Castillo-Segura, P., Fernández-Panadero, C., Alario-Hoyos, C., Muñoz-Merino, P. J., Delgado Kloos, C. (2021). Objective and automated assessment of surgical technical skills with IoT systems: A systematic literature review. *Artificial Intelligence in Medicine*, 112, 102007.

DOI: https://doi.org/10.1016/j.artmed.2020.102007

Objective and automated assessment of surgical technical skills with IoT systems: A systematic literature review

Pablo Castillo-Segura, Carmen Fernández-Panadero, Carlos Alario-Hoyos, Pedro J. Muñoz-Merino, Carlos Delgado Kloos

Universidad Carlos III de Madrid, Av. Universidad 30, 28911 Leganés (Madrid), Spain
{pabcasti, mcfp, calario, pedmume, cdk}@it.uc3m.es

* Corresponding author info: Carlos Alario-Hoyos (email: calario@it.uc3m.es)

Acknowledgements: This work was supported in part by the FEDER/Ministerio de Ciencia, Innovación y Universidades–Agencia Estatal de Investigación, through the Smartlet Project under Grant TIN2017-85179-C3-1-R, and in part by the Madrid Regional Government through the e-Madrid-CM Project under Grant S2018/TCS-4307, a project which is co-funded by the European Structural Funds (FSE and FEDER). Partial support has also been received from the European Commission through Erasmus+ Capacity Building in the Field of Higher Education projects, more specifically through projects LALA (586120-EPP-1-2017-1-ES-EPPKA2-CBHE-JP), InnovaT (598758-EPP-1-2018-1-AT-EPPKA2-CBHE-JP), and PROF-XXI (609767-EPP-1-2019-1-ES-EPPKA2-CBHE-JP). This publication reflects the views only of the authors and funders cannot be held responsible for any use which may be made of the information contained therein.

Highlights

- Systematic literature review on objective and automated assessment of surgical technical skills
- 537 papers published after 2013 screened and 101 analyzed in detail
- Main sensors: mechanical/electromagnetic for tool tracking and IMU for body tracking
- Indicators (e.g., path length, smoothness) to distinguish between levels of expertise
- SVM and Neural Networks are the main methods/algorithms for processing the data

Abstract: The assessment of surgical technical skills to be acquired by novice surgeons has been traditionally done by an expert surgeon and is therefore of a subjective nature. Nevertheless, the recent advances on IoT (Internet of Things), the possibility of incorporating sensors into objects and environments in order to collect large amounts of data, and the progress on machine learning are facilitating a more objective and automated assessment of surgical technical skills. This paper presents a systematic literature review of papers published after 2013 discussing the objective and automated assessment of surgical technical skills. 101 out of an initial list of 537 papers were analyzed to identify: 1) the sensors used; 2) the data collected by these sensors and the relationship between these data, surgical technical skills and surgeons' levels of expertise; 3) the statistical methods and algorithms used to process these data; and 4) the feedback provided based on the outputs of these statistical methods and algorithms. Particularly, 1) mechanical and electromagnetic sensors are widely used for tool tracking, while inertial measurement units are widely used for body tracking; 2) path length, number of sub-movements, smoothness, fixation, saccade and total time are the main indicators obtained from raw data and serve to assess surgical technical skills such as economy, efficiency, hand tremor, or mind control, and distinguish between two or three levels of expertise (novice/intermediate/advanced surgeons); 3) SVM (Support Vector Machines) and Neural Networks are the preferred statistical methods and algorithms for processing the data collected, while new opportunities are opened up to combine various algorithms and use deep learning; and 4) feedback is provided by matching performance indicators and a lexicon of words and visualizations, although there is considerable room for research in the context of feedback and visualizations, taking, for example, ideas from learning analytics.

Keywords: technical skills, surgery, IoT, sensors, statistical methods, algorithms, literature review.

1. Introduction

The acquisition of surgical technical skills is fundamental for any surgeon, both in open surgery and in less invasive techniques, such as robot-assisted laparoscopic surgery [1]. Traditionally, the assessment of surgical technical skills has been done by direct observation and feedback from an expert surgeon [2]. This method, although still the most used today, is a subjective method that presents many problems of homogeneity and can be influenced by the interrelationships between the trainee (novice surgeon / student), the mentor (expert surgeon / teacher) and the environment in which the learning takes place, as well as by personality traits of the trainee and the mentor [3].

A number of rating scales have been defined with the aim to structure and make the assessment process of surgical technical skills more objective. Some of these rating scales are used to assess surgical technical skills in general, such as GRS (Global Rating Scales) [4], OSATS (Objective Structured Assessment of Technical Skills) [5], and GEARS (Global Evaluative Assessment of Robotic Skills) [6], while some others are used to assess specific surgical technical skills by procedure, such as GAGES (Global Assessment of Gastrointestinal Endoscopic Skills) [7], in the case of digestive endoscopy. These rating scales have been validated in numerous surgical procedures and have helped to homogenize the assessment process of surgical technical skills, by reducing subjectivity in the scores that expert surgeons give to novice surgeons [8]; however, the application of these rating scales is not always systematic, and there is a major problem due to the low teacher-student ratio (one-to-one for certification purposes and somewhat higher, e.g., one-to-fifteen in training workshops). In addition, in order to increase objectivity and avoid unbiased assessment, not only it is important to standardize the rating scales, but also the types of exercises on which these are applied, especially when the assessment is used for certification purposes; this is the case with e.g., the 5 FLS (Fundamentals of Laparoscopic Surgery) exercises endorsed by the American College of Surgeons (ACS) in the specific domain of laparoscopic surgery [9]. Therefore, it is necessary to evolve the assessment of surgical technical skills towards more objective and automated processes in order to optimize the teacher-student ratio, both in the case of summative assessment (for certification purposes), and in the case of formative assessment (for the continuous improvement of novice surgeons) [8].

Training novice surgeons in surgical technical skills faces similar problems to those of assessing surgical technical skills [10]. Traditionally, a novice surgeon was trained alongside an expert surgeon by seeing the technique as a first step. Then, the novice surgeon helped the expert surgeon by doing minor tasks in some procedures (e.g., handling the camera in laparoscopic surgery). Finally, the novice surgeon participated directly in certain procedures, progressively increasing the degree of difficulty of the tasks assigned. This approach has evolved over time to improve patient safety by introducing earlier stages in the training process that must be completed by the novice surgeon before any contact with the patient. This evolution has led to safer training models that are widely used around the world [11], such as the pyramid training model in the case of laparoscopic surgery [12], where novice surgeons begin their training with simulators (mainly box trainers and in some cases virtual simulators) [13][14][9][15], then continue their training with animals, and finally go to the operating room to do surgery with real patients. Nevertheless, despite this evolution and improvement of the training process through several stages (i.e., simulator, animal, patient), each of these stages still presents the same recurrent problems that also appear in the assessment of surgical technical skills: the difficulty of measuring and assessing progress in the training process of a novice surgeon; the need to improve the teacher-student ratio; and the need for a sequence of standardized exercises in the curriculum to facilitate objective assessment and certification at each learning stage [8].

The technological advances in recent years, in particular the development of the so-called Internet of Things (IoT), have made it possible to incorporate sensors into all kinds of objects and environments with the aim to design new training settings and collect large amounts of data, which can be then processed with artificial intelligence techniques and machine learning algorithms [16]. In the case of training and assessment in surgery, IoT offers a great opportunity since there are various types of sensors which can be added to the surgeon, the instruments or the environment, and that can help to better understand, assess, and optimize a surgical procedure, not only during training but also in professional surgical practice. Nonetheless, although it is easier to add sensors to the more standard parts in surgical settings (surgeon and instruments) there are more difficulties in adding sensors to non-standard parts in surgical settings such as the patient or the environment. Advances in the use of IoT in surgery are rapid and promising with different systems and architectures proposed which could be useful as a support in the training of novice surgeons, as well as machine learning algorithms which could be useful to automatically process the large amounts of data collected and, in this way, assess the surgical technical skills in a more objective and automated manner [17].

There are already some systematic literature reviews (SLRs) addressing the objective and/or automated assessment of surgical technical skills [18][19][20][21][22][23]. Nevertheless, from a clinical point of view, existing SLRs focused on particular techniques (e.g., robotic surgery [18] or laparoscopic surgery [19][20][21]), particular procedures (e.g., mastoidectomy [22]), or specifically in the analysis of the methods used for each procedure (e.g., tool/hand/eye motion tracking [23]). From a technical point of view, existing SLRs analyzed the sensors (but without making explicit references to IoT), the data collected, and, in some cases, even the statistical techniques, but the circle is not closed including the feedback mechanisms that are used to provide the conclusions to the end user. Therefore, there is currently no comprehensive SLR on objective and automated assessment of surgical technical skills with a focus on IoT and that covers the full life cycle, which includes the type of sensors used, the data collected by these sensors (metrics and indicators), the statistical methods and algorithms used to process these data, and the feedback provided to the trainees based on the outputs of these statistical methods and algorithms.

In this context, this SLR is built upon the following four research questions (RQs).

- RQ1: Which sensors have been used to measure surgical technical skills?
- RQ2: Which data have been collected by sensors and served to differentiate between levels of expertise when assessing surgical technical skills?
- RQ3: Which statistical methods and algorithms have been applied to data in relation to the assessment of surgical technical skills?
- RQ4: How has feedback been provided to trainees in the context of assessment of surgical technical skills?

This SLR aims to provide insights on how IoT can help to automate and make the assessment of surgical technical skills more objective, thus tackling also the problem of the low teacher-student ratio in traditional assessment approaches. This SLR can be useful for researchers and developers so that they can replicate similar IoT-supported scenarios or create new ones for the training of surgical technical skills, know what indicators may be useful for differentiating between surgeons' levels of expertise, as well as identify possible drawbacks and future research directions.

This paper is structured as follows. Section 2 presents the methodology used to identify the relevant publications on automated assessment of surgical technical skills with IoT systems. Section 3 answers the four research questions based on the analysis of the publications identified in the previous section. Section 4 discusses the limitations of this literature review and suggests future research directions. The conclusions of this work are drawn in section 5.

2. Methods

2.1. Eligibility Criteria

This systematic literature review (SLR) was done following the recommendations of PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) statement [27]. For this SLR, six terms need to be defined: 1) "***surgical technical skill***" is defined as a manual ability required to perform the surgery in the operating room, which includes dexterity aspects such as efficiency, economy of movement, bimanual dexterity or tissue handling, among others; 2) "***assessment***" is defined as the evaluation of the surgical student's performance in his or her technical tasks (the assessment can be subjective, if an expert surgeon assesses the student's technical skills typically through rating scales, or objective, if automated methods, typically based on the collection of large amounts of quantitative data, are used); 3) "***metrics***" are defined as low-level quantitative data (crude measures) collected by IoT systems in scenarios where surgical students' technical skills are assessed; 4) "***indicators***" are defined as high-level quantitative data (calculated measures) obtained from low-level data, and used to assess surgical students' technical skills; 5) "***IoT***" is defined as a system of interrelated devices (including sensors) that collect and exchange data between them without requiring human interaction; and 6) "***machine learning***" is defined as the set of advanced techniques and methods that allow the processing of data obtained from IoT systems in order to objectively and automatically assess surgical technical skills.

2.2. Search strategy

The first step was the review of three doctoral dissertations (and their references) on automated methods for the assessment of surgical technical skills (two of them written in Spanish and the third one in English) [19][20][21]. These three doctoral dissertations included literature reviews until 2014 and focused on various aspects in the specific field of laparoscopic surgery. Sánchez Margallo [19] analyzed the motion-based metrics and indicators used in the literature to assess surgical technical skills before 2014. Enciso Sanz [20] analyzed statistical approaches and metrics used to assess surgical technical skills before 2014. Kyaw [21] analyzed statistical methods and algorithms used to assess surgical technical skills before 2013. Although these two publications refer specifically to laparoscopic surgery, most of the problems they address are common to other surgical procedures. The second step was the review of three more recent related SLRs (and their references) [18][22][23]. These recent SLRs focused on particular techniques [18], particular procedures [22], or specifically in the methods used to assess surgical technical skills [23]. The review of all these SRLs served to delve deeper into the field and identify the key terms that constitute the query presented and justified in Table 1. This query was run in late 2019 on the scientific database Scopus, which includes several medical scientific databases, such as PubMed and Medline, resulting in 534 articles obtained. It is noteworthy that the three doctoral dissertations [19][20][21] did not appear in the results of running the query (as they are in Spanish and/or were published before 2014), while the three more recent literature reviews [18][22][23] did appear in the results of the query.

Table 1. Query and justification.

	Objective	Query*	Justification
Context	To review articles from the literature on the field of surgical technical skills	**("surgery" OR "surgical" OR "surgeon") AND**	- General terms such as "medical", "medicine", or "clinical" are discarded because they are too generic and do not refer specifically to studies on surgical technical skills [23].
RQ1	To identify types of sensors used to measure surgical technical skills	**("motion" OR "tracking" OR "force") AND**	- Terms are selected based on the most common actions by surgeons [19]. - General terms such as "sensors", "IoT" or "Internet of Things" are discarded because of their low adoption in the field [24][25]. - Specific sensors, such as "accelerometer" or "electromagnetic", are discarded since they are typically included in the action of the surgeon to be measured [26].

