

## Automated skills assessment in open surgery: A scoping review

Hawa Hamza, Dehlela Shabir, Omar Aboumarzouk, Abdulla Al-Ansari, Khaled Shaban, Nikhil V. Navkar

### Item type

Journal Contribution

### Terms of use

This work is licensed under a [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/) license

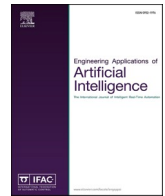
### This version is available at

[https://manara.qnl.qa/articles/journal\\_contribution/Automated\\_skills\\_assessment\\_in\\_open\\_surgery\\_A\\_scoping\\_review/28829579/](https://manara.qnl.qa/articles/journal_contribution/Automated_skills_assessment_in_open_surgery_A_scoping_review/28829579/)

Access the item on Manara for more information about usage details and recommended citation.

Posted on Manara – Qatar Research Repository on

2025-04-18



# Automated skills assessment in open surgery: A scoping review

Hawa Hamza <sup>a,1</sup>, Dehlela Shabir <sup>a,1</sup>, Omar Aboumarzouk <sup>a</sup>, Abdulla Al-Ansari <sup>a</sup>, Khaled Shaban <sup>b</sup>, Nikhil V. Navkar <sup>a,\*</sup>

<sup>a</sup> Department of Surgery, Hamad Medical Corporation, Doha, Qatar

<sup>b</sup> Computer Science and Engineering Department, Qatar University, Doha, Qatar

## ARTICLE INFO

### Keywords:

Automation  
Skills assessment  
Open surgery  
Machine learning  
Surgical education

## ABSTRACT

Surgical skills proficiency lowers the incidence of adverse clinical outcomes during surgeries. Artificial intelligence (AI) has been applied for surgical skills assessment, especially in the field of minimally invasive surgeries (MIS). This paves the way for integrating AI for skills assessment in open surgeries as well. An overview of its applications can inform the scientific community and facilitate further developments. In this scoping review, we present the open surgeries and clinical settings where AI-based skill assessment has been applied, the kind of surgical data acquired for the AI-based algorithms, and the types of AI-based models used for automated skills assessment. A total of 40 articles were identified and included. Majority of the articles focused on macrosurgical suturing (45 %,  $n = 18$ ). Most of the studies acquired data by capturing surgeon's hands (50 %,  $n = 20$ ). About 35 % utilized deep learning algorithms, specifically convolutional neural networks (CNN) ( $n = 14$ ). The assessment input for the automation algorithms were predominantly hand movement. Around 37.5 % ( $n = 15$ ) of the studies assessed algorithm performance using classification accuracy. In the review, we compare conventional methods such as statistical modeling and custom algorithms with the emerging AI-based approaches. We also explore the utilization of object detection and temporal information for surgical skills assessment. We highlight the progress in automated skills assessment during open surgery with advancements in sensor technology, and AI algorithms with high prediction accuracies. Further developments in data acquisition and processing methods are essential to facilitate clinical implementation of such technologies.

## 1. Introduction

For a surgeon to be deemed competent, they must be proficient in three main domains, namely, knowledge, skills, and attitudes (Pakkasjärvi et al., 2024). A core aspect of surgical training focuses on acquisition of technical skills that involve effective manipulation of surgical instruments requiring fine psychomotor ability (Bell, 2009). Technical skills largely consist of knot tying, suturing, and procedure specific techniques. Suturing is an essential surgical skill which is the basis for moving towards other advanced surgical procedures (Singh et al., 2024). Other operative skills such as tactile and visual-spatial awareness are also important. Such skills are acquired through experience by repeated practice (Bell, 2009). Inadequate training in surgical skills can significantly compromise the quality and safety of patient care. For example, insufficient suturing skills can result in tissue damage and bleeding (Singh et al., 2024). Overall surgical skill levels have a

profound impact on clinical outcomes, with lower skill levels associated with increased complications and mortality rates (Birkmeyer et al., 2013). Therefore, ensuring proficiency in surgical skills through assessments conducted by experienced on-site surgeons, based on established guidelines, is crucial (Glossop et al., 2023; Jaffer et al., 2009). Learning by doing, which is a fundamental part of surgical training, occurs in simulation settings (such as virtual trainers, synthetic phantoms, animals, or cadavers) and in the operating room (Pakkasjärvi et al., 2024). However, this process is challenged by the limited availability of experts to assess large cohorts of trainees (Shabir et al., 2021, 2022a; Shayan et al., 2023). Addressing this shortage is critical, as patient outcomes, including readmissions, are influenced by the technical skill level of the operating surgeon (Birkmeyer et al., 2013).

A potential solution for enhancing surgical skill training lies with the integration of objective skills assessment through automated systems using data capturing devices and artificial intelligence (AI) based

\* Corresponding author. Department of Surgery, Surgical Research Section, Hamad General Hospital, Hamad Medical Corporation, PO Box 3050, Doha, Qatar.  
E-mail address: [nnavkar@hamad.qa](mailto:nnavkar@hamad.qa) (N.V. Navkar).

<sup>1</sup> The authors contributed equally to the manuscript.

algorithms for surgical skills assessment (Melton, 2010). Automated systems, utilizing methods such as machine learning (ML), deep learning (DL), or statistical modeling (Ahmadi et al., 2015), have the potential to deliver accurate assessments of surgical skills comparable to evaluations conducted by expert surgeons (Titov et al., 2023). This has been extensively explored in minimally invasive and robot-assisted surgeries, where data capture is relatively straightforward due to the use of camera scopes and robotic systems (Abdurahiman et al., 2022; Ismail Fawaz et al., 2019; Khorasani et al., 2023). However, the application of automated skill assessment in open surgery remains limited, despite it representing a significant portion of procedures performed worldwide (Deng et al., 2021; Shaharan et al., 2017).

Studies have reported various data acquisition and analysis techniques aimed at providing feedback during open surgery. In these studies, different features are extracted from the operative field to identify traits of skilled and unskilled surgical maneuvers (Lavanchy et al., 2021). One of the pioneering applications in open surgery involved the use of electromagnetic (EM) sensors to track the surgeon's hand motions (Datta et al., 2001). Subsequent studies have focused on motion analysis of surgical instruments and operative tissue to determine surgical dexterity (Hamza et al., 2023; Jardine et al., 2015; Sun et al., 2016). In this context, the overall surgical performance of a participant can be assessed and summarized with labels such as novice/expert, pass/fail, or through a scoring system. Advancements in research, development, and application of automated skills assessment during open surgery require a comprehensive understanding of existing technologies employed.

## 2. Literature review

### 2.1. Artificial intelligence for surgical education

Existing literature reviews explore general applications of AI in surgical education (Bilgic et al., 2022; Kirubarajan et al., 2022), including AI's potential in curriculum development and instructional material enhancement. While it is useful for educators and researchers interested in applying AI for general surgical knowledge acquisition, such reviews do not cover assessment of technical skills in real-time. An extensive overview of ML models used in surgical phase detection, which is the identification of high level activities in the procedure, during minimally invasive surgery has been published (Garrow et al., 2021). The survey does not cover skills assessment, and additionally, implementations in minimally invasive surgery may not necessarily apply to open surgery. A thorough examination of AI applications in orthopedic surgery has also been presented (Geda et al., 2024). Nevertheless, assessment of surgeon's skills was not the focus of the review.

### 2.2. Automated skills assessment during surgeries

Specific to surgical skills assessment (Kawka et al., 2022), reviewed ML models for intraoperative video analysis (such as recognition of instrument, gesture, or anatomy) in minimally invasive surgery, while another review focused on metrics for automated skills assessments specifically for robot-assisted laparoscopic surgeries (Guerin et al., 2022). The methods for skills assessment heavily rely on laparoscopic videos and kinematic data from robotic systems. These types of data are not applicable to open surgeries, and therefore, the findings of such reviews cannot be generalized. A review on the use of objective computer-aided technical skill evaluation (Vedula et al., 2017) did not specifically address applications for open surgery. Systematic review published on ML models (Lam et al., 2022) mainly focused on endoscopic, laparoscopic and robot-assisted surgeries. Other reviews of automated methods (Levin et al., 2019) and deep neural networks (Yanik et al., 2022) for surgical skills assessment did not explore solutions to difficulties in data acquisition during open surgeries. Similarly (Dick et al., 2024), focused on automated analysis of surgical videos, and

did not consider other data capture methods such as inertial measurement units (IMU) or EM sensors. It is essential to investigate various data acquisition solutions since skills assessment during open surgery is largely based on surgeon's hand motions, which is hard to capture in a conventional operating room setting. Another systematic review presented Internet of Things (IoT) systems for surgical skills assessment (Castillo-Segura et al., 2021), however, it did not examine the different open surgeries and clinical settings they have been applied to. On the other hand (Titov et al., 2023), reviewed ML models used in virtual reality training, microsurgery, and endoscopies. However, this review was limited to neurosurgical procedures.

To the best of our knowledge, no comprehensive reviews on automated skills assessment have been conducted that specifically focus on open surgeries. To address this gap, we conducted a systematic search of scientific literature. Application of automated skills assessment during open surgery is challenging due to the nature of such procedures, whereby data acquisition becomes harder as compared to minimally invasive surgery. We believe that a comprehensive overview of different automation algorithms utilized for skills assessment during open surgery is much needed to gain a clear insight into approaches that have already been applied. Such a synthesis has the potential for driving advancements while ensuring that future research and developments stay relevant and applicable.

## 3. Methods

This review adheres to the guidelines established by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) (Tricco et al., 2018). While we did not publicly register the protocol previously, we have ensured the guidelines for scoping review methods were thoroughly followed to promote reproducibility.

### 3.1. Research questions

This scoping review aims to summarize existing automated skills assessment technologies used in open surgery. The specific research questions (RQ) and the rationale for each are provided in Table 1.

**Table 1**

Research questions we aim to address through this review and the rationale for each.

#	Research Question (RQ)	Rationale
RQ1	What are the different open surgeries and clinical settings where automated skills assessment has been applied?	Identifying the surgical specialties, procedures, steps, and the clinical settings (such as synthetic phantom, animal, or human) where automated skills assessment have been applied can provide an understanding of the automation systems that are generalizable and thereby inform best practice.
RQ2	What data is captured during open surgery for automated skills assessment?	Examining the most frequently captured type of data as well as the various devices used for data acquisition during open surgery will provide insights into future improvements in data capturing methods for automation skills assessment systems.
RQ3	What are the algorithms used for automated skills assessment during open surgery?	Highlighting current automation methods used, types of assessment inputs from the open surgery data captured, the outputs provided (in terms of skills classification or scoring), and performance of the algorithms can inform future work towards clinical implementations of automated skills assessment.

Distinct from earlier reviews, we present the types of open surgeries utilizing automated skills assessment, the various sensors used for data acquisition, the automation algorithms employed to process the data and predict surgical skill levels, the evaluation criteria used to measure accuracy, and a comparative analysis of these algorithms. In addition, we further explore the challenges in open surgery data acquisition, and provide deeper insights into the common algorithms used, assessment inputs, and performance evaluation.

3.2. Eligibility criteria

To be considered in this review, articles must satisfy the inclusion criteria which were: (i) article reporting original research, (ii) written in English (iii) discussing technologies for automated assessment of surgical technical skills (iv) during open surgery (including microsurgery). Although data capture is more straightforward in microsurgery due to the use of microscopic cameras, we included microsurgical procedures in our review as instrument manipulation occurs through direct hand and wrist motions (Fattahi Sani et al., 2021). We also considered non-AI based papers reporting automated systems using statistical modelling, computer vision, and other custom algorithms, with which the AI algorithms were compared.

The exclusion criteria consisted of (i) articles not on surgery, (ii) articles focusing on minimally invasive surgery using laparoscopes, endoscopes, or robot-assistance, where instrument manipulation is not direct, (iii) articles on skills assessment which is not automated, (iv) articles related to virtual reality training where kinematic data capture is effortlessly done (Titov et al., 2023), (v) articles on methods for data acquisition and metric extraction without incorporating automation algorithms for skill level prediction, (vi) assessment of only non-technical skills (such as communication, teamwork, and situational awareness), and (vii) review articles.

3.3. Search

A comprehensive literature search was conducted across the PubMed, Scopus, and IEEE Xplore databases without any restrictions on publication year. The most recent searches were conducted on September 01, 2024. The search terms included but were not limited to “open surgery”, “microsurgery”, “automation”, “artificial intelligence”, “machine learning”, “deep learning”, “convolutional neural networks”, “skills assessment”, and “proficiency”. Search limits were applied to exclude articles that mentioned laparoscopes, endoscopes, or robotic

**Table 2**  
Boolean search strings used for different databases.

Database	Boolean Search Strings
PubMed	((("Surgical Procedures, Operative"[Mesh] or "surgery" [Subheading] or "open surgery" or "surgical procedure" or "open approach" or "microsurgery") and (("Teaching"[Mesh] or "Education"[Mesh] or "education" [Subheading]) or ("surgical skills" or "surgical training" or "skills assessment" or "skills classification" or "proficiency" or "surgical technique")) and ("Pattern Recognition, Automated"[Mesh] or "Artificial Intelligence"[Mesh] or "Machine Learning"[Mesh] or "Unsupervised Machine Learning"[Mesh] or "Supervised Machine Learning"[Mesh] or "Deep Learning"[Mesh] or "Automation"[Mesh] or "Neural Networks, Computer"[Mesh] or "convolutional neural networks" or "automated" or "computer vision")) NOT ((endoscop* [Title/Abstract] OR laparoscop* [Title/Abstract] OR robot* [Title/Abstract]))
Scopus	(ALL ("open surgery") AND ALL ("skills" OR "proficiency") AND ALL ("artificial intelligence" OR "machine learning" OR "automated") AND NOT TITLE-ABS-KEY ("robot*" OR "endoscop*" OR "laparoscop*"))
IEEE Xplore	("Full Text & Metadata":surgery) AND ("Full Text & Metadata": automated or artificial intelligence or machine learning or deep learning or neural networks) AND ("Full Text & Metadata":skills or proficiency or assessment or classification) NOT ("Publication Title": robot* or endoscop* or laparoscop*)

surgery in the title. Table 2 provides the Boolean search strings used for the different databases. The complete electronic search strategy used for PubMed is provided in Supplementary Content 1. Additional records were identified through the examination of review articles and citation searches.

3.4. Study selection

A total of 926 scientific records were identified through the searches. The results were imported into the Rayyan web app (<https://www.rayyan.ai/>) for duplicate removal, title/abstract, and full-text screening to ensure articles met the eligibility criteria. A total of 908 records were screened by title/abstract to exclude non-relevant studies according to the exclusion criteria. A total of 131 records underwent full-text screening. Two independent reviewers [HH and DS] were responsible for title/abstract and full-text screening process. Any disagreements on the eligibility of an article were resolved through discussions with a third reviewer [NN] when necessary. At the end of the screening process, 40 articles that met the inclusion criteria were selected for the review.

3.5. Data extraction

Data items were extracted from the included papers and summarized based on the type of open surgery performed, the phase of clinical trial, the sensor device used, the data captured, the automation algorithm employed, the predicted output, and the type of evaluation conducted. Two reviewers [HH and DS] jointly developed the data extraction form with the variables to be extracted from each eligible article for this scoping review. The reviewers [HH and DS] were independently responsible for extracting the data from each study. Resolution of disagreements on results of the data extraction was achieved through discussions with a third reviewer [NN] when necessary.

3.6. Quality assessment

The medical education research study quality instrument (MERSQI) was used to assess the studies and provide a score between 5 and 18. The articles were evaluated in terms of study design, sampling, type of data, validity of the evaluation instrument, data analysis, and outcome (Cook and Reed, 2015; Reed et al., 2007). A higher score given to an article indicates a superior study design. Same as the study selection and data extraction process, two reviewers [HH and DS] were responsible for quality assessment of the included studies, while disagreements were resolved through discussions with a third reviewer [NN] when necessary.

3.7. Data synthesis

The data extracted from the articles were presented using a descriptive table with information on open surgery specialty, procedure, clinical trial, type of data captured, sensor used to capture the data, automation algorithm used, predicted output, and primary findings of the study. In addition, visual summary of study characteristics such as surgical procedure, clinical trial, type of data captured, sensors utilized, automation algorithms, assessment inputs, and performance metrics were provided. A meta-analysis was not performed due to the wide range of clinical settings, type of data captured, and outcomes measured.

