashi M, et al. Automated bleeding identification in surgical videos using deep learning. *Tokyo Women Med Univ J.* 2022;6:117-25. [DOI](#)
30. Ban Y, Rosman G, Eckhoff JA, et al. SUPR-GAN: SURgical PRediction GAN for event anticipation in laparoscopic and robotic surgery. *IEEE Robot Autom Lett.* 2022;7:5741-8. [DOI](#)
31. Beyersdorffer P, Kunert W, Jansen K, et al. Detection of adverse events leading to inadvertent injury during laparoscopic cholecystectomy using convolutional neural networks. *Biomed Tech.* 2021;66:413-21. [DOI PubMed](#)
32. Chen G, Xie Y, Yang B, et al. Artificial intelligence model for perigastric blood vessel recognition during laparoscopic radical gastrectomy with D2 lymphadenectomy in locally advanced gastric cancer. *BJS Open.* 2024;9. [DOI PubMed PMC](#)
33. Chen H, Gou L, Fang Z, et al. Artificial intelligence assisted real-time recognition of intra-abdominal metastasis during laparoscopic gastric cancer surgery. *NPJ Digit Med.* 2025;8:9. [DOI PubMed PMC](#)
34. Cheng K, You J, Wu S, et al. Artificial intelligence-based automated laparoscopic cholecystectomy surgical phase recognition and analysis. *Surg Endosc.* 2022;36:3160-8. [DOI PubMed](#)
35. den Boer RB, Jaspers TJM, de Jongh C, et al. Deep learning-based recognition of key anatomical structures during robot-assisted minimally invasive esophagectomy. *Surg Endosc.* 2023;37:5164-75. [DOI PubMed PMC](#)
36. Derathé A, Reche F, Moreau-Gaudry A, Jannin P, Gibaud B, Voros S. Predicting the quality of surgical exposure using spatial and procedural features from laparoscopic videos. *Int J Comput Assist Radiol Surg.* 2020;15:59-67. [DOI PubMed](#)
37. Endo Y, Tokuyasu T, Mori Y, et al. Impact of AI system on recognition for anatomical landmarks related to reducing bile duct injury during laparoscopic cholecystectomy. *Surg Endosc.* 2023;37:5752-9. [DOI PubMed PMC](#)
38. Fer D, Zhang B, Abukhalil R, et al. An artificial intelligence model that automatically labels roux-en-Y gastric bypasses, a comparison to trained surgeon annotators. *Surg Endosc.* 2023;37:5665-72. [DOI PubMed](#)
39. Fujinaga A, Endo Y, Etoh T, et al. Development of a cross-artificial intelligence system for identifying intraoperative anatomical landmarks and surgical phases during laparoscopic cholecystectomy. *Surg Endosc.* 2023;37:6118-28. [DOI PubMed](#)
40. Furube T, Takeuchi M, Kawakubo H, et al. Usefulness of an artificial intelligence model in recognizing recurrent laryngeal nerves during robot-assisted minimally invasive esophagectomy. *Ann Surg Oncol.* 2024;31:9344-51. [DOI PubMed](#)
41. Ghobadi V, Ismail LI, Wan Hasan WZ, et al. Real-time robust liver and gallbladder segmentation during laparoscopic cholecystectomy using convolutional neural networks: an analysis. *Art Int Surg.* 2024;4:278-87. [DOI](#)

42. Golany T, Aides A, Freedman D, et al. Artificial intelligence for phase recognition in complex laparoscopic cholecystectomy. *Surg Endosc.* 2022;36:9215-23. [DOI PubMed PMC](#)
43. Guédon ACP, Meij SEP, Osman KNMMH, et al. Deep learning for surgical phase recognition using endoscopic videos. *Surg Endosc.* 2021;35:6150-7. [DOI PubMed](#)
44. Han F, Zhong G, Zhi S, et al. Artificial intelligence recognition system of pelvic autonomic nerve during total mesorectal excision. *Dis Colon Rectum.* 2025;68:308-15. [DOI PubMed PMC](#)
45. Hashimoto DA, Rosman G, Witkowski ER, et al. Computer vision analysis of intraoperative video: automated recognition of operative steps in laparoscopic sleeve gastrectomy. *Ann Surg.* 2019;270:414-21. [DOI PubMed PMC](#)
