

Title Page

Title: Use of machine learning to assess cataract surgery skill level with tool detection

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

Meeting Presentation: Canadian Ophthalmological Society, Canada, 2020

Financial Support: None

Conflict of Interest: No conflicting relationship exists for any author

Running Head: Cataract surgery assessment using machine learning

Key words: cataract surgery, artificial intelligence, education

1

[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

Corresponding Author: Christine Law, Department of Ophthalmology, Kingston Health Sciences Centre, 166 Brock Street, Kingston, Ontario, Canada, K7L 5G2. Email: christine.law@queensu.ca

Abstract

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 labelled accordingly 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.

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 (accuracy 54.3%) for a single snippet; AUC was 0.570 (accuracy 57.8%) for a single surgery, and AUC was 0.692 (accuracy 63.3%) for a single user given all of 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 level of training or expertise.

Manuscript

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 (CBD) 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, optical coherence tomography, 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 two studies from the Wilmer Eye Institute, Johns Hopkins University, Baltimore, Maryland, USA 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 ten designated phases. Deep learning 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 two capsulorrhexis indices in a standard structured rating scale, 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 grading 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 distinguish between expert and trainee surgical movements without pre-annotation via appointment status.

Methods

Institutional Review Board (IRB)/Ethics Committee approval was obtained through the Health Sciences and Affiliated Teaching Hospitals Research Ethics Board at Queen's University, Kingston, Ontario, Canada.

Consecutive cataract surgeries performed by a staff and/or trainee surgeon 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 resolution 1920 x 1080. At our institution, only trainee surgeons in their last (5^{th}) or second last (4^{th}) 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 other countries. All patients provided informed consent for cataract surgery and intraocular lens (IOL) 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 IOL implant; hypermature cataract requiring VisionBlue; Malyugin ring; iris hooks; capsular tension ring (CTR) insertion; and posterior capsular rupture (PCR).

All videos were individually reviewed to ensure video quality and complete recordings. Videos of poor quality and/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, trainee, or both expert and trainee. Surgical techniques performed during surgery and visible in the videos were labelled. The steps included the following: clear corneal incisions/Wong incision; dilating cocktail used; continuous curvilinear capsulorrhexis (CCC); and nuclear disassembly.

Video analysis was conducted using deep neural networks involving three 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 Figure 1).

First, each microscope video frame was encoded individually 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 pre-trained on ImageNet and fine-tuned it on the Cataracts Grand Challenge dataset for tool detection in cataract surgery.[16] 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. This tool detection network was validated on the Cataracts Grand Challenge dataset using hold-out cross-validation.

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, that is not discernable from a single video frame encoding. To this end, we cut each video into overlapping snippets 100 frames in length. We trained a long short-term memory (LSTM) autoencoder using the 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 vs. expert) for each snippet independently.

We validated our skills assessment pipeline using five-fold 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. 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 for performance measures are computed using a normal approximation, assuming each test fold is an independent sample. These measures of performance were computed for three different evaluation scenarios: 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?. Trialwise and userwise skill levels were computed as a mean over all snippets present for the trial or user.

Results

In total, 391 consecutive cases were recorded. Of these, 310 cases were classified as straightforward (79%), and 81 cases as complex (21%) (see Figure 2). Seven faculty surgeons (ranging from 1-14 years of practice after a 5 year resident program) and five 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[16] (see Table 1).

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

Discussion

Teaching tools such as didactic teaching, access to surgical simulation labs, and operating room teaching, provides trainees with theoretical and practical training in cataract surgery. Surgical simulators can offer quantitative information, allowing trainees to compare their skills relative to averages. However, a simulator's ability to provide direct feedback on how to improve in a real-world 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 two 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 varioius training curriculum or different cataract centres. 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 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 three 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 completed by a single user (userwise), followed by when given the entire video from one surgery (trialwise), 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 CBD training approach that a small sample of evaluations is often insufficient, and multiple observations are required for proper assessment.

The lower AUCs for skill classification in comparison to tool detection may be explained by the difference in training of the two networks. The tool detection network was trained to explicitly detect tools used in the surgery. However, the snippet encoding network was not trained explicitly 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. A future study examining skill classification by using a network that is trained explicitly to assess skills may be warranted. 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. The algorithms for skill assessment are not influenced by surgeon-specific style.

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.[17] 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[16]; 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.