4. Results

As illustrated in Fig. 1, the screening of search results yielded 40 articles covering automated skills assessment for open surgery, published between 2008 and 2024. The articles included had an average MERSQI score of 12.35 out of 18. The articles covered objective data measurement (as opposed to self-reported data) and utilized appropriate

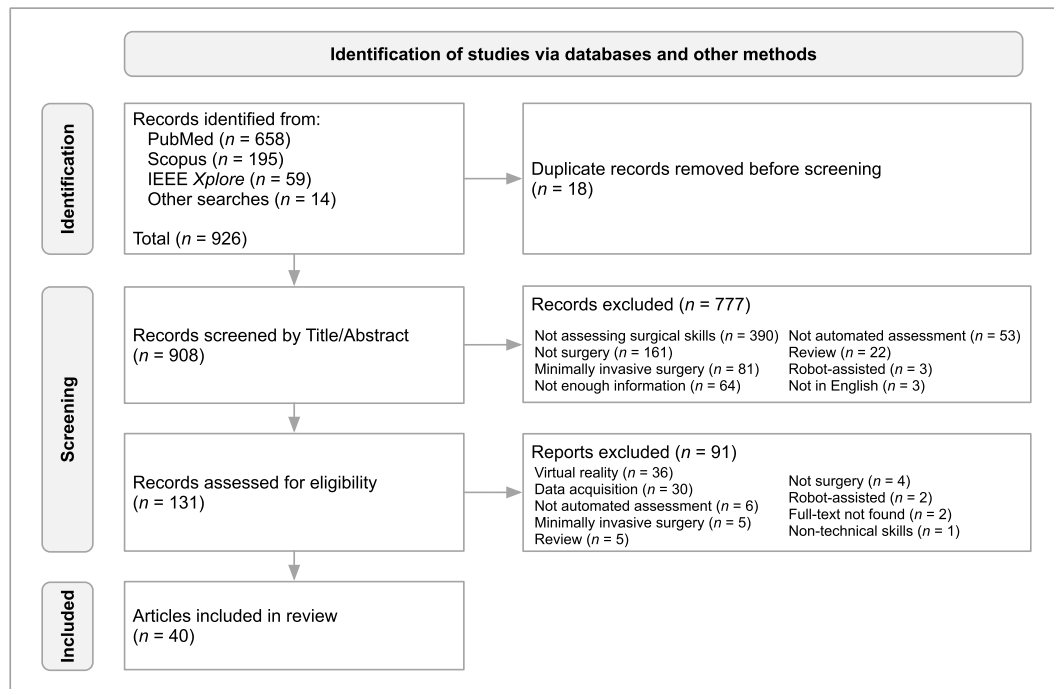


Fig. 1. PRISMA flowchart depicting the identification, screening, and inclusion of articles on automated skills assessment in open surgery.

data analysis which went beyond descriptive analyses resulting in a positive contribution to the MERSQI score. However, most of the studies were designed as single group tests that were not randomized and were all conducted at a single institution. Not all studies consistently reported the validity of the evaluation instruments utilized. In addition, the outcome measures were mostly limited to knowledge and skills. These factors greatly limited the overall MERSQI score obtained.

A summary of the search findings is presented in Table 3. In the *Automation Algorithms* column, an asterisk (\*) indicates a standalone algorithm, and an arrow (→) indicates a sequential algorithm where the output of one is fed into the next. For each algorithm, the contents within brackets specify the network architecture used. The *Results* column provides details about the assessment task achieved using the automation algorithm, with evaluation results presented as measured values of assessment metrics in brackets. These are followed by the assessment inputs on which the algorithm was based. The subsequent sections of this paper provide a detailed description of the table elements, including the types of open surgeries and corresponding surgical skills (section 4.1), the types of data acquired from the operative field in these open surgeries (section 4.2), and the various algorithms utilized to analyze the acquired data for automated assessment of surgical skills (section 4.3).

#### 4.1. Open surgeries (RQ1)

The articles reporting automated skill assessment for open surgery were classified into two categories based on the level of magnification under which the procedure was performed: macrosurgery (including cardiothoracic, colorectal, general, head & neck, plastic, and non-specific specialties) and microsurgery (including neurosurgery and ophthalmology), as depicted in Fig. 2. Various surgical procedures were identified under these specialties, such as mitral valve repair (Tozzi et al., 2022), pilonidal cystectomy (Goodman et al., 2024), appendectomy (Goodman et al., 2024), cholecystectomy (Rittenhouse et al., 2014), septoplasty (Ahmadi et al., 2015), thyroidectomy (Goodman et al., 2024), liposuction (Yibulayimu et al., 2022), knot tying (Azari et al., 2019, 2021a, 2021b; Bkheet et al., 2023; Kasa et al., 2022; Nagaraj et al., 2023; Nguyen et al., 2019; Shaharan et al., 2016, 2017; Sun et al.,

2016; Ying-Ying and Shulruf, 2019; Zia et al., 2018), suturing (Azari et al., 2019, 2021a, 2021b; Bkheet et al., 2023; Handelman et al., 2020; Hoffmann et al., 2024; Kil et al., 2024; Nagaraj et al., 2023; Nguyen et al., 2019; Sbernini et al., 2018; Shaharan et al., 2017; Singh et al., 2024; Yamada et al., 2022; Ying-Ying and Shulruf, 2019; Zia et al., 2018; Zuckerman et al., 2024), venous anastomoses (Watson, 2014), cerebrovascular procedures (Davids et al., 2021; Oliveira et al., 2022; Sugiyama et al., 2018), temporal lobectomy (Sugiyama et al., 2018), tumor resection (Baghdadi et al., 2023; Sugiyama et al., 2018), and cataract surgery (Hira et al., 2022; Kim et al., 2019; Ruzicki et al., 2023). Where specified, surgical steps, such as stitch placement, tissue dissection, liposuction strokes, arteriotomy & microsuture, coagulation & dissection, microvascular anastomosis and capsulorhexis, were also identified. The majority of articles evaluated the automated skill assessment technology for suturing (n = 18 articles) and/or knot tying (n = 13 articles) using low-fidelity synthetic phantoms (Azari et al., 2021a, 2021b; Bkheet et al., 2023; Handelman et al., 2020; Kasa et al., 2022; Nagaraj et al., 2023; Nguyen et al., 2019; Sbernini et al., 2018; Shaharan et al., 2016, 2017; Sun et al., 2016; Yamada et al., 2022; Ying-Ying and Shulruf, 2019; Zia et al., 2018). This was followed by automated skills assessment for the capsulorhexis step during live cataract surgeries (n = 3 articles) (Hira et al., 2022; Kim et al., 2019; Ruzicki et al., 2023). Fewer articles (n = 2 each) utilized automated skill assessment during live surgeries for knot tying and suturing (Azari et al., 2019, 2021a) as well as the coagulation and dissection steps during tumor resection (Baghdadi et al., 2023; Sugiyama et al., 2018). Similarly, low-fidelity synthetic phantoms were used for microvascular anastomosis (Sugiyama et al., 2024; Tang et al., 2024) and ophthalmic suturing (Franco-González et al., 2021; Handelman et al., 2020) (n = 2 articles each). The remaining surgical procedures were described individually (n = 1 article for each procedure/step).

#### 4.2. Data acquisition (RQ2)

The articles present different types of data captured from the operative field during open surgery for automated skills assessment. The operative field consists of the tissue being operated on by the surgeon using surgical instruments (Fig. 3a). Therefore, the data captured from



**Table 3**

Description of open surgery articles on automated skills assessment, by year.

Reference	MERSQI <sup>a</sup> Score	Open Surgery			Data Acquisition		Automated Skills Assessment		
		Surgical Specialty	Surgical Procedure & Skill	Clinical Trial	Data Captured <sup>b</sup>	Sensor Used	Automation Algorithm	Predicted Output	Results
Goodman et al. (2024)	12.5	Colorectal General Head & neck	Pilonidal cystectomy Appendectomy Thyroidectomy	Phantom (high fidelity), human (live)	H	Video camera	*LR <sup>c</sup>	Novice/Expert	Skill classification achieved (with custom evaluation, $p < 0.02$ ) based on <i>hand movement</i>
Hoffmann et al. (2024)	13.5	Not specific	Suturing	Phantom (low fidelity)	H, I, T	Video camera	*CNN <sup>d</sup> (TSN) → CNN (I3D) *CNN (TSN) → ViT (Video Swin)	Novice/Expert	Skill classification achieved (with accuracy 71 %, F <sub>1</sub> score 72 %) based on OSATS <sup>e</sup> scoring by expert raters
Kil et al. (2024)	12.5	Not specific	Suturing	Phantom (low fidelity)	I	Video camera	*Custom CV <sup>f</sup> algorithm	Novice/Expert	Skill classification achieved (with custom evaluation, $p < 0.01$ ) based on <i>instrument motion data in 2D</i>
Singh et al. (2024)	12.5	Not specific	Suturing	Phantom (low fidelity)	I, T	IMU <sup>g</sup> , EM <sup>h</sup> tracking system	*Custom non-CV algorithm	Novice/Expert	Skill classification achieved (with custom evaluation, $p < 0.05$ ) based on <i>instrument motion data in 3D and placement of knot/suture</i>
Sugiyama et al. (2024)	13.5	Neurosurgery Microsurgery	Cerebrovascular procedure - microvascular anastomosis	Phantom (low fidelity)	I	Video camera	*CNN (YOLOv2)	Novice/Expert	Skill classification achieved (with custom evaluation, $p < 0.01$ ) based on <i>instrument motion data in 2D and OSATS scoring</i>
Tang et al. (2024)	11.5	Neurosurgery Microsurgery	Cerebrovascular procedure - microvascular anastomosis	Phantom (low fidelity)	T	Video camera	*CNN (ResNet-50)	Novice/Expert	Skill classification achieved (with custom evaluation, $p < 0.01$ ) based on <i>placement of knot/suture</i>
Baghdadi et al. (2023)	11.5	Neurosurgery Microsurgery	Tumor resection - coagulation & dissection	Human (live)	I	Force sensor	*CNN (U-Net) → CNN (Inception-v4)	Novice/Expert	Skill classification achieved (with AUC <sup>i</sup> 81 %, F <sub>1</sub> score 71 %) based on <i>force from sensor on instrument</i>
Bkheet et al. (2023)	11.5	Not specific	Knot tying, suturing	Phantom (low fidelity)	H, I	Video camera	*CNN (YOLOX) → CNN (ResNet) → CNN (TCN) → Estimation	Novice/Expert	Skill classification achieved (with custom evaluation, $p < 0.05$ ) based on <i>hand movement</i>
Nagaraj et al. (2023)	11.5	Not specific	Knot tying, suturing	Phantom (low fidelity)	H, I	Video camera	*CNN (EfficientNet) *CNN (X3D)	Pass/Fail Pass/Fail	Skill classification achieved (with accuracy 83 %, F <sub>1</sub> score 69 %) based on <i>hand movement and placement of knot/suture</i>
Ruzicki et al. (2023)	11.5	Ophthalmology Microsurgery	Cataract surgery - capsulorhexis	Human (live)	I	Video camera	*CNN (ResNet-152) → RNN <sup>j</sup> (LSTM) → RF <sup>k</sup>	Novice/Expert	Skill classification achieved (with accuracy 63.3 %, AUC 69.2 %) based on <i>instrument motion data in 2D</i>
Xu et al. (2023)	10.5	Neurosurgery Microsurgery	Cerebrovascular procedure - dissection	Phantom (low fidelity)	H	Force sensor	*CNN (TCN)	Novice/Expert	Skill classification achieved (with accuracy 97.45 %) based on <i>hand movement</i>
Goldbraikh et al. (2022)	11.5	Not specific	Suturing	Phantom (low fidelity)	H, I	Video camera	*CNN (YOLOv3) → Estimation	Novice/Expert	Skill classification achieved (with custom evaluation, $p < 0.01$ ) based on <i>hand movement</i>

(continued on next page)

Table 3 (continued)

Reference	MERSQI <sup>a</sup> Score	Open Surgery			Data Acquisition		Automated Skills Assessment		
		Surgical Specialty	Surgical Procedure & Skill	Clinical Trial	Data Captured <sup>b</sup>	Sensor Used	Automation Algorithm	Predicted Output	Results
Hira et al. (2022)	13.5	Ophthalmology <i>Microsurgery</i>	Cataract surgery - capsulorhexis	Human (live)	I	Video camera	*CNN (TCN) → CNN (ResNet) → RNN (LSTM)	Novice/Expert	Skill classification achieved (with AUC 78 %, sensitivity 84.3 %, specificity 75 %) based on <i>instrument motion data in 3D</i>
Kasa et al. (2022)	13.5	Not specific	Knot tying	Phantom (low fidelity)	H, T	Video, depth video, image camera	*CNN (ResNet-18) → RNN (LSTM) *CNN (ResNet-50)	Score Score	OSATS scores predicted: Respect for tissue (MSE <sup>1</sup> 0.48, ICC <sup>in</sup> 30 %), time and motion (MSE 0.35, ICC 59 %), quality of final product (MSE 0.18, ICC 90 %), overall performance (MSE 0.19, ICC 74 %). The predictions were made based on <i>hand movement, placement of knot/suture</i> and OSATS scoring by expert raters
Oliveira et al. (2022)	11.5	Neurosurgery <i>Microsurgery</i>	Cerebrovascular procedure -arteriotomy & microsuture	Human (cadaver)	H, I	Video camera	*Custom CV algorithm	Score	The work did not evaluate the algorithm. It was proof-of-concept. The score ( <i>custom scoring</i> ) was based on <i>hand movements</i>
Soangra et al. (2022)	10.5	Not specific	Knot tying	Phantom (low fidelity)	H	EMG <sup>n</sup> sensor, IMU	*RF	Novice/Expert	Skill classification achieved (with accuracy 61 %) based on <i>hand movement</i>
Tozzi et al. (2022)	11.5	Cardiothoracic	Mitral valve repair - stitch placement	Phantom (high fidelity)	T	Video camera, embedded tissue sensor	*Custom non-CV algorithm	Score	Score ( <i>custom scoring</i> ): Experts rated system 9 on a scale of 1–10, based on <i>placement of knot/suture</i>
Yamada et al. (2022)	13.5	Not specific	Suturing	Phantom (low fidelity)	T	Image camera	*Custom CV algorithm	Score	Score ( <i>custom scoring</i> ): Showed correlation to expert OSATS (p < 0.001), based on accuracy of <i>placement of knot/suture</i>
Yibulayimu et al. (2022)	12.5	Plastic	Liposuction - liposuction strokes	Phantom (high fidelity), human (live)	I	Force sensor, optical tracking system	*RF	Novice/Expert	Skill classification achieved (with accuracy 92.93 %, sensitivity 92.92 %) based on <i>force from sensor on instrument, instrument motion data in 3D</i>
Azari et al. (2021b)	13.5	Not specific	Knot tying, suturing	Phantom (low fidelity)	H	Video camera	*GAM <sup>o</sup> *LR	Score Score	OSATS scores predicted: Fluidity of motion (R <sup>2</sup> 77 %), motion economy (R <sup>2</sup> 66 %), tissue handling (R <sup>2</sup> 57 %), hand coordination (R <sup>2</sup> 63 %). The predictions were made based on <i>hand movement</i> and OSATS scoring by expert raters
Azari et al. (2021a)	13.5	Not specific	Knot tying, suturing	Phantom (low fidelity),	H	Video camera	*GAM *LR	Score Score	OSATS scores predicted: Fluidity of motion (R <sup>2</sup> 55 %),

(continued on next page)

Table 3 (continued)

Reference	MERSQI <sup>a</sup> Score	Open Surgery			Data Acquisition		Automated Skills Assessment		
		Surgical Specialty	Surgical Procedure & Skill	Clinical Trial	Data Captured <sup>b</sup>	Sensor Used	Automation Algorithm	Predicted Output	Results
				human (live)					motion economy ( $R^2$ 49 %). The predictions were made based on <i>hand movement</i> and <i>OSATS scoring</i> by expert raters
<a href="#">Davids et al. (2021)</a>	13.5	Neurosurgery <i>Microsurgery</i>	Cerebrovascular procedure - dissection	Phantom (high fidelity)	I	Video camera	*CNN (Mask-RCNN)	Novice/Expert	Skill classification achieved (with accuracy 84.21 %, AUC 97.7 %) based on <i>instrument motion data in 2D</i>
<a href="#">Franco-González et al. (2021)</a>	11.5	Not specific <i>Microsurgery</i>	Suturing	Phantom (low fidelity)	I	Depth video camera	*Custom CV algorithm	Novice/Expert	Skill classification achieved (with custom evaluation, $p < 0.02$ ) based on <i>instrument motion data in 3D</i>
<a href="#">Handelman et al. (2020)</a>	10.5	Not specific Ophthalmology <i>Microsurgery</i>	Suturing - linear & circular suturing	Phantom (low fidelity), human (cadaver)	T	Embedded tissue sensor, image camera	*Custom CV algorithm	Score	The work did not evaluate the algorithm. The score ( <i>custom scoring</i> ) was based on <i>placement of knot/suture</i> and <i>force from sensor inside tissue</i>
<a href="#">Pérez-Escamirosa et al. (2020)</a>	12.5	Not specific	Knot tying, suturing	Phantom (low fidelity)	T	Embedded tissue sensor	*Custom non-CV algorithm	Novice/Expert	Skill classification achieved (with custom evaluation, $p < 0.05$ ) based on <i>force from sensor inside tissue</i>
<a href="#">Azari et al. (2019)</a>	13.5	Not specific	Knot tying, suturing	Human (live)	H	Video camera	*LR	Score	OSATS scores predicted for suturing, knot-tying: Fluidity of Motion ( $R^2$ 86 %, 54 %), motion economy ( $R^2$ 88 %, 64 %), tissue handling ( $R^2$ 69 %, 52 %). The predictions were made based on <i>hand movement</i> and <i>OSATS scoring</i> by expert raters
<a href="#">Kim et al. (2019)</a>	13.5	Ophthalmology <i>Microsurgery</i>	Cataract surgery - capsulorhexis	Human (live)	I	Video camera	*CNN (TCN)	Novice/Expert	Skill classification achieved (with accuracy 84.8 %, AUC 86.3 %, sensitivity 82.4 %, specificity 70.8 %) based on <i>instrument motion data in 3D</i>
<a href="#">Nguyen et al. (2019)</a>	10.5	Not specific	Knot tying, suturing	Phantom (low fidelity)	H	Video camera, IMU	*CNN → RNN (LSTM)	Novice/Expert	Skill classification achieved (with accuracy 98.2 %) based on <i>hand movement</i>
<a href="#">Ying-Ying and Shulruf (2019)</a>	15.5	Not specific	Knot tying, suturing	Phantom (low fidelity)	T	Video camera, embedded tissue sensor	Custom CV and non-CV algorithms	Score Pass/Fail	The study did not evaluate the algorithm. The algorithm provided OSATS scores on: safety, quality, efficiency. The algorithm's pass/fail classification was based on <i>placement of knot/</i>

