GDAŃSK UNIVERSITY OF TECHNOLOGY

TASK Quarterly 30 (3) 2026

Glaucoma Detection Using Intelligent Analysis of Fundus Images: AI Pipeline, Mobile Application, and Low-Cost Fundus Camera Prototype

Martyna Borkowska, Karolina Glaza *, Mateusz Dobry , Agnieszka Pawłowska, Amila Amarasekara, Antoni Naczke

Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, Narutowicza 11/12, 80-233 Gdańsk, Poland

*s198193@student.pg.edu.pl https://doi.org/10.34808/tq2026/30.3/a

Abstract

Glaucoma is the second leading cause of irreversible blindness worldwide, affecting over 80 million people, with projections reaching 111 million by 2040. Early detection is critical, yet up to 50% of patients remain undiagnosed. This paper presents an integrated system for automated glaucoma screening combining a deep learning image analysis pipeline, an iOS mobile application, and a low-cost fundus camera prototype. The AI pipeline follows a two-stage architecture: a YOLOv9 model for region-of-interest (ROI) detection, followed by a UNet++ segmentation model for optic disc and cup delineation. The Cup-to-Disc Ratio (CDR) is computed from the resulting masks and used to classify glaucoma risk. Five YOLO model variants (v8, v9, v11, v12, v26) were evaluated on an augmented dataset of over 6100 fundus images; YOLOv9 achieved the best overall balance with a precision of 98%. UNet++ achieved a mean Dice of 0.92 for joint optic disc and cup segmentation. On a balanced multi-source validation set, the end-to-end system reached 90% specificity but only 55% sensitivity; our analysis shows that this sensitivity ceiling arises from the single area-CDR decision threshold—which limits even perfect physician masks to comparable sensitivity—rather than from segmentation quality, motivating a shift towards multi-feature classification. The iOS application, built in Swift with an MVVM architecture and a FastAPI backend hosted on Microsoft Azure, streams real-time analysis progress to prevent frozen-screen effects. The fundus camera prototype pairs a Volk 20D lens with a 3D-printed, continuously adjustable rail mount on a smartphone, and is designed for retinal image capture at a cost below 1300 PLN; it remains a proof-of-concept that has not yet been used for on-eye image capture. The system targets clinical screening workflows while remaining accessible to individual users.

Keywords:

glaucoma detection, fundus image segmentation, deep learning

1. Introduction

Glaucoma is a progressive optic neuropathy characterized by structural damage to the optic nerve and corresponding visual field loss. It ranks as the second most common cause of irreversible blindness globally, with current estimates of more than 80 million affected individuals and projections of 111 million by 2040 [1]. The disease is frequently referred to as the "silent thief of sight" because it progresses asymptomatically until a significant, irreversible portion of the visual field has been lost. Approximately 50% of patients are unaware of their diagnosis, underscoring the urgent need for efficient, scalable screening tools.

The primary structural biomarker used in clinical glaucoma assessment is the Cup-to-Disc Ratio (CDR), which quantifies the relative size of the optic cup with respect to the optic disc [2]. Two variants are commonly employed: the vertical CDR (vCDR), defined as the ratio of the vertical diameter of the cup to the diameter of the disc, and the area CDR (aCDR), defined as the ratio of the corresponding areas. The aCDR is generally considered more robust for patients with large optic discs. Clinically, a CDR value of ≥ 0.65 raises suspicion of glaucoma, while a value of ≥ 0.70 indicates high risk.

Existing AI-assisted glaucoma screening systems have been criticized for generating excessive false-positive results, leading to unnecessary patient anxiety and downstream healthcare costs [3]. The work described in this paper directly addresses this limitation by targeting a specificity of at least 90% alongside a sensitivity of at least 85%.

The contribution of this work is threefold. First, we present a two-stage deep learning pipeline that combines a YOLOv9 ROI detector with a UNet++ segmentation model to compute CDR with high accuracy. Second, we describe a native iOS mobile application built around this pipeline, providing real-time streaming feedback during AI inference. Third, we document the design and construction of a low-cost fundus camera prototype that enables retinal image acquisition using an off-the-shelf smartphone, bringing the cost of a functional capture device to under 1300 PLN.

2. Related Work

Automated glaucoma assessment from fundus photographs has been an active research area for over a decade, and several recent reviews survey the rapid migration from hand-crafted descriptors to deep learning [4–6]. Early approaches relied on hand-crafted features such as the ISNT rule, which analyzes the relative thickness of the inferior, superior, nasal, and temporal neural rim, combined with classical machine learning classifiers. Such feature-based pipelines remain attractive where interpretability and data efficiency are paramount; MacCormick et al. [7] model the full cup-to-disc profile at 15° intervals and report 91.0% accuracy on external validation while using roughly two orders of magnitude less training data than typical deep models. More recent methods leverage convolutional neural networks (CNNs) and transformer-based architectures. Diaz-Pinto et al. [8] provide an extensive validation of five ImageNet-pretrained CNNs (VGG16, VGG19, InceptionV3, ResNet50, Xception) for whole-image glaucoma classification, establishing strong transfer-learning baselines on public fundus datasets.

The REFUGE Challenge [9] established a unified evaluation framework for optic disc and cup segmentation from fundus photographs, significantly advancing the comparability of proposed methods. Encoder–decoder architectures based on U-Net [10] became the dominant approach for pixel-level segmentation of retinal structures, owing to their ability to preserve fine spatial detail through skip connections.

Object detection networks from the YOLO family [11, 12] have been proposed as lightweight ROI pre-processors that crop the optic disc region before passing the sub-image to a segmentation model. This two-stage strategy concentrates model capacity on the diagnostically relevant region and has been shown to improve CDR accuracy compared to whole-image segmentation. The same detect-then-segment principle underpins the work of Kim et al. [13], who combine a Mask R-CNN ROI detector with a Multiscale Average Pooling Network (MAPNet) for disc and cup delineation, reaching Dice coefficients of 0.968 for the disc and 0.900 for the cup on 1099 fundus images. Our pipeline follows this two-stage paradigm but substitutes a YOLOv9 detector and a UNet++ segmentation backbone, and couples it with an end-to-end mobile capture-to-result workflow.