RQ2	To identify data collected (raw and high-level data) on surgical technical skills	**(“skill” OR “dexterity”) AND**	- Terms in the query refer to “skill” and the most common related term [19][20][21]. - Other related terms, such as “competence”, “competency” or “expertise” [23], are discarded because they do not add relevant papers.
RQ3	To identify statistical methods and algorithms used to process the data on surgical technical skills	**(“classification” OR “assess*” OR “predict*”)**	- Terms in the query are selected based on the most common purposes for the analysis of the data collected [19][20][21]. - General terms, such as “Training” and “education*”, are discarded as they do not focus on assessment of surgical technical skills. - Specific terms, such as “data mining”, “machine learning” or “Learning Analytics” are discarded because of their low adoption in the field.
RQ4	To identify feedback mechanisms for the trainee on surgical technical skills	-	- General terms such as “feedback” are discarded because the terms used to describe the feedback are highly dependent on the type of exercises which typically suffer from lack of standardization. In addition, the general term does not add relevant papers. - Feedback is analyzed from the papers found with the query, without including additional terms for specific feedback.
***Query applied on title, abstract and keywords for papers published after 2013.**			

2.3. Study selection

Figure 1 presents the PRISMA flow chart [27] followed in this SLR. A total of 534 articles were obtained after running the query. After adding the three above mentioned doctoral dissertations [19][20][21] an initial set with 537 publications was obtained. A researcher reviewed each publication with the support of three additional senior researchers to reach consensus in case of uncertainty. Exclusion criteria were: 1) non-journal or non-conference publications (conference reviews, editorial commentaries and errata were discarded); 2) non-English language publications; 3) publications not discussing metrics, indicators or tools for the objective and/or automated assessment of surgical technical skills; and 4) publications not using IoT systems for the objective and/or automated assessment of surgical technical skills (e.g., articles assessing surgical technical skills through video/image processing, or through Virtual Reality –VR– simulation games with specific VR indicators such as blood loss or volume of tumor removed were discarded). The title and abstract of each publication were screened in the first phase according to the exclusion criteria; this narrowed the publication set from 537 to 165. The full articles were assessed in the second phase according to the exclusion criteria; this narrowed the publication set from 165 to 101, which is the final number of articles included in the analysis. All 101 publications are listed in the references section of this paper.

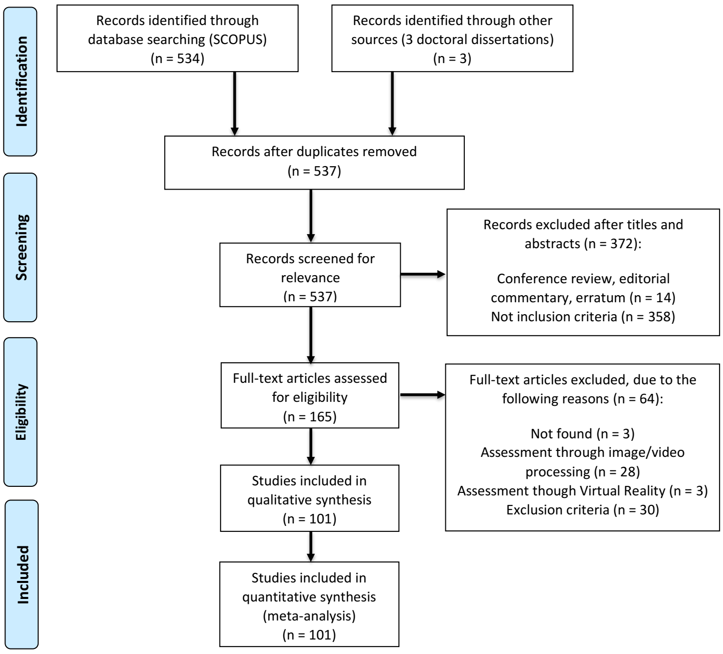
Identification
Screening
Eligibility
Included
Records identified through database searching (SCOPUS) (n = 534)
Records identified through other sources (3 doctoral dissertations) (n = 3)
Records after duplicates removed (n = 537)
Records screened for relevance (n = 537)
Records excluded after titles and abstracts (n = 372):
Conference review, editorial commentary, erratum (n = 14)
Not inclusion criteria (n = 358)
Full-text articles assessed for eligibility (n = 165)
Full-text articles excluded, due to the following reasons (n = 64):
Not found (n = 3)
Assessment through image/video processing (n = 28)
Assessment though Virtual Reality (n = 3)
Exclusion criteria (n = 30)
Studies included in qualitative synthesis (n = 101)
Studies included in quantitative synthesis (meta-analysis) (n = 101)

Figure 1. PRISMA flow chart of the systematic literature review.

3. Results

After agreeing upon all articles that were to be included in the analysis, relevant information was systematically extracted from each of them, being shared through a common spreadsheet. The extracted information was aimed at answering the four Research Questions (RQs), and thus the focus was on: 1) sensors used; 2) data collected; 3) statistical methods and algorithms used; and 4) feedback provided. Figure 2 summarizes the analysis items from the identified articles. Next subsections present the results for each of the four RQs.

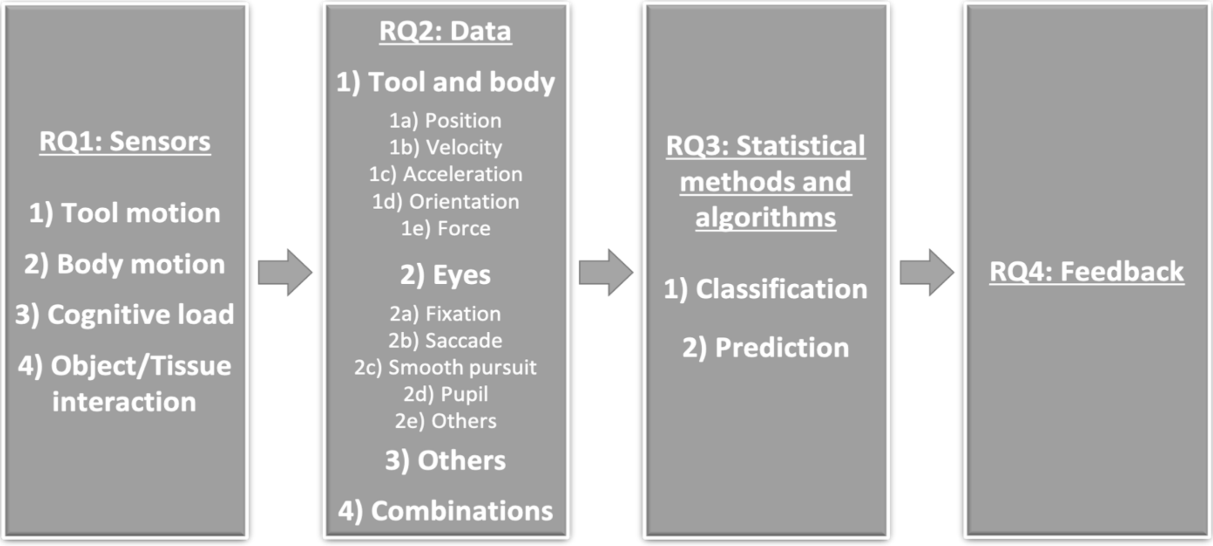

Figure 2. Analysis items from identified articles.

3.1. RQ1: Which sensors have been used to measure surgical technical skills?

IoT systems used in surgery include sensors that collect quantitative data that can be used for the objective and automated assessment of surgical technical skills. There are several types of sensors that are used alone or in combination with other sensors. Table 2 shows the sensors found in the 101 identified articles. For ease of explanation, the sensors are organized into four main categories: **1) tool motion**; **2) body motion**; **3) cognitive load**; and **4) object/tissue interaction**.

Table 2. Sensors used to collect data and the corresponding references (one paper may include several sensors).

Category	Sensors	Advantages	Disadvantages	# Papers	References
Tool motion	Mechanical	- Good accuracy - Robust against interference - Sterilizable	- Ergonomic problems - Limited range of motion due to physical connection - Less portable	26	[28][29][30][31][32][33][34][35][36][37][38][39][40][41][42][43][44][45][46][47][48][49][50][51][52][53]
	Electromagnetic (EM)	- Good accuracy in low ranges - Reduced size - Intermediate cost - Line of sight not required	- Rapid decrease in precision and resolution with distance - Affected by the environment (metal and magnetic objects) - Sensors are wired - Low sample rate	18	[23][44][54][55][56][57][58][59][60][61][62][63][64][65][66][67][68][69]
	Inertial Measurement Units (IMU)	-High sample rate	- Cumulative error - Wired to provide power	10	[42][46][70][71][72][73][74][75][76][77]
	Optical markers	-Accuracy -Robustness (No dependence on objects in its environment) -Large Range - Wireless position markers	- Requires line of sight - Optical markers are relatively high	7	[71][78][79][80][81][82][83]
	Active Infrared Sensors (AIR)	- Large range - Good accuracy - High resolution	- Requires line of sight - Affected by the environment (light, some materials) - Commercial devices are too expensive	1	[84]
	Acoustic	- High range - Low cost	- Slow update rates - Speed of sound affected by environmental conditions	1	[85]
	Flexion sensor for tool opening bending	- Low cost - Easy to use - Detects high range of bending angles	- Errors over time due to changes in sensor flexibility - Limitation of movement by wiring	1	[63]
Body motion	Inertial Measurement Units (IMU)	- High sample rate	- Cumulative error - Wired to provide power	14	[8][23][24][25][26][86][87][88][89][90][91][92][93][94]
	Electromagnetic (EM)	- Good accuracy in low ranges - Reduced size - Intermediate cost - Line of sight not required	- Rapid decrease in precision and resolution with distance - Affected by the environment (metal and magnetic objects) - Instrument sensors are wired - Low sample rate	6	[8][95][96][97][98][99]
	Optical markers	- Accuracy - Robustness (No dependence on objects in its environment) - Large Range - Wireless position markers	- Requires line of sight - Optical markers are relatively high	5	[51][77][100][101][102]

	Active Infrared Sensors (AIR)	- Large range - Good accuracy - High resolution	- Requires line of sight - Affected by light/objects in the environment - Commercial devices are too expensive	2	[23][103]
	Flexion sensor for fingers joint bending	- Low cost - Easy to use - Detects high range of bending angles	- Errors over time due to changes in sensor flexibility - Limitation of movement by wiring	1	[26]
Cognitive load	Smart glasses (eye tracking)	- Natural movement and unobstructive	- Expensive - Some people cannot work with the device - Calibration required (and this process takes some time).	9	[23][29][104][105][106] [107][108][109][110]
	Electroencephalogram (EEG)	- High precision in time - High sample rate - Rapid signal changes can be detected	- Poor spatial precision - Difficult to determine source origin - Limitation of movement by wiring	3	[87][94][111]
	Electromyography (EMG)	- High sample rate	- Noisy, weak signals with interference caused by other muscles (difficult to filter) - Limitation of movement by wiring	2	[8][87]
	Electrodermal activity (EDA)	- High sample rate - Easy to use	- Sometimes time affects the polarization in the electrodes, and therefore the precision - Limitation of movement by wiring	2	[8][87]
	Heart Rate Variability (HRV)	-Easy to use	- Low precision due to movement and skin types - Limitation of movement by wiring	1	[87]
Object / Tissue interaction	Force sensor	- Very thin and flexible construction	- Accuracy could vary depending on device and range of force applied	10	[31][54][64][66][71][75] [112][113][114][115]

The main purpose for using sensors in a surgical procedure is to track **tool motion** (59 different articles); this is done by incorporating sensors that capture motion in surgical instruments, such as catheters, needles, graspers or scissors [31][55][70][75][80]. Sensors are also used to track **motion in the surgeon's body** during a surgical procedure (25 different articles); this is done by incorporating sensors which capture motion typically on the surgeon's hands, wrists, arms or head [86][91][94][99][103]. In addition, sensors are also used to measure the **cognitive load of the surgeon** (13 different articles), incorporating sensors that collect activity on the surgeon's eyes, brain, muscles, heart or skin [8][87][110]. Finally, it is also possible to add sensors on other **objects and/or tissues** (10 articles) typically to measure the force used by the surgeon during the surgical procedure (e.g., when removing a gauze, staples, or the retrieval bag used when removing the appendix) [54][64][113] .

Sensors can be used for different purposes related to the assessment of surgical technical skills and have both advantages and disadvantages. The most common type of sensors are **mechanical sensors** (26 articles). Mechanical sensors have a high precision in movements, are immune to interferences and can be sterilized. However, their ergonomics can affect the surgeon's movement during the execution of the surgical procedure. Moreover, mechanical sensors have a limited range of movement due to their wired connection and are more difficult to move due to their size [28]. **Electromagnetic (EM) sensors** are also very common (24 articles) due to their reduced size and cost compared to mechanical sensors. However, their low latency, wired connection and interferences with metallic objects can affect the quality of the data collected [65]. **Inertial Measurement Units (IMUs)** are also very popular (24 articles) since they can be found in small and inexpensive devices. IMUs combine gyroscopes and accelerometers and allow the measurement of g-forces, angular rates and orientations, taking samples at a high rate. However, IMUs may need wires for power supply and have accumulative errors over time, which may affect the quality of the data collected [71]. **Optical markers** (12 articles) are a type of passive infrared sensors which are frequently used in motion capture since they can be accurately detected and have a reasonable range and precision in movements, although they present problems when there is not a straight line of sight; this same problem is also found in **Active infrared sensors (AIR)** (3 articles) [84]. Other popular sensors are **force sensors** (10 articles), although they require the isolation of friction forces and involuntary vibrations for a better precision [54], and **smart glasses** (9 articles), which are specifically used to analyze the surgeon's eye activity (gaze patterns, eye blinking, etc.) and contain more expensive sensors [104][106]. Other sensors used to collect physiological information of the surgeon during the surgical procedure are **electroencephalogram (EEG)** (brain activity), **electromyography (EMG)** (muscle activity), **electrodermal activity (EDA)** (sweat), or **heart rate variability (HRV)** [87]. Nevertheless, the wires in these sensors limit the surgeon' s range of movement. In addition, the data they collect is not very useful for the objective and automated assessment of surgical technical skills [8]. One last type of sensor, but with a very low use and for a very specific purpose, are **flexion sensors** (to measure surgeon's fingers joint bending [26] and tool opening bending [63]).