(continued on next page)



Table 3 (continued)

Reference	MERSQI <sup>a</sup> Score	Open Surgery			Data Acquisition		Automated Skills Assessment		
		Surgical Specialty	Surgical Procedure & Skill	Clinical Trial	Data Captured <sup>b</sup>	Sensor Used	Automation Algorithm	Predicted Output	Results
Sbernini et al. (2018)	10.5	Not specific	Suturing	Phantom (low fidelity)	H	IMU, force sensor	*ANN <sup>p</sup>	Novice/Expert	<i>suture and force from sensor inside tissue</i> Skill classification achieved (with accuracy 90 %) based on <i>hand movement</i>
Sugiyama et al. (2018)	11.5	Neurosurgery <i>Microsurgery</i>	Cerebrovascular procedures, temporal lobectomy, tumor resection - coagulation & dissection	Human (live)	I	Force sensor	*CDA <sup>q</sup>	Novice/Expert	Skill classification achieved (with accuracy of 87.5 %) based on <i>force from sensor on instrument</i>
Zia et al. (2018)	13.5	Not specific	Knot tying, suturing	Phantom (low fidelity)	H, I	Video camera, IMU	*KNN <sup>r</sup>	Novice/Expert	Skill classification achieved (with accuracies of 94 % and 93.2 % for knot-tying and suturing tasks) based on <i>instrument motion data in 3D, hand movement</i> (and OSATS)
Shaharan et al. (2017)	13.5	Not specific	Knot tying, suturing	Phantom (low fidelity)	H	EM tracking system	*Custom non-CV algorithm	Novice/Expert	Skill classification achieved (with custom evaluation, $p < 0.0001$ ) based on <i>hand movement</i>
Shaharan et al. (2016)	13.5	Not specific	Knot tying	Phantom (low fidelity)	H	Video camera, EM tracking system	*Custom non-CV algorithm	Score	Score ( <i>custom scoring</i> ): Showed correlation to expert OSATS ( $p < 0.05$ ), based on <i>hand movement</i>
Sun et al. (2016)	10.5	Not specific	Knot tying	Phantom (low fidelity)	H	Video, depth video camera	*HMM <sup>s</sup>	Novice/Expert	Skill classification achieved (with accuracy 100 %) based on <i>hand movement</i>
Ahmidi et al. (2015)	10.5	Head & neck	Septoplasty - tissue dissection	Human (live)	I	Depth video camera, EM tracking system	*Custom non-CV algorithm	Novice/Expert	Skill classification achieved (with accuracy 91 %, sensitivity 88.45 %) based on <i>instrument motion data in 3D</i>
Rittenhouse et al. (2014)	13.5	General	Cholecystectomy	Phantom (high fidelity)	H	Video camera, EM, optical tracking systems	*Custom non-CV algorithm	Novice/Expert	Skill classification achieved (with custom evaluation, $p < 0.05$ ) based on <i>hand movement</i> (and OSATS)
Watson (2014)	11.5	Not specific	Venous anastomoses	Phantom (low fidelity)	H	IMU	*SVM <sup>t</sup>	Novice/Expert	Skill classification achieved (with accuracy 83 %, sensitivity 86 %, specificity 80 %) based on <i>hand movement</i>
Frischknecht et al. (2013)	11.5	Not specific	Suturing	Phantom (low fidelity)	T	Image camera	*Custom CV algorithm	Novice/Expert	Skill classification achieved (with custom evaluation, $p < 0.05$ ) based on <i>placement of knot/suture</i>
Solis et al. (2008)	15.5	Not specific	Knot tying, suturing	Phantom (low fidelity)	T	Video camera, embedded tissue sensor	*DA <sup>u</sup>	Novice/Expert	Skill classification achieved (with custom evaluation, $p < 0.001$ ) based on <i>placement of knot/suture and force from sensor inside tissue</i>

<sup>a</sup> MERSQI stands for Medical Education Research Study Quality Instrument.

- <sup>b</sup> H – Surgeon’s hand, I – Surgical instrument, T – Operative tissue.  
<sup>c</sup> LR stands for Linear Regression.  
<sup>d</sup> CNN stands for Convolutional Neural Network.  
<sup>e</sup> OSATS stands for Objective Structured Assessment of Technical Skills.  
<sup>f</sup> CV stands for Computer Vision.  
<sup>g</sup> IMU stands for Inertial Measurement Unit.  
<sup>h</sup> EM stands for Electromagnetic.  
<sup>i</sup> AUC stands for Area Under the Curve.  
<sup>j</sup> RNN stands for Recurrent Neural Network.  
<sup>k</sup> RF stands for Random Forest.  
<sup>l</sup> MSE stands for Mean Squared Error.  
<sup>m</sup> ICC stands for Interclass Correlation.  
<sup>n</sup> EMG stands for Electromyography.  
<sup>o</sup> GAM stands for Generalized Additive Model.  
<sup>p</sup> ANN stands for Artificial Neural Network.  
<sup>q</sup> CDA stands for Canonical Discriminant Analysis.  
<sup>r</sup> KNN stands for *k*-Nearest Neighbor.  
<sup>s</sup> HMM stands for Hidden Markov Model.  
<sup>t</sup> SVM stands for Support Vector Machine.  
<sup>u</sup> DA stands for Discriminant Analysis.

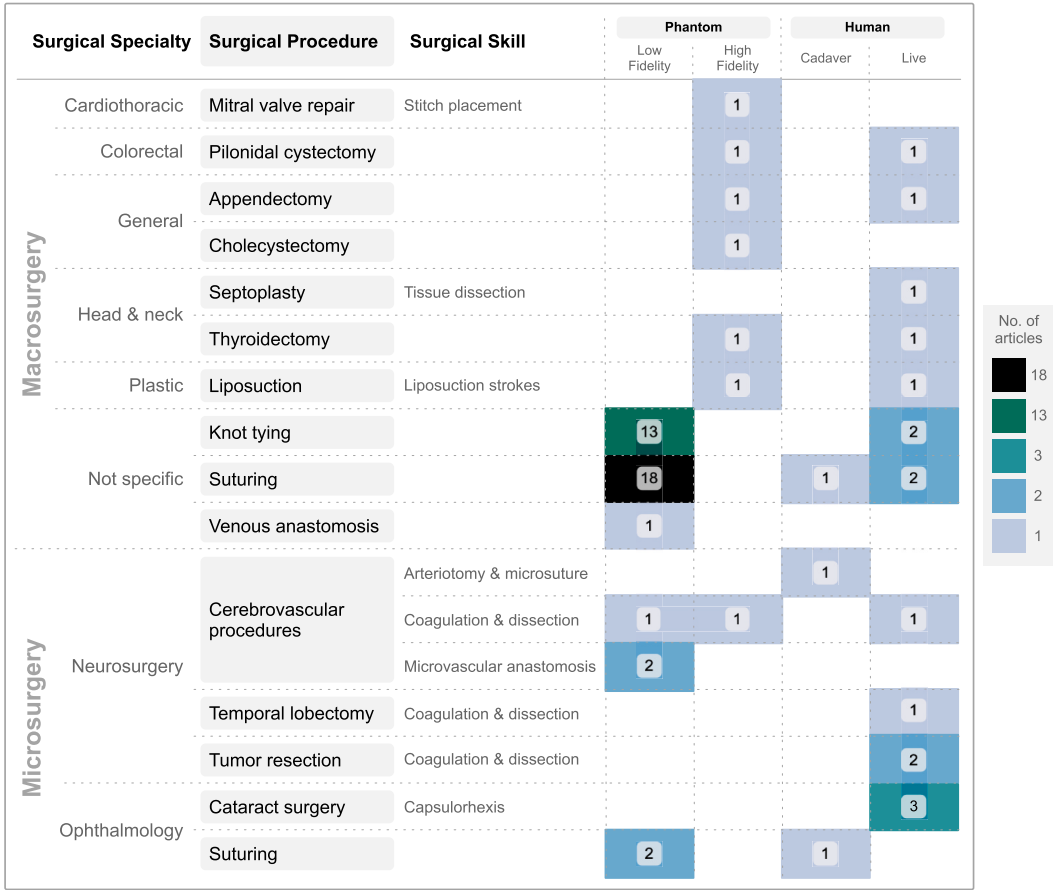
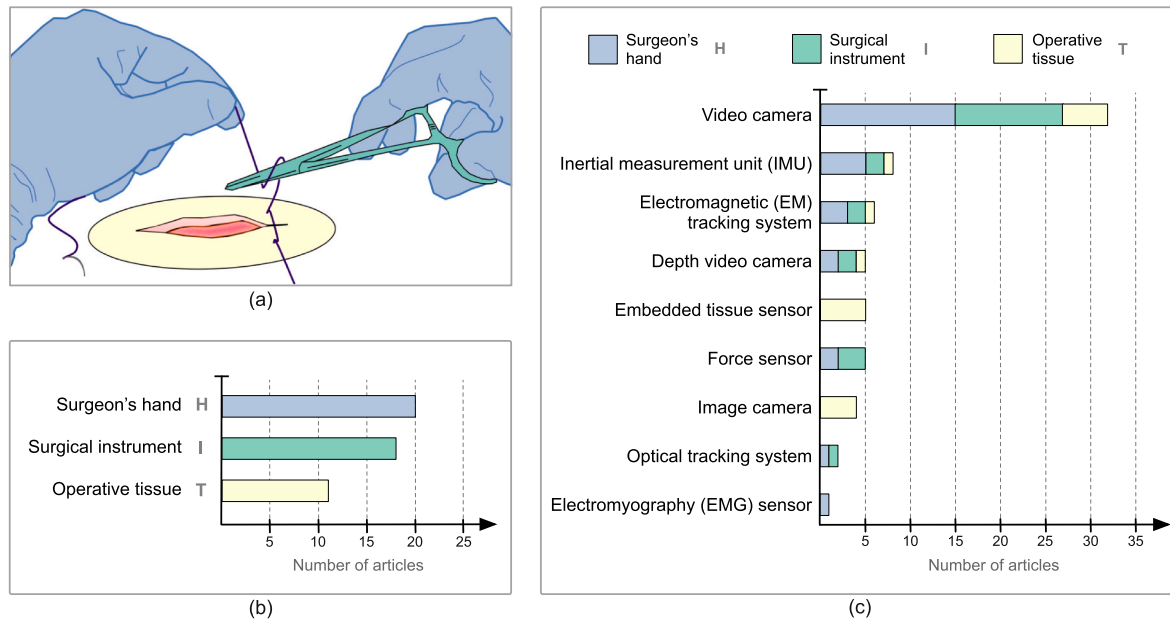


Fig. 2. Number of open surgery articles using automated surgical skill assessment under different phases of clinical trials.

the operative field was classified into three categories: (i) surgeon’s hand motion (denoted by ‘H’), (ii) surgical instrument motion (denoted by ‘I’), and (iii) operative tissue interaction (denoted by ‘T’). As shown in Fig. 3b, studies focusing on hand motion (H) and surgical instrument motion (I) were significantly more common compared to those focusing on operative tissue (T). This is because hand motion metrics effectively distinguish between experience levels of the surgeon, such as in differentiating skilled and unskilled suturing maneuvers (Bkheet et al., 2023; Nagaraj et al., 2023; Shayan et al., 2023). Experienced surgeons exhibit smoother hand motions, higher average roll angles, and fewer roll movements compared to novices, making hand

rotational metrics valuable for skills assessment (Shayan et al., 2023). Similarly, surgical instrument motion metrics, such as, time taken, path length, collision, unnecessary movements, and force exerted, can also differentiate experts from novices (Guerin et al., 2022; Poursartip et al., 2017). Fewer studies utilized operative tissue data due to challenges in capturing accurate data: (i) the final image of the tissue post-procedure may not accurately reflect surgical performance, and (ii) measuring the force exerted on the tissue by surgical instruments is difficult without specialized phantoms (Kasa et al., 2022; Tozzi et al., 2022; Yamada et al., 2022). As depicted in Fig. 3c, various types of sensors were used to acquire



**Fig. 3.** Depiction of (a) the operative field during open surgery, (b) number of articles capturing data in the form of surgeon's hand motion - H, surgical instrument motion - I, and operative tissue interaction - T, and (c) number of articles using various sensors to capture the data.

data from the open surgical field. Two-dimensional (2D) video cameras, inertial measurement units (IMU), electromagnetic (EM) tracking systems, and three-dimensional (3D) depth cameras were used to capture all the three data formats (H, I, and T). The 2D video cameras were extensively utilized due to their affordability and ease of use (Deng et al., 2021). Wearable IMUs equipped with accelerometers provided 3D hand motion measurements (Nguyen et al., 2019; Sbernini et al., 2018). Studies utilizing EM tracking systems placed the sensors either on surgeon's hands (Shaharan et al., 2016, 2017) or the surgical instruments (Ahmidi et al., 2015; Singh et al., 2024). Depth video cameras included Microsoft Kinect, which captured the operative field (Ahmidi et al., 2015), and Leap Motion, which was used to capture 3D hand motion (Kasa et al., 2022; Sun et al., 2016). Embedded tissue sensors (Nguyen et al., 2019; Sbernini et al., 2018; Watson, 2014; Zia et al., 2018) and image cameras (Ahmidi et al., 2015; Shaharan et al., 2016, 2017; Singh et al., 2024) were mostly used for T data. Image cameras captured the outcome in high resolution for evaluation, while embedded sensors used in synthetic phantoms detected incorrect stitches (Tozzi et al., 2022) and measured strain applied to the tissue (Ahmidi et al., 2015; Shaharan et al., 2016, 2017). Force sensors (Sbernini et al., 2018) and optical tracking systems (Shaharan et al., 2016, 2017) were used to capture H and I. Force sensors, such as bipolar forceps (Baghdadi et al., 2023; Sugiyama et al., 2018) and modified instrument handles (Yibulayimu et al., 2022), provided data on the resistance exerted, whereas optical trackers placed on hands and instruments (Rittenhouse et al., 2014; Yibulayimu et al., 2022). Electromyography (EMG) sensors captured H data through wearable components that recorded muscle activity (Soangra et al., 2022).

