AMERICAN ACADEMY OF OPHTHALMOLOGY®

Use of Machine Learning to Assess Cataract Surgery Skill Level With Tool Detection

Jessica Ruzicki, MD, FRCSC,[1] *Matthew Holden, PhD,*[2] *Stephanie Cheon, MD,*[3] *Tamas Ungi, MD, PhD,*[4] *Rylan Egan, PhD,*[5] *Christine Law, MD, FRCSC*[1]

Purpose: To develop a method for objective analysis of the reproducible steps in routine cataract surgery.
Design: Prospective study; machine learning.
Participants: Deidentified faculty and trainee surgical videos.
Methods: Consecutive cataract surgeries performed by a faculty or trainee surgeon in an ophthalmology residency program over 6 months were collected and labeled according to degrees of difficulty. An existing image classification network, ResNet 152, was fine-tuned for tool detection in cataract surgery to allow for automatic identification of each unique surgical instrument. Individual microscope video frame windows were subsequently encoded as a vector. The relation between vector encodings and perceived skill using k-fold user-out cross-validation was examined. Algorithms were evaluated using area under the receiver operating characteristic curve (AUC) and the classification accuracy.
Main Outcome Measures: Accuracy of tool detection and skill assessment.
Results: In total, 391 consecutive cataract procedures with 209 routine cases were used. Our model achieved an AUC ranging from 0.933 to 0.998 for tool detection. For skill classification, AUC was 0.550 (95% confidence interval [CI], 0.547−0.553) with an accuracy of 54.3% (95% CI, 53.9%−54.7%) for a single snippet, AUC was 0.570 (0.565−0.575) with an accuracy of 57.8% (56.8%−58.7%) for a single surgery, and AUC was 0.692 (0.659−0.758) with an accuracy of 63.3% (56.8%−69.8%) for a single user given all their trials.
Conclusions: Our research shows that machine learning can accurately and independently identify distinct cataract surgery tools in videos, which is crucial for comparing the use of the tool in a step. However, it is more challenging for machine learning to accurately differentiate overall and specific step skill to assess the level of training or expertise.
Financial Disclosure(s): The author(s) have no proprietary or commercial interest in any materials discussed in this article.

Surgical competence is a fundamental component of ophthalmology training programs. Cataract surgery is one of the most fundamental procedures residents are taught and expected to competently execute. Nonetheless, cataract surgery is technically challenging, especially for trainees, so assessment optimization is essential to ensure future clinical safety. With the shift to competency by design training, expanding valid and reliable quantitative methods to teach and evaluate learners are required. Currently, trainees are learning the procedure by self-directed reading, didactic lectures, videos, simulation lab practice, and surgical stimulators, as well as through step-by-step instruction during surgeries.[1-4] Surgical simulators and simulation labs have gained significant interest within residency programs. However, these simulations often lack improvement-centered feedback from the program itself. A resident may practice steps in the surgery, but if this is done incorrectly without feedback and appropriate supervision, the resident may develop poor surgical techniques.[5]

Research using deep neural networks has garnered increased publicity in the field of ophthalmology. At present, most applications of deep learning algorithms in ophthalmology mainly exist in detection and diagnostic modalities, including digital photographs, OCT, and visual fields.[6] Several disease processes are being assessed through automated image analysis, especially diabetic retinopathy, age-related macular degeneration, glaucoma, and cataract grading.[6-9] Emerging artificial intelligence platforms are currently being applied to other diseases such as retinopathy of prematurity, corneal ectasia, choroidal neovascularization, macular edema, drusen, geographic atrophy, epiretinal membrane, vitreomacular traction, macular hole, and central serous retinopathy.[8-12]