In summary, there are some research opportunities regarding the use of sensors for the objective and automated assessment of surgical technical skills. In relation to the **movement of both instruments and surgeon's body,** although there are a number of related publications, there are still research gaps on the reduction of the weight of sensors, the increase in the range of movements (for instruments and surgeons' body), and the creation of wireless IoT systems [28]; sensors can also be enhanced to have a higher number of Degrees of Freedom (DoF), in order to collect more data with one single sensor, as it is the case of Genovese et al. [72], who used a sensor with nine DoF to collect three-axis orientation, as well as three-axis linear and angular velocities. In relation to **biometric signals** aimed at representing the cognitive load of the surgeon, it is necessary to create IoT systems that allow the collection of more useful data, perhaps by combining multiple biometric signals at once [8]. In relation to **forces**, more research work is needed to reduce friction forces and vibrations during surgical procedures in order to obtain higher quality data. Finally, it is important to note that some data are currently obtained using cameras and image processing, rather than sensors, such as perpendicular error or needle angle from plane [119]. Therefore, it is important to create sensor networks that allow the incorporation of data, not only from physical sensors but also from other less invasive sources of information such as image processing. All in all, more research is needed to integrate sensors in IoT systems to collect different types of data, such as tool motion, surgeon's body motion, surgeon's cognitive load, and forces applied to objects or tissues, and share all that quantitative data within an IoT network; in that context, it is worth noting the challenge of matching timestamps in order to combine data when using a large variety of sensors [24]. In addition, it should be noted that there is considerable research on the use of sensors in the elements that are common in all procedures (e.g., instruments and surgeon's motion), but there is very little research on the use of sensors on the elements used in specific exercises (all the related efforts have focused on measuring the force applied to objects/tissues), hence the importance of standardizing first the types of exercises (as is the case with the 5 FLS exercises) [9].

3.2. RQ2: Which data have been collected by sensors and served to differentiate between levels of expertise when assessing surgical technical skills?

Sensors collect quantitative data that can be used for the objective and automated assessment of surgical technical skills. The raw data collected by sensors are called metrics, while the calculated measures obtained after processing raw data are called indicators. Indicators are important when making decisions, for example, to differentiate between various levels of expertise when assessing trainees' performance (e.g., novice surgeon, intermediate surgeon, advanced/expert surgeon). The indicators found in the 101 identified articles have been organized into three categories: **1) Tool and body motion tracking indicators**; **2) Eye-tracking indicators**; and **3) Other indicators**. Two of the most popular rating scales, OSATS [5] and GEARS [6], have been taken as a reference to identify the main eight surgical technical skills:

- **Bimanual dexterity (BD)**: Independent use of both hands.
- **Depth perception (DP)**: Instruments handling to the target plane.
- **Economy (EC)**: Smooth progress and economy of movement.
- **Efficiency (EF)**: Optimization of movements made per time unit.
- **Hand tremor (HT)**: Minimization of hand tremor in movements.
- **Mind control (MC)**: Management of the mental workload in a surgical procedure.
- **Precision (P)**: Precise movements from one point to the next one.
- **Respect for tissue (RT)**: Tissue handling to minimize damage.

3.2.1. Tool and body tracking indicators

There are numerous indicators related to tracking the tools the surgeon uses during a surgical procedure (e.g., needles or scissors) and the surgeon's own body (e.g., hands or arms). For a better organization, the indicators found in the 101 identified articles related to tool and body tracking are organized into five types: **a) Position**; **b) Velocity**; **c) Acceleration**; **d) Orientation**; and **e) Force**. It is important to note that some indicators, such as those related to velocity and acceleration, are usually calculated from the position values. Table 3 (position indicators), Table 4 (velocity indicators), Table 5 (acceleration indicators), Table 6 (orientation indicators) and Table 7 (force indicators) detail the indicators of each type, pointing out the surgical technical skills that can be assessed with each indicator and the articles that used each indicator. In addition, Figure 3 summarizes the indicators of each type that can be used to distinguish between two or three levels of expertise (novice surgeon, intermediate surgeon and advanced/expert surgeon) as detailed throughout this subsection.

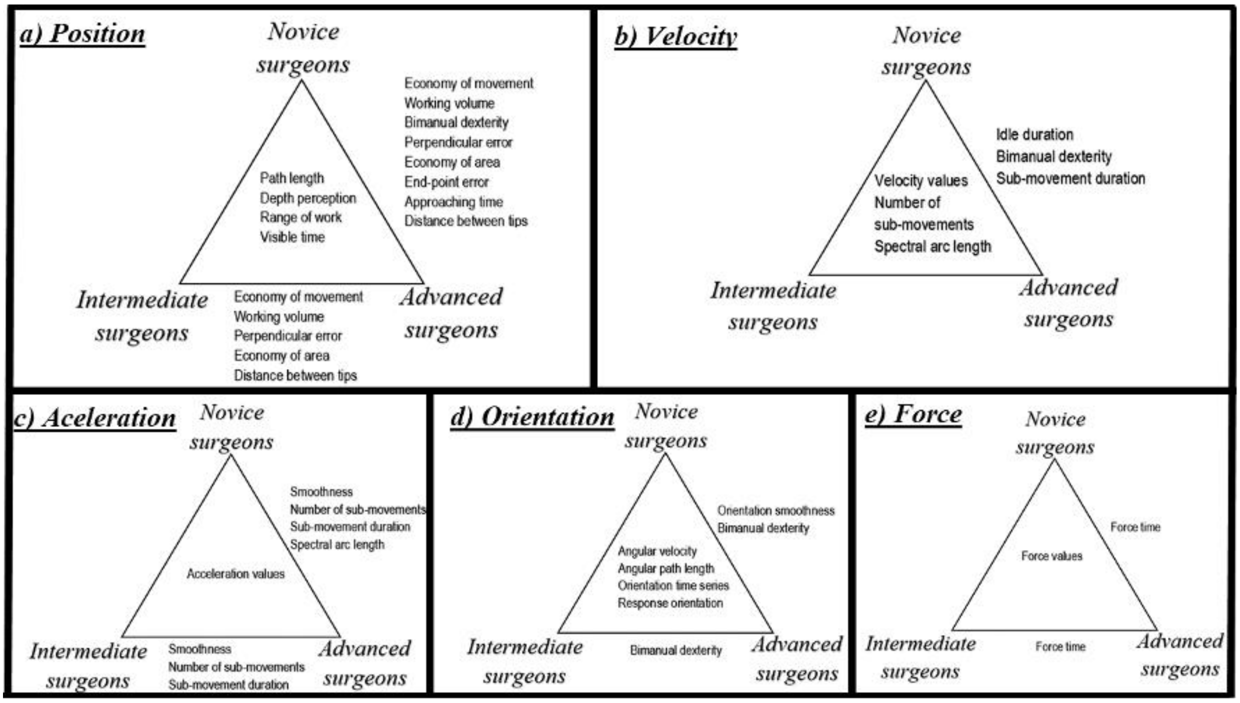

Figure 3. Indicators with significant differences between levels of expertise (novice surgeons, intermediate surgeons and advanced surgeons): a) position; b) velocity; c) acceleration; d) orientation; and e) force (see Table 3 - 7) for a full list of indicators). Those indicators that can be used to distinguish between the three levels of expertise are in the center of the corresponding triangle. Those indicators that can be used to distinguish between two levels of expertise are on the side of the corresponding triangle. Indicators for which no significant differences between levels of expertise were found are not shown in the corresponding triangle.

Regarding indicators related to **position** (see Table 3), *path length* is the most common indicator used (45 articles) [61][64][65][71][78][80][100][101][103][116]. This indicator has shown **significant differences between novice, intermediate and advanced surgeons** when assessing the surgical technical skill called economy of movement (EC) in complex tasks and with the non-dominant hand [79], and between intermediate and advanced surgeons with the dominant hand [66]. Other indicators that have shown significant differences between the three levels of expertise are: *depth perception* [46] (8 articles) to assess the surgical technical skill also called Depth Perception (DP); *range of work* (5 articles) [92] to assess EC, especially with the right part of the body; and *visible time* (5 articles) [79] to assess the surgical technical skill called Efficiency (EF). Indicators that have shown **significant differences between novice and advanced surgeons** are: *economy of movement* (10 articles) to assess EC [60], *working volume* (7 articles) to assess EC [82][101], *bimanual dexterity* obtained through position (6 articles) to assess the surgical technical skill also called Bimanual Dexterity (BD) [44][46], *perpendicular error* (6 articles) to assess EC and Precision (P) [60], *economy of area* (5 articles) to assess EC [82], *end-point error* (3 articles) to assess EC and P [32][52], *approaching time* (2 articles) [96] to assess DP and EF, and *distance between tips* (1 article) [79] to assess EC. In addition, five of these eight indicators (*economy of movement*, *working volume*, *perpendicular error*, *economy of area* and *distance between tips*) have also shown **significant differences between intermediate and advanced surgeons**. Nevertheless, the distinction between levels of expertise for position indicators may depend on the task (preferably complex for a better distinction) and simulator (preferably with more actions from the surgeon for a batter distinction), and in some cases there are articles who did not find significant differences for some of the above-mentioned indicators [31][65][72][81][83][102]. For example, it is more difficult to find differences between novice and advanced surgeons in simple tasks such as cutting, unlike in the case of more complex tasks, such as suturing [81]. Figure 3a summarizes the best results obtained for position indicators in the identified papers.

Table 3. Descriptions, skills and references for position indicators related to tool and body tracking (one paper may include several indicators).

	Indicator	Description	Skill	# Papers	References
Position	Path length	Length of the path travel	EC	45	[8][18][23][25][28][29][30][31][33][35][37][38][41][42][44][45][47][48][57][61][63][64][65][66][71][72][74][78][79][80][81][82][83][84][85][89][95][98][99][100][101][102][103][116][117]
	Economy of movement	Ratio between path length and theoretical shortest path, and/or number of movements per second	EC	10	[25][32][34][35][41][46][52][60][85][88]
	Depth perception	Total distance travelled along the depth axis	DP	8	[28][38][44][46][80][82][83][85]
	Working volume	Relationship between the maximum volume occupied by the instrument (spherical and/or cubic volume) and the total path length	EC	7	[46][79][80][82][85][95][101]
	Bimanual dexterity (position)	Ability to handle two instruments at the same time through position	BD	6	[8][18][30][44][46][63]
	Perpendicular error	Deviations from the ideal course in the direction of the ideal path	EC/P	6	[31][32][52][60][85][95]
	Economy of area	Relationship between the maximum area occupied by the instrument and total path length	EC	5	[34][45][80][82][85]
	Range of work	Amplitude between position limits in one axis	EC	5	[18][36][37][76][92]
	Visible time	Percentage of time spent in the “search zone” and/or out of the camera view	EF	5	[34][48][63][79][85]
	Position time series	Position in the three Cartesian axes at different times	EC	4	[24][40][55][91]

	End-point error	Euclidean distance between the final point and the objective point	EC/P	3	[31][32][52]
	Approaching time	Time taken to reach the target point	DP/EF	2	[85][96]
	Distance between tips	Mean distance between tooltips within the area of interest	EC	1	[79]
	Transit profile	Transit path projected in the 2D plane	EC	1	[85]

Regarding indicators related to **velocity** (see Table 4), *velocity values*, understood as the path traveled on each of the three Cartesian axes at a given time (including mean, maximum value, minimum value or standard deviation), is the most common indicator used (26 articles). This indicator has shown **significant differences between novice, intermediate and advanced surgeons** when assessing the surgical technical skill EF [44][64][80][88]. Other indicators that have shown significant differences between the three levels of expertise are: *number of sub-movements* obtained through velocity (18 articles) when assessing EC [31][56][57][61][64][74]; and *spectral arc length* obtained through Fourier spectrum of velocity (1 article) when assessing also EC [118] (distinguishing smooth movements through low-frequency components from non-smooth movements through high-frequency components). Indicators that have shown **significant differences between novice and advanced surgeons** are: *idle duration* (8 articles) when assessing EF [44][80]; *bimanual dexterity* obtained through velocity (5 articles) when assessing BD [44]; and sub-movement duration obtained through velocity (3 articles) when assessing EF [56]. In addition, *idle duration* has also shown **significant differences between novice and intermediate surgeons** [82]. Nevertheless, once again, the distinction between levels of expertise for velocity indicators may depend on the task and environment, and some authors did not find significant differences for their specific settings in indicators such as *velocity parameters* [60][81][72], *bimanual dexterity* [82]; *number of sub-movements* [44][72]; or *idle duration* [100]. It is noteworthy that faster performances do not mean a better quality, as a trade-off between speed, accuracy and stability is needed [60]. Sánchez-Margallo et al. [82] suggested that some studies which found significant differences in velocity indicators may have committed an error in the post-filtering stage, which could have reduced the motion information and, hence, the correlation between hands. In contrast, Hofstad et al. [44] suggested that some studies did not observe significant differences because they used data from each hand independently, instead of processing data from both hands together. In any case, Figure 3b summarizes the best results obtained for velocity indicators in the identified papers.

Table 4. Descriptions, skills and references for velocity indicators related to tool and body tracking (one paper may include several indicators).

	Indicator	**Description**	**Skill**	**# Papers**	**References**
Velocity	Velocity values	Total path traveled on each of the three Cartesian axes at a given time	EF	26	[8][18][23][25][28][29][38][39][43][44][48][49][50][53][57][60][63][64][72][76][80][81][82][85][95][118]
	Number of sub-movements (velocity)	Number of movements made to complete the task through velocity	EC	18	[25][28][31][33][36][44][45][47][56][57][61][64][72][74][89][99][103][118]
	Idle duration	Percentage of time where the instrument is still	EF	8	[28][44][80][82][85][95][96][100]
	Bimanual dexterity (velocity)	Ability to handle two instruments at the same time through velocity	BD	5	[18][28][44][63][82]
	Sub-movement duration (velocity)	Percentage of time where the velocity exceeds a threshold	EF	3	[56][63][118]
	Spectral arc length (velocity)	Image of the smooth movements composed mainly of low-frequency components and non-smooth movements composed by high frequency components in the Fourier spectrum of velocity.	EC	1	[118]

Regarding indicators related to **acceleration** (see Table 5), *smoothness* (sometimes referred to as *jerk*), which is defined as changes in acceleration, is the most common indicator used (23 articles) when assessing the surgical technical skill called Hand Tremor (HT). This indicator has shown significant differences between novice and advanced surgeons [82], as well as between intermediate and advanced surgeons [60]. The second most common indicator used is *acceleration values* in surgical instruments (22 articles) (including mean, maximum value, range, standard deviation, root-mean-square and total-sum-of-square). This indicator has shown **significant differences between novice, intermediate and advanced surgeons** when assessing the surgical technical skill EF [46][56][64][76][80][86][93][94][102], specially non-linear acceleration parameters, such as kurtosis (a measure of the combined weight of the tails of a distribution relative to the center of the distribution) and power spectral density (a measure of the power content of the signal relative to the frequency) [86]. Indicators that have shown **significant differences between novice and advanced surgeons** beyond *smoothness* are: *number of sub-movements* obtained through acceleration (9 articles) when assessing EC [31][52][102]; *sub-movement duration* obtained through acceleration (4 articles) when assessing EF [102]; and *spectral arc length* obtained through Fourier spectrum of acceleration (2 articles) when assessing EC [56]. In addition, two of these three indicators beyond *smoothness* (*number of sub-movements and sub-movement duration*) have also shown **significant differences between intermediate and advanced surgeons**. Other indicators with a minor use, such as *curvature*, *turning angle*, *tortuosity*, *velocity gain factor* or *dimensional squared jerk*, do not report significant differences between levels of expertise. Nevertheless, once again, the distinction between levels of expertise for acceleration indicators may depend on the task and environment [94]. For example, Sánchez-Margallo et al. [81] did not find significant differences between levels of expertise for smoothness and acceleration values. In addition, Viriyasiripong et al. [94] showed that although novice surgeons tend to make more unnecessary movements, it is also possible to find involuntary movements in expert surgeons. Furthermore, as the surgeon gets older, the chances of having more hand tremor increase [52]. In any case, Figure 3c summarizes the best results obtained for acceleration indicators in the identified papers.

