tvst

Review

Artificial Intelligence in Cataract Surgery: A Systematic Review

Simon Müller[1], Mohit Jain[2], Bhuvan Sachdeva[2,3], Payal N. Shah[3], Frank G. Holz[1], Robert P. Finger[1,4], Kaushik Murali[3], Maximilian W. M. Wintergerst[1,6,*], and Thomas Schultz[5,7,*]

[1] University Hospital Bonn, Department of Ophthalmology, Bonn, Germany
[2] Microsoft Research, Bengaluru, India
[3] Sankara Eye Hospital, Bengaluru, Karnataka, India
[4] Department of Ophthalmology, University Medical Center Mannheim, Heidelberg University, Mannheim, Germany
[5] B-IT and Department of Computer Science, University of Bonn, Bonn, Germany
[6] Augenzentrum Grischun, Chur, Switzerland
[7] Lamarr Institute for Machine Learning and Artificial Intelligence, Dortmund, Germany

Correspondence: Maximilian W. M. Wintergerst, University Hospital Bonn, Department of Ophthalmology, Venusberg-Campus 1, Bonn 53127, Germany. e-mail: maximilian.wintergerst@ukbonn.de

Received: November 15, 2023
Accepted: February 12, 2024
Published: April 15, 2024

Keywords: cataract surgery; phacoemulsification; deep learning; convolutional neural networks; recurrent neural networks; video analysis; tracking; ophthalmology; vision transformers; risk assessment; surgical outcomes; review

Citation: Müller S, Jain M, Sachdeva B, Shah PN, Holz FG, Finger RP, Murali K, Wintergerst MWM, Schultz T. Artificial intelligence in cataract surgery: A systematic review. Transl Vis Sci Technol. 2024;13(4):20, https://doi.org/10.1167/tvst.13.4.20

Purpose: The purpose of this study was to assess the current use and reliability of artificial intelligence (AI)-based algorithms for analyzing cataract surgery videos.

Methods: A systematic review of the literature about intra-operative analysis of cataract surgery videos with machine learning techniques was performed. Cataract diagnosis and detection algorithms were excluded. Resulting algorithms were compared, descriptively analyzed, and metrics summarized or visually reported. The reproducibility and reliability of the methods and results were assessed using a modified version of the Medical Image Computing and Computer-Assisted (MICCAI) checklist.

Results: Thirty-eight of the 550 screened studies were included, 20 addressed the challenge of instrument detection or tracking, 9 focused on phase discrimination, and 8 predicted skill and complications. Instrument detection achieves an area under the receiver operator characteristic curve (ROC AUC) between 0.976 and 0.998, instrument tracking an mAP between 0.685 and 0.929, phase recognition an ROC AUC between 0.773 and 0.990, and complications or surgical skill performs with an ROC AUC between 0.570 and 0.970.

Conclusions: The studies showed a wide variation in quality and pose a challenge regarding replication due to a small number of public datasets (none for manual small incision cataract surgery) and seldom published source code. There is no standard for reported outcome metrics and validation of the models on external datasets is rare making comparisons difficult. The data suggests that tracking of instruments and phase detection work well but surgical skill and complication recognition remains a challenge for deep learning.

Translational Relevance: This overview of cataract surgery analysis with AI models provides translational value for improving training of the clinician by identifying successes and challenges.

Introduction

Cataract surgeries are the most frequently performed procedures worldwide and are often digitally recorded.[1–3] The availability of video material and the standardized nature of the surgery presents a huge opportunity for automatic analysis for quality management, teaching, and training. Recent advances in artificial intelligence (AI), especially deep learning (DL), further enable this automation.

The DL algorithms have demonstrated unparalleled potential in revolutionizing various aspects of cataract surgery, from pre-operative diagnostics and planning to intra-operative guidance and postoperative care.[4] The integration of AI in cataract surgery holds the promise of improved surgical precision, enhanced patient outcomes, and increased efficiency for healthcare practitioners. One example of this in general surgery is the work by Pierre Jannin et al. demonstrating the usefulness of AI for skills assessment by predicting values on an established skill metric for surgeon training programs.[5]

This review article aims to provide a comprehensive assessment of the current state of DL algorithms in analyzing cataract surgeries. It groups available algorithms according to the task that they solve, and compares them with respect to the reported quality of results. We also investigate the reproducibility and reliability of the prior study results due to the known problems in replication of DL algorithms.[6]

Treatment for cataract is surgical and a variety of surgical techniques are available for this, including phacoemulsification (PE), manual small-incision cataract surgery (MSICS), and femto laser-assisted cataract surgery (FLACS).[7] Hence, the type of cataract surgery needs to be considered in addition to a number of additional parameters in any of the available datasets for training of DL architectures.

The earliest publications started out with the recognition of coarse phases in cataract surgeries and detection of a limited number of instruments. Current techniques offer finer recognition of object classes, their position, and more sophisticated stages. Thus, besides providing a narrative overview, we will critically examine the main challenges and limitations faced by the presented approaches and their data sources and discuss potential avenues for future research to overcome these hurdles.

Methodology

The overall methodology of this review is based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) and the Cochrane recommendations for systematic reviews.[8,9]

Eligibility Criteria for Considering Studies for This Review

Search Methods for Identifying Studies

To search relevant studies, we used the search engines available on PubMed, Clarivate Web of Science, Elicit.org, and dblp computer science bibliography, encompassing the following databases that include the most important journals in the fields of medical and computer science research: MEDLINE, MEDLINE In-Process, Science Citation Index Expanded (SCIE), Conference Proceedings Citation Index - Science, Book Citation Index - Science, Emerging Sources Citation Index (ESCI), Scientific Electronic Library Online Citation Index, and arXiv.org for published studies up to July 2023.

Detailed information about the search terms and formulas can be found in Supplementary Table S1. The initial search across the 4 search engines yielded 550 articles. After screening all paper abstracts for eligibility, 49 references were included in the full-text review (Fig. 1). Study authors were contacted to provide additional data if required. Reference lists of manuscripts reviewed in full were hand searched for additional relevant articles.

Study Selection

For the inclusion in our review, we only considered algorithms that were evaluated on cataract extraction surgery datasets with quantitative performance metrics. Datasets must comprise of real surgery videos, and not on surgeries performed on plastic models or be artificially generated. We did not consider studies published more than 10 years ago due to the relative recency of successful computer vision solutions in the field of ophthalmology. Additionally, a huge body of publications utilizes convolution neural networks to diagnose the presence and severity of cataract in patient populations. We excluded this specific application for our review because it is covered thoroughly in prior work[4,10] and instead focused our review on the surgery video analysis and postoperative quality insurance. The final set of included papers covers the detection, segmentation, and tracking for surgical instruments, the recognition of surgical phases, and the assessment of surgical skill and risk for complications.

In case of any uncertainty regarding the in- or exclusion of a specific research paper, either a senior computer scientist (author T.S.) or ophthalmologist (author M.W.W.M.) were consulted.