However, there have been few published studies demonstrating the efficacy of computer-based machine learning as an ophthalmology surgical training tool. Recently, there have been 2 studies from the Wilmer Eye Institute, Johns Hopkins University, Baltimore, Maryland in 2019 that have looked at this concept.[13,14] Yu et al describe a cross-sectional study investigating deep learning techniques for automatic identification of pre-segmented phases in videos of cataract surgery. One hundred cataract surgery videos performed by faculty and trainee surgeons were used and examined in 10 designated phases. Deep learning

https://doi.org/10.1016/j.xops.2022.100235
ISSN 2666-9145/22

algorithms accurately detected unique phases of cataract surgery through recognition of the surgical instruments.[13] Kim et al examined deep learning techniques for automated objective assessment of technical skills in capsulorrhexis. One expert surgeon first annotated 99 videos of capsulorrhexis as expert or novice performance through 2 capsulorrhexis indices in a standard structured rating scale, and then deep neural networks were used to model intraoperative surgical tool movement to identify technical skill level. They conclude that algorithms were able to effectively predict binary (expert or novice) capsulorrhexis technical skill classes.[14] However, pre-segmenting and pre-annotating videos prior to computer-based analysis may inherently introduce human bias into the objective analysis process. For our study, we refer to pre-segmentation as splicing of videos prior to computer analysis, and pre-annotation as specifying the ground-truth skill level prior to computer analysis.

The aim of our study is to investigate whether a deep neural network can correctly identify different surgical tools within cataract surgery without requiring pre-segmentation in an unsupervised approach, and secondly, to distinguish between expert and trainee surgical movements without pre-annotation via appointment status.

Methods

Institutional Review Board/Ethics Committee approval was obtained through the Health Sciences and Affiliated Teaching Hospitals Research Ethics Board at Queen's University, Kingston, Ontario, Canada. All research adhered to the tenets of the Declaration of Helsinki.

Consecutive cataract surgeries performed by a staff, trainee surgeon, or both at Hotel Dieu Hospital, Kingston Health Sciences Centre, Queen's University, Kingston, Ontario, Canada, between October 2018 and March 2019 were video recorded. Videos were recorded at 30 frames per second with a resolution of 1920 × 1080. At our institution, only trainee surgeons in their last (fifth) or second last (fourth) year of residency perform cataract surgery under direct supervision of faculty surgeons. None of the trainees at our institution had completed ophthalmology training elsewhere or in other countries. All patients provided informed consent for cataract surgery and intraocular lens implantation with the possibility of trainee involvement. Prior to participation in the study, informed consent for video recording was obtained from all staff and trainee surgeons involved in the cataract surgeries. Microscope video recording had no patient identifying features.

Following each surgical case, the responsible resident collected identifying data by completing a tracking form noting the surgeons (resident and faculty) and complexity of each case in order to ensure accurate annotation during data analysis. Cases were identified as either straightforward or complex. Complex cases consisted of the following: toric intraocular lens implant, hypermature cataract requiring VisionBlue, Malyugin ring, iris hooks, capsular tension ring insertion, posterior capsular rupture, and suturing of the cornea.

All videos were individually reviewed to ensure video quality and complete recordings. Videos of poor quality or incomplete cases were excluded from the dataset. Each included video was then appropriately annotated with the skill level of the surgeon(s) involved in the surgery, surgical techniques, and case specifics. Skill level consisted of either expert or trainee, or both expert and trainee. This was based only on appointment status, as this does not introduce human bias into the labels. Surgical techniques performed during surgery and those visible in the videos were labeled. The steps included the following: clear corneal incisions/Wong incision, dilating cocktail used, continuous curvilinear capsulorrhexis, nuclear disassembly, cortex removal, lens insertion, and manipulation and wound hydration.

Video analysis was conducted using deep neural networks involving 3 major components: (1) encoding each frame individually as a vector, (2) encoding video snippets as a vector using an unsupervised approach, and (3) classifying the skill level of each snippet (see Fig 1).

