Front. Med. 2020, 14(4): 369–381
https://doi.org/10.1007/s11684-020-0784-7

REVIEW

Current applications of artificial intelligence for intraoperative decision support in surgery

Allison J. Navarrete-Welton[1], Daniel A. Hashimoto (✉)[1,2]

[1]*Surgical Artificial Intelligence and Innovation Laboratory, Massachusetts General Hospital, Boston, MA 02114, USA;* [2]*Harvard Medical School, Boston, MA 02114, USA*

Abstract **Research into medical artificial intelligence (AI) has made significant advances in recent years, including surgical applications. This scoping review investigated AI-based decision support systems targeted at the intraoperative phase of surgery and found a wide range of technological approaches applied across several surgical specialties. Within the twenty-one ($n = 21$) included papers, three main categories of motivations were identified for developing such technologies: (1) augmenting the information available to surgeons, (2) accelerating intraoperative pathology, and (3) recommending surgical steps. While many of the proposals hold promise for improving patient outcomes, important methodological shortcomings were observed in most of the reviewed papers that made it difficult to assess the clinical significance of the reported performance statistics. Despite limitations, the current state of this field suggests that a number of opportunities exist for future researchers and clinicians to work on AI for surgical decision support with exciting implications for improving surgical care.**

Keywords artificial intelligence; decision support; clinical decision support systems; intraoperative; deep learning; computer vision; machine learning; surgery

Introduction

In 1978, the cardiovascular surgeon Dr. Frank Spencer wrote that "a skillfully performed operation is about 75% decision making and 25% dexterity" [1]. While the exact split between technical skill and cognitive decisions can be debated and these domains often overlap, surgical practice requires complex decision making at each phase of care. The literature supports the importance of decision making (in both technical and non-technical aspects of care) in the outcome of a patient. In one recent study of surgical errors, cognitive errors were identified as a contributing factor to over half of the adverse events recorded [2]. Despite the relationship of the decision-making process to patient outcome, decision making skills are less emphasized than technical skills during surgical training, perhaps due to the difficulty of teaching decision making [3]. Furthermore, additional research suggests that decision-making skills vary with surgeon experience [4]. Thus, finding ways to improve the quality of surgical decision making could help improve outcomes by optimizing surgical care.

Intraoperative decision making has been well-studied — though predominantly through structured qualitative methods such as cognitive task analysis [4,5]. Flin *et al*. (2007) presented an excellent framework from which to consider intraoperative decision making, emphasizing processes of naturalistic decision making, i.e., the process of making decision under "conditions of high uncertainty, inadequate information, shifting goals, high time pressures and risk, usually working in teams and subject to organisational constraints." [6] In this framework, surgeons are thought to make decisions through a three-step process that includes situational assessment, action-taking, and reevaluation of the action's consequences.

Artificial intelligence (AI) has been proposed as a decision-making aid in a wide variety of fields, including medicine. Over the past 20 years, there has been an explosion of research in medical AI, facilitated by the increasing availability of medical data. While interest and research on AI applications in surgery is increasing [7,8], much of the focus on medical AI has been in other specialties such as radiology, pathology, or dermatology

Received September 4, 2019; accepted March 14, 2020
Correspondence: Daniel A. Hashimoto, dahashimoto@mgh.harvard.edu

[9–11]. Based on Flin *et al.*'s framework, AI could potentially affect the manner in which surgeons assess a given situation (e.g., through better data about a clinical scenario), the types of actions that are taken (e.g., through decision suggestion), and the process of re-evaluating the impact of an action.

To better understand the landscape of AI-supported intraoperative decision making, we conducted a scoping review that investigated AI technologies intended to support intraoperative decision making. We summarized the literature in three broad categories based on the authors' cited motivations for applying AI technology to surgical decision support: (1) increasing the information available to surgeons, including retrieving similar cases from a database and compensating for the loss of sensory input during minimally invasive surgery; (2) accelerating intraoperative pathology, including tumor margin mapping, tumor classification, and tissue identification; and (3) recommending surgical steps as a form of decision support.

Methods

The Medline (Ovid), IEEE Xplore, Web of Science, and PubMed databases were searched for papers that included a keyword from each of three categories: surgery (keywords: surgery, surgeon, intraoperative, operative, postoperative complication), decision support (keywords: decision, decision making, real-time systems, clinical decision-making, decision support system) and artificial intelligence (keywords: artificial intelligence, machine learning, neural network, algorithm, computer-assisted, computational modeling, biomedical computing, data mining, optimization, data models, computerized monitoring, expert systems). These keywords were selected based on an initial overview of the literature from a few prior review papers on the topic of artificial intelligence in surgery and related fields [8,12]. While this review is focused on intraoperative decision making, our initial search strategy captured titles and abstracts that may have included preoperative and postoperative decision making in case such decisions were related to intraoperative decision support.

Only papers focused on the design or application of AI-based algorithms for *intraoperative* decision support for surgeons were included in the final analysis. Artificial intelligence was defined as "the study of algorithms that give machines the ability to reason and perform cognitive functions." Papers related to anesthesia, surgeon training, and surgeon skill evaluation were excluded, as were papers related to robotic surgery in which the decision support applications were specific to the operation of the robot (for instance, warning systems to avoid robotic tool collisions) rather than on supporting clinical decisions during the operation.

During the abstract screening phase, papers related to the preoperative and postoperative phases of surgery were excluded if they did not apply to intraoperative decision support. All English language peer-reviewed published literature and peer-reviewed conference proceeding papers to May 25, 2019 were eligible for inclusion. Narrative review papers, editorials, letters to the editor, and abstracts were excluded, as were any studies involving animals or fewer than 10 patients. Two reviewers screened articles for inclusion/exclusion using Covidence (Melbourne, Australia). Reference lists of included papers were hand-searched by one reviewer and included if the inclusion criteria were met. Inclusion and exclusion criteria are summarized in Table 1.

Results

A total of 2397 titles were identified from the database search, and 543 abstracts were eligible for screening (including papers dealing with preoperative and postoperative phases). Of these, 9 manuscripts dealing with the intraoperative phase were selected for inclusion. Review of the citations produced an additional 6 qualifying papers; reviews of the citations in the additional papers produced another 6 papers. A total of 21 reviewed papers were included in the final analysis (Fig. 1).