A parallel line of research targets the acquisition side of the screening problem. Bragança et al. [14] assembled a new dataset of 2000 fundus images captured with a smartphone coupled to a handheld ophthalmoscope and trained an ensemble of CNNs to 90% classification accuracy, demonstrating that affordable mobile imaging combined with deep learning is viable for large-scale screening in resource-limited settings. This finding directly motivates the low-cost smartphone-based capture device described in Section 6.

Several publicly available datasets support this line of research. The G1020 dataset [15] provides 1020 annotated fundus photographs, the RIM-ONE DL dataset [16] supplies standardized disc-centered crops with expert segmentation masks, and REFUGE2, ORIGA, and DRISHTI-DB contribute additional annotated samples from diverse imaging conditions. Most of these datasets carry "educational use only" licensing, motivating a dual-model strategy in which a research model is trained on public data while a potential clinical deployment model will be trained on proprietary data from the partner ophthalmology clinic.

3. System Architecture

The proposed system consists of three interacting subsystems: an AI inference pipeline, a FastAPI backend service, and an iOS mobile application, as illustrated in Fig. 1.

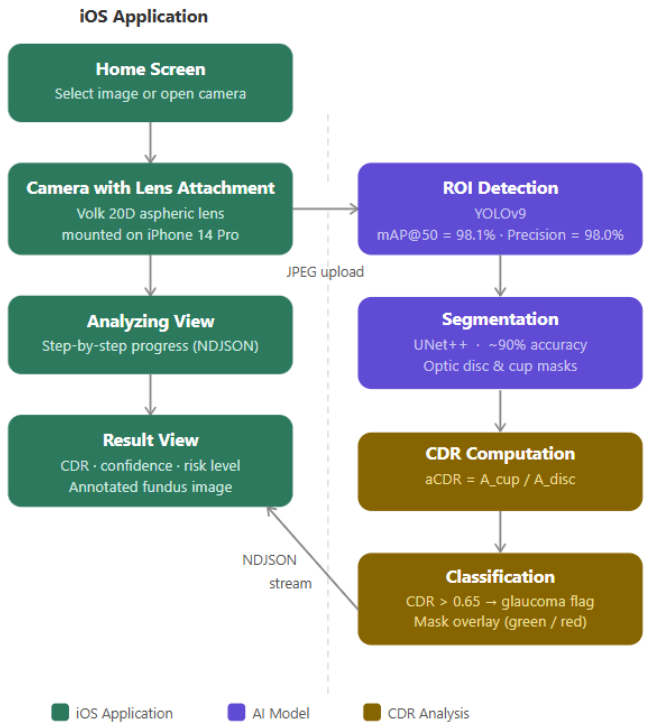

Figure 1: End-to-end glaucoma detection pipeline. A fundus photograph is captured by the camera prototype, uploaded to the backend, passed through YOLOv9 for ROI cropping, segmented by UNet++, and the resulting CDR value is returned to the mobile application.

3.1. AI Pipeline

The processing pipeline operates in five sequential steps:

1. **Image ingestion** – the fundus photograph is received by the backend as a multipart JPEG upload.
2. **ROI detection** – a trained YOLOv9 model localizes the optic disc within the full-resolution image and returns a bounding box.
3. **Cropping** – the image is cropped to the detected bounding box, isolating the region of diagnostic interest.
4. **Segmentation** – the cropped region is passed to a UNet++ model, which produces binary masks for the optic disc and optic cup.
5. **CDR computation and classification** – the CDR is derived from the predicted masks. The resulting masks are overlaid onto the cropped image using color-coded transparency (green for disc, red for cup) and the annotated image is returned alongside the numeric CDR and confidence score.

The CDR is computed from the area of each mask, i.e. as the aCDR, to reduce sensitivity to segmentation boundary noise [2]. Although a vertical CDR of 0.65 is the classical clinical suspicion level, the area CDR operates on a different numeric scale, and its decision threshold was therefore calibrated empirically on the validation set (aCDR ≈ 0.32, Section 8.1) rather than borrowed from the vertical CDR. Images whose aCDR exceeds this calibrated operating point are flagged as indicative of elevated glaucoma risk, and the system recommends the user undergo specialist ophthalmic examination.

3.2. Backend Service

The backend is implemented using FastAPI [17] and hosted on Microsoft Azure. To avoid blocking the asynchronous event loop during computationally intensive inference, model execution is dispatched to a separate thread via `run_in_threadpool`. Analysis progress is streamed to the client using Newline Delimited JSON (NDJSON), with one JSON object emitted per pipeline step. This streaming approach eliminates the "frozen screen" effect that would otherwise occur during the inference time.

The final NDJSON event carries the fields `has_glaucoma`, `confidence`, `cup_to_disc_ratio`, and `image_base64`, where the last field contains a Base64-encoded JPEG of the annotated fundus crop.

3.3. API Response Schema

The final NDJSON event emitted by the backend conforms to the following schema:

- `status` (string): `"success"` or `"error"`.
- `step` (integer, 1–5): current pipeline step index.
- `has_glaucoma` (boolean): glaucoma classification result.
- `confidence` (float, $[0, 1]$): model confidence score.
- `cup_to_disc_ratio` (float, $[0, 1]$): computed aCDR value.
- `image_base64` (string): Base64-encoded JPEG of the annotated fundus crop.

The Swift `GlaucomaResult` struct maps these snake_case fields to camelCase properties via a `CodingKeys` enumeration, maintaining idiomatic naming conventions on both sides of the API boundary.

3.4. iOS Mobile Application

The mobile application is implemented in Swift using the Swift Playgrounds environment, targeting iOS 18.1 and later. The architecture follows the Model–View–ViewModel (MVVM) pattern. A single `GlaucomaViewModel` class, decorated with `@MainActor` to guarantee UI updates on the main thread, owns the application state and coordinates communication with the backend service.

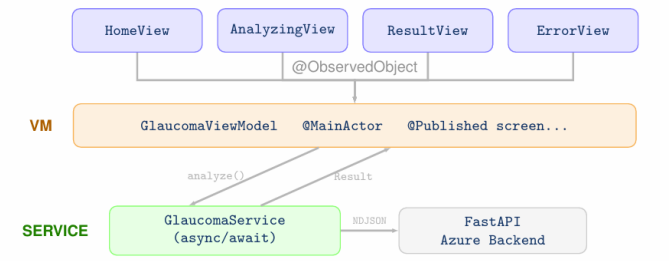

Figure 2: MVVM architecture of the iOS application.