First, each microscope video frame was individually encoded as a vector (called "frame-level encodings"). This video frame encoding is intended to capture information about the entire frame, with emphasis on tool presence and location. To this end, we used the ResNet 152 network pretrained on ImageNet and fine-tuned it on the Cataracts Grand Challenge dataset for tool detection in cataract surgery.[15] We used the output of the second last layer of the network as an encoding of the frame (2048 element vector). The encoding is expected to contain information about instrument presence and pose. Prior work has validated this tool detection network on the Cataracts Grand Challenge dataset[15]; we validate our implementation on the same dataset using hold-out cross-validation with 3 surgeries held out for testing.

Second, video snippets were encoded in an unsupervised way (called "snippet-level encodings"). This snippet encoding is intended to capture temporal information about changes to the surgical scene with emphasis on tool motion, which is not discernable from a single video frame encoding. To this end, we cut each video into overlapping snippets of 100 frames in length (overlapping by up to 99 frames). We trained a long short-term memory autoencoder using the length 100 sequence of frame-level encodings to learn an encoding of video snippets. Subsequently, the encoder component was used to create snippet-level encodings of each video snippet (64 element vector).

Third, we trained a classifier to assess skill from video snippet-level encodings. We used a random forest classifier on the snippet encodings with 100 trees and balanced subsampling. The classifier was trained to predict binary skill label (novice versus expert) independently for each snippet.

We validated our skills assessment pipeline using a fivefold user-out cross-validation. The user-out cross-validation protocol ensures that whenever data from a given user appears in the testing set, data from that user never appears in the training or validation sets. The hyperparameters of the random forest classifier were manually tuned using the validation folds. Performance is reported on the test folds.

To measure performance of our methods for skill classification, we used area under the receiver operating characteristic curve (AUC) and the classification accuracy, which was trained with a balanced dataset. Confidence intervals (CIs) for performance measures are computed using a normal approximation, assuming each test fold is an independent sample. These measures of performance were computed for 3 different evaluation scenarios as follows: a) snippetwise, given a single snippet of video from one surgery, how well can we classify the skill level of the operator performing in that clip?; b) trialwise, given the entire video from one surgery, how well can we classify the skill level of the operator performing in that video?; and c) userwise, given all videos of surgeries completed by a single user, how well can we classify the skill level of the operator performing in those videos? The random forest classifier computes the probability that the input snippet encoding is from each of the novice and expert classes. To compute the trialwise skill of an operator, we computed the mean over all snippets within a trial of the novice and expert class probabilities. To compute the userwise skill of an operator, we computed the

