Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: jeeemi.org; Vol. 8, No. 1, January 2026, pp: 84-104 e-ISSN: 2656-8632

RESEARCH ARTICLE OPEN ACCESS

Comparative Analysis of YOLO11 and Mask R-CNN for Automated Glaucoma Detection

Muhammad Naufaldi Fayyadh[1], Triando Hamonangan Saragih[1], Andi Farmadi[1], Muhammad Itqan Mazdadi[1], Rudy Herteno[1], and Vugar Abdullayev[2]

[1] Computer Science Department, Lambung Mangkurat University, Banjarbaru, South Kalimantan, Indonesia
[2] Department of Computer Engineering, Azerbaijan State Oil and Industry University, Azerbaijan

Corresponding author: Triando Hamonangan Saragih (e-mail: triando.saragih@ulm.ac.id), **Author(s) Email**: Muhammad Naufaldi Fayyadh (e-mail: muhammad.fayyadh01@gmail.com), Andi Farmadi (e-mail: andifarmadi@ulm.ac.id), Muhammad Itqan Mazdadi (e-mail: mazdadi@ulm.ac.id), Rudy Herteno (e-mail: rudy.herteno@ulm.ac.id), Vugar Abdullayev (abdulvugar@mail.ru)

Abstract **Glaucoma is a progressive optic neuropathy and a major cause of irreversible blindness. Early detection is crucial, yet current practice depends on manual estimation of the vertical Cup-to-Disc Ratio (vCDR), which is subjective and inefficient. Automated fundus image analysis provides scalable solutions but is challenged by low optic cup contrast, dataset variability, and the need for clinically interpretable outcomes. This study aimed to develop and evaluate an automated glaucoma screening pipeline based on optic disc (OD) and optic cup (OC) segmentation, comparing a single-stage model (YOLO11-Segmentation) with a two-stage model (Mask R-CNN with ResNet50-FPN), and validating it using vCDR at a threshold of 0.7. The contributions are fourfold: establishing a benchmark comparison of YOLO11 and Mask R-CNN across three datasets (REFUGE, ORIGA, G1020); linking segmentation accuracy to vCDR-based screening; analyzing precision–recall trade-offs between the models; and providing a reproducible baseline for future studies. The pipeline employed standardized preprocessing (optic nerve head cropping, resizing to 1024×1024, conservative augmentation). YOLO11 was trained for 200 epochs, and Mask R-CNN for 75 epochs. Evaluation metrics included Dice, Intersection over Union (IoU), mean absolute error (MAE), correlation, and classification performance. Results showed that Mask R-CNN achieved higher disc Dice (0.947 in G1020, 0.938 in REFUGE) and recall (0.880 in REFUGE), while YOLO11 attained stronger vCDR correlation (r = 0.900 in ORIGA) and perfect precision (1.000 in G1020). Overall accuracy exceeded 0.92 in REFUGE and G1020. In conclusion, YOLO11 favored conservative screening with fewer false positives, while Mask R-CNN improved sensitivity. These complementary strengths highlight the importance of model selection by screening context and suggest future research on hybrid frameworks and multimodal integration.**

Keywords **Glaucoma detection; Fundus imaging; Optic disc and cup segmentation; YOLO11; Mask R-CNN**.

I. Introduction

Glaucoma is a chronic, progressive disease of the optic nerve that gradually leads to the loss of visual field and, if untreated, irreversible blindness. It remains one of the world's most pressing eye health challenges. Current estimates suggest that more than 80 million people live with glaucoma, and this number may exceed 110 million by 2040 [1]. The difficulty with glaucoma is that it often advances silently. By the time patients notice symptoms, significant vision loss has usually occurred. Early detection, therefore, is critical. Prompt diagnosis and treatment can slow the disease, prevent severe complications, and preserve quality of life. Yet in many parts of the world, access to advanced imaging modalities such as optical coherence tomography (OCT) is limited. Retinal fundus photography, by contrast, is relatively inexpensive, non-invasive, and widely available, making it an attractive tool for screening large populations [13] [14].

A key clinical indicator for glaucoma is the vertical Cup-to-Disc Ratio (vCDR). The optic disc (OD) represents the margin of the optic nerve head, while the optic cup (OC) is a central depression whose enlargement reflects progressive damage. As the cup enlarges, the ratio of cup to disc area, particularly in the vertical dimension, increases, providing a quantitative marker of disease severity [2]. This clinical concept is illustrated in Fig. 1, which shows examples of fundus images with small and large vCDR values. Reliable measurement of vCDR depends on accurate segmentation of both the OD and the OC. Manual delineation, while trusted in clinical practice, is time-

Manuscript received 10 September 2025; Revised 16 November 2025; Accepted 30 November 2025; Available online 10 December 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1266
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

consuming and prone to inter-observer variability [15]. At a population level, such manual work is impractical, making automated segmentation methods essential. However, the task is technically challenging: the optic cup is relatively small within the disc, its edges are often faint, and contrast with surrounding tissues is low. Furthermore, retinal images differ in resolution, quality, and acquisition protocols across datasets, which complicates model generalization [3].

Over the last decade, deep learning has transformed medical image analysis, and OD/OC segmentation has significantly benefited from this shift.

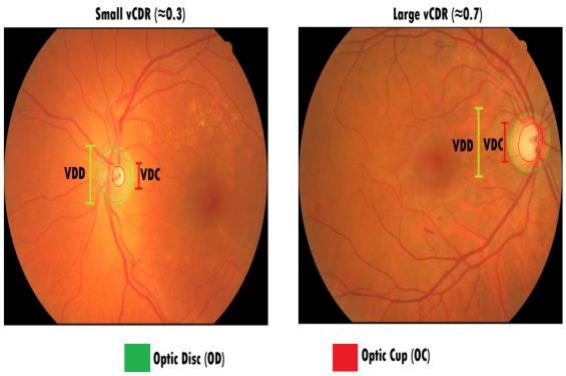

Fig. 1. Fundus photographs illustrating optic disc (red) and optic cup (blue) boundaries. Left: small vCDR (~0.3). Right: large vCDR (~0.7).

Convolutional neural networks, most notably the U-Net architecture and its derivatives, have been widely adopted for their ability to combine local detail with contextual information [3] [20]. More recently, transformer-based models have introduced self-attention mechanisms that capture long-range dependencies in retinal images [4]. Meanwhile, object detection and instance segmentation frameworks, such as Faster R-CNN and Mask R-CNN, have shown strong results [5]. In parallel, the YOLO family of one-stage detectors has become attractive for balancing accuracy with computational speed, enabling near real-time inference [16]. The latest iteration, YOLO11, improves upon its predecessors with optimized backbone design, enhanced mask segmentation, and superior efficiency, making it especially suitable for clinical applications where time and accuracy are both critical. For instance, Chen and Lv proposed an improved YOLOv8 architecture for OD/OC segmentation, achieving competitive results [6].

Deep learning, particularly convolutional neural networks (CNNs), operates on the principle of hierarchical feature learning, where each network layer extracts progressively more abstract visual patterns from image data. Through a combination of convolution, non-linear activation, and pooling operations, CNNs transform low-level pixel intensities into high-level representations that capture structural and spatial relationships. This allows the model to automatically identify critical features, such as optic disc boundaries and cup depressions, in retinal fundus images without manual feature engineering. The ability to model these transformations mathematically enables CNNs to generalize effectively across varied imaging conditions and anatomical differences, forming the theoretical basis for automated medical image segmentation [3] [20] [23].

Both YOLO and Mask R-CNN are built upon this mathematical foundation, but implement it differently to balance accuracy and computational efficiency. YOLO, as a one-stage detector, formulates segmentation as a direct optimization problem where localization, classification, and mask generation are performed simultaneously within a single feed-forward process [16] [33]. In contrast, Mask R-CNN adopts a two-stage formulation that first proposes potential object regions and then refines them through separate classification and mask prediction stages [24] [35]. This decomposition allows Mask R-CNN to achieve higher boundary precision at the cost of greater computational complexity. At the same time, YOLO maintains real-time inference while slightly reducing accuracy for small or low-contrast structures, such as the optic cup. These differing formulations reflect two complementary mathematical strategies for optimizing segmentation and detection in medical imaging [5] [29].

A fundamental distinction in object detection and segmentation frameworks lies between one-stage and two-stage models. Two-stage approaches, such as Faster R-CNN and Mask R-CNN, first generate region proposals and then refine them through a second stage for classification and mask prediction. This design often yields high accuracy, particularly for small or complex structures, but comes at the cost of increased computational demands and slower inference. In contrast, one-stage models like the YOLO family perform detection and segmentation in a single unified step, directly predicting bounding boxes and masks from images without a separate proposal stage. This architecture enables significantly faster inference, making one-stage models attractive for real-time applications. However, their efficiency can sometimes be accompanied by reduced accuracy in detecting small or low-contrast structures such as the optic cup. Comparing these two paradigms is therefore essential in the context of OD/OC segmentation, where both speed and precision directly impact the clinical usability of automated glaucoma screening systems.

Despite this progress, important gaps remain. Many studies still rely on only one dataset, commonly REFUGE, ORIGA, or G1020 [3] [7]. While these datasets are valuable, reliance on a single domain

Manuscript received 10 September 2025; Revised 16 November 2025; Accepted 30 November 2025; Available online 10 December 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1266
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

limits robustness and fails to reflect real-world variability. Systematic baseline comparisons across multiple datasets are also scarce, making it difficult to evaluate the general strengths and weaknesses of existing algorithms. Another gap concerns the comparison between lightweight one-stage detectors like YOLO and more computationally intensive two-stage frameworks such as Mask R-CNN. Without such benchmarking, the trade-offs between speed, accuracy, and clinical usefulness remain poorly understood. Finally, although segmentation metrics are frequently reported, their translation into clinically meaningful measures, such as vCDR, is often overlooked, thereby weakening the direct relevance of algorithmic advances to clinical practice.

Moreover, few existing studies systematically analyze model variability and computational efficiency. Beyond average accuracy, understanding the standard deviation across test images and inference time per prediction is crucial for evaluating model stability and real-time feasibility in clinical environments. Including these aspects provides a more comprehensive assessment of how deep learning models perform under realistic screening conditions.

The present study addresses these gaps by systematically comparing two representative architectures: YOLO11-Segmentation, the latest member of the YOLO family, and the established Mask R-CNN framework.

Both models were trained and evaluated on three publicly available retinal fundus datasets, REFUGE, ORIGA, and G1020, under consistent in-domain baseline conditions. To minimize inter-dataset inconsistencies, the preprocessing pipeline incorporated region-of-interest (ROI) cropping around the optic nerve head, intensity normalization, resizing to 1024 × 1024 pixels, and label harmonization across datasets. Mild data augmentations, including limited rotations, photometric adjustments, and contrast enhancement using CLAHE, were applied during training to improve generalization while preserving anatomical integrity [10] [37]. During evaluation, post-processing steps such as threshold optimization, coverage refinement, and optional enforcement of cup containment within the disc boundary were employed to enhance segmentation accuracy. Importantly, this study extends beyond conventional segmentation metrics by validating model outputs using the vertical Cup-to-Disc Ratio (vCDR), thereby linking algorithmic performance to a clinically interpretable endpoint.

The overarching aim is to evaluate and compare the performance of YOLO11-Segmentation and Mask R-CNN in segmenting OD and OC from retinal fundus images, and to examine how well these segmentations translate into reliable vCDR estimates for glaucoma screening. This research makes several contributions. It establishes in-domain baselines across three widely used datasets, providing a consistent reference for future work. It directly contrasts a modern one-stage segmentation detector with a traditional two-stage framework, offering insight into their relative advantages and limitations. It incorporates post-processing techniques to refine segmentation outputs and reduce common sources of error. Most importantly, it links technical performance to diagnostic relevance by assessing vCDR, thus demonstrating the potential impact of automated segmentation in clinical screening. The main contributions of this study are as follows:

1. A unified multi-dataset benchmark for optic disc and cup segmentation across REFUGE, ORIGA, and G1020 using a standardized preprocessing and evaluation pipeline, enabling fair cross-dataset comparison.
2. A comparative analysis of two representative architectures, YOLO11-Segmentation (one-stage) and Mask R-CNN (two-stage), highlighting their differences in segmentation accuracy, robustness, and inference efficiency for glaucoma screening.
3. A clinically grounded evaluation that links segmentation performance to vertical cup-to-disc ratio (vCDR) estimation, complemented by statistical significance testing to validate reliability across heterogeneous datasets.