Table 5. Descriptions, skills and references for acceleration indicators related to tool and body tracking (one paper may include several indicators).

	Indicator	**Description**	**Skill**	**# Papers**	**References**
Acceleration	Smoothness / Jerk	Motion analysis parameter based on the third derivative of position (change in acceleration)	HT	23	[8][25][28][32][35][38][41][42][44][46][50][56][57][60][64][70][80][81][82][85][95][97][118]
	Acceleration values	Rate of change of the instrument velocity	EF	22	[8][24][26][33][50][62][63][64][70][73][75][76][77][80][81][82][85][86][87][90][94][102]
	Number of sub-movements (acceleration)	Number of movements made to complete the task through acceleration	EC	9	[8][23][31][52][62][63][74][85][102]
	Sub-movement duration (acceleration)	Time of the sub-movements made to complete the task through acceleration	EF	4	[56][62][63][118]
	Bimanual dexterity (acceleration)	Ability to handle two instruments at the same time through acceleration	BD	2	[63][73]
	Spectral arc length (acceleration)	Arc length of the Fourier magnitude spectrum within an adaptive frequency range	EC	2	[56][62]
	Curvature	Calculated based on straightness of the path computed at each point	EC	2	[35][38]
	Turning angle	Direction of movement with respect to the previous and subsequent steps	EC	1	[38]
	Tortuosity	Property of the curve traveled which depends on the number of turns	EC	1	[38]

	Velocity gain factor	Gain factor relating velocity and radius of curvature	EF	1	[41]
	Dimensional squared jerk	Calculated based on deliberate hand movements	HT	1	[102]

Regarding indicators related to **orientation** (see Table 6), *angular velocity*, understood as the travel angle on each of the three axes of rotation in a given time, is the most common indicator used (14 articles). This indicator has shown **significant differences between novice, intermediate and advanced surgeons** when assessing the surgical technical skill EF [76]. Other indicators that have shown significant differences between the three levels of expertise are: *angular path length* (10 articles) when assessing EC [46][70][74]; *orientation time series* (9 articles) (including mean and standard deviation) when assessing EC; and *response orientation* (4 articles) when assessing also EC [46]. Indicators that have shown **significant differences between novice and advanced surgeons** are: *orientation smoothness* (2 articles) when assessing HT [96]; and *bimanual dexterity* obtained through angular velocity (1 article) when assessing BD [70]. In addition, **significant differences between intermediate and advanced surgeons** were also found in *bimanual dexterity*. Nevertheless, once again, the distinction between levels of expertise for orientation indicators may depend on the task and environment, and also on the use of the dominant and non-dominant hand [74], and some authors did not find significant differences for their specific settings in indicators such as angular velocity [70]. In any case, Figure 3d summarizes the best results obtained for orientation indicators in the identified papers.

Table 6. Descriptions, skill and references for orientation indicators related to tool and body tracking (one paper may include several indicators).

	Indicator	Description	Skill	# Papers	References
Orientation	Angular velocity	Total travel angle, on each of the three axes of rotation in a given time	EF	14	[8][24][29][39][43][48][49][53][70][73][76][86][90][103]
	Angular path length	Change in the angle of the instrument in the plane perpendicular to the instrument axis	EC	10	[28][44][46][57][63][70][73][74][77][87]
	Orientation time series	Orientation in three axes of rotation at different times	EC	9	[24][26][39][43][48][49][53][55][86]
	Response orientation	Rotation on the instrument axis (in radians)	EC	4	[28][44][46][85]
	Orientation smoothness	Derivative of the orientation acceleration	HT	2	[96][97]
	Bimanual dexterity (angular velocity)	Ability to handle two instruments at the same time through angular velocity	BD	1	[70]
	Number of sub-movements (angular velocity / acceleration)	Number of movements made to complete the task, using angular velocity or angular acceleration	EC	1	[63]

Regarding indicators related to **forces** (see Table 7), *force values*, understood as the force applied on each of the three Cartesian axes at a given time (including maximum value, mean, median, deviation, root-mean-square or total-sum-of-square), is the most common indicator used (14 articles); this force can be applied to a tissue or any other object needed to perform the surgical task, such as gauzes, staples, or retrieval bags. This indicator has shown **significant differences between novice, intermediate and advanced surgeons** when assessing the surgical technical skill called Respect for Tissue (RT) [64][71][77][79][88][113][117]. *Force time*, which is the time in which the force exceeds a certain threshold has shown differences **between novice and advanced surgeons** [115] and **between intermediate and advanced surgeons** [79] when assessing EF and RT. Nevertheless, once again, the distinction between levels of expertise for force indicators may depend on the task and environment [64], and some authors did not find significant differences for their specific settings in force values [66][71][112]. Figure 3c summarizes the best results obtained for force indicators in the identified papers.

Table 7. Descriptions, skill and references for force indicators related to tool and body tracking (one paper may include several indicators).

	Indicator	Description	Skill	# Papers	References
Force	Force values	Force applied on the three Cartesian axes at a given time	RT	14	[16][23][64][66][71][73][75][77][79][112][113][114][115][117]
	Force time	Time the force exceeds a certain threshold	EF/RT	3	[79][88][115]

3.2.2. Eye-tracking indicators

Tool and body tracking indicators can be useful to assess surgical technical skills like BD, DP, EC, EF, HT, P, and RT. Nevertheless, the surgeon's performance in the operating room also depends on the surgeon's Mind Control (MC). Eye tracking considers the task-evoked pupillary response based on eye movements and changes in the pupil to observe the cognitive process [29], and, therefore, can be used to assess MC. In fact, there are a number of eye-tracking indicators that have been used to assess MC. For a better organization, the indicators found in the 101 identified articles related to eye tracking are organized into five types: **a) fixation** (i.e., eyes being focused on the area of interest); **b) saccade** (i.e., rapid eye movement between two points of interest); **c) smooth pursuit** (i.e., movements of the eye when following a moving object); **d) pupil** (including size and changes in size); and **e) others** (including blinking and combined indicators). Table 8 details the indicators of each type.

The most common indicators are related to *fixation*: *duration* (5 articles), *rate* (4 articles), and *number* (3 articles). These three indicators have shown **significant differences between novice, intermediate and advanced surgeons** [107], or at least between novice and advanced surgeons [104][106][108][109]. Other indicators that have been shown **significant differences between novice and advanced surgeons** are: *saccade duration* (3 articles), *number* (3 articles) and *amplitude* (2 articles) [104][108]; *smooth pursuit number* (1 article) and *duration* (1 article) [108]; *pupil size* (diameter) (4 articles), *rate of change* (2 articles), *index of pupillary activity* (2 articles), and *entropy* (1 article) [29][106][108][109]; *blink rate* (1 article) [106], and *dwell time* (1 article) [109]. Once again, the distinction between levels of expertise for eye-tracking indicators may depend on the task and environment. For example, in the case of saccade duration, number, and rate, significant differences were found between novice and advanced surgeons in complex tasks, such as suturing, but not in basic ones [104]; this means that advanced surgeons may change their visual search strategy for a better information processing in more difficult tasks [104]. As another example, the surrounding light in the simulation environment may affect pupil size and change in pupil diameter [107]; in contrast, the index of pupillary activity allows a good distinction between levels of expertise, obtaining a pupil dilation response caused by cognition and not by the surrounding light [107]. Furthermore, in the particular case of saccade amplitude, differences may be due to the fact that the object is covered by tools or hands and not so much to the surgeon's level of expertise [104]. Differences in saccade duration are typically due to the fact that advanced surgeons know the exact locations where they have to look at each moment, while novice surgeons use more visual feedback to perform the proper movements, so they need to follow more the objects [108]. Advanced surgeons also have less fixation duration than novice surgeons, which indicates a greater ability to obtain the appropriate information in the area of interest in a shorter time [108].

Table 8. Description and references for eye tracking indicators (one paper may include several indicators).

Type	Indicator	Description	Skill	# Papers	References
Fixation	Duration	Cumulative time in which the gaze is kept in the area of interest	MC	5	[104][105][106][107][108]
	Rate	Number of observations in the area of interest per second	MC	4	[104][105][106][107]

	Number	Total number of observations made in the area of interest	MC	3	[107][108][109]
Saccade	Duration	Cumulative time of rapid eye movements between several points in the area of interest.	MC	3	[104][105][108]
	Number	Total number of rapid eye movements between several points in the area of interest.	MC	3	[104][105][108]
	Rate	Number of rapid eye movements in the area of interest per second	MC	2	[104][105]
	Amplitude	Extent of rapid eye movements between several points in the area of interest	MC	2	[104][105]
Smooth pursuit	Number	Total number of eye movements that occur when following a dynamic object in the area of interest	MC	1	[108]
	Duration	Cumulative time of eye movements that occur when following a dynamic object in the area of interest	MC	1	[108]
Pupil	Size	Diameter of the pupil	MC	4	[106][107][108][109]
	Rate of change	Pupil dilation rate (mm/s)	MC	2	[106][109]
	Index of activity	Discontinuities in the signal created from a continuous recording of the pupil diameter	MC	2	[29][107]
	Entropy	Predictability of pupil change [bits]	MC	1	[109]
	Change in diameter	Modifications in pupil diameter compared to the baseline	MC	1	[107]
Others	Blink rate	Number of blinks per second	MC	1	[106]
	Dwell time	Sum of durations from fixations and saccades in the area of interest	MC	1	[109]

3.2.3. Other indicators

There are some other indicators related to the surgeon’s performance that have been used in the literature (see Table 9). The most common indicator is the *total time* spent on the development of a task (49 articles) when assessing EF. This indicator typically serves to differentiate between novice, intermediate and advanced surgeons [47], although it is not always easy to differentiate between intermediate and advanced surgeons when performing very simple tasks [79]. The remaining indicators have been used only in a few articles. For example, the *number of hits* made (3 articles) when assessing EF has shown significant differences between novice and advanced surgeons [65], However, Gong et al. [88] carried out an experiment in which surgeons had to hit some markers with one tool and perform the next step with another tool in less than 0.5 seconds, without finding significant differences between levels of expertise for this particular experiment. In addition, four indicators have been used to assess MC, through the cognitive load associated with performing a surgical procedure and measured in various parts of the surgeon’s body: 1) *electrical muscle activity* collected through EMG (2 articles) has been used to classify surgeons in four levels of expertise [8]; 2) *brain bioelectrical activity* collected through EEG (2 articles) has been used to estimate surgeons’ workload [87], and has shown significant differences between novice and advanced surgeons [111]; 3) *skin conductivity* collected through EDA (2 articles) typically increases with stress and has been used to estimate performance, expertise and workload [87]; 4) *pulsation rate* collected through HRV parameters (e.g., Low Frequency power, High Frequency power, or ratio of Low Frequency power to High Frequency power) (1 article) also tends to increase with stress and has also been used to estimate performance, expertise and workload [87]. Finally, *fingers joint flexion* (1 articles) when assessing RT has been used to distinguish between two levels of expertise in surgeons [26]. In conclusion, with the exception of the total time, the most common indicators used in the literature are related to tool, body and eye tracking, although it may be interesting to consider other possible indicators.

Table 9. Descriptions, skills and references for other indicators (one paper may include several indicators).

Indicator	Description	Skill	# Papers	References
Total time	Total time invested in the development of a task	EF	49	[18][23][25][28][29][30][31][33][34][36][37][38][44][45][46][47][57][60][61][63][64][65][66][70][71][72][74][78][79][80][81][82][83][84][85][88][89][90][93][94][95][99][100][101][102][103][116][117]
Number of hits	Number of markers successfully hit	EF	3	[65][88][116]
Electrical muscle activity	Muscle activity collected through EMG	MC	2	[8][87]
Brain bioelectrical activity	Brain activity collected through EEG	MC	2	[87][111]
Skin conductivity	Electrodermal activity collected through EDA	MC	2	[33][87]
Pulsation rate	Heart activity collected through HRV	MC	1	[87]
Fingers joint flexion	Bending angle of the finger joints	RT	1	[26]

3.2.4. Combination of indicators

Each indicator has so far been examined in isolation, but it is also important to consider the effect of **combining indicators** when assessing surgical technical skills. Ideally, a skilled surgeon can be expected to perform surgical procedures with higher smoothness, lower velocity values, acceleration values, force values, and total time, following a shorter path, and doing a reduced number of movements [64]. There are widespread rating scales that have been used to assess surgical technical skills, such as GRS [4], OSATS [5] and, GEARS [6]. These scales use multiple-choice questions with rubrics in which the influence of several of the above-mentioned indicators are combined to calculate an overall score of surgical ability. These three scales have been widely validated in the literature in different procedures, and therefore, can be taken as reference to assess the reliability of combinations of indicators.