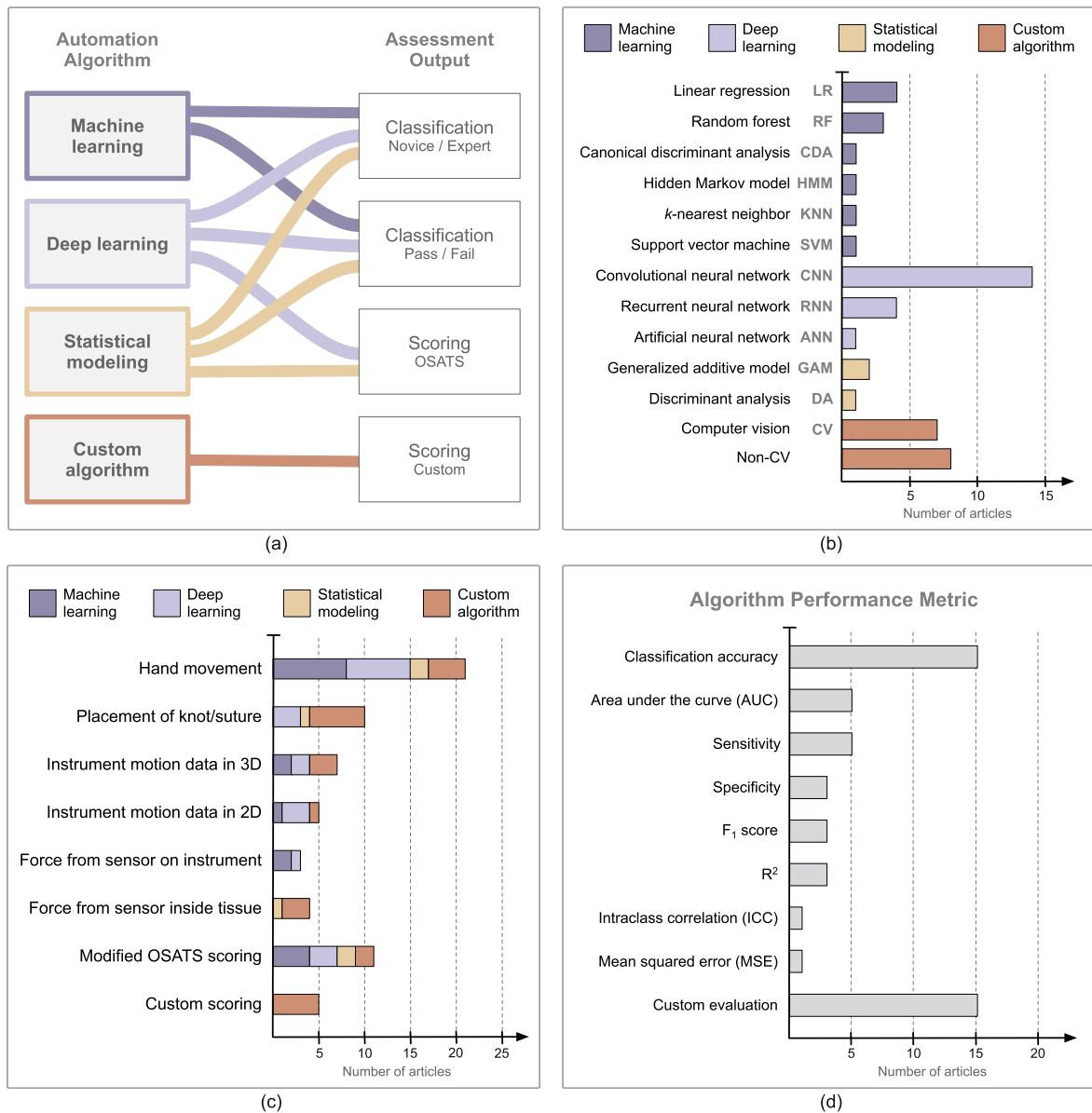
#### 4.3. Automated skills assessment (RQ3)

The automation algorithms in the reviewed articles can be categorized into: (i) machine learning (ML), (ii) deep learning (DL), (iii) statistical modeling, and (iv) custom algorithms.

Automation algorithms refer to the technologies used for automated feedback or assessment of surgical skills. These algorithms perform skill assessments and provide outputs in the form of: (i) classification as "novice" or "expert," (ii) classification as "passed" or "failed," (iii) scoring based on Objective Structured Assessment of Technical Skills (OSATS), or (iv) custom scoring that rates performance during training.

A mapping of the automation algorithms to their skill assessment outputs is depicted in Fig. 4a. The algorithms and their corresponding models (shown in Fig. 4b) are described below.

- (i) **ML algorithms** involve several pre-processing steps, such as extracting features from input data before learning or predicting from the data. They are effective for simpler tasks and smaller datasets. The ML algorithms used in the articles include linear regression (LR) (Azari et al., 2019, 2021a, 2021b), random forest (RF) (Ruzicki et al., 2023; Yibulayimu et al., 2022), canonical discriminant analysis (CDA) (Sugiyama et al., 2018), hidden Markov model (HMM) (Sun et al., 2016), *k*-nearest neighbor (KNN) (Zia et al., 2018), and support vector machine (SVM) (Watson, 2014).
- (ii) **DL algorithms** consist of neural networks that process data through multiple layers to extract information and improve predictions over time. Unlike ML, DL eliminates the need for extensive pre-processing by automatically extracting information through data abstraction. DL is effective for complex tasks, large datasets, and unstructured data like images and text. The DL algorithms used in the articles include convolutional neural networks (CNN) (Baghdadi et al., 2023; Bkheet et al., 2023; Davids et al., 2021; Hira et al., 2022; Kasa et al., 2022; Kim et al., 2019; Nagaraj et al., 2023; Nguyen et al., 2019; Ruzicki et al., 2023), recurrent neural networks (RNN) (Hira et al., 2022; Kasa et al., 2022; Nguyen et al., 2019; Ruzicki et al., 2023), and artificial neural networks (ANN) (Sbernini et al., 2018).
- (iii) **Statistical modeling algorithms** are similar to ML in terms of analyzing data, drawing inferences from patterns, and making predictions. However, while ML is more algorithm-driven and focuses on predictive accuracy, statistical modeling focuses on formal data modeling and understanding underlying data distributions. The statistical modeling algorithms utilized in the articles include generalized additive model (GAM) (Azari et al., 2021a, 2021b) and discriminant analysis (DA) (Solis et al., 2008).
- (iv) **Custom algorithms** refer to automation methodologies that do not fall under any of the above categories. These algorithms produce estimations rather than predictions. In this review, custom algorithms are classified as computer vision (CV)-based (if their outputs are calculated based on captured image data).



**Fig. 4.** (a) Depiction of automation algorithms and the skills assessment outputs. Number of articles utilizing (b) the algorithms and their corresponding models, (c) various assessment inputs for the automation algorithms, and (d) performance metrics to evaluate the algorithm.

(Handelman et al., 2020; Oliveira et al., 2022; Yamada et al., 2022; Ying-Ying and Shulruf, 2019) or as non-CV-based algorithms (Ahmidi et al., 2015; Shaharan et al., 2016, 2017; Tozzi et al., 2022; Ying-Ying and Shulruf, 2019).

The automation algorithms and corresponding models utilized for different surgical specialties and procedures are shown in Fig. 5. In some cases, the raw data extracted from sensors based on H, I, or T was further processed to generate input for the automation algorithms. A description of the processed data (assessment input) is given in Table 4. These assessment inputs and the corresponding automation algorithms utilized are shown in Fig. 4c. The automation algorithms used for skills assessment were evaluated using various performance metrics. A detailed description of these metrics is presented in Table 5 and Fig. 4d.

## 5. Discussion

This review aimed to outline the various technologies available for automated skills assessment during open surgery. The number of

published articles has notably increased since 2021, reflecting a growing interest in this topic (Fig. 6). Automated skills assessment during open surgery was mostly applied for suturing and knot tying on low-fidelity synthetic phantoms. The surgeon's hand motion captured using video cameras was the most frequently used type of data. Deep learning algorithms using hand movements as assessment inputs were predominantly reported, while the algorithm performance was mainly assessed using classification accuracy. The subsequent sections discuss the surgical procedures and clinical trial used in testing the technology (section 5.1), challenges in acquiring data during open surgery (section 5.2), common automation algorithms (section 5.3) with detailed applications of CNNs (section 5.3.1), RNNs (section 5.3.2) and transformers (section 5.3.3), performance evaluation and assessment inputs (section 5.4), and limitations of the review (section 5.5).

### 5.1. Common surgical procedures and clinical trials

A low complexity level of the surgical procedure was observed in most studies, which focused on basic surgical skills such as knot tying

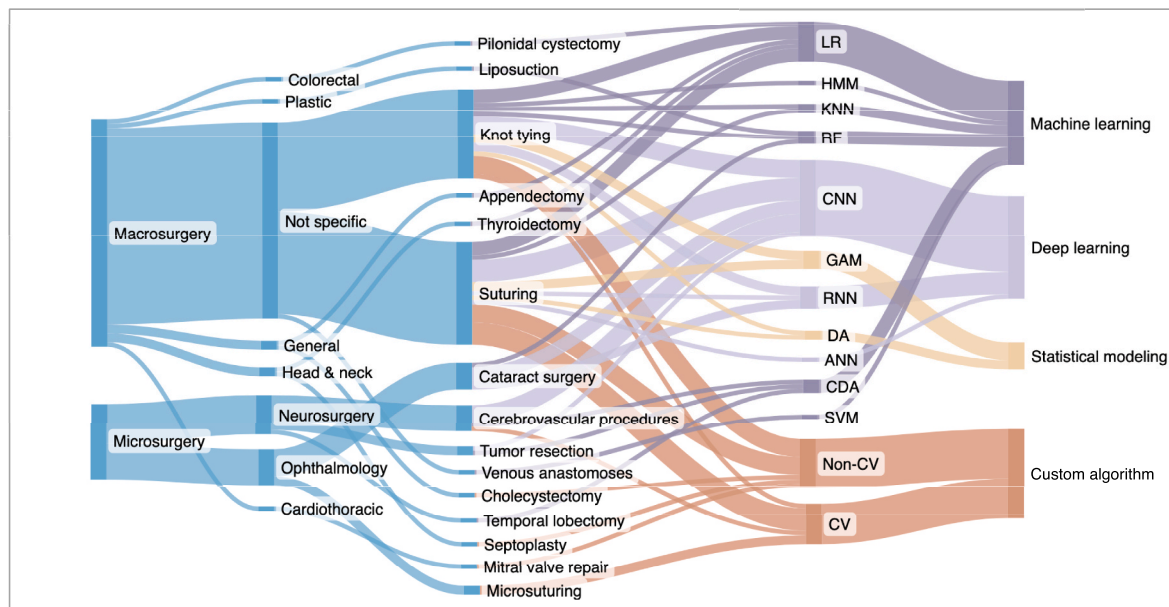


Fig. 5. Distribution of automation algorithms and models used for various surgical specialties and procedures.

and suturing (Bkheet et al., 2023; Nagaraj et al., 2023). Although considered simple tasks, they are fundamental skills requiring precise maneuvers (Davids et al., 2021; Huffman et al., 2020). As a result, these procedures are extensively practiced on low-fidelity synthetic phantoms, providing an inexpensive solution for surgical training with haptic feedback (Feifer et al., 2011). Applying automated skills assessment in such training setups requires minimal effort, as it allows for repeated trials and data acquisition (Watson, 2014). Similarly, multiple articles covered live cataract surgeries, facilitated by the microscope camera used during microsurgery, which provided video recordings for skills assessment (Franco-González et al., 2021; Titov et al., 2023). The coagulation and dissection steps in cerebrovascular procedures, temporal lobectomy, and tumor resection were also extensively used for applying automated skills assessment, with data easily collected from the sensorized bipolar forceps (SmartForceps System) (Baghdadi et al., 2023; Sugiyama et al., 2018). The current research landscape focuses on implementation of automated skills assessment in less complex procedures. While basic surgical skills practiced on simulators prepares a trainee towards the operating room experience, it does not translate into proficiency for increasingly complex procedures (Irfan et al., 2019). Requirements for general surgical trainees around the world often stipulate a variety of procedure-specific surgeries. For example, the Intercollegiate Surgical Curriculum Programme (ISCP) in the United Kingdom requires experience in a minimum number of index procedures including appendectomy, inguinal hernia repair, cholecystectomy, segmental colectomy, emergency laparotomy, and Hartmann's procedure (Elsey et al., 2017). Although current automated skill assessment efforts often focus on relatively basic tasks (e.g., knot tying or suturing on synthetic models), future developments must expand into evaluation of more complex surgical steps or even entire surgical procedures from start to finish. Such implementations would require incorporation of accurate surgical phase detection and tool recognition (Dick et al., 2024). A comprehensive assessment of surgeon's capabilities can be achieved by combining effective surgical workflow recognition along with hand and instrument motion analysis (Guerin et al., 2022). Combining kinematic, force, and video data may reveal the subtle cues that differentiate proficiency levels in intricate, multi-step procedures. Furthermore, advanced surgical skills assessment may be accomplished through the integration of real-time estimation of operating time (Kawka et al., 2022) as well as prediction of surgical outcomes (such as blood loss estimation) (Pangal et al., 2022).

The majority of studies were conducted on synthetic phantoms (with at least 67.5 % using low-fidelity phantoms), with a few on cadavers (5 %) or live patients (25 %). Synthetic phantoms are widely used for surgical training since they are readily available, easy to assemble, and cheaper to maintain. Some phantoms can also mimic both the hard and soft tissue of the human body depending on the synthetic material used in its making. Since they are easily reproducible, they also provide a standardized training platform for surgical skills acquisition (Raeker-Jordan et al., 2022). On the other hand, cadaveric models can be difficult to maintain, prone to degradation, limited to single usage depending on the surgical procedure, and can be relatively expensive to procure and maintain. This explains the low adoption of cadavers in the automated skills assessment studies. About 25 % of the studies captured metrics associated with operative tissue. Nevertheless, they were all limited to synthetic phantoms. Sensors embedded to the synthetic tissue is used to collect data such as strain and force exerted on the tissue (Solis et al., 2008). Such embedded sensors include optical fibers on simulated wound to measure strain applied while suturing (Handelman et al., 2020) and accelerometers placed under a tissue platform to measure the movement of the tissue during interaction with the tool (Pérez-Escamiroso et al., 2020). However, it is challenging to incorporate such embedded sensors during live surgeries due to safety and sterility concerns. Although training on animal models is known to improve surgical skills by providing a highly realistic environment (DeMasi et al., 2016), none of the included articles utilized animal models for automated skills assessment. This may be explained by ethical considerations and public approval concerns regarding the use of live animal models for routine surgical training (as opposed to their broader acceptance for testing novel surgical techniques) (Bergmeister et al., 2020; Ruan et al., 2020). Future work should prioritize the application of automated skills assessment during live surgeries, ensuring that operating theatres are well-equipped for this purpose (Lam et al., 2022). Integrating automated skills assessment into training is expected to reduce the teaching load of expert surgeons (Titov et al., 2023).

Some of the barriers to clinical translation of automated skills assessment in surgical practice include the lack of annotated datasets and high-quality studies. A cultural shift in clinical management is required to promote data acquisition, annotation, and storage in the operating rooms. Evaluation of the studies based on MERSQL revealed that most of them followed a single group study design without



**Table 4**  
Detailed description of assessment inputs for the automation algorithms.

Assessment Input	Description
Hand movement	Hand movement data (including motion of wrists, fingertips, finger joints, palm orientations, hand gestures) can be used to determine patterns in surgeon's hand motion (Sun et al., 2016; Zia et al., 2018) quality of surgeon's technique (Watson, 2014), manual expertise (Sbernini et al., 2018), duration of a certain gesture, subtlety in hand movements, and handling of tools, for example, during suturing (Bkheet et al., 2023)
Placement of knot/suture	Placement of knot/suture on a tissue can be used to determine quality of surgical subtask (Handelman et al., 2020; Tozzi et al., 2022; Yamada et al., 2022; Ying-Ying and Shulruf, 2019) and deformations to tissue (Tang et al., 2024)
Instrument motion data in 3D	Surgical instrument's motion data (including tooltip position, rotation, velocity, acceleration as well as tool path and trajectory) extracted from sensors in 3D space can be used to determine length of tool strokes, regularity of tool trajectories (Kim et al., 2019; Yibulayimu et al., 2022), smoothness of tool motion, and area covered by tooltip (e.g. septoplasty (Ahmadi et al., 2015))
Instrument motion data in 2D	Tracking the positions of the tooltips in video frames of the operative field can be used to determine patterns of tool usage (Ruzicki et al., 2023), movement with respect to tissue, and tool velocity profiles (Davids et al., 2021; Kil et al., 2024; Sugiyama et al., 2024)
Force from sensor on instrument	The forces measured by the sensors on the tool can be used to determine active periods of utilization of tool during the surgery (Baghdadi et al., 2023), strength exerted during tool strokes (e.g. liposuction surgery (Yibulayimu et al., 2022)), finesse during surgical task (e.g. dissection (Baghdadi et al., 2023)), and variability in contrast to an expert handling the tool (Sugiyama et al., 2018)
Force from sensor inside tissue	Force measured from sensors embedded within artificial skin/tissue can be used to determine distributed strain around a wound while performing surgery (Handelman et al., 2020) and quality of task (e.g. quality domain of OSATS for suturing/ligature skill) (Ying-Ying and Shulruf, 2019)
Modified OSATS scoring	Modified Objective Structural Assessment of Technical Skills (OSATS) scores are used to assess surgical skills in new domains such as fluidity of motion (Azari et al., 2019, 2021a, 2021b), motion economy (Azari et al., 2019, 2021a, 2021b), tissue handling (Azari et al., 2019, 2021a, 2021b), hand coordination (Azari et al., 2021b), quality of final product (Kasa et al., 2022; Ying-Ying and Shulruf, 2019), respect for tissue (Kasa et al., 2022), time and motion (Kasa et al., 2022), overall performance (Kasa et al., 2022), safety (Ying-Ying and Shulruf, 2019), efficiency (Ying-Ying and Shulruf, 2019), and Global Rating Score (GRS) (Hoffmann et al., 2024; Sugiyama et al., 2024)
Custom scoring	A study-specific customized scoring is used (based on the type of developed automation algorithm) to determine a proficiency index (Oliveira et al., 2022), errors and out-of-bound actions during surgery (Tozzi et al., 2022), and quality of surgical technique (Handelman et al., 2020; Yamada et al., 2022)

randomization. This limitation in study quality further restricts the clinical translation of automated skills assessment in surgical practice. The quality of research in this area can be improved by encouraging multiple institutional collaborations to ensure generalizability of the findings. Standardized evaluation instruments must be utilized to improve the credibility of the assessment. Beyond perceptions and knowledge, the outcomes measured should also include the surgeon's behavior in a live setting as well as the effect on patient outcomes. Clinical validation of skill assessment technologies can be achieved by designing and conducting appropriate randomized controlled trials (RCT). This will pave the way for conducting future meta-analyses of RCTs, which is the gold standard for informing clinical practice. As advancements in the field continue to develop, integration with current surgical workflow must also be investigated to facilitate clinical