Navigation is modelled as a Swift `enum AppScreen` with four cases: `.home`, `.analyzing`, `.result(GlaucomaResult)`, and `.error(String)`. State transitions are driven exclusively by the

ViewModel, with no NavigationStack involved, giving complete programmatic control over the navigation flow. Screen transitions are animated using .animation(.easeInOut).

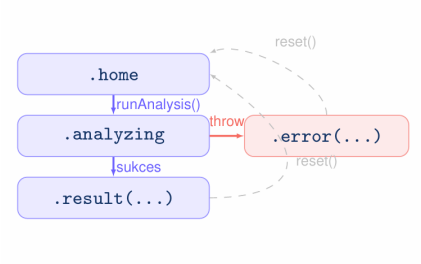

Figure 3: Navigation state machine based on enum AppScreen. Transitions are driven exclusively by GlaucomaViewModel; no NavigationStack is used.

The GlaucomaService class performs the asynchronous HTTP request using:
URLSession.shared.bytes(for:)
which provides line-by-line access to the NDJSON stream via asyncBytes.lines. Each received line is decoded by a JSONDecoder into a StreamProgress value, and the UI is updated on every step event.

The application displays four screens:
(1) HomeView, where the user selects a fundus image from the photo library or camera;
(2) AnalyzingView, which shows an animated progress indicator updated by the streaming steps;
(3) ResultView, displaying the annotated fundus image, a diagnostic banner, and numeric metric cards for CDR, confidence, and risk level;
(4) ErrorView, which provides retry and home navigation options.

The visual design employs a dark medical theme with a consistent color token system. CDR-dependent risk levels are encoded as low (CDR < 0.4, green), moderate ($0.4 \leq$ CDR < 0.6, amber), and high (CDR ≥ 0.6, red).

4. ROI Detection: YOLO Model Comparison

4.1. Dataset and Augmentation

The ROI detection models were trained on the G1020 dataset [15], which originally contained 1020 annotated fundus images. The original images were partitioned with a 70/30 split into a training partition of 714 images and a held-out test set of 306 images (30%); the test images were never seen during training and were kept in their original, un-augmented form. To improve model generalization across imaging conditions, data augmentation was applied to the training partition only, expanding it to over 6100 images. The augmentation configuration included horizontal flipping, rotations of $\pm 15^{\circ}$, hue variation of $\pm 20^{\circ}$, saturation variation of $\pm 25\%$, brightness variation of $\pm 24\%$, blur up to 2.5 px, and random noise affecting up to 1.4% of pixels.

4.2. Evaluated Models and Metrics

Five models from the YOLO family were evaluated: YOLOv8, YOLOv9 [12], YOLOv11, YOLOv12, and YOLOv26. All models were assessed on four standard object detection metrics: precision, recall, F1-score, and mean average precision at an IoU threshold of 0.50 (mAP@50).

4.3. Quantitative Results

Table 1 reports the performance of all evaluated models on the test set.

Table 1: Comparison of YOLO model performance on the fundus ROI detection task.

Model	mAP@50	Precision	Recall	F1
YOLOv8	97.0%	97.5%	96.5%	97.0%
YOLOv12	97.6%	97.5%	96.9%	97.2%
YOLOv11	98.1%	97.6%	97.1%	97.3%
YOLOv26	97.5%	98.0%	98.0%	98.0%
YOLOv9	**98.1%**	**98.0%**	**97.0%**	**97.5%**

All models exceeded 97% mAP@50, confirming that the dataset and augmentation strategy were sufficient for learning robust ROI representations. YOLOv9 and YOLOv11 jointly achieved the highest mAP@50 of 98.1%, while YOLOv26 attained the highest F1-score of 98.0%. However, YOLOv9 surpassed YOLOv11 in both precision and F1-score, and outperformed YOLOv26 in mAP@50. Since mAP@50 more directly reflects the quality of bounding box localization at a defined intersection-over-union threshold—an important property when the cropped region is subsequently passed to a segmentation model—YOLOv9 was selected as the ROI detector for the final pipeline.

Training curve analysis further confirmed the suitability of YOLOv9; its training losses decreased in a regular and consistent manner, and no degradation of validation metrics was observed at the end of training, indicating good generalization without overfitting.

5. Optic Disc and Cup Segmentation: UNet++

The segmentation stage employs a UNet++ architecture [10] with an EfficientNet-B4 encoder backbone. UNet++ extends the original U-Net with a nested encoder–decoder structure and dense skip connections that allow the model to combine features at multiple scales. This property is particularly valuable for optic disc segmentation, where the disc boundary can be subtle and its scale varies across patients.

The segmentation model was trained on the multi-source benchmark dataset described in Section 8.1: a class-balanced collection of approximately 1,300 fundus images carrying expert optic disc and cup masks, partitioned 80/20 into training and validation subsets. Each training image was first reduced to the disc-centred region of interest by the YOLOv9 detector, and UNet++ was trained on these crops against the ground-truth disc and cup masks. The 267-image validation split (Section 8.1) was never seen during training and was used to monitor per-epoch validation. On these held-out validation crops the model achieved a mean Dice of 0.922, comprising a disc Dice of 0.953 (IoU 0.910, pixel accuracy 97.9%) and a cup Dice of 0.892 (IoU 0.804, pixel accuracy 98.6%); the cup is the harder structure owing to its lower contrast against the surrounding neuroretinal rim. Qualitative inspection of the predictions confirmed accurate delineation of the disc boundary on high-quality images, with residual errors concentrated on challenging cases with blurred edges or low image contrast.

Two separate model heads predict the disc mask and the cup mask independently, as shown in Fig. 4.

$$\text{aCDR} = \frac{A_{\text{cup}}}{A_{\text{disc}}} \tag{1}$$

where A_{cup} and A_{disc} denote the number of pixels assigned to each class by the respective mask.