One possible strategy for analyzing the effect of combining several indicators is to study the **correlation** among them or with the scores obtained by surgeons through rating scales. It is important to point out that sensors provide low-level indicators. These indicators are objective information, but this information cannot be directly extrapolated to surgical technical skills. In contrast, rating scales provide high-level indicators directly related with surgical ability, but this information is partly subjective because it is collected through an expert surgeon who completes these scales using a rubric through direct observation of the surgeon being evaluated. A high correlation of low-level indicators obtained through sensors with specific rating scales completed by expert surgeons could be the basis for replacing (or complementing) these rating scales with more objective and automated assessments based on data collected from sensors and transformed into high-level indicators. Regarding correlations between several indicators, Binkley et al. [33] found a high correlation between total time, number of sub-movements, path length and level of expertise. In contrast, Hofstad et al. [44] found that bimanual dexterity and EDA do not correlate with any other indicator. Regarding correlations between indicators and rating scales, Estrada et al. [118] found that the number of sub-movements, sub-movement duration, smoothness and spectral arc length correlate well with the score obtained with GRS. In addition, Varas et al. [98] observed a strong correlation between path length and the score obtained with OSATS. Moreover, Dubin et al. [36] found strong correlations between total time and GEARS Efficiency, as well as between economy of movement and GEARS Depth Perception. It is worth noting that no correlation has been found between acceleration values and rating scales [44].

Another possible strategy is to create **classifiers** or **prediction models**. Classifiers are typically used to differentiate between levels of expertise in surgeons, while prediction models are typically used to get an objective score for the surgeon's performance. Regarding **classifiers** and the indicators used (next subsection will specifically address the statistical methods and algorithms used), spectral arc length, bimanual dexterity, number of sub-movements, curvature, turning angle, tortuosity and velocity gain factor have shown to be useful in the classification of surgeons' levels of expertise [8][38][41][56][63]. In contrast, fixation duration, fixation rate, fixation number, saccade number, saccade duration, and saccade amplitude did not show very good results to classify levels of expertise [105]. In addition, other indicators that are not derived from data collected by sensors can be included in classification models, such as years of experience [8] or years of study of the surgeon [25], GRS real score [47], OSATS real score [24], GEARS real score [36], number of procedures made [28], and time and number of errors made [86]. Regarding **prediction models** and the indicators used, indicators can be combined with some weights to obtain scores about surgical technical skills. For example, Brown et al. [73] used force and acceleration indicators to predict the score obtained with GEARS. Linear and angular path lengths, bimanual dexterity from position, velocity values, acceleration values, economy of area, linear and angular velocity, number of sub-movements from velocity and acceleration, sub-movement duration from velocity and acceleration, spectral arc length from acceleration, smoothness, forces values, total time, and visible time have shown to be useful to predict scores obtained with GRS [25][62][75][89][99], OSATS [45][63], and GEARS [73].

3.3. RQ3: Which statistical methods and algorithms have been applied to data in relation to the assessment of surgical technical skills?

The data collected by the sensors and transformed into indicators are subjected to advanced processing, as seen in the previous subsection, typically for two purposes related to the objective and automated assessment of surgical technical skills: **1) classification** (between two or more surgeons' levels of expertise), and **2) prediction** (of scores). There are several statistical methods and (machine learning) algorithms applied on the data collected (low-level data/metrics and high-level data/indicators). Table 10 shows the statistical methods and algorithms found in the 101 articles identified organized according to the two above mentioned purposes.

It is important to note that some of the statistical methods and algorithms for classification and prediction rely on preliminary steps related to pre-processing the data collected. These include: 1) *feature extraction* (to collect the most important data and reduce the dimensionality of the data) using e.g., PCA (Principal Component Analysis) [26]; 2) *data normalization* (in case data are scaled differently) using e.g., Z-normalization [86]; and 3) *feature selection* (to collect the most relevant features/indicators, excluding those which may cause noise), using e.g., mRMR (Minimum Redundancy Maximum Relevance) [21] or Linear Correlation Coefficients [26]. The pre-processing of the data, although may be of interest, is not the focus of this article.

Regarding **classification**, *SVM (Support Vector Machine)* (12 articles, including also *SVR – Support Vector Regression*), *neural networks* (11 articles), *discriminant analysis* (9 articles), *Hidden Markov models* (6 articles), *logistic regression* (5 articles), *k-NN (k-nearest neighbors)* (5 articles), *Naïve Bayes* (3 articles), *Random forest* (3 articles), and *SAX (Symbolic Aggregate approximation)* (3 articles) are the most used statistical methods and algorithms. The use of these statistical methods and algorithms is subject to four main challenges that may have an impact on the accuracy of the results obtained [16] (this accuracy is typically calculated through scores obtained in rating scales or through the level of expertise reported by the surgeon in a survey). The first challenge refers to the **quality of the input data**, which depends on the type of sensors used (as these may contain errors and even affect the surgeon's motion range) and the indicators selected (as these may be irrelevant to assess a surgical technical skill or contain errors derived from the computation and normalization processes) [16]. For example, for the same acceleration and orientation indicators, Brown et al. [73] obtained an accuracy of 50% in their classifier with *random forest*, collecting data from IMUs and force sensors with 3 DoF, while Nguyen et al. [90] obtained an accuracy of 97.3% with *neural networks*, and 80.3% with *SVM*, collecting data from IMUs with 6 DoF and EM sensors with 3 DoF. The second challenge refers to the **dimensionality of the dataset**. There are indicators which may reduce the accuracy of the classifier by adding noise, while some others may overfit the dataset causing an accuracy larger that the true one of the algorithm [16], hence the importance of proper feature selection. For example, Ershad et al. [8] used indicators such as path length, velocity values, number of sub-movements, angular velocity, acceleration values, smoothness, EMG and EDA, and obtained an accuracy of 97% with *SVM* and 100% with *Naïve Bayes* when classifying surgeons between two levels of expertise. This accuracy was reduced to 84% with *SVM* and 89% with *Naïve Bayes* when classifying surgeons between four levels of expertise. As another example, Zhou et al. [87] used multiple biometric signals obtained from EEG, EMG, EDA, HRV, as well as wrist motion, to increase the dimensionality of the dataset, obtaining a normalized Median Absolute Error (nMAE) for accuracy of 11.05% using a variation of *SVM* called *SVR (Support Vector Regression)*. The third challenge refers to the **dependency on the task performed** (hence the importance of standardizing the types of exercises performed by the surgeon). For example, Winkler et al. [50] used indicators based on position, velocity and acceleration, obtaining an accuracy of 90% with *k-NN*, 84% with *Naïve Bayes*, 78% with *discriminant analysis* and 76% with *SVM*, with the problem that their indicators were obtained from a Virtual Reality (VR) simulation game and cannot be used in real surgical settings. As another example, Gomez et al. [75] used indicators based on acceleration and force, obtaining an accuracy of 90% and even 100% (depending on the task) with *linear regression*. Finally, Tien et al. [110] found a relevant study in which pupil indicators could classify with an accuracy of 91.9% on simulated tasks and of 81% on in-vivo simulators using *neural networks*. The fourth challenge refers to the **training and testing sets used** [16]. For example, Oropesa et al. [80] used indicators based on time, position, velocity, and acceleration and obtained an accuracy of 78.2% with *SVM*, 71.7% with *adaptive neuro fuzzy inference system* (ANFIS), and 71% with *linear discriminant analysis* (LDA), although they used the performance from 10 surgeons as training set, and thus their results may be affected by the skills of these 10 surgeons. In the case of the testing set used, cross-validation is a common technique to estimate the accuracy of many classification models [87], although it is not an exact measurement in real scenarios.

Regarding **prediction**, *SVM* (3 articles, including also *SVR – Support Vector Regression*), *Linear Regression* (2 articles) and *Regression tree* (2 articles) are the most used statistical methods and algorithms. The use of these statistical methods and algorithms is also subject to three out of the four main challenges identified in the case of classification. The first challenge refers, once again, to the **quality of the input data**. For example, there are several research studies which obtained good results in their models when predicting GRS scores, OSATS scores and GEARS scores using different algorithms and indicators. Concerning the prediction of GRS scores with promising results, Zia et al. [53] used *SVR* and indicators based on position, velocity, and orientation to generate automated scale score reports that correlate with GRS scores available in a public dataset (0.61 average Spearman correlation coefficient); Ziesmann et al. [99] used *linear regression* and indicators such as path length (0.59 Pearson correlation) and number of sub-movements (0.52 Pearson correlation); and Kirby et al. [89] used *k-means* and also indicators such as path length and number of sub-movements collected from surgeons' wrists and elbows (0.85 Pearson correlation in the best case). Concerning the prediction of OSATS scores with promising results, Oquendo et al. [63] used LASSO (Least Absolute Shrinkage and Selection Operator) and regression tree in several models with indicators such as total time, path length, velocity values, angular path length, angular velocity, visible time, idle time, acceleration values and number of sub-movements (0.85 Pearson correlation in the best case); and Kowalewski et al. [24] obtained a 10% error with *neural networks* using indicators such as acceleration values, angular velocity and orientation time series. Concerning the prediction of GEARS scores with promising results, Dubin et al. [36] used *linear regression* and indicators such as total time, economy of movement, and range of work to predict not only the total score of the GEARS scale (0,69 Pearson correlation), but also each one of the six domains in which the scale is divided, obtaining very good results for some of them, such as depth perception (0.81 Pearson correlation) or efficiency (0.91 Pearson correlation). The second challenge refers to the **dependency on the task performed**. For example, Brown et al. [73] used indicators based on orientation, acceleration and force to predict GEARS scores with a regressor composed of *SVM*, *Elastic Net Regression*, *Regression* tree and *k-NN*, reaching an accuracy of

0.93 on GEARS Total Score, 0.93 on GEARS Efficiency, and 0.9 on GEARS force sensitivity, although they ran the experiment for one task and thus the results are difficult to extrapolate to real surgical settings [73]. As another example, Malpani et al. [57] used *linear regression* and indicators such as total time, path length, economy of area and number of sub-movements to predict OSATS scores (0.47 Spearman correlation in the best case), but once again for specific tasks (suture throw in two steps, suture throw in one step, first knot, any other knot in the task). The third challenge refers to the **training and testing sets used**. For example, for the training set to be used in prediction models, some surgeons declare to be experts, but they are uncomfortable with the sensors used while performing the tasks, which may eventually affect the outcomes of prediction models [25].

Table 10. Purpose, statistical methods / algorithms, and references (one paper may include several statistical methods / algorithms)).

Purpose	Statistical method / algorithm	# Papers	References
Classification	SVM	12	[8][16][24][26][38][41][50][55][64][80][87][90]
	Neural network	11	[16][24][26][48][49][59][68][80][86][90][110]
	Discriminant analysis	9	[16][26][35][41][50][79][80][95][114]
	Hidden Markov Models	6	[16][40][53][54][55][103]
	Logistic regression	5	[16][38][41][48][62]
	k-NN (k-nearest neighbors)	5	[35][38][43][50][53]
	Naïve Bayes	3	[8][16][50]
	Random forest	3	[35][73][105]
	SAX (Symbolic Aggregate approximation)	3	[39][40][43]
	Fuzzy	2	[16][80]
	Decision tree	2	[16][24]
	Linear regression	2	[70][75]
	DTW (Dynamic time warping)	2	[24][58]
	k-means	1	[69]
Prediction	SVM	3	[53][73][87]
	Linear regression	3	[37][45][99]
	Regression tree	2	[63][73]
	LASSO (Least Absolute Shrinkage and Selection Operator)	1	[63]
	Elastic Net Regression	1	[73]
	k-NN (k-nearest neighbors)	1	[73]
	Logistic regression	1	[37]
	k-means	1	[89]
	Neural network	1	[24]

3.4. RQ4: How has feedback been provided to trainees in the context of assessment of surgical technical skills?

Feedback, either real-time feedback or delayed feedback, is crucial in the assessment of surgical technical skills, since it allows surgeons to increase their awareness on the procedure undertaken and make changes to improve their (on-going and/or future) performance. Feedback closes the full life cycle of objective and automated assessment of surgical technical skills, providing (in an understandable manner) the output of the statistical methods and algorithms applied to the data collected by the sensors in IoT systems. Nevertheless, feedback is not the main focus of this article and, therefore, no additional specific related terms were included in the query used to perform this SLR. Nonetheless, the way in which the 101 articles identified deal with feedback was analyzed. Table 11 summarizes the feedback mechanisms found and the corresponding references. It can be seen that very few of the articles identified have addressed the issue of feedback.

Table 11. Type, mechanisms and references on feedback (one paper may include several feedback mechanisms).

Type	Mechanisms	# Papers	References
Visualization	Trajectories	5	[39][40][43][67][68]
	Bar charts	1	[68]
	Radar chart	1	[114]
Text	Lexicon of Words	1	[8]

The most common way of providing feedback is by offering visualizations to trainees. The most popular visualization shows the **trajectories** (5 articles) followed by novice and expert surgeons in 2D [68] and 3D [39][40][43][67] indicating with a different color the specific sub-movements of the novice surgeon. In addition, Uemura et al. [68] used **bar charts** showing the scores for several indicators, including the range in which each score was considered "good". Sugiyama et al. [114] used a triangle **radar chart** with three force parameters on the vertices to provide a visual feedback to trainees. Nevertheless, these charts are typically static and there is a need for research on how the indicators showed in these charts evolve over time with the number of repetitions by the same surgeon. A more dynamic representation would help not only to see the evolution of a surgeon with the number of repetitions but also the moment in which more repetitions do not lead to an improvement thus better calibrating the number of training sessions required.

As for the summary feedback in text format, Ershad et al. [8] matched indicators on surgeon's performance and a **lexicon of words**. They related number of sub-movements and angular velocity with "fluid", acceleration and smoothness with "smooth", smoothness in depth axis and maximum smoothness with "jittery", total time with "swift", EMG with "relaxed", EDA with "calm", velocity parameters with "wavering", and path length of each hand with "coordinated".

In view of the few articles identified which dealt with feedback, the articles discarded during the assessment of full texts as part of the article selection process (see Figure 1) were further reviewed with the aim to look for additional feedback mechanisms. Interestingly, Payne et al. [120] studied the effectiveness of including real-time vibrations in the surgical instruments when the surgeon exceeds a threshold force while performing a task. Moreover, Yamaguchi et al. [121] used a hexagon radar chart showing an indicator collected from surgeon's performance in each vertex (e.g., task time, left-hand instrument velocity, right-hand instrument velocity, etc.); these authors also generated a timeline showing certain events which occurred in the course of the procedure, such as instrument intersection.