The papers featured a wide range of surgical subspecialties, including gynecologic surgery, neurosurgery,

Table 1 Inclusion and exclusion criteria for this scoping review

Included	Excluded
Artificial intelligence applied to decision support for the surgeon during the intraoperative phase of surgery	Decision support during the preoperative and postoperative surgical phases Anesthesia, surgical training, and surgeon skill evaluation when unrelated to clinical decision support
Studies with at least 10 human patients	Studies with fewer than 10 patients
Published prior to May 25, 2019	Animal studies
Peer-reviewed published literature and conference proceedings papers	Narrative review papers, editorials, letters to the editor, abstracts
All geographical areas but written in English	Languages other than English

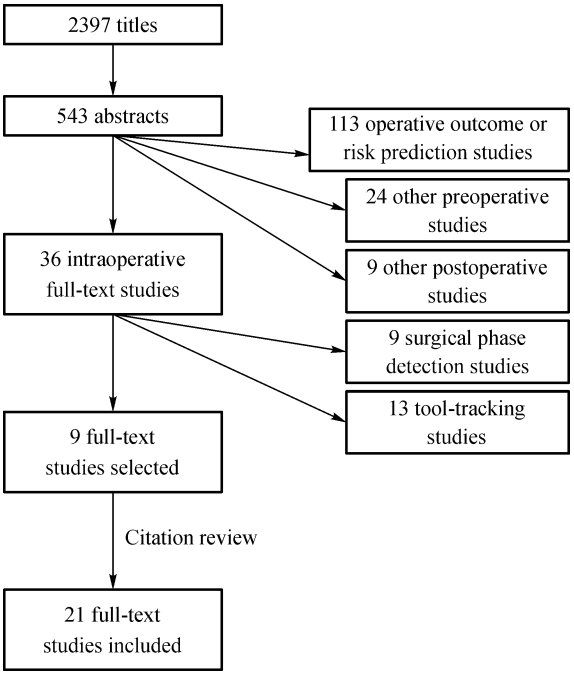

Fig. 1 Modified PRISMA diagram for this scoping review.

general surgery, ophthalmologic surgery, and endocrine surgery. Data in these studies were multimodal and included surgical videos, imaging modalities (e.g., hyperspectral imaging, optical coherence tomography, etc.), and intraoperative variables such as heart rate and blood pressure. A diversity of artificial intelligence techniques were used, including convolutional and recurrent neural networks, support vector regression, support vector machines, k-nearest neighbors classification, and graph-based methods. While this review was focused on intraoperative decision making, initial screening included preoperative and postoperative decision making in case such decisions were related to intraoperative decision support. Though none met inclusion criteria, a list of the preoperative and postoperative phase papers that were considered are available in Supplementary Material. Summarization of these papers was outside the scope of this review.

Given the diversity of methods and topics, the motivations cited for the introduction of AI to the surgical setting were the most unifying features within the reviewed body of work. The papers summarized here have been grouped by the three main motivations we identified: (1) providing extra information to surgeons during operations, (2) accelerating intraoperative pathological diagnoses, and (3) direct recommendation of surgical steps in cases where experts capable of making those judgments are scarce or unreliable (Fig. 2). In the summaries, we focus on four areas in each paper: motivation, methods, data set characteristics, and evaluation process.

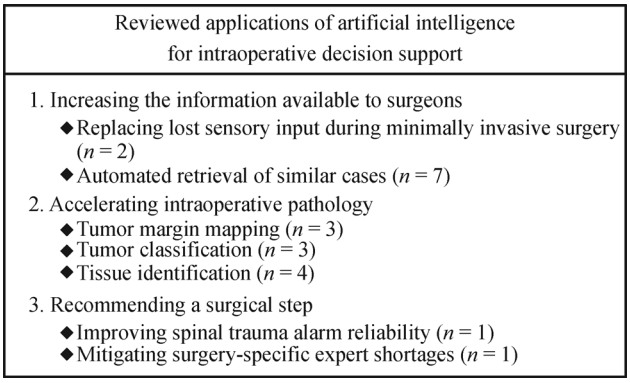

Fig. 2 Motivations cited by reviewed papers for developing artificial intelligence-based intraoperative decision support systems.

Increasing the information available to surgeons during operations

Two main approaches for providing extra information to surgeons intraoperatively were identified in this body of work. Two papers suggested methods to augment the limited sensory information available to surgeons during minimally invasive surgery [13,14]. Seven papers proposed algorithms for retrieving similar cases from a database using surgical video and images [15–20].

While minimally invasive surgery has many benefits for patients, the format can limit the sensory information available to surgeons compared to open surgeries. To address this challenge, Udelsman *et al*. (2014) used preoperative and intraoperative parathyroid hormone (PTH) levels to predict the probability of a cure during minimally invasive parathyroidectomy surgery [13]. During this procedure, the targeted approach of the minimally invasive technique focuses on a single adenoma that is localized preoperatively through imaging studies. In some patients, multiple adenomas may exist that may not have been identified preoperatively; thus, markers such as PTH can help surgeons determine whether all adenomas responsible for hyperparathyroidism have been resected. Udelsman *et al*. mathematically transformed the PTH level data and fed the results into a final stage logistic regression model to produce predicted probabilities of cure. They tested the model on an unscreened population of 100 patients, none of whom had indications of remaining hyperparathyroidism in either short-term or long-term follow-up. The model correctly predicted a cure in 78 of 81 (96.3%) patients with single adenoma and 17 of 19 (89.4%) patients with multiple adenomas.

Another application of AI to augment the intraoperative information during minimally invasive surgeries came from Harangi *et al*. (2017), who developed an artificial neural network (ANN) to distinguish the uterine artery from the ureter during laparoscopic hysterectomy [14].

During laparoscopic surgery, there is minimal tactile feedback from the instruments and palpation of specific structures with one's hands is not possible; therefore, identifying anatomic structures can become more difficult. In this study, a human would review an endoscopic image captured during laparoscopic hysterectomy and then draw a line over an anatomic structure of interest. The subimage along the line would then be fed into the ANN (a modified GoogLeNet architecture), which would classify the sub-images as either uterine artery or ureter. Harangi *et al*. used 2500 images taken from 35 patient videos. After data augmentation, they created a training set of 8000 images and a testing set of 2000 images, on which they obtained 94.2% accuracy.

Seven papers focused on case retrieval, where AI techniques were used to present surgeons with images or video of cases similar to the case being actively reviewed or performed. Several motivations were cited for using case retrieval as an intraoperative aid. Quellec *et al*. (2011) suggested that automated case retrieval could help provide intraoperative warnings or recommendations based on real-time surgical video [21]. Focusing on the relatively new field of probe-based confocal laser endomicroscopy (pCLE), André *et al*. (2009) noted that "the taxonomy of pathologies for pCLE is still under active construction by the physicians" and suggested that retrieving cases with existing annotations and corresponding histopathological diagnoses could help surgeons make real-time decisions, such as whether or not to biopsy tissue [17].

Six of the case retrieval papers focused on pCLE. Four of the pCLE papers, published between 2009 and 2014, were published by Barbara André's group and utilized a colonic polyp data set from Mayo Clinic in Jacksonville, FL [15–18]. The other two pCLE papers were published by Yun Gu of Shanghai Jiao Tong University using a breast tissue data set [19,20]. An overview of the data set characteristics and algorithm performance measurements for the pCLE case retrieval papers are presented in Table 2.

