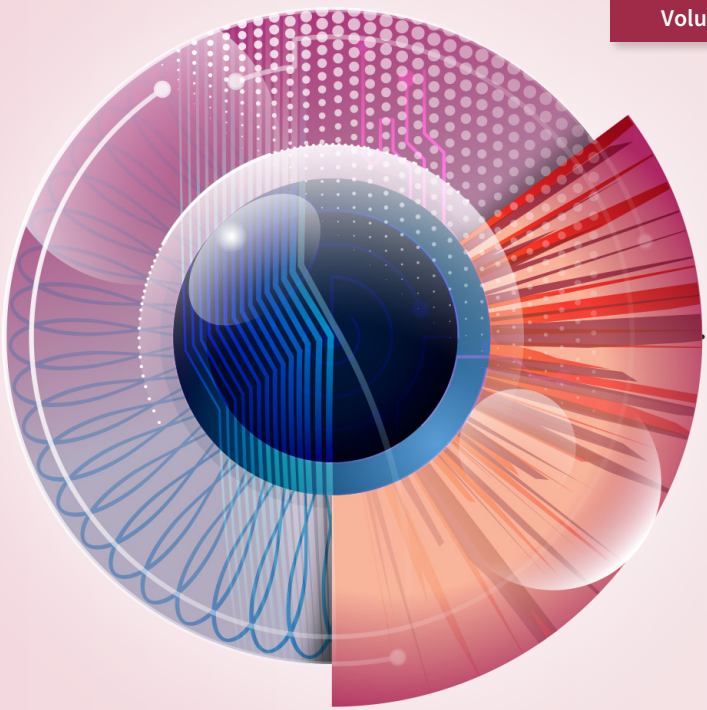
aivojournal.com
Volume 2| Supplement



Abstracts from SAIVO 2026 Annual Meeting
Where AI Meets Vision

Artificial Intelligence in Vision & Ophthalmology

Official Journal of the Society for Artificial Intelligence in Vision and Ophthalmology (SAIVO)

ISSN
Print: 3051-2328 | Online: 3117-4035

SAIVO

Society for AI in Vision and Ophthalmology

Established 2025

SAIVO - the Society for Artificial Intelligence in Vision and Ophthalmology - is the world's first professional society formally dedicated to AI and big data in eye care. SAIVO stands at the intersection of medicine, technology, and scientific innovation, with a shared commitment to improving patient outcomes through responsible, patient-centered AI.

WWW.SAIVO.ORG

AIVO

Artificial Intelligence in Vision and Ophthalmology

While the rapid advance of imaging technologies in ophthalmology is making available a continually increasing number of data, the interpretation of such data is still very challenging and this hinders the advance in the understanding of ocular diseases and their treatment. Interdisciplinary approaches encompassing ophthalmology, physiology, mathematics, engineering, and computer science have shown great capabilities in data analysis and interpretation for advancing basic and applied clinical sciences. Artificial Intelligence in Vision and Ophthalmology (AIVO) was created with the aim of providing a forum for interdisciplinary approaches integrating mathematical and computational methods with experimental and clinical studies to address open problems in ophthalmology. AIVO welcomes articles that investigate questions related to the anatomy, physiology and function of the eye in health and disease.

For further information on AIVO's focus and scope as well as manuscript submissions:

www.aivojournal.com
aivo@aivojournal.com

Copyright

Publisher
Kugler Publications
P.O. Box 20538
1001 NM Amsterdam
The Netherlands

info@kuglerpublications.com
www.kuglerpublications.com

ISSN
Online: 3117-4035
Print: 3051-2328

Manuscript submissions
Author guidelines and templates are available via the website, through which all manuscripts should be submitted. For inquiries please contact us via e-mail.

Publication frequency
AIVO uses the Continuous Article Publication (CAP) model. Articles are published online as soon as they are ready.

Advertising inquiries
AVIO offers online and in print sponsorship and advertising opportunities. Please contact Kugler Publications to for inquiries.

Submit your article now:

Open access policy

AIVO is fully open access without requiring any publication fee from the authors. Publication fees will be introduced in 2026.

SAIVO – Society for Artificial Intelligence in Vision and Ophthalmology

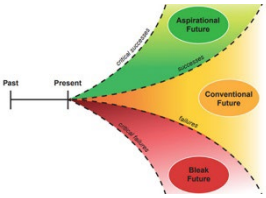
AIVO is the official journal of SAIVO. SAIVO is the first society in the world formally dedicated to artificial intelligence and big data in the fields of vision and ophthalmology, brings together experts in medicine, science, and technology to advance the safe and effective use of artificial intelligence in eye care. SAIVO supports innovation, education, and collaboration to improve diagnosis, treatment, and outcomes for patients around the world. SAIVO is committed to advancing the development, validation, and clinical integration of artificial intelligence technologies in eye care. More information on how to join: https://www.saivo.org/

SAIVO
Society for AI in Vision and Ophthalmology
Established 2025

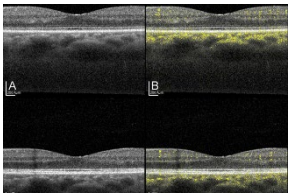
Disclaimers

Table of Contents

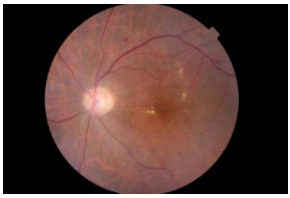
Recently Published Articles



Development and planning for future scenarios in ophthalmology: content generation using a modified Delphi process
Timothy P. Mayotte, Rafid Q. Farjo, Krystal D. Kao, Harrison Wong, Christopher Nagata, Paul Salow, Shahzad I. Mian, K. Thiran Jayasundera



Analysis of the anatomical and functional ocular changes related to spaceflight
Gal Antman, Irit Bahar, Alon Tiosano, Alon Harris, Yamit Cohen-Tayar, Yair Zimmer, Amoy Fraser, Mehul Patel, Iftach Yassur, Itay Gabbay, Yehonatan Weinberger, Keren Wood, Orly Gal-Or



Multimodal large language models for use in diabetic retinopathy screening
S. Saeed Mohammadi, Sahana Aggarwal, Kavina Aggarwal, Grant Wiarda, Kayla Nguyen, Emmanuel A. Sarmiento, Quan Nguyen, Manjot K. Gill

Submit your article

AIVO uses the Continuous Article Publication (CAP) model. Articles are published online as soon as they are ready.

AIVO
Artificial Intelligence in Vision and Ophthalmology

SAIVO 2026 Annual Meeting Abstracts

V3xV: A Video-Based Computer Vision Algorithm for Real-Time Volumetric Dosing Precision in Subretinal Gene Therapy

David Almeida[1], Vinit Mahajan[2], Michael Singer[3]

[1]The Centers for Advanced Surgical Exploration (CASEx), ERIE, United States; [2]Stanford University, Palo Alto, United States; [3]The Centers for Advanced Surgical Exploration (CASEx), Miami, United States

Purpose: Subretinal delivery of viral vectors is currently plagued by a qualitative problem: surgeons rely on visual estimation for dosing, resulting in a ± 30-50 µL variance (approx. 25% deviation) from target volumes. This lack of precision leads to efficacy dilution, safety risks from reflux, and significant protocol deviations in clinical trials. We introduce SurgVECTORai, a software-as-a-medical-device (SaMD) that utilizes a novel Video-based Viral Vector Volumetric (V3xV) analysis engine to provide real-time, hardware-agnostic dosing intelligence.

Methods: The V3xV algorithm processes standard 2D surgical video feeds without the need for sequential frame-by-frame analysis: 1) Background Subtraction to isolate the injection site; 2) Gaussian Blur for noise reduction; 3) Canny Edge Detection to define bleb contours; and 4) Geometric Characterization using an ellipsoid approximation model to calculate volume. The system classifies anatomical delivery location (Subretinal vs. Vitreous) using a machine learning classifier trained on reflux patterns. Real-time feedback is displayed via a traffic light indicator (Green: Subretinal >95% probability; Red: Reflux Detected).

Results: In validation testing, the SurgVECTORai platform demonstrated a volumetric measurement precision of ±0.1 µL, significantly outperforming traditional visual estimation. The system successfully generated automated Dosing Logs and Audit Trails, creating an objective temporal profile of the injection (Early, Mid, Late phases). The anatomical classifier achieved high confidence (0.816) in distinguishing successful bleb formation from vitreous dispersion, triggering immediate alerts upon reflux detection to prevent wasted therapeutic payload.

Conclusion: SurgVECTORai transforms subretinal gene therapy from a subjective art to a quantitative science. By leveraging the V3xV computer vision pipeline, the platform provides the first objective metric for dosing compliance, ensuring protocol fidelity, and enabling pay-for-performance verification for high-cost gene therapies. This technology offers a scalable, equipment-agnostic solution to standardize dosing precision across clinical trial sites globally.

CLEAR: Consensus Layer Evaluation for Artifial Intelligence Algorithm Reporting, A Framework for Standardized Evaluation in Age-Related Macular Degeneration

Omer Trivizki

Tel Aviv Medical Center, Tel Aviv, Israel

Purpose: Artificial intelligence (AI) algorithms are increasingly used to derive quantitative imaging biomarkers for age-related macular degeneration (AMD). While reading center manual annotations remain the gold standard for measurements, they are limited by scalability, cost, and inter-grader variability. AI offers the potential for rapid and scalable analysis, while enabling adoption of newer imaging endpoints. However, the lack of standardized definitions and reporting practices has limited the acceptance of AI as a reliable clinical trial grading tool. The CLEAR (Consensus Layer Evaluation for AI Algorithm Reporting) initiative was established to address these challenges by developing a consensus framework for evaluating and reporting AI-derived retinal layer biomarkers.

Methods: CLEAR is a multi-stakeholder initiative involving clinicians, reading center experts, AI developers and industry partners. The framework focuses on AI algorithms designed to quantify retinal endpoints relevant to AMD beyond GA area growth, including retinal pigment epithelium loss, ellipsoid zone loss, hypertransmission defects, hyperreflective foci area, calcified drusen progression and others. Key elements include standardized anatomical definitions, alignment with reading center grading conventions, transparent reporting of algorithm design, and validation against expert human annotations. The initiative further explores how AI can support clinical trials and the creation of predictive models for visual outcomes.

Results: The initiative identified substantial variability across existing AI algorithms in how anatomical biomarkers are defined, segmented, and quantified, limiting comparability and regulatory acceptance. The CLEAR framework highlights distinctions between eligibility tools and validated clinical trial endpoints, emphasizing that AI algorithms must demonstrate performance equivalent to expert human graders to support endpoint adoption. Preliminary applications illustrate the potential of AI to improve recruitment efficiency, support structure–function analyses, and enhance disease progression prediction, while preserving the need for standardized definitions.

Conclusion: The CLEAR initiative proposes a structured approach to evaluating and reporting AI-derived retinal biomarkers. By standardizing AI outputs and reports with reading center standards and emphasizing transparent validation, CLEAR aims to enable trustworthy, scalable imaging endpoints that can accelerate clinical trial conduct and support earlier intervention in AMD. Validation is essential for this process – the CLEAR study sets the benchmark for reliable AI-based clinical trial endpoints.

Global Benchmark for AMD Algorithm Validation

Amitha Domalpally, Barbara Blodi, Roomasa Channa

University of Wisconsin, Madison, United States

Purpose: Regulatory approval of artificial intelligence (AI) algorithms requires prospective validation against reference standards. However, conducting independent prospective studies for each algorithm is costly and burdensome, creating a major barrier to clinical translation. We describe the design and early progress of a multicenter prospective study establishing a benchmark dataset intended to support consistent performance evaluation, demonstrated with the use case of age-related macular degeneration (AMD) screening.

Methods: The AMD Benchmark Imaging Dataset (ABID) study is an ongoing, prospective, global multicenter observational study, spanning healthy to advanced AMD spectrum. Participants undergo standardized retinal imaging, including dilated and undilated color fundus photography and optical coherence tomography (OCT). Images are graded at a centralized reading center using reference standard labels with Beckman Scale. Clinical metadata includes best-corrected visual acuity, demographics, and clinical diagnosis. The dataset is intended for algorithm performance benchmarking rather than model training.

Results: As of the current data cutoff, institutional review board approval has been obtained for U.S. and international sites (including EU, Africa and Asia), with 10 of 21 planned sites activated. A total of 166 participants were enrolled. The cohort includes 60% female and 40% male participants, with 76% aged ≥65 years and 24% aged 50–64 years. Based on centralized grading using the Beckman AMD scale, 24% had no AMD, 2% had early AMD, 49% had intermediate AMD, 15% had geographic atrophy, and 10% had neovascular AMD.

Conclusions: This prospective benchmark study directly addresses an important gap in facilitating deployment of AMD screening algorithms. Continued enrollment toward a target of approximately 1,000 participants is expected to support regulatory relevant benchmarking and accelerate clinical translation.

Ultra-widefield Diabetic Retinopathy Prescreening Using a Patch-Based Attention Multiple Instance Learning Model

Varsha Satish[1,2], Robert Slater[1,2], Nancy Barrett[2], Tom Saunders[2], Rachel Linderman[1,2], Barbara Blodi[2], Amitha Domalpally[1,2]

[1]A-Eye Research Unit, Department of Ophthalmology and Visual Sciences, University of Wisconsin-Madison, Madison, United States; [2]Wisconsin Reading Center, Ophthalmology & Visual Sciences, University of Wisconsin-Madison, Madison, United States

Purpose/Background: Ultra-widefield (UWF) color fundus imaging provides extensive retinal coverage but introduces a trade-off between field of view, resolution, and computational cost for automated analysis. Reliable prescreening at the ETDRS ≥47 threshold depends on detecting

small, spatially sparse lesions that are easily lost with naive down-sampling. This study evaluates a weakly supervised, patch-based Multiple Instance Learning (MIL) framework with attention-based aggregation designed to preserve lesion-level detail while using only image-level ETDRS labels.

Methods: A total of 835 UWF images from multiple diabetic retinopathy (DR) studies were included, labeled as ETDRS <47 (345 eyes) versus ≥47 (318 eyes) based on double grading with adjudication. Images were cropped into the mid-periphery to remove lashes and background, then tiled into 224×224 patches to maintain local spatial resolution. Each patch was encoded using a pre-trained EfficientNet_B0 backbone, producing a 1280-dimensional feature vector; all patch embeddings from an image formed a bag of instances. An attention-based MIL pooling module learned patch-level importance weights and aggregated features into a single image-level representation, which was passed to a linear classifier to predict ETDRS <47 versus ≥47. Model development used 5-fold cross-validation on 663 eyes, with 172 eyes held out as an independent test set. Attention heatmaps were generated by projecting attention weights back to patch locations for qualitative assessment of alignment with DR pathology.

Results: Across 5-fold cross-validation, mean out-of-fold performance was: accuracy 0.69, AUROC 0.74, sensitivity 0.52, specificity 0.82, and F1 score 0.59. On the independent test set, the model achieved accuracy 0.61, AUROC 0.76, sensitivity 0.40, specificity 0.87, and F1 score 0.65. The model correctly identified 87% of ETDRS ≥47 eyes (65/75) and 40% of ETDRS <47 eyes (39/97), indicating a high-specificity, lower-sensitivity operating point. Attention maps frequently emphasized patches containing DR-related abnormalities, supporting clinically plausible instance weighting under weak supervision.

