Automated Quantification of Neovascular AMD Biomarkers: *Bridging the Gap Between Machine Learning Metrics and Clinical Reality*

East Lancashire Hospitals NHS Trust
A University Teaching Trust

Mahmoud Elbalawi, Yusuf Mubarak, Salwa Abugreen | East Lancashire Hospitals NHS Trust, Blackburn

1 The Problem

Neovascular age-related macular degeneration remains the leading cause of irreversible central vision loss in the developed world. Optical coherence tomography is the cornerstone of disease monitoring, with treatment decisions to inject, extend, or observe depending on the accurate identification of fluid compartments, retinal layer integrity, and structural biomarkers across serial visits.

In current clinical practice, this assessment is performed qualitatively. Clinicians visually interpret each B-scan, mentally integrating information about subretinal fluid, intraretinal fluid, pigment epithelial detachment, layer disruption, and hyperreflective material to form a treatment decision. This process is time-consuming, inherently subjective, and vulnerable to inter-observer variability. Two clinicians reviewing the same scan may reach different conclusions about fluid status and treatment need.

Existing Artificial Intelligence (AI) tools for OCT analysis typically address a single biomarker in isolation, offering either fluid detection or layer segmentation, but never both together. Clinicians are still left to mentally piece together the complete clinical picture.

2 What We Built

One scan in, complete analysis out.

We developed an AI system that performs simultaneous segmentation of retinal layers, fluid compartments, and structural biomarkers from standard OCT B-scans. The system delivers a unified, colour-coded overlay with quantitative measurements in a single pass.

The system handles the full spectrum of AMD presentation, from early dry disease through to active neovascular pathology, providing consistent analysis regardless of disease stage.

OCT B-scan → **AI System** → **Labelled Output**

Biomarkers

- SRF — Subretinal Fluid
- IRF — Intraretinal Fluid
- PED — Pigment Epithelial Detachment
- SHRM — Subretinal Hyperreflective Material
- GA — Geographic Atrophy

Retinal Layers

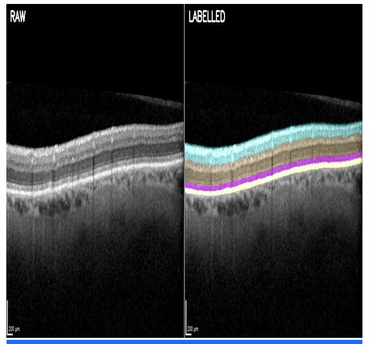
- L1 — ILM – IPL/INL
- L2 — IPL/INL – RPE
- L3 — RPE – BM
- EZ — Ellipsoid Zone

3 Results

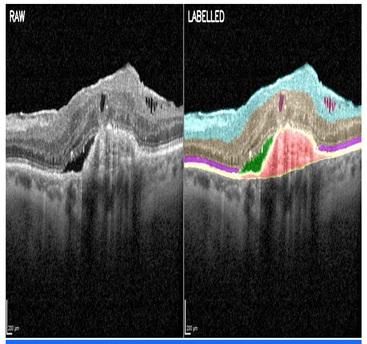
The system was validated against expert-annotated ground truth across multiple independent datasets. These figures reflect the clinically preferred model configuration, selected through clinical validation rather than optimisation for benchmark performance alone. Performance is reported using the Dice coefficient, a standard measure of overlap between AI prediction and expert annotation where 1.0 represents perfect agreement.