Data Collection and Quality Assessment

After the initial screening and eligibility assessment, the following characteristics of the included studies were extracted:

- Study design, type of cataract surgery, type of machine learning algorithm and neural network (if applicable), reported measurement metrics of the algorithm, recording equipment, dataset details,

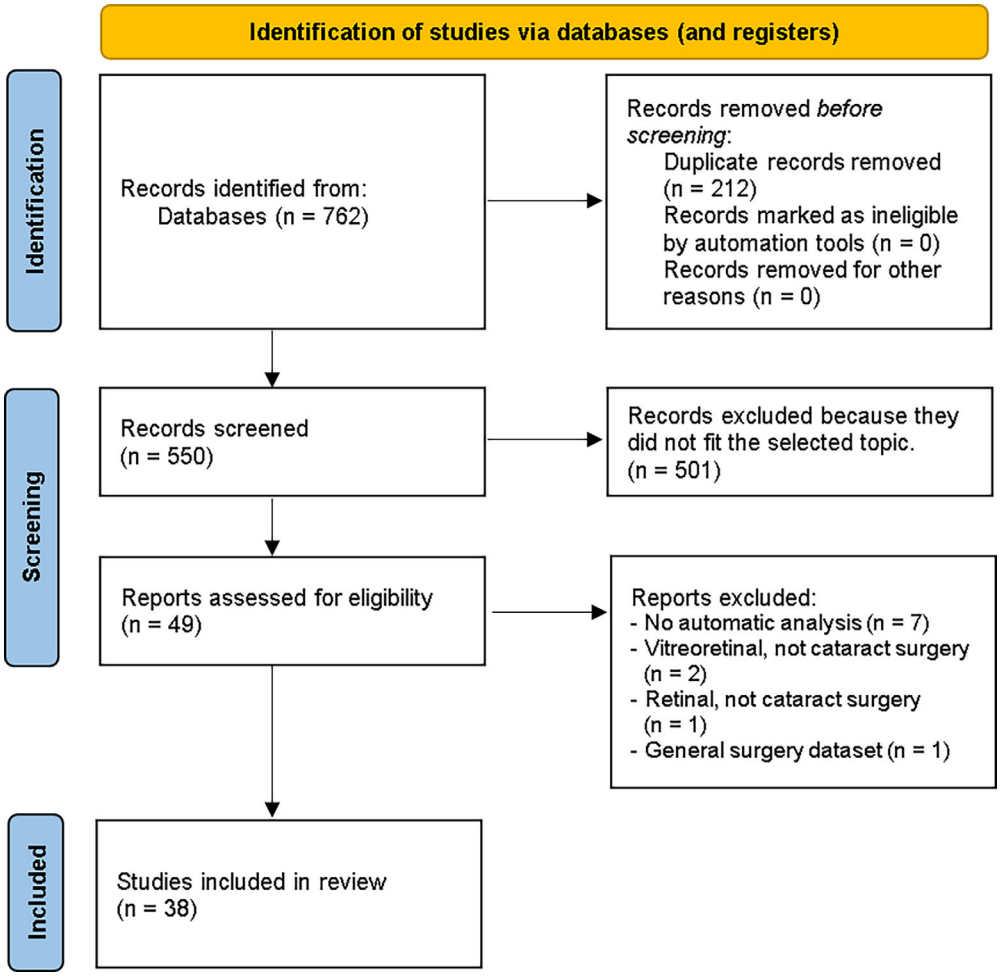

Figure 1. Flowchart based on PRISMA 2020 diagram for new systematic reviews depicting the literature search on algorithms for analysis of cataract extraction surgery.[9]

which phases, instruments and/or skill/complication types were tracked, validity, reliability, limitations, key conclusions of the authors, and other relevant information.

One independent assessor (author S.M.) used a modified version of the International Conference on Medical Image Computing and Computer-Assisted Interventions (MICCAI) Reproducibility Checklist to assess the reproducibility of the studies (see Supplementary Table S1 for details). Wintergerst et al. raised the issue of replicating results from automated ophthalmic image analysis in independently collected data and we aim to assess this risk in cataract surgery using this list.[6] The checklist was adapted to the requirements of cataract surgery analysis by including the following: description of the mathematical setting, algorithm and model, characteristics of the study cohort making up the dataset, description of data collection and parameters of used devices or tools, publication of the dataset, pre-trained models and codebase, reporting of hyperparameters and their selection procedure, details on hardware used and its computing performance, and whether there was a statistical analysis of results.

Data Synthesis and Analysis

Algorithms were categorized based on which aspect(s) of the surgery they analyzed (instruments, phases, and/or surgical skills), and they were further subdivided on detection, segmentation, and/or tracking. Quality of the algorithms was assessed by comparison of reported metrics (accuracy, sensitivity, specificity, precision, area under the receiver operating characteristic curve [ROC AUC], area under the precision-recall curve, and the F1 or Dice score), and quality of the validation procedure (dataset split, use of k-fold cross validation, and evaluation on external datasets). When possible, metrics in a category of surgery analysis (e.g. instrument detection) were

compared and displayed in the form of boxplots. Correlation between the most reported performance metric (e.g. ROC AUC) per category and dataset sizes were calculated with the Spearman's rank correlation coefficient (denoted r_s). Significance probabilities were then determined with a permutation test on this statistic.

Results

We identified a total of 38 algorithms for automated analysis of cataract extraction surgery (see Fig. 1). These include 20 publications about the detection (12), segmentation or tracking (8) of surgical tools, 17 about the recognition of the current surgical phase, and 9 about the assessment of surgical skills (6), and complication risk (3). Some of these publications overlap with each other, for example, algorithms that use the presence of instruments to predict the current phase of the surgery (overlap shown in Fig. 2). The algorithm characteristics are summarized in Tables 1–4.

For each category, we briefly discuss publications with interesting aspects or insights.

Public Datasets

For training the DL architectures, surgical videos were used that were either recorded in a local, cooperating hospital or part of a public dataset. The most comprehensive open dataset is the CATARACTS Grand Challenge Dataset. It consists of 50 videos of cataract surgeries performed at the Brest University Hospital in 2015.[11] Patients had a mean age of 61 years and the surgeries lasted for an average of 10 minutes and 56 seconds, were performed by 3 specialists and recorded in a resolution of 1920 × 1080 through an OPMI Lumera T microscope. The presence of the following 21 instruments is labeled in the dataset: biomarker, Charleux cannula, hydrodissection cannula, Rycroft cannula, viscoelastic cannula, cotton, capsulorhexis cystotome, Bonn forceps, capsulorhexis forceps, Troutman forceps, needle holder, irrigation/aspiration handpiece, phacoemulsifier handpiece, vitrectomy handpiece, implant injector, primary incision knife, secondary incision knife, micromanipulator, suture needle, Mendez ring, and Vannas scissors (Fig. 3).

The 50 videos are split into training and test datasets with 25 videos each.[11]

On top of this grand challenge data, the Cataract Dataset for Image Segmentation (CaDis) segmentation is available and consists of 4670 images sampled randomly from the 25 videos in CATARACTS' training set. Each pixel in each image is labeled with its respective instrument or anatomic class from a set of 36 identified classes (see Fig. 3).

Another dataset is the 101-Cataracts video collection which consists of 101 cataract surgeries performed by 4 different surgeons over a period of 9 months in the Klinikum Klagenfurt in Austria, with annotations for different surgery phases.[12] The four operating surgeons were grouped into two different levels of experience based on position and hours operated. The total length of all videos amounts to 14 hours, 2 minutes, and 5 seconds (1,263,116 frames) with a resolution of 720 × 540 pixels.

All described datasets in the study selection consists of recordings of the phacoemulsification procedure but other approaches like the MSICS are not considered.