Conclusion: A patch-based attention MIL framework enables DR severity classification on large UWF images while preserving fine lesion detail typically lost with global down-sampling. Future work will focus on improving sensitivity via hybrid global–local architectures and task-aligned loss functions.

Using Natural Language Processing to Evaluate Screening Practice Compliance and Prevalence of Hydroxychloroquine Retinal Toxicity Among Patients Using Hydroxychloroquine at Loyola Chicago

Samantha Dreyer[1], Meriam Ben Hadj Tahar[2], Sarah Tajran[2], Pablo Vendrell[1], Baraa Hussein[1], Mohammad Saad[3], Jhansi Raju[2]

[1]Loyola University Chicago Stritch School of Medicine, Maywood, United States; [2]Loyola University Medical Center, Department of Ophthalmology, Maywood, United States; [3]Loyola University Medical Center, Maywood, United States

Purpose/Background: Hydroxychloroquine (HCQ), a commonly used therapy for many autoimmune conditions, carries a known risk of retinal toxicity. The American Academy of Ophthalmology (AAO) published screening guidelines in 2016 (which were recently updated in

2025). This retrospective study aims to evaluate compliance with the 2016 AAO guidelines among HCQ users at Loyola University Medical Center (LUMC), determine the feasibility of using NLP systems to monitor screening efforts, and examine the prevalence of retinal toxicity in both screened and unscreened populations.

Methods: We performed a retrospective electronic health record analysis at LUMC (2012–2025) using an NLP system to identify patients prescribed HCQ. Ophthalmic screening information was extracted from both structured data fields and unstructured clinical notes. The NLP platform identified medication exposure, duration of therapy, ophthalmology visits, screening methods (OCT, visual field testing, and fundus examinations), and documentation of retinal toxicity. Initial challenges included inconsistent terminology, ambiguous language, and incomplete billing data. These limitations were addressed by expanding keyword libraries, incorporating diagnostic and billing codes, and validating manual chart review. Patients with at least five years of continuous HCQ use were stratified as having guideline-concordant, partial, or no screening according to the 2016 AAO guidelines.

Results: 809 out of the 5,474 patients taking HCQ met the criteria for long-term exposure. Using the NLP system, we efficiently categorized screening patterns and identified gaps across a large cohort, showing that only 37.2% of patients received appropriate screening with both OCT and visual field testing. Of those who were screened, retinal toxicity was identified in 0.6% of patients. Higher toxicity rates among partially screened and unscreened patients were identified. Incomplete or absent screening was associated with a markedly higher proportion of retinal toxicity when compared with full adherence.

Conclusion: This study highlights the practicality and limitations of using NLP for the evaluation of screening adherence in a large cohort. The results suggest that NLP-based analysis can help facilitate quality improvement efforts and inform system-level strategies to limit irreversible retinal toxicity. Future studies will analyze the effect that these methods have on HCQ screening practice, along with other ophthalmic measures.

Evaluating Large Language Models' Confidence on Predicting Glaucoma Progression to Surgery using Free-Text Clinical Notes

Jawwad Javeed, Sophia Wang

Byers Eye Institute, Stanford University School of Medicine, Palo Alto, CA, United States

Purpose

The rise of large language models (LLMs) in performing many general tasks has prompted interest in evaluating their utility in healthcare applications. The purpose of this study was to benchmark the ability of open-source LLMs in predicting patients' progression to glaucoma surgery using free-text clinical ophthalmology notes. We evaluated the accuracy of each LLMs' confidence through prompting (extrinsic) or via probabilities associated with the yes/no tokens generated (intrinsic).

Methods: Ophthalmic free-text clinical notes from 2008-2020 for a previously identified cohort of adult patients with glaucoma with >= 120 days of follow-up were extracted from Stanford's STRIDE. We prompted 8 open-source LLMs (Llama 70B, Qwen 3 32B, Qwen 2.5 72B, LEME 70B-SFT; Quantized: Qwen 2.5 72B-AWQ, Qwen 2.5 32B-AWQ, Gemma 3 27B 4-bit, MedGemma 27B 4-bit) to predict whether patients would progress to require glaucoma surgery from clinical notes. Models were prompted to answer yes or no and to estimate the probability of progression to surgery (extrinsic probability). The probability associated with the generated yes or no token was also extracted (intrinsic probability). Standard evaluation metrics included area under the receiver operating curve (AUROC), area under the precision-recall curve (AUPRC), F1 score, precision, recall, and accuracy on a held-out test set, evaluated on intrinsic/extrinsic probabilities.

Results: We included 4512 glaucoma patients, of whom 748 progressed to glaucoma surgery (16.6%). From the test set, 500 patients were included, of whom 84 progressed to glaucoma surgery (16.8%). For classification performance, Qwen 2.5 72B-AWQ had the highest F1 score (0.324) and precision (.3052), while Qwen 2.5 32B-AWQ had the highest accuracy (.772) and Gemma 3 27B 4-bit had the highest recall (.7738). Of the extrinsic probabilities generated, Qwen 32B had the highest AUROC (.642). The intrinsic probabilities of the LEME 70B-SFT model had the highest AUROC (.612).

Conclusions: LLMs were able to predict glaucoma progression to surgery from clinical notes with modest performance. Compared to prior work, LLM performance could outperform an ophthalmologist's predictions (F1=29.9%) but not previously trained models that were expressly trained for this task (AUROC=0.734). Further research may be needed to optimize LLM performance on clinical prognostic tasks.

Venture Capital Investment in Ophthalmology Artificial Intelligence

Victoria Gu[1,2], Jipeng Yue[3], Austin Gu[4,5], Jingli Guo[2], Amir Akhavanrezayat[2], Negin Yavari[2], Ngoc Than[2], Gunay Uludag[2], Dalia Elfeqi[2], Osama Elaraby[2], Gunjan Awatramani[2], David Xu[2], Bin Mo[2], Mohammad Rajabi[2], Azadeh Mobasserian[2], Anh Tran[2], Ishaan Iyer[2], Alan Sherif[2], Isaac Sanchez[2], Aubrey Nguyen[2], Henry Moshfeghi[2], Jia-Horung Hung[2], Quan Nguyen[2]

[1]Georgetown University School of Medicine, Washington, United States; [2]Stanford Byers Eye Institute, Palo Alto, United States; [3]University of Connecticut School of Medicine, Farmington, United States; [4]Columbia Business School, New York, United States; [5]Fu Foundation School of Engineering and Applied Science, New York, United States

Purpose: Venture capital (VC) funding is a principal driver of the development and commercialization of AI-enabled medical devices and software, and investment trajectories have direct implications for the clinical tools available to ophthalmologists. Despite growing interest in these technologies, the landscape of VC investment in ophthalmology AI, including the clinical applications attracting capital, funding maturity, and regulatory outcomes, has not been systematically characterized.

Methods: A retrospective analysis of VC investments in US ophthalmology AI companies from January 2016 to December 2025 was conducted using the PitchBook database, with AI and machine learning (ML) applied as an industry classification filter. Companies were manually classified by clinical category (Autonomous Diagnosis & Screening, Clinical Decision Support, Surgical & Procedural AI, Systemic Biomarkers, Practice Operations) and relevant subspecialty and use case. Deal flow, capital investment, funding stage distribution, and temporal trends were assessed. FDA regulatory status was determined through cross-referencing the 510(k) and De Novo databases and the FDA's AI/ML-enabled medical device list.

Results: A total of 133 VC deals were made in 51 ophthalmology AI companies, totaling $613.3M (median $1.60M [IQR $0.20–$5.00M]). 'Autonomous Diagnosis & Screening' captured 69.4% of capital ($425.4M). Later-stage VC accounted for 51.0% of investment, followed by early-stage (30.8%) and seed (9.8%). Investment peaked in 2021 at $167.6M before declining through 2023, coinciding with a shift in capital composition as seed funding fell from 23.4% to 6.8% and later-stage funding rose from 27.8% to 64.2% between the first (2016–20) and second (2021–25) halves of the study period. Of the five FDA-approved AI/ML-enabled ophthalmic devices, four were VC-backed, raising a combined $277.5M (45.3%). All four were cleared between 2018 and 2022 for autonomous diabetic retinopathy screening, with the majority of capital deployed in later-stage rounds post-clearance, consistent with FDA approval functioning as a de-risking event for subsequent investment.

Conclusion: VC investment in ophthalmology AI was concentrated in autonomous diagnostic applications, with multiple companies and funding volume converging on retinal screening indications. Post-2021 investment contraction may reflect both macroeconomic trends observed across healthcare AI sectors and implementation barriers, including reimbursement and workflow integration, that remain unresolved despite regulatory progress.

AI-inferred fluorescein angiography using generative adversarial networks or diffusion models

Toshinori Murata[1], Takao Hirano[1], Hideaki Mizobe[2], Shuhei Toba[2]

[1]Shinshu University, Matsumoto, Japan; [2]Canon, Tokyo, Japan

Purpose: Optical coherence tomography angiography (OCTA) delineates retinal vessels without the need for contrast dye, eliminating concerns about allergic reactions. However, a significant drawback of OCTA is its low detection rate of microaneurysms (MAs) and vascular leakage compared to fluorescein angiography (FA). We have previously reported a novel system for generating FA-like images (AI-inferred FA) from OCTA (Murata T et al., Biomed Opt Express, 2023). We evaluated the ability of this new OCTA-based AI-inferred FA using GANs or diffusion models to detect changes in MAs and vascular leakage before and after anti-VEGF injections for diabetic macular edema (DME).

Methods: We obtained OCTA images (OCT-S1, Canon, Tokyo, Japan) and FA images from five eyes of five patients with DME before and after three intravitreal injections of faricimab. The number of MAs within the 3 mm central area of each FA (early phase), OCTA, and OCTA-based AI-inferred FA (early phase) image was evaluated before and after anti-VEGF therapy. Additionally, the macular leakage area within the 3 mm central area of each FA (late phase) and OCTA-based AI-inferred FA (late phase) image was assessed.

Results: The number of MAs was significantly reduced after treatment compared to baseline in FA (early phase), OCTA, and OCTA-based AI-inferred FA (early phase) images (19.0 ± 10.8, 6.8 ± 3.1, 16.6 ± 9.2 at baseline; 5.8 ± 5.3, 1.2 ± 1.8, 3.0 ± 3.5 after treatment; $P < 0.05$, $P < 0.01$, $P < 0.05$, respectively). The total number of MAs identified by OCTA was only 32.2% (40/124) of that detected by FA, whereas OCTA-based AI-inferred FA identified 86.3% (107/124). The macular leakage area was significantly reduced after treatment in FA (late phase) (2.86 ± 0.95 mm^2 to 0.35 ± 0.12 mm^2, $P < 0.01$). A significant reduction was also observed in OCTA-based AI-inferred FA (2.49 ± 1.09 mm^2 to 0.27 ± 0.17 mm^2, $P < 0.05$).

Conclusion: OCTA-based AI-inferred FA more accurately depicted the reduction in MAs and macular leakage area following anti-VEGF injections than OCTA. AI-inferred FA can non-invasively depict retinal circulatory dynamics and so has the potential to be helpful in the management of DME.

Artificial Intelligence Analysis of Home Optical Coherence Tomography (OCT): A Longitudinal Variation

Anat Loewenstein

Tel Aviv Medical Center, Tel Aviv, Israel

Background: Longitudinal validation of computed trajectories tracking changes in key biomarkers in neovascular age-related macular degeneration (nAMD).

Methods: The analysis was performed on longitudinal home OCT data from nAMD patients self-acquired for an approximately 5-week period. An expert grader classified the eyes as stable or changed. The longitudinal trajectories were scored by the Signal to Noise Ratio (SNR). A Receiver Operating Characteristic curve (ROC) of the SNR compared to the reference provided the optimal threshold (OT) value for the identification of total retinal hyporeflective (TRO) volume change; these were compared to a reference change value (RCV) and population-based threshold value.

The main outcome measures were area under the curve (AUC), sensitivity, specificity and accuracy at the optimal threshold.

Results: The data from 180 participants that initiated testing at home was reviewed. A total of 296 trajectories were included in analysis, 107 (36.1%) were classified as having a change, and 189 (63.9%) trajectories classified as stable.

The SNR ROC curve had an AUC of 0.9811 and an optimal SNR score of 2.42, for which the sensitivity was 99.1%, specificity 89.4%, and accuracy 94.2%. For reference, similar results were observed with a common RCV of 2.8, for which the sensitivity was 97.2%, specificity 89.9%, and accuracy 93.6%. The ROC for population-based thresholds, had an AUC of 0.9687, an optimal threshold of 3.88 VU, sensitivity of 94.4%, specificity 89.4%, and accuracy 91.9%.

Conclusions: The AI-based algorithm was able to accurately track and differentiate between stable home OCT trajectories and those with changes over time. The threshold values using the individualized RCV approach were more sensitive than the population-based approach in detecting changes.

To Evaluate the Efficacy of Five Large Language Models using Zero-Shot Prompting in the Extraction of Microbial Keratitis Descriptors

Lokeshwari Aruljyothi[1], Shruthi Banerjie[2], Tushar Mungle[2], Maria A Woodward[3], Prajna Venkatesh[4], Nambi Nallasamy[3]

[1]Aravind Eye Hospital, Salem, India; [2]Stanford University, Stanford, United States; [3]Kellogg Eye Center, Ann Arbor, United States; [4]Aravind Eye Hospital, Madurai, India

Purpose: To evaluate the efficacy of 5 different Large Language Models (LLMs) in the extraction of microbial keratitis (MK) descriptors from clinician notes in electronic health records (EHRs) using a zero-shot prompting approach.

Methods: 4999 patients with culture-proven fungal MK seen between 2019 and 2024 at 5 tertiary care centers of Aravind Eye Hospital (AEH), India were gathered. Free-text clinical notes from each patient's first encounter corneal examination were obtained. Each of the three MK descriptors—centrality, infiltrate depth, and thinning—was annotated by expert consensus and coded as 1 (present), 0 (absent), or 9 (details unavailable). GPT-4o and GPT-4o mini, Llama3.1, Owen, and MedGemma were prompted to extract the three MK descriptors. The models' performance was determined by comparing with expert human annotations which is considered gold standard.

Results: GPT-4o demonstrated accuracy (agreement rate) of 95%, 82%, and 90%, for centrality, depth, and thinning, respectively, Similarly GPT-4o mini demonstrated accuracy of 83%, 68%, and 90%. The accuracy rates of Llama, Qwen and MedGemma ranged between 37-40% for centrality, 42-60% for depth and 44-87% for thinning.

Conclusions: Both GPT-4o and GPT-4o mini showed good agreement with human annotations in extracting MK descriptors when compared with the other three models with GPT-4o demonstrating the highest performance. Detection of MK descriptors was influenced by limitations in the quality and consistency of EHR documentation.