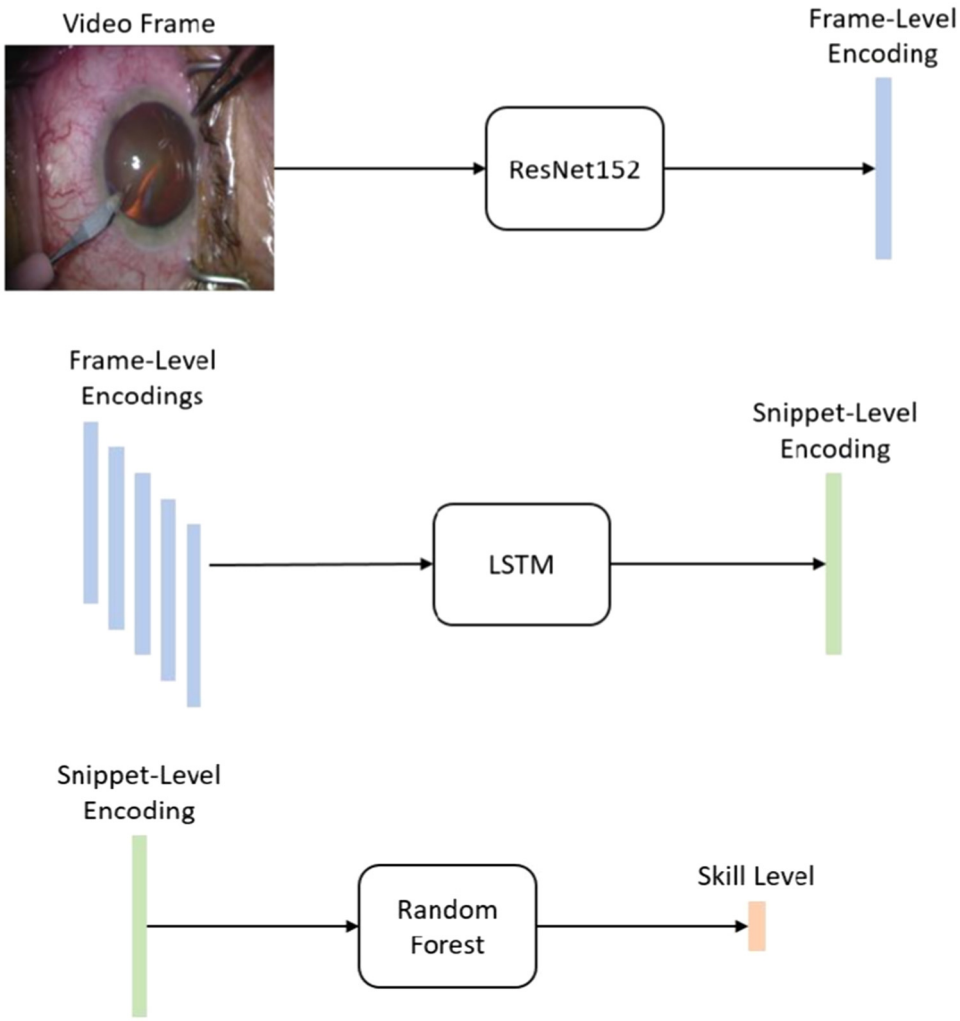

Figure 1. Components of skill classification model: frame-level encoding (top), snippet-level encoding (middle), skill level assessment (bottom). Each component is trained separately. LSTM = long short-term memory.

mean over all snippets performed by a user of the novice and expert class probabilities. We take the larger of the novice or expert probability as the most likely class.

Results

In total, 391 consecutive cases were recorded. Of these, 310 cases were classified as straightforward (79%) and 81 cases as complex (21%) (see Fig 2). Seven faculty surgeons (ranging from 1−14 years of practice after a 5 year resident program) and 5 trainee surgeons were involved in the surgeries, with the primary operating surgeon varying by case. As per our method criteria, we included straightforward cases performed by expert or trainee alone resulting in the inclusion of 209 cataract surgeries. All cases were done under topical anesthesia.

A few representative frames from our dataset and an illustration of their corresponding frame-level encodings from the tool detection network are demonstrated in Figure 3. Our model achieved an AUC ranging from 0.933 to 0.998 for 11 distinct tool detections on the Cataracts Grand Challenges dataset and their corresponding step of surgery[15] (see Table 1).

For skill classification of a single snippet (snippetwise), the AUC was 0.550 (95% CI, 0.547−0.553) and accuracy was 54.3% (95% CI, 53.9%−54.7%). For skill classification of a single surgery (trialwise), AUC was 0.570 (95% CI, 0.565−0.575) and accuracy was 57.8% (95% CI, 56.8%−58.7%). For skill classification of a single user given all of their trials (userwise), the AUC was 0.692 (0.659−0.758) and accuracy was 63.3% (56.8%−69.8%).