André's work on pCLE case retrieval focused on methods for representing surgical videos and images. To identify similar cases, the surgical videos and images must be represented in a numerical form that allowed for the calculation of a distance metric (i.e., similarity) between cases. In 2009, André first used the Bag of Visual Words (BoW) method from computer vision to represent pCLE images [17]. In this method, established feature descriptor algorithms such as Scale-Invariant Feature Transform (SIFT) were used to extract visual features from the images, and "visual words" were identified through k-nearest neighbors clustering that grouped the extracted features from the data set. Taking the clinical context into account, André *et al*. selected feature extraction methods

Table 2 Data set characteristics and algorithm performance on pCLE case retrieval

Paper	Data set	Data set processing	Data set distribution	Validation	Performance[a]
André *et al*. 2009 [17]	54 patients; colonic polyps	1036 images; videos discarded if histology and pCLE diagnoses did not match	Roughly equal (2 classes)	Leave-n-out cross validation (patient held-out)	80.1% weighted k-NN classification accuracy (benign vs. pathological)
André *et al*. 2010 [15]	68 patients; colonic polyps	121 single-polyp videos	Roughly equal (2 classes)	Leave-one-patient-out cross-validation	94.2% weighted k-NN classification accuracy (benign vs. pathological)
André *et al*. 2012 [16]	66 patients; colonic polyps	118 single-polyp videos	Unspecified (8 binary classes)	30 × 3-fold cross-validation (patient segregated)	96.7% AUC for the highest-performing semantic concept ("elongated crypt") 49.4% Kendall τ rank correlation coefficient (Likert similarity scale)
Kohandani Tafresh *et al*. 2014 [18]	66 patients; colonic polyps	118 videos (mostly single-tissue)	Unbalanced (35 benign, 83 neoplastic)	Leave-one-patient-out cross-validation	89.9% k-NN classification accuracy 48.8% Spearman ρ correlation coefficient (Likert similarity scale)
Gu *et al*. 2016 [19]	50 patients; breast tissue	Unspecified	Unspecified (3 coarse classes, 8 sub-classes)	10-fold cross validation (not patient-segregated)	96.6% SVM classification accuracy (coarse class)
Gu *et al*. 2017 [20]	45 patients; breast tissue	700 pCLE mosaics, 144 matched with histology images	Unspecified (3 coarse classes, 8 sub-classes)	10-fold cross validation (not patient-segregated)	89.2% top-1 retrieval accuracy (coarse class) 96.2% top-5 retrieval accuracy (coarse class)

[a]Performance of the main algorithm described in the Methods section of each paper, not the tested baseline algorithms.

that were likely to register the blob-like shapes of goblet cells and crypts that physicians use to make colonic polyp diagnoses. André *et al*. also adopted the method to be translation and rotation invariant but not scale invariant, since the size of the features could inform a diagnosis.

In 2010, André *et al*. published an approach to improving this method through mosaicing [15]. Mosaicing refers to the process of combining multiple subsequent frames from a pCLE video related by viewpoint changes into a single image with an expanded field of view. In 2012, André *et al*. evolved the representation of pCLE mosaics from visual words into "semantic signatures" intended to reflect diagnostically relevant information [16]. To do so, André *et al*. identified eight binary "semantic concepts" that physicians use for *in vivo* colonic polyp pCLE diagnoses and then used the signed Fisher's criterion to estimate the expressive power of each visual word in the images, where the visual words were extracted with the BoW method. Using these Fisher weights, they transformed the visual words into semantic signatures that represented each video. A distance adjustment was then made to further distinguish the "very similar" videos from the other pairs, where the identification of "very similar" videos in the training data set came from expert ratings of pCLE video pairs using the four-point Likert scale ("very dissimilar," "rather dissimilar," "rather similar," or "very similar"). The 2012 paper also proposed visualizing each pCLE video as an intuitive "star plot" with eight vertices, where the edge lengths corresponded to the presence or absence of the eight binary diagnostic concepts.

Finally, in 2014, researchers from the same group presented a semi-automated method to speed the process of building a case retrieval query, noting that manual query construction was time-intensive and required expert knowledge [18]. To automate the query process, the algorithm proposed by Kohandani Tafresh *et al*. first temporally segmented the video based on the kinematic stability of the clips based on the assumption that the endoscopist spends more time in meaningful regions. The endoscopist was then intended to manually select a subset of the clips for case retrieval, which was accomplished using André *et al*.'s previously published methods. The overall goal of this work was to assist endoscopists in identifying colonic polyps by presenting them with prior similar cases of confirmed polyps, to assist them in deciding whether polypectomy was indicated.

Gu *et al*. (2016) built on the André group's work on pCLE video feature representation with the innovation of incorporating information from histology images into the pCLE mosaic representations, a method known as multimodal embedding [19]. The group utilized a mapping function to transform visual features extracted from breast tissue pCLE mosaics into a latent space by maximizing the semantic correlation between the mosaics and histology images. Feature extraction was accomplished using the SIFT, Texton, and histogram of oriented gradients (HoG) methods. To incorporate semantic meaning, the mapping function was trained using both coarse labels (neoplastic vs. non-neoplastic) and fine labels (i.e., tissue type and lesion characteristics) created by histopathological analysis. In 2017, the same group presented a novel graph-based approach for learning the pCLE features [20]. This method was intended to circumvent the difficulty of maintaining one-to-one pCLE-histology registration in the training set, since identifying the tiny pCLE field-of-view on the histological slide was time-consuming and difficult. In the graph-based method, only some of the pCLE mosaics had been directly paired to histology images. These registered mosaic-histology pairs formed "anchor nodes" in the graph while the unregistered histology images formed "patch nodes." Directed edges were created between nodes if the second node belonged in the k-nearest neighbors of the first, where the edge weight was the Euclidean distance between the visual features extracted by SIFT. For each anchor node, an n-order directed cycle was found, generating n positive pairs of pCLE mosaic-histology images matches. Negative pairs were found by calculating the geodesic distance (accumulated edge weights) along the shortest paths from each anchor node to each patch node and identifying those pairs whose geodesic distance was larger than a threshold. Gu *et al*. then mapped the pCLE and histology images to a latent feature space, where the mapping function was learned to preserve the positive and pair distances.

Evaluating the success of case retrieval algorithms was not straightforward. One approach was to use classification as a proxy evaluation method, although this was an imperfect measure of case similarity and was further complicated by the availability of different classifiers. For instance, the André group used k-nearest neighbors (k-NN) classification (i.e., majority vote of the top-n) [15,17,18], while Gu *et al*. (2016) used a Support Vector Machine (SVM) with the learned image features as input [19]. Another approach was to find the percentage of the top-n retrieved cases that belong to the starting case class [20]. A third approach was to measure the correlation of the algorithmic distances with expert judgments of case similarity [16,18].

For the single non-pCLE intraoperative case retrieval paper, Quellec *et al*. (2011) developed a fast method for retinal surgery video retrieval [21]. They did so by representing video clips as a single feature vector averaged from vector representations of individual video frames. After dimensionality reduction, the fixed vector length allowed fast calculations of inter-case distances within the video clip database. The algorithm was tested on 23 epiretinal membrane surgery videos divided into 69 clips that each contained one of three surgical steps. The area under the receiver operating characteristic (ROC) curve was assessed on a test set comprised of half the data set,

with an evaluation metric of the percentage of the five "closest" retrieved videos that showed the same surgical step as the test case.

