Taher *et al. Art Int Surg* 2022;2:144-58
DOI: 10.20517/ais.2022.11

Artificial Intelligence Surgery

Systematic Review

Open Access

Check for updates

The challenges of deep learning in artificial intelligence and autonomous actions in surgery: a literature review

Heba Taher[1], Vincent Grasso[2], Sherifa Tawfik[3], Andrew Gumbs[4]

[1]Pediatric Surgery Department, Cairo University, Cairo 11441, Egypt.
[2]Family Christian Health Center, Department of Information Systems, Science and Technology, Harvey, IL 60426, USA.
[3]Department of Pathology, Egypt Ministry of Health and Population, Cairo 11441, Egypt.
[4]Département de Chirurgie Viscéral, Center Hospitalier Intercommunal de Poissy/Saint-Germain-en-Laye, Poissy 78300, France.

Correspondence to: Dr. Heba Taher, Pediatric Surgery Department, Cairo University, Aly Ibrahim bash, El Sayda zeinab, Cairo 11441, Egypt. E-mail: Hebatallah.Taher@kasralainy.edu.eg

How to cite this article: Taher H, Grasso V, Tawfik S, Gumbs A. The challenges of deep learning in artificial intelligence and autonomous actions in surgery: a literature review. *Art Int Surg* 2022;2:144-58. https://dx.doi.org/10.20517/ais.2022.11

Received: 2 May 2022 **First Decision:** 24 Jun 2022 **Revised:** 10 Aug 2022 **Accepted:** 9 Sep 2022 **Published:** 23 Sep 2022

Academic Editor: Marialuisa Lugaresi **Copy Editor:** Peng-Juan Wen **Production Editor:** Peng-Juan Wen

Abstract

Aim: Artificial intelligence (AI) is rapidly evolving in healthcare worldwide, especially in surgery. This article reviews important terms used in machine learning and the challenges of deep learning in surgery.

Methods: A review of the English literature was carried out focused on the terms "challenges of deep learning" and "surgery" using Medline and PubMed between 2018 and 2022.

Results: In total, 54 articles discussed the challenges of deep learning in general. We include 25 articles from various surgical specialties discussing challenges corresponding to their respective specialties.

Conclusion: The increased utilization of AI in surgery is faced with a wide variety of technical, ethical, clinical, and business-related challenges. The best way to expedite its expansion in surgery in the safest and most cost-efficient manner is by ensuring that as many surgeons as possible have a clear understanding of basic AI concepts and how they can be applied to the preoperative, intraoperative, postoperative, and long-term follow-up phases of the surgical patient care.

Keywords: Challenges of deep learning, surgery, artificial intelligence, surgical data science, image annotation, data standardization

INTRODUCTION

Nobel Prize Laureate and Professor Stephen Hawking stated in 2014 that AI could *"spell the end of the human race, as once humans develop AI, it will take off on its own and redesign itself at an ever increasing rate and humans, who are limited by slow biological evolution, couldn't compete and would be superseded."*

These words of caution are shared by Tesla and SpaceX founder Elon Musk, history professor Yuval Noah Hariri, and Microsoft founder Bill Gates, all expressing serious concerns about the possibility of artificial intelligence (AI) spinning out of control from our incessant attempts for machines to acquire complete autonomy[1].

There are many significant and complex obstacles that must be overcome for AI to become autonomous and hyper-intelligent. This article reviews one of the main pillars of AI, machine learning (ML), and focuses on the challenges of achieving strong AI or full autonomy in surgery. ML includes algorithms such as classification and regression decision trees (CART) and support vector machines. Both models can analyze categorical (classification) or continuous (regression) variables.

Deep learning (DL) is an architecture of ML that was designed to mimic the neurological structure of the human brain and is currently the most promising method for us to get to more autonomous actions in surgery[1]. Thus far, narrow or "weak" AI is being utilized, but it is limited to algorithms that perform singular (unique) or extremely limited human tasks[2]. Artificial neural networks (ANNs) are architectures of DL. They require higher computational power and more annotated data. Data analysis is improved by way of enabling increased data input and simplifying data acquisition[3].

As mentioned above, ANNs are compared to a biological neural network, since training in an ANN is similar to learning in the brain[4,5]. Furthermore, ANNs are the most popular DL model. ANNs process an input (e.g., an image), passing it through many layers of interconnected nodes [called DNN (deep neural network)] that loosely resemble the structure of biological neurons. These neurons progressively detect features to finally provide an output[2], typically a classification (e.g., "dog" or "not a dog").

The building block of the ANN is the node or perceptron. Each perceptron has multiple inputs (1 to a definite number), where inputs are weighted differently depending on their predictive value. Neurons fire when their input sums arrive at the neuron activation threshold (resulting in output production). A bias function can be introduced to change the likelihood of reaching this threshold by making it easier or harder to reach. Regarding input weight, it is randomly determined, and the model calculates the predictive error by comparing its prediction to the known outputs in the test data. Input weights are then iteratively adjusted to improve performance in a process called backpropagation.