**Table 5**  
Description of performance metric used to evaluate automation algorithms.

Performance Metric	Description	Assessment Output & Reference
Classification accuracy	Classification accuracy is the ratio of correct predictions to the total number of predictions. It is a measure of how often an ML model correctly predicts the class.	Novice/Expert classification (Ahmadi et al., 2015; Davids et al., 2021; Kim et al., 2019; Nguyen et al., 2019; Ruzicki et al., 2023; Sbernini et al., 2018; Sugiyama et al., 2018; Sun et al., 2016; Watson, 2014; Yibulayimu et al., 2022; Zia et al., 2018) Pass/Fail classification (Nagaraj et al., 2023)
Area under the curve (AUC)	Receiver Operating Characteristic (ROC) curve is a graphical plot that determines the diagnostic ability of a classification model. It is created by plotting true positive rate (recall) against false positive rate (also known as probability of false alarm). The total area under this curve (AUC) gives a probability estimation of how well an ML model can distinguish between two classes.	Novice/Expert classification (Baghdadi et al., 2023; Davids et al., 2021; Hira et al., 2022; Kim et al., 2019; Ruzicki et al., 2023)
Sensitivity	Sensitivity calculates the percentage of correct classes detected. It measures the ability of a model to identify all actual positive cases. It is the ratio of true positives to the sum of true positives and false negatives (positive cases that were missed). We report the average sensitivity across all classes.	Novice/Expert classification (Ahmadi et al., 2015; Hira et al., 2022; Kim et al., 2019; Watson, 2014; Yibulayimu et al., 2022)
Specificity	Specificity assesses how well a model correctly identifies negative instances. It measures the proportion of true negative predictions (correctly identified negative cases) relative to all actual negative cases in the dataset. We report the average specificity across all classes.	Novice/Expert classification (Hira et al., 2022; Kim et al., 2019; Watson, 2014)
F <sub>1</sub> score	F <sub>1</sub> score provides a balance between how many relevant items are retrieved (recall) and how many of the retrieved items are relevant (precision). A higher F <sub>1</sub> score indicates a better overall performance of the classification model.	Novice/Expert classification (Baghdadi et al., 2023; Hoffmann et al., 2024) Pass/Fail classification (Nagaraj et al., 2023)
R <sup>2</sup>	R <sup>2</sup> , also known as the coefficient of determination, is a statistical measure used to assess how well a numerical prediction model fits the observed data.	OSATS scoring (Azari et al., 2019, 2021a, 2021b)
Intraclass correlation (ICC)	ICC is a metric used in statistics to assess the consistency or agreement between measurements made by different observers or methods on the same set of subjects or items.	OSATS scoring (Kasa et al., 2022)
Mean squared error (MSE)	MSE is the difference between actual and predicted numerical predictions which is then summed and averaged over total number of predictions.	OSATS scoring (Kasa et al., 2022)

(continued on next page)



Table 5 (continued)

Performance Metric	Description	Assessment Output & Reference
Custom evaluation	The metrics used to evaluate performance of custom automation methodologies fall under this category.	Novice/Expert classification (Bkheet et al., 2023; Shaharan et al., 2017) Pass/Fail classification (Ying-Ying and Shulruf, 2019) Custom scoring (Handelman et al., 2020; Oliveira et al., 2022; Shaharan et al., 2016; Tozzi et al., 2022; Yamada et al., 2022; Ying-Ying and Shulruf, 2019)

translation of automated skills assessment technologies (Maier-Hein et al., 2022).

## 5.2. Challenges in open surgery data acquisition

### 5.2.1. Overview of challenges in data acquisition

Limited application of automated skills assessment in the operating room during live open surgery has been noted. This limitation is likely due to difficulty in obtaining accurate data from the operative field. For instance, existing data acquisition based on optical tracking is limited by clear line-of-sight requirements, which are difficult to maintain due to frequent obstructions by hand, instruments, and human motion during live procedures. The data acquired through EM tracking is susceptible to interference from the presence of metallic or other ferromagnetic materials near the tracking sensor, including the instrument itself. IMU can capture hand movements to differentiate between skilled and unskilled suturing (Shayan et al., 2023). However, the sensor must be placed on the surgeon's hand and wrist (Singh et al., 2024) which is impractical during open surgery due to sterility issues. Furthermore, the magnetometer within IMUs faces interference issues similar to those faced by EM sensors, despite gyro-assisted compensation (Ren and Kazanzides, 2009). Depth sensors require expensive hardware and are severely affected by occlusions and clutter. Moreover, for robustness they require multiple depth capturing setups that typically rely on computationally complex background subtraction algorithms (Kadkhodamohammadi et al., 2017). Embedded tissue sensors, which measure strain and force applied to the tissue (Solis et al., 2008), also face challenges in being used on the patient's body during live open surgery due to safety and sterility concerns. Force sensors used for automated skills assessment

during open surgery consists of two types: sensorized surgical instrument that can record tool-tissue force data (Baghdadi et al., 2023; Sugiyama et al., 2018), and sensors placed on surgeon's hands capturing the force exerted on the instrument held (Sbernini et al., 2018; Xu et al., 2023). The use of sensorized forceps is typically limited to neurosurgery and may not be applicable for majority of other open surgical procedures. Additionally, the force sensors placed on the surgeon's hands raise sterility concerns. Ensuring reliable placement of piezoresistive sensors for optimal measurements across different instruments can also be difficult (Xu et al., 2023). EMG sensors are placed on surgeon's thumb, arm, and shoulder to capture muscle activity (Soangra et al., 2022). However, this is also not practical due to sterility concerns during live surgery.

Concerns about safety, sterility, and ease of acquisition can be overcome with the use of image (RGB)-based cameras placed at a distance from the surgeon's hands and patient body. However, acquiring high-quality video footage during open surgeries involving large incisions can present its own difficulties (Deng et al., 2021; Rafiq et al., 2004). Unlike minimally invasive procedures where the camera scope captures the full operative view with relative ease, video footages from open surgeries can be partially obstructed with the operating surgeon's head or body (Bilgic et al., 2022). Moreover, image cameras can capture the final result of knot tying and suturing tasks (Kasa et al., 2022; Yamada et al., 2022). However, they do not record hand or instrument movements. This makes them unsuitable for advanced surgical procedures where the process must also be captured to gauge the surgeon's skills. To address this, various camera setups (such as head-mounted, body-mounted, tripod, or overhead surgical lights) have been tested. However, these setups can have drawbacks, including unnecessary motion causing blur and low quality footage, short battery life, unintentional occlusion by operating room staff, and excessive light exposure (Kajita et al., 2020).

### 5.2.2. Deep learning solutions to data acquisition challenges

**5.2.2.1. Robustness of algorithms to suboptimal data.** Challenges within image-based data predominantly affect computer vision-based (image processing) algorithms and traditional machine learning algorithms. These algorithms rely on inflexible assumptions about objects based on their appearance, structure and motion. Additionally, they exhibit inability to interpret meaningful information and handle significant appearance changes, due to dependence on manually designed features such as Histogram of Oriented Gradients, Color Names, and Scale-

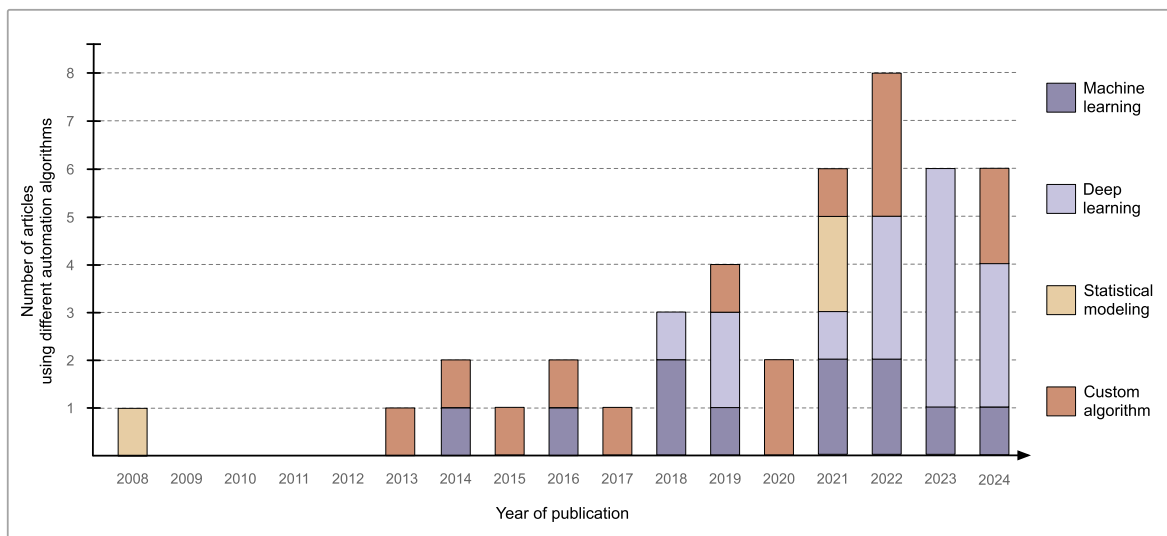


Fig. 6. Number of articles using various automation algorithms, by year of publication.

Invariant Feature Transforms (Marvasti-Zadeh et al., 2021). In contrast, deep learning techniques have demonstrated robustness to such limitations, by learning hierarchical representations from large datasets, making it more adaptable to variations in motion blur (Zamir et al., 2021; Zhang et al., 2018), lighting (Silwal et al., 2021; Windrim et al., 2016) and partial occlusions (Zhu et al., 2019). To overcome limited availability of large datasets, deep learning uses transfer learning, in which knowledge from prior tasks and diverse datasets enhances learning for new or specialized tasks. This allows a model to focus on the specific task while leveraging other well-generalized models to handle common challenges such as illumination variations and blur (Iman et al., 2023). This approach is not possible in ML as they do not learn hierarchical representations of objects from images as DL does, making them less adaptable across tasks (Djolonga et al., 2021).

**5.2.2.2. Synthetic data generation.** Advancements in DL have provided solutions to overcome challenges pertaining to low resolution of image-based data, by enhancing video quality during real-time surgical procedures. Models like Generative Adversarial Networks (GAN) have potential to enhance image resolution through techniques like deblurring (Kupyn et al., 2019), and dehazing (Zheng et al., 2023). Transformers have shown potential in video restoration, also called video super resolution, through feature extraction followed by reconstruction of high-quality frames from low quality frames, based on temporal self-attention from multiple adjacent frames (Liang et al., 2024). Such image generation-based models have also shown potential in addressing scarcity of data, by generating reliable and accurate synthetic datasets including photo-realistic high resolution images (Ledig et al., 2017). These datasets have been validated to show high concordance to real surgical data (Azizi et al., 2021), suggesting that synthetic data can serve as effective substitutes and additionally address privacy concerns (detailed in section 5.2.5).

**5.2.2.3. Enhancing sensor-based data acquisition.** Challenges in accurate data acquisition pertaining to sensors have been addressed using DL techniques. Presence of metallic objects around EM sensors causes reflection, scattering and absorption of EM signals, distorting the received signals and reducing their tracking accuracy. By training deep neural networks on correctly labeled data, it has shown potential to capture the complex interactions within the scene and account for the effects of metallic interferences (Li et al., 2020). Similarly, deep learning has been applied to improve data acquisition for IMU sensors (Chen et al., 2018), and depth sensors (Wang and He, 2023; Yoneyama et al., 2021). These technological advancements highlight the potential to improve data acquisition in open surgeries. Deep learning-based neural network training and synthetic datasets can address the challenge of limited accurate data acquisition in open surgical settings.

### 5.2.3. Limitations of virtual reality trainers

Effective acquisition of hand motion and surgical instrument data is crucial, and immersive virtual reality (VR) simulators, which allow effortless data capture and feedback generation, may present a promising solution (Shabir et al., 2022b; Titov et al., 2023; Velazco Garcia et al., 2019). VR has been extensively applied for orthopedic procedures involving tools for drilling, sawing, or screw placement, as such movements can be simulated using haptic devices (Syamlan et al., 2022). However, replicating such a virtual environment with a realistic human-computer interface is challenging for open procedures using standard surgical instruments. Furthermore, intricate hand movements and deformations resulting from tissue manipulation in open surgeries may not be accurately represented (Seymour et al., 2006). Additionally, VR may only provide approximations of the tasks performed (Titov et al., 2023). Therefore, automated skills assessment for most open surgeries relies on reliable data capture during surgical training.

### 5.2.4. Non-technical skills

Most studies did not consider non-technical skills in automated assessment, which can potentially have a significant impact on intra-operative performance (Nagyné Elek and Haidegger, 2021). According to the Non-Technical Skills for Surgeons (NOTSS) rating tool, key elements assessed include situational awareness, decision-making, leadership, communication and teamwork (Abahuje et al., 2022). These elements all have an impact on the surgical care received by the patient. In fact, surgical errors do not occur exclusively due to lack of technical skills. Inadequate decision-making, situational awareness, and communication errors are also significant factors in the incidence of intra-operative adverse events. An earlier study revealed a substantial portion of errors occurring in the operating room attributed to surgeon's behavior and decision making (Gawande et al., 2003). Such insights question the reliability of traditional surgical skills assessment, and calls for a more comprehensive method that incorporates assessment of non-technical skills as well (Khan and Begum, 2021). Future developments focusing on integrating automated assessment of non-technical skills would be very useful (Nagyné Elek and Haidegger, 2022). This may be achieved through analysis of intraoperative video recordings (Dick et al., 2024). Expert surgeons may assess non-technical skills using the NOTSS tool by reviewing multiple camera angle views depicting discussion among the operating team. The video segments may be labeled for training algorithms based on the expert surgeon's assessment (Likosky et al., 2021). Alternatively, video data combined with force input data may also be utilized for automated non-technical skills assessment. In the study by (Nagyné Elek and Haidegger, 2022), sensory data was combined with Surgery Task Load Index (SURG-TLX) results as class labels. In addition to mental, physical, and temporal demands, the SURG-TLX also evaluates task complexity, situational stress and distractions. However, it can be prone to bias since they are self-reported results. Future research on automated non-technical skill assessment may incorporate objective measurements such as eye movement, heart rate, and electroencephalogram (EEG) recordings of the operating surgeon.

### 5.2.5. Ethical implications

Ethical implications must be taken into consideration while utilizing data acquired from live surgeries for automated skill assessment. Legal frameworks such as Health Insurance Portability and Accountability Act (HIPAA) in the United States and General Data Protection Regulation (GDPR) in the European Union provide guidelines that ensure protection of patient privacy (Walsh et al., 2023). Risk of data breaches can be reduced through anonymization (removal of patient identifiers), data encryption (such as hashing mechanism), and the use of firewall protection (Filicori and Addison, 2022; Godfrey et al., 2019). Prior approvals must be obtained from the relevant ethical boards before conducting research with surgical data. In the subsequent sections, we focus on the applications of the commonly used automation algorithms, namely, CNN, RNN, and transformers.

## 5.3. Common automation algorithms

### 5.3.1. Applications of convolutional neural networks

The automation algorithms used within the articles are diverse, with most utilizing AI for skill assessment and standardized evaluation metrics such as classification accuracy and mean-squared-error. Notably, the most frequently used algorithm was CNN. A CNN (LeCun et al., 2015) is a type of deep learning model specifically designed to process and analyze visual input through automatic recognition and learning of hierarchical patterns within images. It has been widely used in image recognition, classification, and computer vision tasks. Prominent CNN architectures include U-Net (Ronneberger et al., 2015), which is designed for image segmentation in the medical domain and uses an encoder-decoder structure with skip connections for high accuracy even with limited data, as demonstrated by (Baghdadi et al., 2023). ResNet

models (ResNet (Bkheet et al., 2023; Hira et al., 2022), ResNet-18 (Kasa et al., 2022), ResNet-50 (Kasa et al., 2022; Tang et al., 2024), ResNet-152 (Ruzicki et al., 2023)) utilize residual learning with shortcut connections to effectively train deep networks by overcoming challenges typically associated with very deep architectures (He et al., 2016). Although more than 20 % of the DL-based automations followed the ResNet architecture (Bkheet et al., 2023; Hira et al., 2022; Kasa et al., 2022; Ruzicki et al., 2023) for skill classification or score prediction using various assessment inputs, their performances were average (mean values of 63.3 % accuracy, 73.6 % AUC, 75 % specificity, 84.3 % sensitivity, 0.26 MSE, and 72 % ICC), indicating that this architecture may not be the best choice for automated assessment in open surgery. Inception-v4 (Szegedy et al., 2015) improves feature extraction by using scalable convolution filters rather than single-sized filters within the network. This approach allows for the capture of both fine and global features from images, balancing accuracy and computational efficiency. However, when combined with U-Net for analyzing time-series data on *forces from sensors on tool*, as demonstrated by (Baghdadi et al., 2023), the model yielded average performance (71 % F<sub>1</sub> score, 81 % AUC), falling short of other approaches in skill classification. Similarly, a variant of the Inception-v4 network, I3D (or Inflated 3D ConvNets), which extends scalability into the third dimension and is pre-trained for human action recognition, was used by (Hoffmann et al., 2024) for skill classification, yielding moderate results (71 % F<sub>1</sub> score and 72 % AUC). Having described the general applications of CNN, we now explore the utilization of two main aspects, namely, object detection and temporal information for skills assessment in the following subsections.