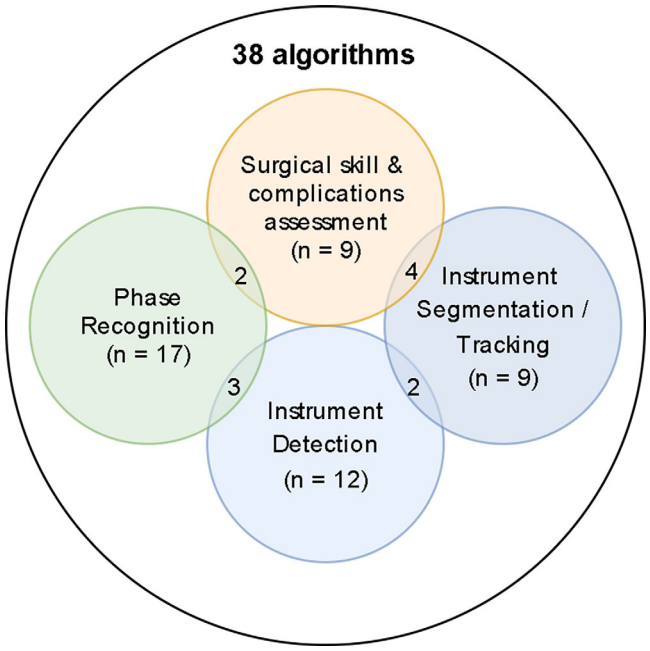

Figure 2. Overview of the different cataract surgery analysis algorithms.

Instrument Analysis

This subset of algorithms aims to either detect the presence of surgical instruments (detection) or quantitatively track their position (tracking) during the procedure.

Instrument Detection

Of the 19 instrument-related publications, 12 specifically address detection of surgical instruments. Here, the main goal of the algorithm is to recognize for every frame or time step in a recording the presence or absence of a prespecified list of surgical instruments

Table 1. Reviewed Studies That Detect the Presence of an Instrument From Cataract Surgical Videos

Author	Algorithm Type	ROC AUC	Sen/Spec	W x H	Dataset [Train/Val/Test], # Classes	Public Data?	HP Details?	Statistical Analysis?
Matton, 2022[13]	ResNet, DenseNet, RNN, LSTM	0.998	0.9372/-	480 × 270	[114/38/38], 15	✖	✔	Yes**
Ruzicki, 2022[21]	ResNet, LSTM, Random Forest	0.976	—	1920 × 1080	[209/167 /42], 20	✖	✖	95%-CI
Al Hajj, 2018[19]	VGG, ResNet, Inception V4, NASNet, LSTM	0.996	—	331 × 184	[25/25, 40/40], 40	✔	✔	Yes**
Yu, 2019[14]	LTSM, SqueezeNet, SVM	0.773	0.797/0.977		[60/20/20], 25	✖	✖	Yes**
Al Hajj, 2017[20]	Custom fusion CNN	0.977	0.950/0.820	576 × 720	[24/6/-], 15	✖	✖	Yes**
Whitten, 2022[15]	ResNet-52, Inception V3	0.976	0.959/0.997	224 × 224	[22/7/-], 22	✖	✔	✖
Al Hajj, 2019[11]	NASNet, ResNet, Inception V3	0.9971	—	540 × 960	[25/25/-], 41	✔*	✔	95% CI
Lalys, 2013[40]	BoW, SVM	—	—	720 × 756	[16/4/-], 29	✖	✖	✖
Banerjee, 2019[17]	AlexNet, VGG, ResNet	0.82	—	224 × 224	[25/-/25], 41	✔	✔	✖
Zisimopoulos, 2018[18]	ResNet 152, LSTM	0.959	—	224 × 224	[20/3/2], 5	✔	✔	✖
Al Hajj, 2016[16]	Nearest Neighbor	0.947	—	1920 × 1080	Unknown	✖	✔	95% CI
Sokolova, 2023[22]	ResNet-50, Inception V3, NASNet Mobile	—	0.086/0.942	Depending on architecture	[17/4/4; 71/15/15], 22	✔	✔	✖

BoVW, Bag of Visual Words; 95% CI, 95% confidence interval; CNN, Convolutional Neural Network; HP, hyperparameter; LSTM, Long Short-Term Memory network; N, size of the dataset; ROC AUC, area under receiver operator characteristic curve; Sen/Spec, sensitivity/specificity; W x H, weight times height of input video frame data.

*Public dataset and public code.

**Statistically significant, $p < 0.05$.

Table 2. Reviewed Studies That Track the Surgical Instruments in the Recorded Video Material (the Last Two Rows Represent Segmentation Approaches)

Author	Algorithm Type	mAP	IoU	W x H	Dataset [Train/Validation/Test], # Classes	Public Data?	HP Details?	Statistical Analysis?
Kim, 2019[26]	Temporal CNN, optical flow	—	—	640 × 480	[79/20/-], 2	✖	✔	95% CI
Yeh, 2023[27]	YOLOCAT	0.782	0.820	456 × 256	F: [3550/108/105], 15	✖	✖	✖
Morita, 2020[29]	Inception V3, scSE-FC-DenseNet	—	0.915	256 × 168	[211/30/61], 6	✖	✔	✖
Lin, 2022[28]	YOLO v4-tiny, Kallman filter	—	—	604 × 604	F: [960/320/320], 1	✖	✔	✖
Tian, 2015[30]	Hough transform, SVM, Tracking Learning Detection	0.876	—	—	[1/4/-], 29	✖	✔	✖
Zang, 2019[25]	EF-PNet	0.929	—	448 × 448	[45/7/-], 9	✖	✔	✖
Ni, 2019[46]	Resnet34 encoder, decoder with Attention	—	0.956	1920 × 1080	F: [3582/542/614], 29	✖	✔	✖
Fox, 2020[23]	Mask R-CNN, ResNet 101 backbone	0.685	—	224 × 224	[19/6/-], 16	✔	✔	✖
Pissas, 2021[24]	ResNet, UPerNet, OCRNet, DeepLab v3	—	0.864	1920 × 1080	[79/20/-], 2	✔*	✔	✖

BoVW, Bag of Visual Words; 95% CI, 95% confidence interval; CNN, Convolutional Neural Network; F, only number of frames is reported; HP, hyperparameter; IoU, intersection over union; LSTM, Long Short-Term Memory network; mAP, mean average precision; N, size of the dataset; W x H, weight times height of input data.

*Public dataset and public code.