"Iteration" is used to train DL networks, and it is the process of running and rerunning networks with continuous optimization of neuronal parameters to optimize performance and minimize errors. This improvement in output is the result of analyzing massive datasets, the development of better algorithms, and faster computational hardware[1].

In addition to ANNs, convolutional neural networks (CNNs) are increasingly being used in image processing as they are designed to process pixel data[6] and find efficacy in the field of computer vision (CV). Interestingly, CNNs structurally resemble the neural architecture used by the human visual cortex, which is situated in the occipital lobe[7].

How does a machine learn?

Pioneering computer scientist Arthur Samuel, in 1959, defined ML as giving computers the ability to learn without being definitively programmed. In other words, ML is the potential for algorithms to make decisions and predictions by interpreting data on their own. ML utilizes learning tasks that include classification, regression, clustering, and dimensionality reduction[8]. Models carry out analysis and predict outcomes based on input features. The process of ML involves model selection, training, evaluation and optimization[8], data collection, feature engineering, and data preprocessing.

Data collection and preprocessing are key steps in ML. Additional aspects of ML code can lead to shortened computer times and improved accuracy by generating features that determine algorithm performance. After data extraction, new datasets are constructed and selected, which are divided into training sets and others used for testing of models that are then validated[8]. ML, and in particular DL, is ideal for processing big data, and its utilization has been rising dramatically over the last few years.

The emergence of surgical data science

The quality of the resulting algorithm depends largely on the data on which the model is trained. Poor quality datasets, which have duplicates, missing data, and inconsistencies, weaken the accuracy of algorithms even when large datasets are used.

The importance of data in ML, particularly in surgery, has led to the emergence of a new field called surgical data science (SDS)[9]. The first building block of SDS is *perception*, during which the relevant data are perceived by the system, whether by humans (surgeons, operating team, or medical staff), devices, robots, or sensors used to capture patient data. These data provide the basis for the next building block, interpretation, in which the perceived data are interpreted in a context-aware manner to provide real-time assistance, which is the third building block in SDS[9].

The interpretation comprises data annotation, which is a cornerstone of SDS, and data analytics. There are several classes of data analytics tools. These include descriptive analytic tools, which provide a comprehensive summary of data made available through simple reports. There are also diagnostic analytic tools, which clinicians use to assess the effectiveness of a treatment protocol. In addition, predictive analytic tools assist healthcare providers in making a decision, while prescriptive analytic tools help them arrive at a decision[9].

Clinical translation, which is occasionally referred to as "from bench to bedside" or "the valley of death", provides real-time assistance ranging from surgical education to various clinical tasks such as early detection, diagnosis, and assistance with therapy. An example of this is a DL algorithm created to aid in the diagnosis of diabetic retinopathy[9].

Big data and its four attributes

Big data describes very large volumes of structured and unstructured data and is defined by the following four attributes: volume, variety, velocity, and veracity. Concerning the healthcare industry, approximately 80% of its data are unstructured[4].

Volume refers to large datasets that continue to expand from terabytes to petabytes and beyond. Variety refers to structured data such as patient laboratory data or demographic details in electronic health records (EHRs) or unstructured data such as radiological images from MRI scans. Velocity refers to the speed that data are generated and how quickly they move and are transmitted. Veracity refers to the accuracy, consistency, quality, and trustworthiness of data[4].

Types of machine learning

Supervised learning (SL), the commonest type of ML[10], is also referred to as predictive learning. In SL, the underlying truth is known and provided to the machine that then uses labeled data to learn a general prediction rule in the derivation of an algorithm[4,11]. Retinopathy screening algorithms are an example of supervised deep learning[12].

Unsupervised learning (UL), also referred to as descriptive learning, does not have a ground truth label, and the algorithm finds structure within the data even if there is no training dataset. Data are fed into the algorithm without an explanation of what to do. By comparing similarities within the dataset, a structure in the dataset consisting of unknown inputs can be ascertained[4].

Reinforcement (incremental) learning (RL) is where a program attempts to accomplish a task while learning from its own success and mistakes, essentially through trial and error[13]. It can be used in the customization of anti-retroviral and anti-epileptic drugs, sepsis management[4], and insulin administration in artificial pancreatic systems[13]. It is also used to develop autonomous actions in the interventional fields of surgery, interventional radiology, and endoscopy[14-17].

Other types of emerging ML methods in surgery include semi-supervised learning, which is a type of ML that combines a small number of labeled data with a large number of unlabeled data during training. Learning from demonstration is a novel method in which robots acquire new skills by learning to imitate an expert. Federated learning is an ML method that enables models to get experience from different datasets located in different sites (e.g., a central server or local data centers) without sharing data. Finally, meta-learning refers to learning algorithms that learn from other learning algorithms[18].

The performance of ML is ranked according to its level of discrimination (probability of predicting outcomes accurately) and calibration (the degree of over- or underestimating the predicted versus true outcome)[7].