A Multimodal Deep Learning Framework for the Prediction of Subclinical Retinal Vasculitis

Ling Chen

Department of Ophthalmology, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China; Department of Ophthalmology, Eye and ENT Hospital of Fudan University, Shanghai, China

Background: Subclinical retinal vasculitis—defined as fluorescein angiography (FA) evidence of inflammation in the absence of clinical signs on fundus photography— invasive procedures for diagnosis and carries risks of potential adverse effects. Furthermore, precisely differentiating Behçet's disease (BD) uveitis from idiopathic retinal vasculitis (RV) remains difficult with conventional imaging. This study aimed to develop and validate non-invasive multimodal imaging artificial intelligence (AI) models using paired fundus photography and optical coherence tomography (OCT) to: (1) detect subclinical retinal vasculitis and (2) distinguish BD from idiopathic RV.

Methods: This multicenter study aggregated data from five clinical centers. Retinal vasculitis was annotated based on fluorescein leakage or vascular occlusion. For subclinical detection (Part 1), the development cohort encompassed 53,077 OCT and 34,704 fundus images (2,222 patients), with independent external validation at two centers. For BD-RV differentiation (Part 2), the development cohort included 1,615 patients with BD and 607 with RV, while four external centers provided validation data. Deep learning models were trained via supervised learning using single-modal (OCT or fundus) and multimodal (OCT + fundus) architectures. Performance was quantified using the area under the receiver operating characteristic curve (AUC).

Results: In Part 1, a pilot model trained on a 10% data subset achieved an AUC of 0.80 for detecting subclinical vasculitis. Similarly, for Part 2, the single-modal pilot models (OCT and fundus photography) both yielded an AUC of approximately 0.82 in diagnostic differentiation. The multimodal fusion pilot model significantly enhanced performance, achieving an AUC of 0.85 in the development cohort and retaining a robust AUC of 0.83 during external validation.

Conclusion: Deep learning-based multimodal AI models integrating fundus photography with OCT reliably identify subclinical retinal vasculitis and differentiate BD uveitis from idiopathic RV. The synergistic integration of multimodal data significantly outperforms single-modality approaches, offering a non-invasive, objective tool that may reduce the clinical reliance on invasive angiography in the management of retinal inflammatory diseases.

Automated Surgical Planning for Epiretinal Membrane: Validation of the SurgMAP AI Predictive Retinal Epiretinal Surgical Stratification (PRESS) System

David Almeida[1], Vinit Mahajan[2]

[1]The Centers for Advanced Surgical Exploration (CASEx), ERIE, United States; [2]Stanford University, Palo Alto, United States

Purpose: Epiretinal membrane (ERM) surgery is historically challenged by variable membrane-retina adherence and limited intraoperative stereopsis, often leading to inefficient peeling and iatrogenic retinal trauma. We aimed to validate the logic engine of SurgMAPai, a novel platform that utilizes high-resolution Optical Coherence Tomography (OCT) to identify the Submembrane Space (SuMS). This study evaluates the system's proprietary classification algorithm, Predictive Retinal Epiretinal Surgical Stratification (PRESS), to predict optimal surgical entry points and stratify case complexity.

Methods: This prospective, consecutive interventional case series analyzed 137 eyes undergoing pars plana vitrectomy for ERM. Preoperative high-resolution volume scans were analyzed to identify SuMS—a distinct hyporeflective space between the ERM and Internal Limiting Membrane (ILM). The SurgMAP AI logic categorized membranes into four PRESS classes: Type 1 (Broad/Deep SuMS), Type 2 (Focal/Shallow SuMS), Type 3 (Adherent/Opaque SuMS), and Type 4 (Vitreomacular Traction associated). Surgical technique was modified based on these AI-driven predictive maps to target pre-identified SuMS for initial engagement. Primary endpoints included first-attempt peeling success rate, operative efficiency, and complication rates (petechial hemorrhage).

Results: The PRESS classification successfully stratified surgical behavior across all phenotypes. Type 1 membranes demonstrated the highest first-attempt success rate at 95.3%, followed by Type 2 (82.1%), Type 4 (81.8%), and Type 3 (71.4%). Visual acuity improvement at 3 months correlated with PRESS class complexity, with Type 1 eyes gaining a mean of +3.4 ETDRS lines compared to +2.3 lines in Type 4. Importantly, the system effectively predicted safety risks; Type 1 membranes had a significantly lower rate of petechial hemorrhage (23.3%) compared to the more adherent Type 3 membranes (52.4%).

Conclusion: The integration of SuMS identification via the PRESS system provides a robust, predictive framework for ERM surgery. These data validate the core algorithms of the SurgMAP AI platform, demonstrating that preoperative OCT stratification can accurately predict intraoperative membrane behavior. By transitioning ERM surgery from subjective visual assessment to data-driven surgical planning, SurgMAP AI offers a pathway to standardize outcomes and reduce surgical trauma.

Virtual Reality Simulation to Improve Ophthalmic Knowledge and Diagnostic Skills in Internal Medicine Residents

Aarun Devgan[1], Jhansi Raju[1], Joel VandeLune[2]

[1]Loyola Stritch School of Medicine, Chicago, United States; [2]Roy J. and Lucille A. Carver College of Medicine, Iowa City, United States

Purpose and Background: Internal medicine residents frequently encounter ocular complaints but receive limited formal ophthalmology training, contributing to low confidence and missed

opportunities for timely diagnosis and referral. Virtual reality (VR) offers an immersive approach to teaching complex ophthalmic anatomy and clinical reasoning through interactive 3D visualization and simulated patient encounters. This study evaluated the effectiveness of a VR-based simulation curriculum in improving residents' ophthalmic knowledge, examination skills, and diagnostic accuracy.

Methods: Eighty PGY-1 and PGY-2 internal medicine residents participated in a faculty-led VR simulation lab incorporating (1) stereoscopic 3D anatomy instruction, (2) interactive holographic case-based training, and (3) immersive headset-based modules focused on ocular anatomy, pupillary/cranial nerve abnormalities, and glaucoma pathophysiology. Participants completed identical pre- and post-session assessments measuring ocular terminology knowledge, anatomical correlation of pupillary and cranial nerve findings, and diagnostic accuracy using virtual patients. A post-session survey assessed perceived skill gains and clinical relevance. Paired t-tests compared pre- and post-intervention scores.

Results: Sixty-three participants completed both pre- and post-assessments (63/80; 78.8%). Mean (SD) ocular terminology knowledge scores increased from 4.8 (1.0) to 5.5 (0.9) ($P<0.001$). Scores assessing anatomical correlation of pupillary and cranial nerve findings improved from 1.6 (1.1) to 2.5 (0.7) ($P<0.001$). Diagnostic accuracy on virtual patient cases increased from 2.4 (1.5) to 3.9 (1.3) ($P<0.001$). Most participants reported improved understanding of pupillary dysfunction (95.2%) and cranial nerve dysfunction (93.5%), and improved ability to perform pupil (90.3%) and ocular motility exams (91.9%). Residents rated the VR curriculum as clinically relevant across modalities (82.3%–91.9%).

Conclusions: VR-based simulation curriculum significantly improved internal medicine residents' ophthalmic knowledge, examination skill performance, and diagnostic accuracy. VR simulation represents a scalable, technology-enabled approach to strengthening ophthalmic competency in non-ophthalmology trainees and may support earlier recognition of vision-threatening conditions in clinical practice.

Calibration for Color Aberration in Confocal Scanning Laser Ophthalmoscopy by Generative Artificial Intelligence

Yi-Ting Hsieh[1,2], Hsu-Hang Yeh[1]

[1]National Taiwan University Hospital, Taipei, Taiwan. [2]National Taiwan University Hospital, Hsin-Chu Branch, Hsinchu, Taiwan

Purpose

Color scanning laser ophthalmoscopy (cSLO) provides wide-field high-resolution retinal images, but often exhibits unrealistic color representation due to the limited wavelengths of its scanning laser sources. This study aimed to develop a deep generative learning model to calibrate cSLO images so that their color appearance more closely resembles that of color fundus photography (CFP).

Methods: We retrospectively collected 27,580 cSLO images and 31,302 CFP images from patients at one medical center. A CycleGAN framework was employed, consisting of a forward generator to produce color-calibrated images, a backward generator to reconstruct the original cSLO images, and a discriminator to differentiate generated images from real CFP. Model performance was evaluated on a test set of 65 cSLO images. Color similarity between the generated images and corresponding CFP was independently assessed by retina specialists.

Results: The model effectively corrected the characteristic greenish hue in the posterior pole of cSLO images across various lesion types and locations, and enhanced the contrast between optic disc cups and neural rims. No spatial distortion was observed, and consistent performance was maintained in eyes with media opacity.

Conclusions: Deep generative models can selectively correct pseudo-color artifacts in cSLO images, particularly the greenish hue resulting from disproportionate short-wavelength laser reflectance. This color calibration approach has the potential to reduce diagnostic ambiguity and improve clinical interpretation of cSLO imaging.

AI-Driven multi-omics integration in Vogt-Koyanagi-Harada disease reveals microbial, proteo-metabolomic, and immune pathways predicting disease activity

Jingli Guo, Jia-Horung Hung, Victoria Gu, Ngoc Trong Tuong Than, Amir Akhavanrezayat, Gunay Uludag Kirimli, Negin Yavari, Azadeh Mobasserian, Dalia El Feky, Gunjan Awatramani, Bin Mo, Alan Sherif, Cristian de los Santos, David Chanthan, Isaac Sanchez, David Xu, Aubrey Nguyen, Ishan Suresh Iyer, Mohammad Rajabi, Christopher Or, Diana V Do, Quan Dong Nguyen

Byers Eye Institute at Stanford, Palo Alto, United States

Purpose: To elucidate the molecular signatures of Vogt-Koyanagi-Harada (VKH) disease through integrative analysis of gut microbiome, metabolomic, proteomic, and single-cell transcriptomic data, and to evaluate whether artificial intelligence (AI) can accurately classify disease activity by linking microbiome-metabolome-immune interactions in this retrospective cross-sectional study.

Methods: Multi-omics datasets of VKH were harmonized, including gut microbiome metagenomics (71 VKH, 67 controls), urine metabolomics (26/26), plasma metabolomics (55/30), aqueous humor metabolomics (15/15), sweat proteo-metabolomics (30/30), and PBMC single-cell RNA-seq (3/3). After quality control and batch correction, standardized feature matrices were integrated using multi-view learning. Ensemble machine-learning models (XGBoost, Random Forest) and graph neural networks were developed for three predictive tasks: distinguishing VKH from controls, classifying active vs inactive VKH, and predicting immunosuppressive treatment response. Transfer learning with external autoimmune-disease datasets enhanced model generalizability. Model explainability was assessed by SHAP and pathway enrichment analyses.

Results: Integrated modeling achieved AUC 0.92 for VKH vs controls and 0.90 for active vs inactive VKH. Discriminative features included enriched taxa (Pediococcus, Rhodococcus) and reduced

Lachnospiraceae, together with amino-acid and lipid metabolic alterations. Proteins ENPP2 and FUCA1, as well as ISG15+ monocyte subclusters, contributed to higher prediction accuracy. Mediation analysis suggested that approximately 30% of microbiome effects on VKH risk were mediated through metabolite changes. Transfer-learning models showed improved stability with minimal performance drop across cohorts.

Conclusions: AI-driven multi-omics integration reveals interconnected microbial, proteo-metabolomic, and immune pathways in VKH, supporting a microbiome-to-metabolome-to-immunity axis underlying disease activity. This integrative framework enables molecular phenotyping and provides a foundation for precision diagnosis, and activity monitoring in VKH diseases.

Retinal Encoding for Neuromorphic and AI-Driven Vision Systems Using Difference of Gaussians-Based Spiking Representations

Hannah Rana[1,2], Mohammad Eslami[1], Michael Morley[1], Nazlee Zebardast[1,3], Mengyu Wang[1], Tobias Elze[1]

[1]Harvard Ophthalmology AI Lab, Schepens Eye Research Institute of Massachusetts Eye and Ear, Harvard Medical School, Boston, Massachusetts, United States; [2]Department of Neurosurgery, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, United States; [3]Department of Ophthalmology, Massachusetts Eye and Ear, Harvard Medical School, Boston, Massachusetts, United States

Purpose: Artificial retinal prostheses and neuromorphic vision systems require encoding strategies that efficiently translate visual scenes into spike-based representations while preserving key aspects of retinal structure. While spiking neural networks (SNNs) provide a powerful AI framework for vision, the design of retinal-inspired front-end encoders that simultaneously address biological plausibility, computational efficiency, and prosthetic constraints remains an open challenge. This work presents a Difference of Gaussians (DoG)-based retinal encoding framework that functions as a biologically grounded front-end encoding layer for AI-driven and prosthetic vision systems.

Methods: A DoG spatial filter was implemented to simulate ON-center and OFF-surround retinal ganglion cell responses using center and surround standard deviations of 2σ and 5σ, respectively, on an 11×11 pixel kernel. Visual inputs included synthetic stimuli and Poisson-encoded subsets of the NE15-MNIST dataset. Filter outputs were converted into spike-based representations compatible with SNN inference and training pipelines, and mapped onto a simulated electrode grid, to reflect retinal prosthetic stimulation constraints. Encoding performance was evaluated using sparsity (percentage of inactive electrodes) and total spike count as quantitative metrics of efficiency and biological plausibility, alongside qualitative analysis of spatial and temporal activation patterns.

Results: The encoding framework generated spiking activity that preserved spatial contrast features across both synthetic and benchmark datasets. Mean sparsity was 78.5% (±3.2%), indicating highly energy-efficient encoding, while spike counts ranged from approximately 1300 to 2100 spikes per image. Encoded outputs exhibited clear contrasting ON-OFF response patterns consistent with retinal ganglion cell behavior, supporting the interpretability and biological relevance of the spiking representations.

Conclusions: This study demonstrates that DoG-based retinal encoding provides an effective and computationally efficient front-end for neuromorphic and AI-driven vision systems, particularly for retinal prosthetic stimulation. By combining biologically inspired contrast processing with sparse, spike-based representations, the framework supports real-time visual encoding and integration with downstream SNN architectures. This work contributes a biologically grounded front-end representation for SNNs, addressing a critical gap between retinal ganglion cell contrast encoding and energy-efficient AI vision systems. This approach offers a practical pathway toward adaptive, AI-enabled retinal prostheses and generalizable neuromorphic vision pipelines that support artificial vision.

Generative Artificial Intelligence for Ethical Reasoning in Ophthalmology: Exploring Comparisons Between a Large Language Model and a Human Ethicist

Daniel Kelly, Ishan Chillikatil, Rebika Khanal, Andrew Trippiedi, Chelsea-Jane Arcalas, Matthew Claxton, Tochukwu Ndukwe, Elizabeth Pogrebniak, Joshua Barnett, Jacquelyn O'Banion, Jeremy Jones, Rebecca F. Neustein

Department of Ophthalmology, Emory University School of Medicine, Atlanta, United States

Purpose/Background: Large language models (LLMs) have garnered significant attention as potential adjuncts to aid clinical decision-making. However, there is a dearth of evidence describing the ability of AI-based applications to contemplate ethical dilemmas in ophthalmology. This study explored the ability of a LLM to address common ethical scenarios faced by ophthalmologists.