46. Hegde SR, Namazi B, Iyengar N, et al. Automated segmentation of phases, steps, and tasks in laparoscopic cholecystectomy using deep learning. *Surg Endosc.* 2024;38:158-70. [DOI PubMed](#)
47. Honda R, Kitaguchi D, Ishikawa Y, et al. Deep learning-based surgical step recognition for laparoscopic right-sided colectomy. *Langenbecks Arch Surg.* 2024;409:309. [DOI PubMed](#)
48. Horita K, Hida K, Itatani Y, et al. Real-time detection of active bleeding in laparoscopic colectomy using artificial intelligence. *Surg Endosc.* 2024;38:3461-9. [DOI PubMed](#)
49. Igaki T, Kitaguchi D, Kojima S, et al. Artificial intelligence-based total mesorectal excision plane navigation in laparoscopic colorectal surgery. *Dis Colon Rectum.* 2022;65:e329-33. [DOI PubMed](#)
50. Igaki T, Kitaguchi D, Matsuzaki H, et al. Automatic surgical skill assessment system based on concordance of standardized surgical field development using artificial intelligence. *JAMA Surg.* 2023;158:e231131. [DOI PubMed PMC](#)
51. Jin Y, Long Y, Chen C, Zhao Z, Dou Q, Heng PA. Temporal memory relation network for workflow recognition from surgical video. *IEEE Trans Med Imaging.* 2021;40:1911-23. [DOI PubMed](#)
52. Kawamura M, Endo Y, Fujinaga A, et al. Development of an artificial intelligence system for real-time intraoperative assessment of the Critical View of Safety in laparoscopic cholecystectomy. *Surg Endosc.* 2023;37:8755-63. [DOI PubMed](#)
53. Khalid MU, Laplante S, Masino C, et al. Use of artificial intelligence for decision-support to avoid high-risk behaviors during laparoscopic cholecystectomy. *Surg Endosc.* 2023;37:9467-75. [DOI PubMed](#)
54. Khojah B, Enani G, Saleem A, et al. Deep learning-based intraoperative visual guidance model for ureter identification in laparoscopic sigmoidectomy. *Surg Endosc.* 2025;39:3610-23. [DOI PubMed](#)
55. Kinoshita K, Maruyama T, Kobayashi N, et al. An artificial intelligence-based nerve recognition model is useful as surgical support technology and as an educational tool in laparoscopic and robot-assisted rectal cancer surgery. *Surg Endosc.* 2024;38:5394-404. [DOI PubMed PMC](#)
56. Kirtac K, Aydin N, Lavanchy JL, et al. Surgical phase recognition: from public datasets to real-world data. *Appl Sci.* 2022;12:8746. [DOI](#)
57. Kitaguchi D, Takeshita N, Matsuzaki H, et al. Automated laparoscopic colorectal surgery workflow recognition using artificial intelligence: experimental research. *Int J Surg.* 2020;79:88-94. [DOI PubMed](#)
58. Kitaguchi D, Takeshita N, Matsuzaki H, et al. Real-time automatic surgical phase recognition in laparoscopic sigmoidectomy using the convolutional neural network-based deep learning approach. *Surg Endosc.* 2020;34:4924-31. [DOI PubMed](#)
59. 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:e2120786. [DOI PubMed PMC](#)
60. Kitaguchi D, Lee Y, Hayashi K, et al. Development and validation of a model for laparoscopic colorectal surgical instrument recognition using convolutional neural network-based instance segmentation and videos of laparoscopic procedures. *JAMA Netw Open.* 2022;5:e2226265. [DOI PubMed PMC](#)
61. Kitaguchi D, Takeshita N, Matsuzaki H, et al. Real-time vascular anatomical image navigation for laparoscopic surgery: experimental study. *Surg Endosc.* 2022;36:6105-12. [DOI PubMed](#)
62. Kitaguchi D, Harai Y, Kosugi N, et al. Artificial intelligence for the recognition of key anatomical structures in laparoscopic colorectal surgery. *Br J Surg.* 2023;110:1355-8. [DOI PubMed](#)