Accelerating intraoperative pathology

Eight papers demonstrated potential for AI to provide an intraoperative pathological recommendation using one of four imaging modalities (Table 3). Three of the four modalities — pCLE, hyperspectral imaging (HSI), and optical coherence tomography (OCT) — are relatively new imaging methods for which clinical interpretation is still evolving (though OCT has been studied extensively in ophthalmology). However, all three have potential to be used for "optical biopsy" (*in vivo* diagnostic imaging), which could circumvent the need for the current time-consuming model of resection and frozen histopathological examination. The fourth modality, contrast-enhanced ultrasound (CEUS), has been clinically established in some fields. However, the researchers chose to focus on less-established neurosurgical applications of CEUS. The papers are summarized here, grouped by imaging modality.

Two papers attempted to localize residual glioblastoma tumor remaining after resection using intraoperative ultrasound (US) images. Both included CEUS, a technique in which a hyperechoic contrast agent is injected intravenously and the resulting distribution of the contrast agent is then recorded on US. The greatest concentrations of contrast agent appear in the regions of highest perfusion, which suggests the presence of tumor.

Ritschel *et al*. (2015) classified residual tumor regions by first fitting an equation modeling the distribution of contrast over time to the images and then running a classification algorithm using the equation parameters found through the fitting process [22]. They tested four different equations (gamma variate function model, bolus kinetic function model, bolus method model, and combined sigmoid function) and three different classifiers (linear discriminant analysis, soft-margin SVM with a Gaussian and soft-margin SVM with an ARD kernel). In addition, for each parameter, Ritschel *et al*. presented an image of the brain surface with colors to reflect the value of the parameter. This allowed surgeons to visually interpret the results. The data set consisted of 16 patients, but as no data acquisition protocol was agreed upon before the operations, three of the cases were discarded for part of the study due to problems with data acquisition (e.g., changing parameters during the reading, loss of surface contact).

Table 3 Summary of papers with the aim of accelerating intraoperative pathology

Paper	Imaging modality	Aim	Method
Tumor margin mapping (n = 3)			
Ritschel *et al*. (2015) [22]	CEUS	Localize glioblastoma tumor residuals	Latent Dirichlet analysis, support vector machine
Ilunga-Mbuyamba *et al*. (2017) [23]	CEUS	Localize glioblastoma tumor residuals	Data fusion
Fabelo *et al*. (2019) [31]	HSI	Map locations of glioblastoma tissue, normal tissue, hypervascularized tissue, and background material	Support vector machine, convolutional neural network/deep neural network joint architecture
Tumor classification (n = 3)			
Wan *et al*. (2015) [25]	pCLE	Distinguish between glioblastoma and meningioma	Feature descriptors, bag-of-visual-words dimensionality reduction, support vector machine
Kamen *et al*. (2016) [26]	pCLE	Distinguish between glioblastoma and meningioma	Feature descriptors, sparse coding with locality constraint to reduce dimensionality, support vector machine
Li *et al*. (2018) [27]	pCLE	Distinguish between glioblastoma and meningioma	Convolutional neural network, long short-term memory neural network
Tissue identification (n = 4)			
Couceiro *et al*. (2012) [28]	pCLE	Classify low or high probability of inflammatory bowel disease, based on intestinal crypts	Feature descriptors, support vector machine
Halicek *et al*. (2017) [29]	HSI	Classify normal or cancerous thyroid and aerodigestive tract tissues	Convolutional neural network
Halicek *et al*. (2018) [30]	HSI	Distinguish thyroid carcinoma from normal tissue	Convolutional neural network
Hou *et al*. (2019) [32]	OCT	Distinguish metastatic lymph nodes from normal lymph nodes	Artificial neural network

The algorithms were trained in two ways. First, the classifiers were trained using a leave-one-patient-out method. To provide the labels, three tumor and three non-tumor regions were manually identified 1 cm adjacent to the tumor border. The lowest classification error rate was produced using the combined sigmoid function, while the SVMs and the LDA did not produce significantly different results in this case. Second, on the 13 cases without data acquisition errors, the SVMs were trained on two tumor and two non-tumor regions from a single patient and then tested on the two other regions from the same patient. This was meant to simulate training on a patient during an operation. The authors suggested that the training process on four samples could be completed quickly enough (approximately 3 s) to be accomplished intraoperatively. The within-patient training method was used to segment the images, which achieved a mean precision of 0.71 with a standard deviation of 0.13 compared to manual segmentation completed by one neurosurgeon.

Ilunga-Mbuyamba *et al*. (2017) approached the same problem using a data fusion approach in which they combined the 3D CEUS images with 3D B-mode US imaging [23]. First, both sets of images were segmented to identify the border of the resection cavity in the B-mode images and the high-perfusion likely tumor residual regions in the CEUS images. Segmentation was accomplished using the Otsu thresholding method to identify different intensity classes (3 classes for B-mode, 4 classes for CEUS); in each case, the highest intensity class was preserved. The resulting images were then fused and the intersection of the high-intensity regions was classified as tumor residual. This was based on the expert knowledge that tumor residuals should be both found at the border of the resection cavity and supplied by denser-than-average vasculature. This study used a 23-patient data set, 19 of which contained residual tumor and four of which did not. Because this was a rules-based method based on expert knowledge without any need for algorithmic training, the method was evaluated on the entire 23-patient data set although experiments undertaken to set the ideal number classes for the thresholding algorithms were done on the full data set.

To test the algorithm, four neurosurgeons and scientists with more than seven years of experience with intraoperative brain tumor US manually segmented the images, aided by postoperative magnetic resonance imaging. The authors did not specify if input from multiple experts was used on each image; if this was the case, no statistics on variability between expert judgments were included. To qualitatively evaluate the method, they used the Overlap coefficient proposed by Dollar *et al*. [24] with a threshold of Overlap $\geqslant 0.5$ to consider the method a success. With this qualitative metric, they recorded successful tumor residual identification in 15 of the 19 patients with tumor residuals and false positives in 5 of the 23 patients, including 2 of the patients without any tumor residuals. On a voxel-wise basis, of the 15 patients in which success was achieved according to the Overlap coefficient, the average accuracy was 0.9507, average AUC was 0.7351, and average error was 0.0493. The authors argued that the Overlap coefficient was an appropriate metric because of the uncertainty underlying the manual annotations. They also noted that using shape descriptors, unprocessed 2D US images (unavailable at their institution) rather than the processed 3D images, or a semi-automatic method in which the surgeon manually annotates the tumor resection cavity could improve the segmentation scores.