References

1. Alwadani S. Cataract surgery training using surgical simulators and wet-labs: course description and literature review. Saudi J Ophthalmol 2018 Oct-Dec;32(4):324-9.
2. Bozkurt Oflaz A, Ekinci Köktekir B, Okudan S. Does cataract surgery simulation correlate with real-life experience? Turkish J Ophthalmol 2018 Jun;48(3):122-6.
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 Nov;44(11):1344-9.
4. Tzamalis A, Lamprogiannis L, Chalvatzis N, Symeonidis C, Dimitrakos S, Tsinopoulos I. 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 Jan;22(1):64-7.
6. Rahimy E. Deep learning applications in ophthalmology. Curr Opin Ophthalmol 2018 May;29(3):254-60.
7. Grewal PS, Oloumi F, Rubin U, Tennant MTS. Deep learning in ophthalmology: a review. Can J Ophthalmol 2018 Aug;53(4):309-13.
8. Du X-L, Li W-B, Hu B-J. Application of artificial intelligence in ophthalmology. Int J Ophthalmol 2018 Sep;11(9):1555-61.
9. Lu W, Tong Y, Yu Y, Xing Y, Chen C, Shen Y. Applications of artificial intelligence in ophthalmology: general overview. J Ophthalmology 2018 Nov;2018:5278196.
10. Ting DSW, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, Tan GSW, Schmetterer L, Keane PA, Wong TY. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol 2019 Feb;103(2):167-75.

11. Kapoor R, Walters SP, Al-Aswad LA, Lee AG, Raab E. The current state of artificial intelligence in ophthalmology. Surv Ophthalmol 2019 Mar-Apr;64(2):233-40.

12. Hogarty DT, Mackey DA, Hewitt AW. Current state and future prospects of artificial intelligence in ophthalmology: a review. Clin Experiment Ophthalmol 2019 Jan;47(1):128-39.

13. Yu F, Silva Croso G, Kim TS, Song Z, Parker F, Hager GD, Reiter A, Vedula S, Ali H, Sikder S. Assessment of automated identification of phases in videos of cataract surgery using machine learning and deep learning techniques. JAMA Netw Open 2019 Apr;2(4):e191860.

14. 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 Jun;14(6):1097-105.

15. Lee A, Taylor P, Kalpathy-Cramer J, Tufail A. Machine learning has arrived! Ophthalmology 2017 Dec;124(12):1726-8.

16. Al Hajj H, Lamard M, Conze PH, Roychowdhury S, Hu X, Maršalkaitė G, Zisimopoulos O, Dedmari MA, Zhao F, Prellberg J, Sahu M, Galdran A, Araújo T, Vo DM, Panda C, Dahiya N, Kondo S, Bian Z, Vahdat A, Bialopetravičius J, Flouty E, Qiu C, Dill S, Mukhopadhyay A, Costa P, Aresta G, Ramamurthy S, Lee S-W, Campilho A, Zachow S, Xia S, Conjeti S, Stovanov D, Armaitis J, Heng, P-A, Macready WG, Cochener B,, Quellec G CATARACTS: challenge on automatic tool annotation for cataRACT surgery. Med Image Anal 2019 Feb;52:24-41.

17. Puri S, Sikder S. Cataract surgical skill assessment tools. J Cataract Refract Surg 2014 Apr;40(4):657-65.

Figure Legends

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

Figure 2: Breakdown of the consecutive cataract surgery cases.

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.

Figure 1

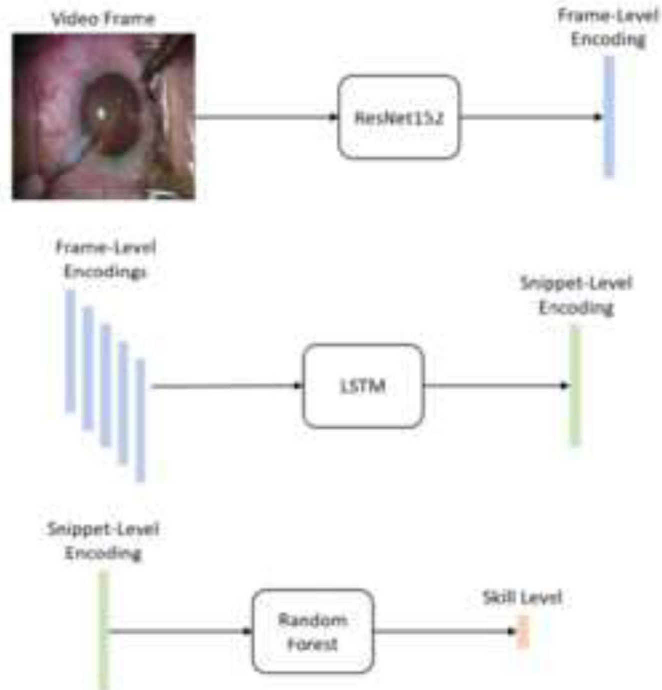
Video Frame
Frame-Level
Encoding
ResNet152
Frame-Level
Encodings
Snippet-Level
Encoding
LSTM
Snippet-Level
Encoding
Random
Forest
Skill Level