**5.3.1.1. Utilizing object detection for skill assessment.** Region-based Convolutional Neural Network (R-CNN) focuses on object detection in images by first generating region proposals using Selective Search and then classifying each region through a two-stage process. Faster R-CNN enhances this approach by performing the classification in a single stage, and Mask R-CNN builds on Faster R-CNN by predicting segmentation masks (He et al., 2017; Ren et al., 2015; Davids et al., 2021) used Mask R-CNN to track surgical tool behavior from *instrument motion data in 2D* captured from 2 cameras. The model demonstrated strong performance (97.7 % AUC, 84.21 % accuracy) in automated skill classification. However, the broader applicability of this deep learning network for skill assessment remains uncertain due to the limited amount of supporting research. The You Only Look Once (YOLO) architecture, renowned for real-time object detection, predicts bounding boxes and class probabilities in a single stage (Redmon et al., 2016). Its newer versions are recognized for delivering sub-millisecond speed and high performance. The works by (Bkheet et al., 2023; Goldbraikh et al., 2022) demonstrate reliable performance ( $p < 0.05$ ) in terms of speed and accuracy for skill classification based on *hand movement*. However, further research, particularly with more recent YOLO architectures like YOLOv7 and YOLOv8, is necessary to fully assess their potential for automated skill assessment in open surgery.

**5.3.1.2. Utilizing temporal information for skill assessment.** More recent networks used, like EfficientNet (Tan and Le, 2019), systematically scales network depth, width, and resolution to optimize performance with fewer parameters. Similarly, X3D (Feichtenhofer, 2020) expands a small 2D network along multiple axes, adding a temporal dimension, specifically targeting video classification with enhanced accuracy. However, the use of these networks is limited in number as well as performance in the current scope (only one work (Nagaraj et al., 2023) with average performance of 69 % F<sub>1</sub> score, 83 % accuracy) and requires further research to understand their full potential. Temporal Convolutional Networks (TCNs) (Bai et al., 2018) are effective for capturing information from a broader temporal context due to 1D convolutions that process sequences in parallel, allowing them to capture long-range dependencies efficiently without significantly increasing computational

load. The temporal segment network (TSN), extends this through segment-based aggregation, suitable for efficient action recognition tasks that was applied in combination with I3D, for skill classification by (Hoffmann et al., 2024).

### 5.3.2. Applications of recurrent neural networks

Recurrent Neural Networks (RNNs) are designed for sequential data processing by maintaining a hidden state that captures information from previous time steps, making them effective for tasks like time-series prediction. However, standard RNNs struggle with long-term dependencies, a challenge overcome by the Long Short-Term Memory (LSTM) network architecture, which uses mechanisms like gates to better capture and retain information over longer sequences (Hochreiter and Schmidhuber, 1997). Its application for automated skill assessment demonstrates good performance (98.2 % accuracy) when integrated with the CNN-based architecture proposed by (Nguyen et al., 2019). However, when combined with ResNet (Hira et al., 2022; Kim et al., 2019; Xu et al., 2023), the performance appears average (63.3 % accuracy, 69.2 % AUC, 0.26 MSE, and 72 % ICC). In contrast, TCNs, which similarly utilizes temporal information, demonstrate superior performance (91.2 % accuracy, 82.2 % AUC) (Bkheet et al., 2023; Nguyen et al., 2019), even when integrated with ResNet ( $p < 0.05$ ) (Kiyasseh et al., 2023; Matsoukas et al., 2021; Xie et al., 2021). Existing studies have primarily focused on hand movement for skill classification. However, TCNs hold the potential to be applied effectively to various other assessment inputs across different forms of skill assessment.

As evident from the above analyses, integrating CNN architectures with temporal feature extractors improves surgical skill assessment, as it provides insights on variation of surgeon actions and motions throughout consecutive time frames. When integrating temporal information, the appropriate choice of input sequence length (also called window size) is important. A small window size may enhance computational efficiency but may demonstrate reduced performance accuracy. However, for complex tasks like knot-tying, a higher window size could demonstrate improved performance (Wang and Majewicz Fey, 2018). Specific modifications to CNN architectures, like utilization of Scaled Exponential Linear Unit (SELU) instead of Rectified Linear Unit (ReLU) as the activation function within the neurons may benefit from self-normalization properties, providing the effect of batch normalization and sufficient regularization to maintain robust learning (Castro et al., 2019). Additionally, for multi-variate time-series like force sequence data and kinematics data (like position, velocity), the use of global mixed pooling strategy has shown to improve generalization capabilities for skill classification by taking advantage of both global average pooling and global max pooling strategies. The use of Adaptive Synthetic (ADASYN) sampling method for enhancing classification on minority classes in imbalanced datasets (Peng et al., 2024, 2025) and encoding predictable patterns in surgical motion using approximate entropy (ApEn) have shown to demonstrate good results in other domains (Zia and Essa, 2018), whose potential could be explored in open surgical skill assessment using CNNs.

### 5.3.3. Applications of transformers

More recent algorithms, such as vision-based transformers (ViT) have emerged to show results on par with CNNs when initialized with pre-trained information, and even outperform CNNs in tasks like object detection and semantic segmentation in the natural image domain (Kiyasseh et al., 2023; Matsoukas et al., 2021; Xie et al., 2021). Although they have been widely used for skill assessment in other surgical domains like MIS and robotic surgeries (Kiyasseh et al., 2023), only one work out of 40 in the current scope, employed a vision transformer (ViT), Microsoft's Shifted Windows (Swin) Transformer (Hoffmann et al., 2024; Liu et al., 2021), achieving moderate results on skill classification (71 % F<sub>1</sub> score). This highlights a gap in research utilizing such algorithms for the automation of skill assessment within open surgery. Having discussed the applications of commonly used automation



algorithms, the following section discusses some of the methods used for evaluating the algorithms, and the different assessment inputs utilized by the automation systems.

#### 5.4. Performance evaluation and assessment inputs

Evaluation of AI algorithms in the studies was done through standardized metrics pertaining to the task. Classification tasks utilized accuracy, F1 score, AUC, etc., while regression tasks utilized MSE or  $R^2$ . However, not all studies had the same list of metrics used for evaluation, i.e., some used a combination of accuracy and F1, while others included AUC, sensitivity and so on. This variability in use of metrics poses a challenge when it comes to comparison of algorithm performance between studies. AUC of a model describes how well the algorithm can distinguish between two classes. In skill classification, this metric can be considered more useful than other metrics due to its direct distinguishing capability between a novice and an expert. However, it is noteworthy that for the skill classification task, 15 studies used accuracy as their metric to evaluate performance, while only 4 studies made use of AUC. Although measuring performance in terms of accuracy helps understand the algorithm's overall predictive ability, it overlooks the algorithm's distinguishing ability – which is particularly important for classification algorithms. For example, consider two scenarios A and B: estimated probabilities of 0.51 and 0.99 are both classified as Expert. If the ground truths were Novice and Expert respectively, then the accuracy metric has evaluated the model performance as 50 %. In contrast, AUC measures the probability that a randomly chosen negative example will have a smaller estimated probability of belonging to the positive class than a randomly chosen positive example (Hanley and McNeil, 1982). This means that AUC evaluates the algorithm based on its prediction probability directly, rather than post-thresholding result. Since the estimated probability of scenario A (randomly selected negative sample Novice) is 0.51 and it is lesser than scenario B (randomly selected positive sample), AUC would help evaluate the algorithm as 100 % estimation of performance. Hence, AUC is generally a better estimate of classification performance than accuracy (Huang and Ling, 2005), and is recommended to use in combination with accuracy to provide a more complete performance evaluation of AI algorithms for skill classification.

Beyond learning-based algorithms, custom algorithms were also prevalent among the discussed articles. While AI algorithms often adhere to standardized evaluation metrics for performance comparison, custom algorithms are frequently evaluated using varied or non-standardized methods. For instance, some studies employed statistical analyses, such as *t*-tests and correlation coefficients, to assess the algorithm's correlation to OSATS scoring (Yamada et al., 2022). Others relied on expert evaluations, such as a study where seven experts rated the overall reliability of the automation system with a score of 9/10 (Tozzi et al., 2022). However, the reliability of such variable and subjective evaluations raises concerns, particularly given the variability in the quantity and quality of raters. This variability poses challenges when translating these automation algorithms into clinical settings. For such algorithms, the correlation to well-established scoring benchmarks like OSATS or Procedure-Based Assessment (PBA) demonstrates highest reliability (Beard et al., 2011). To ensure a dependable and comparable evaluation of such algorithms, reliability and validity are crucial factors to consider. The MERSQI score includes an aspect that is dedicated to the validity of the evaluation instrument used in a study. A higher score would indicate that the study has reported evidence of content and reliability of the instrument used. In addition, to provide a reliability estimate, every relevant factor should be sampled as widely and representatively as possible, based on generalizability theory (Andersen et al., 2021). This includes the number of evaluations of the algorithm, number of assessors, assessor designations, types of surgical cases assessed, and most importantly, the inter-rater reliability between the assessors. The assessment algorithm should also undergo validity evaluation through

predictive validity, internal content validity, and construct validity. Predictive validity measures the correlation between the prediction/outcome (i.e., whether the trainee successfully completed the procedure or demonstrated proficiency in technical skills), and the scores estimated by the algorithm. Internal content validity evaluates the correlation between each criterion within the assessment algorithm; and construct validity explores the correlation of these scores with age, experience and demographics of the assessors. Additionally, a user satisfaction and acceptability assessment by the assessors is recommended to understand the usefulness of the algorithm for providing feedback, for its summative purpose, and its importance in surgical education. To obtain a comparative analysis between algorithms, it is recommended to use pooled reliability tests based on Generalizability-coefficients (or G-coefficients) (Mitchell, 1979), and conduct concurrent validity that estimates the correlation between tools, i.e., tools measuring the same construct should have high correlation estimates; and finally, inter-procedural differences should be accounted for, i.e., algorithms designed for different surgical procedures may not be comparable (Beard et al., 2011).

As observed in Fig. 4c, *hand movement* was the most frequently used assessment input across all algorithms, followed by *modified OSATS scoring*, *knot/suture placement*, and *instrument motion in 3D*, in decreasing order. When *hand movement* was used for skill classification, 77 % of the automation algorithms showed good results (i.e., significantly high performance with  $p < 0.05$  (Bkheet et al., 2023; Goldbraikh et al., 2022; Rittenhouse et al., 2014; Shaharan et al., 2017) or above 90 % accuracy (Nguyen et al., 2019; Sbernini et al., 2018; Sun et al., 2016; Xu et al., 2023; Zia et al., 2018)). Notably, half of these successful applications involved DL algorithms, and the rest constituted the other types of algorithms. This indicates that using DL to assess a surgeon's skills based on hand motion or manual expertise has yielded the most promising results compared to other combinations of assessment inputs and algorithms.

The use of *3D instrument motion data* has also produced notable results for skill classification. All studies using this data showed good results in skill classification ( $p < 0.05$  (Franco-González et al., 2021) or above 90 % accuracy (Ahmadi et al., 2015; Yibulayimu et al., 2022; Zia et al., 2018)), except for two (Hira et al., 2022; Kim et al., 2019), both of which involved DL algorithms. This suggests that using DL on *3D instrument motion data* may not be the most effective approach for skill classification and warrants further research. Additionally, it is worth noting that existing works have used *3D instrument motion data* exclusively for skill classification; no studies have applied this data for generating skill assessment scores (like OSATS).

It is also important to highlight that only 10 out of 40 studies focused on automated skill assessment using scores. Out of these, 3 studies utilized ML techniques (Azari et al., 2019, 2021a, 2021b) and 1 utilized DL techniques for assessing surgeon skills (Kasa et al., 2022). There is a lack of studies focusing on direct output of skill assessment through scores, in comparison to classification, indicating a significant research gap in scoring-based skill assessment. The main reason for this gap is the unavailability of ground truth. Acquisition of detailed ground truth information like OSATS scores is difficult as this requires a minimum number of expert evaluators assessing in-person or video recordings of the user performance (Martin et al., 1997). There is a lack of public datasets for such ground truth, and there are usually restrictions on making any such collected data public (Yanik et al., 2022). For example, according to GDPR, videos recorded for the purpose of performance metric scoring (like OSATS) should be deleted right after scoring (Filicori and Addison, 2022). This provides scoring datasets, however with no corresponding source videos to train on. Comparatively, classification ground-truth is easy to acquire from self-proclaimed skill levels based on hours of experience (Wang and Majewicz Fey, 2018).

### 5.5. Limitations of the review

One limitation of this review is that the search strategy did not include databases specifically for grey literature (such as dissertations and unpublished clinical trials). Although an effort was made to be as comprehensive as possible with the search strategy used, articles describing automated skill assessment systems for open surgery that did not use certain terminologies might have been overlooked. Nonetheless, this scoping review presents an overall picture of the field of automated skills assessment during open surgery, which is still in its early stages.

## 6. Conclusion

Overall, automation of skills assessment during open surgery is making significant progress. The comprehensive analysis of 40 articles indicated a steady evolution of sensors, including wearable devices, for acquiring data from the open surgical field. Various automation algorithms have been applied, resulting in high accuracies for predicting the skill levels of participants. Improvements in methods for data acquisition from the surgical field and the adoption of standardized surgical skills assessments will be crucial for facilitating the clinical translation of this technology.

Despite these advancements, challenges remain, particularly in the reliable capture of data during live open surgeries. Although 65 % of the studies utilized video cameras, setbacks such as variability in camera setups and difficulties in obtaining high-quality footage and accurate measurements remain. This highlights the need for further refinement in data acquisition methods. As such, operating rooms must be equipped adequately for the collection and storage of intraoperative data during open surgeries. Additionally, about 45 % of the studies focused on suturing which was not specific to a surgical procedure. This indicates an underrepresentation of complex surgical procedures and limited focus on comprehensive skill assessment. Future research must be directed towards advanced surgical procedures. In addition, automated assessment of non-technical skills must also be taken into consideration. To facilitate these developments, efforts should be devoted to building annotated open surgery research datasets.

While DL algorithms, especially CNNs, have shown promise in skill classification based on hand and instrument motions, there is still a significant gap in research exploring their application for scoring systems like OSATS. Only 10 % of the studies included in the review had OSATS as a predicted score. To improve reliability and validity of automated assessment, further development on OSATS score predictions is needed. Moreover, the potential of newer architectures, such as vision-based transformers and TCNs, remains underexplored in the context of open surgery, suggesting that future work should investigate these avenues to fully leverage their capabilities.

In general, there is an optimistic outlook for further research and development in the field of automated skills assessment in open surgery. Continued innovation in sensor technology, algorithm development, and data acquisition methods will be key to overcoming existing challenges and advancing the clinical implementation of these systems, ultimately enhancing the training and performance of surgeons.

### CRedit authorship contribution statement

**Hawa Hamza:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation. **Dehlela Shabir:** Writing – review & editing, Writing – original draft, Formal analysis, Data curation. **Omar Aboumarzouk:** Supervision, Resources, Project administration. **Abdulla Al-Ansari:** Supervision, Resources, Project administration. **Khaled Shaban:** Writing – review & editing, Supervision. **Nikhil V. Navkar:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Formal analysis, Conceptualization.

## Funding

Research reported in this publication was supported by the Qatar Research, Development and Innovation (QRDI) Council Academic Research Grant (ARG) awards ARG02-0315-240013 and ARG01-0430-230047, and Graduate Sponsorship Research Award GSRA11-L-1-0319-24006. All opinions, findings, conclusions, or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of our sponsors. Open access funding was provided by the Qatar National Library.

## Declaration of competing interest

The authors of this submission Hawa Hamza, Dehlela Shabir, Omar Aboumarzouk, Abdulla Al-Ansari, Khaled Shaban, and Nikhil V. Navkar have no conflict of interest or financial ties to disclose.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.engappai.2025.110893>.

## Data availability

No data was used for the research described in the article.