Three papers reported algorithms that could distinguish between glioblastoma and meningioma in pCLE images and videos [25–27]. All three used the same data set from Merheim Hospital in Germany. Wan *et al*. (2015) [25] and Kamen *et al*. (2016) [26] used the full data set with 86 glioblastoma patient videos and 29 meningioma patient videos. Citing data ownership reasons, Li *et al*. (2018) [27] used a subset of the data with 16 glioblastoma videos and 17 meningioma videos. The videos were taken on excised stained brain tissue and histopathological analysis served as the ground truth. Both Wan *et al*. and Kamen *et al*. extracted image features using existing feature descriptor methods such as SIFT and then encoded the features into a reduced dimensionality representation. For the encoding method, Wan *et al*. (2015) modified the existing BoW method (see discussion in "case retrieval" related section above) while Kamen *et al*. (2016) developed a sparse coding method that incorporated a locality constraint. Wan *et al*. classified each video frame using an SVM. Kamen *et al*. also used an SVM, but instead used the majority vote prediction of surrounding frames (i.e., frames within a certain time range) to produce the final label. Li *et al*. instead used CNN and LSTM models to interpret the image and video data. Of the feature descriptor and coding method combinations tested by Wan *et al*., the Oriented FAST and Rotated BRIEF (ORB) feature descriptor with the Linear Locality Constraint coding method achieved the best performance, reported as 90% accuracy (no details of the validation or testing method were published). Kamen *et al*.'s locality-constrained sparse coding method produced an 84% accuracy on a test data set, the highest accuracy of the methods tested although it came at a significant computational cost. Li *et al*.'s best model achieved a 99.49% accuracy on the test set, using a 67% training, 17% validation, and 17% test set.

The fourth example of intraoperative pCLE pathology AI came from Couceiro *et al*. (2012), who classified low and high probability cases of inflammatory bowel disease (IBD) [28]. Couceiro *et al*.'s method took advantage of the diagnostic importance of intestinal crypt shapes. First, possible crypts were segmented using a frequency-based

approach to identify symmetric objects and an ellipse-fitting equation. After normalization and affine transformation to account for variation in endoscope perspective, image features were extracted using SIFT, texture analysis, and hand-crafted methods. An SVM was then used to determine whether or not each segmented object was a crypt. Once the crypts were detected, the images were represented as feature histograms and fed into a second SVM to determine whether or not the image suggested disordered or ordered crypt arrangement, which corresponded to high or low probability of IBD pathology, respectively. The data set consisted of a roughly equally distributed data set of 192 images from 18 patients. 10-fold cross-validation was used. For crypt detection, recall was high (0.99) but precision was very low (0.33 for the non-pathological and 0.07 for the pathological data subsets), indicating over-detection. Using the hand-crafted features achieved higher performance than the texture analysis methods and similar performance to the SIFT descriptors. On image classification, Couceiro *et al*. achieved 0.89 accuracy compared to 0.71 accuracy using the BoW method discussed above.

Three papers focused on HSI [29–31]. Two related papers from Halicek *et al*. used convolutional neural networks to distinguish between hyperspectral images of head and neck cancers. In both cases, the imaging was done *ex vivo* on fresh surgical specimens. In 2017, Halicek *et al*. reported 80% accuracy at distinguishing cancerous from normal tissues in a database of thyroid and aerodigestive tract hyperspectral images [29]. When the data set was divided according to tissue type, the CNN achieved 77% accuracy on aerodigestive tract tissues and 90% accuracy on thyroid tissues. The data set consisted of 88 excised tissue samples from 50 patients, 29 of whom had squamous cell carcinoma and 21 of whom had thyroid carcinoma. A head-and-neck specialized pathologist created the gold standard labels. A leave-one-patient-out external validation scheme was used. In 2018, Halicek *et al*. used a modified CNN architecture to again distinguish thyroid carcinoma from normal tissue [30]. They reported a 92% accuracy using an 11-patient data set with a leave-one-out validation scheme.

Fabelo *et al*. (2019) compared the abilities of an SVM and a multi-step deep learning framework to distinguish between four classes in HSI images: glioblastoma tumor tissue, normal tissue, hypervascularized tissue (i.e., blood vessels), and background (i.e., bone/dura/skin/surgical material) [31]. First, three HSI spectral channels (wavelengths) were selected to create a gray-scale blood vessel map. Second, a training data set of 20 manually-segmented images, augmented by eight using rotations and reflections, was used to train a 2D-CNN to identify brain parenchymal regions; on an eight-image test set, a Dice similarity coefficient of 86.5% was achieved. Third, a 1D-DNN then classified the hyperspectral data into the four classes. Finally, the four-class map was merged with the blood vessel map and parenchymal map to create the final product. The authors compared this methodology to the existing spatial-spectral classification algorithm. The spatial-spectral algorithm relied on an SVM classifier after principal component analysis (PCA) was used to reduce the dimensionality of the hyperspectral cubes. Various kernel methods for the SVM were tested, as was a binary classification task (tumor vs. normal tissue) on a subset of the data. The data set was comprised of tumor images from 6 patients, normal and hypervascularized images from 16 patients, and background images from 15 patients. Bootstrapping was used to balance class distribution and obtain a confidence range. On the binary classification test, the 1D-DNN achieved the highest performance with 94% accuracy calculated with leave-one-patient-out cross-validation.

Finally, one paper used AI to classify OCT images. Hou *et al*. (2019) used an ANN to distinguish metastatic lymph nodes from normal tissues in thyroidectomy patients on images of resected neck tissues taken using OCT [32]. Texture features were extracted and then ranked by the ratio of in-class to between-class scatter. The 14 top-ranked features were used as input for the ANN with results demonstrating 90.1% accuracy from a data set of 573 images from 28 patients.

Recommending a surgical step

While the technologies discussed above were crafted with the intent of providing surgeons with more context and information to incorporate into their intraoperative decision making process, two papers framed their work as capable of making a direct recommendation for an intraoperative decision. In one case, the researchers were motivated by an interest in mitigating the shortage of experts in a particular surgery and thereby expanded access to surgical care [33]. In the other case, the proposal to rely on algorithmic rather than human judgment arose partly in response to the difficulty of training experts capable of interpreting a nuanced intraoperative signal and partly in response to the need to mitigate certain human factors that can negatively impact surgeon decision making [34].

Fan *et al*. (2016) developed a model that uses a set of non-surgical factors affecting spinal cord function to predict the amplitude of the somatosensory evoked potential (SEP) during spinal surgeries [34]. Lowered SEP amplitudes can indicate spinal trauma during surgery, but the SEP amplitude may also drop in response to anesthesia and other variables that vary in the operating room context without reflecting spinal trauma (a "false alarm").When a SEP amplitude below 50% of the baseline is detected, the common practice is to terminate the operation, awaken the patient from anesthesia, and assess their neurological function. If the low SEP was due to a

false alarm, the patient now has to undergo induction of anesthesia again to complete the operation. Although human experts can interpret the SEP with greater nuance, Fan *et al*. argued that such experts are hard to train and are also vulnerable to fatigue (with corresponding inattention) and emotion. By predicting the SEP, the algorithm created a dynamic SEP baseline that could potentially help decrease the number of prematurely truncated spinal surgeries. To create the baseline prediction, Fan *et al*. developed a method called Multi-Support Vector Regression. Using training data from non-traumatic and non-false alarm surgical cases, the authors combined several Least Squares Support Vector Regression models separately trained on sub-datasets, which were created through a clustering and resampling process from the training data. The clustering and resampling process was intended to make the models more robust to noise. Fan *et al*. proposed using a SEP amplitude three standard deviations below the dynamic baseline as the threshold for indicating spinal trauma, where the standard deviation of the SEP amplitude was determined using variability within the training data. The training dataset included 10 successful (true-negative) surgeries, four false alarm (false-positive) cases, and one trauma (true-positive) case. The algorithm was trained via a leave-one-out methodology on the 10 successful surgeries; each testing round was completed on the left-out successful case as well as the false alarm and trauma cases. The multi-support vector regression model achieved the best performance with a low amplitude warning rate of 6.79% on the false positive cases, 3.27% on the successful cases, and 71.43% on the trauma case compared to respective warning rates of 30.42%, 5.70%, and 50% using the baseline method.