Table 3. Reviewed Studies That Predict the Current Surgical Phase or Step From Cataract Surgical Videos

Author	Algorithm Type	ROC AUC	Accuracy	Sen./Spec.	W x H	Dataset [Train/Validation/Test], #Classes	Public Data?	HP Details?	Statistical Analysis?
Nespolo, 2022[33]	Faster R-CNN	0.961	—	—	640 × 360	[6/2/2], 15	✖	✖	✖
Touma, 2022[41]	AutoML	—	0.960	0.61/0.962	720 × 540	F: [896/384/144], 20	✔	✔	✖
Yeh, 2021[34]	VGG, LSTM	0.990	0.978	0.84/0.997	256 × 456	[211/26/31], 25	✖	✖	✖
Yu, 2019[14]	LTSM, SqueezeNet, SVM	0.773	0.959	0.797/0.977	—	[60/20/20], 25	✖	✖	Yes**
Quellec, 2014[37]	Conditional random fields	0.832	0.793	—	—	[93/93/-], 2	✖	✔	Yes**
Quellec, 2014[36]	Nearest Neighbor	0.794	0.870	—	720 × 756	[93/ 93/-], 1	✖	✔	Yes**
Nespolo, 2022[33]	R-CNN built on ResNet-50	0.961	—	—	640 × 320	[6/2/2], 14	✖	✔	95% CI
Morita, 2019[44]	Inception V3, moving average	—	0.965	—	256 × 168	[245/10/48], 5	✖	✔	✖
Lecuyer, 2020[39]	VGG19, Inception V3, ResNet50	—	0.957	—	299 × 299	[30/20/-], 41	✔	✔	✖
Quellec, 2015[38]	Spatio- temporal Polynomials	0.856	—	—	720 × 576	[93/93/-], 9	✖	✖	Yes**
Lalys, 2013[40]	BovW, SVM	—	0.645	0.549/0.763	720 × 756	[16/4/-], 29	✖	✖	✖
Zisimopoulos, 2018[18]	ResNet 152, LSTM	—	0.782	—	224 × 224	[20/3/2], 5	✔	✔	✖
Charrière, 2014[31]	Nearest Neighbor, HOG, BoW	0.815	—	—	720 × 756	[15/15/-], 8	✔	✔	✖
Charriere, 2016[35]	Histograms, BoVW, Bayesian network	0.986	—	—	720 × 756	[25/5/-], 5	✖	✔	✖
Primus, 2018[47]	GoogLeNet	—	—	0.72	224 × 224	[17/4/-], 22	✔	✔	✖
Ghamsarian, 2020[43]	ResNet50/101, Mask R-CNN	—	0.910	0.91	—	[18/-/4], 16	✔	✔	✖
Ghamsarian, 2021[42]	Custom R-CNN, ResNet50, VGG-19	—	0.990	0.99	512 × 512	[85 /15/-], 2	✖	✔	✖

BoVW, Bag of Visual Words; 95% CI, 95% confidence interval; CNN, Convolutional Neural Network; F, only number of frames is reported; HP, hyperparameter; LSTM, Long Short-Term Memory network; N, size of the dataset; ROC AUC, area under receiver operator characteristic curve; Sen/Spec, sensitivity/specificity; W x H, weight times height of input data.

**Statistically significant, $P < 0.05$.

Table 4. Reviewed Studies that Estimate Risk or Recognize Possible Complications During Surgery

Author	Algorithm Type	ROC AUC	Other Metric	W x H	Dataset [Train/Validation/Test], # Classes	Public Data?	HP Details?	Statistical Analysis?
Hira, 2022[45]	SVM, LSTM, ResNet	0.720	—	640 × 480	[59/20/20], 4	✖	✔	95% CI
Ruzicki, 2022[21]	ResNet, LSTM, Random Forest	0.570	—	1920 × 1080	[209/167/42], 2	✖	✖	95% CI
Kim, 2019[26]	Temporal CNN, optical flow	0.863	—	640 × 480	[79/20/-], 4	✖	✔	95% CI
Nespolo, 2022[33]	R-CNN built on ResNet-50	0.961	—	640 × 320	[6/2/2], 14	✖	✔	95% CI
Yeh, 2023[27]	YOLOCAT	—	R: 0.753	456 × 256	F: [3550/108/105], 3	✖	✖	✖
Tabuchi, 2022[48]	Custom model built on Inception V3	—	Δ Risk score: 0.124	299 × 168	[9/9/-], 2	✖	✔	Yes**
Morita, 2020[29]	Inception V3, scSE-FC-DenseNet	0.970	—	256 × 168	[211/30/61], 2	✖	✔	✖
Gu, 2020[49]	ResNet, ResUnet	—	std. pred./truth: 4.93px	224 × 224	[69/-/29], 3	✖	✔	✖
Ghamsarian, 2021[42]	Custom R-CNN, backbone: ResNet50/VGG-19	—	F1: 0.990	512 × 512	[85/15; 34/9], 2	✖	✔	✖

BoVW, Bag of Visual Words; 95% CI, 95% confidence interval; CNN, Convolutional Neural Network; F, only number of frames is reported; HP, hyperparameter, LSTM, Long Short-Term Memory network; N, size of the dataset; R, correlation; ROC AUC, area under receiver operator characteristic curve; W x H, weight times height of input data.

**Statistically significant, $P < 0.05$.

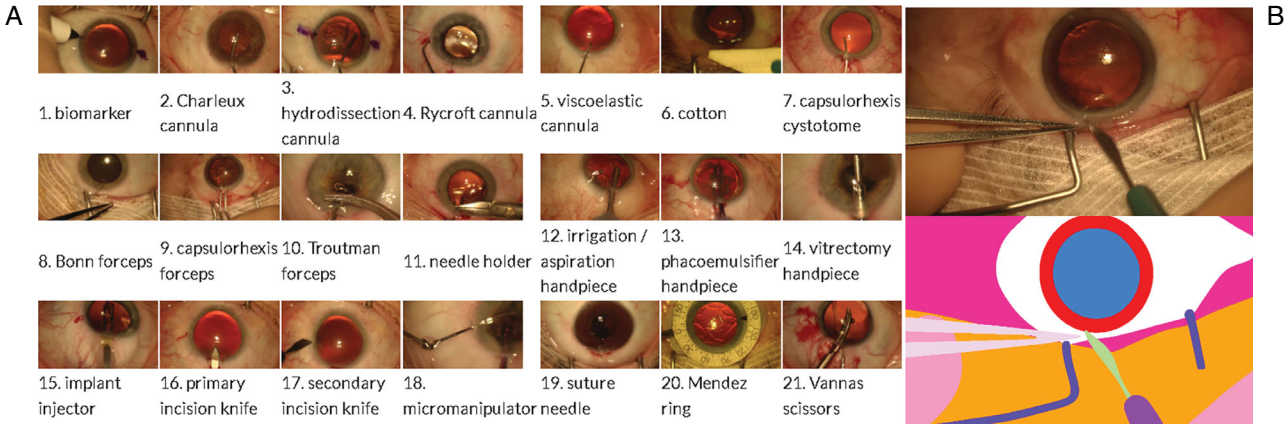

Figure 3. Dataset overview with (**A**) detection labels from CATARACTS grand challenge and (**B**) segmentation of the CaDis subset.[11]

(see Fig. 3). Various performance metrics were reported to assess the algorithms with the most common being ROC AUC. Overall, a high performance can be seen ranging from 0.976 to 0.998.[13–21] Sensitivity and specificity are only reported in 5 studies and range from 0.797 to 0.959 and from 0.820 to 0.997, respectively.[13–15,20,22] The dataset sizes of these detection studies range between 25 and 209 videos with a median of 50 videos.[18,21] There was no statistically significant Spearman's correlation between the size of the dataset and the performance metric ROC AUC with a coefficient r_s of 0.16 and a P value of 0.75.

The majority of proposed algorithms ($n = 8$) used a setup that consists of a convolutional neural network (CNN) to extract image features from the ingested video frames and a recurrent neural network (RNN) to interpret these features across time frames to detect the presence of instruments. For the first part, various architectures were explored (such as ResNet-52, DenseNet, Inception v4, VGG16, and NASNet), and for the second step, variations of the popular long short-term memory (LSTM) network were used, achieving overall similar results. The remaining algorithms (2) designed custom networks that fused the features inside the CNN, and used non-neural network-based machine learning approaches for the complete ($n = 1$) or part of the prediction process ($n = 1$).[20,21] One of the research papers detected the presence and removal of instruments from the instrument table which could be useful to verify the presence during the surgery video.[16]

Specifically, we want to highlight the paper by Natalia Sokolova et al., where networks were trained on one dataset and then evaluated on another.[22] This led to a massive degradation in sensitivity from 0.834 down to 0.285. The authors theorize that this could be due to the fact that the CATARACTS Grand Challenge dataset has an unusually high quality which does not generalize well to other circumstances.