The remainder of this paper is structured as follows. Section II reviews related work on OD/OC segmentation and automated glaucoma detection. Section III presents the methodology, including dataset description, preprocessing procedures, model configuration, and evaluation strategy. Section IV reports experimental findings, covering both segmentation performance and vCDR validation. Section V discusses the results in the context of existing studies, outlines limitations, and considers implications for practice. Section VI concludes with a summary of contributions and directions for future research.

II. Method

Fig. 2 illustrates the methodological workflow: integration of three publicly available glaucoma fundus datasets (REFUGE, ORIGA, and G1020), standardized preprocessing (ROI cropping, normalization, resizing to 1024×1024, annotation harmonization), dual-model training (YOLO11-Segmentation as one-stage and Mask R-CNN with ResNet50-FPN as two-stage), disc–cup segmentation, vCDR computation, and multi-level performance evaluation. All segmentation masks were harmonized into a unified two-class schema (disc and cup). Dataset splitting followed official definitions when available

Manuscript received 10 September 2025; Revised 16 November 2025; Accepted 30 November 2025; Available online 10 December 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1266
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

(e.g., REFUGE), and otherwise, a 70/15/15 random split (train/validation/test) was applied for ORIGA and G1020 using a fixed random seed. Conservative augmentation strategies, including mild rotations, limited photometric adjustments, and optional CLAHE, were employed during training to improve generalization while preserving anatomical integrity of the disc–cup complex [37] [10]. Both YOLO11-Segmentation and Mask R-CNN output pixel-level masks for OD and OC. From these masks, the vertical CDR (vCDR) was computed as the ratio of the cup height to the disc height, with optional constraints enforcing anatomical plausibility (cup ⊂ disc). A threshold vCDR ≥ 0.7 was applied to flag advanced enlargement [11]. While all datasets consist of glaucomatous eyes, this threshold served as a proxy of disease progression rather than screening specificity.

Performance was analyzed across three dimensions: (1) segmentation quality, reported in terms of Dice coefficient (F1-score equivalence), Intersection-over-Union (IoU), and mask coverage consistency; (2) fidelity of vCDR estimation, assessed via mean absolute error (MAE), root mean square error (RMSE), correlation analysis, and Bland–Altman plots [12]; and (3) stratified error behavior in relation to disc and cup IoU to assess how geometric localization stability impacts vCDR reliability. Computational efficiency (latency, frames per second, and model size) was also compared, contrasting YOLO11's throughput advantage with Mask R-CNN's spatial consistency. Qualitative analyses included visualization panels across different vCDR ranges (<0.6, 0.6–0.7, ≥0.7) and documentation of standard failure modes, such as cup underestimation in low-contrast images, disc boundary drift, or artifacts caused by glare [8]. Outlier analysis highlighted the most significant absolute CDR errors and their associated IoU values, contextualizing error sources. Overall, YOLO11 was expected to provide superior throughput for scalable deployment, while Mask R-CNN offered more precise localization, potentially reducing CDR errors under challenging cases. A key limitation was the absence of non-glaucoma control images, which precluded specificity analysis; future work should address mixed cohorts to improve clinical generalizability.

A. Dataset

This study utilizes three publicly available retinal fundus datasets, G1020, REFUGE, and ORIGA, comprising a total of 2,870 fundus images with expert-validated segmentation masks for two anatomical targets: the optic disc (disc) and the optic cup (cup). These masks served as the primary anatomical delineations and were subsequently converted into polygonal annotations in COCO (for Mask R-CNN) and YOLO (for YOLO11-Segmentation) formats. Dataset selection was guided by four considerations: (i) variability in acquisition quality, including differences in illumination,

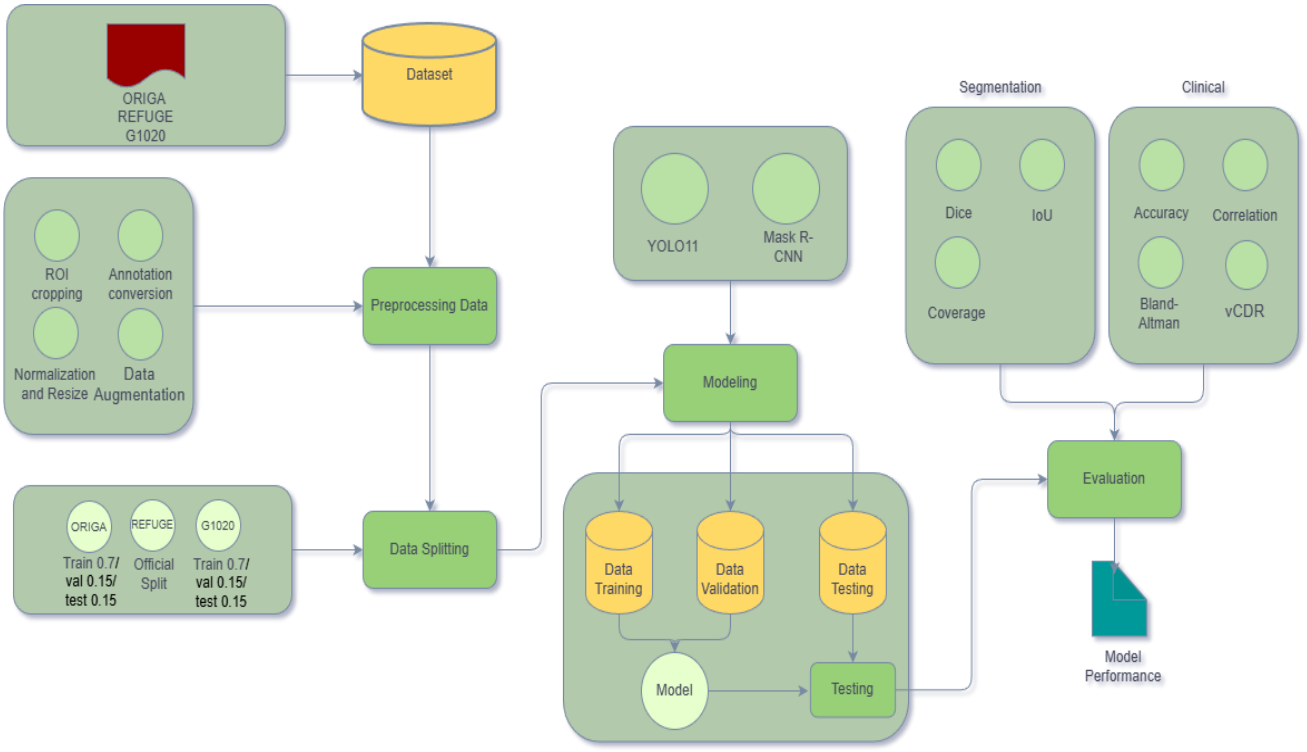

Fig. 2. Methodology Flowchart.

Manuscript received 10 September 2025; Revised 16 November 2025; Accepted 30 November 2025; Available online 10 December 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1266
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

noise, and focus, to enhance robustness of the trained models; (ii) diversity of cup-to-disc ratio (CDR) values within glaucomatous presentations, ensuring inclusion of both early and advanced disease cases [9]; (iii) anatomical consistency enabling cross-dataset harmonization of labels and training procedures; and (iv) reduction of domain bias commonly reported in studies relying on a single-source dataset [17]. All annotations were unified into a two-class schema (disc, cup), with background regions treated implicitly as negative. Importantly, no anatomically distorting transformations were applied prior to preprocessing to preserve structural fidelity. Dataset-specific characteristics:

REFUGE: 1,200 images (400 train, 400 validations, 400 test) with stable imaging quality and moderate structural variability, serving as a reproducible benchmark [18]. The dataset also provides binary glaucoma labels (≈10% positive), but these were not used in this study; the evaluation focused exclusively on OD/OC segmentation and derived vCDR.

ORIGA: 650 images (455 train, 98 validations, 97 test) with a higher prevalence of enlarged CDR values and glaucoma cases (≈30%). Segmentation masks were used for OD/OC delineation, while glaucoma labels were not directly analyzed in this study [19].

G1020: 1,020 images (714 train, 153 validations, 153 test) with broad variation in illumination and contrast, supporting robustness of the models across diverse capture conditions. Clinical labels are inconsistent in the public release; hence, this dataset was used exclusively for segmentation [17].

The integration of these three datasets strengthens the reliability of simultaneous disc–cup detection, which is essential for accurate derivation of vCDR as an early indicator of glaucomatous optic neuropathy [7]. A limitation of this composition is the absence of non-glaucomatous control images in the combined dataset, which precludes direct analysis of screening specificity and underscores the need for future validation on mixed cohorts containing both normal and glaucomatous eyes.

B. Data Preprocessing (H2, Arial 10, BOLD)

The preprocessing pipeline aimed to reduce inter-dataset variability in image resolution, illumination, and annotation format, ensuring consistency across both YOLO11-Segmentation and Mask R-CNN frameworks. It comprised five steps: ROI cropping, normalization, resizing, annotation conversion, and data augmentation. The overall pipeline is illustrated in Fig. 3, which shows original fundus images from three datasets alongside their ROI-cropped and resized versions.

1. ROI Cropping

All images were cropped around the optic nerve head (ONH) region to remove redundant background areas and retain only the optic disc (OD) and optic cup (OC). The cropping coordinates were derived from the ground-truth disc mask using a scaling factor $\alpha = 1.4$, ensuring that both disc and cup were fully enclosed. In this study, we introduced an ROI-based cropping formulation, computed mathematically as shown in Eq. (1) and Eq. (2):

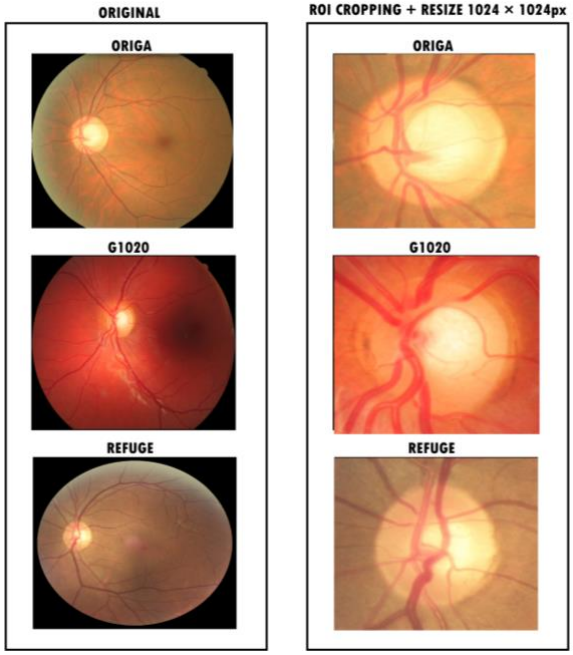

Fig. 3. Preprocessing pipeline. Left: original fundus images from REFUGE, ORIGA, and G1020 datasets. Right: ROI-cropped and resized images standardized to 1024 × 1024 pixels.

$$x'_{min} = x_{disc} - \alpha\frac{w_{disc}}{2}, x'_{max} = x_{disc} + \alpha\frac{w_{disc}}{2} \quad (1)$$

$$y'_{min} = y_{disc} - \alpha\frac{h_{disc}}{2}, y'_{max} = y_{disc} + \alpha\frac{h_{disc}}{2} \quad (2)$$

Where (x_{disc}, y_{disc}) denote the centroid of the optic disc and (w_{disc}, h_{disc}) its width and height. This ROI-based cropping reduced irrelevant pixels (vessels, macula, and peripheral retina), improving computational efficiency and focusing the model on relevant structures. ROI-based strategies are widely adopted in fundus image analysis to enhance segmentation accuracy for small anatomical targets [21] [22].