0.877
Fluid Segmentation Dice

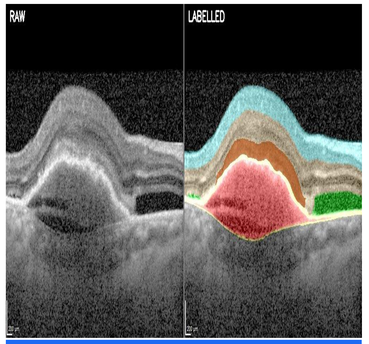
0.972
Layer Segmentation Dice



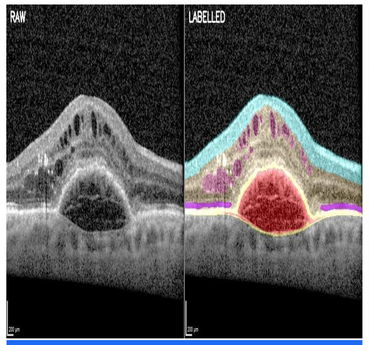
AI labelling all layers of normal retina



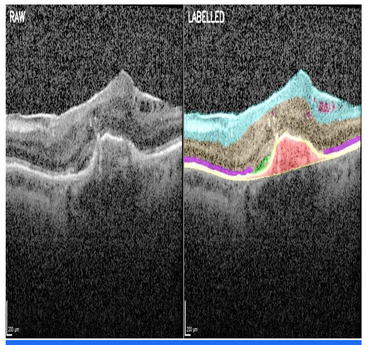
AI labelling mixed pathology (IRF, SRF, PED, and EZ disruptions)



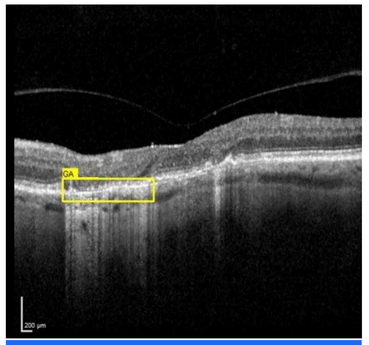
AI labelling SHRM with PED, SRF, and EZ disruptions



AI labelling multiple IRFs with PED and EZ disruptions



AI labelling suboptimal scan quality with IRF, SRF, PED, and EZ disruptions

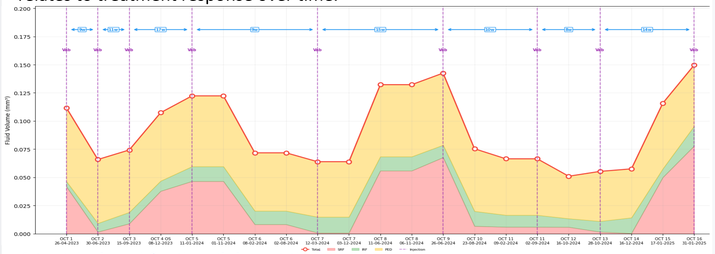


AI labelling geographic atrophy

4 Clinical Value

Treatment interval decisions currently rely on qualitative fluid assessment, which is subjective and vulnerable to inter-observer variability. Using the AI-generated labels, fluid volumes are calculated automatically for each compartment, enabling objective comparison between visits and providing a consistent, observer-independent metric to inform treatment decisions.

Below: fluid volume tracking across sixteen visits for a single patient. Each data point is computed directly from the AI segmentation of a standard OCT B-scan, with injection timing and inter-injection intervals shown to illustrate how fluid behaviour relates to treatment response over time.



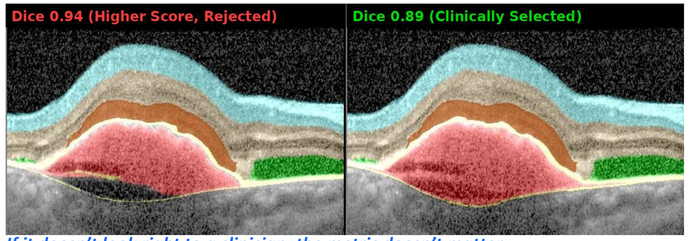
A tool designed for the clinic, not the leaderboard.

5 The Insight

Not all accuracy is equal.

A model achieving Dice 0.94 for pigment epithelial detachment produced segmentations that conflicted with known retinal anatomy, generating boundaries that no clinician would consider plausible.

A lower-scoring model (Dice 0.89) consistently produced anatomically faithful segmentations that aligned with clinical expectations.
The higher-scoring model was rejected in favour of anatomical fidelity.



If it doesn't look right to a clinician, the metric doesn't matter.