63. Kojima S, Kitaguchi D, Igaki T, et al. Deep-learning-based semantic segmentation of autonomic nerves from laparoscopic images of colorectal surgery: an experimental pilot study. *Int J Surg.* 2023;109:813-20. [DOI PubMed PMC](#)
64. Kolbinger FR, Rinner FM, Jenke AC, et al. Anatomy segmentation in laparoscopic surgery: comparison of machine learning and human expertise - an experimental study. *Int J Surg.* 2023;109:2962-74. [DOI PubMed PMC](#)
65. Kolbinger FR, Bodenstedt S, Carstens M, et al. Artificial Intelligence for context-aware surgical guidance in complex robot-assisted oncological procedures: an exploratory feasibility study. *Eur J Surg Oncol.* 2024;50:106996. [DOI PubMed](#)
66. Komatsu M, Kitaguchi D, Yura M, et al. Automatic surgical phase recognition-based skill assessment in laparoscopic distal gastrectomy using multicenter videos. *Gastric Cancer.* 2024;27:187-96. [DOI PubMed](#)

67. Korndorffer JR Jr, Hawn MT, Spain DA, et al. Situating artificial intelligence in surgery: a focus on disease severity. *Ann Surg.* 2020;272:523-8. DOI PubMed
68. Kumazu Y, Kobayashi N, Kitamura N, et al. Automated segmentation by deep learning of loose connective tissue fibers to define safe dissection planes in robot-assisted gastrectomy. *Sci Rep.* 2021;11:21198. DOI PubMed PMC
69. Kumazu Y, Kobayashi N, Senya S, et al. AI-based visualization of loose connective tissue as a dissectable layer in gastrointestinal surgery. *Sci Rep.* 2025;15:152. DOI PubMed PMC
70. Laplante S, Namazi B, Kiani P, et al. Validation of an artificial intelligence platform for the guidance of safe laparoscopic cholecystectomy. *Surg Endosc.* 2023;37:2260-8. DOI PubMed
71. Lavanchy JL, Zindel J, Kirtac K, et al. Automation of surgical skill assessment using a three-stage machine learning algorithm. *Sci Rep.* 2021;11:5197. DOI PubMed PMC
72. Lavanchy JL, Ramesh S, Dall'Alba D, et al. Challenges in multi-centric generalization: phase and step recognition in Roux-en-Y gastric bypass surgery. *Int J Comput Assist Radiol Surg.* 2024;19:2249-57. DOI PubMed PMC
73. Leifman G, Golany T, Rivlin E, Khoury W, Assalia A, Reissman P. Real-time artificial intelligence validation of critical view of safety in laparoscopic cholecystectomy. *Intell-Based Med.* 2024;10:100153. DOI
74. Liu K, Zhao Z, Shi P, Li F, Song H. Real-time surgical tool detection in computer-aided surgery based on enhanced feature-fusion convolutional neural network. *J Comput Des Eng.* 2022;9:1123-34. DOI
75. Liu R, An J, Wang Z, Guan J, Liu J, et al. Artificial intelligence in laparoscopic cholecystectomy: does computer vision outperform human vision? *Art Int Surg.* 2022;2:80-92. DOI
76. Liu Z, Zhou Y, Zheng L, Zhang G. SiNet: a hybrid deep CNN model for real-time detection and segmentation of surgical instruments. *Biomed Signal Process Control.* 2024;88:105670. DOI
77. Loukas C, Frountzas M, Schizas D. Patch-based classification of gallbladder wall vascularity from laparoscopic images using deep learning. *Int J Comput Assist Radiol Surg.* 2021;16:103-13. DOI PubMed
78. Madani A, Namazi B, Altieri MS, et al. Artificial intelligence for intraoperative guidance: using semantic segmentation to identify surgical anatomy during laparoscopic cholecystectomy. *Ann Surg.* 2022;276:363-9. DOI PubMed PMC
79. Mascagni P, Alapatt D, Urade T, et al. A computer vision platform to automatically locate critical events in surgical videos: documenting safety in laparoscopic cholecystectomy. *Ann Surg.* 2021;274:e93-5. DOI PubMed