Instrument Tracking

Among the eight papers in this section, two use a full semantic segmentation approach for the tracking of surgical instruments, wherein each pixel in a given image frame is classified to a certain class. They use the popular U-Net architecture and a CNN-based feature extractor in combination with a mask recurrent neural network. Fox et al.[23] returns a segmentation class for each image pixel, whereas Pissas et al.[24] returns masks for the different classes.

The remaining instrument tracking papers propose various technical approaches, the most common being the "You only look once" (YOLO) network architecture that is able to localize trained classes in various sizes after only ingesting an inference image once. It outputs bounding boxes for the recognized classes with position and size information.[25–30]

For the evaluation of these approaches, intersection over union (IoU) and mean average precision (mAP), which is calculated from a cutoff value of the former, were used. Overall, there is a moderate success for this application with an mAP that ranges from 0.685 to 0.929[23,25,27,30] and an IoU from 0.820 to 0.956.[24,27,29,46] One author reports an impressive mean absolute error (MASRE) between prediction and ground truth of 21.26 µm while using a combination of an attitude and heading reference system (AHRS) and a CNN architecture.[28] The approaches vary widely in dataset size with two examples only using 5 videos, one 300 videos, and, overall, a median video count of 38.5.[28–30] Also in this scenario, there was no statistically

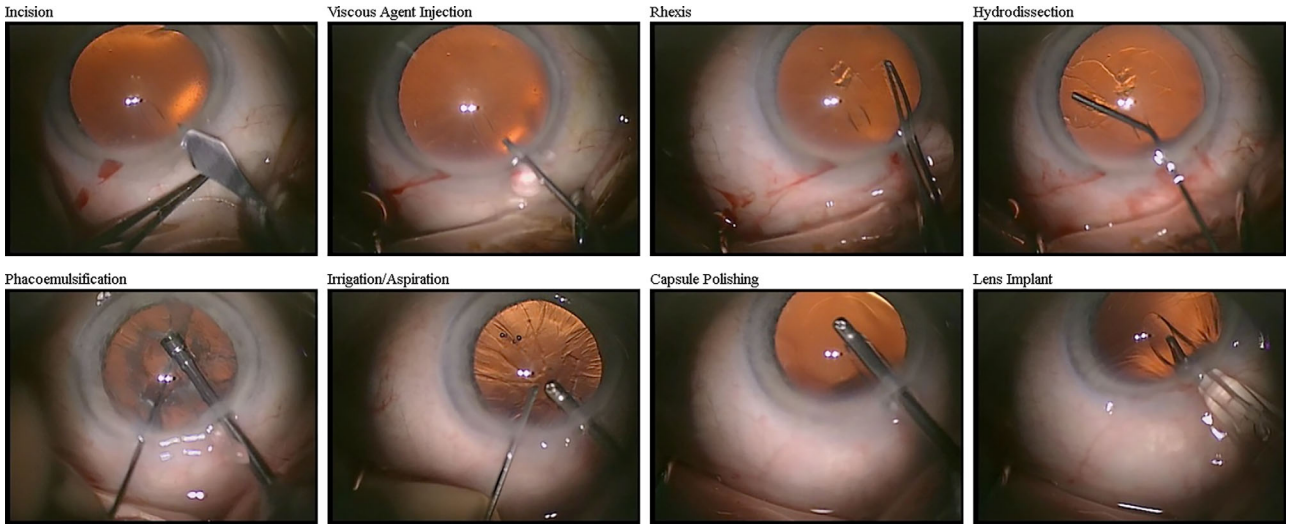

Figure 4. The 101-Cataracts dataset overview showing the different phases in cataract surgery.[12]

significant correlation between the outcome metric mAP and the dataset size ($r_s = 0.00$, $P = 1.00$).

Surgical Phase Recognition

Seventeen publications address the problem of matching video segments with their respective surgical phases in cataract surgery. Examples of surgical phases in a publicly available dataset can be seen in Figure 4. From these, nine papers report an ROC AUC metric between 0.773 and 0.990[14,31–38] and 11 report an accuracy between 0.645 and 0.978.[14,18,27,36,37,39–43,47] The results with lower accuracy or ROC AUC are older studies that construct hand-crafted features and feed them into machine learning algorithms like Support Vector Machines or Nearest-Neighbor matching.

The more recent publications utilize state-of-the-art neural networks and construct various CNN feature extractors in combination with a recurrent neural network for integration across time. Here, the most common type is the previously mentioned LSTM to predict a fitting surgical phase for every video frame. The minimal amount of videos utilized by these approaches is 10, the maximum amount is 303, the median corresponds to 50 videos, and there was no significant correlation between ROC AUC and the number of training videos ($r_s = -0.07$, $P = 0.84$).[33,44]

Interestingly, Yu et al. fed only manually curated instrument labels corresponding to the current timestep as input data into an RNN and achieved high results with an accuracy, sensitivity, and specificity of 0.959, 0.797, and 0.977, respectively.[14]

Surgical Skill Assessment and Complication Recognition

The remaining nine studies go beyond the typical computer vision application of instrument detection and surgical phase segmentation and try to extract clinically more relevant information about surgical performance and complication risk. Overall, they achieve an ROC AUC between 0.570 and 0.970.[21,26,29,45] Dataset sizes vary again greatly between 10 and 302 videos with an median of 99 videos.[29,33] The size and ROC AUC metric have no statistically significant correlation ($r_s = -0.36$, $P = 0.75$).

The simplest form of surgical skill determination (4 of the available papers) trained a binary classifier on differentiating between a novice and expert surgeon using the operating recordings as input, and achieved an accuracy between 0.578 and 0.848. To highlight this difference, Hitoshi Tabuchi et al. predicted a continuous risk score that significantly ($P < 0.01$) differed between experienced and novice surgeons.[48]

Gu et al. combines tracking information with a surgical guidance system to calculate trajectories with only a few pixels of error.[49] Another approach estimated items specific for phacoemulsification from the ICO-Ophthalmology Surgical Competency Assessment Rubric (ICO-OSCAR:phaco) utilizing properties calculated from the instrument tracking procedure. For example, Morita et al. uses the tooltip data to predict the handling of the rhexis formation during the surgery.[29] Similarly, Tae Soo Kim et al. estimates ICO-OSCAR:phaco items by predicting instrument trajectory velocities with a temporal convolutional neural

network (TCN) achieving an accuracy of 0.728 and an ROC AUC of 0.773.[26] The scores on the rubric can be translated into a simpler novice/expert differentiation as in the previous examples.

Specific complications can be recognized by focusing on certain characteristics of the surgery.[29,42] In the LensID paper, the implanted lens is recognized by a segmentation network, and location plus deformation parameters is calculated to estimate lens instability which is hypothesized to lead to dislocation after the surgery.[42] Shoji Morita et al. trained a CNN on different adverse outcomes by accumulating the individual frame risk in a moving average.[28] They were able to predict problems during the surgery even before a human surgeon with an accuracy of 0.902 and an ROC AUC of 0.970.