Methods: Ten publicly available ethical inquiries on the *Ask the Ethicist* column of the American Academy of Ophthalmology's website were randomly selected. ChatGPT 5.1 (OpenAI) was prompted to respond to each scenario, matching the word count of the corresponding human ethicist response. The following readability index scores were generated for all responses: Flesch Reading Ease (FRE); Flesch-Kincaid Grade Level; Gunning Fog Index (FOG); SMOG Index; Coleman-Liau Index; and the Automated Readability Index. A physician panel (6 ophthalmologists; 4 ophthalmology trainees) independently rated all responses, using 6-point and 5-point Likert scales, for two outcomes: (1) likelihood-of-use and (2) perceived patient impact.

Results: AI-generated responses demonstrated a lower FRE (10.9 ± 9.6 vs 25.4 ± 10.9, $p = 0.002$) and a higher FOG (21.3 ± 1.9 vs 19.5 ± 2.9, $p = 0.037$). For the likelihood-of-use outcome, the panel

assigned higher ratings to AI-generated responses (median [IQR]: 5.0 [4.7-5.2] vs 4.1 [3.8-4.3], Z = -2.49, Wilcoxon signed-rank p = 0.013). Similar results were observed for the perceived patient impact outcome (3.9 [3.8-4.4] vs 3.6 [3.3-3.7], Z = -2.67, p = 0.008). At the physician rater level, larger differences in AI-generated and human ethicist response ratings were associated with fewer ties for both outcomes of likelihood-of-use (Spearman ρ = 0.68, p = 0.03) and perceived patient impact (ρ = 0.94, p < 0.001).

Conclusion: Human ethicist responses were more readable, but physician raters displayed a preference for AI-generated responses when evaluating for both outcomes. In the concordance analysis, physician raters were most decisive evaluating for likelihood-of-use, and the results indicated that ties were due to perceived similarity between AI-generated and human ethicist responses rather than response randomness. The results of this study advocate for potentially integrating AI into the ophthalmologist's toolbox; nonetheless, more research is required in the realm of AI-based ethical reasoning.

Eye Rubbing Assessment Metrics (ERAMs) for Objective Keratoconus Progression Analysis Using AI-based Wearable Detection System

Binh Duong Giap[1], Jefferson Lustre[1], Joshua Ong[1], Anitha Venugopal[2], Nambi Nallasamy[1]

[1]University of Michigan, Ann Arbor, United States; [2]Aravind Eye Hospital, Tirunelveli, India

Purpose: Eye rubbing (ER) is considered a modifiable behavioral factor in keratoconus (KCN) development and progression, yet current clinical assessment relies solely on self-report. This study introduces Eye Rubbing Assessment Metrics (ERAMs), computed by a proposed AI-based wearable system designed to detect eye rubbing behavior from kinematic sensor data and generate standardized exposure metrics suitable for future linkage with KCN progression.

Methods: A three-stage analysis system was developed consisting of 1) wearable sensor data acquisition and preprocessing, 2) AI-based ER detection, and 3) ERAM metric computation. Motion signals with 13 data channels were segmented into 80% overlapping fixed-length windows of 256 samples (100Hz sampling) and normalized prior to inference. ER events were identified using a 1D CNN-LSTM architecture trained on window-level classification. The ERAM computation component quantifies ER behaviors using: 1) ER frequency, 2) duration (mean and cumulative rubbing time), and 3) estimated intensity computed as the mean acceleration magnitude within each detected window. A dataset including 8,640 windows collected from 20 participants was established to train and fine-tune the AI-based ER detection model with 5-fold cross validation on the training set of 6,912 windows from 16 subjects. The proposed system was finally evaluated on the synthetic timelines established using 1,728 ER and Non-ER windows of four subjects in the testing subset.

Results: The ER detection model demonstrated strong performance, achieving an F1-score of 92.54% and an AUC of 96.70% on the testing set. Timeline-based event matching on synthetic sequences of the testing set yielded onset MAE/MedAE of 0.91/0.51 seconds, offset MAE/MedAE of

0.84/0.51 seconds, and duration MAE/MedAE of 1.44/1.02 seconds. Measured window intensity differed between classes (ER: 0.24 +/- 0.11 vs non-ER: 0.33 +/- 0.33, P<0.001), demonstrating statistically meaningful separation between rubbing and non-rubbing behavior.

Conclusions: This work introduces the first integrated AI-based wearable framework capable of objectively detecting and quantifying ER behavior using simplified and clinically interpretable exposure metrics. The proposed ERAM framework establishes a foundation for future studies investigating the role of ER in KCN development and progression.

MAE-SAM2: Mask Autoencoder-Enhanced SAM2 for Clinical Retinal Vascular Leakage Segmentation

Xin Xing[1], Irmak Karaca[2], Amir Akhavanrezaya[3], Samira Badrloo[1], Quan Dong Nguyen[3], Mahadevan Subramaniam[1]

[1]University of Nebraska Omaha, Omaha, United States; [2]Columbia University Irving Medical Center, New York, United States. [3]Stanford University, Stanford, United States

Purpose/Background: Retinal vascular leakage is a key hallmark of inflammation in non-infectious retinal vasculitis and is commonly evaluated using fluorescein angiography (FA). Accurately segmenting leakage regions is critical for disease evaluation and follow-up. However, this study remains challenging due to the small size of lesions, dense and irregular distribution, imaging artifacts, and the limited availability of expert-annotated clinical data. Existing AI models, such as convolutional architecture and transformer architecture methods perform unsatisfied under these conditions.

Methods: To address these challenges, we propose MAE-SAM2, a segmentation framework that combines self-supervised learning with a foundation model for retinal vascular leakage segmentation. Specifically, we enhance MedSAM2 by introducing masked autoencoder (MAE)–based self-supervised pretraining on the SAM2 image encoder, allowing the model to learn robust representations from unlabeled FA images. The pretrained encoder is then fine-tuned end-to-end for leakage segmentation using a task-specific loss function that combines Dice loss and binary cross-entropy to better handle class imbalance and small lesion characteristics. We evaluate our method on a clinically curated FA dataset consisting of 74 images from 38 patients with expert-annotated leakage masks.

Results: Experimental results show that MAE-SAM2 consistently outperforms state-of-the-art baseline models, including U-Net, U-Net++, DeepLabV3+, Swin-UNet, and MedSAM2. Our method achieves the highest Dice score (0.5593) and IoU (0.4348). Ablation studies further demonstrate that both MAE-based pretraining and the combined loss function play important roles in the observed performance gains. Qualitative results indicate improved robustness to overexposure artifacts and fewer false positive predictions.

Conclusion: MAE-SAM2 demonstrates the effectiveness of integrating self-supervised learning with foundation models for clinically challenging medical image segmentation tasks. This work

highlights the potential of self-supervised foundation-model-based approaches to improve segmentation performance in settings with limited annotations and small lesion targets.

A Deep Learning Classifier for Full-Field Electroretinogram Interpretation

Peter Zhao, Binh Giap, Keeley Likosky, Jeff Lustre, Karthik Srinivasan, Naheed Khan, Nambi Nallasamy

University of Michigan, Ann Arbor, United States

Purpose: Full-field electroretinography (ERG) has important clinical applications, including diagnosing inherited retinal dystrophies (IRDs) and providing functional assessment in unexplained vision loss. ERG interpretation requires specialized knowledge, creating a resource bottleneck that limits accessibility and delays diagnosis. We developed ERGNet, a deep learning classifier to automate ERG interpretation.

Methods: We retrospectively analyzed clinical ERG recordings from 3,208 eyes of 1,622 patients at a tertiary academic center obtained between 2016 and 2023. Each patient-eye had a standard six-ERG recording sequence (0.01DA, 3.0DA, 3.0OP, 10.0DA, 3.0LA, and 30 Hz Flicker) performed according to International Society for Clinical Electrophysiology of Vision guidelines. An expert electrophysiologist classified each patient-eye as normal (NM), rod-cone dysfunction (RCD), cone-rod dysfunction (CRD), cone dysfunction (CD), or advanced retinal degeneration (ARD). The dataset was split at the patient level into training (60%), validation (20%), and test (20%) sets. Training was augmented using random scaling, Gaussian noise, and delayed random walk to replicate common artifacts. The model was trained for 100 epochs using batch size of 16, Adam optimizer, and initial learning rate of 0.00001. 95% confidence intervals were calculated with bootstrapping.

Results: ERGNet, a 1D residual network, achieved weighted F1 score of 86.9% (84.1-89.4%), precision of 87.0% (84.2-89.6%), and recall of 87.3% (84.7-89.8%). AUROC was 0.977 (0.968-0.984) and AUPRC was 0.924 (0.901-0.944). Inference time was 1.7 milliseconds. ERGNet performed well at classifying NM (F1=93.4%, 91.2-95.3%), ARD (F1=93.0%, 89.8-95.8%), and RCD (F1=78.9%, 69.0-87.1%). Performance for classifying CRD (F1=66.4%, 52.8-78.1%) and CD (F1=70.2%, 62.2-77.4%) was lower, with CRD most frequently misclassified as CD (13/39), and CD most frequently misclassified as NM (28/100) and CRD (4/100). With GradCAM, the NM class attention pattern was focused on the initial 5-50 milliseconds of the ERG, aligning with physiologic a-wave and b-wave timings. In contrast, the ARD and RCD classes had more diffuse attention patterns. The CRD and CD classes contained examples with both NM-like and ARD/RCD-like attention patterns.

Conclusion: Automated ERG classification is possible with strong performance for classifying major diagnostic categories. Misclassifications mirrored known clinical phenotypic overlap. Prospective validation on an external dataset is needed to generalize classifier performance for clinical use.

Automated Quantification of Decreased FAF in Stargardt Disease Using Starguage: Validation Against Manual Grading Standards

Mauro Campigotto[1], Mohamed I. Ahmed[1], Hikmet Yucel[1], Rubbia Afridi[1], Thales A.C. de Guimarães[1,2], Isabel Sendino-Tenorio[1,3], Nam V. Nguyen[1], Romaisa Kiran[1], Sidrah Khan[1], Ufaq Khan[1], Syeda Sharaf un Nisa[1], Amir Hariri[1], Michel Michaelides[4,5], Hendrik P.N. Scholl[6,7,8,9], Nathan Mata[6], Quan D. Nguyen[10], Yasir J. Sepah[1]

[1]OIRRC, Sunnyvale, United States; [2]Department of Ophthalmology, Faculdade São Leopoldo Mandic, Campinas, Brazil; [3]Department of Ophthalmology, University Hospital of Leon, Leon, Spain; [4]UCL Institute of Ophthalmology, University College London, London, United Kingdom; [5]Moorfields Eye Hospital NHS Foundation Trust, London, United Kingdom; [6]Belite Bio Inc, San Diego, United States; [7]Medical University of Vienna, Department of Clinical Pharmacology, Vienna, Austria; [8]Pallas Kliniken AG, Pallas Klinik Zürich, Zurich, Switzerland; [9]European Vision Institute, Basel, Switzerland; [10]Ophthalmology, Stanford University School of Medicine, Palo Alto, United States

Purpose: To evaluate the repeatability and reproducibility of *Starguage*, a novel automated method compared with manual segmentation for measuring decreased autofluorescence (DAF) and definitely decreased autofluorescence (DDAF) in fundus autofluorescence (FAF) images of patients with Stargardt disease.

Methods: DAF and DDAF lesion areas were independently quantified by five certified graders using either manual delineation with Heidelberg RegionFinder or a threshold-based automated algorithm, with automated quantification and cross-method agreement analyses restricted to a prespecified central 6-mm fovea-centered region. Agreement and repeatability were assessed using intraclass correlation coefficients (ICC), standard error of measurement (SEM), minimal detectable change (MDC), Lin's concordance correlation coefficient (CCC), Bland–Altman plots, and Passing–Bablok regression. Both raw and square-root-transformed lesion areas were evaluated.

Results: The automated method achieved excellent intra-grader repeatability for both DAF and DDAF (ICCs ≥0.988, SEM ≤0.71 mm^2, MDC ≤1.98 mm^2), with minimal operator influence. Manual measurements showed variable repeatability (DAF ICCs 0.909–0.974; DDAF ICCs as low as 0.837), with square-root transformation reducing SEM and MDC. Inter-grader reproducibility was highest for automated methods (ICC = 0.989–0.992), whereas manual methods ranged from 0.764–0.939 (raw) and 0.867–0.922 (transformed). Cross-method agreement was strong (CCC = 0.91–0.96), though minor proportional and constant bias was observed in raw DAF data.

Conclusions: The automated approach provides near-perfect repeatability and high agreement with manual grading, offering a scalable, objective alternative for quantifying hypo-autofluorescent lesions in Stargardt disease. Manual methods are generally reliable but more variable, especially for DDAF, and benefit from square-root transformation. Findings reflect a pediatric/adolescent single-trial predominantly Asian cohort.

Implementation of Artificial Intelligence for RPE Loss Annotation in OCT Imaging

Robert Slater[1], Reeva Faisal[1], Rushi Mankad[1], Rachel Linderman[1], Jeong Pak[1], Roomasa Channa[1,2], Barbara Blodi[1,2], Amitha Domalpally[1,2]

[1]Wisconsin Reading Center, Madison, United States; [2]University of Wisconsin - Madison, Madison, United States

Purpose: Measurement of RPE loss on OCT provides a practical approach for assessing area of geographic atrophy. The purpose of this study is to develop, validate and deploy an Artificial Intelligence (AI) model that can annotate Retina Pigment Epithelium (RPE) loss.

Methods: The AI model was trained on 395 OCT volume scans from the AREDS 10-year study annotated at the WRC on internal and independent test sets The model was deployed as an application ("dockerized") in our Imaging Database that allowed it to be used by non-AI personnel, allowing a grader to run the model by selecting it through a menu drive interface. The application generated a human readable report that summarized the AI prediction and annotations. A subset of test cases was further assessed in the deployment platform using a qualitative grading framework that categorized AI outputs based on the extent of human edits required.

Results: The average Dice Score was 0.92 on the internal test set for the en-face area of RPE loss. An independent testing dataset found a Dice Score of 0.72 and a mean difference of 0.3 mm^2 difference in measured area between human and AI annotation. Qualitative assessments of 399 cases from the independent test set showed that 376 were found to have at least adequate agreement, requiring limited human corrections. Deployment of the AI workflow reduced annotation time from approximately 1–2 hours per case to 5–10 minutes, without compromising grading quality.

Conclusion: This study demonstrates feasibility of deploying AI based annotation models into routine reading center workflows as supportive tool for graders. By substantially reducing annotation time and requiring only basic operational skill from graders, the system enables significant throughput gains while preserving human oversight and annotation quality.

AI-based algorithm for post-surgical IOL positioning evaluation: validation against human approach

Mauro Campigotto, Mohamed I. Ahmed, Hikmet Yucel, Yasir J. Sepah

Ocular Imaging Research and Reading Center (OIRRC), Sunnyvale, CA, United States

Purpose: To evaluate the repeatability and reproducibility of a novel AI-based algorithm for post-surgical IOL positioning evaluation compared with a human approach by measuring IOL decentration in degrees versus Pupil and Cornea center, including its direction.