80. Mascagni P, Alapatt D, Laracca GG, et al. Multicentric validation of EndoDigest: a computer vision platform for video documentation of the critical view of safety in laparoscopic cholecystectomy. *Surg Endosc.* 2022;36:8379-86. DOI PubMed
81. Mascagni P, Vardazaryan A, Alapatt D, et al. Artificial intelligence for surgical safety: automatic assessment of the critical view of safety in laparoscopic cholecystectomy using deep learning. *Ann Surg.* 2022;275:955-61. DOI PubMed
82. Mascagni P, Alapatt D, Lapergola A, et al. Early-stage clinical evaluation of real-time artificial intelligence assistance for laparoscopic cholecystectomy. *Br J Surg.* 2024;111:znad353. DOI PubMed
83. Mita K, Kobayashi N, Takahashi K, et al. Anatomical recognition of dissection layers, nerves, vas deferens, and microvessels using artificial intelligence during transabdominal preperitoneal inguinal hernia repair. *Hernia.* 2024;29:52. DOI PubMed PMC
84. Nakajima K, Kitaguchi D, Takenaka S, et al. Automated surgical skill assessment in colorectal surgery using a deep learning-based surgical phase recognition model. *Surg Endosc.* 2024;38:6347-55. DOI PubMed
85. Nakajima K, Takenaka S, Kitaguchi D, et al. Artificial intelligence assessment of tissue-dissection efficiency in laparoscopic colorectal surgery. *Langenbecks Arch Surg.* 2025;410:80. DOI PubMed PMC
86. Nakanuma H, Endo Y, Fujinaga A, et al. An intraoperative artificial intelligence system identifying anatomical landmarks for laparoscopic cholecystectomy: a prospective clinical feasibility trial (J-SUMMIT-C-01). *Surg Endosc.* 2023;37:1933-42. DOI PubMed
87. Nakamura T, Kobayashi N, Kumazu Y, et al. Precise highlighting of the pancreas by semantic segmentation during robot-assisted gastrectomy: visual assistance with artificial intelligence for surgeons. *Gastric Cancer.* 2024;27:869-75. DOI PubMed
88. Narihiro S, Kitaguchi D, Hasegawa H, Takeshita N, Ito M. Deep learning-based real-time ureter identification in laparoscopic colorectal surgery. *Dis Colon Rectum.* 2024;67:e1596-9. DOI PubMed
89. Oh N, Kim B, Kim T, Rhu J, Kim J, Choi GS. Real-time segmentation of biliary structure in pure laparoscopic donor hepatectomy. *Sci Rep.* 2024;14:22508. DOI PubMed PMC
90. Orimoto H, Hirashita T, Ikeda S, et al. Development of an artificial intelligence system to indicate intraoperative findings of scarring in laparoscopic cholecystectomy for cholecystitis. *Surg Endosc.* 2025;39:1379-87. DOI PubMed PMC
91. Ortenzi M, Rapoport Ferman J, Antolin A, et al. A novel high accuracy model for automatic surgical workflow recognition using artificial intelligence in laparoscopic totally extraperitoneal inguinal hernia repair (TEP). *Surg Endosc.* 2023;37:8818-28. DOI PubMed PMC
92. Owen D, Grammatikopoulou M, Luengo I, Stoyanov D. Automated identification of critical structures in laparoscopic cholecystectomy. *Int J Comput Assist Radiol Surg.* 2022;17:2173-81. DOI PubMed

93. Petracchi EJ, Olivieri SE, Varela J, et al. Use of artificial intelligence in the detection of the critical view of safety during laparoscopic cholecystectomy. *J Gastrointest Surg.* 2024;28:877-9. [DOI PubMed](#)
94. Pradeep CS, Sinha N. Surgical phase classification and operative skill assessment through spatial context aware CNNs and time-invariant feature extracting autoencoders. *Biocybern Biomed Eng.* 2023;43:700-24. [DOI](#)