Tian *et al*. (2015) developed a system named VeBIRD to track and classify cataract grade on videos of phacoemulsification surgeries [33]. They proposed that the system could eventually be used to control the amount of ultrasonic energy released to emulsify the cataracts, a decision currently made by experienced ophthalmological surgeons. Tian *et al*. noted that this method could help increase access to phacoemulsification surgeries, especially in small and rural hospitals that lack experienced specialists. Eye detection was accomplished using ellipse detection with a modification of the Hough transform to increase robustness to noise and shape distortion. A version of the tracking-learning-detection algorithm was used for probe-tracking and an SVM classified the cataract grade. Using a 50-50 split of 2000 annotated frames, the authors reported a 92.3% accuracy at eye detection and 96.3% classification accuracy on cataract grade. No steps were taken to segregate the frames by patient, and the probe-tracking performance was only evaluated on 5 videos. The relationship between cataract grade and ultrasonic energy release was left for future work.

Discussion

This scoping review found that research into artificial intelligence for intraoperative decision support is still in its infancy. No AI method or surgical specialty yet dominates the field, nor is it likely that any single topic or method will emerge as the dominant methodology, given the complexity and diversity of surgery and the rapid evolution of surgical AI. Successfully integrating the proposed AI tools into the operating room will depend on the ability of researchers to identify clear points in the surgical decision making process where the insertion of technology could helpfully augment human capabilities. Additionally, in some papers we observed misalignment between the application of AI technologies and current clinical practice standards that could limit the safety and efficacy of such technologies in real-world implementation. Our analysis also uncovered key methodological shortcomings in several papers that hinder fair evaluation and interpretation of the algorithms and the data on which they were trained and tested.

With regard to the papers that cited the goal of directly influencing decisions about the next surgical step [33,34], it is important to note that the described technologies would not accomplish decision support by explicitly directing the next step; rather, the potential role of these technologies was to provide additional data on which a surgeon could act. These studies demonstrated that the utilization of machine learning to analyze complex data streams such as SEP and surgical video could provide surgeons with access to valuable — otherwise potentially inaccessible or less interpretable — data to incorporate into their decisions. This suggests that a framework of decision *augmentation* rather than *automation* is likely to be the first, and perhaps most effective, route to achieving incorporation of AI into surgical decision making.

For AI technologies seeking to influence surgical decision making, explainable AI, algorithms that provide evidence to support their predictions, will be a critical component of translating research to clinical applications [35]. Conceding judgment to an algorithm without understanding its interpretations or implications for patients does nothing to improve access to quality surgical care. Without some explanation of an AI system's predictions/recommendations, physicians would be asked to place blind trust in the algorithmic recommendations, without human clinical experience and judgment to contextualize the recommendations. Ritschel's approach on residual tumor classification using CEUS was an example of providing data to neurosurgeons in an explainable manner (i.e., presenting color maps reflecting the different probabilities of predictions) that allows for decision augmentation for responsible clinical practice [22].

Some of the papers rationalized their focus on new

imaging modalities as a way to help physicians make correct diagnoses despite a lack of expertise in the imaging modality at hand [16,25,27]. Exploration of new imaging modalities (pCLE) or new applications of existing modalities (e.g., OCT, CEUS) may lead to new understandings of disease pathology, which could be an important basis for the development of new therapies. Research should focus on settings where AI can expand human capabilities or overcome systemic limitations. Finally, from a methodological standpoint, developing these technologies using established imaging modalities would allow researchers to have more confidence in their algorithms' performance. For instance, the expert-created similarity scores fundamental to the case retrieval algorithms would be more reliable if pCLE images were more thoroughly understood.

Notably, during the title review phase of our review process, a number (n = 24) of intraoperative AI-related titles were found on topics that were not directly related to surgical decision making. Specifically, 16 papers on surgical tool tracking and eight papers on surgical phase detection were identified with decision support-related keywords. While these topics may have downstream applications in surgical robot development, surgical simulation, or surgeon training, they did not meet inclusion criteria for current decision support in the real-time operating room setting. For instance, identification of a tool on the screen during laparoscopic surgery may be unlikely to affect decision making with that tool as the surgeon has selected the tool based on decisions made before it ever appears on the screen; however, tool identification could play a role in downstream selection of instruments or in inventory management.

For researchers who aim to help surgeons perform operations more safely and efficiently in real-time, careful analyses of surgical workflows and decision making processes could help identify points where technological supplementation could be useful [6,36]. The majority of papers identified in this review focused on improving surgeons' situational assessment by providing additional, quantitative data to assist in making decisions on which action to take (e.g., additional dissection or selecting an area of resection). Some of these applications can also help with re-evaluation (e.g., providing data on possible margins). We would suggest that while more work in assisting situational assessment should be encouraged, a particular area of need is in assisting surgeons with decision making for actions.

Flin *et al*.'s framework describes the intraoperative decision making process of surgeons as being intuitive, rule-based, comparative, or creative, where creative thinking is particularly applicable to rare, high pressure situations where novel decisions must be made on limited data [6]. AI technologies can be investigated to assess the potential for intervention and augmented decision making in each of these types of decision processes, and researchers may find that different surgeons respond best to different types of decision-support based on the clinical scenario at hand. There is anecdotal interest within the clinical community in improving access to data for scenarios where creative decision making is needed — rare clinical scenarios where the cost of an imperfect decision can be high (e.g., bile duct injury). However, generating sufficient data for AI to be helpful in these scenarios requires greater access to data.

While recognizing that this field is still in its infancy, we noted several methodological shortcomings across this body of work. First, most of the algorithms were developed and tested on small data sets taken from single institutions. No analysis was done to determine whether the data sets were representative of the patient populations the researchers intended to treat. Larger, multi-institutional training data sets — while recognizing the difficulty in developing or accessing such data sets — could help improve the accuracy and generalizability of these machine learning algorithms.

Second, most groups did not account for the uncertainty and variability of medical diagnoses and decision making because they relied on single experts to produce the gold standard. Within other areas of medical AI development, it is common practice to involve multiple physicians in labeling the data sets and reporting inter-observer variability [7,37–39]. As an example, Hashimoto *et al*. (2019) found that expert surgeons annotating video within the same surgical practice differed on their conceptualization of steps of laparoscopic sleeve gastrectomy [7]. Without a clinically applicable ground truth, the reported statistics for algorithm performance are difficult to compare or translate to the clinical setting.

