Performance of a Novel Artificial Intelligence Model for Cataract Surgery Phase Identification using a Real World Cataract Surgery Video Dataset

Kirsty Clarke[1],Ayushi Gupte[1], Joanna Li[2], Asmita Anand[3], Azeem Siddique[4], Ajay Mohite[1], Marco Anastasi[1], Franciso Porto Guerra E Vasconcelos[3], Sophia Bano[3], Aleksandra Goch[5], Marinko Sarunic[5], Riaz Asaria[1]
[1]Royal Free London NHS Foundation Trust, [2]UCL Medical School, [3]UCL Dept of Computer Science, [4]Buckinghamshire Healthcare NHS Trust, [5]UCL Department of Medical Physics and Biomedical Engineering

Introduction

Artificial Intelligence (AI) has the potential to revolutionise cataract surgery training and self-evaluation by enabling automated analysis of surgical videos—a common tool used by surgeons to refine their technique.

However, most existing AI models have been trained on small, homogenous datasets from expert surgeons operating under ideal conditions, limiting their accuracy and generalisability in real-world clinical settings.

To bridge this gap, we developed and tested a deep learning model designed to analyse cataract surgery videos recorded under routine operating conditions. Our dataset included cases performed by consultants, senior trainees, and junior trainees, reflecting real-world variation in surgical technique, case complexity, and video quality.

The foundation of such analysis is accurate phase recognition—the identification of key procedural steps such as capsulorhexis, hydrodissection, and phacoemulsification. This study evaluates our model's ability to detect 12 distinct cataract surgical phases, and examines how surgeon experience, phase difficulty, and phase duration influence model performance.

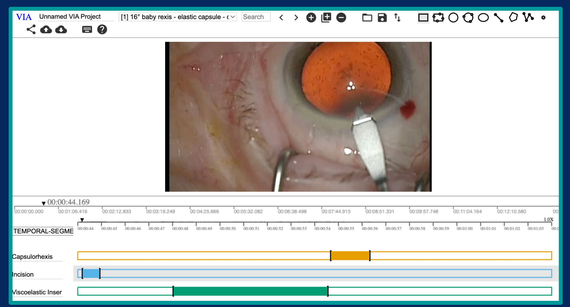


Figure 1: Video annotation software to label ground truth
www.robots.ox.ac.uk/~vgg/software/via/app/via_video_annotator.html

Methods

We manually annotated 100 cataract surgery videos with ground truth timestamps for 12 distinct surgical phases using the VGG Image Annotator (Figure 1). Videos were evenly distributed across three groups: consultants, senior trainees, and junior trainees. Each surgical phase was independently graded as standard or difficult by two senior ophthalmologists.

An AI model was trained on the publicly available Cataract-101 dataset using a ResNet-50 architecture to classify video frames into surgical phases. The trained model was then tested on our heterogeneous dataset, which included surgeries performed under varying conditions (e.g., microscope type, video quality, surgeon experience).

Model performance was assessed by comparing predicted phase timelines to manual annotations:

- Confusion matrices evaluated phase classification accuracy.
- Confidence scores (range: 0.0–1.0) were used to measure model certainty.
- Independent t-tests compared mean confidence scores across surgeon grades.
- Frequency of phase prediction was used to assess recognition accuracy per phase.
- Spearman's rank correlation tested the relationships between phase difficulty, phase length, and prediction frequency.
- All analyses were performed using IBM SPSS Statistics (v30).

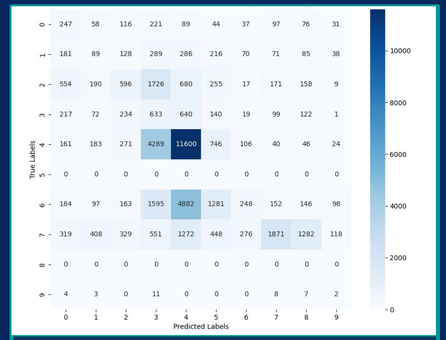


Figure 2: Confusion Matrix
Model predictions across all 9 phases of cataract surgery. Most misclassifications occur between visually similar phases like Phase 2 Hydrodissection vs Phase 3, Capsulorhexis.

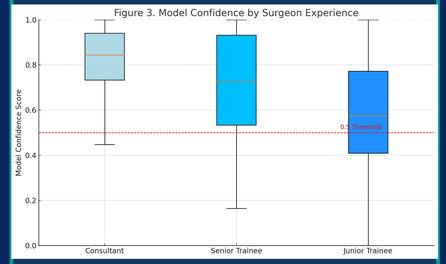


Figure 3: Box plot showing the distribution of AI model confidence scores stratified by surgeon experience
Consultant-performed surgeries yielded higher median confidence scores with tighter distributions, while junior trainee videos showed lower median scores and greater variability, reflecting reduced model certainty and more inconsistent visual input.

Results & Conclusion

As predicted our model showed higher phase classification accuracy on the standardised Cataract 101 dataset than on our real-world surgical video dataset. In our dataset, misclassifications often occurred between visually similar phases (e.g., Hydrodissection vs Viscous Agent Injection). Model performance decreased with increasing phase difficulty, confirming that complex or ambiguous phases are harder to label (Figure 2).

- Average model confidence was 0.72 ± 0.35.
- Consultant videos had the highest confidence (0.85 ± 0.15), followed by senior (0.72 ± 0.25 and junior trainees (0.60 ± 0.32) (Figure 3)
- Phacoemulsification had the highest prediction frequency; Irrigation & Aspiration the lowest.
- A negative correlation was found between phase difficulty and prediction frequency ($\rho = -0.6$, $P \leq 0.05$), (Figure 4).
- As expected, a positive correlation was found between phase difficulty and phase length ($\rho = 0.8$, $P < 0.05$).

Overall, this study serves as a proof of concept that an AI model can reliably identify cataract surgery phases from real-world operative videos. Performance is influenced by phase complexity and surgeon experience, highlighting areas for refinement.
Future improvements should target enhanced model resilience to surgical variability to better support trainee learning and performance feedback. As larger datasets are constructed we anticipate the model performance will improve.

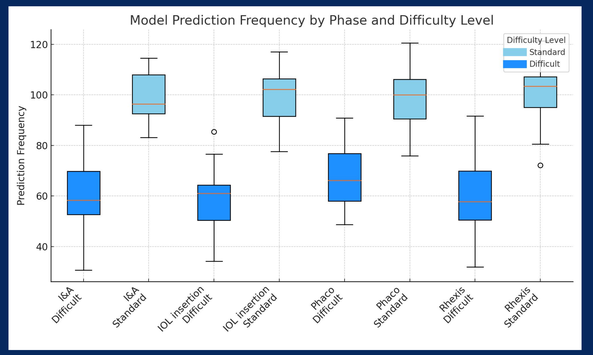


Figure 4: Box plots comparing model performance by phase and difficulty level across four surgical steps. Prediction frequency was significantly lower in Difficult cases across all phases.

UCL

NHS
Royal Free London
NHS Foundation Trust