95. Ramesh S, Dall'Alba D, Gonzalez C, et al. Multi-task temporal convolutional networks for joint recognition of surgical phases and steps in gastric bypass procedures. *Int J Comput Assist Radiol Surg.* 2021;16:1111-9. [DOI PubMed PMC](#)
96. Ryu S, Goto K, Kitagawa T, et al. Real-time artificial intelligence navigation-assisted anatomical recognition in laparoscopic colorectal surgery. *J Gastrointest Surg.* 2023;27:3080-2. [DOI PubMed PMC](#)
97. Ryu K, Kitaguchi D, Nakajima K, et al. Deep learning-based vessel automatic recognition for laparoscopic right hemicolectomy. *Surg Endosc.* 2024;38:171-8. [DOI PubMed](#)
98. Ryu S, Imaizumi Y, Goto K, et al. Feasibility of simultaneous artificial intelligence-assisted and nir fluorescence navigation for anatomical recognition in laparoscopic colorectal surgery. *J Fluoresc.* 2025;35:6755-61. [DOI PubMed](#)
99. Ryu S, Imaizumi Y, Goto K, et al. Artificial intelligence-enhanced navigation for nerve recognition and surgical education in laparoscopic colorectal surgery. *Surg Endosc.* 2025;39:1388-96. [DOI PubMed PMC](#)
100. Sasaki K, Ito M, Kobayashi S, et al. Automated surgical workflow identification by artificial intelligence in laparoscopic hepatectomy: experimental research. *Int J Surg.* 2022;105:106856. [DOI PubMed](#)
101. Sasaki S, Kitaguchi D, Takenaka S, et al. Machine learning-based automatic evaluation of tissue handling skills in laparoscopic colorectal surgery: a retrospective experimental study. *Ann Surg.* 2023;278:e250-5. [DOI PubMed](#)
102. Sato K, Fujita T, Matsuzaki H, et al. Real-time detection of the recurrent laryngeal nerve in thoracoscopic esophagectomy using artificial intelligence. *Surg Endosc.* 2022;36:5531-9. [DOI PubMed](#)
103. Sengun B, Iscan Y, Tataroglu Ozbulak GA, et al. Artificial intelligence in minimally invasive adrenalectomy: using deep learning to identify the left adrenal vein. *Surg Laparosc Endosc Percutan Tech.* 2023;33:327-31. [DOI PubMed](#)
104. Sengun B, Iscan Y, Yazici ZA, et al. Utilization of artificial intelligence in minimally invasive right adrenalectomy: recognition of anatomical landmarks with deep learning. *Acta Chir Belg.* 2024;124:492-8. [DOI PubMed](#)
105. Sharma S, Vannucci M, Pestana Legori L, et al. Early operative difficulty assessment in laparoscopic cholecystectomy via snapshot-centric video analysis. *Int J Comput Assist Radiol Surg.* 2025;20:1185-93. [DOI PubMed](#)
106. Shi P, Zhao Z, Liu K, Li F. Attention-based spatial-temporal neural network for accurate phase recognition in minimally invasive surgery: feasibility and efficiency verification. *J Comput Des Eng.* 2022;9:406-16. [DOI](#)
107. Shi J, Cui R, Wang Z, et al. Deep learning HRNet FCN for blood vessel identification in laparoscopic pancreatic surgery. *NPJ Digit Med.* 2025;8:235. [DOI PubMed PMC](#)
108. Shinozuka K, Turuda S, Fujinaga A, et al. Artificial intelligence software available for medical devices: surgical phase recognition in laparoscopic cholecystectomy. *Surg Endosc.* 2022;36:7444-52. [DOI PubMed PMC](#)
109. Smithmaitrie P, Khaonualsri M, Sae-Lim W, Wangkulangkul P, Jearanai S, Cheewatanakornkul S. Development of deep learning framework for anatomical landmark detection and guided dissection line during laparoscopic cholecystectomy. *Heliyon.* 2024;10:e25210. [DOI PubMed PMC](#)
110. Strong JS, Furube T, Takeuchi M, et al. Evaluating surgical expertise with AI-based automated instrument recognition for robotic distal gastrectomy. *Ann Gastroenterol Surg.* 2024;8:611-9. [DOI PubMed PMC](#)