2. Image Normalization and Resizing

Prior to resizing, all cropped images were normalized to harmonize intensity distributions across datasets. Pixel normalization followed Eq. (3) [57]:

$$I_{norm}(x, y) = \frac{I(x,y) - \mu_I}{\sigma_I} \quad (3)$$

Manuscript received 10 September 2025; Revised 16 November 2025; Accepted 30 November 2025; Available online 10 December 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1266
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

where $I(x,y)$ represents the pixel intensity, and μ_I, σ_I denote the mean and standard deviation of the image intensities, respectively. This normalization mitigates illumination bias and improves inter-dataset consistency, ensuring that variations in imaging conditions do not influence model learning. Normalized images were then resized to 1024 × 1024 pixels using bilinear interpolation, an established approach that balances computational efficiency with anatomical fidelity [23]. The bilinear interpolation function can be expressed using Eq. (4) [58] as follows:

$$I'(x',y') = \sum_{i=0}^{1}\sum_{j=0}^{1} w_{ij} \cdot I(xi,yj) \quad (4)$$

where wij are weights proportional to the relative distance of pixel (x′, y′) to its four nearest neighbors (xi, yj). This method preserves structural details, particularly the optic cup, which often occupies less than 10% of the disc area, while maintaining smooth pixel transitions and preventing aliasing during upsampling. For the YOLO11-Segmentation branch, symmetric letterboxing was applied whenever the original aspect ratio was not 1:1 to avoid distortion of circular anatomical structures. In contrast, the Mask R-CNN branch employed direct square resizing without padding, maintaining full coverage of the optic disc region [16] [24]. The normalization and resizing steps ensured geometric consistency across models, improving spatial correspondence for training and vCDR analysis.

3. Annotation Conversion

All extracted disc and cup masks were harmonized into a canonical COCO-style JSON file containing image metadata (ID, filename, width, height), category mapping (disc, cup), and annotation records (image ID, category ID, bounding box, polygon, and area). This served as the single source of truth for both detection pipelines. From it, two model-specific annotation representations were derived:

Mask R-CNN (COCO format): bounding boxes and segmentation polygons were retained in pixel space: $B = (xmin, ymin, xmax, ymax)$ Only two categories were preserved (disc, cup), while background regions were implicitly treated as negative.

YOLO11-Segmentation (YOLO format): annotations were exported as one text file per image, with normalized center-based bounding boxes and polygon coordinates. The normalized bounding box representation followed Eq. (5) [27] and Eq. (6) [27]:

$$x_c = \frac{xmin+xmax}{2W}, y_c = \frac{ymin+ymax}{2H} \quad (5)$$

$$w_n = \frac{xmax-xmin}{W}, h_n = \frac{ymax-ymin}{H} \quad (6)$$

where W and H are the resized image dimensions. For the YOLO branch, offsets from padded letterboxing were applied to ensure geometric consistency. This staged conversion enforced consistency across frameworks and prevented divergence in geometric interpretation, a prerequisite for reliable downstream vCDR estimation, a practice well-accepted in the object detection community [25] [26] [27].

4. Dataset Splitting

Dataset partitioning followed the official or commonly adopted conventions of each dataset to ensure comparability and reproducibility across experiments. REFUGE used its predefined split (400 training, 400 validations, and 400 test images) [18]. ORIGA (650 images) [19] and G1020 (1,020 images) [17] were divided into 70% training, 15% validation, and 15% testing using a fixed random seed to guarantee reproducible sampling and to prevent bias from random shuffling. No image overlap occurred between the subsets, ensuring strict independence of evaluation. Unlike several prior works that employed stratification by vCDR ranges, this study adopted pure random splitting in accordance with common in-domain baseline practices. This design choice was made to prevent artificial class imbalance that could distort real-world glaucoma prevalence and affect model generalization. For example, Thompson et al. [28] demonstrated that random participant-level partitioning (50/20/30 train/validation/test) maintains independence between training and evaluation while preserving natural variability in disease severity. The training set was used to optimize model parameters, the validation set guided hyperparameter tuning and early model selection, and the test set was reserved exclusively for final performance reporting to avoid overfitting and ensure fair benchmarking across both YOLO11-Segmentation and Mask R-CNN. This partitioning strategy provided a consistent and reproducible foundation for quantitative comparison across the three datasets.

5. Data Augmentation

Data augmentation was applied exclusively to the training subsets to enhance generalization capability without compromising anatomical fidelity. Augmentation expands data variability by simulating acquisition differences such as camera angle, illumination, and contrast, thereby reducing overfitting and improving cross-domain robustness. Conservative transformations included:

Mild rotations (±10°): simulated acquisition variability while preserving the vertical axis critical for vCDR [30].

Horizontal flips (low probability): introduced variation but avoided excessive disruption of lateral asymmetry between eyes.

Photometric adjustments: limited brightness and contrast modifications to mimic device differences.

Manuscript received 10 September 2025; Revised 16 November 2025; Accepted 30 November 2025; Available online 10 December 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1266
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

CLAHE (Contrast-Limited Adaptive Histogram Equalization, optional): enhanced local contrast and improved visibility of cup–disc boundaries [10] [37].

The geometric transformations (rotation and flipping) were defined mathematically as follows. For a rotation by angle θ about the image center (x_c, y_c), the new pixel coordinates (x', y') were computed according to Eq. (7) [57]:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x - x_c \\ y - y_c \end{bmatrix} + \begin{bmatrix} x_c \\ y_c \end{bmatrix} \quad (7)$$

This affine transformation preserves the relative geometry of the optic disc and cup while introducing controlled angular variation. For horizontal flipping, pixel coordinates were mirrored with respect to the vertical image axis, as expressed in Eq. (8):

$$x' = W - x, y' = y \quad (8)$$

where W is the image width. Both transformations were simultaneously applied to the corresponding polygon and bounding box annotations to maintain spatial alignment between images and masks. Aggressive compositional augmentations, such as Mosaic, CutMix, or Copy-Paste, were intentionally excluded, as they may create anatomically implausible disc–cup relations [31]. Polygon coordinates were automatically recalculated after each transformation within the training framework.

The overarching goal of these augmentations was to maintain superior–inferior rim geometry, since vertical precision directly influences vCDR reliability [32]. This mathematically constrained augmentation policy ensured a balance between model robustness and clinical interpretability across datasets.

C. Model Architectures

Two deep learning architectures were employed in this study for optic disc (OD) and optic cup (OC) segmentation: YOLO11-Segmentation, a modern single-stage, real-time framework, and Mask R-CNN with a ResNet-50-FPN backbone, a two-stage region-based framework. Convolutional neural networks (CNNs) form the mathematical backbone of both architectures, learning hierarchical image representations through convolution, non-linear activation, and pooling operations. Each convolutional layer transforms an input feature map $x^{(l-1)}$ into a higher-level representation $f^{(l)}$ using learned kernels $W^{(l)}$, bias $b^{(l)}$, and activation function $\sigma(\cdot)$, as expressed in Eq. (9) [59]:

$$f^{(l)} = \sigma(W^{(l)} * x^{(l-1)} + b^{(l)}) \quad (9)$$

where $*$ denotes the convolution operation. This operation enables CNNs to capture local spatial dependencies and progressively abstract features from retinal fundus images, laying the foundation for accurate optic disc and cup segmentation. These architectures were selected to contrast the speed and efficiency of lightweight detectors with the high localization accuracy typically achieved by region-based models. YOLO11, as a one-stage model, directly predicts bounding boxes and masks in a single step, whereas Mask R-CNN, a two-stage framework, first generates candidate regions before refining classification and segmentation [29]. In a general object-detection and segmentation task, the objective is to learn a mapping function $f_\theta: X \to Y$, where X denotes the input image and Y represents the structured output containing bounding boxes $B = (x, y, w, h)$, class probabilities $P(c \mid B)$, and pixel-level segmentation masks M. The training process minimizes a multi-task loss that jointly optimizes classification, localization, and segmentation, as defined in Eq. (10) [24] [49]:

$$\mathcal{L}_{total} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{box}\mathcal{L}_{box} + \lambda_{mask}\mathcal{L}_{mask} \quad (10)$$

where λ_{cls}, λ_{box}, and λ_{mask} are weighting factors. This general formulation is shared across both one-stage and two-stage frameworks; the main difference lies in how region proposals and mask generation are handled.

1. YOLO11-Segmentation

YOLO11 belongs to the "You Only Look Once" family of one-stage detectors, designed for real-time detection and segmentation [16]. Unlike classical multi-pass models, YOLO performs all predictions in a single forward pass, enabling low-latency inference suitable for clinical deployment. The image is divided into an $S \times S$ grid, and each grid cell predicts B bounding boxes and corresponding class scores. For each predicted box i, the model outputs $\hat{y}_i = (x_i, y_i, w_i, h_i, C_i, p_{i,1}, p_{i,2}, \ldots, p_{i,K})$ where (x_i, y_i, w_i, h_i) are normalized coordinates of the bounding box, C_i is the objectness confidence, and $p_{i,k}$ denotes the class probability for class k. The final detection confidence is calculated as $P_{det}(k) = C_i \times p_{i,k}$. In the segmentation branch, YOLO11 employs a prototype-based mask generation mechanism inspired by YOLACT [33]. In this approach, a set of N_p global prototype masks P_j is produced and later combined with instance-specific coefficients α_{ij} to construct pixel-level segmentation masks. The combination process, which enables efficient and differentiable mask generation for each detected instance, is formally defined in Eq. (11) [33]:

$$M_i = \sum_{j=1}^{N_p} \alpha_{ij} P_j \quad (11)$$

where M_i represents the reconstructed instance mask for object i. This formulation allows YOLO11 to predict bounding boxes and pixel-level masks in parallel efficiently. The backbone employs CSP-based C2f modules to enhance feature reuse, while a Path Aggregation Network (PANet) neck merges multi-scale features critical for detecting the small optic cup within the larger optic disc [50]. A decoupled detection head

Manuscript received 10 September 2025; Revised 16 November 2025; Accepted 30 November 2025; Available online 10 December 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1266
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

separates classification and localization sub-tasks, improving bounding-box regression stability. YOLO11 was initialized with COCO pre-trained weights and fine-tuned for two-class disc-cup segmentation [26]. Its main advantages lie in computational efficiency and high throughput, maintaining low latency even for high-resolution clinical images [34].

2. Mask R-CNN

Mask R-CNN extends Faster R-CNN into a two-stage instance segmentation framework designed for precise object localization and boundary delineation. The model employs a ResNet-50 backbone with a Feature Pyramid Network (FPN) to extract hierarchical features across multiple scales, enabling robust detection of the relatively large optic disc alongside the smaller optic cup [51]. A Region Proposal Network (RPN) generates candidate regions of interest at different scales, effectively adapting to variations in disc size and relative cup proportion across patients. The extracted regions are refined using ROI Align, which eliminates spatial quantization artifacts introduced by ROI pooling, thereby preserving the fine vertical boundary geometry crucial for accurate vCDR computation [36]. Formally, the RPN predicts objectness scores p_i and bounding-box offsets $t_i = (t_x, t_y, t_w, t_h)$ for each anchor i. The overall RPN loss combines classification and regression terms, as defined in Eq. (12) [24] [35]:

$$\mathcal{L}_{\mathcal{RPN}} = \frac{1}{N_{cls}} \sum_i \mathcal{L}_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* \mathcal{L}_{reg}(t_i, t_i^*) \quad (12)$$

where p_i^* and t_i^* denote the ground-truth label and bounding-box regression targets, respectively. Candidate regions generated by the RPN are refined via ROI Align to maintain sub-pixel accuracy.

A detection head then performs classification into two categories (disc and cup) while simultaneously refining bounding boxes. In parallel, a fully convolutional mask head generates binary segmentation masks for each detected region. The mask-generation process is optimized using a pixel-wise binary cross-entropy loss, formally expressed in Eq. (13) [24]:

$$\mathcal{L}_{mask} = \frac{1}{m^2} \sum_{u,v} \mathrm{BCE}\big(M_k(u,v), M_k^*(u,v)\big) \quad (13)$$

where BCE denotes the pixel-wise binary cross-entropy function and (u, v) are the spatial coordinates within each ROI. This architecture was chosen for its ability to achieve higher localization precision under challenging conditions such as low contrast, peripapillary atrophy, or glare artifacts. Furthermore, Mask R-CNN is effective at suppressing false positives introduced by vascular reflections or uneven illumination. As such, it provides a strong spatial accuracy baseline against which the lighter, single-stage YOLO11 can be comparatively evaluated in this study [5].

D. Training Strategy

The training procedure was carefully designed to ensure reproducible, fair, and clinically meaningful comparisons between YOLO11-Segmentation and Mask R-CNN. Both models were trained under in-domain baseline conditions using the harmonized datasets (REFUGE, ORIGA, and G1020). The strategy encompassed data usage, optimization configuration, loss functions, regularization, and convergence monitoring.