Methods: IOL, Pupil, and Cornea contours were evaluated by using digital slit lamp images. A set of 30 images was used with a resolution of 5 megapixels (MP). Each image shows the entire anterior

view of the eye, including at least one of the two canthi. Two certified graders use a dedicated software tool integrated in EyeSol, an EMR with viewing capabilities developed by OIRRC. The tool permits the placement of at least 10 points on each element's contour (IOL, Pupil, Cornea). The points were placed only on the visible portion, avoiding extrapolation. Based on their coordinates, the manual tool extrapolated the contour to determine the center of each element and the IOL's orientation. The same analysis was performed by the AI-based algorithm, using properly trained machine learning classifiers for each element, and the same kind of information was extracted as output.

Results: The results from each certified grader were compared with those from the AI-based algorithm. For each image, the average analysis time for the graders was 343 ± 141 seconds, while the AI-based algorithm took 0.56 ± 0.08 seconds. The IOL direction results were comparable. The major difference was in the extrapolation of each element's centres. The AI-based extraction of each element's contour enabled better alignment with their true edges, leading to more precise center calculations.

Conclusion: The AI-based algorithm provided a near-perfect evaluation of each element's contour, surpassing the limited precision of each grader. This precision will strongly affect results when comparing baseline and follow-up analyses over time within the same eye. The significantly reduced processing time of the AI-based algorithm compared with the human-based software tool is a strong advantage when dealing with an extensive set of images. More variability was observed in case of white-saturated or too dark images, especially in the IOL-visible portion. Proper image selection could be considered a limitation of the AI-based algorithm compared to a manual approach.

Identification of IRMA on Color Images

Reeva Faisal[1], Mozhdeh Bahrainian[2], Robert Slater[1], Amitha Domalpally[1,3], Roomasa Channa[1]

[1]A-EYE Research Unit, Department of Ophthalmology and Visual Sciences, University of Wisconsin-Madison, Madison, WI, United States, Madison, United States; [2]Department of Ophthalmology, University of Wisconsin-Madison, Madison, WI, United States, Madison, United States; [3]Wisconsin Reading Center, Department of Ophthalmology and Visual Sciences, University of Wisconsin-Madison, Madison, WI, United States, Madison, United States

Purpose: Intraretinal microvascular abnormalities (IRMA) are an important feature of classifying severity of non-proliferative diabetic retinopathy but are difficult to identify on routine color fundus photographs (CFP) without specialized training. This study evaluates the feasibility of using expert-annotated CFP data to train an AI model to identify the presence of IRMA.

Methods: A total of 477 macular field CFP from DRCR dataset were annotated for IRMA by trained graders from the Wisconsin Reading Center. Of these, 401 were initially allocated for training and 76 for testing. After excluding images with significant quality issues such as focus and sharpness, 356 images remained in the training dataset. A combination of CLAHE, Gamma, and Unsharpening

filters was applied, along with a cropping mechanism to maintain consistent backgrounds amongst them as part of the preprocessing. A SWIN-B Transformer model was trained and validated using Stratified-Kfold cross validation to classify images as IRMA present or absent. AI predictions were compared to grader derived ground truth labels and disagreements were evaluated.

Results: Across the 5 folds, the performance metrics are reported as mean ± standard deviation. The AI model achieved an accuracy of 75.56% ± 4.79%, precision 56.91% ± 10.49%, recall 49.42% ± 7.25%, and F1 score 52.53% ± 7.28%. A review of the false positives (10.67%) and false negatives (13.76%) revealed that non-uniform illumination of the image often caused the model to inaccurately classify it as IRMA or shifted its attention away from the actual presence of IRMA entirely. Other factors included image quality limitations or the presence of only very small, dot-level IRMA lesions. Model attention maps demonstrated predominant focus within the macular region and surrounding areas where IRMA was present, though precise lesion-level co-localization was not achieved.

Conclusion: An AI classifier trained on expert-annotated CFP achieved moderate performance for detecting the presence of IRMA. Future efforts include further refining the model to produce better cross-validation results before validating model performance on the hold-out test set.

Automated capillary-level optical coherence tomography angiography phenotyping in patients with glaucoma and diabetes

Yash Lahoti[1], Giovanna Guidoboni[2], Jason Greenfield[1], Gal Jacob Cohen[1], Samuel Potash[1], Brent Siesky[1], Alice Verticchio Vercellin[1], Keren Wood[1], Minwoo Kwon[1], Alon Harris[1]

[1]Icahn School of Medicine at Mount Sinai, New York, United States; [2]Maine College of Engineering and Computing, University of Maine, Orono, United States

Purpose: To evaluate whether a synthetic-data-trained optical coherence tomography angiography (OCTA) segmentation pipeline could capture disease-specific microvascular signatures across glaucoma and diabetes in the Artificial Intelligence Ready and Equitable Atlas for Diabetes Insights (AI-READI) dataset.

Methods: We used previously developed synthetic vascular network simulation framework to create a high-resolution dataset. The trained model processed 2,081 macular 6x6 mm OCTA scans from the AI-READI dataset, generating capillary-level binary segmentations without manual labeling. The cohort included patients with diabetes (n=797), glaucoma (n=79), and comorbid glaucoma plus diabetes (G+D; n=122), and healthy controls (n=1,083). Several global and sectoral features were engineered, including vessel density (VD), vessel length density (VLD), VD/VLD ratios, foveal avascular zone (FAZ) metrics, and vascular morphology measures (fragmentation). Linear mixed-effects (LME) models assessed diagnosis effects with age as a covariate and patient ID as a random effect to account for inter-eye correlation.

Results: Glaucoma was primarily characterized by structural vessel dropout via significant reductions in Global VD (β= -0.012, p=0.002) compared to controls while diabetics showed no reduction in Global VD (p=0.061). Conversely, diabetes manifested distinct markers of network discontinuity and foveal ischemia. VLD was significantly reduced in diabetes (β = -0.204, p<0.001) and FAZ metrics revealed enlarged area (β= 0.074, p<0.001) and reduced circularity (β= -0.016, p=0.002). These foveal remodeling changes were absent in the glaucoma-only group. Morphological analysis revealed a stepwise increase in vascular fragmentation from healthy to diabetes (β= 0.134, p=0.013), glaucoma (β= 0.379, p=0.005), and G+D (β= 0.596, p<0.001). Notably, the combined G+D group showed a supra-additive synergistic effect in VLD reduction (β= -0.790) which exceeded the sum of individual diabetes and glaucoma effects.

Conclusion: Automated capillary phenotyping trained on synthetic data reveals distinct vascular signatures between glaucoma and diabetes. The comorbidity of glaucoma and diabetes may exacerbate microvascular damage beyond either condition alone.

Comparing Pointwise Versus Garway-Heath Neural Network Performance in Humphrey Visual Field Predictions from Previous and Current Quantitative Spectral-Domain OCT

Karim Dirani, Justin Bennie, Ossama Mahmoud, Daniel Blessing, Victor Tawansy, Mark Juzych

Kresge Eye Institute/Wayne State University School of Medicine, Detroit, United States

Purpose: Quantitative spectral-domain (SD) OCT and 24-2 Humphrey visual fields (HVFs) are widely used to assess glaucoma and have been extensively studied for predicting overall visual field status and mean deviation (MD). However, pointwise HVF prediction is often limited by high test–retest variability, and MD alone may not capture the spatial patterns that characterize glaucomatous progression. Garway–Heath sectorization may reduce pointwise variance while preserving region-specific progression metrics.

Methods: We analyzed 11,952 paired observations consisting of prior SD OCT/24-2 HVF data and subsequent SD OCT/24-2 HVF data. A neural network model was trained on 10,142 samples and evaluated on 1,810 randomly selected test samples. The dataset included 4,681 unique patients and encompassed 25 years of SD OCT and HVF testing from a multicenter academic institution in the Midwest serving a predominantly urban population. Exclusion criteria included SD OCT signal strength <6 and unreliable HVFs based on standard reliability parameters.

Results: The pointwise model achieved a mean test R^2 of 0.4254 and a mean test mean absolute error (MAE) of 3.06. The Garway–Heath model achieved a mean test MAE of 2.09, with the central sector demonstrating the highest accuracy (MAE 1.66). In the pointwise model, the two least accurate locations were statistical outliers and corresponded to physiologic blind-spot regions.

Conclusion: Garway–Heath sector-based modeling may provide a practical and more stable alternative to pointwise prediction for forecasting visual field change, while retaining clinically meaningful spatial information. Because visual field testing remains the gold standard for

functional assessment in glaucoma, accurate projection of future fields could support earlier identification of rapid progression and more timely intervention. Further work is needed to evaluate external validity and generalizability across diverse populations and imaging platforms.

Accelerating Pediatric Glaucoma Specialist Evaluation Using PATH-PCG™: A Paired Analysis of Clinical Pathways

Christine L. Xu[1], David N. Xu[1], Ngoc Trong Tuong Than[1], Amir Akhavanrezayat[1], Gunay Uludag Kirimli[1], Jia-Horung Hung[1], Christine Xu[2], Joyce Kang[2], Andrew DesLauriers[1], Jason Xiao[1], Aniket Ramshekar[1], Luke Leidy[1], James D. Brandt[2], Darius Moshfeghi[1], Ann Shue[1], Quan Dong Nguyen[1]

[1]Stanford, Palo Alto, United States; [2]UC Davis, Sacramento, United States

Purpose/Background: Delays in access to pediatric glaucoma specialists remain common and may adversely affect outcomes in suspected pediatric congenital glaucoma (PCG). We evaluated whether application of a structured triage platform, the Pediatric Assessment and Triage Hub for PCG (PATH-PCG™), accelerates specialist evaluation compared with observed clinical care.

Methods: We performed a retrospective paired, within-case analysis comparing observed time from documented symptom onset to pediatric glaucoma specialist evaluation with a PATH-PCG™–guided triage timeline. PATH-PCG™ is a symptom-first platform designed for non-specialist use that incorporates structured red-flag detection, tiered urgency recommendations with explicit timeframes, and embedded educational guidance, including instructions to differentiate concerning findings from similar non-pathologic presentations. Triage inputs emphasize hallmark features of PCG, including components of the classic triad (epiphora, photophobia, and blepharospasm), as well as inter-eye asymmetry in ocular appearance or measurements, which is weighted more heavily than single absolute findings. The application provides rapid access to an integrated online photo library of representative clinical findings and includes AI-assisted guidance that suggests the most informative additional findings or tests to consider, supporting structured data entry and triage without providing diagnoses. Eligible cases required documented symptom onset, pediatric glaucoma specialist evaluation date, and assignable PATH-PCG™ parameters. The primary endpoint was time (days) to specialist evaluation. PATH-PCG™ timelines were constrained to be no slower than observed care. Paired differences were assessed using Wilcoxon signed-rank tests.

Results: Nineteen cases met inclusion criteria. Median observed time to pediatric glaucoma specialist evaluation was 30 days (IQR 14–85), compared with 15 days (IQR 0–41) using PATH-PCG™. The median within-case reduction in time to specialist evaluation was 3 days (IQR 3–12), with a 95% bootstrap confidence interval of 3–11 days and was statistically significant on paired non-parametric testing (Wilcoxon $p = 0.000158$). Time-to-evaluation analysis demonstrated earlier specialist access under PATH-PCG™ compared with observed care. Delay-category analysis showed redistribution from prolonged delays into earlier evaluation intervals with PATH-PCG™

guidance. PATH-PCG™ accelerated evaluation in 18 of 19 cases (94.7%), with no cases experiencing delay.

Conclusion: By enabling symptom-first triage with risk-stratified urgency and decision-support guidance, PATH-PCG™ addresses system-level delays in PCG care.

Pathology-Aware Latent Recomposition for Mitigating Class Imbalance in Diabetic Retinopathy Classification: Validation on the EyePACS Dataset

Anthony Dongchau[1], Andrew Dongchau[1], Truc Dinh[1], David Xu[2]

[1]Texas A&M University, College Station, United States; [2]Stanford University, Palo Alto, United States

Purpose: Data scarcity remains a major challenge when applying artificial intelligence to medical imaging. In diabetic retinopathy (DR) classification, advanced disease stages are underrepresented, limiting accurate severity grading. Many existing augmentation and generative approaches of generating synthetic data rely on manual annotation, segmentation, or extensive preprocessing, reducing scalability and clinical applicability. This study introduces Pathology-Aware Latent Recomposition (PLAR), an automated, data-centric framework that mitigates class imbalance due to limited data by generating representative images of pathology without human-provided segmentation or lesion labeling.

Method: An object-centric decomposition model was trained on Mild DR images to decompose retinal fundus images into four feature groups learned directly from raw data, without annotations or manual region selection. A refiner model was then trained to recombine different features from pairs of images and correct boundary inconsistencies introduced, producing anatomic and pathologically coherent synthetic images in which pathology appears across varied retinal contexts. Using this pipeline, 2,000 synthetic images were added to the EyePACS training set. To isolate the effect of image recomposition, a classifier was retrained under identical training conditions using the augmented dataset. Performance was evaluated on a held-out EyePACS test set using accuracy, macro–F1 score, one-vs-rest AUC, and quadratic weighted kappa (QWK) metrics before and after incorporation of synthetic images.

Results: At full image resolution (512×512), PLAR-augmented training improved all clinically relevant metrics. QWK increased from 0.7654 to 0.7938, indicating improved agreement with expert severity grading. Macro–F1 improved from 0.5683 to 0.6029, reflecting better performance on underrepresented classes, while AUC increased from 0.8716 to 0.9052, demonstrating enhanced class separability. The largest per-class improvement occurred in Mild DR, where accuracy increased from 0.2742 to 0.3638. Due to computational constraints, exhaustive significance testing at full resolution was not feasible. However, at reduced image resolution (224×224), paired evaluation demonstrated a statistically significant increase in QWK (0.7273 to 0.7453, $p < 0.05$).

Conclusion: These results demonstrate that PLAR-augmentation can improve clinically aligned DR classification without manual annotation. Reduced performance at lower resolution underscores the importance of high-resolution inputs for detecting small, high-frequency retinal lesions, and suggests broader applicability to other medical imaging domains.

AI for OCT Corneal Map-Based Keratoconus Detection

Yan Li, Jiachi Hong, Afshan Nanji, Richard Stutzman, Xubo Song, David Huang

Oregon Health & Science University, Portland, United States

Purpose: Early detection and accurate diagnosis of keratoconus are needed to provide timely treatment and for preventing post-LASIK ectasia.

Methods: We constructed and optimized a convolutional neural network (CNN) model to combine optical coherence tomography (OCT) corneal topography and thickness maps to differentiate keratoconus of various disease stages (manifest, subclinical, and forme fruste keratoconus) from non-keratoconus eyes. All OCT maps were down-sampled to a size of 16 × 16 pixels to reduce the number of features to be trained on. Each map type was treated as a different color channel in the neural network. A grid search hyperparameter optimization resulted in a streamlined CNN architecture. We included 131 eyes of 78 patients with keratoconus and 148 eyes of 74 volunteers without keratoconus (normal or warpage) in the analyses. Repeated 5-fold cross-validation was used to evaluate the model performance.