111. Strong JS, Yura M, Takeuchi M, et al. External validation of an automated surgical step recognition model for robotic distal gastrectomy (RDG) using a multicenter dataset. *Ann Gastroenterol Surg.* 2025;9:1155-62. [DOI PubMed PMC](#)
112. Sunakawa T, Kitaguchi D, Kobayashi S, et al. Deep learning-based automatic bleeding recognition during liver resection in laparoscopic hepatectomy. *Surg Endosc.* 2024;38:7656-62. [DOI PubMed](#)
113. Takeuchi M, Collins T, Ndagijimana A, et al. Automatic surgical phase recognition in laparoscopic inguinal hernia repair with artificial intelligence. *Hernia.* 2022;26:1669-78. [DOI PubMed](#)
114. Takeuchi M, Kawakubo H, Saito K, et al. Automated surgical-phase recognition for robot-assisted minimally invasive esophagectomy using artificial intelligence. *Ann Surg Oncol.* 2022;29:6847-55. [DOI PubMed](#)
115. Takeuchi M, Collins T, Lipps C, et al. Towards automatic verification of the critical view of the myoeptineal orifice with artificial intelligence. *Surg Endosc.* 2023;37:4525-34. [DOI PubMed](#)
116. Takeuchi M, Kawakubo H, Tsuji T, et al. Evaluation of surgical complexity by automated surgical process recognition in robotic distal gastrectomy using artificial intelligence. *Surg Endosc.* 2023;37:4517-24. [DOI PubMed PMC](#)
117. Tao H, Li B, Zeng X, et al. Artificial intelligence-based anatomy recognition in laparoscopic hepatectomy: multicentre study. *Br J Surg.* 2024;111:znae254. [DOI PubMed](#)
118. Tashiro Y, Aoki T, Kobayashi N, et al. Novel navigation for laparoscopic cholecystectomy fusing artificial intelligence and indocyanine green fluorescent imaging. *J Hepatobiliary Pancreat Sci.* 2024;31:305-7. [DOI PubMed](#)

119. Tokuyasu T, Iwashita Y, Matsunobu Y, et al. Development of an artificial intelligence system using deep learning to indicate anatomical landmarks during laparoscopic cholecystectomy. *Surg Endosc*. 2021;35:1651-8. [DOI PubMed PMC](#)
120. Tomioka K, Aoki T, Kobayashi N, et al. Development of a novel artificial intelligence system for laparoscopic hepatectomy. *Anticancer Res*. 2023;43:5235-43. [DOI PubMed](#)
121. Twinanda AP, Shehata S, Mutter D, Marescaux J, de Mathelin M, Padoy N. EndoNet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE Trans Med Imaging*. 2017;36:86-97. [DOI PubMed](#)
122. Wagner M, Müller-Stich BP, Kisilenko A, et al. Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the HeiChole benchmark. *Med Image Anal*. 2023;86:102770. [DOI PubMed](#)
123. Ward TM, Hashimoto DA, Ban Y, Rosman G, Meireles OR. Artificial intelligence prediction of cholecystectomy operative course from automated identification of gallbladder inflammation. *Surg Endosc*. 2022;36:6832-40. [DOI PubMed](#)
124. Wu S, Chen Z, Liu R, et al.; Youth Committee of Pancreatic Disease of Sichuan Doctor Association (YCPD). SurgSmart: an artificial intelligent system for quality control in laparoscopic cholecystectomy: an observational study. *Int J Surg*. 2023;109:1105-14. [DOI PubMed PMC](#)
125. Yamazaki Y, Kanaji S, Matsuda T, et al. Automated surgical instrument detection from laparoscopic gastrectomy video images using an open source convolutional neural network platform. *J Am Coll Surg*. 2020;230:725-732e1. [DOI PubMed](#)
126. Yang HY, Hong SS, Yoon J, et al. Deep learning-based surgical phase recognition in laparoscopic cholecystectomy. *Ann Hepatobiliary Pancreat Surg*. 2024;28:466-73. [DOI PubMed PMC](#)