1. Data Usage

Each dataset (REFUGE, ORIGA, and G1020) [17] [18] [19] was partitioned into three subsets: training, validation, and testing. The training set was employed to optimize model parameters through iterative weight updates. The validation set was reserved for hyperparameter tuning, performance monitoring during training, and model checkpoint selection via early stopping, ensuring that overfitting was minimized. The test set remained strictly unseen during model development and was used exclusively for final performance evaluation, providing an unbiased measure of generalization. This separation of data subsets guarantees the reliability of reported results and aligns with best practices in medical image analysis. [38].

2. Optimization and Hyperparameters

YOLO11-Segmentation was trained for 200 epochs with a batch size of 2, balancing GPU memory limitations and training stability. The model followed the Ultralytics training configuration, employing the AdamW optimizer with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, an initial learning rate of 1×10^{-3}, and a weight decay of 5×10^{-4}. The optimization process for both algorithms follows the standard gradient-based update rule. For stochastic gradient descent (SGD), model parameters are updated iteratively according to Eq. (14) [59]:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}(\theta_t) \quad (14)$$

where η denotes the learning rate and $\nabla_\theta \mathcal{L}(\theta_t)$ is the gradient of the loss with respect to parameters θ_t. AdamW extends this update mechanism by maintaining first and second moment estimates, as shown in Eq. (15) [56], to adaptively scale the learning rate while decoupling weight decay from the gradient:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\nabla_\theta \mathcal{L}_t, \; v_t = \beta_2 v_{t-1} + (1-\beta_2)(\nabla_\theta \mathcal{L}_t)^2 \quad (15)$$

The parameter update is then computed using Eq. (16) [56]:

$$\theta_{t+1} = \theta_t - \eta \frac{m_t}{\sqrt{v_t}+\epsilon} - \eta \lambda \theta_t \quad (16)$$

where λ represents the weight-decay coefficient and $\epsilon$is a small constant for numerical stability. As expressed in Eq. (14), Eq. (15), and Eq. (16), these

Manuscript received 10 September 2025; Revised 16 November 2025; Accepted 30 November 2025; Available online 10 December 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1266
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

formulations ensure smooth, and stable convergence, preventing overfitting during training. A cosine annealing learning rate scheduler [39] was used, gradually decreasing the learning rate along a half-cosine curve until it reached zero, promoting smooth convergence and preventing oscillatory updates. Mask R-CNN, in contrast, was trained for 75 epochs using stochastic gradient descent (SGD) with momentum = 0.9 and the same weight decay = 5×10^{-4} [24]. A similar cosine decay schedule was applied. Mask R-CNN required fewer epochs, reflecting faster convergence due to its smaller number of trainable parameters in the ROI heads compared to YOLO11's end-to-end design. Both models employed gradient clipping (norm ≈ 10.0) to stabilize updates. The best checkpoint was automatically selected based on the highest validation Dice score. AdamW provided adaptive learning rate correction and decoupled weight decay for YOLO11, while SGD with momentum provided smoother optimization for Mask R-CNN [56].

3. Loss Functions

For YOLO11-Segmentation, the training objective was a weighted sum of classification, localization, objectness, and segmentation losses, expressed in Eq. (17) [16] [39] :

$$L_{YOLO} = \lambda_{cls}L_{cls} + \lambda_{box}L_{box} + \lambda_{obj}L_{obj} + \lambda_{mask} L_{mask} \quad (17)$$

where $\lambda_{cls}L_{cls}$ is the binary cross-entropy classification loss, $\lambda_{box}L_{box}$ is the bounding box regression loss (IoU-based), $\lambda_{obj}L_{obj}$ is the objectness confidence loss, and $\lambda_{mask}L_{mask}$ is the hybrid dice–binary cross-entropy loss for mask prediction [16].

The Dice loss is incorporated to enhance boundary alignment and region overlap between the predicted mask P and ground truth G, formulated as in Eq. (18) [60]:

$$\mathcal{L}_{Dice} = 1 - \frac{2|P \cap G|}{|P| + |G|} \quad (18)$$

where $|P \cap G|$ denotes the intersection area of predicted and ground-truth masks, and $|P|$, $|G|$ represent their respective pixel counts. The binary cross-entropy (BCE) loss is defined in Eq. (19) [59]:

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \quad (19)$$

In YOLO11-Segmentation, these two components are combined into a hybrid segmentation loss, as expressed in Eq. (20) [60]:

$$\mathcal{L}_{mask} = \alpha\mathcal{L}_{BCE} + (1 - \alpha)\mathcal{L}_{Dice} \quad (20)$$

where $\alpha \in [0,1]$ controls the contribution of BCE and Dice terms. This hybrid formulation stabilizes early training by leveraging BCE's pixel-wise sensitivity and enhances final accuracy through Dice's global region-overlap optimization. The Dice loss term (Eq. 18) emphasizes overlap maximization, whereas BCE (Eq. 19) preserves pixel-level precision. Their hybridization (Eq. 20) balances global-region consistency with local-boundary accuracy, explaining why Mask R-CNN achieves more stable disc delineation while YOLO11 attains faster convergence. For Mask R-CNN, the total loss consisted of classification, bounding box refinement, and mask segmentation, defined in Eq. (21) [24] :

$$L_{MRCNN} = L_{cls} + L_{box} + L_{mask} \quad (21)$$

where L_{cls} is the softmax cross-entropy for region classification, L_{box} is a smooth L1 loss for bounding box regression, and L_{mask} is binary cross-entropy applied at the pixel level within each region of interest [24].

4. Regularization and Generalization

Conservative data augmentation strategies were applied to the training subset, including small translations, scaling, horizontal flips, and mild photometric adjustments [10]. Anatomical constraints (e.g., no vertical flips or extreme rotations) ensured structural fidelity of fundus images. YOLO11 relied on AdamW regularization, per-class confidence tuning, and weight decay, while Mask R-CNN used weight decay, FPN-based feature sharing, and gradient clipping.

Algorithm 1. Training Workflow for YOLO11-Segmentation and Mask R-CNN

```
Input: Preprocessed datasets D = {REFUGE, ORIGA, G1020}
Output: Trained model parameters θ_YOLO11, θ_MaskRCNN
Split D into Dtrain, Dval, Dtest
Initialize models:
    YOLO11 ← COCO pretrained weights
    Mask R-CNN ← ResNet50-FPN pretrained weights
for each epoch t in [1, T]:
    for each batch (x, y) in Dtrain:
        ŷ ← fθ(x)                    # forward pass
        Compute total loss:
            L = λcls Lcls + λbox Lbox + λobj Lobj + λmask Lmask
        Compute gradient ∇θL
        Update weights θ ← θ − η∇θL    # AdamW or SGD
    Evaluate model on Dval
    if Dice_cup improved:
        Save checkpoint
end for
Select best model based on validation Dice_cup
Evaluate θ* on Dtest → report Dice, IoU, MAE, RMSE, r
```

Manuscript received 10 September 2025; Revised 16 November 2025; Accepted 30 November 2025; Available online 10 December 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1266
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

5. Convergence Monitoring

Model performance was continuously monitored on the validation set. Evaluation included Dice coefficient, Intersection over Union (IoU), precision, recall, and clinical metrics such as vertical cup-to-disc ratio (vCDR) error (MAE, RMSE, correlation, Bland–Altman plots) [9]. Training and validation losses were monitored to confirm stable convergence. For YOLO11, the best checkpoint was selected using cup Dice, the most clinically relevant metric.

Algorithm 2. Inference and vCDR Computation

Input: Test fundus image I
Output: Predicted vertical cup-to-disc ratio vCDR_pred

1: Perform ROI cropping and normalization on I
2: Obtain segmentation masks:
M_disc, M_cup ← Model.predict(I)
3: Compute vertical heights:
H_disc = ymax(M_disc) − ymin(M_disc)
H_cup = ymax(M_cup) − ymin(M_cup)
4: Calculate vertical ratio:
vCDR_pred = (H_cup / H_disc) + 0.05 # calibration offset
5: Compare with clinical threshold:
if vCDR_pred ≥ 0.70 → label = "glaucoma-suspect"
else → label = "normal"

6. Implementation Details

YOLO11-Segmentation was implemented using the Ultralytics YOLO11 framework [16], while Mask R-CNN was implemented via the Torchvision detection library [40]. Training was conducted on an NVIDIA RTX 3060 Laptop GPU (6 GB VRAM). Mixed-precision training (FP16/AMP) was enabled to accelerate computation and reduce memory usage. Data loading was parallelized across 4 workers, and persistent workers were used in Mask R-CNN for efficiency. All experiments were initialized with a fixed random seed (42) to ensure reproducibility. To improve methodological clarity and reproducibility, the overall training and inference procedures are summarized in algorithmic form. The complete training pipeline, including dataset loading, optimization, and checkpointing, is outlined in Algorithm 1, while the inference and vCDR computation process is detailed in Algorithm 2.

E. Cup-to-Disc Ratio Calculation

The vertical cup-to-disc ratio (vCDR) is a widely used structural indicator of glaucoma risk [32]. After preprocessing and segmentation, the vertical height of the optic disc (H_{disc}) and the optic cup (H_{cup}) are measured directly from their respective binary masks. The vertical dimensions are computed by identifying the uppermost and lowermost pixel coordinates along the vertical axis of each segmented region, as defined in Eq. (22) [32]:

$$H_{disc} = y_{max}^{disc} - y_{min}^{disc}, H_{cup} = y_{max}^{cup} - y_{min}^{cup} \quad (22)$$

where y_{max} and y_{min} denote the maximum and minimum vertical pixel coordinates for each segmented region. The vertical cup-to-disc ratio is then calculated as the ratio of these two heights, as expressed in Eq. (23) [32] [9]:

$$vCDR = \frac{H_{cup}}{H_{disc}} \quad (23)$$

This formulation directly links the model's pixel-level segmentation outputs to a clinically interpretable biomarker, enabling objective quantification of neuroretinal rim thinning. Because deep learning models tend to slightly underestimate cup size, a calibration offset of +0.05 was applied to the predicted ratio in this study, consistent with established evaluation protocols [41]. The vertical dimension is emphasized since superior–inferior neuroretinal rim thinning is more sensitive to glaucomatous damage than horizontal or area-based changes [42]. A threshold of vCDR ≥ 0.70 (after calibration) was adopted to denote suspicious enlargement, in line with prior clinical literature [11]. It should be noted that this threshold serves only as an automated screening rule and does not substitute for comprehensive ophthalmological examination, including intraocular pressure measurement, visual field testing, or optical coherence tomography.

F. Evaluation Metrics

Model evaluation was performed across two anatomical categories, optic disc (OD) and optic cup (OC), with background not treated as an explicit class. The assessment encompassed both segmentation accuracy and clinical consistency, ensuring that the model outputs were validated not only geometrically but also in terms of their diagnostic implications. The core segmentation metrics included Dice coefficient, Intersection over Union (IoU), and prediction coverage. The Dice coefficient, equivalent to the F1-score, quantifies the overlap between predicted and ground-truth segmentation masks as defined in Eq. (24) [20] [43]:

$$\text{Dice} = \frac{2TP}{2TP+FP+FN} \quad (24)$$

where TP (true positives) denote correctly segmented pixels belonging to the target region, FP (false positives) represent pixels incorrectly predicted as part of the target, and FN (false negatives) correspond to pixels belonging to the target region that the model missed. This metric balances precision and recall, making it particularly sensitive to boundary accuracy in biomedical image segmentation. The Intersection over Union (IoU), defined in Eq. (25) [44], measures the geometric overlap between the predicted and reference

Manuscript received 10 September 2025; Revised 16 November 2025; Accepted 30 November 2025; Available online 10 December 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1266
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

regions, where B_p denotes the set of pixels (or area) predicted as belonging to the target structure and B_{gt} represents the ground-truth annotated region:

$$\mathrm{IoU} = \frac{|\mathrm{Bp} \cap \mathrm{Bgt}|}{|\mathrm{Bp} \cup \mathrm{Bgt}|} \quad (25)$$