All in all, there is an important research opportunity on how to provide (real-time and delayed) feedback as part of the assessment of surgical technical skills. Although this feedback may simply represent the indicators of the novice surgeon's performance in a textual or visual manner, it is advisable to compare his/her performance with that of an expert surgeon or against reference values. Furthermore, it is worth noting that the lack of standardization of exercises makes it difficult to provide feedback to the surgeon as feedback is highly dependent of the type of exercise (the standardization of the curriculum in several phases with models such as the pyramid training model [12] or the 5 FLS exercises [9] are some steps forward). Beyond the standardization of exercises, the identification of common errors per exercise, and the detection of common errors from sensors and indicators are still challenges to be addressed from a research perspective.

4. Discussion

The assessment of surgical technical skills, such as bimanual dexterity, depth perception, economy, efficiency, hand tremor, mind control, precision or respect for tissue, among others, must evolve from the subjective and non-scalable assessment by an expert surgeon who assigns a score based on rating scales (e.g., GRS, OSATS, GEARS) to novice surgeons, to a more objective and automated assessment of surgical technical skills through IoT systems that can track e.g., surgical instruments, surgeon's body and surgeon's eyes. This SLR has revised 101 articles and has detected important differences in the literature in relation to the sensors used, the data collected (and the corresponding indicators obtained from low-level data) and the statistical methods and algorithms applied. In addition, another research gap has been identified in terms of closing the full life cycle by providing appropriate feedback to the trainees, especially real-time and action-oriented feedback. Moreover, the challenge of objectively and automatically assessing surgical technical skills is even more complex due to the lack of standardization in the surgical tasks to be performed during this assessment [80] leading to a disparity in the results obtained in the literature which are sometimes difficult to compare or extrapolate to other tasks and surgical settings. More standardization efforts are needed, as is the case, for example, with the programs for specific procedures known as Fundamentals of Laparoscopic Surgery (FLS), Fundamentals of Endoscopic Surgery (FES) and Fundamentals Use of Surgical Energy (FUSE) [9], which try to extrapolate simulations to real scenarios.

Regarding the sensors used, there are some limitations which need to be addressed. First, it is necessary to **improve ergonomics** of the tools and sensors used to reach a wider range of motions [28], limited sometimes by the particular sensors used or by the need for wires to get power supply [71]. Second, it is necessary to use tools and sensors without problems with **line of sight** (in the case of e.g., optical sensors) [84], **interferences with metallic objects** (in the case of e.g., electromagnetic sensors) [71][65], or **friction forces and involuntary vibrations** (in the case of e.g., IMUs or force sensors) [54]. Third, it is important to get a **higher DoF** in sensors although this may also require more complex IoT systems due to the need of collecting and combining data of different types and with possibly different timestamps. Fourth, it is necessary to explore the potential of combining motion data with **biometric signals** (such as those collected from EEG, EMG, EDA or HRV), for which there are currently few research articles published. Finally, it is also necessary to explore the use of IoT systems to collect **position correction data**, which is currently obtained mainly through cameras and image processing [119]. All these lines of research are aimed at obtaining more and better-quality raw data through IoT systems.

Regarding the indicators used, the articles published in the literature have focused mainly on motion of surgical instruments/tools and surgeon's body (position, velocity, acceleration, orientation and force), surgeon's eyes (fixation, saccade, smooth pursuit, pupils, and others such as blinking), and others (e.g., total time and biometric indicators). Several of the indicators analyzed in the literature can be used to distinguish between levels of expertise, sometimes between three levels of expertise (novice, intermediate, advanced), and sometimes between two of them. An important challenge is that some indicators in isolation may present problems due to, for example, uncontrolled muscle movements (e.g., related to hand tremor in older expert surgeons [52][92]), high amplitude movements from changes in position [99], surrounding light (and thus affecting pupil dilation) [107], etc. Therefore, it is necessary to look for **more controlled surgical settings**, and possibly **combine several indicators** to better distinguish between levels of expertise. Another important challenge is that the results obtained are very dependent on the task. It has been observed that the significant differences between levels of expertise are more frequently found with complex tasks [83] (rather than with basic tasks [113]), and in realistic surgical environments. Therefore, it is necessary to perform **more complex tasks in real scenarios** with patients to better distinguish between levels of expertise [64]. All these lines of research are aimed at (1) defining better experimental settings and (2) better low-level indicators (and combinations of these), to feed the statistical methods and algorithms in charge of classifying the surgeons in levels of expertise for different surgical technical skills and predicting their scores in rating scales. Low-level indicators mentioned in this SLR have shown to be useful to predict certain skills in simple tasks, but more complex exercises are necessary to collect other type of low-level indicators for more complex measurements, such as the approach angle of a certain tool or the tension of the thread in a suture, that allow predicting specific high-level skills required for complex interventions.

Regarding the statistical methods and algorithms used, some of them get a high accuracy between the classified and actual level of expertise [8][90], and some of them get a high accuracy between the predicted score of a surgeon and the real one according to rating scales [73], but, in general, there is a lack of comparison between the classification and prediction models used in the literature. The statistical methods and algorithms used to classify surgeons into levels of expertise (e.g., novice/advanced surgeons) were predominant over the algorithms used to predict a quantitative variable such as scores (GRS/OSATS/GEARS); although knowing the score is more precise, it is usually enough to classify the surgeon according to his/her level of expertise. There are a number of related limitations which need to be addressed from a research perspective. Firstly, a great number of studies use **small sample sizes** (number of participants in the study), which makes it difficult to generalize the models [86][107]. Secondly, in some cases, **samples are unbalanced**, meaning that there are e.g., too many novice surgeons and a few expert surgeons [75]; in order to get more samples some studies used data from several performances of the same surgeons, although there is a risk that participants perform with higher accuracies and lower reproducibility in subsequent attempts [80]. Thirdly, the **actual level of expertise of the surgeon** (to be compared with the output of classification models) is based on tags and thresholds [109], and this is not always a reliable approach (e.g., the number of cases performed is not always the best method to assess the level of expertise of the surgeon). Fourthly, **distinguishing between intermediate surgeons and advanced/expert surgeons** for a certain surgical technical skill is complicated because surgeons from both levels of expertise could have reached the highest point in their learning curve [98]. Fifthly, surgeons can suffer the **Hawthorne effect** (modifications in their behavior due to being observed) during assessment activities, which may have a negative influence in their results [106]. All in all, the lines of research should focus on improving the quality of the input data (both the data collected by sensors and the actual data used to compare the accuracy of classifiers and prediction models), and the comparisons between the proposed classification and prediction models. For example, the application of deep learning techniques could improve the accuracy of predictions, although very few articles have used deep learning techniques [48][49][90].

Finally, regarding feedback, it is necessary to prioritize research studies on feedback provision as part of the assessment of surgical technical skills. There are only a few examples of delayed feedback, such as **summaries** of the actions taken, either through visualizations [39][40][43][68] or texts [8] and some pioneering work in **real-time feedback** through vibrations in the surgical instruments when exceeding a threshold force [120]. All in all, the lines of research on feedback should focus on real-time feedback related to task correction during the realization of a surgical procedure, and summary visual feedback including group comparative results. For example, the assessment of surgical technical skills can benefit from previous research on learning analytics related to feedback provision and visualizations, adapting advances from other domains.

5. Conclusions

The objective and automated assessment of surgical technical skills is a growing research field motivated by the rapid evolution of technologies, in particular of IoT systems that include sensors which collect different types of data to be processed with statistical methods and (machine learning) algorithms. In this context, this SLR is relevant since it presents a comprehensive and updated overview of the sensors used, the indicators extracted from raw data (metrics), and the statistical methods and algorithms applied on these metrics and indicators, based on the analysis of 101 related articles published after 2013. Regarding sensors, there are research opportunities, for instance, related to the improvement of motion sensors (e.g., reduction of weight or increase of degrees of freedom) and force sensors (e.g., reduction of friction forces and vibrations), as well as in the realization of more studies focused on the use of biometric sensors to measure the surgeon's cognitive load during a surgical procedure. Regarding indicators, there are research opportunities, for instance, related to the realization of more studies in which surgeons perform different tasks in different environment (as both tasks and environments can have an effect on the distinction between levels of expertise and there is currently a lack of standardized surgical simulation tasks), and in the combination of several indicators to create better classifiers and prediction models. Regarding statistical methods and algorithms, there are research opportunities, for instance, related to improving the quality of the input data and the metrics to evaluate the algorithms, such as accuracy in classifiers and prediction models. Finally, regarding feedback, there is considerable room for improvement related to the provision of real-time and delayed feedback through appropriate visualizations and/or texts in order to improve the learning experience.

In addition, there are two important lessons learned from this SLR. First, most of the research conducted so far has focused on adding sensors on instruments and surgeons rather on sensing the specific elements of the surgical task, which is key to objectively and automatically assess surgeon's performance in both formative and summative assessment. However, this requires standardizing the tasks, identifying relevant indicators and adding sensors that serve to obtain these relevant indicators. For example, FLS has taken a step forward in the standardization of tasks but the identification of relevant indicators and their automatic calculation is still missing [9]. Second, the assessment of surgical technical skills is dynamic. The absolute value of indicators is not as relevant as their evolution in time with the number of repetitions. There is a research opportunity in the use of algorithms that use time series, such as dynamic time warping [58][122], and the provision of feedback through appropriate visualizations.

It is worth noting that this work has some limitations. Firstly, it focuses on a very specific field of surgery which is the (summative) assessment of surgical skills. Nevertheless, most of the findings related to the use of sensors, indicators, statistical methods and algorithms can also be used in the training (formative assessment) of novice surgeons and in the professional practice or intermediates and advanced surgeons. In addition, this SLR is limited by the keywords that comprise the query used, which could have left out relevant articles, and by the fact that articles not published after 2013 were not included in the SLR (as these had been previously analyzed in three doctoral dissertations [19][20][21]). In any case, we believe that the conclusions obtained from an SLR in which 537 were screened will contribute to advance the research in the field and draw attention to scientific debates on the objective and automated assessment of surgical technical skills.

Conflict of interest statement

Declarations of interest: none

Acknowledgements

This work was supported in part by the FEDER/Ministerio de Ciencia, Innovación y Universidades–Agencia Estatal de Investigación, through the Smartlet Project under Grant TIN2017-85179-C3-1-R, and in part by the Madrid Regional Government through the e-Madrid-CM Project under Grant S2018/TCS-4307, a project which is co-funded by the European Structural Funds (FSE and FEDER). Partial support has also been received from the European Commission through Erasmus+ Capacity Building in the Field of Higher Education projects, more specifically through projects LALA (586120-EPP-1-2017-1-ES-EPPKA2-CBHE-JP), InnovaT (598758-EPP-1-2018-1-AT-EPPKA2-CBHE-JP), and PROF-XXI (609767-EPP-1-2019-1-ES-EPPKA2-CBHE-JP). This publication reflects the views only of the authors and funders cannot be held responsible for any use which may be made of the information contained therein.

References

[1] Seale, J., Knoetze, M., Phung, A., Prior, D., & Butchers, C. (2019). Commencing Technical Clinical Skills Training in the Early Stages of Medical Education: Exploring Student Views. *Medical Science Educator, 29*(1), 173-179.

[2] Walter, A. J. (2006). Surgical education for the twenty-first century: beyond the apprentice model. *Obstetrics and Gynecology Clinics, 33*(2), 233-236.

[3] Moorthy, K., Munz, Y., Sarker, S. K., & Darzi, A. (2003). Objective assessment of technical skills in surgery. *British Medical Journal, 327*(7422), 1032-1037.

[4] Reznick, R., Regehr, G., MacRae, H., Martin, J., & McCulloch, W. (1997). Testing technical skill via an innovative "bench station" examination. *The American Journal of Surgery, 173*(3), 226-230.

[5] Niitsu, H., Hirabayashi, N., Yoshimitsu, M., Mimura, T., Taomoto, J., Sugiyama, Y., ... & Takiyama, W. (2013). Using the Objective Structured Assessment of Technical Skills (OSATS) global rating scale to evaluate the skills of surgical trainees in the operating room. *Surgery today, 43*(3), 271-275.

[6] Goh, A. C., Goldfarb, D. W., Sander, J. C., Miles, B. J., & Dunkin, B. J. (2012). Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. *The Journal of urology, 187*(1), 247-252.

[7] Vassiliou, M. C., Kaneva, P. A., Poulose, B. K., Dunkin, B. J., Marks, J. M., Sadik, R., ... & Hazey, J. W. (2010). Global Assessment of Gastrointestinal Endoscopic Skills (GAGES): a valid measurement tool for technical skills in flexible endoscopy. *Surgical endoscopy, 24*(8), 1834-1841.

[8] Ershad, M., Rege, R., & Fey, A. M. (2018). Meaningful assessment of robotic surgical style using the wisdom of crowds. *International journal of computer assisted radiology and surgery, 13*(7), 1037-1048.

[9] Sroka, G., Feldman, L. S., Vassiliou, M. C., Kaneva, P. A., Fayez, R., & Fried, G. M. (2010). Fundamentals of laparoscopic surgery simulator training to proficiency improves laparoscopic performance in the operating room—a randomized controlled trial. *The American Journal of Surgery, 199*(1), 115-120.

[10] Nguyen, L., Brunicardi, F. C., DiBardino, D. J., Scott, B. G., Awad, S. S., Bush, R. L., & Brandt, M. L. (2006). Education of the modern surgical resident: novel approaches to learning in the era of the 80-hour workweek. *World Journal of Surgery, 30*(6), 1120-1127.

[11] Douglas, H. E., & Mackay, I. R. (2011). Microvascular surgical training models. *Journal of Plastic, Reconstructive & Aesthetic Surgery, 64*(8), e210-e212.

[12] Usón-Gargallo, J., Pérez-Merino, E. M., Usón-Casaús, J. M., Sánchez-Fernández, J., & Sánchez-Margallo, F. M. (2013). Pyramid training model in laparoscopic surgery. *Cirugía y cirujanos, 81*(5), 420-430.

[13] Tan, S. S. Y., & Sarker, S. K. (2011). Simulation in surgery: a review. *Scottish Medical Journal, 56*(2), 104-109.

[14] de Montbrun, S. L., & MacRae, H. (2012). Simulation in surgical education. *Clinics in colon and rectal surgery, 25*(3), 156-165.

[15] Liss, M. A., & McDougall, E. M. (2013). Robotic surgical simulation. *The Cancer Journal, 19*(2), 124-129.