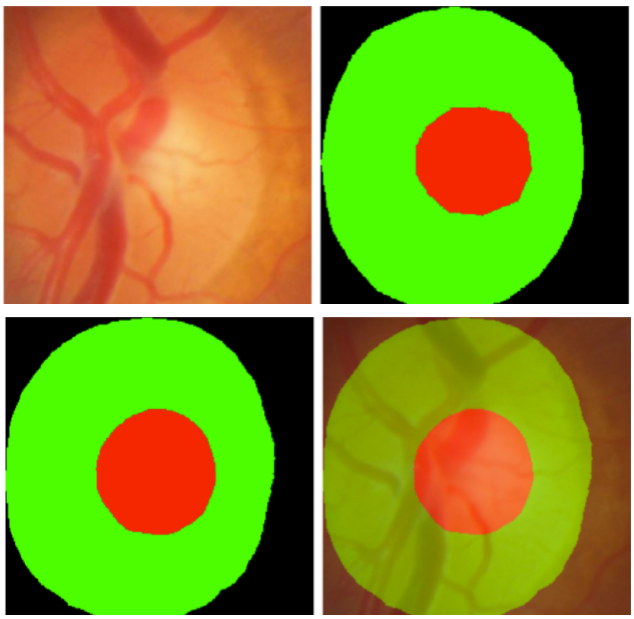

Figure 4: UNet++ segmentation output for a sample fundus image. Top row (left to right): original fundus image, ground truth mask (GT) annotated by a physician. Bottom row (left to right): AI model prediction (Disc IoU = 0.93, Cup IoU = 0.84), color overlay (green – optic disc, red – optic cup).

The DeepLab segmentation model was also evaluated during the exploratory phase but was discarded after achieving only approximately 86% accuracy and exhibiting difficulties in distinguishing the disc boundary from the surrounding retinal tissue on low-contrast images. YOLO-based segmentation was explored for disc and cup delineation, but its inferior shape precision compared to UNet++ made it unsuitable for this task; YOLO was therefore retained exclusively for the ROI detection step.

6. Fundus Camera Prototype

A key design goal of the project was to provide an affordable image acquisition device that removes the dependency on professional fundus cameras costing tens of thousands of PLN. Two hardware iterations were developed: an initial fixed-length tube that served as a proof-of-concept (V1), and the current adjustable, fully 3D-printed rail system (V2) informed by the lessons learned from the proof-pf-concept. Both reuse the same Volk 20D aspheric condensing lens and an off-the-shelf smartphone as the imaging sensor and illumination source.

6.1. First Iteration: Fixed-Tube Prototype

The first prototype consisted of a 25 cm PVC tube (50 mm inner diameter) with the Volk 20D lens mounted at the distal end and a dedicated iPhone 14 Pro protective case bonded to the proximal end. The tube interior was sanded with P100-grade abrasive paper and coated with several layers of ultra-flat black matte spray paint to suppress internal reflections, and the lens was secured with industrial repair tape to permit coarse positional adjustment.

This fixed-geometry device was assembled and evaluated on an optical bench together with an experienced ophthalmologist. Two findings emerged. First, the tube length proved critical to the focal path; the initial 25 cm configuration produced unsatisfactory focus, while shortening the tube to 20 cm improved focusing but introduced internal light reflections that obscured the simulated fundus view. Second, lens choice and internal reflection control were identified as the dominant factors limiting image quality. Above all, the fixed tube offered no means of finely tuning the lens-to-eye distance—which is essential for focusing the aerial retinal image across patients with differing refractive errors—so the device could not reach a usable image quality. This iteration was therefore discontinued in favor of a design with continuously adjustable optics.

6.2. Current Design: Adjustable 3D-Printed Rail System

The current prototype replaces the fixed tube with a linear rail system that lets the operator vary the lens-to-eye distance continuously and in real time. It operates on the principle of indirect ophthalmoscopy; the Volk 20D condensing lens (focal length ≈50 mm) is held between the smartphone camera and the patient's eye, where it forms a real, inverted, aerial image of the retina that is captured by the smartphone's rear camera. The smartphone's built-in LED flash provides illumination, and the device is intended for non-mydriatic use, i.e. without pharmacological pupil dilation.

The mechanical structure comprises three unique 3D-printed parts (FDM, PLA)—one of which is printed twice—connected by four 15 cm stainless-steel guide rods of 5 mm diameter that act as linear

rails:

- **Phone Adapter** (Fig. 5): an L-shaped bracket whose rear face forms a cradle shaped to the smartphone body, with a dedicated cutout for the protruding camera island so the phone seats flush and a central aperture aligned with the rear camera. Four corner bores anchor the guide rods, and a perpendicular wall acts as a backstop.
- **Lens Holder Ring**, printed twice (Fig. 6, left): two identical rings clamp the Volk 20D lens by its rim, applying no stress to the optical surfaces. Four tabs at 90° intervals carry the rod bores, and one ring is fitted with four LM5UU linear ball bearings (5 mm ID / 10 mm OD) so the lens carriage slides smoothly while remaining perpendicular to the optical axis.
- **Eye Cup Stabilizer** (Fig. 6, right): a thick toroidal part that anchors the distal ends of the rods and holds a standard rubber ophthalmoscopy eye cup. The eye cup rests against the patient's orbital rim, blocking ambient light and maintaining a stable working distance.

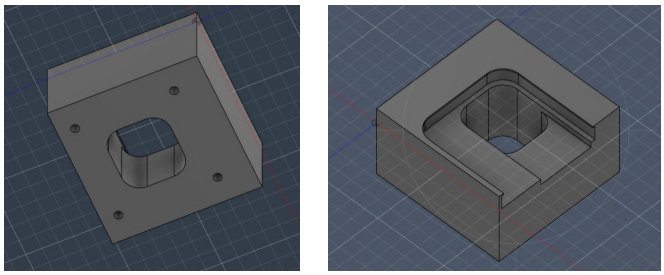

Figure 5: Phone Adapter (CAD model). Left: rod-mounting side, showing the central camera aperture and the four corner bores for the guide rods. Right: phone-cradle side, with the recessed pocket and camera-island cutout that let the smartphone seat flush.

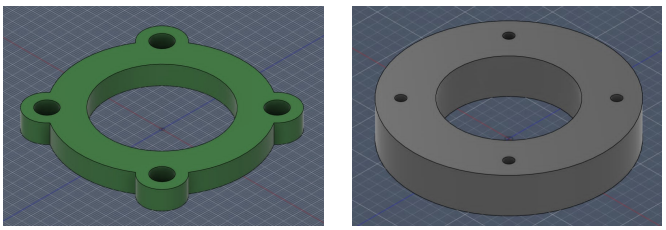

Figure 6: Left: Lens Holder Ring (printed twice); the paired rings clamp the Volk 20D lens by its rim and one carries the four LM5UU linear bearings. Right: Eye Cup Stabilizer, which anchors the distal rod ends and mounts the rubber eye cup.