Discussion

Teaching tools such as didactic teaching, access to surgical simulation labs, and operating room teaching provide trainees with theoretical and practical training in cataract surgery. Surgical simulators can offer quantitative

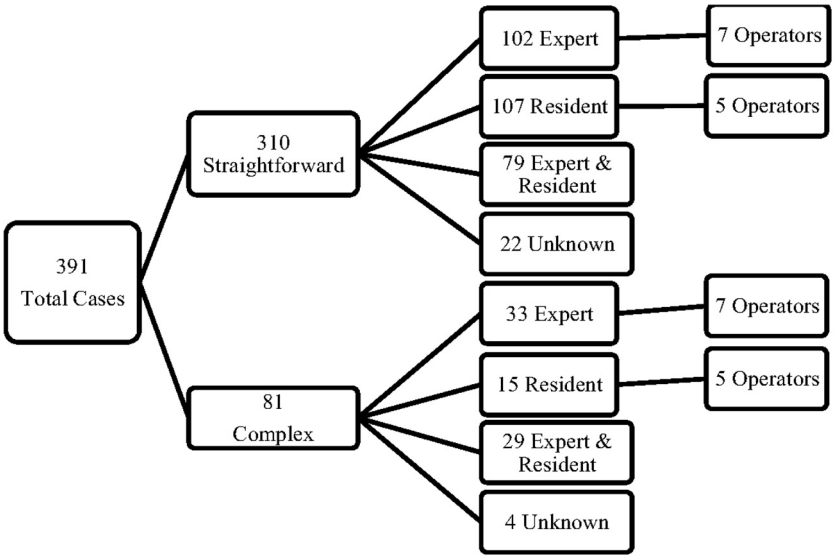

Figure 2. Breakdown of the consecutive cataract surgery cases.

information, allowing trainees to compare their skills relative to averages. However, a simulator's ability to provide direct feedback on how to improve in an operating room scenario is limited. Our research aims to provide an objective method whereby individual trainee's intraoperative cataract surgery steps can be analyzed and compared to expert norms.

We elected to use a "late supervision" approach to train our network. That is, we trained the first 2 components of our skills assessment network to encode video snippets without ground-truth skill labels. We only use the ground-truth skill labels in the final component of the approach. We conjecture that the snippet-level encodings will contain information about the surgeon's skill level that is robust to the particular criteria used to generate the ground-truth skill labels. While this "late supervision" approach may reduce performance for our particular task, it makes our model widely applicable across different cataract surgery centers, as only the final component must be retrained to new ground-truth skill labels. This reduces time, technical expertise, compute resources, and data requirements when deploying the model within various training curriculum or different cataract centers. This also removes the need for expert structured rating scales with the inherent variability and biases associated with human-based grades.

Our model achieved high accuracy in tool detection and the corresponding surgical step, being able to identify whether or not a tool was in the video frame. This indicates that the video frame encodings contain information about tool usage and position, which is an important indicator of skill. As for skill classification, using our "late supervision" approach, there was low accuracy in all 3 scenarios. However, there was some evidence that our model was able to classify operators by skill level. The skill level of the operating surgeon was most accurately classified when given all videos of surgeries that were completed by a single user (userwise), followed by when given the entire video from one surgery (trialwise), and then finally when given a single small clip of videos from one surgery (framewise). This suggests that in order to accurately classify an operator's skill level, videos of many of their trials may be needed for analysis; a small sample of frames may be insufficient. This is consistent with the competency by design training approach, that a small sample of evaluations is often insufficient, and multiple observations are required for proper assessment.

While our skill classification approach outperforms the zero-rule classifier, it has lower AUCs for skill classification in comparison to tool detection. This may be explained by the difference in training of the 2 networks. The tool detection network was trained to explicitly detect tools used in the surgery. However, the snippet encoding network was not explicitly trained to assess skills for our study as we used a "late supervision" approach. This network was trained to produce a representation that may be indicative of skill level (using an unsupervised approach), accounting for the lower AUCs. This is a trade-off for the added flexibility of the "late supervision" approach. A future study examining skill classification by using a network that is explicitly trained to assess skills may be warranted (similar to the work of Kim et al[14]). Furthermore, video classification methods have not been as well developed as methods for object detection in images. Lastly, machine learning for skill classification poses greater difficulty than tool detection. As opposed to the relatively straightforward process of determining whether a particular tool is present or absent in an image, the training it takes to understand the nuances of skill in surgery is lengthy and complex.