Results: The CNN class activation maps illustrated a difference between the keratoconus and non-keratoconus eyes. For the keratoconus eyes, the inferotemporal region of the map was most important. The center of the map was the most important region for the non-keratoconus eyes. The precision was 98 ± 3% for the AI model, and the recall was 91 ± 4%. The area under the receiver operating characteristic curve (AUC) was excellent (93 ± 2%). The AI model was able to detect all the manifest and subclinical keratoconus cases and 56% of the forme fruste keratoconus cases. The classification accuracy was above 96% for the non-keratoconus (normal or warpage) cases using the AI model.

Conclusion: The AI model demonstrated excellent accuracy differentiating keratoconus from non-keratoconus eyes.

DR.GRPO: Diabetic Retinopathy Grading through Group Relative Policy Optimization

Hae-Won Hwang[1], Igor Kozak[2], Eungjoo Lee[2]

[1]Inha University, Incheon, Republic of Korea; [2]University of Arizona, Tucson, United States

Diabetic retinopathy (DR) is a leading cause of preventable blindness worldwide, requiring accurate grading for clinical management. While deep learning models have achieved high accuracy in DR classification, their lack of interpretability limits clinical adoption. Clinicians

require not only accurate predictions but also transparent reasoning that aligns with established diagnostic criteria such as the International Clinical Diabetic Retinopathy (ICDR) severity scale. We propose a reasoning-enhanced Vision-Language Model (VLM) for DR grading that generates explicit chain-of-thought (CoT) reasoning alongside predictions. Our approach employs a three-stage training pipeline: (1) Medical domain adaptation through supervised fine-tuning on Messidor-2, (2) Reasoning capability injection via rule-based CoT annotations leveraging IDRiD lesion segmentation masks, and (3) Reasoning quality enhancement through Group Relative Policy Optimization (GRPO) with soft reward based on grade proximity. We utilize a compact 2B-parameter model (Qwen3-VL-2B) trained with only 750 CoT-annotated samples for SFT and 3,363 samples for GRPO, without relying on LLM-generated annotations. Our key contributions are threefold. First, we demonstrate that lesion-aware CoT generation using pixel-level segmentation masks significantly improves reasoning quality compared to image-only approaches. Second, we show that GRPO with soft reward combining format correctness and graded accuracy effectively enhances model performance while encouraging diverse reasoning. Third, we introduce a multi-source training strategy combining IDRiD, DDR, and APTOS datasets to improve generalization across different imaging conditions. We evaluate our model on the APTOS 2019 test set (587 images) using accuracy and quadratic weighted kappa (QWK). Our GRPO-enhanced model achieves 33.73% accuracy and 0.424 QWK, compared to 29.47% and 0.391 for CoT-SFT alone, and 17.21% and 0.101 for the base Qwen3-VL-2B. Ablation studies show GRPO improves accuracy by 4.26% and QWK by 0.033 over CoT-SFT alone. The QWK improvement represents a shift from “slight” to “moderate agreement,” indicating that our model avoids severe misclassification (e.g., confusing Grade 0 with Grade 4), which can be clinically more critical than raw accuracy. Our work demonstrates that combining lesion-guided CoT supervision with GRPO reinforcement learning enables compact VLMs to perform interpretable DR grading with improved reasoning quality, providing a promising direction for trustworthy medical AI systems. Code is available at https://github.com/jhsea99/DR.GRPO

From Fundus to Diagnosis: End-to-End Binocular-Monocular Framework for Ocular Disease Classification with Evidence-Based Reasoning

Jekyung Lee[1], Igor Kozak[2], Eungjoo Lee[2]

[1]Kyungpook National University, Daegu, Republic of Korea; [2]University of Arizona, Tucson, United States

Ocular diseases are a leading cause of preventable blindness worldwide, and early detection with accurate diagnosis is critical to preserving vision. In clinical practice, fundus examination involves capturing separate images of the left and right eyes, describing per-eye findings, and synthesizing them into a final diagnosis. In the ODIR-5K dataset, per-eye findings are expressed as 94 distinct compound clinical descriptions combining severity, stage, location, etiology, and disease name (e.g., “moderate non proliferative retinopathy”, “dry age-related macular degeneration”), which must then be consolidated into eight diagnostic categories (Normal, Diabetic Retinopathy,

Glaucoma, Cataract, AMD, Hypertension, Myopia, Other). Notably, 42.2% of patients exhibit asymmetric findings between the left and right eyes, making accurate diagnosis difficult from a single eye alone. Despite this complexity, most existing automated approaches are limited to single-image classification without integrating binocular context or providing the interpretable reasoning that clinicians require. We propose an end-to-end framework that infers clinical findings from binocular fundus images and performs multi-label disease classification based on that reasoning. A pretrained retinal-specialized vision encoder extracts and fuses features from left and right eye images, while a learnable eye position embedding enables a single model to handle both binocular and monocular inputs. The extracted visual features and disease classification probabilities are projected into the input space of a large language model (LLM) through dedicated projectors, and the parameter-efficient fine-tuned LLM generates laterality-specific findings, diagnostic conclusions, and clinical recommendations. The entire pipeline is jointly optimized end-to-end through combined classification and language modeling losses. On a held-out test set of 304 patients, the model achieves macro-AUC/F1 of 97.0% / 75.4% (binocular) and 91.9% / 57.1% (monocular) for classification. The quality of the inferred diagnostic reasoning is evaluated at BLEU-4 of 0.718, ROUGE-L of 0.802, and METEOR of 0.796. By performing interpretable classification grounded in clinical finding inference, mirroring how clinicians reason through diagnosis, this work presents a practical direction for AI-assisted ophthalmologic diagnosis.

Efficient MoE-Enhanced Vision Transformer with Adaptive Token Sampling for Cross-Scanner OCT Classification

Hansol Ko, Igor Kozak, Eungjoo Lee

University of Arizona, Tucson, United States

Purpose/Background: Optical coherence tomography (OCT) classifiers can degrade under scanner/domain shift, limiting clinical deployment. Such distributional discrepancies arise from differences in hardware optics and acquisition protocols, which can severely compromise diagnostic reliability across devices. While domain adaptation methods exist, they often require target-domain data or impose computational overhead unsuitable for resource-constrained settings. We propose an efficiency-aware Vision Transformer (ViT) that emphasizes cross-scanner generalization while reducing inference burden.

Methods: A ViT-Base (patch16/224) encoder was initialized from Masked Autoencoder (MAE) self-supervised pretraining and fine-tuned on the Optical Coherence Tomography Dataset for Image-Based Deep Learning Methods (OCTDL) for three-class diagnosis (Normal, AMD, DME). For accuracy gains without widening the backbone, we replaced the Feed-Forward Network (FFN) in the last six transformer blocks with a dense Mixture-of-Experts (MoE) FFN with E=4 experts, where a learned gate forms a convex combination of expert FFN outputs. To mitigate MoE overhead, we applied Adaptive Token Sampling (ATS) at inference, which ranks patch tokens by their cosine similarity to the CLS embedding and retains a specified fraction of patch tokens (minimum 49) before the transformer encoder.

Results: We evaluated in-domain on OCTDL-TEST (n=361) and cross-scanner on DUKE (n=3,231). Dense MoE improved macro-AUC over a non-MoE MAE–ViT baseline (DUKE: 0.966→0.974; OCTDL-TEST: 0.996→0.999). It achieved 0.882 accuracy, 0.884 macro-F1, 0.974 macro-AUC on DUKE and 0.986 accuracy, 0.986 macro-F1, 0.999 macro-AUC on OCTDL-TEST. With retaining 50% of patch tokens, ATS largely maintained classification performance (DUKE macro-AUC 0.971, macro-F1 0.883; OCTDL-TEST macro-AUC 0.999, macro-F1 0.959) while reducing peak memory from 1.58 GB to 1.31 GB (−17%) with comparable latency (~6.5 ms; batch=1, AMP). This enables the deployment on resource-constrained clinical hardware.

Conclusion: A dense MoE-FFN ViT paired with ATS achieves cross-scanner performance (DUKE macro-AUC≈0.97) while reducing inference memory. This supports practical OCT deployment under scanner shift and resource constraints. Future work will benchmark against domain adaptation methods (e.g., domain adversarial training), evaluate throughput at batch>1, and explore quantization-aware deployment (FP16/INT8) for further efficiency gains.

From Risk to Progression: Genetic Insights into Atrophy Progression in AMD Using AI-Driven OCT Annotations

Shlomit Jaskoll[1], Adi Kramer[1], Liran Tiosano[1], Jaime Levy[1], Sandro De Zanet[2], Stefanos Apostolopoulos[2], Carlos Ciller[2], Itay Chowers[1]

[1]Hadassah Medical Center, Jerusalem, Israel; [2]RetinAI Ikerian AG, Bern, Switzerland

Purpose: AI-based automated OCT segmentation has opened new avenues for studying retinal degeneration dynamics at scale, enabling reproducible, high-throughput phenotyping across large patient cohorts. Age-related macular degeneration (AMD) is the leading cause of irreversible visual loss in older adults. While genome-wide association studies (GWAS) have identified numerous susceptibility loci, their influence on macular atrophy (MA) progression remains unclear. We examined whether AMD-associated genetic variants modulate progression rates in macular atrophy.

Methods: We analyzed 632 AMD patients (1,131 eyes) with MA using optical coherence tomography (OCT). Automated AI-based segmentation by RetinAI (Ikerian AG, Bern, Switzerland) quantified photoreceptor (PR) and retinal pigment epithelium (RPE) atrophy. Linear mixed-effects models (LMMs) were employed and tested 113 variants as the primary fixed effect of interest.

Results: Mean age was 87.4 ± 9.2 years, and 59.7% were female. Variants in RDH5/CD63 (β = –0.03 mm/year, 95% CI –0.05 to –0.005, p=0.01) and TGFBR1 (β = –0.02 mm/year, 95% CI –0.05 to –0.002, p=0.03) were associated with slower PR layer atrophy progression, with similar trends observed for the RPE layer. In contrast, variants in CETP, PLA2G12A, COL10A1, ARHGAP21 and VARS loci, involved in lipid metabolism, collagen regulation, visual cycle and protein synthesis, were associated with faster RPE layer atrophy progression (β range: 0.04-0.09 mm/year; p range: 0.001-0.02). A known association at the ARMS2 locus (β = 0.04 mm/year, 95% CI 0.02–0.06, p=0.001), one of the strongest risk alleles for AMD onset, was replicated. Other variants showed no significant

effect on progression, supporting a distinct genetic architecture underlying atrophy dynamics compared to disease susceptibility.

Conclusions: AI-based OCT segmentation enabled this large-scale genotype-phenotype analysis, revealing specific AMD risk loci influence the rate of retinal degeneration beyond initial disease susceptibility. Identifying these progression-associated variants provides a genetic framework for patient stratification and therapeutic targeting, bridging molecular risk to clinical trajectory in AMD.

Artificial Intelligence Augmentation of the Physician-Scientist in Visual Field Analysis

Yujia Zhou, Alvaro Mejia-Vergara, Paul Clifford

University of Florida, Gainesville, United States

Background: Quantitative analysis of kinetic perimetry is historically limited by the lack software in volumetric modeling. While methods like the Christoforidis volumetric estimation exist, implementing them requires specialized coding skills typically outside the scope of physician training. This project evaluates the feasibility of utilizing Large Language Model agents to bridge this technical gap, enabling a physician with minimal coding experience to develop a browser-based, standalone application for sophisticated visual field analysis.

Methods: The objective was to build simple client-side application capable of digitizing Goldmann visual field images, performing distortion correction, and calculating isopter volumes using the Christoforidis method and steradian-based retinal mapping. The development process utilized an iterative "prompt engineering" workflow, where the physician acted as the software architect and the agent acted as the junior developer. The effectiveness of the AI was evaluated across three domains: statistical logic, interface design, and mathematical implementation.

Results: The agent successfully generated approximately 90% of the functional codebase, excelling in syntax generation and boilerplate logic including file handling and canvas user interface. It significantly accelerated the implementation of complex mathematical formulas, specifically the Jacobian transformations required for mapping 2D Cartesian coordinates to 3D steradian retinal surface area, a significant computational shortcut. However, the agent struggled with domain-specific nuance and deep debugging. Iterative design also created technological debt, a phenomenon seen in human coding teams.

Conclusion: This project demonstrates that LLMs can effectively extend the capabilities of physician-scientists, allowing for the rapid prototyping of bespoke research tools without a dedicated engineering team. However, the role of the physician remains critical: success depends not on coding fluency, but on the ability to deconstruct clinical problems into logical constraints that the AI can execute.

Retinopathy of Prematurity Screening in the Era of Artificial Intelligence: Interpretable Models and Clinical Adoption

Hadas Ben-Eli[1,2], Ayelet Goldstein[2], Edna Granit[2], Smadar Eventov-Friedman[1], Noa Ofek-Shlomai[1], Sinan Abu-Leil[1], Milka Matanis-Suidan[1], Hadas Mechoulam[1]

[1]Hadassah-Hebrew University Medical Center, Jerusalem, Israel; [2]Jerusalem Multidisciplinary College, Jerusalem, Israel

Purpose/Background: Retinopathy of prematurity (ROP) is a major cause of preventable childhood blindness. Current screening guidelines, including those of the American Academy of Ophthalmology (AAO) and the Growth and ROP (G-ROP) criteria, are designed to maximize sensitivity but lead to substantial over-screening. This study assessed whether a locally developed, interpretable machine learning (ML) model could reduce screening burden while maintaining 100% sensitivity for detecting Early Treatment for ROP (ETROP) Type 1 disease.

Methods: We performed a retrospective cohort study of infants admitted to Hadassah Medical Center between January 2023 and February 2025. Of 2,159 infants, 522 met inclusion criteria (gestational age ≤33 weeks or birth weight ≤1500 g), and 193 had complete data for both G-ROP and ML analyses. A logistic regression–based ML model was developed at the infant level using gestational age, birth weight, and early postnatal weight gain. Model performance was evaluated using leave-one-out cross-validation. Model interpretability was examined using SHapley Additive exPlanations (SHAP).

Results: Among the 193 infants analyzed, six developed ETROP Type 1 ROP. Both AAO and G-ROP criteria achieved 100% sensitivity. G-ROP improved specificity to 28% and reduced the screening burden by 26% relative to local practice (97.9% to 72.5%). The ML model also achieved 100% sensitivity while substantially improving specificity to 82.9%. This reduced the number of infants requiring screening from 179 to 38 compared with AAO (−78.8%), from 140 to 38 compared with G-ROP (−72.9%), and from 189 to 38 compared with local practice (−79.9%). SHAP analysis showed that the model integrated established clinical risk factors into a transparent, clinically interpretable risk score.

Conclusion: An interpretable ML model preserved patient safety while markedly reducing unnecessary ROP screenings. These findings support population-specific validation and demonstrate the potential of transparent AI tools to optimize neonatal screening protocols and reduce clinical burden.

AI-Simulated Ophthalmology Residency Interview Preparation: A Mixed Methods Study

Kiran Depala, BS[1], Howard Zhang, BS[1], Eric Brown, MD PhD[2], Reid Longmuir, MD[2], Janice Law, MD[2]

[1]Vanderbilt University School of Medicine, Nashville, TN, United States; [2]Vanderbilt University Medical Center, Department of Ophthalmology, Nashville, TN, United States

Background/Purpose: To evaluate the quality, relevance, and alignment with ophthalmology-specific core competencies of simulated residency interview questions and suggested responses generated by multiple AI chatbots.