Sliding the lens carriage along the rods gives a continuously adjustable working distance of roughly 25–50 mm (rail travel up to 150 mm), replacing the coarse, fixed focus of the tube design with fine real-time focus control. In use, the operator places the eye cup against the patient's orbital rim in a dimly lit room, enables the LED flash, and slides the carriage until a focused fundus image appears on the live camera feed before capturing it.

The Bill of Materials is summarized in Table 2. The Volk 20D condensing lens dominates the cost; all other components—the four 3D-printed parts, the guide rods, the linear bearings, and the eye cup—together cost under 120 PLN, keeping the total close to that of the earlier tube design and roughly three orders of magnitude below a commercial fundus camera. As the lens is a standard, reusable clinical instrument, the recurring cost of reproducing the device is only that of the sub-120 PLN mechanical assembly.

Table 2: Bill of Materials for the fundus camera prototype (current 3D-printed rail design). Eye-cup and filament entries are approximate market prices.

Item	Qty	Price (PLN)
Volk 20D Aspheric Lens	1	1194.04
LM5UU Linear Bearing	4	67.20
Stainless Steel Guide Rod (15 cm, 5 mm)	4	6.87
Rubber Eye Cup	1	22.00
PLA Filament (all four printed parts)	–	8.00
Total		**1298.11**

6.3. Current Status and Limitations

Beyond the complete CAD design and assembly specification (Figs. 5–6), the device has not yet been validated for on-eye image capture; no fundus images have been acquired from human subjects, and no clinical or volunteer capture trial has been conducted. It therefore remains a proof-of-concept rather than a validated, certified imaging instrument, and the AI pipeline reported in this paper was consequently developed and validated on established public datasets rather than on prototype-captured images. Achieving a clinically usable on-eye capture is the immediate next step for the hardware work stream.

Several limitations of the current design are already anticipated. The smartphone LED flash sits adjacent to, rather than coaxial with, the camera lens, so the off-axis illumination can introduce uneven lighting and reflections. Focusing is manual, with no autofocus or distance read-out. The phone cradle is geometry-specific and must be re-modelled for smartphones with different camera-island layouts. Finally, the non-mydriatic operation, while improving patient comfort, may limit the achievable field of view in patients with small pupils.

Planned improvements follow directly from these limitations: parametric phone adapters for additional smartphone models, a graduated distance scale along the guide rods for repeatable positioning, threaded fasteners to replace the adhesive bond between the lens-holder rings (enabling tool-free lens cleaning), and a ring-LED module for more uniform, coaxial illumination. The team also intends to consult an optical physicist on lens and illumination optimization before on-eye acquisition is attempted.

7. Testing and Validation

7.1. iOS Unit Tests

The presentation and state-management logic of the iOS application is verified by the `GlaucomaViewModelTests` suite, which comprises 11 unit test cases written with Apple's XCTest framework. Because the `GlaucomaViewModel` is annotated with `@MainActor`, every test method is likewise executed on the main actor, allowing the suite to read and mutate the view model's published state directly and deterministically. The suite deliberately targets the view model in isolation; it exercises the pure, synchronous mapping and state-transition logic that drives the user interface, while the networking and paths are left to the backend tests (Section 7.2). This separation keeps the unit tests independent of a running backend.

7.2. Backend Tests

The backend test suite uses pytest and is organized into four modules. Unit tests (`test_api.py`) mock the `GlaucomaPipeline.run()` method and verify correct CDR-to-classification mapping, boundary behaviour at CDR (which should return `has_glaucoma = False` under the strict inequality threshold), and handling of a null detection result. Validation tests (`test_validation.py`) confirm that the endpoint returns HTTP 422 for missing file uploads and appropriate error codes for text files, empty JPEGs, and corrupted images. Streaming tests (`test_streaming.py`) verify that the NDJSON stream returns HTTP 200, emits at least one event with a `step` field, delivers steps in monotonically increasing order, terminates at step 5, and closes with `status = "success"`. Integration tests (`test_integration.py`) exercise the full endpoint against a running server instance and validate the CDR and confidence values against $[0, 1]$, the Base64 decodability of the returned image, and the HTTP 422 response for missing input.

8. Results and Discussion

The full system achieves the following headline results:

- **YOLOv9 ROI detector:** mAP@50 = 98.1%, precision = 98.0%, ;recall = 97.0% on the G1020 detection test set.
- **UNet++ segmentation model:** mean Dice = 0.922 (disc Dice = 0.953, IoU = 0.910; cup Dice = 0.892, IoU = 0.804), with ROI pixel accuracy of 97.9% (disc) and 98.6% (cup).
- **End-to-end glaucoma classification** on the balanced validation set (267 images, optimised aCDR threshold of 0.323): accuracy = 72.3%, sensitivity = 54.8%, specificity = 90.2%, precision = 85.1%, F1 = 66.7%, AUROC = 0.704.

The two-stage pipeline strategy proved effective; delegating ROI localization to YOLO before passing the cropped sub-image to UNet++ reduced the area of the segmentation task and concentrated model capacity on the diagnostically relevant region. The choice of YOLOv9 over YOLOv11 was motivated by the marginally higher precision and F1-score despite equal mAP@50, which is advantageous for a clinical pre-processing step where incorrect bounding boxes would compromise downstream segmentation.

A clear gap separates the two parts of the result; segmentation quality is high, whereas the end-to-end screening sensitivity (54.8%) falls well short of the $\geq 85\%$ design target stated in Section 1, even though the specificity target ($\geq 90\%$) is met. The following subsection analyzes the origin of this gap.

The NDJSON streaming architecture successfully eliminated frozen-screen effects during the inference window, improving the perceived responsiveness of the mobile application. Streaming five discrete progress events (image ingestion, model loading, AI inference, mask processing, success) gives users meaningful feedback at each stage.

8.1. End-to-End Clinical Validation

To evaluate the full pipeline on the diagnostic task rather than on segmentation alone, we used the held-out validation split of the multi-source benchmark dataset [18] on which the segmentation model was developed. This benchmark aggregates the DRISHTI-GS, ORIGA, REFUGE1, G1020, CRFO, and PAPILA collections; retaining only images that carry expert optic disc and cup masks and balancing the two classes yields a collection of approximately 1,300 images, partitioned 80/20 into training and validation. The 267-image validation split analyzed here (132 healthy, 135 glaucomatous; G1020 80, ORIGA 68, REFUGE1 65, PAPILA 37, DRISHTI-GS 11, CRFO 6) was excluded from segmentation training. Crucially, whereas UNet++ was trained and monitored per epoch on regions already cropped by YOLOv9, for this end-to-end evaluation the same validation images were submitted to the pipeline as complete, uncropped fundus photographs and processed in full: YOLOv9 ROI detection and cropping, UNet++ (EfficientNet-B4 backbone) disc and cup segmentation, area-CDR (aCDR) computation, and thresholding. This measures the system exactly as deployed—including the ROI-detection stage—rather than assuming an oracle crop.