[16] Winkler-Schwartz, A., Bissonnette, V., Mirchi, N., Ponnudurai, N., Yilmaz, R., Ledwos, N., ... & Del Maestro, R. F. (2019). Artificial intelligence in medical education: best practices using machine learning to assess surgical expertise in virtual reality simulation. *Journal of surgical education, 76*(6), 1681-1690.

[17] Ray, P. P. (2018). A survey on Internet of Things architectures. *Journal of King Saud University-Computer and Information Sciences, 30*(3), 291-319.

[18] Chen, J., Cheng, N., Cacciamani, G., Oh, P., Lin-Brande, M., Remulla, D., ... & Hung, A. J. (2019). Objective assessment of robotic surgical technical skill: a systematic review. *The Journal of urology, 201*(3), 461-469.

[19] Sánchez Margallo, J. A. (2014). *Desarrollo y validación de sistemas de seguimiento de instrumental laparoscópico para la evaluación objetiva de destrezas quirúrgicas* (*Development and validation of monitoring systems for laparoscopic instruments for objective assessment of surgical skills*) (Doctoral dissertation).

[20] Enciso Sanz, S. (2013). *Evaluación de la adquisición de destrezas y habilidades quirúrgicas durante la formación en cirugía laparoscópica* (*Evaluation of the acquisition of surgical skills and abilities during training in laparoscopic surgery*) (Doctoral dissertation).

[21] Kyaw, T. Z. (2013). *Surgical Skills and Ergonomics Evaluation for Laparoscopic Surgery Training* (Doctoral dissertation).

[22] Al-Shahrestani, F., Sørensen, M. S., & Andersen, S. A. W. (2019). Performance metrics in mastoidectomy training: a systematic review. *European Archives of Oto-Rhino-Laryngology, 276*(3), 657-664.

[23] Levin, M., McKechnie, T., Khalid, S., Grantcharov, T. P., & Goldenberg, M. (2019). Automated methods of technical skill assessment in surgery: A systematic review. *Journal of surgical education, 76*(6), 1629-1639.

[24] Kowalewski, K. F., Garrow, C. R., Schmidt, M. W., Benner, L., Müller-Stich, B. P., & Nickel, F. (2019). Sensor-based machine learning for workflow detection and as key to detect expert level in laparoscopic suturing and knot-tying. *Surgical Endoscopy, 33*(11), 3732-3740.

[25] Arbelaez-Garces, G., Joseph, D., Camargo, M., Tran, N., & Morel, L. (2018). Contribution to the objective assessment of technical skills for surgery students: An accelerometer based approach. *International Journal of Industrial Ergonomics, 64*, 79-88.

[26] Sbernini, L., Quitadamo, L. R., Riillo, F., Di Lorenzo, N., Gaspari, A. L., & Saggio, G. (2018). Sensory-glove-based open surgery skill evaluation. *IEEE Transactions on Human-Machine Systems, 48*(2), 213-218.

[27] Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Prisma Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med, 6*(7), e1000097, 1-6.

[28] Våpenstad, C., Fagertun Hofstad, E., Eivind Bernstein, T., Aadahl, P., Johnsen, G., & Mårvik, R. (2020). Optimal timing of assessment tasks depending on experience level of surgical trainees. *Minimally Invasive Therapy & Allied Technologies, 29*(3), 161-169.

[29] Nguyen, J. H., Chen, J., Marshall, S. P., Ghodoussipour, S., Chen, A., Gill, I. S., & Hung, A. J. (2020). Using objective robotic automated performance metrics and task-evoked pupillary response to distinguish surgeon expertise. *World journal of urology, 38*(7), 1599-1605.

[30] Rose, K., & Pedowitz, R. (2015). Fundamental arthroscopic skill differentiation with virtual reality simulation. *Arthroscopy: The Journal of Arthroscopic & Related Surgery, 31*(2), 299-305.

[31] Ahmmad, S. N. Z., Ming, E. S. L., Fai, Y. C., & Narayanan, A. L. T. (2014). Experimental Study of Surgeon's Psychomotor Skill Using Sensor-Based Measurement. *Procedia Computer Science, 42*, 130-137.

[32] Ahmmad, S. N. Z., Ming, E. S. L., Fai, Y. C., Sood, S., & Gandhi, A. (2016). Objective measurement for surgical skill evaluation. *Jurnal Teknologi*, 78(7-5), 145-152.

[33] Binkley, J., Bukoski, A. D., Doty, J., Crane, M., Barnes, S. L., & Quick, J. A. (2019). *Surgical simulation: markers of proficiency. Journal of surgical education, 76*(1), 234-241.

[34] Connolly, M., Seligman, J., Kastenmeier, A., Goldblatt, M., & Gould, J. C. (2014). Validation of a virtual reality-based robotic surgical skills curriculum. *Surgical endoscopy, 28*(5), 1691-1694.

[35] Dockter, R. L., Lendvay, T. S., Sweet, R. M., & Kowalewski, T. M. (2017). The minimally acceptable classification criterion for surgical skill: intent vectors and separability of raw motion data. *International journal of computer assisted radiology and surgery, 12*(7), 1151-1159.

[36] Dubin, A. K., Smith, R., Julian, D., Tanaka, A., & Mattingly, P. (2017). A comparison of robotic simulation performance on basic virtual reality skills: simulator subjective versus objective assessment tools. *Journal of Minimally Invasive Gynecology, 24*(7), 1184-1189.

[37] Dubin, A. K., Julian, D., Tanaka, A., Mattingly, P., & Smith, R. (2018). A model for predicting the GEARS score from virtual reality surgical simulator metrics. *Surgical endoscopy, 32*(8), 3576-3581.

[38] Fard, M. J., Ameri, S., Darin Ellis, R., Chinnam, R. B., Pandya, A. K., & Klein, M. D. (2018). Automated robot‐ assisted surgical skill evaluation: Predictive analytics approach. *The International Journal of Medical Robotics and Computer Assisted Surgery, 14*(1), e1850, 1-10.

[39] Forestier, G., Petitjean, F., Senin, P., Despinoy, F., & Jannin, P. (2017, June). Discovering discriminative and interpretable patterns for surgical motion analysis. In *Conference on Artificial Intelligence in Medicine in Europe* (pp. 136-145). Springer, Cham.

[40] Forestier, G., Petitjean, F., Senin, P., Despinoy, F., Huaulmé, A., Fawaz, H. I., ... & Jannin, P. (2018). Surgical motion analysis using discriminative interpretable patterns. *Artificial intelligence in medicine, 91*, 3-11.

[41] French, A., Lendvay, T. S., Sweet, R. M., & Kowalewski, T. M. (2017). Predicting surgical skill from the first N seconds of a task: value over task time using the isogony principle. *International journal of computer assisted radiology and surgery, 12*(7), 1161-1170.

[42] Ghasemloonia, A., Maddahi, Y., Zareinia, K., Lama, S., Dort, J. C., & Sutherland, G. R. (2017). Surgical skill assessment using motion quality and smoothness. *Journal of surgical education, 74*(2), 295-305.

[43] Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. A. (2018, September). Evaluating surgical skills from kinematic data using convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 214-221). Springer, Cham.

[44] Hofstad, E. F., Våpenstad, C., Bø, L. E., Langø, T., Kuhry, E., & Mårvik, R. (2017). Psychomotor skills assessment by motion analysis in minimally invasive surgery on an animal organ. *Minimally Invasive Therapy & Allied Technologies, 26*(4), 240-248.

[45] Malpani, A., Vedula, S. S., Chen, C. C. G., & Hager, G. D. (2014, June). Pairwise comparison-based objective score for automated skill assessment of segments in a surgical task. In *International Conference on Information Processing in Computer-Assisted Interventions* (pp. 138-147). Springer, Cham.

[46] Rudderow, J., Bansal, J., Wearne, S., Lara-Torre, E., Paget, C., & Ferrara, J. (2014). Development of a web-based laparoscopic technical skills assessment and testing instrument: a pilot study. *Journal of surgical education, 71*(6), e73-e78.

[47] Vedula, S. S., Malpani, A., Ahmidi, N., Khudanpur, S., Hager, G., & Chen, C. C. G. (2016). Task-level vs. segment-level quantitative metrics for surgical skill assessment. *Journal of surgical education, 73*(3), 482-489.

[48] Wang, Z., & Fey, A. M. (2018). Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. *International journal of computer assisted radiology and surgery, 13*(12), 1959-1970.

[49] Wang, Z., & Fey, A. M. (2018, July). SATR-DL: improving surgical skill assessment and task recognition in robot-assisted surgery with deep neural networks. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 1793-1796). IEEE.

[50] Winkler-Schwartz, A., Yilmaz, R., Mirchi, N., Bissonnette, V., Ledwos, N., Siyar, S., ... & Del Maestro, R. (2019). Machine learning identification of surgical and operative factors associated with surgical expertise in virtual reality simulation. *JAMA network open, 2*(8), e198363, 1-16.

[51] Zahedi, E., Rahmat-Khah, H., Dargahi, J., & Zadeh, M. (2017, March). Virtual reality based training: Evaluation of user performance by capturing upper limb motion. In *2017 IEEE Virtual Reality (VR)* (pp. 255-256). IEEE.

[52] Ahmmad, S. N. Z., Ming, E. S. L., Fai, Y. C., Sood, S., Gandhi, A., Mohamed, N. S., ... & Burdet, E. (2019). Objective assessment of surgeon's psychomotor skill using virtual reality module. *Indonesian Journal of Electrical Engineering and Computer Science, 14*(3), 1533-1543.

[53] Zia, A., & Essa, I. (2018). Automated surgical skill assessment in RMIS training. *International journal of computer assisted radiology and surgery, 13*(5), 731-739.

[54] Rafii-Tari, H., Payne, C. J., Liu, J., Riga, C., Bicknell, C., & Yang, G. Z. (2015, May). Towards automated surgical skill evaluation of endovascular catheterization tasks based on force and motion signatures. In *2015 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 1789-1794). IEEE.

[55] 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. *International journal of computer assisted radiology and surgery, 10*(6), 981-991.

[56] Duran, C., Estrada, S., O'Malley, M., Lumsden, A. B., & Bismuth, J. (2015). Kinematics effectively delineate accomplished users of endovascular robotics with a physical training model. *Journal of Vascular Surgery, 61*(2), 535-541.

[57] Fuerst, D., Hollensteiner, M., & Schrempf, A. (2015, August). Assessment parameters for a novel simulator in minimally invasive spine surgery. In *2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (pp. 5110-5113). IEEE.

[58] Jiang, J., Xing, Y., Wang, S., & Liang, K. (2017). Evaluation of robotic surgery skills using dynamic time warping. *Computer Methods and Programs in Biomedicine, 152*, 71-83.

[59] Kramer, B. D., Losey, D. P., & O'Malley, M. K. (2016). SOM and LVQ classification of endovascular surgeons using motion-based metrics. In *Advances in Self-Organizing Maps and Learning Vector Quantization* (pp. 227-237). Springer, Cham.

[60] Liang, K., Xing, Y., Li, J., Wang, S., Li, A., & Li, J. (2018). Motion control skill assessment based on kinematic analysis of robotic end- effector movements. *The International Journal of Medical Robotics and Computer Assisted Surgery, 14*(1), e1845, 1-9.

[61] Oh, C. J., Tripathi, P. B., Gu, J. T., Borden, P., & Wong, B. J. F. (2019). Development and evaluation of rhinoplasty spreader graft suture simulator for novice surgeons. *The Laryngoscope, 129*(2), 344-350.

[62] O'Malley, M. K., Byrne, M. D., Estrada, S., Duran, C., Schulz, D., & Bismuth, J. (2019). Expert surgeons can smoothly control robotic tools with a discrete control interface. *IEEE Transactions on Human-Machine Systems, 49*(4), 388-394.

[63] Oquendo, Y. A., Riddle, E. W., Hiller, D., Blinman, T. A., & Kuchenbecker, K. J. (2018). Automatically rating trainee skill at a pediatric laparoscopic suturing task. *Surgical endoscopy, 32*(4), 1840-1857.

[64] Rafii-Tari, H., Payne, C. J., Bicknell, C., Kwok, K. W., Cheshire, N. J., Riga, C., & Yang, G. Z. (2017). Objective assessment of endovascular navigation skills with force sensing. *Annals of biomedical engineering, 45*(5), 1315-1327.

[65] Ross, P. D., Steven, R., Zhang, D., Li, H., & Abel, E. W. (2015). Computer-assessed performance of psychomotor skills in endoscopic otolaryngology surgery: construct validity of the Dundee Endoscopic Psychomotor Otolaryngology Surgery Trainer (DEPOST). *Surgical endoscopy, 29*(11), 3125-3131.

[66] Takazawa, S., Ishimaru, T., Harada, K., Deie, K., Fujishiro, J., Sugita, N., ... & Iwanaka, T. (2016). Pediatric thoracoscopic surgical simulation using a rapid-prototyped chest model and motion sensors can better identify skilled surgeons than a conventional box trainer. *Journal of Laparoendoscopic & Advanced Surgical Techniques, 26*(9), 740-747.

[67] Uemura, M., Tomikawa, M., Kumashiro, R., Miao, T., Souzaki, R., Ieiri, S., ... & Hashizume, M. (2014). Analysis of hand motion differentiates expert and novice surgeons. *Journal of surgical research, 188*(1), 8-13.

[68] Uemura, M., Tomikawa, M., Miao, T., Souzaki, R., Ieiri, S., Akahoshi, T., ... & Hashizume, M. (2018). Feasibility of an AI-based measure of the hand motions of expert and novice surgeons. *Computational and mathematical methods in medicine, 2018*, 1-6.

[69] Weede, O., Möhrle, F., Wörn, H., Falkinger, M., & Feussner, H. (2014, November). Movement analysis for surgical skill assessment and measurement of ergonomic conditions. In *2014 2nd International Conference on Artificial Intelligence, Modelling and Simulation* (pp. 97-102). IEEE.

[70] Evans, R. L., Partridge, R. W., & Arvind, D. K. (2014). Demonstration Paper: A Comparative Study of Surgical Skills Assessment in a Physical Laparoscopy Simulator Using Wireless Inertial Sensors. In *Proceedings of the Wireless Health 2014 on National Institutes of Health* (pp. 1-8).