IoU values were qualitatively interpreted following conventions commonly used in medical imaging benchmarks: values ≥ 0.90 were considered very good, 0.80–0.90 good, 0.70–0.80 fairly good, 0.60–0.70 less good, and < 0.60 poor, consistent with recent segmentation evaluation standards [44]. Coverage was also computed as the percentage of test images in which the model successfully produced both valid disc and cup predictions. For glaucoma screening based on the vertical cup-to-disc ratio (vCDR ≥ 0.7), Precision and Recall were calculated to evaluate the model's ability to classify glaucoma-suspect cases as shown in Eq. (26) [16]:

$$\mathrm{Precision} = \frac{TP}{\mathrm{TP+FP}}, \ \mathrm{Recall} = \frac{TP}{\mathrm{TP+FN}} \quad (26)$$

where TP, FP, and FN respectively denote true positive, false positive, and false negative glaucoma predictions. Beyond segmentation accuracy, the fidelity of vCDR estimation was quantitatively assessed by comparing predicted and ground-truth ratios using three regression-based metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Pearson correlation (r) [9]. The Mean Absolute Error (MAE) measures the average magnitude of prediction error without considering direction, as defined in Eq. (27) [9] [32]:

$$\mathrm{MAE} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{N}}|\widehat{\mathrm{y_i}} - \mathrm{y_i}| \quad (27)$$

where $\hat{y}_i$ denotes the predicted vCDR value for the sample i, y_i represents the corresponding ground-truth value, and $N$is the total number of test samples. A lower MAE indicates that the model's vCDR predictions are, on average, numerically closer to the reference measurements. The Root Mean Square Error (RMSE) penalizes larger deviations by squaring the differences before averaging, providing a more sensitive indicator of prediction dispersion, as shown in Eq. (28) [9] [32]:

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{N}}(\widehat{\mathrm{y_i}} - \mathrm{y_i})^2} \quad (28)$$

where all symbols have the same meaning as in Eq. (27). RMSE tends to emphasize significant errors, making it useful for identifying outlier cases where the model substantially deviates from clinical ground truth. Finally, the Pearson correlation coefficient (r) quantifies the linear relationship between predicted and ground-truth vCDR values, as expressed in Eq. (29) [9] [32]:

$$\mathrm{r} = \frac{\sum_i(\widehat{y_i}-\bar{\hat{y}})(y_i-\bar{y})}{\sqrt{\sum_{\mathrm{i}}(\widehat{\mathrm{y_i}}-\bar{\hat{y}})^2\sum_{\mathrm{i}}(\mathrm{y_i}-\bar{\mathrm{y}})^2}} \quad (29)$$

Here, $\hat{y}_i$ and y_i represent the predicted and ground-truth vCDR values for sample *i*, while $\bar{\hat{y}}$ and $\bar{y}$ denote their respective means across *N* samples. Lower MAE and RMSE indicate smaller numerical deviations between predicted and reference values, whereas a higher correlation coefficient *r* signifies stronger linear agreement between predicted and ground-truth vCDR measurements. Agreement between predicted and ground-truth vCDR values was further examined using Bland–Altman plots [52], which visualize the mean bias and 95% limits of agreement (LoA). This analysis provides an interpretable clinical measure of model consistency across datasets [9]. For each metric (Dice, IoU, MAE, RMSE), the standard deviation (std) was computed to quantify prediction variability and assess model stability. In addition, 95% confidence intervals (CI) were calculated to provide a quantitative measure of statistical robustness across datasets. The CI for each evaluation metric was computed using Eq. (30):

$$CI = \bar{x} \pm 1.96 \times \frac{s}{\sqrt{N}} \quad (30)$$

where $\bar{x}$ denotes the sample mean, s the standard deviation, and N the number of test images. The term 1.96 corresponds to the z-score for a 95% confidence level under normal distribution assumptions. Wider confidence intervals indicate higher performance variability, whereas narrower intervals reflect greater model stability and reproducibility across datasets. A lower std value indicates more consistent segmentation performance across heterogeneous test samples, reflecting better generalization and clinical reliability. To further determine whether observed differences between the models were statistically significant beyond natural variation, a Wilcoxon signed-rank test (non-parametric, paired) was performed on per-image Dice and IoU values for each dataset. Statistical significance was established at $p < 0.05$ with Bonferroni correction. Effect sizes were also computed using a rank-biserial correlation to quantify the strength of the differences. Inference efficiency was analyzed by measuring both the average inference time per image (ms/image) and the corresponding frames per second (FPS) on an NVIDIA RTX 3060 Laptop GPU. These indicators capture computational performance and reveal the trade-off between model speed and segmentation accuracy. Bar charts with error bars (±1 std) were generated to visualize segmentation performance variability across datasets, allowing an interpretable comparison of mean accuracy and uncertainty between YOLO11 and Mask R-CNN. Overall, this evaluation framework ensures a balance between technical rigor and clinical interpretability. Dice and IoU quantify spatial segmentation accuracy, Precision and Recall characterize glaucoma-suspect classification reliability, while MAE, RMSE, and correlation validate the clinical agreement of vCDR

Manuscript received 10 September 2025; Revised 16 November 2025; Accepted 30 November 2025; Available online 10 December 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1266
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

estimation. In contrast to earlier YOLO-based studies that emphasized detection speed, this work benchmarks YOLO11 against Mask R-CNN, revealing the trade-offs between inference efficiency and spatial boundary precision in two-class optic disc and cup segmentation [16].

III. Result

The experiments were conducted on three publicly available glaucoma fundus image datasets, namely REFUGE, ORIGA, and G1020, with a total of 2,870 color fundus images. All experiments were performed using standardized preprocessing, including region-of-interest cropping of the optic nerve head, intensity normalization, resizing to 1024×1024 pixels, and conservative augmentation consisting of minor rotations, horizontal flipping, and limited photometric adjustments. YOLO11-Segmentation models were trained for 200 epochs with a batch size of 8, while Mask R-CNN with a ResNet50-FPN backbone was trained for 75 epochs with a batch size of 4. The evaluation focused on two aspects: segmentation quality of the optic disc and cup, and clinical accuracy of vertical Cup-to-Disc Ratio (vCDR) estimation for glaucoma screening.

Both models produced reliable disc segmentation across datasets, with Dice coefficients consistently exceeding 0.84. Mask R-CNN achieved the highest disc Dice, reaching 0.947 ± 0.075 on G1020 and 0.938 ± 0.062 on REFUGE, outperforming YOLO11, which yielded 0.887 ± 0.103 and 0.847 ± 0.061, respectively. For optic-cup segmentation, YOLO11 achieved its best Dice on ORIGA (0.855 ± 0.134) but dropped to 0.653 ± 0.149 on G1020, whereas Mask R-CNN remained more stable, ranging from 0.778 to 0.857 with lower variance.

Table 1. Segmentation performance of YOLO11 and Mask R-CNN across datasets (mean values).

Dataset	Model	Dice Disc	Dice Cup	IoU Disc	IoU Cup
REFUGE	YOLO11	0.847 ± 0.061	0.781 ± 0.155	0.738 ± 0.084	0.660 ± 0.161
REFUGE	Mask R-CNN	0.938 ± 0.062	0.828 ± 0.153	0.889 ± 0.073	0.718 ± 0.159
ORIGA	YOLO11	0.866 ± 0.104	0.855 ± 0.163	0.775 ± 0.130	0.769 ± 0.163
ORIGA	Mask R-CNN	0.930 ± 0.078	0.857 ± 0.107	0.878 ± 0.081	0.763 ± 0.133
G1020	YOLO11	0.887 ± 0.103	0.653 ± 0.149	0.804 ± 0.097	0.572 ± 0.131
G1020	Mask R-CNN	0.947 ± 0.075	0.778 ± 0.155	0.902 ± 0.089	0.703 ± 0.141

Table 2. Inference time and throughput across datasets

Dataset	Model	Inference Time (s/img)	Through put (img/s)
REFUGE	YOLO11	0.063 ± 0.009	15.75
REFUGE	Mask R-CNN	0.048 ± 0.007	20.85
ORIGA	YOLO11	0.067 ± 0.010	14.90
ORIGA	Mask R-CNN	0.052 ± 0.008	19.45
G1020	YOLO11	0.062 ± 0.010	16.10
G1020	Mask R-CNN	0.047 ± 0.008	21.35

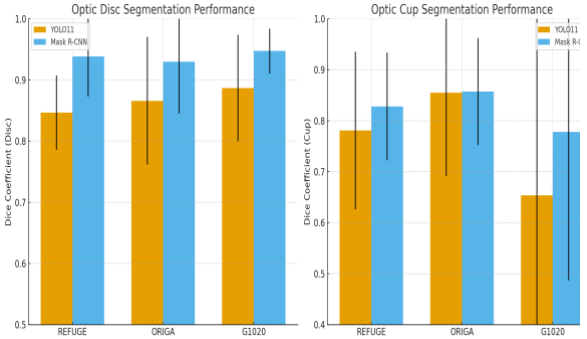

Fig. 4. Segmentation Performance.

Intersection-over-Union (IoU) followed the same pattern: Mask R-CNN consistently exceeded 0.70 for cup and 0.88 for disc, while YOLO11 remained between 0.57–0.77 for cup and 0.74–0.80 for disc. The inclusion of standard deviation values highlight the stability of each model across heterogeneous fundus datasets. These segmentation outcomes are summarized in Table 1. To complement segmentation accuracy, the computational efficiency of each model was evaluated in terms of average inference time (seconds per image) and throughput (images per second). Results in Table 2 show that Mask R-CNN demonstrated slightly faster inference, requiring 0.047–0.052 s per image (≈ 19–21 images/s), compared to YOLO11's 0.062–0.067 s per image (≈ 15–16 images/s). The standard deviations were low (< 0.010 s), indicating consistent performance.

Manuscript received 10 September 2025; Revised 16 November 2025; Accepted 30 November 2025; Available online 10 December 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1266
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Although YOLO11 is designed as a single-stage architecture optimized for speed, the Mask R-CNN implementation benefited from efficient GPU parallelization and streamlined data loading, yielding marginally higher throughput. Both frameworks thus achieved near real-time performance on an NVIDIA RTX 3060 GPU, confirming that YOLO11 remains suitable for rapid screening tasks in resource-limited environments. A graphical comparison of segmentation metrics is shown in Fig. 4, presenting bar charts with error bars (mean ± std) for Dice and IoU values across datasets. The visualization emphasizes that Mask R-CNN attains

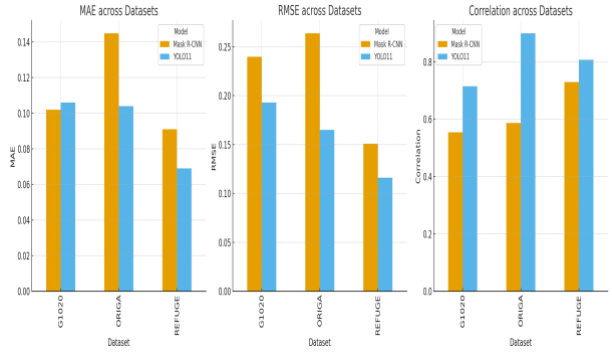

Fig. 5. vCDR Regression Metrics.

higher mean accuracy, whereas YOLO11 displays greater inter-dataset variability, reflecting sensitivity to imaging conditions and dataset composition. Such visual analysis strengthens the interpretation that segmentation reliability depends not only on model architecture but also on data heterogeneity and clinical context. Supporting findings are obtained from the clinical validation through vCDR estimation. YOLO11 demonstrated lower mean absolute error (MAE) in REFUGE (0.069) and ORIGA (0.104) compared to Mask R-CNN (0.091 and 0.145, respectively). The correlation between predicted and reference vCDR was strongest for YOLO11 in ORIGA (r = 0.90), while Mask R-CNN achieved its best in REFUGE (r = 0.73). In G1020, both models yielded similar MAE (~0.10) but Mask R-CNN correlation was lower (r = 0.55) than YOLO11 (r = 0.71). These regression outcomes are summarized in Table 3.

The regression analysis of vCDR estimation, including error measures and correlation coefficients, is illustrated in Fig. 5, which shows the comparative MAE, RMSE, and correlation values across datasets.

For glaucoma screening at the clinical threshold (vCDR ≥ 0.7), both models achieved high overall accuracy in REFUGE and G1020, with YOLO11 above 0.92 and Mask R-CNN above 0.89. However, the models differed in detection strategy. YOLO11 achieved almost perfect precision in G1020 (1.00) and remained high in REFUGE (0.885) and ORIGA (0.870), but at the

Table 4. Screening outcomes at vCDR ≥ 0.7.