Similarly, quantified error analysis will be important for directing future research and helping physicians interpret the outputs of AI tools. Most of the papers that we reviewed overlooked the question of error analysis. In contrast, Tian *et al*. noted that most of the errors made by VeBIRD were between adjacent cataract grades, which they suggested was also a source of disagreement among expert surgeons [33]. While this was an important assertion, the failure to quantify this assertion or compare it to inter-expert variability made it impossible to determine how closely VeBIRD approximated human expert judgment.

Third, only some researchers segregated the training and test data by patient. This poses the risk of artificially inflating algorithm performance, since the training and test data sets were not independent. In fact, Halicek *et al*. (2017) found that adding a known tissue type sample from the test patient into the training data set would significantly increase the classifier's performance, and Ritschel *et al*. (2015) proposed training an algorithm using a small data set (4 samples) of known tumor and known non-tumor

tissue from the same patient during the operation [22,29]. While obtaining a known sample from a patient might be a valid way to achieve more accurate results in the research and development setting, it may be impractical in some applications within the clinical environment. For instance, Halicek's proposal required adding an additional invasive procedure (a biopsy) into the preoperative workflow if it was not already performed. The benefits of the technology would need to be carefully weighed against the additional risks incurred by the patient. While Ritschel's proposal is less invasive, a balance would need to be struck between training time and sufficient performance. While these proposals are interesting, it is unlikely that patient-specific training will be workable in all surgical scenarios. Such tools need to clearly specify their approaches, applications, and proposed integration into the clinical workflow. Outside of these special cases, the training and test data should be segregated at the patient level, and cross-validation methods should be used to determine hyper-parameters in order to avoid overfitting.

This review has several limitations. First and foremost, we conducted a scoping review with the intent of better understanding the variety of technologies that have been investigated for AI-based intraoperative decision support. By design, scoping reviews do not compare different methods or results so no determination was made about which approaches, if any, might be better or worse than another. We maintained strict inclusion/exclusion criteria; thus, some papers may have been missed. Disappointingly, two papers on interesting intraoperative decision support applications were excluded from our review because they did not specify the number of patients in their database. While simply recording the number of images may be sufficient in other fields, patient-level information should be included for medical computer vision studies to better understand how generalizable results might be to a wider patient population. While we ultimately limited ourselves to intraoperative applications for this work, we identified a significant number of preoperative and postoperative decision support AI studies. Analyses of these studies will be important for gaining a more complete understanding of the current state of research on AI-based surgical decision support.

Despite these limitations, our review highlighted several key findings that we hope can help guide the future of research in AI-based decision support. Much of the intraoperative AI research has thus far been focused on technology without immediate translational applications in surgical decision support. We also identified important methodological shortcomings, including frequent failure to segregate training and testing data at the patient level and a general lack of attention to the applicability of the data sets that were used to represent the patient populations of interest. Interestingly, much of the image- and video-related research focused on newer imaging modalities that are not yet widespread in clinical practice. While these studies provide novel approaches and applications to AI-based decision support, more research is needed in well-established and frequently utilized imaging modalities to maximize clinical impact.

This review also further highlighted the need to create large, multi-institutional data sets with standardized annotation and data storage frameworks. Currently, most surgical data sets are sourced from single institutions using proprietary annotations, and few of these data sets are easily combined. While some nationwide efforts such as the American College of Surgeons National Surgical Quality Improvement Program (NSQIP) provide the framework for multi-institutional, standardized data capture of health record information, significant effort is still required to create data sets of surgical images and videos. Such an effort will require addressing important fundamental concerns on data privacy, ownership, and ethical use [40]. From a logistical perspective, an agreed upon framework or standards of annotation and data storage need to be developed to allow concatenation or combination of data sets across institutions. Researchers will need to consider whether or not their population of interest is actually included (and accurately represented) in their selected data set so data transparency (with prioritization of patient privacy) will need to be addressed. Ultimately, if AI-based intraoperative decision support is to be implemented clinically, collaboration across institutions will likely be required to provide the large and balanced data sets necessary to ensure that the benefits of the technology will be both equitable and reliable.

Conclusions

The goal of intraoperative AI should be the improvement of patient care. In this review, we uncovered several papers that did an excellent job of finding a suitable point of intervention within the surgical workflow where the addition of AI could provide value to the patient. However, the field is in its infancy, and future work can be structured to maximize the potential for clinical applicability. Future research should focus on ensuring that data sets are representative of the patient populations of interest, have appropriate and clinically applicable ground truth, and are validated in ways that are representative of clinical use.

Compliance with ethics guidelines

Allison J. Navarrete-Welton has received research support from Olympus Corporation for projects outside of this paper. Daniel A. Hashimoto is an independent consultant for Verily Life Sciences and the Johnson & Johnson Institute. He serves on the clinical advisory board of Worrell, Inc. He has received research support from Olympus Corporation for projects outside of this paper. This

manuscript is a review article and does not involve a research protocol requiring approval by the relevant institutional review board or ethics committee.

Electronic Supplementary Material Supplementary material is available in the online version of this article at https://doi.org/10.1007/s11684-020-0784-7 and is accessible for authorized users.