METHODS

Search terms and ideas were discussed with the Editorial Team of Artificial Intelligence Surgery. The research was carried out on PubMed (https://pubmed.ncbi.nlm.nih.gov/), accessed on 20 January 2021, using the words "challenges in deep learning and surgery" for the years between 2018 and 2022. All studies discussing deep learning challenges in non-surgical specialties and non-English language articles were excluded. Four independent reviewers were involved in the process, and any disagreement was resolved by discussion.

This article reviews the systematic reviews and primary peer-reviewed articles published in the literature regarding the challenges of DL in surgery (PRISMA 2020 flow diagram, Figure 1). Secondary literature, consisting of narrative reviews, is also included. This review was submitted to PROSPERO, and it is currently awaiting approval. The PRISMA checklist was used for the manuscript design (PRISMA 2020 checklist, Table 1) and drafting of the manuscript. An explanation of some of the basics of AI is presented in

Table 1. PRISMA 2020 checklist[34]

Section and topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review	1
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist	1
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge	2
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses	2
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses	6
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted	6
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used	6-7
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process	6
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process	6-7
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect	6
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information	6
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process	6
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results	6
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)	6
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions	6
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses	6-7
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used	-
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression)	-
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results	-
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases)	6
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome	6
RESULTS			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram	7
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded	7

Study characteristics	17	Cite each included study and present its characteristics	7
Risk of bias in studies	18	Present assessments of risk of bias for each included study	-
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots	7
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies	-
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect	-
	20c	Present results of all investigations of possible causes of heterogeneity among study results	-
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results	7
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed	-
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed	-
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence	7
	23b	Discuss any limitations of the evidence included in the review	8-15
	23c	Discuss any limitations of the review processes used	15
	23d	Discuss implications of the results for practice, policy, and future research	14-15
OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered	6
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared	6
	24c	Describe and explain any amendments to information provided at registration or in the protocol	6
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review	15
Competing interests	26	Declare any competing interests of review authors	15
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review	6-7

the Introduction to better explain the analysis and delineate the outcomes of the studies.

RESULTS

When the search was done on PubMed using the words "challenges in deep learning" and "surgery" between 2018 and 2022, 54 manuscripts were identified (PRISMA 2020 flow diagram, Figure 1) and 25 studies were included: 5 systematic reviews, 2 prospective studies, and 18 narrative reviews. Systematic reviews and prospective studies are considered primary sources, while narrative reviews are considered secondary sources.

The manuscripts focus on the challenges of deep learning in surgery, with 18 on general surgery, 3 on cardiothoracic specialty, 1 on orthopedic surgery, 3 on neurosurgery, and 1 on otolaryngology.

The seven systematic reviews focus on the most important ML and computer models to realize phase recognition in general surgery. Secondary aims include presenting commonly used concepts of AI and ML and discussions of data types and limitations of currently used model systems[3,4,7,10], variance analysis for predictive ability[8], and overall accuracy of those models[7,10].