Methods: Four AI chatbots - ChatGPT 5.0, Negotiator (custom ChatGPT), Microsoft Copilot, and Google Gemini (2.5 Flash) - were each prompted three times to simulate residency interview questions and suggested responses across four domains: (1) ice-breaker, (2) ophthalmology-specific, (3) career-objective, and (4) behavioral. Three ophthalmologists involved in residency selection independently evaluated outputs using a six-item Likert-scale (1–5), assessing relevance, clarity, practicality, realism, alignment with ophthalmology core competencies, and overall effectiveness. Inter-rater agreement and model consistency were assessed using Krippendorff's alpha and intraclass correlation coefficients (ICC), respectively. Differences in model performance were analyzed using the Kruskal–Wallis test. Free-text comments underwent thematic analysis.

Results: Consistency between iterations was poor across AI models (overall ICC=0.435), with Gemini demonstrating the highest stability (ICC=0.638). Inter-rater agreement was low (overall alpha=0.0576). Model performance scores were similar (Kruskal–Wallis p=0.285, η^2=0.099), though ChatGPT achieved the highest composite score (median 4/5, IQR 1). Thematic analysis identified strengths of AI, including authentic question framing and realistic conversational flow. Limitations included insufficient question depth and occasional misalignment with interview objectives.

Conclusion: Despite variability in evaluator scoring and model output across iterations, AI chatbots received generally favorable qualitative assessments (≥3/5 across all domains). Evaluators viewed these tools as useful for interview preparation, while emphasizing the need for improved model stability and goal-aligned questioning.

Unseen Insights: An AI-Powered Exploration of Secure Patient Messages in Ophthalmology

Isabel Sendino-Tenorio[1], Jiyeong Y Kim[2], Zoha Z Fazal[2], Sophia Y Wang[2], Robert T Chang[2], Eleni Linos[2], Yasir J Sepah[2]

[1]Hospital de Leon, Leon, Spain; [2]Stanford University School of Medicine, Palo Alto, United States

Background: Our study sought to identify the common themes of inquiries and associated demographics of patients approaching with ocular complaints on the electronic health record (EHR) patient message portal using a large-language model (LLM).

Methods: Patient messages reaching the Stanford Department of Ophthalmology between June 2014 and July 2024 were programmatically extracted from the Epic EHR and automatically de-identified. Messages were analyzed at the thread level and topic labels were generated using LLM. Messages were categorized into thematic clusters through keyword matching. A two-proportion z-test was used with $p<0.05$ considered statistically significant to assess the demographic factors associated with message frequency and specific ophthalmic concerns.

Results: Among 4,817 patients messaging ophthalmology, 48.7% were White, 85.7% non-Hispanic, 55.5% women, and 56.9% aged 50 years or older. Of 30,390 messages, appointment schedules, medication prescriptions, ophthalmic procedures, and visual symptoms were the most discussed topics. Based on ICD-10 codes, visual disturbance (26.3%), refractive errors (17.1%), and lens disorders (14.0%) were the most frequent diagnoses. Message frequency by topic significantly differed ($p<0.05$) by patient characteristics in subgroup analysis: non-White patients had more messages on pharmacy refill, glaucoma, insurance, and disability while Whites had more messages on surgery. Non-Hispanics had more messages on vision and cornea. Female patients had more messages on complications and swelling/infection, and unmarried patients sought more advice on vision or disability while married patients had more issues on glaucoma, swelling/infection and cornea.

Conclusion: Our findings suggest opportunities for AI-assisted triage and personalized patient outreach to improve care efficiency and equity. Such data-driven insights can guide workflow optimization and equitable healthcare resource allocation.

ML-Derived Ellipsoid Zone (EZ) and Geographic Atrophy Segmentation Compared to Manual Grading in Non-Exudative Macular Degeneration Eyes

Hanna Coleman[1,2], Jason Slakter[3,2], Daniel Russakoff[2], Jonathan Oakley[2], Vlad Diaconita[1]

[1]Columbia University, New York, United States; [2]Voiant Clinical, Waltham, United States; [3]Vitreous Retina Macula Consultants of New York, New York, United States

Objective/Purpose: To determine if automated, machine learning (ML) assessments of Ellipsoid Zone (EZ) and RPE (Retinal Pigment Epithelium) thickness, volume and loss are accurate and can be deployed at-scale for assessment in clinical trials of non-exudative age-related macular degeneration.

Methods: Heidelberg Spectralis (Heidelberg Engineering Gmbh) OCT images of 84 eyes (42 subjects) with dry macular degeneration, were analyzed by a fully automated ML algorithm [Orion, Voiant Clinical, Inc.] to measure EZ-RPE thickness, volumes and the calculated size of cRORA Geographic Atrophy. The software performs a deep learning-based semantic segmentation of the volume cubes. These values were compared to those measured by two independent retina specialist graders using a semi-automatic method in the same software [Orion, Voiant Clinical, Inc.]. The two graders assessed and manually edited the segmentation of EZ-RPE lines and cRORA on each OCT scan.

Results: There was no statistical difference ($p= 0.32$) between the automated reported average area of cRORA Geographic Atrophy (5.77 mm^2) and the manual measurement (5.85 mm^2). Similarly, although the average EZ-RPE thickness and volume within the ETDRS grid was under-reported by the automated algorithm, this difference was not statistically or clinically significant ($p=0.12$).

Conclusions: OCT is the most clinically relevant imaging modality in the monitoring of treatments for dry AMD. This data shows the ability of the Orion platform to accurately deploy ML-assisted

tools for cRORA and for EZ measurements. This technology can streamline workflows, reduce grader burden and enhance the consistency of structural endpoints in a clinical trial setting.

EyeLecture.com: AI-Assisted Development of Lecture-Based Active Learning Content for BCSC-Mapped Ophthalmology Education

Lindsey Fields[1], Claire Abraham[1], Daniel Wisotsky[1], Rachel Zhang[1], Sally Park[2], Anurag Shrivastava[2]

[1]Albert Einstein College of Medicine, Bronx, United States; [2]Department of Ophthalmology, Montefiore Medical Center, Bronx, United States

Background: Ophthalmology medical education is highly variable across US residency programs and medical schools. This indicates utility in the creation of online-based third party educational resources. EyeLecture.com was developed to provide a centralized ophthalmology lecture platform curated by ophthalmologists and mapped to the American Academy of Ophthalmology's Basic and Clinical Science Course (BCSC) curriculum. Each lecture includes multiple choice questions (MCQs) and flashcards to facilitate active learning. This project describes platform development within a large academic training environment.

Methods: Publicly accessible lectures were collected using predefined search terms mapped to topics in the BCSC curriculum using Google Search and YouTube. Lectures were auto transcribed or underwent GoogleGemini transcription. ChatGPT (OpenAI) was then trained to generate MCQs and flashcards using preexisting content created by board certified ophthalmologists. Prompts were standardized to require that each MCQ and flashcard be grounded in the lecture content and written in an Ophthalmology Knowledge Assessment Program (OKAP)-style format. The questions then underwent an extensive peer-review process in which residents and attendings approved and modified them to best align with the OKAP standards.

Results: An initial pool of 441 publicly accessible lectures was collected (48 glaucoma, 50 cornea, 83 retina, 53 lens/cataract, 60 neuro-ophthalmology, 90 pediatric ophthalmology, 49 oculoplastics, and 8 medical-student lecture series). Forty lectures were included on the platform to date. Using the AI-assisted, human-reviewed workflow, 165 multiple-choice questions and 108 flashcards were created. Learner ratings (1–5 stars) demonstrated a mean of 4.0 and a median of 4.0 stars across lectures.

Conclusion: EyeLecture.com's mission to develop a centralized ophthalmology lecture platform required input from many individuals across learning levels. Thus, the use of OpenAI to develop active learning questions was imperative in ensuring alignment with resident-level knowledge. Future research can analyze the success of using artificial intelligence to develop these questions, such as whether certain instructions had better outcomes and required less modification. Future evaluation will also assess cross-institutional adoption and educational impact to support EyeLecture.com as a shared resource for trainees at multiple institutions.

A Staged EHR Pipeline for AI-Driven Automated Identification and Characterization of Hydroxychloroquine Retinopathy Using OphthoACR (Automated Chart Review) Tool

Karen Chen[1], Ruoqi Yang[2], Reid Weisberg[2], Stanley Chang[2], Elizabeth Park[2], Leejee Suh[2], Vlad Diaconita[2]

[1]New York, New York, United States; [2]Columbia University, New York, United States

Purpose/Background: Hydroxychloroquine (HCQ) retinopathy is a vision-threatening adverse effect that is difficult to identify reliably at scale using structured electronic health record (EHR) data alone. We applied OphthoACR (https://doi.org/10.1167/tvst.14.10.8) our in-house automated chart review (ACR) natural language processing tool to identify patients with HCQ retinopathy and to characterize HCQ exposure and indications within EHR data.

Methods: From an initial cohort of 1,531 patients with documented HCQ exposure since 2020, we first applied a keyword-matching strategy to clinical notes to identify patients with potential HCQ retinopathy, yielding 196 candidates. Billing data was cross-referenced within this subset, identifying 5 patients with specific positive HCQ retinopathy-related billing codes. The full 196-patient cohort was subsequently processed using an automated chart review (OphthoACR) pipeline to extract HCQ retinopathy status and supporting clinical evidence from unstructured notes, as well as HCQ exposure variables including daily dose, duration of use, and estimated cumulative lifetime dose.

Results: Among the 196 patients identified through keyword screening, OphthoACR identified 11 patients as having HCQ retinopathy. Of the retinopathy-positive patients, 10 had systemic lupus erythematosus and 1 had Sjögren syndrome as the underlying indication for HCQ therapy. Human review of the extracted evidence fields supported the presence of HCQ retinopathy in these patients. Patients identified as retinopathy-positive demonstrated higher HCQ exposure compared with retinopathy-negative patients. Mean daily HCQ dose was 366.7 mg in retinopathy-positive patients versus 342.8 mg in retinopathy-negative patients. Mean duration of HCQ use was 95.6 months (7.97 years) in retinopathy-positive patients compared with 76.0 months (6.33 years) in retinopathy-negative patients. Mean estimated cumulative lifetime HCQ dose was 331.2 g in retinopathy-positive patients versus 300.1 g in retinopathy-negative patients.

Conclusion: In a large HCQ-exposed cohort, a staged approach combining keyword screening, billing data review, and automated chart review identified a subset of patients with HCQ retinopathy and enabled efficient extraction of clinically relevant exposure metrics and underlying diagnoses. Retinopathy-positive patients had higher HCQ dose and longer duration of therapy compared with retinopathy-negative patients. This workflow supports scalable descriptive analyses of HCQ toxicity using unstructured EHR data.

Evaluation of Ophthalmology Residents and Large Language Models on Real-World Neuro-Ophthalmology Clinical Cases from the NOVEL Database

Ryan Shean[1], Jenay Yuen[1,2], Rahul Dhodapkar[1,2], Sarah Pike[1,2], Jayanth Mallapu[2], Benjamin Xu[1,2], Melinda Chang[1,2,3]

[1]Keck School of Medicine, University of Southern California, Los Angeles, United States; [2]Roski Eye Institute, University of Southern California, Los Angeles, United States; [3]Division of Ophthalmology, Department of Surgery, Children's Hospital Los Angeles, Los Angeles, United States

Purpose/Background: Prior evaluations of large language models (LLMs) in ophthalmology have largely relied on board-style questions that may not reflect real-world diagnostic reasoning. We evaluated multiple-choice questions (MCQs) derived from neuro-ophthalmology cases in the Neuro-Ophthalmology Virtual Education Library (NOVEL) assessing imaging interpretation, neuroanatomic localization, diagnostic synthesis, and management decision-making, improving clinical realism while reducing bias from familiarity with standardized question banks. We compared performance between LLMs and ophthalmology residents.

Methods: One hundred MCQs were developed by a board-certified neuro-ophthalmologist using imaging and clinical information from real patient cases contained in the NOVEL database, each including a clinical vignette and associated image. Three LLMs (Gemini 3 Pro, Claude Opus 4.5, and GPT-5.2) and three ophthalmology residents (one PGY-2 and two PGY-3) answered all questions. Residents additionally classified whether the image was required to answer each question. LLM responses were generated using low-stochasticity settings (temperature=0.1) to reduce sampling variability across runs. Accuracy was compared using paired McNemar's tests performed at the question level with Benjamini–Hochberg correction. Majority consensus accuracy (≥2/3 agreement) was calculated separately for residents and LLMs.

Results: Overall accuracy ranged from 68.0%-73.0% among residents and 76.0%-87.0% among LLMs. Gemini 3 Pro achieved the highest accuracy (87.0%), outperforming all residents ($p \leq 0.01$ for all) and GPT-5.2 ($p=0.04$). Claude Opus 4.5 (81.0%) and GPT-5.2 (76.0%) performed comparably to residents, with no statistically significant differences after correction ($p \geq 0.11$ for all). Differences between image-required and non-image-required question accuracy were not statistically significant for residents, GPT-5.2, or Gemini 3 Pro ($p \geq 0.18$ for all), whereas Claude Opus 4.5 demonstrated higher accuracy on non-image-required questions (95.2% vs 70.7%; $p=0.03$). LLM majority consensus accuracy (85.0%) was borderline higher than resident majority consensus accuracy (74.0%; $p=0.05$).

Conclusion: When evaluated on clinically grounded MCQs derived from real neuro-ophthalmology cases, LLMs matched or exceeded ophthalmology resident performance, with Gemini 3 Pro demonstrating the highest overall accuracy. Performance differences on image-required questions should be interpreted cautiously, as image-dependence classification is subjective and

may represent intrinsically more complex diagnostic tasks. These findings support further investigation into LLMs as clinical reasoning support tools in ophthalmic education and practice.

Class-Aware Channel Pruning with Resource-Aware Optimization for Efficient Retinal OCT Classification

Minju Hyun[1], Igor Kozak[2], Eungjoo Lee[3,4]

[1]Department of Electrical and Computer Engineering, Inha University, Incheon, Republic of Korea; [2]Department of Ophthalmology and Vision Science, University of Arizona, Tucson, United States; [3]Department of Electrical and Computer Engineering, University of Arizona, Incheon, Republic of Korea; [4]Department of Ophthalmology and Vision Science, Tucson, United States

Optical coherence tomography (OCT) is one of the key imaging modalities for retinal disease diagnosis, where automated identification of pathological conditions such as age-related macular degeneration (AMD) and diabetic macular edema (DME) is essential for clinical decision-making. However, OCT images contain structural redundancy, leading deep convolutional neural networks to employ more channel capacity than necessary for this task and incur unnecessary computational cost. This inefficiency poses a significant barrier to deployment in resource-constrained clinical settings such as mobile devices or low-power edge systems.