127. Yen HH, Hsiao YH, Yang MH, et al. Automated surgical action recognition and competency assessment in laparoscopic cholecystectomy: a proof-of-concept study. *Surg Endosc*. 2025;39:3006-16. [DOI PubMed](#)
128. Yin SM, Lien JJ, Chiu IM. Deep learning implementation for extrahepatic bile duct detection during indocyanine green fluorescence-guided laparoscopic cholecystectomy: pilot study. *BJS Open*. 2025;9. [DOI PubMed PMC](#)
129. Yoshida M, Kitaguchi D, Takeshita N, et al. Surgical step recognition in laparoscopic distal gastrectomy using artificial intelligence: a proof-of-concept study. *Langenbecks Arch Surg*. 2024;409:213. [DOI PubMed](#)
130. You J, Cai H, Wang Y, et al. Artificial intelligence automated surgical phases recognition in intraoperative videos of laparoscopic pancreatoduodenectomy. *Surg Endosc*. 2024;38:4894-905. [DOI PubMed](#)
131. Zang C, Turkcan MK, Narasimhan S, et al. Surgical phase recognition in inguinal hernia repair-ai-based confirmatory baseline and exploration of competitive models. *Bioengineering*. 2023;10. [DOI PubMed PMC](#)
132. Zhai Y, Chen Z, Zheng Z, et al. Artificial intelligence for automatic surgical phase recognition of laparoscopic gastrectomy in gastric cancer. *Int J Comput Assist Radiol Surg*. 2024;19:345-53. [DOI PubMed](#)
133. Zygomalas A, Kalles D, Katsiakakis N, Anastasopoulos A, Skroubis G. Artificial intelligence assisted recognition of anatomical landmarks and laparoscopic instruments in transabdominal preperitoneal inguinal hernia repair. *Surg Innov*. 2024;31:178-84. [DOI PubMed](#)
134. Meireles OR, Rosman G, Altieri MS, et al.; SAGES Video Annotation for AI Working Groups. SAGES consensus recommendations on an annotation framework for surgical video. *Surg Endosc*. 2021;35:4918-29. [DOI PubMed](#)
135. Maier-Hein L, Reinke A, Kozubek M, et al. BIAS: transparent reporting of biomedical image analysis challenges. *Med Image Anal*. 2020;66:101796. [DOI PubMed PMC](#)
136. Bossuyt PM, Reitsma JB, Bruns DE, et al.; STARD Group. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ*. 2015;351:h5527. [DOI PubMed PMC](#)
137. Sounderajah V, Ashrafian H, Golub RM, et al.; STARD-AI Steering Committee. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open*. 2021;11:e047709. [DOI PubMed PMC](#)
138. Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ*. 2007;335:806-8. [DOI PubMed PMC](#)
139. Ogrinc G, Davies L, Goodman D, Batalden P, Davidoff F, Stevens D. SQUIRE 2.0 (Standards for Quality Improvement Reporting Excellence): revised publication guidelines from a detailed consensus process. *BMJ Qual Saf*. 2016;25:986-92. [DOI PubMed PMC](#)
140. Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell*. 2020;2:e200029. [DOI PubMed PMC](#)
141. Tejani AS, Klontzas ME, Gatti AA, et al.; CLAIM 2024 Update Panel. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. *Radiol Artif Intell*. 2024;6:e240300. [DOI PubMed PMC](#)
142. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med*. 2020;26:1320-4. [DOI PubMed PMC](#)
143. Vasey B, Nagendran M, Campbell B, et al.; DECIDE-AI expert group. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ*. 2022;377:e070904. [DOI PubMed PMC](#)

144. Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ; SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Lancet Digit Health.* 2020;2:e549-60. DOI PubMed PMC
145. Liu X, Rivera SC, 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. *BMJ.* 2020;370:m3164. DOI PubMed PMC
146. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ.* 2024;385:e078378. DOI PubMed PMC
147. Lekadir K, Frangi AF, Porras AR, et al.; FUTURE-AI Consortium. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ.* 2025;388:e081554. DOI PubMed PMC
148. Park SH, Han K, Jang HY, et al. Methods for clinical evaluation of artificial intelligence algorithms for medical diagnosis. *Radiology.* 2023;306:20-31. DOI PubMed
149. Rex DK, Kahi C, O'Brien M, et al. The American Society for Gastrointestinal Endoscopy PIVI (Preservation and Incorporation of Valuable Endoscopic Innovations) on real-time endoscopic assessment of the histology of diminutive colorectal polyps. *Gastrointest Endosc.* 2011;73:419-22. DOI PubMed
150. Mascagni P, Alapatt D, Sestini L, et al. Applications of artificial intelligence in surgery: clinical, technical, and governance considerations. *Cir Esp.* 2024;102 Suppl 1:S66-71. DOI
151. Kitaguchi D, Takeshita N, Hasegawa H, Ito M. Artificial intelligence-based computer vision in surgery: recent advances and future perspectives. *Ann Gastroenterol Surg.* 2022;6:29-36. DOI PubMed PMC
152. Mascagni P, Padoy N. OR black box and surgical control tower: recording and streaming data and analytics to improve surgical care. *J Visc Surg.* 2021;158:S18-25. DOI PubMed
153. McGivern KG, Drake TM, Knight SR, et al. Applying artificial intelligence to big data in hepatopancreatic and biliary surgery: a scoping review. *Art Int Surg.* 2023;3:27-47. DOI
154. Adegbesan A, Akingbola A, Aremu O, Adewole O, Amamdikwa JC, Shagaya U. From scalpels to algorithms: the risk of dependence on artificial intelligence in surgery. *J Med Surg Public Health.* 2024;3:100140. DOI
155. Hwang N, Chao PP, Kirkpatrick J, Srinivasa K, Koea JB, Srinivasa S. Educational quality of Robotic Whipple videos on YouTube. *HPB.* 2024;26:826-32. DOI PubMed
156. Byrd IV TF, Tignanelli CJ. Artificial intelligence in surgery - a narrative review. *J Med Artif Intell.* 2024;7:29. DOI
157. Ergina PL, Barkun JS, McCulloch P, Cook JA, Altman DG; IDEAL Group. IDEAL framework for surgical innovation 2: observational studies in the exploration and assessment stages. *BMJ.* 2013;346:f3011. DOI PubMed PMC
158. McCulloch P, Cook JA, Altman DG, Heneghan C, Diener MK; IDEAL Group. IDEAL framework for surgical innovation 1: the idea and development stages. *BMJ.* 2013;346:f3012. DOI PubMed PMC
159. Marcus HJ, Ramirez PT, Khan DZ, et al.; IDEAL Robotics Colloquium. The IDEAL framework for surgical robotics: development, comparative evaluation and long-term monitoring. *Nat Med.* 2024;30:61-75. DOI PubMed
160. Roodbeen SX, Lo Conte A, Hirst A, et al. Evolution of transanal total mesorectal excision according to the IDEAL framework. *BMJ Surg Interv Health Technol.* 2019;1:e000004. DOI PubMed PMC
161. Cook JA, McCulloch P, Blazeby JM, Beard DJ, Marinac-Dabic D, Sedrakyan A; IDEAL Group. IDEAL framework for surgical innovation 3: randomised controlled trials in the assessment stage and evaluations in the long term study stage. *BMJ.* 2013;346:f2820. DOI PubMed PMC
162. Li X, Li L, Jiang Y, et al. Vision-Language Models in medical image analysis: from simple fusion to general large models. *Inform Fusion.* 2025;118:102995. DOI
163. Li X, Sun Y, Lin J, Li L, Feng T, Yin S. The synergy of seeing and saying: revolutionary advances in multi-modality medical vision-language large models. *Artif Intell Sci Eng.* 2025;1:79-97. DOI

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



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