

Classifying and Segmenting Multiple Ocular Diseases Using Deep Learning Based on Retinal Imaging

1st Felipe M. Panugan III

School of Information Technology

Mapúa University

Makati, Philippines

fimpanugan@mymail.mapua.edu.ph

2nd Liandro E. Refulle

School of Information Technology

Mapúa University

Makati, Philippines

lerefulle@mymail.mapua.edu.ph

3rd Prince Jeffery Villamil

School of Information Technology

Mapúa University

Makati, Philippines

pjvillamil@mymail.mapua.edu.ph

4th Nicko Gabriel Baldo

School of Information Technology

Mapúa University

Makati, Philippines

ngabaldo@mymail.mapua.edu.ph

Abstract—This study presents a deep learning-based framework for the classification and segmentation of multiple ocular diseases using retinal images. The system employs the YOLOv12 segmentation architecture, trained on the RetinaVision dataset composed of curated fundus images annotated for cataract, age-related macular degeneration (AMD), and pathologic myopia. The dataset was divided into an 80-10-10 split for training, validation, and testing, respectively. After 606 epochs, the model achieved a mean Average Precision (mAP@0.5) of 0.93 and mAP@0.5:0.95 of 0.767, demonstrating both excellent localization accuracy and strong generalization across ocular disease classes. These results highlight YOLOv12's capacity for simultaneous detection and segmentation, indicating its potential for real-time ophthalmic screening applications.

Index Terms—neural network, segmentation, classification, retinal imaging

I. INTRODUCTION

Ocular diseases remain a leading cause of vision impairment worldwide, affecting millions annually. Conditions such as cataract, AMD, and pathologic myopia often progress silently, underscoring the importance of early detection. Traditional diagnosis relies on manual interpretation of retinal fundus images by ophthalmologists, which can be slow and subjective [4], [5], [9], [12]. Automated image analysis using deep learning (DL) has emerged as a powerful diagnostic complement, capable of accelerating screening processes and improving consistency [4], [5], [9], [12].

Earlier research demonstrated strong results using convolutional neural networks (CNNs) and encoder-decoder architectures such as U-Net and SegNet. However, these models struggle to balance segmentation accuracy with inference speed, limiting their clinical usability. The YOLO (You Only Look Once) family of

models revolutionized real-time object detection, and subsequent variants (YOLOv5–v8) introduced segmentation heads to localize fine structures. YOLOv12, introduced in 2024, integrates attention-based feature extraction and cross-scale fusion, offering high accuracy and efficiency [11].

This paper applies YOLOv12 to multi-disease retinal imaging using the custom-curated RetinaVision dataset. It aims to evaluate the model's ability to perform both classification and segmentation simultaneously and to identify design and training strategies that optimize generalization in medical imaging contexts.

II. METHODOLOGY

A. Dataset Acquisition and Preparation

1) Data Source and Filtering

The study employed a large collection of retinal fundus images consolidated from multiple publicly available datasets cited in the reference section, including the Mendeley Combined Fundus Dataset [10] and several Kaggle repositories [6], [7], [8]. These datasets collectively contain labeled fundus images depicting various ocular pathologies, ensuring diverse image quality and clinical representation [6], [7], [8], [10].

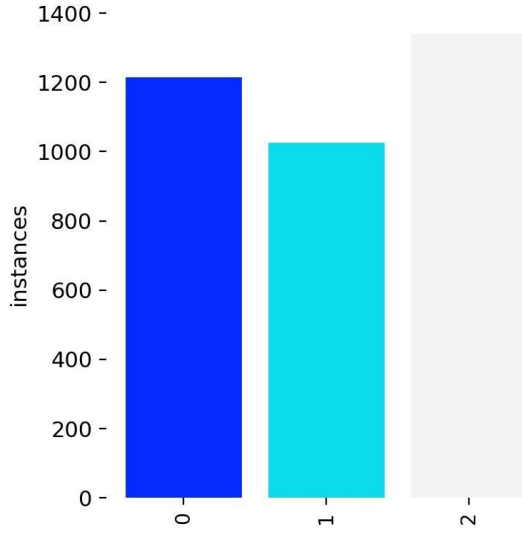


Fig. 1. YOLOv12-seg model architecture

The combined dataset contains around 3,000 images in total, distributed into around 1,000 images for class balancing shown in Fig. 1. To align with the study's diagnostic objectives, the data was filtered to focus on three primary diseases: Age-related Macular Degeneration (AMD) (Label 0), Pathologic Myopia (Label 1), and Cataract (Label 2). All samples associated with other ocular conditions, such as hypertensive retinopathy, were excluded to maintain a targeted and clinically coherent dataset. This filtering approach ensured that the model's training remained specific to the intended pathologies.