The large number of surgical videos collected was a strength of our study. Previous studies that examine the use of computer-based machine learning as an ophthalmology surgical training tool employ a total of approximately 100 videos.[13,14] Having a vast databank of multiple expert surgeons' techniques, including variation in instruments and their use in different phases across surgeons, allows for heterogeneity in data across settings to be captured.

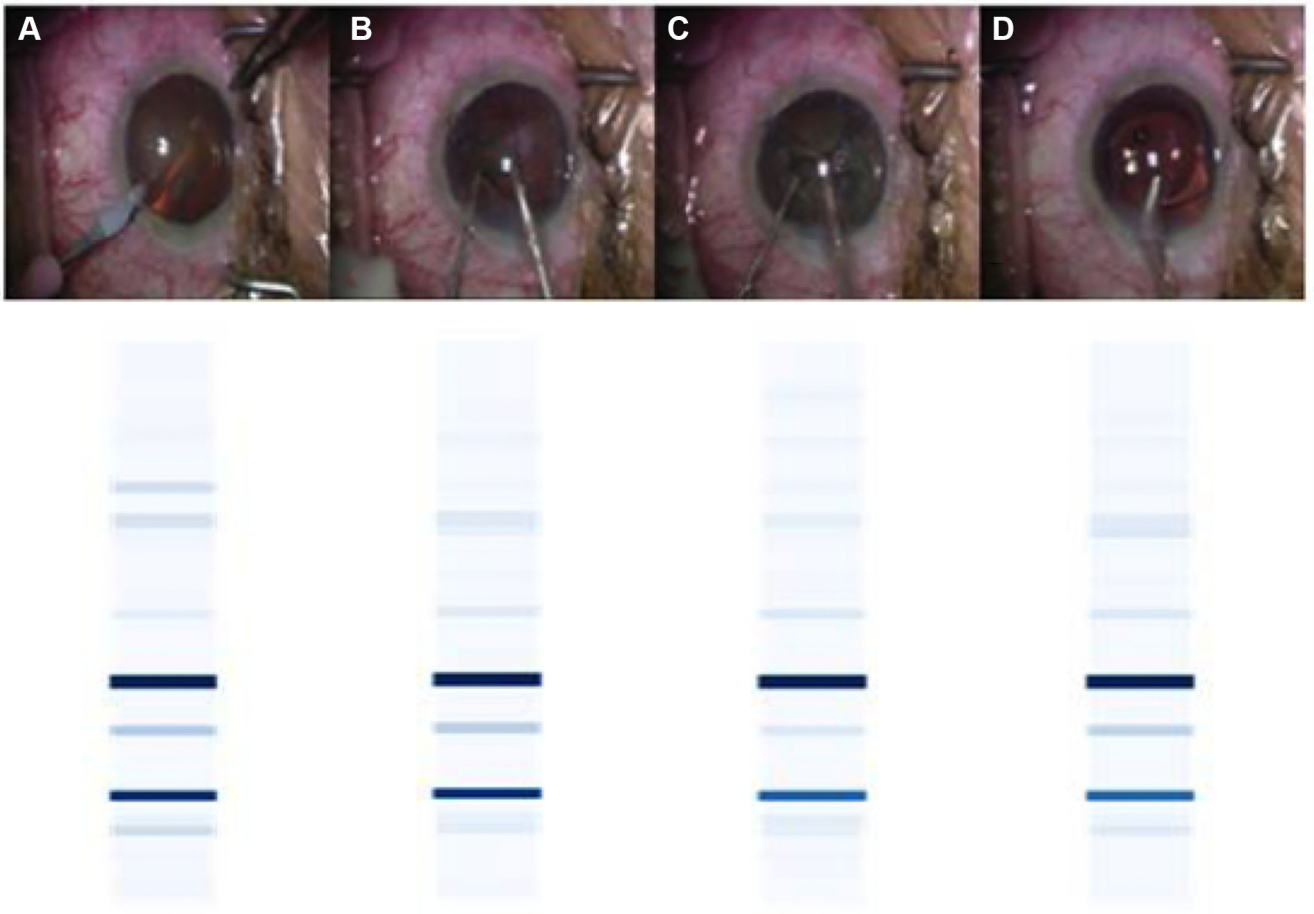

Figure 3. Representative cataract surgery video frames and their corresponding encodings from the neural networks. The shaded bars are visual representations of encodings of the frames from the videos (i.e. darkness is proportional to the magnitude of the element in the vector encoding): **A**, Creation of a main corneal incision with a keratome; **B**, Splitting a nucleus during phacoemulsification; **C**, Emulsification of a nuclear quadrant during phacoemulsification; **D**, Aspiration of viscoelastic with an irrigation and aspiration handpiece.

The algorithms for skill assessment are not influenced by surgeon-specific style.

Table 1. AUC Values for Tool Detection on the Cataracts Grand Challenge Dataset by Surgical Step

Tool	Corresponding Surgical Step	AUC
Paracentesis Blade	Side Incision	0.998
Viscoelastic Cannula	Viscoelastic	0.940
Keratome Blade	Main Incision	0.981
Cystotome	Capsurlorrhexis Creation	0.933
Utrata Forceps	Capsurlorrhexis Completion	0.968
Hydrodissection Cannula	Hydrodissection	0.979
Phacoemulsification Probe	Phacoemulsification	0.991
Irrigation-Aspiration Handpiece	Cortical Removal	0.990
Intraocular Lens Injector	Lens Insertion	0.982
Sinskey Hook*	Lens Manipulation	0.984
Hydration Cannula	Corneal Hydration	0.990

AUC = area under the receiver operating characteristic curve.
*A Lester Hook was used at our institution.