[71] Harada, K., Morita, A., Minakawa, Y., Baek, Y. M., Sora, S., Sugita, N., ... & Mitsuishi, M. (2015). Assessing microneurosurgical skill with medico-engineering technology. *World neurosurgery, 84*(4), 964-971.

[72] Genovese, B., Yin, S., Sareh, S., DeVirgilio, M., Mukdad, L., Davis, J., ... & Benharash, P. (2016). Surgical hand tracking in open surgery using a versatile motion sensing system: are we there yet?. *The American Surgeon, 82*(10), 872-875.

[73] Brown, J. D., O'Brien, C. E., Leung, S. C., Dumon, K. R., Lee, D. I., & Kuchenbecker, K. J. (2016). Using contact forces and robot arm accelerations to automatically rate surgeon skill at peg transfer. *IEEE Transactions on Biomedical Engineering, 64*(9), 2263-2275.

[74] Fahy, A. S., Fok, K. H., Gavrilovic, B., Farcas, M., Carrillo, B., Gerstle, J. T., & Azzie, G. (2018). Refinement in the analysis of motion within low-cost laparoscopic simulators of differing size: implications on assessing technical skills. *Journal of pediatric surgery, 53*(12), 2480-2487.

[75] Gomez, E. D., Aggarwal, R., McMahan, W., Bark, K., & Kuchenbecker, K. J. (2016). Objective assessment of robotic surgical skill using instrument contact vibrations. *Surgical endoscopy, 30*(4), 1419-1431.

[76] Nasr, A., Carrillo, B., Gerstle, J. T., & Azzie, G. (2014). Motion analysis in the pediatric laparoscopic surgery (PLS) simulator: validation and potential use in teaching and assessing surgical skills. *Journal of Pediatric Surgery, 49*(5), 791-794.

[77] Pourkand, A., Salas, C., Regalado, J., Bhakta, K., Tufaro, R., Mercer, D., & Grow, D. (2016). Objective evaluation of motor skills for orthopedic residents using a motion tracking drill system: outcomes of an ABOS approved surgical skills training program. *The Iowa orthopaedic journal, 36*, 13-19.

[78] Fransson, B. A., Chen, C. Y., Noyes, J. A., & Ragle, C. A. (2016). Instrument motion metrics for laparoscopic skills assessment in virtual reality and augmented reality. *Veterinary surgery, 45*(S1), O5-O13.

[79] Horeman, T., Dankelman, J., Jansen, F. W., & van den Dobbelsteen, J. J. (2013). Assessment of laparoscopic skills based on force and motion parameters. *IEEE Transactions on Biomedical Engineering, 61*(3), 805-813.

[80] Oropesa, I., Sánchez-González, P., Chmarra, M. K., Lamata, P., Pérez-Rodríguez, R., Jansen, F. W., ... & Gómez, E. J. (2014). Supervised classification of psychomotor competence in minimally invasive surgery based on instruments motion analysis. *Surgical endoscopy, 28*(2), 657-670.

[81] Sánchez-Margallo, J. A., Sánchez-Margallo, F. M., Carrasco, J. B. P., García, I. O., Aguilera, E. J. G., & del Pozo, J. M. (2014). Usefulness of an optical tracking system in laparoscopic surgery for motor skills assessment. *Cirugía Española (English Edition), 92*(6), 421-428.

[82] Sánchez-Margallo, J. A., Sánchez-Margallo, F. M., Oropesa, I., Enciso, S., & Gómez, E. J. (2017). Objective assessment based on motion-related metrics and technical performance in laparoscopic suturing. *International journal of computer assisted radiology and surgery, 12*(2), 307-314.

[83] Twijnstra, A. R. H., Hiemstra, E., van Zwet, E. W., Balkema, E. I. R., Dankelman, J., & Jansen, F. W. (2014). Intracorporeal knot tying in a box trainer: how proficient is in vitro evaluation in laparoscopic experts?. *Journal of Minimally Invasive Gynecology, 21*(2), 291-295.

[84] Lahanas, V., Loukas, C., Georgiou, K., Lababidi, H., & Al-Jaroudi, D. (2017). Virtual reality-based assessment of basic laparoscopic skills using the Leap Motion controller. *Surgical endoscopy, 31*(12), 5012-5023.

[85] Sánchez-Margallo, J. A., Sánchez-Margallo, F. M., Oropesa, I., & Gómez, E. J. (2014). Systems and technologies for objective evaluation of technical skills in laparoscopic surgery. *Minimally Invasive Therapy & Allied Technologies, 23*(1), 40-51.

[86] Laverde, R., Rueda, C., Amado, L., Rojas, D., & Altuve, M. (2018, July). Artificial neural network for laparoscopic skills classification using motion signals from apple watch. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 5434-5437). IEEE.

[87] Zhou, T., Cha, J. S., Gonzalez, G. T., Wachs, J. P., Sundaram, C., & Yu, D. (2018, March). Joint Surgeon Attributes Estimation in Robot-Assisted Surgery. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 285-286).

[88] Gong, J., & Lach, J. (2015, October). Motion marker discovery from inertial body sensors for enhancing objective assessment of robotic surgical skills. In *2015 International Symposium on Bioelectronics and Bioinformatics (ISBB)* (pp. 215-218). IEEE.

[89] Kirby, G. S., Guyver, P., Strickland, L., Alvand, A., Yang, G. Z., Hargrove, C., ... & Rees, J. L. (2015). Assessing arthroscopic skills using wireless elbow-worn motion sensors. *The Journal of Bone and Joint Surgery, 97*(13), 1119-1127.

[90] 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. *Computer methods and programs in biomedicine, 177*, 1-8.

[91] Overby, D. W., & Watson, R. A. (2014). Hand motion patterns of Fundamentals of Laparoscopic Surgery certified and noncertified surgeons. *The American Journal of Surgery, 207*(2), 226-230.

[92] Rose, M., Curtze, C., O'Sullivan, J., El-Gohary, M., Crawford, D., Friess, D., & Brady, J. M. (2017). Wearable inertial sensors allow for quantitative assessment of shoulder and elbow kinematics in a cadaveric knee arthroscopy model. *Arthroscopy: The Journal of Arthroscopic & Related Surgery, 33*(12), 2110-2116.

[93] Sánchez, A., Rodríguez, O., Sánchez, R., Benítez, G., Pena, R., Salamo, O., & Baez, V. (2014). Laparoscopic surgery skills evaluation: analysis based on accelerometers. *JSLS: Journal of the Society of Laparoendoscopic Surgeons, 18*(4), 1-5.

[94] Viriyasiripong, S., Lopez, A., Mandava, S. H., Lai, W. R., Mitchell, G. C., Boonjindasup, A., ... & Lee, B. R. (2016). Accelerometer measurement of head movement during laparoscopic surgery as a tool to evaluate skill development of surgeons. *Journal of surgical education, 73*(4), 589-594.

[95] Huang, F. C., Mohamadipanah, H., Mussa-Ivaldi, F. A., & Pugh, C. M. (2019). Combining metrics from clinical simulators and sensorimotor tasks can reveal the training background of surgeons. *IEEE Transactions on Biomedical Engineering, 66*(9), 2576-2584.

[96] Mohamadipanah, H., Parthiban, C., Nathwani, J., Rutherford, D., DiMarco, S., & Pugh, C. (2016). Can a virtual reality assessment of fine motor skill predict successful central line insertion?. *The American Journal of Surgery, 212*(4), 573-578.

[97] Mohamadipanah, H., Parthiban, C., Law, K., Nathwani, J., Maulson, L., DiMarco, S., & Pugh, C. (2016, June). Hand smoothness in laparoscopic surgery correlates to psychomotor skills in virtual reality. In *2016 IEEE 13th International Conference on wearable and implantable body sensor networks (BSN)* (pp. 242-246). IEEE.

[98] Varas, J., Achurra, P., León, F., Castillo, R., De La Fuente, N., Aggarwal, R., ... & Montaña, R. (2016). Assessment of central venous catheterization in a simulated model using a motion-tracking device: an experimental validation study. *Annals of surgical innovation and research, 10*(1), 2, 1-5.

[99] Ziesmann, M. T., Park, J., Unger, B., Kirkpatrick, A. W., Vergis, A., Pham, C., ... & Gillman, L. M. (2015). Validation of hand motion analysis as an objective assessment tool for the Focused Assessment with Sonography for Trauma examination. *Journal of Trauma and Acute Care Surgery, 79*(4), 631-637.

[100] D'Angelo, A. L. D., Rutherford, D. N., Ray, R. D., Laufer, S., Kwan, C., Cohen, E. R., ... & Pugh, C. M. (2015). Idle time: an underdeveloped performance metric for assessing surgical skill. *The American Journal of Surgery, 209*(4), 645-651.

[101] D'Angelo, A. L. D., Rutherford, D. N., Ray, R. D., Laufer, S., Mason, A., & Pugh, C. M. (2016). Working volume: validity evidence for a motion-based metric of surgical efficiency. *The American Journal of Surgery, 211*(2), 445-450.

[102] Kholinne, E., Gandhi, M. J., Adikrishna, A., Hong, H., Kim, H., Hong, J., & Jeon, I. H. (2018). The dimensionless squared jerk: an objective parameter that improves assessment of hand motion analysis during simulated shoulder arthroscopy. *BioMed research international, 2018*, 1–8.

[103] Sun, X., Byrns, S., Cheng, I., Zheng, B., & Basu, A. (2017). Smart sensor-based motion detection system for hand movement training in open surgery. Journal of medical systems, 41(2), 24, 1-13.

[104] Eivazi, S., Hafez, A., Fuhl, W., Afkari, H., Kasneci, E., Lehecka, M., & Bednarik, R. (2017). Optimal eye movement strategies: a comparison of neurosurgeons gaze patterns when using a surgical microscope. *Acta neurochirurgica, 159*(6), 959-966.

[105] Eivazi, S., Slupina, M., Fuhl, W., Afkari, H., Hafez, A., & Kasneci, E. (2017, March). Towards automatic skill evaluation in microsurgery. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces Companion* (pp. 73-76).

[106] Erridge, S., Ashraf, H., Purkayastha, S., Darzi, A., & Sodergren, M. H. (2018). Comparison of gaze behaviour of trainee and experienced surgeons during laparoscopic gastric bypass. *British Journal of Surgery, 105*(3), 287–294.

[107] Gunawardena, N., Matscheko, M., Anzengruber, B., Ferscha, A., Schobesberger, M., Shamiyeh, A., ... & Solleder, P. (2019, June). Assessing surgeons' skill level in laparoscopic cholecystectomy using eye metrics. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications* (pp. 1-8).

[108] Menekse Dalveren, G. G., & Cagiltay, N. E. (2018). Insights from surgeons' eye-movement data in a virtual simulation surgical training environment: effect of experience level and hand conditions. *Behaviour & Information Technology, 37*(5), 517-537.

[109] Tien, T., Pucher, P. H., Sodergren, M. H., Sriskandarajah, K., Yang, G. Z., & Darzi, A. (2015). Differences in gaze behaviour of expert and junior surgeons performing open inguinal hernia repair. *Surgical endoscopy, 29*(2), 405-413.

[110] Tien, T., Pucher, P. H., Sodergren, M. H., Sriskandarajah, K., Yang, G. Z., & Darzi, A. (2014). Eye tracking for skills assessment and training: a systematic review. *Journal of surgical research, 191*(1), 169-178.

[111] Shahbazi, M., Poursartip, B., Siroen, K., Schlachta, C. M., & Patel, R. V. (2018, July). Robotics-Assisted Surgical Skills Evaluation based on Electrocortical Activity. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 3673-3676). IEEE.

[112] Ahmmad, S. N. Z., San, C. Z., Ming, E. S. L., & Fai, Y. C. (2015). Force variability as an objective measure of surgical skill. *Jurnal Teknologi, 74*(6), 125-128.

[113] Araki, A., Makiyama, K., Yamanaka, H., Ueno, D., Osaka, K., Nagasaka, M., ... & Yao, M. (2017). Comparison of the performance of experienced and novice surgeons: measurement of gripping force during laparoscopic surgery performed on pigs using forceps with pressure sensors. *Surgical endoscopy, 31(*4), 1999-2005.

[114] 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*(3), 234-242.

[115] Takayasu, K., Yoshida, K., Kinoshita, H., Yoshimoto, S., Oshiro, O., & Matsuda, T. (2018). Analysis of the tractive force pattern on a knot by force measurement during laparoscopic knot tying. *The American Journal of Surgery, 216*(2), 314-318.

[116] Lahanas, V., Loukas, C., Smailis, N., & Georgiou, E. (2015). A novel augmented reality simulator for skills assessment in minimal invasive surgery. *Surgical endoscopy, 29*(8), 2224-2234.

[117] Harada, K., Takazawa, S., Tsukuda, Y., Ishimaru, T., Sugita, N., Iwanaka, T., & Mitsuishi, M. (2015). Quantitative pediatric surgical skill assessment using a rapid-prototyped chest model. Minimally Invasive Therapy & Allied Technologies, 24(4), 226-232.

[118] Estrada, S., Duran, C., Schulz, D., Bismuth, J., Byrne, M. D., & O'Malley, M. K. (2016). Smoothness of surgical tool tip motion correlates to skill in endovascular tasks. *IEEE Transactions on Human-Machine Systems, 46*(5), 647-659.

[119] Kil, I., Singapogu, R. B., & Groff, R. E. (2019, April). Needle Entry Angle & Force: Vision-enabled Force-based Metrics to Assess Surgical Suturing Skill. In *2019 International Symposium on Medical Robotics (ISMR)* (pp. 1-6). IEEE.

[120] Payne, C. J., Marcus, H. J., & Yang, G. Z. (2015). A smart haptic hand-held device for neurosurgical microdissection. *Annals of biomedical engineering, 43*(9), 2185-2195.

[121] Yamaguchi, T., & Nakamura, R. (2018). Laparoscopic training using a quantitative assessment and instructional system. *International Journal of Computer Assisted Radiology and Surgery, 13*(9), 1453-1461.

[122] Forestier, G., Lalys, F., Riffaud, L., Trelhu, B., & Jannin, P. (2012). Classification of surgical processes using dynamic time warping. *Journal of biomedical informatics, 45*(2), 255-264.