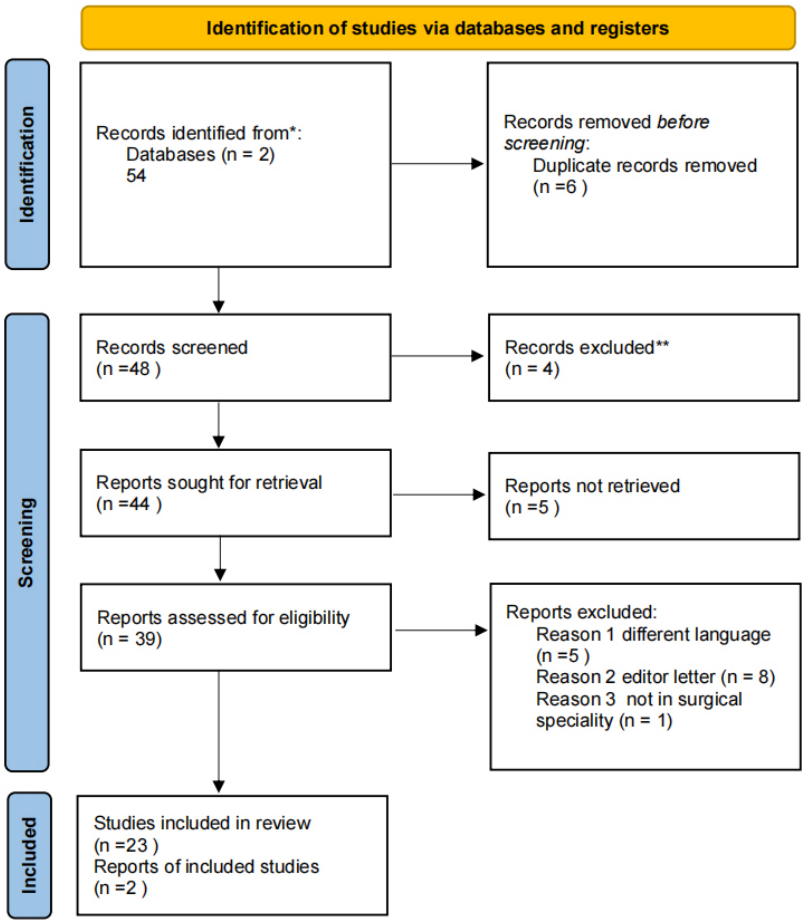

Figure 1. PRISMA 2020 flow diagram for new systematic reviews which included searches of databases and registers only[34]. *Consider, if feasible to do so, reporting the number of records identified from each database or register searched (rather than the total number across all databases/registers). **If automation tools were used, indicate how many records were excluded by a human and how many were excluded by automation tools.

The most recently published systematic review is a comprehensive review on surgical data science, providing evidence[9] of progress in this field since 2004. It includes a list of publicly accessible surgical datasets and registered clinical trials relevant to the field, in addition to 51 expert opinions about open challenges in the field of surgical data science.

Two prospective studies were reviewed. The first focused on methods and challenges related to providing sufficient high-quality surgical video recordings for a machine to be trained in surgical context with sufficient fidelity. The second study evaluated the application of AI in simulation-based training of necessary psychomotor skills involved in neurosurgery.

The remaining narrative reviews the definitions and challenges of DL in the field of surgery.

DISCUSSION

Challenges of deep learning

Data perception

Data structure and storage

There is, at times, limited technical infrastructure for data acquisition, storage, and access because not all acquired data are recorded and permanently stored. This is because of the high resolution of images and videos generated in surgical theaters, which are difficult to store (and retrieve) in healthcare information systems. Storing data in the cloud is becoming much more common, due to the expansion of data storage solutions within cloud environments including Amazon web services (AWS) and Microsoft Azure[9].

However, not all data are digitized and stored in a structured manner, which makes data access and exchange (and interoperability)[9] a challenging process. Advances in healthcare information systems involve increasing the level of interoperability of their systems to enable data to be securely exchanged. The highest level of interoperability is system awareness of the underlying assumptions, models, and processes of the systems involved. Currently, most systems are operating at a low level of interoperability. Adoption of standardized data formats, for example, fast healthcare interoperability resources (FHIR), provides better aggregation of data and systematized nomenclature of medicine/clinical terms (SNOMED CT). However, despite improved interoperability, it does not necessarily fix the problem of inconsistent semantic (vocabulary) coding in EHR data[16].

Data capture and sensors

There are challenges unique to the practice of surgery within its preoperative, intraoperative, and postoperative phases. For example, variables such as patient diversity, symptom complexity, imaging modalities and resolution, *etc*. impact the intraoperative phase[9].

The intraoperative phase is further complicated by tissue deformation, challenges in visual resolution, and environmental factors. In fact, the dynamic nature of tissues presents safety challenges, especially during robotic surgery, where real-time 3D reconstruction of tissue at the highest resolution possible is a prerequisite for safe operative navigation.

A challenging task is projecting an overlay on markerless deformable organs[18]. Differences between medical and internal images can be significant and impede clinical applicability. Therefore, the models used and the resultant outcomes may not be easily interpreted by computers unless a human flags any important issues, such as a potential risk or uncertainty during surgery. As a result, figuring out different transfer techniques to reduce the differences between image modalities and developing more explainable AI algorithms could enhance the performance of intraoperative decision making[18].

Sensors and algorithms need to be advanced to overcome the challenge of tissue deformation and the very dynamic environment encountered intraoperatively. Sensors, in particular, will play an important role in overcoming challenges to the development and implementation of AI by allowing for more accurate perception and capturing of dynamic and complicated environments via the fusion of multimodal data. The placement of sensors at the tips of instruments creates many challenging issues, starting from size constraints to limitations including the need for repetitive sterilization[14].

Traditional SLAM (simultaneous localization and mapping) algorithms[18] are built on the assumption of a rigid environment, unlike in a typical surgical field where the deformation of soft tissue and organs is often continuous. This flawed assumption limits its direct application for surgical tasks. To overcome this problem, a stereo-endoscope has been used within SLAM algorithms. Stereo-endoscopy is when an endoscopic camera is fitted with a sensor unit made of two flat image sensors arranged symmetrically around the shaft's central rotational axis, to help restore the sense of depth lost in Management Information

System (MIS) techniques and allow for the reconstruction of the anatomical structures within *in vivo* video recordings using well-established computer vision techniques.

Another approach to overcoming the challenges of a dynamic environment is the emerging field of biophotonics which uses spectral imaging to extract relevant information on tissue morphology, function, and pathology. The advantages of this approach are that it lacks ionizing radiation, has a low hardware complexity, and is easily integrated into the surgical workflow[9].