A limitation of our study was the lack of use of a structured rating scale to assess surgical skill, in conjunction with the machine learning analysis. The reasoning for our approach was due to the potential layer of bias by having an expert assess another expert's skills. Staff surgeons who are operating without supervision are assumed to be experts in their field and may be using different techniques that lead to identical surgical outcomes. In addition, although established cataract surgical skill assessment tools have shifted from subjective towards largely objective standardized measures, currently validated evaluation tools still involve the evaluators' subjective opinion.[16] Also to note, we chose to group trainees versus experts since there would not be enough video points for a continuous spectrum of expertise. Another limitation of our study was the large range of tools from several manufacturers used in the surgeries. The tool detection component of our model was trained to recognize tools on the Cataracts Grand Challenge dataset[15]; however, our dataset used tools from different manufacturers. Furthermore, our model needed to recognize numerous tools, some of which have similar appearance. Nevertheless, tool detection accuracy was high in our study.

The ultimate goal of creating an objective computer-based analysis system for cataract surgery is to provide valuable feedback to trainees based on intraoperative cases. Further research is required to determine the best network to identify skill classification, whether intermediate skill level stratification is possible, and the minimum number of surgical videos needed to create a reliable, reproducible, and valid network algorithm.

Footnotes and Disclosures

Originally received: March 31, 2022.
Final revision: October 5, 2022.
Accepted: October 18, 2022.
Available online: October 27, 2022. Manuscript no. XOPS-D-22-00065.

[1] Department of Ophthalmology, Kingston Health Sciences Centre, Queen's University, Kingston, Ontario, Canada.

[2] School of Computer Science, Carleton University, Ottawa, Ontario, Canada.

[3] School of Medicine, Faculty of Health Sciences, Queen's University, Kingston, Ontario, Canada.

[4] Laboratory for Percutaneous Surgery, School of Computing, Queen's University, Kingston, Ontario, Canada.