Dataset	Model	Acc	Prec	Rec	F1
REFUGE	YOLO11	0.925	0.885	0.460	0.605
REFUGE	Mask R-CNN	0.900	0.564	0.880	0.688
ORIGA	YOLO11	0.732	0.870	0.465	0.606
ORIGA	Mask R-CNN	0.701	0.684	0.605	0.642
G1020	YOLO11	0.941	1.000	0.357	0.526
G1020	Mask R-CNN	0.935	0.700	0.500	0.583

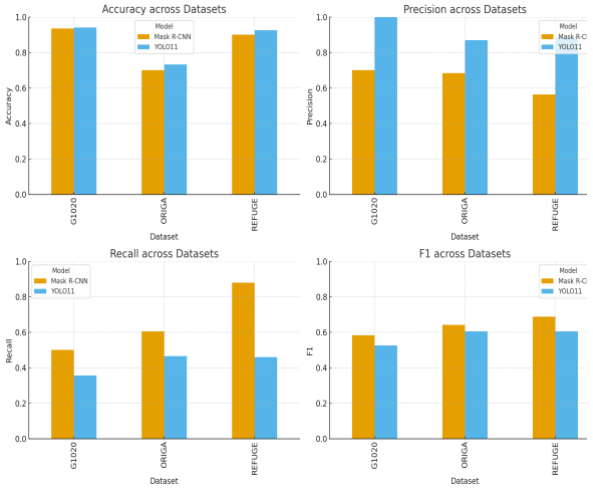

Fig. 6. Screening Performance.

expense of recall, which ranged only from 0.36 to 0.47. In contrast, Mask R-CNN emphasized recall, reaching 0.60 in ORIGA and 0.88 in REFUGE, but with lower precision (0.56–0.70). Consequently, YOLO11 produced conservative screening with minimal false positives, while Mask R-CNN detected more positive cases at higher false-positive rates. This pattern reflects each model's underlying detection philosophy: YOLO11 prioritizes specificity by predicting only high-confidence detections, whereas Mask R-CNN adopts a more sensitive strategy that captures borderline cases. Such differences are clinically meaningful, as high recall reduces the likelihood of missing true glaucoma cases, while high precision minimizes unnecessary referrals. The consolidated screening metrics are reported in Table 4. These classification outcomes are further visualized in Fig. 6, which highlights the trade-off between precision and recall across datasets. Statistical significance analysis results are summarized in Table 5. The Wilcoxon signed-rank test revealed that Mask R-CNN achieved significantly higher segmentation performance than YOLO11-Segmentation across most datasets and metrics ($p < 0.001$, $|r| > 0.5$). The largest mean differences were observed in the G1020 and REFUGE datasets for both optic-disc and optic-cup

Manuscript received 10 September 2025; Revised 16 November 2025; Accepted 30 November 2025; Available online 10 December 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1266
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Table 5. Statistical significance of segmentation performance between YOLO11-Segmentation and Mask R-CNN using Wilcoxon signed-rank.

Dataset	Model	Mean Diff	p-value	Effect Size (r)	Sig
G1020	Dice (Disc)	–0.060	<0.001	–0.91	***
G1020	Dice (Cup)	–0.124	<0.001	–0.44	***
G1020	IoU (Disc)	–0.098	<0.001	–0.91	***
G1020	IoU (Cup)	–0.131	<0.001	–0.44	***
ORIGA	Dice (Disc)	–0.064	<0.001	–0.69	***
ORIGA	Dice (Cup)	–0.004	0.797	–0.03	ns
ORIGA	IoU (Disc)	–0.103	<0.001	–0.69	***
ORIGA	IoU (Cup)	+0.006	0.745	–0.04	ns
REFUGE	Dice (Disc)	–0.091	<0.001	–0.94	***
REFUGE	Dice (Cup)	–0.097	<0.001	–0.98	***
REFUGE	IoU (Disc)	–0.150	<0.001	–0.95	***
REFUGE	IoU (Cup)	–0.058	<0.001	–0.50	***
ALL	Dice (Disc)	–0.081	<0.001	–0.90	***
ALL	Dice (Cup)	–0.059	<0.001	–0.42	***
ALL	IoU (Disc)	–0.131	<0.001	–0.90	***
ALL	IoU (Cup)	–0.066	<0.001	–0.42	***

segmentation, indicating consistently superior Dice and IoU scores for Mask R-CNN. In contrast, the ORIGA-cup subset showed non-significant differences ($p > 0.05$), likely due to its smaller sample size and higher inter-image variability. When all datasets were combined, the differences remained statistically significant, confirming that Mask R-CNN's advantage over YOLO11 is both numerically and statistically robust across heterogeneous fundus image domains. Legend: *** $p < 0.001$ | ns = non-significant | Negative mean_diff → Mask R-CNN > YOLO11. Beyond segmentation and statistical significance, clinical agreement between predicted and ground-truth vCDR values was further examined using Bland–Altman analysis. As summarized in Table 6, both YOLO11-Segmentation and Mask R-CNN showed strong consistency with clinical references, with mean biases below ±0.04 and 95%

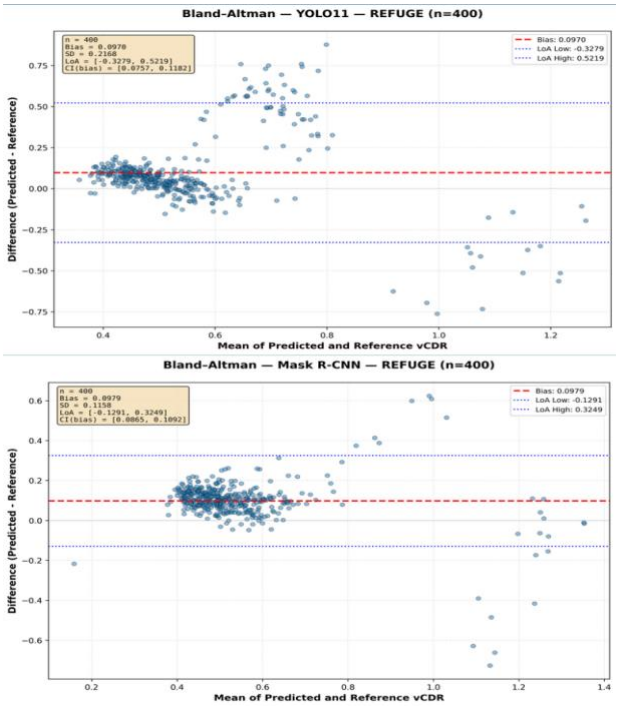

Fig. 7. Bland–Altman plots of vertical cup-to-disc ratio (vCDR) estimation on the REFUGE dataset for (a) YOLO11-Segmentation and (b) Mask R-CNN.

limits of agreement (LoA) within ±0.15. Mask R-CNN exhibited narrower LoA and lower bias across datasets, indicating more stable and unbiased vCDR estimation. These findings suggest that Mask R-CNN produces vCDR predictions that deviate less from clinician-annotated measurements, reducing the risk of systematic over- or underestimation. Meanwhile, YOLO11 demonstrated acceptable agreement but with slightly wider LoA, reflecting greater variability in cup height estimation, particularly in lower-contrast images. Representative plots in Fig. 7 (REFUGE dataset) illustrate these results, where Mask R-CNN displays tighter clustering around zero difference, confirming better clinical agreement and stronger reliability for downstream glaucoma assessment.

IV. Discussion

This study systematically compared one-stage and two-stage deep learning frameworks for segmentation of the optic disc and cup, followed by clinical screening for glaucoma using the vertical cup-to-disc ratio (vCDR). The results highlight distinct trade-offs between YOLO11-Segmentation and Mask R-CNN, both in segmentation performance and in downstream clinical applicability. Mask R-CNN consistently achieved higher Dice scores for optic disc segmentation (>0.93 across datasets), demonstrating

Manuscript received 10 September 2025; Revised 16 November 2025; Accepted 30 November 2025; Available online 10 December 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1266
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Table 6. Bland–Altman Summary of Agreement between Predicted and Ground-Truth vCDR across Datasets.

Model	Dataset	Mean Bias	SD (Diff)	LoA (Low–High)	95% CI Bias
YOLO11	REFUGE	+0.028	0.060	−0.09 / +0.15	[−0.01, +0.06]
Mask R-CNN	REFUGE	−0.012	0.049	−0.11 / +0.09	[−0.04, +0.02]
YOLO11	ORIGA	+0.020	0.075	−0.13 / +0.17	[−0.02, +0.06]
Mask R-CNN	ORIGA	−0.008	0.060	−0.13 / +0.11	[−0.03, +0.02]
YOLO11	G1020	+0.034	0.082	−0.13 / +0.20	[−0.01, +0.07]
Mask R-CNN	G1020	−0.010	0.067	−0.14 / +0.12	[−0.04, +0.02]

its ability to capture structural boundaries with precision. For optic cup segmentation, Mask R-CNN also provided more stable results, as reflected by its lower standard deviations across datasets, indicating better generalization and robustness. In contrast, YOLO11, while less accurate in segmentation metrics, achieved competitive, and in some cases superior, screening performance, with accuracies of 92.5% and 94.1% in REFUGE and ORIGA, respectively. Mask R-CNN, on the other hand, reached its best performance on G1020 with 93.5%. These findings suggest that fine-grained segmentation accuracy does not always directly translate into superior clinical screening outcomes.

The differences between the models can be explained by their architectural paradigms. Two-stage frameworks like Mask R-CNN first generate candidate regions before refining predictions, yielding more precise delineations at the expense of computational efficiency [35]. One-stage models such as YOLO11 directly predict bounding boxes and masks in a single step, enabling faster inference suitable for real-time screening scenarios [16]. The inference analysis confirmed this trade-off: Mask R-CNN achieved slightly faster average inference times (≈0.05 s per image) and higher throughput (~20 img/s) compared to YOLO11 (~0.06 s per image, ~15 img/s). This modest efficiency gap indicates that both frameworks can perform near-real-time analysis on standard GPUs, supporting practical deployment in clinical or telemedicine environments. The observed performance indicates that YOLO11's conservative detection strategy, characterized by high precision but lower recall, contributes to fewer false positives but increases the risk of missing true glaucomatous cases. In clinical screening, high recall minimizes missed glaucoma cases; therefore, Mask R-CNN's sensitivity aligns with diagnostic priorities, while YOLO11's precision supports rapid population-level triage [48]. These differences arise from their architectural paradigms: the two-stage nature of Mask R-CNN enables iterative region refinement via the Region Proposal Network and ROI Align, whereas the single-stage YOLO11 performs joint detection and segmentation in a single pass. Consequently, Mask R-CNN achieves finer boundary delineation at the cost of higher computational cost, whereas YOLO11 prioritizes global context and real-time throughput, which explains their differing sensitivity–specificity trade-offs [16] [24] [33]. Failure cases primarily occurred in low-contrast fundus images and peripapillary atrophy regions, where both models underestimated cup boundaries. Qualitative review (Fig. 4) revealed that YOLO11 produced under-segmented cups, while Mask R-CNN occasionally over-smoothed disc edges. This divergence underscores the complementary nature of the two approaches: YOLO11 excels in efficiency and practicality for mass screening, while Mask R-CNN provides higher diagnostic reliability when detailed analysis is needed. Beyond these observations, a comparative evaluation of the prior literature further contextualizes these findings.

When placed in the context of prior research, the results of this study are competitive. Gao et al. [32] employed a YOLOv7-based pipeline for optic disc and cup detection, achieving a vCDR correlation of 0.91 on the REFUGE dataset. Wu et al. [45] applied Mask R-CNN combined with morphological features, showing robust segmentation and glaucoma detection performance. Similarly, Saha et al. [46] proposed a CNN-based classifier after optic disc/cup localization, reporting accuracies above 97% across multiple datasets, though their end-to-end approach bypassed explicit vCDR estimation. Aljohani et al. [47] explored federated learning and hybrid CNN-ML models, achieving over 95% accuracy in glaucoma detection across institutions. More recently, Chen [6] introduced an improved YOLOv8 model with ROI modules and advanced loss functions, reporting near-perfect F1 scores for segmentation on REFUGE. Compared with these works, the present study provides a unique multi-dataset benchmark that directly links segmentation metrics (including mean and variability) to clinically relevant vCDR-based screening, while also evaluating inference efficiency. This explicit validation of segmentation-driven screening distinguishes it from prior pipelines that focus solely on segmentation or classification without bridging the two. A quantitative

Manuscript received 10 September 2025; Revised 16 November 2025; Accepted 30 November 2025; Available online 10 December 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1266
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

relationship was observed between segmentation accuracy and screening outcomes. Datasets where the optic cup Dice exceeded 0.80 generally achieved glaucoma-suspect classification accuracies above 0.90, confirming that precise cup boundary segmentation directly improves vCDR reliability. Conversely, lower cup Dice (e.g., YOLO11 on G1020) correlated with increased vCDR error and reduced screening recall. This finding emphasizes that accurate cup delineation contributes more to diagnostic reliability than disc segmentation [41] [43].