Segmentation quality was consistent across the constituent databases, with mean Dice scores ranging from approximately 0.89 to 0.96 (Fig. 7). Crucially, the aCDR computed from the AI-predicted masks tracks the aCDR derived from the physician ground-truth masks closely, with the majority of points lying near the ideal-correlation diagonal (Fig. 8). This confirms that the segmentation stage reproduces the diagnostic biomarker faithfully.

Quantitatively, the agreement between the AI-derived aCDR and the physician ground-truth aCDR is strong. Across the 267 validation images the mean absolute aCDR error is 0.044 (median 0.031, RMSE 0.065), the Pearson correlation is $r = 0.91$ ($R^2 = 0.83$), and the mean signed error is only $+0.003$, indicating no systematic over- or under-estimation. The error distribution is markedly right-skewed; the median (0.031) lies well below the mean, so the average is inflated by a small number of difficult images rather than reflecting a broad bias. In total, 92.5% of images are reproduced to within an aCDR of 0.10 of the expert value, only 7.5% exceed 0.10, just 1.5%

exceed 0.20, and the single worst case reaches 0.428.

This level of agreement compares favorably with the reproducibility reported between human graders. Clinical cup-to-disc assessment is only moderately repeatable: in the classic study of Tielsch et al. [19], interobserver agreement for the cup-to-disc ratio graded from stereo disc photographs reached only $\kappa = 0.71$ (horizontal) and $\kappa = 0.74$ (vertical). Evidence specific to the area CDR is scarcer, because clinicians conventionally estimate the ratio from diameters rather than areas; however, when six ophthalmologists independently delineated the optic disc and cup areas in the RIGA dataset [20], the inter-observer spread in the resulting area measurements was large enough that outlying annotations had to be identified and removed before the manual contours could serve as a reference standard. Because these human benchmarks derive partly from diameter-based grading, the comparison is indicative rather than exact; nonetheless, the mean area-CDR error of 0.044 and correlation of $r = 0.91$ achieved by our pipeline are of the same order as—and in several respects better than—this documented human disagreement, supporting the use of the segmentation stage as a reliable estimator of the area-CDR biomarker.

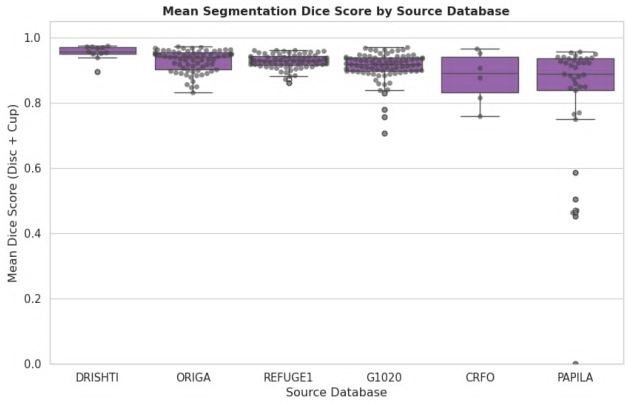

Figure 7: Mean segmentation Dice score (disc + cup) per source database within the validation set.

Despite this agreement, the end-to-end classification performance is modest. Using the operating threshold that maximizes balanced performance on the ROC curve (aCDR > 0.323; AUROC = 0.704, Fig. 9), the system reaches 72.3% accuracy with high specificity (90.2%) and precision (85.1%) but low sensitivity (54.8%). The corresponding confusion matrix is given in Table 3.

Table 3: Confusion matrix for end-to-end AI classification on the balanced validation set ($n = 267$, operating threshold aCDR > 0.323).

	Pred. Healthy	**Pred. Glaucoma**
Actual Healthy	119 (TN)	13 (FP)
Actual Glaucoma	61 (FN)	74 (TP)

The decisive question is whether this limited sensitivity originates from the AI segmentation or from the diagnostic rule itself. To isolate these factors, the same single-threshold decision was applied to the ground-truth physician masks using the standard clinical area-CDR cut-off (aCDR ≥ 0.365). This idealized configuration—equivalent to a perfect segmentation model—yields an even lower sensitivity of 45.9% (accuracy 69.7%, specificity 93.9%, F1 60.5%). In other words, perfect masks do not resolve the problem; the AI pipeline (54.8% sensitivity) actually slightly outperforms the perfect-mask clinical rule because its threshold is calibrated on the data.

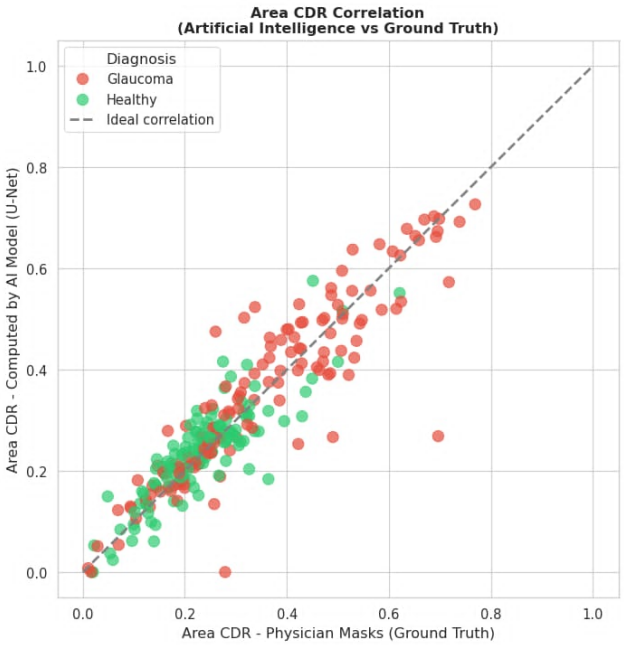

Figure 8: Area CDR computed by the AI pipeline versus the value derived from physician ground-truth masks. Points cluster around the ideal diagonal, indicating close agreement.