Figure 2

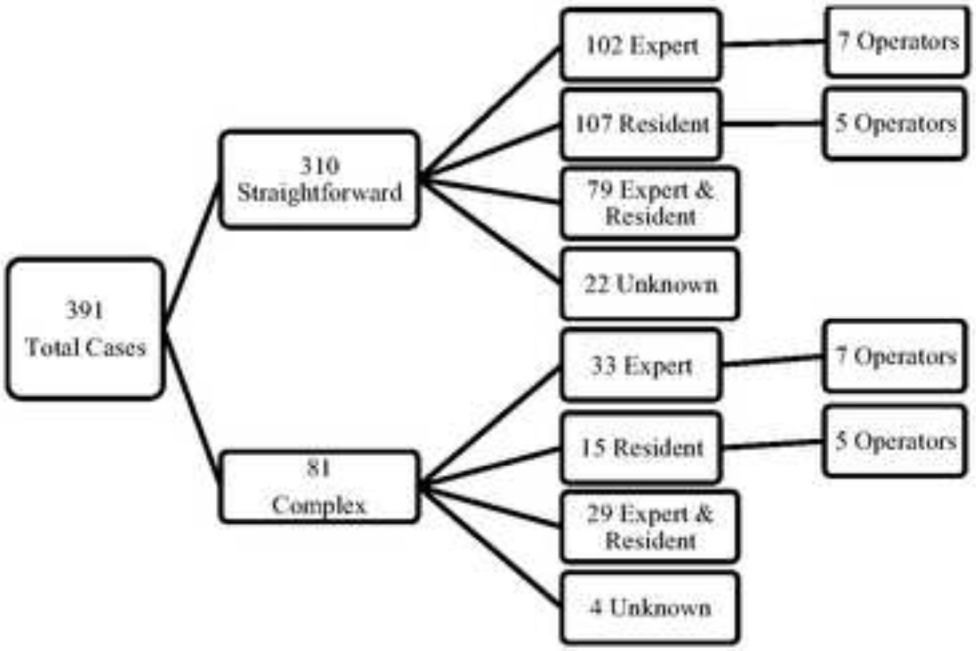
391
Total Cases
310
Straightforward
81
Complex
102 Expert
107 Resident
79 Expert & Resident
22 Unknown
33 Expert
15 Resident
29 Expert & Resident
4 Unknown
7 Operators
5 Operators
7 Operators
5 Operators

Figure 3

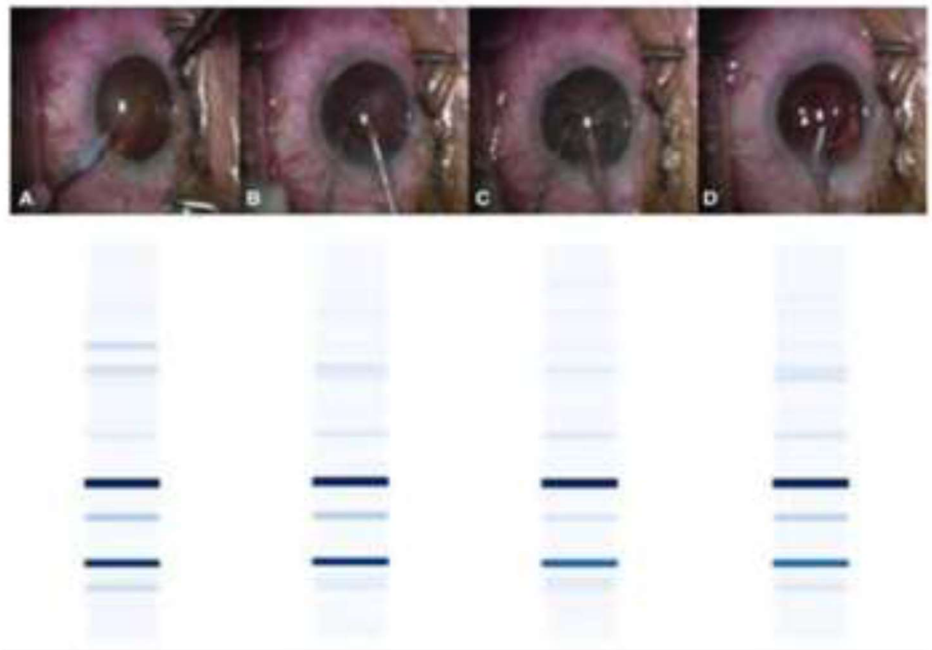

Table 1: Area under the receiver operating characteristic curve (AUC) values for tool detection on the Cataracts Grand Challenge dataset by surgical step.

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