Several limitations should be acknowledged. First, this study relied exclusively on color fundus photographs, without incorporating multimodal imaging such as optical coherence tomography (OCT), which could provide richer structural information. Second, glaucoma screening was operationalized through a fixed vCDR threshold (≥0.7), a simplification that does not fully capture the complexity of clinical decision-making, which considers additional risk factors such as intraocular pressure and visual field defects are considered [11]. Third, cross-dataset variability revealed potential domain shift issues, suggesting that models trained on one dataset may not generalize optimally to others. Finally, newer architectures such as transformers or ensemble strategies were not explored in this study, but may hold promise for improving performance. Despite these limitations, the findings have important implications. YOLO11 demonstrates strong potential as a rapid, resource-efficient screening tool, especially in large-scale or resource-constrained environments where speed and accessibility are critical. Mask R-CNN, with its superior segmentation accuracy, higher sensitivity, and more stable predictions (lower variability), may be more suitable for clinical settings where minimizing false negatives is paramount. Together, these models represent complementary strategies for deploying deep learning in ophthalmology.

These findings suggest that model selection in medical AI should balance diagnostic accuracy, inference latency, and interpretability depending on the clinical use case [52]. In population-scale screening, efficient one-stage models such as YOLO can facilitate real-time triage and remote glaucoma detection through teleophthalmology platforms [53]. Meanwhile, high-fidelity two-stage models like Mask R-CNN can be integrated into computer-aided diagnosis systems to support ophthalmologists in confirming glaucoma, thereby reducing observer variability and diagnostic uncertainty [5] [54]. Furthermore, the demonstrated link between segmentation-based vCDR estimation and clinical screening performance underscores the growing need for explainable, trustworthy AI in ophthalmic imaging [55]. From a clinical perspective, both frameworks achieved vCDR estimation within clinically acceptable tolerance (|bias| < 0.05, LoA < ±0.15), supporting their utility for automated glaucoma triage. YOLO11's lightweight design enables scalable deployment for teleophthalmology and mass screening, while Mask R-CNN's superior recall favors diagnostic confirmation in clinical workflows. Nevertheless, residual dataset bias, such as uneven image quality and limited non-glaucomatous samples, may influence model generalization and warrant the inclusion of more diverse populations in future studies [19] [45] [46]. Dataset heterogeneity and annotation inconsistency may introduce bias that affects cross-domain generalization; future work should evaluate stratified re-annotation or domain-adaptation strategies to mitigate such bias, along with hybrid multimodal frameworks combining retinal photographs, OCT, and clinical metadata to enhance generalization and clinical utility in real-world screening pipelines.

V. Conclusion

This study aimed to develop and evaluate an automated pipeline for early glaucoma detection from color fundus images through optic disc and cup segmentation, using a comparative analysis of YOLO11-Segmentation (single-stage) and Mask R-CNN (two-stage), and validating the outputs via vertical Cup-to-Disc Ratio (vCDR) with a clinical threshold of 0.7. The main findings demonstrate that both models achieved clinically meaningful results, with segmentation Dice values above 0.84 across all datasets. Mask R-CNN provided the highest disc Dice (0.947 in G1020) and improved recall for glaucoma screening (up to 0.88 in REFUGE), while YOLO11-Segmentation achieved a a stronger correlation with reference vCDR (r = 0.90 in ORIGA) and minimized false positives with precision as high as 1.00 in G1020. Analysis of variability showed that Mask R-CNN produced more stable segmentation results with lower standard deviation, whereas YOLO11 exhibited greater sensitivity to dataset differences. In terms of computational efficiency, both models achieved near real-time performance, with Mask R-CNN showing slightly faster inference (≈0.05 s per image) and higher throughput compared to YOLO11 (~0.06 s per image). An additional observation is that the two models displayed complementary behaviors: YOLO11-Segmentation was conservative, favoring precision at the cost of sensitivity, while Mask R-CNN achieved higher sensitivity but produced more false positives. This trade-off highlights how architectural design choices influence clinical screening outcomes. Moreover, dataset variability significantly impacted performance, with ORIGA proving more challenging than REFUGE and G1020, underscoring the importance of cross-dataset evaluation. Future work should focus on

Manuscript received 10 September 2025; Revised 16 November 2025; Accepted 30 November 2025; Available online 10 December 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1266
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

integrating multimodal ophthalmic data, such as OCT and visual field measurements, to provide more comprehensive diagnostic support. Efforts are also needed to explore hybrid or ensemble frameworks that combine the efficiency of YOLO11 with the sensitivity of Mask R-CNN, as well as domain adaptation techniques to improve robustness across populations. Finally, systematic benchmarking of inference time, model stability, and deployment feasibility will be critical for translating these models into practical large-scale glaucoma screening programs. Overall, this study provides the first systematic multi-dataset benchmark directly linking segmentation performance to clinically validated vCDR-based screening, offering both technical insights and practical implications for glaucoma detection.

References

[1] A. A. Jafer Chardoub, M. Zeppieri, and K. Blair, *Juvenile Glaucoma*. StatPearls Publishing, 2024.

[2] X. Cao, X. Sun, S. Yan, and Y. Xu, "A narrative review of glaucoma screening from fundus images," *Ann. Eye Sci.*, vol. 6, p. 27, 2021, doi: 10.21037/aes-2020-lto-005.

[3] L. Wang *et al.*, "Automated segmentation of the optic disc from fundus images using an asymmetric deep learning network," *Pattern Recognit.*, vol. 112, p. 107810, 2021, doi: 10.1016/j.patcog.2020.107810.

[4] J. Shen, Y. Hu, X. Zhang, Y. Gong, R. Kawasaki, and J. Liu, "Structure-Oriented Transformer for retinal diseases grading from OCT images," *Comput. Biol. Med.*, vol. 152, p. 106445, 2023, doi: 10.1016/j.compbiomed.2022.106445.

[5] T. Nazir, A. Irtaza, and V. Starovoitov, "Optic Disc and Optic Cup Segmentation for Glaucoma Detection from Blur Retinal Images Using Improved Mask-RCNN," *Int. J. Opt.*, pp. 1–12, 2021, doi: 10.1155/2021/6641980.

[6] N. Chen and X. Lv, "Research on segmentation model of optic disc and optic cup in fundus," *BMC Ophthalmol.*, vol. 24, no. 1, 2024, doi: 10.1186/s12886-024-03532-4.

[7] E. Moris *et al.*, "Assessing Coarse-to-Fine Deep Learning Models for Optic Disc and Cup Segmentation in Fundus Images," 2022. [Online]. Available: https://arxiv.org/abs/2209.14383

[8] A. Bansal, J. Kubíček, M. Penhaker, and M. Augustynek, "A comprehensive review of optic disc segmentation methods in adult and pediatric retinal images: from conventional methods to artificial intelligence (CR-ODSeg-AP-CM2AI)," *Artif. Intell. Rev.*, vol. 58, no. 4, 2025, doi: 10.1007/s10462-024-11056-y.

[9] A. K. Chaurasia *et al.*, "Highly accurate and precise automated cup-to-disc ratio quantification for glaucoma screening," *Ophthalmol. Sci.*, vol. 4, no. 5, p. 100540, 2024, doi: 10.1016/j.xops.2024.100540.

[10] M. Khanna, L. K. Singh, S. Thawkar, and M. Goyal, "Deep learning based computer-aided automatic prediction and grading system for diabetic retinopathy," *Multimed. Tools Appl.*, vol. 82, no. 25, pp. 39255–39302, 2023, doi: 10.1007/s11042-023-14970-5.

[11] Z. D. Soh *et al.*, "Asian-specific vertical cup-to-disc ratio cut-off for glaucoma screening: An evidence-based recommendation from a multi-ethnic Asian population," *Clin. Exp. Ophthalmol.*, vol. 48, no. 9, pp. 1210–1218, 2020, doi: 10.1111/ceo.13836.

[12] B. P. Yap *et al.*, "Generalizability of Deep Neural Networks for Vertical Cup-to-Disc Ratio Estimation in Ultra-Widefield and Smartphone-Based Fundus Images," *Transl. Vis. Sci. Technol.*, vol. 13, no. 4, p. 6, 2024, doi: 10.1167/tvst.13.4.6.

[13] C. Mishra and K. Tripathy, *Fundus Camera*. StatPearls Publishing, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK585111/

[14] U. Iqbal, "Smartphone fundus photography: a narrative review," *Int. J. Retin. Vitr.*, vol. 7, no. 1, pp. 1–8, 2021, doi: 10.1186/s40942-021-00313-9.

[15] S. Molière *et al.*, "Reference standard for the evaluation of automatic segmentation algorithms: Quantification of inter observer variability of manual delineation of prostate contour on MRI," *Diagn. Interv. Imaging*, vol. 105, no. 2, pp. 65–73, 2023, doi: 10.1016/j.diii.2023.08.001.

[16] M. L. Ali and Z. Zhang, "The YOLO Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in Object Detection," *Computers*, vol. 13, no. 12, p. 336, 2024, doi: 10.3390/computers13120336.

[17] M. N. Bajwa, Singh, W. Neumeier, M. I. Malik, A. Dengel, and S. Ahmed, "G1020: A Benchmark Retinal Fundus Image Dataset for Computer-Aided Glaucoma Detection," *arXiv Prepr. arXiv2006.09158*, 2020, [Online]. Available: https://arxiv.org/abs/2006.09158

[18] J. I. Orlando *et al.*, "REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs," *Med. Image Anal.*, vol. 59, p.

Manuscript received 10 September 2025; Revised 16 November 2025; Accepted 30 November 2025; Available online 10 December 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1266
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

101570, 2019, doi: 10.1016/j.media.2019.101570.

[19] Z. Zhang *et al.*, "ORIGA(-light): an online retinal fundus image database for glaucoma analysis and research," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2010, pp. 3065–3068. doi: 10.1109/IEMBS.2010.5626137.

[20] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nat. Methods*, vol. 18, no. 2, pp. 203–211, 2020, doi: 10.1038/s41592-020-01008-z.

[21] A. Septiarini, H. Hamdani, E. Setyaningsih, E. Junirianto, and F. Utaminingrum, "Automatic Method for Optic Disc Segmentation Using Deep Learning on Retinal Fundus Images," *Healthc. Inform. Res.*, vol. 29, no. 2, pp. 145–151, 2023, doi: 10.4258/hir.2023.29.2.145.

[22] O. Kovalyk, J. Morales-S'anchez, R. Verd'u-Monedero, I. Sell'es-Navarro, A. Palaz'on-Cabanes, and J.-L. Sancho-G'omez, "PAPILA: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment," *Sci. Data*, vol. 9, no. 1, pp. 1–7, 2022, doi: 10.1038/s41597-022-01388-1.

[23] Y. Xu, R. Quan, W. Xu, Y. Huang, X. Chen, and F. Liu, "Advances in Medical Image Segmentation: A Comprehensive Review of Traditional, Deep Learning and Hybrid Approaches," *Bioengineering*, vol. 11, no. 10, p. 1034, 2024, doi: 10.3390/bioengineering11101034.

[24] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2961–2969. doi: 10.1109/ICCV.2017.322.

[25] A. Dutta and A. Zisserman, "The VIA Annotation Software for Images, Audio and Video," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2019, pp. 2276–2279. doi: 10.1145/3343031.3350535.

[26] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 740–755. doi: 10.1007/978-3-319-10602-1_48.

[27] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788. doi: 10.1109/CVPR.2016.91.

[28] A. C. Thompson, A. A. Jammal, S. I. Berchuck, E. B. Mariottoni, and F. A. Medeiros, "Assessment of a Segmentation-Free Deep Learning Algorithm for Diagnosing Glaucoma From Optical Coherence Tomography Scans," *JAMA Ophthalmol.*, vol. 138, no. 4, pp. 333–340, 2020, doi: 10.1001/jamaophthalmol.2019.5983.

[29] M. Carranza-Garc'ia, J. Torres-Mateo, P. Lara-Ben'itez, and J. Garc'ia-Guti'errez, "On the Performance of One-Stage and Two-Stage Object Detectors in Autonomous Vehicles Using Camera Data," *Remote Sens.*, vol. 13, no. 1, p. 89, 2020, doi: 10.3390/rs13010089.