Discussion

Across all studies, we observed significant improvements in various performance metrics in DL methods compared to classical machine learning techniques like Random Forest, Support Vector Machines, or Nearest Neighbor Classifiers.[30,36,40] Especially for the tasks of instrument detection, tracking, and surgical phase recognition, DL approaches achieve results of high quality consistently across multiple datasets.

Availability of Data

Exploration of the available datasets underscored a limitation in the availability of public datasets for cataract extraction surgery. Overall, the research community needs access to more public datasets. Moreover, the two existing sources only include recordings of the phacoemulsification procedure usually performed in high-income countries. MSICS is still the surgery of choice in high volume cataract settings in low- and middle-income countries (LMICs).[50] Because of the lacking data, not even a single of the 38 studies is focused on MSICS. This lack of datasets and studies poses a challenge for algorithms development tailored to MSICS scenarios where performance evaluation and recognition is especially important due to higher chances of complications.[51] Current phacoemulsification datasets and trained models cannot be used without translation on MSICS data because the surgery differs in phases, utilized instruments, and risk for certain complications. Further, it is important to keep in mind that LMICs often do not have access to high-end hardware. This could lead to training data of lower quality and higher importance of selecting efficient neural network architectures. How well current data and architectures can be applied in LMICs should be investigated with an MSCIS dataset. Therefore, releasing a public MSICS dataset of sufficient size in the future would make it possible to address these problems and publishing a programming competition like the CATARACTs challenge, which could contribute to finding a generalizable and well-performing DL architecture.

Many studies use in-house datasets from collaborating hospitals that are not public, but, if made public, would help in training better algorithms and validating new approaches. Additionally, only 2 out of 37 studies (5.4%) have published their codebase or trained models which makes replication difficult. However, 30 studies (81.1%) at least support reproduction by reporting their hyperparameters.

Common Metrics and Validation

Comparison of differences in the performance of cataract surgery analysis algorithms is currently limited by the fact that it is not reported with standardized metrics. An international expert consortium recently highlighted the need to choose the right metric for a specific problem in the biomedical realm to not over- or underestimate an algorithm.[52] In our review, authors often report only one or two metrics, even though models can be better assessed and compared if at least the most common machine learning metrics are present. Following the guidelines from the consortium, we would recommend including the ROC AUC, precision (positive predictive value), recall (sensitivity), and accuracy metrics for classification tasks (instrument detection, phase recognition, surgical skill, and complication determination). Whenever possible, additional metrics like sensitivity and specificity could be reported to make comparisons easier and improve trust in the results. For tracking and segmentation algorithms, we would recommend the usage of mAP, IoU, and Dice Similarity Coefficient (DSC) metrics.

Our confidence in the generalizability of the majority of described studies is reduced by the fact that only 9 out of 38 studies (23.7%) evaluated the algorithm on an additional, external dataset. In regard to validation methods, 10 authors (26.3%) use cross-validation or leave-one-out validation and the rest use a simple training, validation, and test split. Statistical significance is only investigated in 14 studies (36.8%) and only 8 studies (21.1%) achieved a P value < 0.05. Figure 5 highlights how this variability in validation methods and datasets leads to a broad spread of performance metric results.

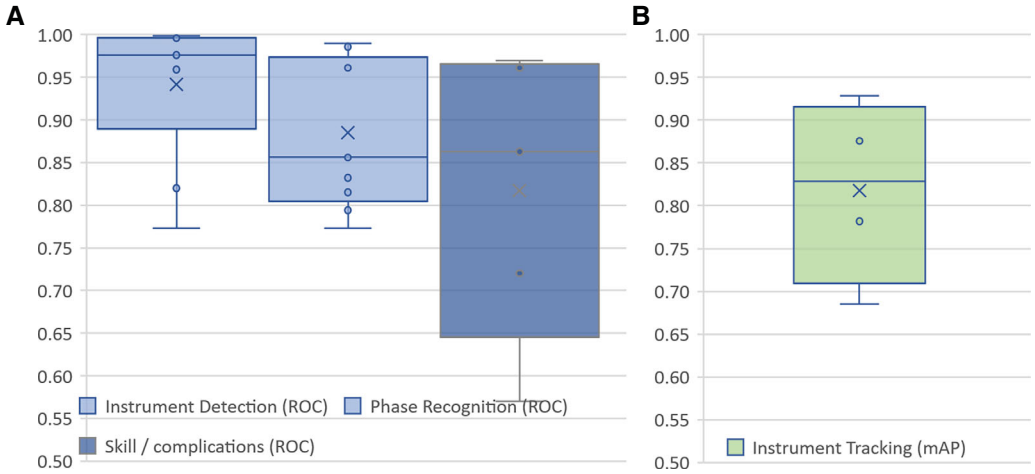

Figure 5. Spread of the results. Boxplots of the area (**A**) under the receiver operator characteristic curve (ROC AUC) for surgical phase, instrument detection, and surgical skill or risk assessment, and (**B**) instrument tracking results reported in mean average precision (mAP).

Furthermore, there is a huge variability in the number of recognized instruments and phases in the studies that do not use one of the public datasets. Similarly, for recognized complications during procedures or surrogates of surgical skill, there is no consensus. Establishing a common list of instruments, phases, and evaluation metrics is pivotal for advancing the field and promoting the reproducibility of results. In addition, patient characteristics of datasets should be included to identify potential confounders that could influence predictions.

Although it is to be expected that learning-based methods trained on larger and more representative data will deliver better results, our correlation analysis between performance of models and dataset sizes did not yield any statistically significant relationship. In particular, several studies based on small datasets report high accuracy.

Therefore, we are not able to recommend a minimal number of videos that are required for training DL approaches. In general, that number should depend on several factors, including the exact task, homogeneity, resolution, and quality of the videos, as well as algorithmic approach. In particular, the increasing availability of foundation models should help to reduce the required amount of training data.[53] However, it is known that studies that evaluate methods on small test sets often overestimate their performance, so we emphasize the relevance of evaluating methods on sufficiently large test datasets.[54]

Another significant challenge highlighted by Sokolova et al. is the difficulty in transferring trained algorithms between datasets and acquisition setups. They trained an instrument detection network on one dataset and then tested it on a second one leading to massive performance losses.[22] The variance in video recording setup, surgical instruments, phases, and complications in different data sources probably contributes to this problem. Developing algorithms that are able to generalize well is imperative for clinical applications. In addition, to even detect this problem, proposed models have to be tested on external data sources.

Conclusions and Outlook

One notable DL design that is missing from the observed studies are vision transformers that have been introduced in 2020 and improved performance in computer vision applications.[55] Kiyasseh et al. demonstrated that vision transformers make it possible to recognize surgical phases and gestures with high accuracy and good generalizability across different datasets.[56,57] In the future, it would be interesting to explore this approach in cataract extraction to overcome the issue of generalizability.

We believe that the potential of automated video analysis in surgical teaching and training is promising. Algorithms could aid surgical education by providing near real-time feedback to improve surgical techniques. Collaborative efforts to curate comprehensive datasets and develop standardized models have the potential to speed up advancements in this field.

In conclusion, our study highlights the ability of DL models to estimate surgical phase, track or detect instruments, and recognize complications. This promising outlook is somewhat lessened by the general lack of publicly available data, public code, and pretrained models (especially for a comprehensive MSICS dataset), and the need for standardized evaluation protocols. Addressing these challenges and leveraging emerging technologies like vision transform-

ers will undoubtedly shape the future landscape of cataract surgery video analysis and hold the potential to significantly enhance surgical training and surgical outcomes.