One of the challenges of sensor development is the special focus on haptics, where the detection of a surgeon's hand tremors is the data input to intraoperative machine processing. However, there is currently a debate as to whether there is a need to invest in haptics since the procedure may ultimately be performed without the need for the clinician to sense and interpret haptic information[14]. This is because surgeons who presently use surgical systems can perform surgery without haptics and rely exclusively on visual cues[14]; this makes the development of more sophisticated image capturing sensors perhaps more relevant.

Rather than haptics, investigators are paying more attention to the noise developed by those instruments when in use and are developing sensors to capture these audio data. The upside of this type of sensor is that it could be placed on the proximal end of the guide wire and not the distal part inside the patient, avoiding issues of size and regulatory constraints[14].

Interpretation

Data annotation

Datasets used in data annotation[9] must include reliable, accurate, efficient, scalable, and representative data of the correct specification. Surgical data such as videos can be distinguished by spatial, temporal, and spatiotemporal annotation. *Spatial annotations* include image-level classification, for example, what tissue, tools and semantic classification, which pixel belongs to which tissue, and tool and numerical regression such as for tissue oxygenation.

Temporal annotation involves surgical workflow, for example, surgical phases, such as suturing and knot tying. Data labeling could be expensive, while gathering poorly labeled data will produce poor results and waste money[3,13,19]. It could also have a high inter-operator variability; for example, the total surface area of a burn injury could present a barrier to effective ML due to varying degrees of accuracy of labeled data[7].

Therefore, the major challenge[9] for data annotation for large-scale surgical applications is access to expert knowledge, where reducing annotation effort is the main target in annotation development. Active learning approaches that decide which unlabeled data would provide the most information and therefore reduce annotation efforts have been proposed, as have error detection methods and methods for annotating data directly during acquisition.

Other solutions are projects aimed at specifying the core ontology[9] of surgical processes, i.e., gathering basic vocabulary to describe surgical actions and instruments. The aim of such projects is to define standards for surgical video annotation from different working groups regarding temporal models, actions and tasks, tissue characterization, and general anatomy, as well as software and data structure.

Fortunately, there are open-source datasets[9] which already exist as examples of publicly accessible and annotated surgical data repositories, e.g., ChlecTriplet21 and HeiSurf for cholecystectomy and GLENDA for laparoscopic gynecology. The challenge is providing registries for less standardized operations such as

hepatic-pancreatic and biliary or colorectal surgery, as this would require more sophisticated algorithms. The main obstacle is the paucity of videos for these procedures, which are significantly longer compared to cholecystectomy and restrictive bariatric surgery.

In SDS, images or videos are typically the main data sources that can be used to capture information at different granularities, ranging from theater room cameras to endoscopic cameras or microscopes. There is no gold standard framework enabling different image/annotation tools with AI-assisted annotation methods in the SDS field[9]. A new industry has emerged[9] offering online dataset annotation services through a large and organized human workforce. It is, however, unclear to what extent medical image data annotation can be outsourced to such companies, where important information could be lost. Furthermore, the high cost of such services may render the cost out of reach for many institutions.

Some technology companies gain access to patient information via their devices (da Vinci Surgical System, Intuitive Surgical, Sunnyvale, CA, USA) by keeping track of every intraoperative gesture carried out by their device[15]. This also raises the concern of advances in the SDS (software-defined storage) sector as commercial ventures (information is power) could enable near-monopolies of relevant SDS data and create tech giants, while surgery should be a public good and not just commercially exploited.

Another challenge faced in annotation is bias awareness, where confounding variables are crucial to the successful predictive model. Unrecognized cofounders or systemic bias in data can lead to flawed algorithms that may include racial biases or ersatz identifiers[19-21]. For instance, an algorithm might identify a pneumothorax because of the presence of a chest tube rather than the actual pathological finding on X-ray. Another example is the diagnosis of skin-based melanoma because the algorithm was fed an image containing a ruler adjacent to the lesion. This is one of the most important short-term accuracy and reliability challenges and is occasionally referred to as *distributional shift*[7].

Data analytics

Computational challenges

Algorithm training can be computationally expensive and depends on advances in computational powers provided by central processing unit (CPU) and graphics processing unit (GPU) resources[8,22]. The rate-limiting step for increasing ANN applications is computational power, with increases in computational power and advances in CPUs and GPUs enabling theoretically limitless improvements[7,20].

Generalizability and external validity

This indicates the degree to which an algorithm can be transferred (transfer means fine-tuning a convoluted neural network (CNN) that is pre-trained on a large dataset of natural images, instead of training a network from the start)[11] from the population being studied to a different setting, where algorithms do not need to be universally generalizable to be pertinent. AI trained to recognize a patient at risk of mortality in a UK NHS ICU might not perform well in another country with a population with different characteristics or intensive care protocols, but it could still be useful in the target population.

Algorithms trained on large datasets might have different levels of performance based on subgroups within the data[20].