To address this challenge, we propose a class-aware channel importance pruning framework applicable to ResNet-18, ResNet-34, and ResNet-50 architectures. The method quantifies each convolutional channel's contribution to inter-class discrimination and selectively removes channels with redundant or low-class relevance. A resource-aware loss function promotes channel sparsity under computational constraints without requiring predefined budgets. Rather than pruning solely based on performance-driven criteria, the proposed formulation explicitly informs the model of computational constraints, guiding the pruning process to jointly preserve classification performance and reduce computation. To mitigate performance degradation, we employ knowledge distillation from the unpruned model (teacher) to the pruned model (student).

Experiments on the Duke retinal OCT dataset (NORMAL, AMD, DME classes) demonstrate substantial efficiency gains while maintaining clinical performance. Under high compression ratios, ResNet-18 reduces parameters by 95% (11.18M→581K) and MACs by 81% (1.82G→341.2M), achieving 92.02% accuracy (F1: 0.926) with per-class precision/sensitivity of 0.99/0.86 (NORMAL), 1.00/0.96 (AMD), and 0.80/0.99 (DME), compared to baseline 96.40% accuracy. ResNet-34 achieves 80% MAC reduction (3.68G→733.8M) and 93% parameter reduction (21.29M→1.59M) with 87.48% accuracy (F1: 0.872). ResNet-50 not only improves efficiency but enhances performance, reducing MACs by 57% (4.13G→1.76G) and parameters by 44% (23.51M→13.22M) while improving accuracy from 85.92% to 94.37% (F1: 0.947), suggesting possible regularization benefits.

Inference-time benchmarking validates the framework's practical efficiency across platforms. On CPU, pruned models achieve 1.50–2.18× latency reductions. On GPU, ResNet-18 and ResNet-34 achieve ~1.04× speedups with modest accuracy drops (4.4 and 2.0 percentage points,

respectively). These results confirm that resource-aware structured pruning yields meaningful efficiency gains in both low-power CPU and GPU-accelerated clinical settings.

High-Order Tensor Decomposition for Fundus Image Enhancement and Retinal Vessel Segmentation

Binh Duong Giap, Nambi Nallasamy

University of Michigan, Ann Arbor, United States

Purpose: Retinal vessel segmentation from color fundus images is essential for the diagnosis of ophthalmic and systemic diseases. However, non-uniform illumination, low contrast, and noise often degrade segmentation performance. This study proposes a tensor-based image enhancement method to improve vessel visibility and segmentation accuracy.

Methods: Each color fundus image is decomposed into RGB channels to form a third-order tensor of size HW3. High-Order Singular Value Decomposition (HOSVD) is applied to obtain three factor matrices (U_1, U_2, U_3) and a core tensor **S**. Image enhancement is performed in the tensor domain by modulating the core tensor using a Gaussian tensor to improve illumination and contrast. High-magnitude coefficients exceeding the mean value of the core tensor are suppressed to reduce noise and artifacts. The enhanced image is reconstructed using the original factor matrices and the compensated core tensor and subsequently used as input to deep learning (DL)-based vessel segmentation networks. The proposed method was evaluated on the STARE, DRIVE, and CHASEDB1 datasets using FPN, LinkNet, and UNet architectures.

Results: Across all three datasets (CHASEDB1, STARE, and DRIVE) and deep learning architectures (FPN, LinkNet, and U-Net), the proposed HOSVD-based enhancement consistently improved retinal vessel segmentation performance. On CHASEDB1, FPN achieved an IoU improvement from 90.18% to 94.01% and an F1-score increase from 94.83% to 96.91%. Notably, U-Net benefited substantially from the enhancement, with IoU increasing from 78.97% to 92.51% and F1-score improving from 88.19% to 96.10%, accompanied by a recall increase from 84.00% to 97.33%. Similar performance gains were observed for LinkNet, demonstrating improved vessel continuity and reduced false negatives across architectures and datasets.

Conclusions: The proposed HOSVD-based tensor enhancement effectively improves fundus image quality and enhances retinal vessel segmentation across multiple datasets and DL architectures. The method is architecture-independent, computationally efficient, and can be readily integrated as a preprocessing step to improve robustness and generalization in vessel segmentation tasks.

Machine Learning Based Estimation of Axial Length in Myopic Eyes from Ocular and Demographic Variables

Neelam Pawar[1], Nambi Nallasamy[2], Giap Binh Duong[2], Meenakshi R[1]

[1]Aravind Eye Hospital, Tirunelveli, India; [2]Kellogg Eye Center, Ann Arbor, United States

Background/Objectives: Axial length (AL) is a key structural biomarker for myopia onset, progression, and the risk of sight-threatening complications; however, direct AL measurement is not universally available, particularly in large-scale screening and resource-limited settings. This study aimed to develop and externally validate machine learning (ML) models to estimate axial length in myopic children and adolescents using readily available demographic, refractive, visual acuity, and corneal parameters, and to evaluate their generalizability across an independent dataset.

Subjects/Methods: ML regression models, including Ridge Regression, Random Forest, Support Vector Machine (SVM), and XGBoost, were developed using age, sex, uncorrected visual acuity (UCVA), cycloplegic spherical equivalent refraction (SER), and average corneal curvature as predictors. Model development and internal testing were performed using data from 5,124 myopic children and adolescents examined at a tertiary eye care center. External validation was conducted on an independent cohort of 600 myopic individuals. Model performance was assessed using mean absolute error (MAE), root mean squared error (RMSE), coefficient of determination (R^2), and evaluation of estimated AL.

Results: On internal testing, XGBoost achieved the highest predictive accuracy (R^2 = 0.83, RMSE = 0.60), while SVM showed the lowest MAE (0.40). All non-linear models outperformed Ridge Regression (R^2 = 0.81). On external validation, model performance declined with R^2 values ranging from approximately 0.50 to 0.55, indicating limited generalizability to unseen populations. Random Forest demonstrated the most stable external performance (R^2 ~ 0.55; MAE~ 0.57), whereas XGBoost showed the greatest reduction in accuracy.

Conclusions: Machine learning models can estimate axial length in pediatric myopia using routinely available clinical and demographic parameters; however, external validation performance was modest, highlighting the impact of dataset heterogeneity and domain shift. These findings emphasize the need for external validation and cautious interpretation before clinical deployment in real-world settings.

COI: Neelam Pawar (Corresponding Author) was a past Research Scholar at Kellogg Eye Centre, University of Michigan, Ann Arbor, MI, USA, supported by an NIH/Fogarty International Centre D43TW012027 grant. Nambi Nallasamy-Funding support NIH K12EY022299, Fogarty/NIH D43TW012027

Comparative accuracy and clinical utility of healthcare-specific versus general-purpose AI models in the management of HLA-B27 associated uveitis

Aubrey Nguyen[1], David Xu[1], Ngoc Than[1], Dalia El Feky[1,2], Cristian de los Santos[1], Jia-Horung Hung[1], Gunay Uludag Kirimli[1], Amir Akhavanrezayat[1], Negin Yavari[1], Gunjan Anil Awatramani[1],

Mohammad Rajabi[1], Osama Elaraby[1], Ishaan Iyer[1], Isaac Sanchez[1], Alan Sherif[1], Jingli Guo[1], Bin Mo[1], Azadeh Mobasserian[1], Anh Tran[1], Victoria Gu[1], Quan Dong Nguyen[1]

[1]Stanford University, Palo Alto, United States; [2]Tanta University, Tanta, Egypt

Purpose: Artificial intelligence (AI)-driven Clinical Decision Support Systems (CDSS) offer expanding clinical utilizations in ophthalmology. Given its prevalence as a uveitic entity, HLA-B27 associated uveitis requires precise and prompt management to minimize visual impairment. This cross-sectional study compares the accuracy and utility of CDSS versus General Purpose AI Systems (GPAIS) in the management of HLA-B27 associated uveitis to determine its broader applicability in clinical practice.

Methods: Redacted case scenarios from 6 patients diagnosed with HLA-B27 associated uveitis at a tertiary uveitis center were included. Data included clinical presentation, history, demographics, and current treatment. Two CDSS (OpenEvidence, Arkangel AI) and three GPAIS (OpenAI GPT-4o, GoogleAI Gemini 2.0, Anthropic AI Claude Sonnet 4) were evaluated. Models were prompted to generate responses about treat/monitor decision, current regimen adjustment, adding new therapies, required testing/imaging, follow-up timing, urgency level, priorities, critical actions, and overall thoughts. These outputs were rated by 3 uveitis-trained fellows on a 1-5 Likert scale, with explanations reported for ratings ≤3. Friedman's test was used to compare platforms on agreement.

Results: Despite no significant inter-platform differences (p=0.13), Arkangel had the highest ratings in 6 categories with the fewest Likert scores ≤3 across all responses, while GPT-4o led in the remaining 3 categories (Figure 1). For overall Likert ratings, Arkangel had higher scores over Gemini 2.0, OpenEvidence, and Claude, and GPT-4o had higher scores over Gemini 2.0, although these differences were not statistically significant. Figure 2 shows the raters' explanations of Likert scores ≤3. Arkangel overestimated disease severity (40%), OpenEvidence had a greater prevalence of nonspecific answers (38%), and GPAIS had more rater disagreement (Gemini 57%, GPT-4o 56%, Claude 54%).

Conclusion: Arkangel and GPT-4o had the highest agreement with raters. Although GPAIS answers were more individualized, raters reported less agreement and greater inaccuracies. CDSS were more conservative and generic, exposing limitations in tailored case responses. Advancement of GPAIS and CDSS can assist in clinical decision-making in the management of HLA-B27 associated uveitis, broadening access to care in underserved populations.

Kugler Publications

Established in 1974, Kugler Publications is a distinguished independent publishing company specializing in scientific medical publications in Ophthalmology, ENT, and related disciplines. With decades of expertise, Kugler has earned a strong reputation for producing high-quality books, journals, proceedings, and other publications in both print and electronic formats.

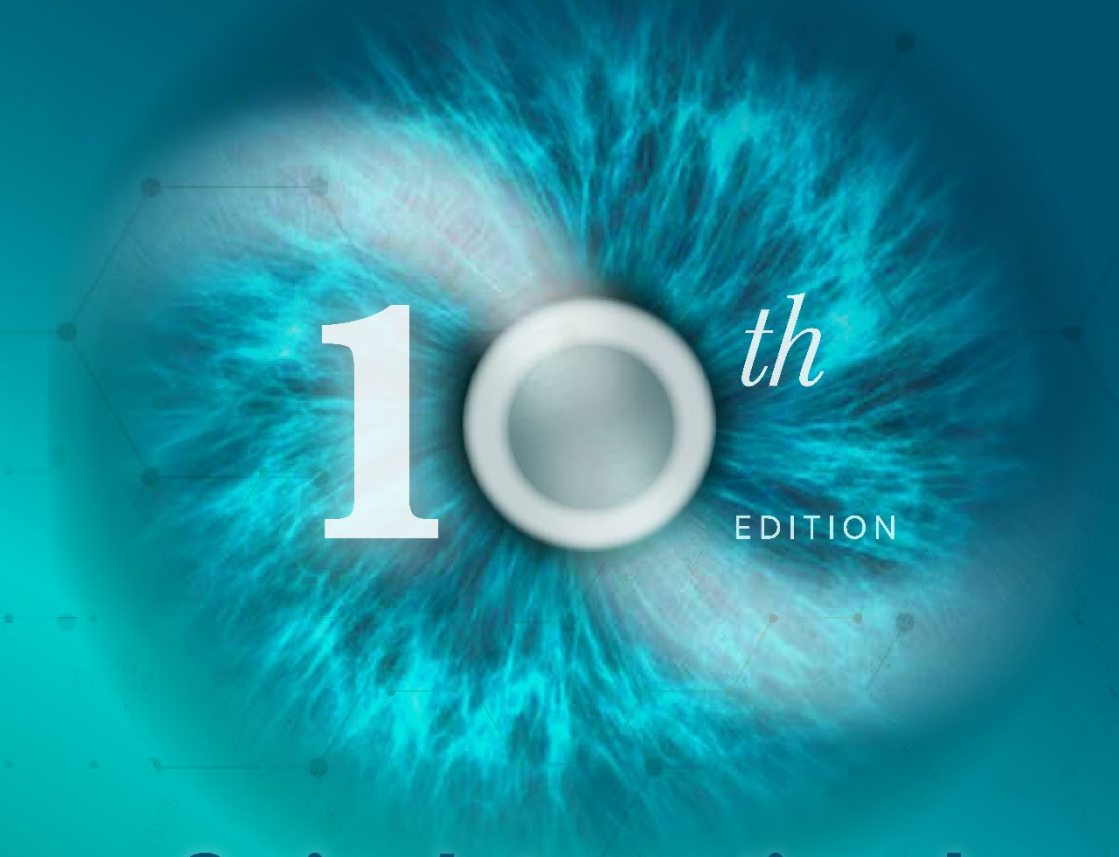
We also organize conferences and build sub-specialty websites, further enriching our commitment to advancing knowledge and innovation in specialized medical fields. With our dedication to excellence, we aim to deliver valuable resources that meet the needs of professionals and researchers worldwide, supporting continual advancements in medical science and practice.

Contact Kugler Publications
Kugler Publications
P.O. Box 20538
1001 NM Amsterdam
The Netherlands

info@kuglerpublications.com
kuglerpublications.com

Publication Highlights

Visit our full catalogue. Receive 20% discount on list prices with coupon EYE25



Swiss International Glaucoma Symposium

JUNE 6, 2026

LAUSANNE

SWISSTECH CONVENTION CENTER
EPFL LAUSANNE

Swiss Glaucoma Research Foundation

SUMMIT FOR INNOVATION IN GLAUCOMA MANAGEMENT IN ASIA

January 22-24, 2027 | Hong Kong, SAR

HKU Med School of Clinical Medicine
Department of Ophthalmology
香港大學眼科學系

KUGLER PUBLICATIONS

Program Chairs

Christopher Leung, MD
Hong Kong, SAR, PR China

Shan Lin, MD
San Francisco, CA, USA

REGISTRATION & ABSTRACT SUBMISSION ARE NOW OPEN

Mission

SIGMA: Collaboration in Glaucoma – Synergy for Better Care

The SIGMA meeting aims to drive advancements in glaucoma care by uniting key stakeholders from across the field—clinicians, researchers, industry leaders, innovators, and investors. Through collaborative discussions, participants will explore the latest developments in glaucoma management and forge new partnerships that push the boundaries of treatment.

GlaucomaSummitAsia.com

Artificial Intelligence in Vision & Ophthalmology

While the rapid advance of imaging technologies in ophthalmology is making available a continually increasing number of data, the interpretation of such data is still very challenging and this hinders the advance in the understanding of ocular diseases and their treatment. Interdisciplinary approaches encompassing ophthalmology, physiology, mathematics, engineering, and computer science have shown great capabilities in data analysis and interpretation for advancing basic and applied clinical sciences. Artificial Intelligence in Vision and Ophthalmology (AIVO) was created with the aim of providing a forum for interdisciplinary approaches integrating mathematical and computational methods with experimental and clinical studies to address open problems in ophthalmology. AIVO welcomes articles that investigate questions related to the anatomy, physiology and function of the eye in health and disease.

SAIVO
Society for AI in Vision and Ophthalmology
Established 2025

Official Journal of the Society for Artificial Intelligence in Vision and Ophthalmology (SAIVO)

www.aivojournal.com
Published by Kugler Publications
www.kuglerpublications.com