Acknowledgments

Disclosure: **S. Müller**, None; **M. Jain**, None; **B. Sachdeva**, None; **P.N. Shah**, None; **F.G. Holz**, Acucela (F), Allergan (F), Apellis (F), Bayer (F), Bioeq/Formycon (F), CenterVue (F), Ellex (F), Roche/Genentech (F), Geuder (F), Kanghong (F), NightStarx (F), Novartis (F), Optos (F), Zeiss (F); Acucela (C), Aerie (C), Allergan (C), Apellis (C), Bayer (C), Boehringer-Ingelheim (C), Ivericbio (C), Roche/Genentech (C), Geuder (C), Grayburg Vision (C), Ivericbio (C), LinBioscience (C), Kanghong (C), Novartis (C), Pixium Vision (C), Oxurion (C), Stealth BioTherapeutics (C), Zeiss (C); **R.P. Finger**, Novartis (F), CentreVue (F), Heidelberg Engineering (F), Zeiss (F); Novartis (C), Bayer (C), Roche/Genentech (C), Ellex (C), Alimera (C), Allergan (C), Santhera (C), Inositec (C), Opthea (C); **K. Murali**, None; **M.W.M. Wintergerst**, ASKIN & CO GmbH (R), Bayer AG (R), Berlin-Chemie AG (R, F), CenterVue SpA (non-financial support), Carl Zeiss Meditec (non-financial support), D-Eye Srl (non-financial support), DigiSight Technologies (F, non-financial support), Eyenuk, Inc. (non-financial support), Eyepress Fachmedien GmbH (R), Glaucare GmbH (owner, C), Heine Optotechnik GmbH (C, R, F, non-financial support), Heidelberg Engineering (R, non-financial support), Novartis Pharma GmbH (R, non-financial support), Optos (non-financial support), Pro Generika e.V. (R), Science Consulting in Diabetes GmbH (R); **T. Schultz**, None

* MWMW and TS contributed equally to this work.

References

1. Bhandarkar P, Gadgil A, Patil P, Mohan M, Roy N. Estimation of the national surgical needs in India by enumerating the surgical procedures in an urban community under universal health coverage. *World J Surg*. 2021;45(1):33–40.
2. Eurostat Statistics Explained. Surgical operations and procedures statistics. 2023. Available at: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Surgical_operations_and_procedures_statistics.
3. Rossi T, Romano MR, Iannetta D, et al. Cataract surgery practice patterns worldwide: a survey. *BMJ Open Ophthalmol*. 2021;6(1):e000464.
4. Rampat R, Deshmukh R, Chen X, et al. Artificial intelligence in cornea, refractive surgery, and cataract: basic principles, clinical applications, and future directions. *Asia-Pacific J Ophthalmol*. 2021;10(3):268–281.
5. Benmansour M, Malti A, Jannin P. Deep neural network architecture for automated soft surgical skills evaluation using objective structured assessment of technical skills criteria. *Int J Comput Assist Radiol Surg*. 2023;18(5):929–937.
6. Wintergerst MWM, Gorgi Zadeh S, Wiens V, et al. Replication and refinement of an algorithm for automated drusen segmentation on optical coherence tomography. *Sci Rep*. 2020;10(1):7395.
7. Bali J, Bali O, Sahu A, Boramani J, Deori N. Health economics and manual small-incision cataract surgery: an illustrative mini review. *Indian J Ophthalmol*. 2022;70(11):3765–3770.
8. Higgins JT, Chandler J. *Cochrane Handbook for Systematic Reviews of Interventions version 6.4*. 2nd ed. Chichester (UK): John Wiley & Sons; 2019.
9. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71.
10. Nyangoh Timoh K, Huaulme A, Cleary K, et al. A systematic review of annotation for surgical process model analysis in minimally invasive surgery based on video. *Surg Endosc*. 2023;37(6):4298–4314.
11. Al Hajj H, Lamard M, Conze PH, et al. CATARACTS: challenge on automatic tool annotation for cataRACT surgery. *Med Image Anal*. 2019;52:24–41.
12. Schoeffmann K. Cataract-101: video dataset of 101 cataract surgeries. *MMSys '18: Proceedings of the 9th ACM Multimedia Systems Conference*. 2018:421–425.
13. Matton N, Qalieh A, Zhang Y, et al. Analysis of cataract surgery instrument identification performance of convolutional and recurrent neural network ensembles leveraging BigCat. *Transl Vis Sci Technol*. 2022;11(4):1.
14. Yu F, Silva Croso G, Kim TS, et al. Assessment of automated identification of phases in videos of cataract surgery using machine learning and deep learning techniques. *JAMA Netw Open*. 2019;2(4):e191860.
15. Whitten J, McKelvie J, Mayo M. Clinically-relevant summarisation of cataract surgery videos using deep learning. In: Szczerbicki E, Wojtkiewicz