Overfitting and underfitting

Even after the model passes internal validation, it might not be generalizable to other populations of interest if the model is *overfitted* to idiosyncrasies or biases in the data using predictors which are not clinically

useful (high variance model), as in the weight of portable X-rays for chest tube examinations. Conversely, when a model performs poorly on internal validation, this can be described as *underfitting* or having high bias. This is caused by insufficient data or an overly simplistic model (e.g., linear relationship between variables when this is not true)[20].

Interpretability, i.e., "black box effect"

As the complexities of models increase, it becomes difficult or impossible to identify the relationship between input and output (the black box conundrum)[23], which is unique to DL ANNs. This creates a challenging dilemma for those individuals responsible for a potential resulting medical error.

Understanding how models reach their predictions allows the researchers to evaluate whether their reasoning is justifiable or was subject to bias or error, such as using rulers to predict melanomas[20]. European legislation mandates that individuals have a right to an explanation of decisions made by automated systems[20]. Moreover, the features used to differentiate between data categories are not translated quickly into verbal or visual rules for any human to easily understand.

Insensitivity to impact

There exists an ethical challenge in ML for the algorithm to be oblivious to the consequences of a false negative or false positive test (predictive tools that underestimate the outcome of a false positive or false negative outcome). While comparing ML systems and an expert dermatologist, both machines and humans found it difficult to discriminate between malignant melanocytic lesions and benign lesions; however, human experts "erred on the side of caution" because humans are aware of the dire impact of a false negative diagnosis on the patient. Herein, ML systems should be trained not only with the end result but also with the costs of both missed diagnosis and over-diagnosis.

Reward hacking

A machine can learn unpredicted methods of achieving an outcome that cheats the system[7,24], necessitating meticulous validation through expert opinion and multicenter studies to evaluate the transferability of expertise to real-life scenarios. This is particularly important in surgical simulations when a loophole would enable the trainee to perform a surgical step successfully in the simulator that is inapplicable to a real-life operative scenario and should be identified by experts to retrain the system and validate it.

Catastrophic interference

A main criticism of AI is catastrophic interference, i.e., forgetting where ANNs may no longer utilize previously learned information during the process of learning new information. It is due to overwriting of the original data memory by the next one. It is hoped that one day models will master the ability of gradually sustained accumulation of knowledge similar to humans. Incremental learning (IL) is the continual process of knowledge extraction from new data and preserving most of the previously obtained knowledge. This learning could overcome historical data occupying storage space. In addition, it fully utilizes all previous results, hence significantly saving time for new training. This learning applies mainly to big data and is particularly suitable to the medical field. Common methods of incremental learning or reinforced learning (RL) include feature extraction, fine-tuning, knowledge distillation, and joint training. Incremental learning has been widely employed in domains such as muscle activity and kinematics, non-convulsive epileptic seizure, segmentation of anatomical structures, and atrial fibrillation[8].

Clinical translation (real-time assistance)

Lack of outcome reporting and accountability

Algorithms can show impressive results on internal validation but poor results when applied clinically (from bench to bedside). Hence, this is occasionally referred to as the "valley of death"[9], where there is a lack of external validation within randomized controlled trials.

A systematic review published in 2020 showed that 38% of studies discussed the need for prospective studies and 9% claimed algorithms could be integrated into clinical practice[1]. Despite this shortcoming, it is worth noting that not all questions that arise in the process of clinical translation of an algorithm need to be addressed by an RCT. For example, a diagnostic algorithm to diagnose diabetic retinopathy based on a pivotal prospective study was recently approved by the FDA[9].

Reporting guidelines implementation

With advances in AI and deep learning, healthcare could be able to emulate aviation as regards safety. Key safety-related domains in aviation involve checklists, training, crew resource management, investigating and reporting incidents, and organizational culture. Attempts to achieve a similar degree of standardization and automation of procedures with multiple backup systems and automatic availability of information are currently underway[25].

Since research into the role of AI in the healthcare sector, and specifically within the surgical domain, has exponentially increased over the past decade, there is a dire need to implement reporting guidelines.

In addition, other requirements need to be met, such as defining the role of AI within clinical care pathways, stress testing the AI in various clinical scenarios, and, just as importantly, the willingness to accept AI by patients and physicians. It is also fundamental to recall that AI systems will always need to update their parameters because of the natural history of diseases and the fact that populations change.

There is an ethical concern about who is responsible when an error occurs. Is it the machine, the physician, or the research group that trained, monitored, and evaluated the algorithm? Is it the technology company that developed the software or the data source? Could it be a combination of them all? As mentioned above, this is further complicated by the need to continuously update algorithms, which necessitates a continuous input of new patient information[20].

Other concerns

There are great concerns about the possible reduction in open surgical skills, particularly with the rise of autonomy in minimally invasive/robotic surgery. Surprisingly this is far from the reality because, despite advances in robotic surgery, surgery is far from high-level autonomy in the operating room[26]. Attanasio *et al.* defined and described six levels of autonomy for surgical robots, where level 0 is no autonomy and level 5 is full autonomy. The vast majority of current surgeries utilizing machines reside at levels 0 and 1, with advances in technology affording gradual increases towards level 5[14,15,17,27].