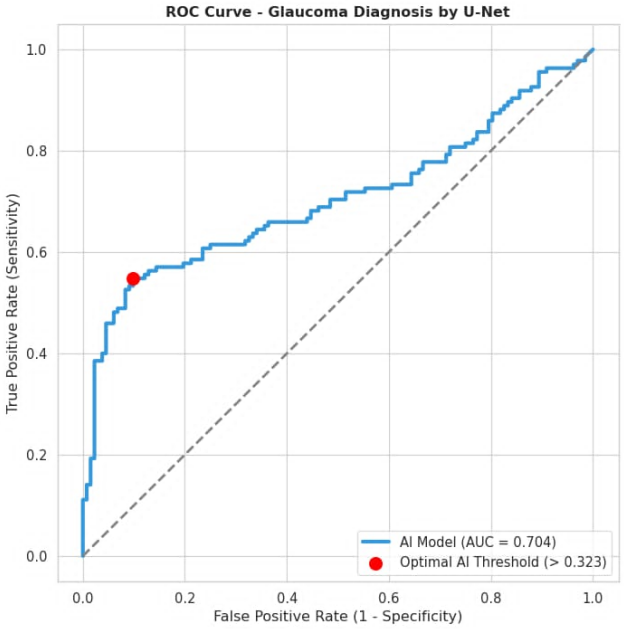

Figure 9: ROC curve for end-to-end glaucoma classification (AUROC = 0.704). The marked point is the selected operating threshold, aCDR > 0.323.

This result localizes the bottleneck precisely. The limiting factor is not the segmentation network but the discriminative ceiling of a single area-CDR threshold; as Fig. 10 shows, the aCDR distributions of healthy and glaucomatous eyes overlap substantially, so no

single cut-off can simultaneously achieve high sensitivity and high specificity. Many true glaucoma cases present with an aCDR below the threshold and are therefore missed. This finding motivates moving beyond a univariate CDR rule towards multi-feature classification and, ultimately, multimodal data, which we identify as the principal direction for future work.

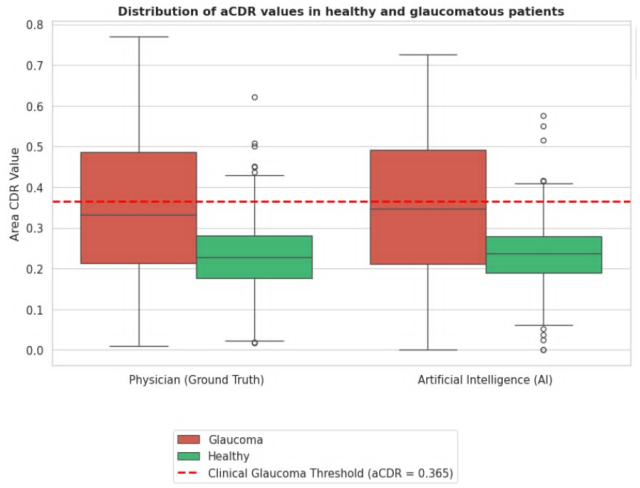

Figure 10: Distribution of aCDR values for healthy and glaucomatous eyes, for physician ground-truth masks and AI predictions. The substantial overlap between classes limits the separability achievable with a single threshold (dashed line: clinical aCDR = 0.365).

8.2. Comparison with Existing Methods

To place the proposed pipeline in context, Table 4 summarizes representative recent approaches to automated glaucoma assessment from fundus images. The comparison spans both whole-image classifiers and detect-then-segment pipelines analogous to ours. We emphasise that these figures are drawn from heterogeneous studies that use different datasets, train/test splits, and evaluation metrics, so the table is intended as indicative context rather than a controlled head-to-head benchmark.

At the segmentation level, the proposed system is competitive with these published methods—its disc and cup Dice scores (0.953 and 0.892) are close to the 0.968 and 0.900 reported by Kim et al. [13], whose Mask R-CNN + MAPNet pipeline is the closest architectural precedent to our detect-then-segment design. At the classification level, however, our end-to-end accuracy (72.3%) is lower than the ≈90–91% reported by classification-oriented studies [7, 14]. As shown in Section 8.1, this gap is not attributable to segmentation quality but to our reliance on a single area-CDR threshold for the diagnostic decision; the cited studies employ richer, learned decision functions rather than a univariate biomarker rule. The distinctive contribution of this work is therefore less a new accuracy record than the integration of a faithful CDR-estimation pipeline with a native mobile streaming application and a low-cost capture device. A definitive, like-for-like accuracy comparison requires evaluation on a shared benchmark with a common split and metric set, which—together with the multi-feature classifier motivated above—is part of the ongoing clinical validation on the partner clinic dataset.

The main limitation of the current system is the image quality requirement; UNet++ accuracy degrades on blurred, low-contrast, or poorly illuminated fundus photographs. Addressing this dependency is one motivation for the dedicated capture device; however, as detailed in Section 6.3, the camera prototype has not yet reached the image quality required for on-eye acquisition, and optimizing its optics is ongoing hardware work. Clinical validation against the annotated dataset from the partner clinic also remains ongoing and will provide sensitivity and specificity estimates on prospectively collected, single-source clinical data to complement the multi-source public validation reported in Section 8.1.

9. Conclusions

This paper has presented an integrated glaucoma screening system comprising a two-stage AI pipeline (YOLOv9 for ROI detection followed by UNet++ for optic disc and cup segmentation), a native iOS mobile application with real-time NDJSON streaming, and a low-cost fundus camera prototype. The ROI detector operates at 98.1% mAP@50 and the segmentation stage reaches a mean Dice of 0.922, reproducing the area-CDR biomarker faithfully (Fig. 8). On a balanced multi-source validation set the end-to-end classifier attains 72.3% accuracy at 90.2% specificity but only 54.8% sensitivity. Our analysis shows this sensitivity ceiling stems from the single-threshold area-CDR decision rule—which limits even perfect physician masks to 45.9% sensitivity—rather than from the segmentation network.

The system is designed to support clinical pre-screening workflows while remaining sufficiently affordable and portable for use outside specialist ophthalmology centres. The camera prototype, constructed for under 1300 PLN, demonstrates that retinal imaging does not require expensive dedicated hardware when combined with the computational photography capabilities of a modern smartphone.