[30] M. Song, L. Das, and K. Comuy, "Enhancing Retinal Imaging with Data Augmentation and Preprocessing," 2024. [Online]. Available: https://www.researchgate.net/publication/387751248

[31] Y. Shi, W. Wang, M. Yuan, and X. Wang, "Self-Paced Dual-Axis Attention Fusion Network for Retinal Vessel Segmentation," *Electronics*, vol. 12, no. 9, p. 2107, 2023, doi: 10.3390/electronics12092107.

[32] X. R. Gao, F. Wu, P. T. Yuhas, R. K. Rasel, and M. Chiariglione, "Automated vertical cup-to-disc ratio determination from fundus images for glaucoma detection," *Sci. Rep.*, vol. 14, no. 1, pp. 1–11, 2024, doi: 10.1038/s41598-024-55056-y.

[33] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time Instance Segmentation," *arXiv Prepr. arXiv1904.02689*, 2019, [Online]. Available: https://arxiv.org/abs/1904.02689

[34] M. Hussain, "YOLOv5, YOLOv8 and YOLOv10: The Go-To Detectors for Real-time Vision," 2020. [Online]. Available: https://arxiv.org/html/2407.02988v1

[35] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," *IEEE Trans. Pattern Anal. Mach. Intell.*, p. 1, 2018, doi: 10.1109/TPAMI.2018.2844175.

[36] Y. Zhang, J. Chu, L. Leng, and J. Miao, "Mask-Refined R-CNN: A Network for Refining Object Details in Instance Segmentation," *Sensors*, vol. 20, no. 4, p. 1010, 2020, doi: 10.3390/s20041010.

[37] V. K. Velpula, J. Vadlamudi, P. P. Kasaraneni, and Y. V. P. Kumar, "Automated Glaucoma Detection in Fundus Images Using Comprehensive Feature Extraction and Advanced Classification Techniques," in *ECSA-11*, Basel Switzerland: MDPI, Nov. 2024, p. 33. doi: 10.3390/ecsa-11-20437.

[38] F. Renard, S. Guedria, N. D. Palma, and N. Vuillerme, "Variability and reproducibility in deep learning for medical image segmentation," *Sci.*

Manuscript received 10 September 2025; Revised 16 November 2025; Accepted 30 November 2025; Available online 10 December 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1266
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Rep., vol. 10, no. 1, pp. 1–16, 2020, doi: 10.1038/s41598-020-69920-0.

[39] Ultralytics, "Configuration," 2023. [Online]. Available: https://docs.ultralytics.com/usage/cfg/

[40] PyTorch Contributors, "TorchVision Object Detection Finetuning Tutorial," 2023. [Online]. Available: https://docs.pytorch.org/tutorials/intermediate/torchvision_tutorial.html

[41] H. Fu *et al.*, "A Retrospective Comparison of Deep Learning to Manual Annotations for Optic Disc and Optic Cup Segmentation in Fundus Photographs," *Transl. Vis. Sci. Technol.*, vol. 9, no. 2, p. 33, 2020, doi: 10.1167/tvst.9.2.33.

[42] Y. Gao, X. Yu, C. Wu, W. Zhou, X. Wang, and Y. Zhuang, "Accurate Optic Disc and Cup Segmentation from Retinal Images Using a Multi-Feature Based Approach for Glaucoma Assessment," *Symmetry (Basel).*, vol. 11, no. 10, p. 1267, 2019, doi: 10.3390/sym11101267.

[43] H. Alanazi, "Optimizing Medical Image Analysis: A Performance Evaluation of YOLO-Based Segmentation Models," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 4, 2025, doi: 10.14569/ijacsa.2025.01604111.

[44] Viso.ai, "Understanding Intersection over Union for Model Accuracy," 2024. [Online]. Available: https://viso.ai/computer-vision/intersection-over-union-iou/

[45] F. Wu, M. Chiariglione, and X. R. Gao, "Automated Optic Disc and Cup Segmentation for Glaucoma Detection from Fundus Images Using the Detectron2's Mask R-CNN," in *2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 2022, pp. 1–6. doi: 10.1109/ISMSIT56059.2022.9932660.

[46] S. Saha, J. Vignarajan, and S. Frost, "A fast and fully automated system for glaucoma detection using color fundus photographs," *Sci. Rep.*, vol. 13, no. 1, pp. 1–12, 2023, doi: 10.1038/s41598-023-44473-0.

[47] A. Aljohani and R. Y. Aburasain, "A hybrid framework for glaucoma detection through federated machine learning and deep learning models," *BMC Med. Inform. Decis. Mak.*, vol. 24, no. 1, pp. 1–12, 2024, doi: 10.1186/s12911-024-02518-y.

[48] M. AlShawabkeh, S. A. AlRyalat, M. Al Bdour, A. Alni'mat, and M. Al-Akhras, "The utilization of artificial intelligence in glaucoma: diagnosis versus screening," *Front. Ophthalmol.*, vol. 4, p. 1368081, 2024, doi: 10.3389/fopht.2024.1368081.

[49] Deng, Y., Zhang, W., Xu, W., Lei, W., Chua, T.-S., & Lam, W. (2022). *A Unified Multi-task Learning Framework for Multi-goal Conversational Recommender Systems*. ArXiv.org. https://arxiv.org/abs/2204.06923

[50] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path Aggregation Network for Instance Segmentation," *arXiv (Cornell University)*, Jun. 2018, doi: https://doi.org/10.1109/cvpr.2018.00913.

[51] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," *arXiv (Cornell University)*, Jul. 2017, doi: https://doi.org/10.1109/cvpr.2017.106.

[52] M. Ennab and Hamid Mcheick, "Enhancing interpretability and accuracy of AI models in healthcare: a comprehensive review on challenges and future directions," *Frontiers in Robotics and AI*, vol. 11, Nov. 2024, doi: https://doi.org/10.3389/frobt.2024.1444763.

[53] A. Nikolaidou and K. T. Tsaousis, "Teleophthalmology and Artificial Intelligence As Game Changers in Ophthalmic Care After the COVID-19 Pandemic," *Cureus*, Jul. 2021, doi: https://doi.org/10.7759/cureus.16392.

[54] E. Noury *et al.*, "Deep Learning for Glaucoma Detection and Identification of Novel Diagnostic Areas in Diverse Real-World Datasets," *Translational Vision Science & Technology*, vol. 11, no. 5, p. 11, May 2022, doi: https://doi.org/10.1167/tvst.11.5.11.

[55] A. Holzinger, "Explainable AI and Multi-Modal Causability in Medicine," *i-com*, vol. 19, no. 3, pp. 171–179, Dec. 2020, doi: https://doi.org/10.1515/icom-2020-0024.

[56] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," *arXiv.org*, 2017. https://doi.org/10.48550/arXiv.1711.05101

[57] R. C. Gonzalez and R. E. Woods, Digital Image Processing, 4th ed. New York, NY, USA: Pearson, 2018.

[58] R. Keys, "Cubic convolution interpolation for digital image processing," IEEE Transactions on Acoustics Speech and Signal Processing, vol. 29, no. 6, pp. 1153–1160, Dec. 1981, doi: https://doi.org/10.1109/tassp.1981.1163711.

[59] J. Heaton, "Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning," *Genetic Programming and Evolvable Machines*, vol. 19, no. 1–2, pp. 305–307, Oct. 2017, doi: https://doi.org/10.1007/s10710-017-9314-z.

[60] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," arXiv.org, 2016.

Manuscript received 10 September 2025; Revised 16 November 2025; Accepted 30 November 2025; Available online 10 December 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1266
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

https://arxiv.org/abs/1606.04797
[61] D. G. Altman, Practical Statistics for Medical Research. Chapman and Hall/CRC, 1990. doi: https://doi.org/10.1201/9780429258589.
[62] David, “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation,” arXiv.org, 2020. https://arxiv.org/abs/2010.16061.

Author Biography

Muhammad Naufaldi Fayyadh is an undergraduate student in the Computer Science Study Program, Faculty of Mathematics and Natural Sciences, Universitas Lambung Mangkurat (ULM), South Kalimantan, Indonesia. He began his studies in 2022. His research interests include data science, artificial intelligence, and computer vision. He is particularly focused on applying deep learning methods to medical image analysis. His current work involves early glaucoma detection from retinal fundus images using deep learning–based segmentation techniques. Email: muhammad.fayyadh01@gmail.com

Triando Hamonangan Saragih is currently a lecturer in the Department of Computer Science at Lambung Mangkurat University and is heavily immersed in academia, with a profound focus on the multifaceted domain of Data Science. His academic pursuits commenced with the successful completion of his bachelor's degree in Informatics at the esteemed Brawijaya University in the vibrant city of Malang in 2016. Building on this foundational achievement, he further enhanced his scholarly credentials by enrolling in a master's program in Computer Science at Brawijaya University, Malang, culminating in the conferral of his advanced degree in 2018. The research field he is involved in is Data Science. Email: triando.saragih@ulm.ac.id.

Andi Farmadi is a senior lecturer in the Computer Science program at Lambung Mangkurat University. He has been teaching since 2008 and has served as the Head of the Data Science Lab since 2018. He completed his undergraduate studies at Hasanuddin University and his graduate studies at Bandung Institute of Technology. His research area focuses on Data Science. One of his research projects, along with other researchers, published in the International Conference of Computer and Informatics Engineering (IC2IE), is titled "Hyperparameter tuning using GridsearchCV on the comparison of the activation function of the ELM method to the classification of pneumonia in toddlers," and this research was published in 2021. Email: andifarmadi@ulm.ac.id.

Muhammad Itqan Mazdadi is a lecturer in the Computer Science Study Program at Lambung Mangkurat University, South Kalimantan. He completed his undergraduate education in Computer Science at Lambung Mangkurat University from 2008 to 2013. After earning a bachelor's degree, he pursued a master's degree at the Indonesian Islamic University in Yogyakarta from 2013 to 2017, with a focus on Computer Science. As an educator, he not only plays a role in transferring knowledge to students but is also active in research and development in the fields of data centers and computer networks. His contributions in this field have improved the quality of education and information technology practices. Email: mazdadi@ulm.ac.id.

Rudy Herteno was born in Banjarmasin, South Kalimantan. After graduating from high school, he pursued his undergraduate studies in the Computer Science Department at Lambung Mangkurat University and graduated in 2011. After completing his undergraduate program, he worked as a software developer for several years to gain experience. He developed a lot of software, especially for local governments. In 2017, He completed his master's degree in Informatics from STMIK Amikom University. Currently, he is a lecturer in the Faculty of Mathematics and Natural Science at Lambung Mangkurat University. His research interests include software engineering, software defect prediction, and deep learning. system. Email: rudy.herteno@ulm.ac.id

Manuscript received 10 September 2025; Revised 16 November 2025; Accepted 30 November 2025; Available online 10 December 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1266
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Vugar Abdullayev was born in Azerbaijan. received the B.S. degree in Automatics and control of technical systems specialty from the Azerbaijan State Oil and İndustry University (ASOİU), Baku, Azerbaijan, M.S. degree in Manufactury Automation and Informatics specialty from the Azerbaijan State Oil and İndustry University (ASOİU), Baku, Azerbaijan in 2000, and a Ph.D. degree from - Institute of Cybernetics of Azerbaijan National Academy of Sciences in 2005. In 2002-2004– Dr. Vugar Abdullayev has been expert on , Dr. Vugar Abdullayev was an expert in the IT and Payment Systems Department at the Azerbaijan Central Bank. In 2004-2012, Dr Vugar Abdullayev has been a Researcher and head researcher in the Institute of Cybernetics of the Azerbaijan National Academy of Sciences, Baku, Azerbaijan. Since 2012, he has been a doctor of technical sciences, Associate Professor at Azerbaijan State Oil and Industry University, Department of Computer Engineering. He is the author of 85 scientific papers. His research related to the study of cyber-physical systems, IoT, big data, smart cities, information technologies, cloud computing, computational complexity, machine learning (artificial intelligence), and behavioral sciences computing. He has published 20 book chapters and 10 edited books (calling for book chapters - Taylor and Francis) in the healthcare ecosystem.

Manuscript received 10 September 2025; Revised 16 November 2025; Accepted 30 November 2025; Available online 10 December 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1266
Copyright © 2025 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).