Concerns related to a decrease in open surgical procedures and their related value to surgical training may not be well founded. A highly informative and exhaustive Lancet publication reports that nearly five billion people do not have access to appropriate surgical care, especially within communities represented by the medically underserved and economically disadvantaged[28]. With increased advocacy about patient confidentiality, there are rising concerns about data security[29] and cybercrime due to the ever-expanding digitization of patient health information and an overreliance on the digital system[23]. Moving away from

fee-for-service models towards models rewarding cost efficiency and prioritizing the treatment of low acuity patient populations may indirectly alienate certain patients from accessing care, particularly when you take into account ethnic or socioeconomic backgrounds and associated co-morbidities and risk factors[10].

Biggest challenge: acceptance

Should we accept AI into our lives and view it as a support for our progress and not as a threat to our existence? Should we view it as a means to understand ourselves and the world around us?

To understand how a certain behavior is generated, scientists can build the neural pathway of this behavior[30] and observe the biological robot in action in their laboratories. They can mimic natural pathways that link optical and acoustic sensors - i.e., eyes and ears - to motors, which can then act as surrogate muscles. This was previously used to simulate a cricket called Khepera. It is an easy creature for which to build the pathway needed for a female Khepera to locate a male mate utilizing only a few neurons and, hence, a small cluster of mathematical processors or nodes (10 nodes) designed to act as nerve cells. The process is called phonotaxis, which proved that there are not only two systems involved in the process, i.e., one to recognize the male's call and the other to locate it, but also a third system that recognizes and moves towards the sound.

With complex creatures rising up to the ant kingdom (Cataglyphis), scientists could take the "black box" approach in building neural networks. Here, they did not simulate the actual firing of the neurons as in Khepera; they only wanted to reproduce the same inputs and outputs to and from the neural network, and it worked and behaved as living ants. The model also helped them understand why the Cataglyphis have three polarized light (POL) neurons despite the model showing that one is enough to sense light.

Finally, building CNNs that structurally resemble the neural architecture used by the human visual cortex situated in the occipital lobe might help us understand more about ourselves and how our brains function and possibly revolutionize certain treatments.

Deep learning could help us understand life processes by recreating them, providing a great opportunity to learn not only about the world around us but also about ourselves.

In conclusion, thanks to advances in deep learning, AI will play a bigger role in surgical specialties and instruments and move us closer to more autonomous actions in surgery. However, to achieve this autonomy and broader application of AI, we need to overcome certain challenges, as highlighted in the reviewed articles. Computational power and technology seem to be the most manageable components, while acquiring high-quality data that accurately represent the target population[31-33] is challenging. High-quality representative data are important during the entire machine training process. In addition to the technical and training challenges described above, all relevant healthcare personnel need to be included in these processes and recognize the valuable role they play within the entire healthcare delivery landscape as it migrates to utilizing all things AI.

DECLARATIONS

Acknowledgments

The authors would like to thank Miss Jane Lee assistant editor for facilitating acquisition of information and support.

Authors' contributions

Conception and design of the study and performed data analysis and interpretation: Taher H
Performed data acquisition, as well as provided administrative, technical, and material support: Gumbs A, Tawfik S, Grasso V

Availability of data and materials

Not applicable.

Financial support and sponsorship

None.

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

REFERENCES

1. Darbari A, Kumar K, Darbari S, Patil PL. Requirement of artificial intelligence technology awareness for thoracic surgeons. *Cardiothorac Surg* 2021. DOI
2. Aung YYM, Wong DCS, Ting DSW. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. *Br Med Bull* 2021;139:4-15. DOI PubMed
3. Garrow CR, Kowalewski KF, Li L, et al. Machine learning for surgical phase recognition: a systematic review. *Ann Surg* 2021;273:684-93. DOI PubMed
4. Raju B, Jumah F, Ashraf O, et al. Big data, machine learning, and artificial intelligence: a field guide for neurosurgeons. *J Neurosurg* 2020. DOI PubMed
5. Bodenstedt S, Wagner M, Müller-Stich BP, Weitz J, Speidel S. Artificial intelligence-assisted surgery: potential and challenges. *Visc Med* 2020;36:450-5. DOI PubMed PMC
6. Alapatt D, Mascagni P, Srivastav V, Padoy N, Hill M. Artificial intelligence in surgery neural networks and deep learning. Available from: https://github.com/CAMMA-public/ai4surgery [Last accessed on 15 Sep 2022].
7. Chang M, Canseco JA, Nicholson KJ, Patel N, Vaccaro AR. The role of machine learning in spine surgery: the future is now. *Front Surg* 2020;7:54. DOI PubMed PMC
8. Ren GR, Yu K, Xie ZY, et al. Current applications of machine learning in spine: from clinical view. *Global Spine J* 2021. DOI PubMed
9. Maier-Hein L, Eisenmann M, Sarikaya D, et al. Surgical data science - from concepts toward clinical translation. *Med Image Anal* 2022;76:102306. DOI PubMed PMC
10. Lopez CD, Gazgalis A, Boddapati V, Shah RP, Cooper HJ, Geller JA. Artificial learning and machine learning decision guidance applications in total hip and knee arthroplasty: a systematic review. *Arthroplast Today* 2021;11:103-12. DOI PubMed PMC
11. Flores AM, Demsas F, Leeper NJ, Ross EG. Leveraging machine learning and artificial intelligence to improve peripheral artery disease detection, treatment, and outcomes. *Circ Res* 2021;128:1833-50. DOI PubMed PMC
12. Simon HA, Newell A. Human problem solving: the state of the theory in 1970. *Am Psychol* 1971;26:145-59. DOI
13. Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial intelligence in surgery: promises and perils. *Ann Surg* 2018;268:70-6. DOI PubMed PMC
14. Gumbs AA, Frigerio I, Spolverato G, et al. Artificial intelligence surgery: how do we get to autonomous actions in surgery? *Sensors (Basel)* 2021;21:5526. DOI PubMed PMC
15. Gumbs AA, Grasso V, Bourdel N, et al. The advances in computer vision that are enabling more autonomous actions in surgery: a systematic review of the literature. *Sensors (Basel)* 2022;22:4918. DOI PubMed PMC
16. Wagner M, Bodenstedt S, Daum M, et al. The importance of machine learning in autonomous actions for surgical decision making. *Art*