[5] Office of Health Sciences Education, Faculty of Health Sciences, Queen's University, Kingston, Ontario, Canada.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The authors have no proprietary or commercial interest in any materials discussed in this article.

HUMAN SUBJECTS:

Human Subjects were included in this study. Institutional Review Board/Ethics Committee approval was obtained through the Health Sciences and Affiliated Teaching Hospitals Research Ethics Board at Queen's University, Kingston, Ontario, Canada. All research adhered to the tenets of the Declaration of Helsinki. All participants provided informed consent.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Ruzicki, Holden, Egan, Law

Analysis and interpretation: Ruzicki, Holden, Cheon, Ungi, Egan, Law

Data collection: Ruzicki, Holden, Cheon, Ungi

Obtained funding: N/A

Overall responsibility: Ruzicki, Holden, Cheon, Ungi, Egan, Law

Meeting Presentation: Canadian Ophthalmological Society, Ottawa, Ontario, Canada, 2020 (Virtual).

Abbreviations and Acronyms:
AUC = area under the receiver operating characteristic curve; **CI** = confidence interval.

Keywords:
Cataract surgery, Artificial intelligence, Education.

Correspondence:
Christine Law, Department of Ophthalmology, Kingston Health Sciences Centre, 166 Brock Street, Kingston, Ontario, Canada K7L 5G2. E-mail: christine.law@queensu.ca.

References

1. Alwadani S. Cataract surgery training using surgical simulators and wet-labs: course description and literature review. *Saudi J Ophthalmol*. 2018;32:324–329.
2. Bozkurt Oflaz A, Ekinci Köktekir B, Okudan S. Does cataract surgery simulation correlate with real-life experience? *Turkish J Ophthalmol*. 2018;48:122–126.
3. Low SAW, Braga-Mele R, Yan DB, El-Defrawy S. Intraoperative complication rates in cataract surgery performed by ophthalmology resident trainees compared to staff surgeons in a Canadian academic center. *J Cataract Refract Surg*. 2018;44:1344–1349.
4. Tzamalis A, Lamprogiannis L, Chalvatzis N, et al. Training of resident ophthalmologists in cataract surgery: a comparative study of two approaches. *J Ophthalmol*. 2015;2015:932043.
5. Ament CS, Henderson BA. Optimizing resident education in cataract surgery. *Curr Opin Ophthalmol*. 2011;22:64–67.
6. Rahimy E. Deep learning applications in ophthalmology. *Curr Opin Ophthalmol*. 2018;29:254–260.
7. Grewal PS, Oloumi F, Rubin U, Tennant MTS. Deep learning in ophthalmology: a review. *Can J Ophthalmol*. 2018;53:309–313.
8. Du X-L, Li W-B, Hu B-J. Application of artificial intelligence in ophthalmology. *Int J Ophthalmol*. 2018;11:1555–1561.
9. Lu W, Tong Y, Yu Y, et al. Applications of artificial intelligence in ophthalmology: general overview. *J Ophthalmol*. 2018;2018:5278196.
10. Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol*. 2019;103:167–175.
11. Kapoor R, Walters SP, Al-Aswad LA, et al. The current state of artificial intelligence in ophthalmology. *Surv Ophthalmol*. 2019;64:233–240.
12. Hogarty DT, Mackey DA, Hewitt AW. Current state and future prospects of artificial intelligence in ophthalmology: a review. *Clin Exp Ophthalmol*. 2019;47:128–139.
13. 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:e191860.
14. Kim TS, O'Brien M, Zafar S, et al. Objective assessment of intraoperative technical skill in capsulorhexis using videos of cataract surgery. *Int J Comput Assist Radiol Surg*. 2019;14:1097–1105.
15. 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.
16. Puri S, Sikder S. Cataract surgical skill assessment tools. *J Cataract Refract Surg*. 2014;40:657–665.