2) Custom Annotation and Class Definition

As the public datasets contained only image-level disease labels, a custom annotation pipeline was implemented to produce precise pixel-level segmentation masks and bounding boxes [1], [2], [3]. This process defined three distinct object classes:

- **AMD (Lesions):** Annotated targets included *Geographic Atrophy (GA)*, *Drusen*, and *Exudates*. These lesions are critical indicators for assessing AMD progression and treatment efficacy [4], [5].
- **Pathologic Myopia (PPA):** The *Peripapillary Atrophy (PPA)* region was segmented as a biomarker of retinal deformation caused by excessive axial elongation [9].
- **Cataract (Blur Artifact):** Regions affected by *lens opacity* and *light scattering* were annotated to represent image degradation due to cataracts, aiding in the assessment of image quality and

potential diagnostic obstruction [12].

This manual annotation procedure ensured reliable ground truth data for both object detection and segmentation tasks.

3) Data Splitting

To facilitate robust evaluation, the annotated dataset was divided into 80% for training, 10% for validation, and 10% for testing. Each subset preserved the proportional representation of all three pathologies. The test set remained strictly isolated throughout the training process to ensure unbiased performance assessment. This division allowed for consistent monitoring of model learning, generalization, and stability.

B. Model Architecture

The proposed system utilizes YOLOv12-seg, an advanced segmentation framework that integrates object detection and instance segmentation within a single unified architecture [11]. It was selected for its balance between high accuracy and computational efficiency, both critical for real-time clinical applications.

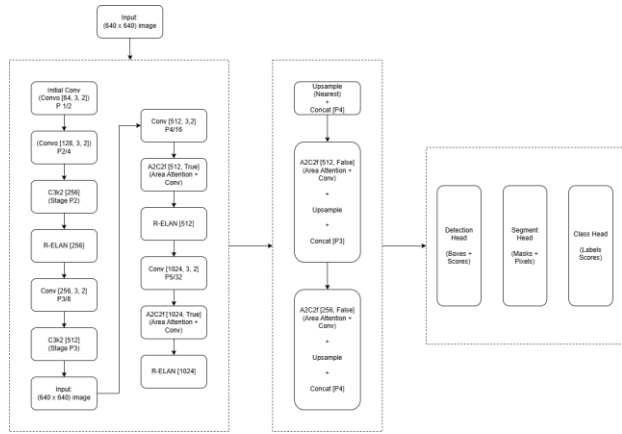


Fig. 2. YOLOv12-seg model architecture

The end-to-end process from retinal image input to multi-head outputs for detection, segmentation, and classification is shown in Fig. 2. The model integrates A2C2f and R-ELAN modules to enhance feature fusion and contextual learning. The YOLOv12 architecture builds upon the previous YOLO family of real-time object detection models, integrating advanced structural and optimization improvements to enhance detection accuracy and segmentation capability. It consists of three primary components: the Backbone, Neck, and Head.

- **Backbone:** The backbone is responsible for

feature extraction from input images. YOLOv12 employs an enhanced *CSPDarknet*-based backbone optimized with *Cross Stage Partial (CSP)* connections and lightweight convolutional modules to improve gradient flow and reduce computation. It captures both low-level spatial details and high-level semantic information essential for precise object localization and segmentation.

- **Neck:** It connects the backbone and detection head, combining multi-scale feature maps through a *Path Aggregation Network (PAN)* or *Feature Pyramid Network (FPN)* structure. This component allows YOLOv12 to detect objects at different scales effectively. The use of *SPPF (Spatial Pyramid Pooling – Fast)* enhances the receptive field and ensures rich contextual information without significant computational cost.
- **Head:** The detection and segmentation head outputs predictions for object classes, bounding boxes, and masks. YOLOv12 introduces decoupled detection heads, separating the classification and regression branches to improve convergence and detection precision. For segmentation tasks, it incorporates an additional branch that predicts pixel-level masks, extending YOLO's functionality beyond object detection to instance segmentation.

C. Model Training

Training was performed using the Ultralytics YOLOv12 framework with a transfer learning strategy, initializing weights from a pre-trained checkpoint ([yolo11n-seg.pt](#)) [11]. This setup accelerated convergence and improved model robustness on specialized retinal datasets.