Int Surg 2022;2:64-79. DOI
17. Entwistle A. What is artificial intelligence? *Eng Mater Des* 1988;32:1-10. DOI
18. Zhou XY, Guo Y, Shen M, Yang GZ. Application of artificial intelligence in surgery. *Front Med* 2020;14:417-30. DOI PubMed
19. Bar O, Neimark D, Zohar M, et al. Impact of data on generalization of AI for surgical intelligence applications. *Sci Rep* 2020;10:22208. DOI PubMed PMC
20. Kuo RYL, Harrison CJ, Jones BE, Geoghegan L, Furniss D. Perspectives: a surgeon's guide to machine learning. *Int J Surg* 2021;94:106133. DOI PubMed
21. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019;17:195. DOI PubMed PMC
22. Loftus TJ, Vlaar APJ, Hung AJ, et al. Executive summary of the artificial intelligence in surgery series. *Surgery* 2022;171:1435-9. DOI PubMed PMC
23. Devabalan Y. The use and challenges of artificial intelligence in otolaryngology. *Authorea* 2020. DOI
24. Mirchi N, Bissonnette V, Yilmaz R, Ledwos N, Winkler-Schwartz A, Del Maestro RF. The virtual operative assistant: an explainable artificial intelligence tool for simulation-based training in surgery and medicine. *PLoS One* 2020;15:e0229596. DOI PubMed PMC
25. Kapur N, Parand A, Soukup T, Reader T, Sevdalis N. Aviation and healthcare: a comparative review with implications for patient safety. *JRSM Open* 2016;7:2054270415616548. DOI PubMed PMC
26. Etienne H, Hamdi S, Le Roux M, et al. Artificial intelligence in thoracic surgery: past, present, perspective and limits. *Eur Respir Rev* 2020;29:200010. DOI PubMed
27. Attanasio A, Scaglioni B, De Momi E, Fiorini P, Valdastri P. Autonomy in surgical robotics. *Annu Rev Control Robot Auton Syst* 2021;4:651-79. DOI
28. Meara JG, Leather AJM, Hagander L, et al. Global surgery 2030: evidence and solutions for achieving health, welfare, and economic development. *Lancet* 2015;386:569-624. DOI PubMed
29. Birkhoff DC, van Dalen ASHM, Schijven MP. A review on the current applications of artificial intelligence in the operating room. *Surg Innov* 2021;28:611-9. DOI PubMed PMC
30. Graham-Rowe D. March of the biobots. *New Scientist* 1998;160:26-30. Available from: https://www.newscientist.com/article/mg16021634-900-march-of-the-biobots/ [Last accessed on 15 Sep 2022].
31. Taher HMA, Fares A, Wishahy AMK. Laparoscopic resurrection of an old technique: a new approach for total urogenital separation and rectal pull-through in patients with long common channel cloacal malformation. *J Endourol* 2022;36:1177-82. DOI PubMed PMC
32. Taher H, Khalil H, Ahmed S, et al. Umbilical hernia repair post umbilical cord graft closure of gastroschisis: a cohort study. *Int J Surg Case Rep* 2022;95:107175. DOI PubMed PMC
33. Taher H, Elboraie A, Fares A, Tawfiq S, Elbarbary M, Abdullateef KS. Laparoscopic inguinal hernia repair in bladder exstrophy, a new modified solution to an old problem: a cohort study. *Int J Surg Case Rep* 2022;95:107252. DOI PubMed PMC
34. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71. DOI PubMed PMC