Future work will focus first on replacing the single area-CDR threshold with a multi-feature classifier (incorporating neuroretinal rim morphology, ISNT-rule features, peripapillary atrophy, and vessel-based cues, and ultimately multimodal data) to raise screening sensitivity above the ceiling imposed by a univariate biomarker. Further directions include clinical validation against proprietary patient data from the partner ophthalmology clinic, refinement of the fundus camera optics based on consultation with an optical physicist, integration of a patient survey module within the mobile application, and extension of the iOS application to support direct integration with professional fundus cameras and a database of application users.

Table 4: Indicative comparison of recent fundus-based glaucoma assessment methods. Metrics are reproduced as reported in each source and are *not* directly comparable, as the studies use different datasets, train/test splits, and evaluation protocols. A dash (—) marks a metric not reported by that study.

Study	Approach	Data	Segmentation (Dice)	Classification
Diaz-Pinto et al. [8]	Five ImageNet CNNs, whole-image transfer learning	Public fundus sets	—	Multi-dataset AUC validation
MacCormick et al. [7]	Spatial cup-to-disc profile + probabilistic model	External validation	—	91.0% acc.
Bragança et al. [14]	CNN ensemble on smartphone images	2000 smartphone img.	—	90.0% acc.
Kim et al. [13]	Mask R-CNN + MAPNet (two-stage)	1099 fundus img.	Disc 0.968, Cup 0.900	—
This work	YOLOv9 + UNet++ (two-stage)	SMDG multi-source, $n = 267$	Disc 0.953, Cup 0.892	72.3% acc. (sens. 54.8%, spec. 90.2%)

Acknowledgements

The authors thank Prof. dr hab. inż. Andrzej Czyżewski of the Department of Multimedia Systems at Gdańsk University of Technology for supervision and guidance. The team also acknowledges the support of ophthalmologist dr n. med. Aneta Harasimiuk for clinical consultation and engineer Daniel Wiśniewski for technical consultation on the camera prototype.

References

[1] V. K. Adithya et al., "A systematic review of deep learning methods for glaucoma detection," Journal of Imaging, vol. 7, no. 10, p. 198, 2021.

[2] Y. Zhang et al., "CDR-based glaucoma detection," Scientific Reports, vol. 13, 2023.

[3] J. M. Ahn et al., "Deep learning-based optic disc and cup segmentation," PMC, 2024.

[4] L. J. Coan, B. M. Williams, K. A. Venkatesh, S. Upadhyaya, A. Al Kafri, S. Czanner, R. Venkatesh, C. E. Willoughby, S. Kavitha, and G. Czanner, "Automatic detection of glaucoma via fundus imaging and artificial intelligence: A review," Survey of Ophthalmology, vol. 68, no. 1, pp. 17–41, 2023.

[5] M. Ashtari-Majlan, M. M. Dehshibi, and D. Masip, "Deep learning and computer vision for glaucoma detection: A review," arXiv preprint, 2023.

[6] S. Yousefi, "Clinical applications of artificial intelligence in glaucoma," Journal of Ophthalmic & Vision Research, vol. 18, no. 1, pp. 97–112, 2023.

[7] I. J. C. MacCormick, B. M. Williams, Y. Zheng, K. Li, B. Al-Bander, S. Czanner, R. Cheeseman, C. E. Willoughby, E. N. Brown, G. L. Spaeth, and G. Czanner, "Accurate, fast, data efficient and interpretable glaucoma diagnosis with automated spatial analysis of the whole cup to disc profile," PLOS ONE, vol. 14, no. 1, p. e0209409, 2019.

[8] A. Diaz-Pinto, S. Morales, V. Naranjo, T. Köhler, J. M. Mossi, and A. Navea, "CNNs for automatic glaucoma assessment using fundus images: an extensive validation," BioMedical Engineering OnLine, vol. 18, no. 1, p. 29, 2019.

[9] J. I. Orlando et al., "REFUGE challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs," Medical Image Analysis, vol. 59, p. 101570, 2020.

[10] Z. Zhou et al., "UNet++: A nested U-Net architecture for medical image segmentation," Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 3–11, 2018.

[11] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO." Software library, available via PyPI and GitHub, 2023. Accessed May 2026.

[12] C.-Y. Wang et al., "YOLOv9: Learning what you want to learn using programmable gradient information," in arXiv preprint, 2024.

[13] J. Kim, L. Tran, T. Peto, and E. Y. Chew, "Identifying those at risk of glaucoma: A deep learning approach for optic disc and cup segmentation and their boundary analysis," Diagnostics, vol. 12, no. 5, p. 1063, 2022.

[14] C. P. Bragança, J. M. Torres, C. P. d. A. Soares, and L. O. Macedo, "Detection of glaucoma on fundus images using deep learning on a new image set obtained with a smartphone and handheld ophthalmoscope," Healthcare, vol. 10, no. 12, p. 2345, 2022.

[15] M. N. Bajwa et al., "G1020: A benchmark retinal fundus image dataset," arXiv preprint, 2019.

[16] A. Diaz-Pinto, S. Morales, V. Naranjo, T. Köhler, J. M. Mossi, and A. Navea, "RIM-ONE DL: A unified retinal image database for assessing glaucoma using deep learning," Translational Vision Science & Technology, vol. 11, no. 1, p. 7, 2022.

[17] S. Ramírez, "FastAPI: A modern, fast web framework for building apis with python 3.8+ based on standard python type hints." Software library, available via PyPI, 2024. Originally released 2018; accessed May 2026.

[18] R. Kiefer, "SMDG: A standardized fundus glaucoma dataset (multi-channel glaucoma benchmark dataset)." Kaggle dataset, aggregating DRISHTI-GS, ORIGA, REFUGE1, G1020, CRFO, PAPILA, and other public collections, 2023. Accessed June 2026.

[19] J. M. Tielsch, J. Katz, H. A. Quigley, N. R. Miller, and A. Sommer, "Intraobserver and interobserver agreement in measurement of optic disc characteristics," Ophthalmology, vol. 95, no. 3, pp. 350–356, 1988.

[20] A. Almazroa, S. Alodhayb, E. Osman, E. Ramadan, M. Hummadi, M. Dlaim, M. Alkatee, K. Raahemifar, and V. Lakshminarayanan, "Retinal fundus images for glaucoma analysis: The RIGA dataset," in Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications, vol. 10579 of Proc. SPIE, p. 105790B, 2018.