References

1. Spencer F. Teaching and measuring surgical techniques: the technical evaluation of competence. Bull Am Coll Surg 1978; 63: 9–12
2. Suliburk JW, Buck QM, Pirko CJ, Massarweh NN, Barshes NR, Singh H, Rosengart TK. Analysis of human performance deficiencies associated with surgical adverse events. JAMA Netw Open 2019; 2(7): e198067
3. Pugh CM, Santacaterina S, DaRosa DA, Clark RE. Intra-operative decision making: more than meets the eye. J Biomed Inform 2011; 44(3): 486–496
4. Hashimoto DA, Axelsson CG, Jones CB, Phitayakorn R, Petrusa E, McKinley SK, Gee D, Pugh C. Surgical procedural map scoring for decision-making in laparoscopic cholecystectomy. Am J Surg 2019; 217(2): 356–361
5. Pugh CM, DaRosa DA. Use of cognitive task analysis to guide the development of performance-based assessments for intraoperative decision making. Mil Med 2013; 178(10 Suppl): 22–27
6. Flin R, Youngson G, Yule S. How do surgeons make intraoperative decisions? Qual Saf Health Care 2007; 16(3): 235–239
7. Hashimoto DA, Rosman G, Witkowski ER, Stafford C, Navarette-Welton AJ, Rattner DW, Lillemoe KD, Rus DL, Meireles OR. Computer vision analysis of intraoperative video: automated recognition of operative steps in laparoscopic sleeve gastrectomy. Ann Surg 2019; 270(3): 414–421
8. Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial intelligence in surgery: promises and perils. Ann Surg 2018; 268 (1): 70–76
9. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018; 18(8): 500–510
10. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. Nat Rev Clin Oncol 2019; 16(11): 703–715
11. Hogarty DT, Su JC, Phan K, Attia M, Hossny M, Nahavandi S, Lenane P, Moloney FJ, Yazdabadi A. Artificial intelligence in dermatology—where we are and the way to the future: a review. Am J Clin Dermatol 2020; 21(1): 41–47
12. Maier-Hein L, Vedula SS, Speidel S, Navab N, Kikinis R, Park A, Eisenmann M, Feussner H, Forestier G, Giannarou S, Hashizume M, Katic D, Kenngott H, Kranzfelder M, Malpani A, März K, Neumuth T, Padoy N, Pugh C, Schoch N, Stoyanov D, Taylor R, Wagner M, Hager GD, Jannin P. Surgical data science for next-generation interventions. Nat Biomed Eng 2017; 1(9): 691–696
13. Udelsman R, Donovan P, Shaw C. Cure predictability during parathyroidectomy. World J Surg 2014; 38(3): 525–533
14. Harangi B, Hajdu A, Lampe R, Torok P. Recognizing ureter and uterine artery in endoscopic images using a convolutional neural network. In: 2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS). 2017. 726–727. doi: 10.1109/CBMS.2017.137
15. André B, Vercauteren T, Buchner AM, Wallace MB, Ayache N. Endomicroscopic video retrieval using mosaicing and visualwords. In: 2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. 2010. doi: 10.1109/isbi.2010.5490265
16. André B, Vercauteren T, Buchner AM, Wallace MB, Ayache N. Learning semantic and visual similarity for endomicroscopy video retrieval. IEEE Trans Med Imaging 2012; 31(6): 1276–1288
17. André B, Vercauteren T, Perchant A, Buchner A, Wallace M, Ayache N. Endomicroscopic image retrieval and classification using invariant visual features. In: 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. 2009. doi: 10.1109/isbi.2009.5193055
18. Kohandani Tafresh M, Linard N, André B, Ayache N, Vercauteren T. Semi-automated query construction for content-based endomicroscopy video retrieval. In: Medical Image Computing and Computer-Assisted Intervention — MICCAI 2014. Springer International Publishing, 2014. 89–96. doi: 10.1007/978-3-319-10404-1_12
19. Gu Y, Yang J, Yang GZ. Multi-view multi-modal feature embedding for endomicroscopy mosaic classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2016. 11–19
20. Gu Y, Vyas K, Yang J, Yang GZ. Unsupervised feature learning for endomicroscopy image retrieval. In: Medical Image Computing and Computer Assisted Intervention — MICCAI 2017. Springer International Publishing, 2017. 64–71 doi: 10.1007/978-3-319-66179-7_8
21. Quellec G, Lamard M, Cazuguel G, Droueche Z, Roux C, Cochener B. Real-time retrieval of similar videos with application to computer-aided retinal surgery. Conf Proc IEEE Eng Med Biol Soc 2011; 2011: 4465–4468
22. Ritschel K, Pechlivanis I, Winter S. Brain tumor classification on intraoperative contrast-enhanced ultrasound. Int J CARS 2015; 10 (5): 531–540
23. Ilunga-Mbuyamba E, Lindner D, Avina-Cervantes J, Arlt F, Rostro-Gonzalez H, Cruz-Aceves I, Chalopin C. Fusion of intraoperative 3D B-mode and contrast-enhanced ultrasound data for automatic identification of residual brain tumors. Appl Sci (Basel) 2017; 7(4): 415
24. Dollar P, Tu Z, Perona P, Belongie S. Integral channel features. In: Procedings of the British Machine Vision Conference. 2009. doi: 10.5244/c.23.91
25. Wan S, Sun S, Bhattacharya S, Kluckner S, Gigler A, Simon E, Fleischer M, Charalampaki P, Chen T, Kamen A. Towards an efficient computational framework for guiding surgical resection through intra-operative endo-microscopic pathology. In: Medical Image Computing and Computer-Assisted Intervention — MICCAI 2015. Springer International Publishing, 2015. 421–429. doi: 10.1007/978-3-319-24553-9_52
26. Kamen A, Sun S, Wan S, Kluckner S, Chen T, Gigler AM, Simon E, Fleischer M, Javed M, Daali S, Igressa A, Charalampaki P. Automatic tissue differentiation based on confocal endomicroscopic

images for intraoperative guidance in neurosurgery. BioMed Res Int 2016; 2016: 6183218

27. Li Y, Charalampaki P, Liu Y, Yang GZ, Giannarou S. Context aware decision support in neurosurgical oncology based on an efficient classification of endomicroscopic data. Int J CARS 2018; 13(8): 1187–1199
28. Couceiro S, Barreto JP, Freire P, Figueiredo P. Description and classification of confocal endomicroscopic images for the automatic diagnosis of inflammatory bowel disease. In: Machine Learning in Medical Imaging. Springer Berlin Heidelberg, 2012. 144–151. doi: 10.1007/978-3-642-35428-1_18
29. Halicek M, Lu G, Little JV, Wang X, Patel M, Griffith CC, El-Deiry MW, Chen AY, Fei B. Deep convolutional neural networks for classifying head and neck cancer using hyperspectral imaging. J Biomed Opt 2017; 22(6): 60503
30. Halicek M, Little JV, Wang X, Patel M, Griffith CC, El-Deiry MW, Chen AY, Fei B. Optical biopsy of head and neck cancer using hyperspectral imaging and convolutional neural networks. Proc SPIE Int Soc Opt Eng 2018; 104690X doi: 10.1117/12.2289023
31. Fabelo H, Halicek M, Ortega S, Shahedi M, Szolna A, Piñeiro JF, Sosa C, O'Shanahan AJ, Bisshopp S, Espino C, Márquez M, Hernández M, Carrera D, Morera J, Callico GM, Sarmiento R, Fei B. Deep learning-based framework for *in vivo* identification of glioblastoma tumor using hyperspectral images of human brain. Sensors (Basel) 2019; 19(4): 920
32. Hou F, Liang Y, Yang Z, Gu W, Yu Y. Automatic identification of metastatic lymph nodes in OCT images. Proceedings Volume 10867, Optical Coherence Tomography and Coherence Domain Optical Methods in Biomedicine XXIII; 108673G. 2019. doi: 10.1117/12.2511588
33. Tian S, Yin XC, Wang ZB, Zhou F, Hao HW. A VidEo-Based Intelligent Recognition and Decision System for the phacoemulsification cataract surgery. Comput Math Methods Med 2015; 2015: 202934
34. Fan B, Li HX, Hu Y. An intelligent decision system for intraoperative somatosensory evoked potential monitoring. IEEE Trans Neural Syst Rehabil Eng 2016; 24(2): 300–307
35. Gordon L, Grantcharov T, Rudzicz F. Explainable artificial intelligence for safe intraoperative decision support. JAMA Surg 2019; 154(11): 1064
36. Lalys F, Jannin P. Surgical process modelling: a review. Int J CARS 2014; 9(3): 495–511
37. Krause J, Gulshan V, Rahimy E, Karth P, Widner K, Corrado GS, Peng L, Webster DR. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology 2018; 125(8): 1264–1272
38. Lee H, Yune S, Mansouri M, Kim M, Tajmir SH, Guerrier CE, Ebert SA, Pomerantz SR, Romero JM, Kamalian S, Gonzalez RG, Lev MH, Do S. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. Nat Biomed Eng 2019; 3(3): 173–182
39. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017; 542(7639): 115–118
40. Safdar NM, Banja JD, Meltzer CC. Ethical considerations in artificial intelligence. Eur J Radiol 2020; 122: 108768