## References

- Abahuje, E., Johnson, J., Halverson, A., Stulberg, J.J., 2022. Intraoperative assessment of non-technical skills for surgeons (NOTSS) and qualitative description of their effects on intraoperative performance. *J. Surg. Educ.* 79, 1237–1245.
- Abdurahiman, N., Padhan, J., Zhao, H., Balakrishnan, S., Al-Ansari, A., Abinayed, J., Velasquez, C.A., Becker, A.T., Navkar, N.V., 2022. Human-computer interfacing for control of angulated scopes in robotic scope assistant systems. 2022 International Symposium on Medical Robotics (ISMIR). IEEE, pp. 1–7.
- Ahmidi, N., Poddar, P., Jones, J.D., Vedula, S.S., Ishii, L., Hager, G.D., Ishii, M., 2015. Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty. *Int. J. Comput. Assist. Radiol. Surg.* 10, 981–991.
- Andersen, S.A.W., Nayahangan, L.J., Park, Y.S., Konge, L., 2021. Use of generalizability theory for exploring reliability of and sources of variance in assessment of technical skills: a systematic review and meta-analysis. *Acad. Med.* 96, 1609–1619.
- Azari, D.P., Frasier, L.L., Miller, B.L., Pavuluri Quamme, S.R., Le, B.V., Greenberg, C.C., Radwin, R.G., 2021a. Modeling performance of open surgical cases. *Simulat. Healthc. J. Soc. Med. Simulat.* 16, e188–e193.
- Azari, D.P., Frasier, L.L., Quamme, S.R.P., Greenberg, C.C., Pugh, C.M., Greenberg, J.A., Radwin, R.G., 2019. Modeling surgical technical skill using expert assessment for automated computer rating. *Ann. Surg.* 269, 574–581.
- Azari, D.P., Miller, B.L., Le, B.V., Greenberg, J.A., Bruskewitz, R.C., Long, K.L., Chen, G., Radwin, R.G., 2021b. A comparison of expert ratings and marker-less hand tracking along OSATS-derived motion scales. *IEEE Transactions on Human-Machine Systems* 51, 22–31.
- Azizi, Z., Zheng, C., Mosquera, L., Pilote, L., El Emam, K., 2021. Can synthetic data be a proxy for real clinical trial data? A validation study. *BMJ Open* 11, e043497.
- Baghdadi, A., Lama, S., Singh, R., Sutherland, G.R., 2023. Tool-tissue force segmentation and pattern recognition for evaluating neurosurgical performance. *Sci. Rep.* 13, 9591.
- Bai, S., Kolter, J.Z., Koltun, V., 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.
- Beard, J.D., Marriott, J., Purdie, H., Crossley, J., 2011. Assessing the surgical skills of trainees in the operating theatre: a prospective observational study of the methodology. *Clin. Govern. Int. J.* 16.
- Bell, R.H., 2009. Why Johnny cannot operate. *Surgery* 146, 533–542.
- Bergmeister, K.D., Aman, M., Kramer, A., Schenck, T.L., Riedl, O., Daeschler, S.C., Aszmann, O.C., Bergmeister, H., Golriz, M., Mehrabi, A., Hundeshagen, G., Enkhbaatar, P., Kinsky, M.P., Podesser, B.K., 2020. Simulating surgical skills in animals: systematic review, costs & acceptance analyses. *Front. Vet. Sci.* 7.
- Bilgic, E., Gorgy, A., Yang, A., Cwintal, M., Ranjbar, H., Kahla, K., Reddy, D., Li, K., Ozturk, H., Zimmermann, E., Quaiattini, A., Abbasgholizadeh-Rahimi, S., Poenaru, D., Harley, J.M., 2022. Exploring the roles of artificial intelligence in surgical education: a scoping review. *Am. J. Surg.* 224, 205–216.
- Birkmeyer, J.D., Finks, J.F., O'Reilly, A., Oerline, M., Carlin, A.M., Nunn, A.R., Dimick, J., Banerjee, M., Birkmeyer, N.J.O., 2013. Surgical skill and complication rates after bariatric surgery. *N. Engl. J. Med.* 369, 1434–1442.
- Bkheet, E., D'Angelo, A.-L., Goldbraikh, A., Laufer, S., 2023. Using hand pose estimation to automate open surgery training feedback. *Int. J. Comput. Assist. Radiol. Surg.* 18, 1279–1285.

- Castillo-Segura, P., Fernández-Panadero, C., Iario-Hoyos, C., Muñoz-Merino, P., Delgado Kloos, C., 2021. Objective and automated assessment of surgical technical skills with IoT systems: a systematic literature review. *Artif. Intell. Med.* 112.
- Castro, D., Pereira, D., Zanchettin, C., Macêdo, D., Bezerra, B.L., 2019. Towards optimizing convolutional neural networks for robotic surgery skill evaluation. 2019 International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 1–8.
- Chen, H., Aggarwal, P., Taha, T.M., Chodavarapu, V.P., 2018. Improving inertial sensor by reducing errors using deep learning methodology. *NAECON 2018 - IEEE National Aerospace and Electronics Conference*, pp. 197–202.
- Cook, D.A., Reed, D.A., 2015. Appraising the quality of medical education research methods: the medical education research study quality instrument and the newcastle-Ottawa scale-education. *Acad. Med.* 90.
- Datta, V., Mackay, S., Mandalia, M., Darzi, A., 2001. The use of electromagnetic motion tracking analysis to objectively measure open surgical skill in the laboratory-based model 1No competing interests declared. *J. Am. Coll. Surg.* 193, 479–485.
- Davids, J., Makariou, S.-G., Ashrafi, H., Darzi, A., Marcus, H.J., Giannarou, S., 2021. Automated vision-based microsurgical skill analysis in neurosurgery using deep learning: development and preclinical validation. *World Neurosurg.* 149, e669–e686.
- DeMasi, S., Katsuta, E., Takabe, K., 2016. Live Animals for Preclinical Medical Student Surgical Training.
- Deng, T., Gulati, S., Rodriguez, W., Dawant, B.M., Langerman, A., 2021. Automated detection of electrocautery instrument in videos of open neck procedures using YOLOv3. 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 2071–2074.
- Dick, L., Boyle, C.P., Skipworth, R.J.E., Smink, D.S., Tallentire, V.R., Yule, S., 2024. Automated analysis of operative video in surgical training: scoping review. *BJS Open* 8, zrae124.
- Djlonga, J., Yung, J., Tschannen, M., Romijnders, R., Beyer, L., Kolesnikov, A., Puigcerver, J., Minderer, M., D'Amour, A., Moldovan, D., 2021. On Robustness and Transferability of Convolutional Neural Networks, pp. 16458–16468.
- Elsej, E.J., Griffiths, G., Humes, D.J., West, J., 2017. Meta-analysis of operative experiences of general surgery trainees during training. *Journal of British Surgery* 104, 22–33.
- Fattahi Sani, M., Ascione, R., Dogramadzi, S., 2021. Mapping surgeons hand/finger movements to surgical tool motion during conventional microsurgery using machine learning. *Journal of Medical Robotics Research* 6, 2150004.
- Feichtenhofer, C., 2020. X3d: expanding architectures for efficient video recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 203–213.
- Feifer, A., Al-Ammari, A., Kovac, E., Delisle, J., Carrier, S., Anidjar, M., 2011. Randomized controlled trial of virtual reality and hybrid simulation for robotic surgical training. *BJU Int.* 108, 1652–1656.
- Filicori, F., Addison, P., 2022. Intellectual property and data ownership in the age of video recording in the operating room. *Surg. Endosc.* 36, 3772–3774.
- Franco-González, I.T., Pérez-Escamirosa, F., Minor-Martínez, A., Rosas-Barrientos, J.V., Hernández-Paredes, T.J., Franco-González, I.T., Pérez-Escamirosa, F., Minor-Martínez, A., Rosas-Barrientos, J.V., Hernández-Paredes, T.J., 2021. Development of a 3D motion tracking system for the analysis of skills in microsurgery. *J. Med. Syst.* 45.
- Frischknecht, A.C., Kasten, S.J., Hamstra, S.J., Perkins, N.C., Gillespie, R.B., Armstrong, T.J., Minter, R.M., 2013. The objective assessment of experts' and novices' suturing skills using an image analysis program. *Acad. Med.* 88.
- Garrow, C.R., Kowalewski, K.-F., Li, L., Wagner, M., Schmidt, M.W., Engelhardt, S., Hashimoto, D.A., Kenngott, G.B., Bodenstedt, S., Speidel, S., Müller-Stich, B.P., Nickel, F., 2021. Machine learning for surgical phase recognition: a systematic review. *Ann. Surg.* 273.
- Gawande, A.A., Zinner, M.J., Studdert, D.M., Brennan, T.A., 2003. Analysis of errors reported by surgeons at three teaching hospitals. *Surgery* 133, 614–621.
- Geda, M.W., Tang, Y.M., Lee, C.K.M., 2024. Applications of artificial intelligence in Orthopaedic surgery: a systematic review and meta-analysis. *Eng. Appl. Artif. Intell.* 133, 108326.
- Glossop, S.C., Bhachoo, H., Murray, T.M., Cherif, R.A., Helo, J.Y., Morgan, E., Poacher, A.T., 2023. Undergraduate teaching of surgical skills in the UK: systematic review. *BJS Open* 7.
- Godfrey, M., Walle, K.V., Rosser, A.A., Quamme, S.P., Greenberg, C., Greenberg, J.A., Jung, S.J., 2019. Overcoming hurdles to video recording in the operating room for surgical education. *J. Am. Coll. Surg.* 229, e190.
- Goldbraikh, A., D'Angelo, A.-L., Pugh, C.M., Laufer, S., 2022. Video-based fully automatic assessment of open surgery suturing skills. *Int. J. Comput. Assist. Radiol. Surg.* 17, 437–448.
- Goodman, E.D., Patel, K.K., Zhang, Y., Locke, W., Kennedy, C.J., Mehrotra, R., Ren, S., Guan, M., Zohar, O., Downing, M., Chen, H.W., Clark, J.Z., Berrigan, M.T., Brat, G. A., Yeung-Levy, S., 2024. Analyzing surgical technique in diverse open surgical videos with multitask machine learning. *JAMA Surgery* 159, 185–192.
- Guerin, S., Huauclmé, A., Lavoue, V., Jannin, P., Timoh, K.N., 2022. Review of automated performance metrics to assess surgical technical skills in robot-assisted laparoscopy. *Surg. Endosc.* 36, 853–870.
- Hamza, H., Baez, V.M., Al-Ansari, A., Becker, A.T., Navkar, N.V., 2023. User interfaces for actuated scope maneuvering in surgical systems: a scoping review. *Surg. Endosc.* 37, 4193–4223.
- Handelman, A., Keshet, Y., Livny, E., Barkan, R., Nahum, Y., Tepper, R., 2020. Evaluation of suturing performance in general surgery and ocular microsurgery by combining computer vision-based software and distributed fiber optic strain sensors: a proof-of-concept. *Int. J. Comput. Assist. Radiol. Surg.* 15, 1359–1367.
- Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 29–36.
- He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778.
- Hira, S., Singh, D., Kim, T.S., Gupta, S., Hager, G., Sikder, S., Vedula, S.S., 2022. Video-based assessment of intraoperative surgical skill. *Int. J. Comput. Assist. Radiol. Surg.* 17, 1801–1811.
- Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. *Neural Comput.* 9, 1735–1780.
- Hoffmann, H., Funke, I., Peters, P., Venkatesh, D.K., Egger, J., Rivoir, D., Röhrig, R., Hölzle, F., Bodenstedt, S., Willemer, M.-C., Speidel, S., Puladi, B., 2024. AIXSuture: vision-based assessment of open suturing skills. *Int. J. Comput. Assist. Radiol. Surg.* 19, 1045–1052.
- Huang, J., Ling, C.X., 2005. Using AUC and accuracy in evaluating learning algorithms. *IEEE Trans. Knowl. Data Eng.* 17, 299–310.
- Huffman, E., Anton, N., Martin, J., Timsina, L., Dearing, W., Breece, B., Mann, I., Stefanidis, D., 2020. Optimizing assessment of surgical knot tying skill. *J. Surg. Educ.* 77, 1577–1582.
- Iman, M., Arabnia, H.R., Rasheed, K., 2023. A review of deep transfer learning and recent advancements. *Technologies* 11, 40.
- Irfan, W., Sheahan, C., Mitchell, E.L., Sheahan, M.G., 2019. The pathway to a national vascular skills examination and the role of simulation-based training in an increasingly complex specialty. *Semin. Vasc. Surg.* 32, 48–67.
- Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., Muller, P.-A., 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.
- Jaffer, A., Bednarz, B., Challacombe, B., Sriprasad, S., 2009. The assessment of surgical competency in the UK. *Int. J. Surg.* 7.
- Jardine, D., Hoagland, B., Perez, A., Gessler, E., 2015. Evaluation of surgical dexterity during the interview day: another factor for consideration. *J. Grad. Med. Educ.* 7, 234–237.
- Kadkhodamohammadi, A., Gangi, A., de Mathelin, M., Padoy, N., 2017. A multi-view RGB-D approach for human pose estimation in operating rooms. 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 363–372.
- Kajita, H., Takatsume, Y., Shimizu, T., Saito, H., Kishi, K., 2020. Overhead multiview camera system for recording open surgery. *Plast. Reconstr. Surg. Glob. Open* 8.
- Kasa, K., Burns, D., Goldenberg, M.G., Selim, O., Whyne, C., Hardisty, M., 2022. Multimodal deep learning for assessing surgeon technical skill. *Sensors*.
- Kawka, M., Gall, T.M.H., Fang, C., Liu, R., Jiao, L.R., 2022. Intraoperative video analysis and machine learning models will change the future of surgical training. *Intelligent Surgery* 1, 13–15.
- Khan, M.R., Begum, S., 2021. Apprenticeship to simulation - the metamorphosis of surgical training. *J. Pakistan Med. Assoc.* 71 (Suppl. 1), S72–S76.
- Khorasani, M., Abdurahman, N., Padhan, J., Zhao, H., Al-Ansari, A., Becker, A.T., Navkar, N., 2023. Preliminary design and evaluation of a generic surgical scope adapter. *Int. J. Med. Robot.* 19, e2475.
- Kil, I., Eidt, J.F., Singapogu, R.B., Groff, R.E., 2024. Assessment of open surgery suturing skill: image-based metrics using computer vision. *J. Surg. Educ.* 81, 983–993.
- Kim, T.S., O'Brien, M., Zafar, S., Hager, G.D., Sikder, S., Vedula, S.S., 2019. Objective assessment of intraoperative technical skill in capsulorhexis using videos of cataract surgery. *Int. J. Comput. Assist. Radiol. Surg.* 14, 1097–1105.
- Kirubakaran, A., Young, D., Khan, S., Crasto, N., Sobel, M., Sussman, D., 2022. Artificial intelligence and surgical education: a systematic scoping review of interventions. *J. Surg. Educ.* 79, 500–515.
- Kiyasseh, D., Ma, R., Haque, T.F., Miles, B.J., Wagner, C., Donoho, D.A., Anandkumar, A., Hung, A.J., 2023. A vision transformer for decoding surgeon activity from surgical videos. *Nat. Biomed. Eng.* 7, 780–796.
- Kupyn, O., Martyniuk, T., Wu, J., Wang, Z., 2019. Deblurgan-v2: Deblurring (Orders-of-magnitude) Faster and Better, pp. 8878–8887.
- Lam, K., Chen, J., Wang, Z., Iqbal, F.M., Darzi, A., Lo, B., Purkayastha, S., Kinross, J.M., 2022. Machine learning for technical skill assessment in surgery: a systematic review. *npj Digit. Med.* 5.
- Lavanchy, J.L., 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.
- LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. *Nature* 521, 436–444.
- Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., 2017. Photo-realistic Single Image Super-resolution Using a Generative Adversarial Network, pp. 4681–4690.
- Levin, M., McKechnie, T., Khalid, S., Grantcharov, T.P., Goldenberg, M., 2019. Automated methods of technical skill assessment in surgery: a systematic review. *J. Surg. Educ.* 76, 1629–1639.
- Li, H.-Y., Zhao, H.-T., Wei, M.-L., Ruan, H.-X., Shuang, Y., Cui, T.J., Del Hougne, P., Li, L., 2020. Intelligent electromagnetic sensing with learnable data acquisition and processing. *Patterns* 1.
- Liang, J., Cao, J., Fan, Y., Zhang, K., Ranjan, R., Li, Y., Timofte, R., Van Gool, L., 2024. Vrt: a video restoration transformer. *IEEE Trans. Image Process.*
- Likosky, D., Yule, S.J., Mathis, M.R., Dias, R.D., Corso, J.J., Zhang, M., Krein, S.L., Caldwell, M.D., Louis, N., Janda, A.M., Shah, N.J., Pagani, F.D., Stakich-Alpirez, K., Manojlovich, M.M., 2021. Novel assessments of technical and nontechnical cardiac surgery quality: protocol for a mixed methods study. *JMIR Res Protoc* 10, e22536.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: hierarchical vision transformer using shifted windows. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9992–10002.