K, Nguyen S.V., Pietranik M., Krótkiewicz M. (eds). *Communications in Computer and Information Science, vol 1716*. Recent Challenges in Intelligent Information and Database Systems, ACIIDS Singapore: Springer; 2022;1716:711–723.
16. Hajj HA. Coarse-to-fine surgical instrument detection for cataract surgery monitoring. *arXiv Preprint*. 2016. arXiv: 1609.05619 [cs.CV]. doi.org/10.48550/arXiv.1609.05619.
17. Banerjee N, Sathish R, Sheet D. Deep neural architecture for localization and tracking of surgical tools in cataract surgery. In: Peter J., Fernandes S., Eduardo Thomaz C., Viriri S. (eds). *Computer Aided Intervention and Diagnostics in Clinical and Medical Images. Lecture Notes in Computational Vision and Biomechanics*, vol 31. Springer, Cham, https://doi.org/10.1007/978-3-030-04061-1_4.
18. Zisimopoulos O, Flouty E, Luengo I, et al. DeepPhase: surgical phase recognition in CATARACTS videos. In: Frangi A., Schnabel J., Davatzikos C., Alberola-López C., Fichtinger G. (eds) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. MICCAI 2018. Lecture Notes in Computer Science, vol 11073. Springer, Cham, https://doi.org/10.1007/978-3-030-00937-3_31.
19. Al Hajj H, Lamard M, Conze PH, Cochener B, Quellec G. Monitoring tool usage in surgery videos using boosted convolutional and recurrent neural networks. *Med Image Anal*. 2018;47:203–218.
20. Al Hajj H, Lamard M, Charriere K, Cochener B, Quellec G. Surgical tool detection in cataract surgery videos through multi-image fusion inside a convolutional neural network. *Annu Int Conf IEEE Eng Med Biol Soc*. 2017;2017:2002–2005.
21. Ruzicki J, Holden M, Cheon S, Ungi T, Egan R, Law C. Use of machine learning to assess cataract surgery skill level with tool detection. *Ophthalmol Sci*. 2023;3(1):100235.
22. Sokolova N, Schoeffmann K., Taschwer M., Putzgruber-Adamitsch D., El-Shabrawi Y. Evaluating the generalization performance of instrument classification in cataract surgery videos. *International Conference on Multimedia Modeling*. 2019. In: Ro Y., et al. MultiMedia Modeling. MMM 2020. Lecture Notes in Computer Science, vol 11962. Springer, Cham, https://doi.org/10.1007/978-3-030-37734-2_51.
23. Fox M, Taschwer M, Schoeffmann K. Pixel-based tool segmentation in cataract surgery videos with mask R-CNN. *2020 Ieee 33rd International Symposium on Computer-Based Medical Systems(Cbms 2020)*. 2020:565–568.
24. Pissas T, Ravasio CS, Da Cruz L, Bergeles C. Effective semantic segmentation in Cataract Surgery: what matters most? In: de Bruijne M., et al. 2021. Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12904. Springer, Cham, https://doi.org/10.1007/978-3-030-87202-1_49.
25. Zang DQ, Bian GB, Wang YL, Li Z. An extremely fast and precise convolutional neural network for recognition and localization of cataract surgical tools. *Lect Notes Comput Sci*. 2019;11768:56–64.
26. Kim TS, O'Brien M, Zafar S, Hager GD, Sikder S, Vedula SS. Objective assessment of intraoperative technical skill in capsulorhexis using videos of cataract surgery. *Int J Comput Assist Radiol Surg*. 2019;14(6):1097–1105.
27. Yeh HH, Jain AM, Fox O, Sebov K, Wang SY. PhacoTrainer: deep learning for cataract surgical videos to track surgical tools. *Transl Vis Sci Technol*. 2023;12(3):23.
28. Lin C, Zheng Y, Guang C, Ma K, Yang Y. Precision forceps tracking and localisation using a Kalman filter for continuous curvilinear capsulorhexis. *Int J Med Robot*. 2022;18(6):e2432.
29. Morita S, Tabuchi H, Masumoto H, Tanabe H, Kamiura N. Real-time surgical problem detection and instrument tracking in cataract surgery. *J Clin Med*. 2020;9(12):3896.
30. 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.
31. Charriere K, Quellec G, Lamard M, Coatrieux G, Cochener B, Cazuguel G. Automated surgical step recognition in normalized cataract surgery videos. *Annu Int Conf IEEE Eng Med Biol Soc*. 2014;2014:4647–4650.
32. Garcia Nespolo R, Yi D, Cole E, Valikodath N, Luciano C, Leiderman YI. Evaluation of artificial intelligence-based intraoperative guidance tools for phacoemulsification cataract surgery. *JAMA Ophthalmol*. 2022;140(2):170–177.
33. Nespolo RG, Yi D, Cole E, Valikodath N, Luciano C, Leiderman YI. Evaluation of artificial intelligence-based intraoperative guidance tools for phacoemulsification cataract surgery. *JAMA Ophthalmol*. 2022;140(2):170–177.
34. Yeh HH, Jain AM, Fox O, Wang SY. PhacoTrainer: a multicenter study of deep learning for activity recognition in cataract surgical videos. *Transl Vis Sci Technol*. 2021;10(13):23.

35. Charriere K, Quellec G, Lamard M, et al. Real-time multilevel sequencing of cataract surgery videos. *2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI)*. 2016. Available at: https://www.semanticscholar.org/paper/Real-time-multilevel-sequencing-of-cataract-surgery-Charri%C3%A8re-Quellec/3c2cfbeef332747b92ab160c0ca0acd4df5d5996 time multilevel sequencing of cataract surgery videos.
36. Quellec G, Charriere K, Lamard M, et al. Real-time recognition of surgical tasks in eye surgery videos. *Med Image Anal*. 2014;18(3):579–590.
37. Quellec G, Lamard M, Cochener B, Cazuguel G. Real-time segmentation and recognition of surgical tasks in cataract surgery videos. *IEEE Trans Med Imaging*. 2014;33(12):2352–2360.
38. Quellec G, Lamard M, Cochener B, Cazuguel G. Real-time task recognition in cataract surgery videos using adaptive spatiotemporal polynomials. *IEEE Trans Med Imaging*. 2015;34(4):877–887.
39. Lecuyer G, Ragot M, Martin N, Launay L, Jannin P. Assisted phase and step annotation for surgical videos. *Int J Comput Assist Radiol Surg*. 2020;15(4):673–680.
40. Lalys F, Bouget D, Riffaud L, Jannin P. Automatic knowledge-based recognition of low-level tasks in ophthalmological procedures. *Int J Comput Assist Radiol Surg*. 2013;8(1):39–49.
41. Touma S, Antaki F, Duval R. Development of a code-free machine learning model for the classification of cataract surgery phases. *Sci Rep*. 2022;12(1):2398.
42. Ghamsarian N, Taschwer M, Putzgruber-Adamitsch D, Sarny S, El-Shabrawi Y, Schoeffmann K. LensID: a CNN-RNN-based framework towards lens irregularity detection in cataract surgery videos. Medical Image Computing and Computer Assisted Intervention - MICCAI 2021, *PT VIII*. 2021;12908:76–86.
43. Ghamsarian N. Relevance detection in cataract surgery videos by spatio- temporal action localization. *25th International Conference on Pattern Recognition*. 2020.
44. Morita S, Tabuchi H, Masumoto H, Yamauchi T, Kamiura N. Real-time extraction of important surgical phases in cataract surgery videos. *Sci Rep*. 2019;9(1):16590.
45. Hira S, Singh D, Kim TS, et al. Video-based assessment of intraoperative surgical skill. *Int J Comput Assist Radiol Surg*. 2022;17(10):1801–1811.
46. Ni ZL, Bian GB, Zhou XH, et al. RAUNet: residual attention U-Net for semantic segmentation of cataract surgical instruments. *Neural Information Processing (ICONIP 2019), PT II*. 2019;11954:139–149.
47. Primus MJ, Putzgruber-Adamitsch D, Taschwer M, et al. Frame-based classification of operation phases in cataract surgery videos. *Multimedia Modeling, MMM 2018, Pt I*. 2018;10704:241–253.
48. Tabuchi H, Morita S, Miki M, Deguchi H, Kamiura N. Real-time artificial intelligence evaluation of cataract surgery: a preliminary study on demonstration experiment. *Taiwan J Ophthalmol*. 2022;12(2):147–154.
49. Gu Y. Construction of Quantitative Indexes for Cataract Surgery Evaluation Based on Deep Learning. *International Workshop on Ophthalmic Medical Image Analysis*. 2020.
50. Gurnani B. *Manual Small Incision Cataract Surgery*. Internet: Treasure Island (FL); 2023.
51. Bernhisel A, Pettey J. Manual small incision cataract surgery. *Curr Opin Ophthalmol*. 2020;31(1):74–79.
52. Maier-Hein L, Reinke A, Godau P, et al. Metrics reloaded: recommendations for image analysis validation. *Nature Methods*. 2024;21:195–212.
53. Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. *Nat Commun*. 2024;15(1):654.
54. Flint C, Cearns M, Opel N, et al. Systematic misestimation of machine learning performance in neuroimaging studies of depression. *Neuropsychopharmacology*. 2021;46(8):1510–1517.
55. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. *arXiv Preprint*. 2020, https://doi.org/10.48550/arXiv.2010.11929.
56. Kiyasseh D, Ma R, Haque TF, et al. A vision transformer for decoding surgeon activity from surgical videos. *Nat Biomed Eng*. 2023;7(6):780–796.
57. Zhang B, Goel B, Sarhan MH, et al. Surgical workflow recognition with temporal convolution and transformer for action segmentation. *Int J Comput Assist Radiol Surg*. 2023;18(4):785–794.