- Maier-Hein, L., Eisenmann, M., Sarikaya, D., März, K., Collins, T., Malpani, A., Fallert, J., Feussner, H., Giannarou, S., Mascagni, P., Nakawala, H., Park, A., Pugh, C., Stoyanov, D., Vedula, S.S., Cleary, K., Fichtinger, G., Forestier, G., Gibaud, B., Grantcharov, T., Hashizume, M., Heckmann-Nötzel, D., Kennigott, H.G., Kikinis, R., Mündermann, L., Navab, N., Onogur, S., Roß, T., Sznitman, R., Taylor, R.H., Tizabi, M.D., Wagner, M., Hager, G.D., Neumuth, T., Padoy, N., Collins, J., Gockel, I., Goedeke, J., Hashimoto, D.A., Joyeux, L., Lam, K., Leff, D.R., Madani, A., Marcus, H. J., Meireles, O., Seitel, A., Teber, D., Ückert, F., Müller-Stich, B.P., Jannin, P., Speidel, S., 2022. Surgical data science – from concepts toward clinical translation. *Med. Image Anal.* 76, 102306.
- Martin, J., Regehr, G., Reznick, R., Macrae, H., Murnaghan, J., Hutchison, C., Brown, M., 1997. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br. J. Surg.* 84, 273–278.
- Marvasti-Zadeh, S.M., Cheng, L., Ghanei-Yakhdan, H., Kasaei, S., 2021. Deep learning for visual tracking: a comprehensive survey. *IEEE Trans. Intell. Transport. Syst.* 23, 3943–3968.
- Matsoukas, C., Haslum, J.F., Söderberg, M., Smith, K., 2021. Is it time to replace CNNs with transformers for medical images? ICCV 2021 Workshop on Computer Vision for Automated Medical Diagnosis (CVAMD).
- Melton, G.B., 2010. Biomedical and Health informatics for surgery. *Adv. Surg.* 44, 117–130.
- Mitchell, S.K., 1979. Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychol. Bull.* 86, 376–390.
- Nagaraj, M.B., Namazi, B., Sankaranarayanan, G., Scott, D.J., 2023. Developing artificial intelligence models for medical student suturing and knot-tying video-based assessment and coaching. *Surg. Endosc.* 37, 402–411.
- Nagyné Elek, R., Haidegger, T., 2021. Non-technical skill assessment and mental load evaluation in robot-assisted minimally invasive surgery. *Sensors* 21, 2666.
- Nagyné Elek, R., Haidegger, T., 2022. Next in surgical data science: autonomous non-technical skill assessment in minimally invasive surgery training. *J. Clin. Med.* 11, 7533.
- Nguyen, X.A., Ljuhar, D., Pacilli, M., Nataraja, R.M., Chauhan, S., 2019. Surgical skill levels: classification and analysis using deep neural network model and motion signals. *Comput. Methods Progr. Biomed.* 177, 1–8.
- Oliveira, M.M., Quittes, L., Costa, P.H.V., Ramos, T.M., Rodrigues, A.C.F., Nicolato, A., Malheiros, J.A., Machado, C., 2022. Computer vision coaching microsurgical laboratory training: PRIME (Proficiency Index in Microsurgical Education) proof of concept. *Neurosurg. Rev.* 45, 1601–1606.
- Pakkasjärvi, N., Anttila, H., Pyhälä, K., 2024. What are the learning objectives in surgical training – a systematic literature review of the surgical competence framework. *BMC Med. Educ.* 24.
- Pangal, D.J., Kugener, G., Zhu, Y., Sinha, A., Unadkat, V., Cote, D.J., Strickland, B., Rutkowski, M., Hung, A., Anandkumar, A., Han, X.Y., Pappan, V., Wrobel, B., Zada, G., Donoho, D.A., 2022. Expert surgeons and deep learning models can predict the outcome of surgical hemorrhage from 1 min of video. *Sci. Rep.* 12.
- Peng, Y., Wang, Y., Hu, F., He, M., Mao, Z., Huang, X., Ding, J., 2024. Predictive modeling of flexible EHD pumps using Kolmogorov–Arnold Networks. *Biomimetic Intelligence and Robotics* 4, 100184.
- Peng, Y., Yang, X., Li, D., Ma, Z., Liu, Z., Bai, X., Mao, Z., 2025. Predicting flow status of a flexible rectifier using cognitive computing. *Expert Syst. Appl.* 264, 125878.
- Pérez-Escamirós, F., Montoya-Alvarez, S., Ordoica-Flores, R.M., Padilla-Sánchez, L., Jiménez-Corona, J.L., Ruiz-Lizarraga, J., Minor-Martínez, A., Pérez-Escamirós, F., Montoya-Alvarez, S., Ordoica-Flores, R.M., Padilla-Sánchez, L., Jiménez-Corona, J. L., Ruiz-Lizarraga, J., Minor-Martínez, A., 2020. Design of a dynamic force measurement system for training and evaluation of suture surgical skills. *J. Med. Syst.* 44.
- Poursartip, B., LeBel, M.-E., McCracken, L.C., Escoto, A., Patel, R.V., Naish, M.D., Trejos, A.L., 2017. Energy-based metrics for arthroscopic skills assessment. *Sensors*.
- Raeker-Jordan, E., Martinez, M., Shimada, K., 2022. 3D printing of customizable phantoms to replace cadaveric models in upper extremity surgical residency training. *Materials* 15, 694.
- Rafiq, A., Moore, J.A., Zhao, X., Doarn, C.R., Merrell, R.C., 2004. Digital video capture and synchronous consultation in open surgery. *Ann. Surg.* 239.
- Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788.
- Reed, D.A., Cook, D.A., Beckman, T.J., Levine, R.B., Kern, D.E., Wright, S.M., 2007. Association between funding and quality of published medical education research. *JAMA* 298, 1002.
- Ren, H., Kazanzides, P., 2009. Hybrid attitude estimation for laparoscopic surgical tools: a preliminary study. 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, pp. 5583–5586.
- Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* 28.
- Rittenhouse, N., Sharma, B., Sonnadara, R., Mihailidis, A., Grantcharov, T., 2014. Design and validation of an assessment tool for open surgical procedures. *Surg. Endosc.* 28, 918–924.
- Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation, Medical image computing and computer-assisted intervention. MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18. Springer, pp. 234–241.
- Ruan, Y., Robinson, N.B., Khan, F.M., Hameed, I., Rahouma, M., Naik, A., Oakley, C.T., Rong, L., Girardi, L.N., Gaudino, M., 2020. The translation of surgical animal models to human clinical research: a cross-sectional study. *Int. J. Surg.* 77.
- Ruzicki, J., Holden, M., Cheon, S., Ungi, T., Egan, R., Law, C., 2023. Use of machine learning to assess cataract surgery skill level with tool detection. *Ophthalmol. Sci.* 3, 100235.
- Sbernini, L., Quitadamo, L.R., Riillo, F., Lorenzo, N.D., Gaspari, A.L., Saggio, G., 2018. Sensory-glove-based open surgery skill evaluation. *IEEE Transactions on Human-Machine Systems* 48, 213–218.
- Seymour, N.E., Rotnes, J.S., Seymour, N.E., Røtnes, J.S., 2006. Challenges to the development of complex virtual reality surgical simulations. *Surg. Endosc.* 20.
- Shabir, D., Abdurahiman, N., Padhan, J., Anbatawi, M., Trinh, M., Balakrishnan, S., Al-Ansari, A., Yaacoub, E., Deng, Z., Erbad, A., Mohammed, A., Navkar, N.V., 2022a. Preliminary design and evaluation of a remote tele-mentoring system for minimally invasive surgery. *Surg. Endosc.* 36, 3663–3674.
- Shabir, D., Abdurahiman, N., Padhan, J., Trinh, M., Balakrishnan, S., Kurer, M., Ali, O., Al-Ansari, A., Yaacoub, E., Deng, Z., Erbad, A., Mohammed, A., Navkar, N.V., 2021. Towards development of a tele-mentoring framework for minimally invasive surgeries. *Int. J. Med. Robot.* 17, e2305.
- Shabir, D., Anbatawi, M., Padhan, J., Balakrishnan, S., Al-Ansari, A., Abinshed, J., Tsiamyrtzis, P., Yaacoub, E., Mohammed, A., Deng, Z., Navkar, N.V., 2022b. Evaluation of user-interfaces for controlling movements of virtual minimally invasive surgical instruments. *Int. J. Med. Robot.* 18, e2414.
- Shaharan, S., Nugent, E., Ryan, D.M., Traynor, O., Neary, P., Buckley, D., 2016. Basic surgical skill retention: can patriot motion tracking system provide an objective measurement for it? *J. Surg. Educ.* 73, 245–249.
- Shaharan, S., Ryan, D.M., Neary, P.C., 2017. Motion tracking system in surgical training. *Motion Tracking and Gesture Recognition*. InTech.
- Shayan, A.M., Singh, S., Gao, J., Groff, R.E., Bible, J., Eidt, J.F., Sheahan, M., Gandhi, S. S., Blas, J.V., Singapogu, R., 2023. Measuring hand movement for suturing skill assessment: a simulation-based study. *Surgery* 174, 1184–1192.
- Silwal, A., Parhar, T., Yandun, F., Baweja, H., Kantor, G., 2021. A robust illumination-invariant camera system for agricultural applications. *IEEE ASME J. Microelectromech. Syst.* 3292–3298.
- Singh, S.P., Shayan, A.M., Gao, J., Bible, J., Groff, R.E., Singapogu, R., 2024. Objective and automated quantification of instrument handling for open surgical suturing skill assessment: a simulation-based study. *IEEE Open Journal of Engineering in Medicine and Biology* 5, 485–493.
- Soangra, R., Sivakumar, R., Anirudh, E.R., Reddy, Y.S.V., John, E.B., 2022. Evaluation of surgical skill using machine learning with optimal wearable sensor locations. *PLoS One* 17, e0267936.
- Solis, J., Oshima, N., Ishii, H., Matsuoka, N., Hatake, K., Takanishi, A., Solis, J., Oshima, N., Ishii, H., Matsuoka, N., Hatake, K., Takanishi, A., 2008. Towards understanding the suture/ligature skills during the training process using WKS-2RII. *Int. J. Comput. Assist. Radiol. Surg.* 3.
- Sugiyama, T., Lama, S., Gan, L.S., Maddahi, Y., Zareinia, K., Sutherland, G.R., 2018. Forces of tool-tissue interaction to assess surgical skill level. *JAMA Surgery* 153, 234–242.
- Sugiyama, T., Sugimori, H., Tang, M., Ito, Y., Gekka, M., Uchino, H., Ito, M., Ogasawara, K., Fujimura, M., 2024. Deep learning-based video-analysis of instrument motion in microvascular anastomosis training. *Acta Neurochir.* 166 (Wien).
- Sun, X., Byrns, S., Cheng, I., Zheng, B., Basu, A., 2016. Smart sensor-based motion detection system for hand movement training in open surgery. *J. Med. Syst.* 41, 24.
- Syاملan, A., Fathurachman, Denis, K., Vander Poorten, E., Pramujati, B., Tjahjowidodo, T., 2022. Haptic/virtual reality orthopedic surgical simulators: a literature review. *Virtual Real.* 26, 1795–1825.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9.
- Tan, M., Le, Q., 2019. EfficientNet: rethinking model scaling for convolutional neural networks. In: Kamalika, C., Ruslan, S. (Eds.), *Proceedings of the 36th International Conference on Machine Learning*. PMLR, *Proceedings of Machine Learning Research*, pp. 6105–6114.
- Tang, M., Sugiyama, T., Takahari, R., Sugimori, H., Yoshimura, T., Ogasawara, K., Kudo, K., Fujimura, M., 2024. Assessment of changes in vessel area during needle manipulation in microvascular anastomosis using a deep learning-based semantic segmentation algorithm: a pilot study. *Neurosurg. Rev.* 47.
- Titov, O., Bykanov, A., Pitskhelauri, D., 2023. Neurosurgical skills analysis by machine learning models: systematic review. *Neurosurg. Rev.* 46, 121.
- Tozzi, P., Solida, A., Siniscalchi, G., Ferrari, E., 2022. A heart surgery simulator with an integrated supervision system for self-learning the key steps and pitfalls of the mitral valve repair: initial investigation. *Simulat. Healthc. J. Soc. Med. Simulat.* 17, 192–197.
- Tricco, A.C., Lillie, E., Zarin, W., O'Brien, K.K., Colquhoun, H., Levac, D., Moher, D., Peters, M.D.J., Horsley, T., Weeks, L., Hempel, S., Akl, E.A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M.G., Garrity, C., Lewin, S., Godfrey, C.M., Macdonald, M.T., Langlois, E.V., Soares-Weiser, K., Moriarty, J., Clifford, T., Tunçalp, Ö., Straus, S.E., 2018. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann. Intern. Med.* 169, 467–473.
- Vedula, S., Ishii, M., Hager, G.D., 2017. Objective assessment of surgical technical skill and competency in the operating room. *Annu. Rev. Biomed. Eng.* 19.
- Velazco Garcia, J.D., Navkar, N.V., Gui, D., Morales, C.M., Christoforou, E.G., Ozcan, A., Abinshed, J., Al-Ansari, A., Webb, A., Seimenis, I., 2019. A platform integrating acquisition, reconstruction, visualization, and manipulator control modules for MRI-guided interventions. *J. Digit. Imag.* 32, 420–432.
- Walsh, R., Kearns, E.C., Moynihan, A., Gerke, S., Duffour, M., Corrales Compagnucci, M., Minssen, T., Cahill, R.A., 2023. Ethical perspectives on surgical video recording for patients, surgeons and society: systematic review. *BJS Open* 7.

- Wang, Q., He, Q., 2023. Deep learning and depth integrated method for visual tracking of object under complicated background. *IEEE ASME J. Microelectromech. Syst.* 1–3.
- Wang, Z., Majewicz Fey, A., 2018. Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. *Int. J. Comput. Assist. Radiol. Surg.* 13, 1959–1970.
- Watson, R.A., 2014. Use of a machine learning algorithm to classify expertise: analysis of hand motion patterns during a simulated surgical task. *Acad. Med.* 89.
- Windrim, L., Melkumyan, A., Murphy, R., Chlingaryan, A., Nieto, J., 2016. Unsupervised feature learning for illumination robustness. *IEEE ASME J. Microelectromech. Syst.* 4453–4457.
- Xie, Y., Zhang, J., Shen, C., Xia, Y., 2021. CoTr: efficiently bridging CNN and transformer for 3D medical image segmentation, medical image computing and computer assisted intervention–MICCAI 2021. 24th International Conference, Strasbourg, France, September 27–October 1, 2021, *Proceedings, Part III* 24. Springer, pp. 171–180.
- Xu, J., Anastasiou, D., Booker, J., Burton, O.E., Layard Horsfall, H., Salvadores Fernandez, C., Xue, Y., Stoyanov, D., Tiwari, M.K., Marcus, H.J., Mazomenos, E.B., 2023. A deep learning approach to classify surgical skill in microsurgery using force data from a novel sensorised surgical glove. *Sensors* 23, 8947.
- Yamada, T., Suda, H., Yoshitake, A., Shimizu, H., 2022. Development of an automated smartphone-based suture evaluation system. *J. Surg. Educ.* 79, 802–808.
- Yanik, E., Intes, X., Kruger, U., Yan, P., Diller, D., Van Voorst, B., Makled, B., Norfleet, J., De, S., 2022. Deep neural networks for the assessment of surgical skills: a systematic review. *The Journal of Defense Modeling and Simulation: applications. Methodology, Technology* 19, 159–171.
- Yibulayimu, S., Wang, Y., Liu, Y., Sun, Z., Wang, Y., Jiang, H., Li, F., 2022. An explainable machine learning method for assessing surgical skill in liposuction surgery. *Int. J. Comput. Assist. Radiol. Surg.* 17, 2325–2336.
- Ying-Ying, Y., Shulruf, B., 2019. An expert-led and artificial intelligence system-assisted tutoring course to improve the confidence of Chinese medical interns in suturing and ligature skills: a prospective pilot study. *J. Educ. Eval. Health Prof.* 16.
- Yoneyama, R., Duran, A.J., Del Pobil, A.P., 2021. Integrating sensor models in deep learning boosts performance: application to monocular depth estimation in warehouse automation. *Sensors* 21, 1437.
- Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.-H., Shao, L., 2021. Multi-stage Progressive Image Restoration, pp. 14821–14831.
- Zhang, S., Shen, X., Lin, Z., Mèch, R., Costeira, J.P., Moura, J.M.F., 2018. Learning to Understand Image Blur, pp. 6586–6595.
- Zheng, Q., Yang, R., Ni, X., Yang, S., Jiang, Z., Wang, L., Chen, Z., Liu, X., 2023. Development and validation of a deep learning-based laparoscopic system for improving video quality. *Int. J. Comput. Assist. Radiol. Surg.* 18, 257–268.
- Zhu, H., Tang, P., Park, J., Park, S., Yuille, A., 2019. Robustness of object recognition under extreme occlusion in humans and computational models. *arXiv preprint arXiv: 1905.04598*.
- Zia, A., Essa, I., 2018. Automated surgical skill assessment in RMIS training. *Int. J. Comput. Assist. Radiol. Surg.* 13, 731–739.
- Zia, A., Sharma, Y., Bettadapura, V., Sarin, E.L., Essa, I., 2018. Video and accelerometer-based motion analysis for automated surgical skills assessment. *Int. J. Comput. Assist. Radiol. Surg.* 13, 443–455.
- Zuckerman, I., Werner, N., Kouchly, J., Huston, E., Dimarco, S., Dimusto, P., Laufer, S., 2024. Depth over RGB: automatic evaluation of open surgery skills using depth camera. *Int. J. Comput. Assist. Radiol. Surg.* 19, 1349–1357.