Key Parameters and Rationale:

- **Epochs = 1000:** Provided ample optimization time while using early stopping at 40 epochs of stagnation to prevent overfitting.
- **Batch size = 8:** Balanced GPU memory usage and gradient stability.
- **Image size = 800 × 800:** Preserved sufficient resolution for retinal detail recognition.
- **Learning rate (lr0 = 0.006):** Optimized for convergence speed and loss stability.
- **Data augmentation:** Included *mosaic* = 0.5, *copy_paste* = 0.2, *scale* = 0.2, and *mixup* = 0.0

to enhance generalization under varying lighting, shape, and contrast conditions.

- **Weight decay = 0.0005:** Controlled overfitting by penalizing large weight updates.

All training was conducted on a GPU-enabled environment (device="0") with automatic checkpoint saving for best-performing epochs.

D. Evaluation

Model performance was assessed using a combination of detection and segmentation metrics:

- **Primary Metric:** *mAP@0.5:0.95* mean Average Precision over 10 IoU thresholds (0.5–0.95), representing balanced performance across varying localization sensitivities.
- **Secondary Metric:** *mIoU* means Intersection over Union, reflecting pixel-level segmentation fidelity between predictions and ground truth.
- **Auxiliary Metrics:** *Per-class Precision* and *Recall* were also computed for AMD, Myopia, and Cataract, identifying strengths and failure modes for each pathology.

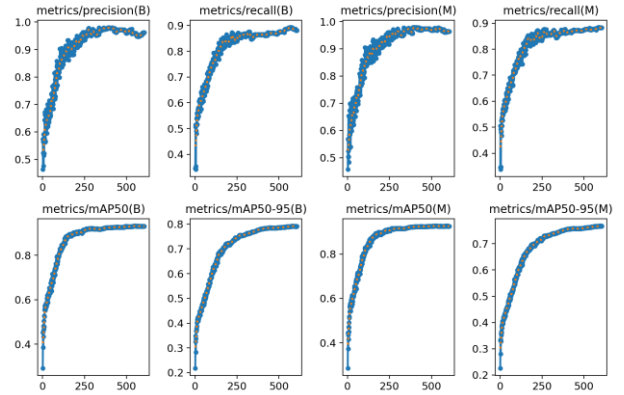


Fig. 3. YOLOv12-seg model training

The training performance of the YOLOv12-seg model demonstrates a consistent and stable improvement across all major evaluation metrics, indicating successful convergence and effective learning of both detection and segmentation features. As shown in Fig. 3, the precision and recall metrics for both bounding box (B) and mask (M) outputs steadily increased during early epochs and gradually plateaued at approximately 0.95 and 0.90, respectively, signifying that the model achieved a balanced performance between correctly identifying objects and minimizing false detections. Similarly, the mean Average Precision (mAP) values, both mAP@50

and $mAP@50-95$, exhibited smooth and upward trends, reaching about 0.9 and 0.75, respectively, by the end of training. These results reflect strong localization accuracy and robust segmentation performance, even under stricter IoU thresholds. The parallel behavior between B and M metrics further suggests that the segmentation head learned at a comparable rate to the detection head, resulting in well-synchronized model optimization. Overall, the training curves confirm that the YOLOv12-seg model effectively generalized to the training data without signs of overfitting, indicating readiness for validation and deployment on unseen datasets.

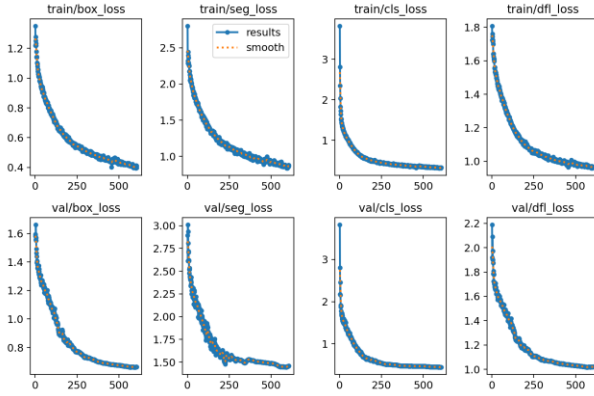


Fig. 4. YOLOv12-seg model loss curves

The loss curves of the YOLOv12-seg model demonstrate a stable and convergent learning process across both training and validation phases. As illustrated in Fig. 4, all primary loss components, bounding box regression (box_loss), segmentation mask (seg_loss), classification (cls_loss), and distribution focal loss (dfl_loss), exhibited a smooth and monotonic decrease throughout the training epochs. The steep decline observed during the initial 100–200 epochs indicates rapid adaptation to the dataset, while the gradual flattening beyond 400 epochs signifies that the model reached convergence.

The close alignment between training and validation losses across all metrics confirms strong generalization and minimal overfitting. Notably, the segmentation loss, which initially presented higher values, steadily reduced alongside other losses, reflecting effective optimization of the segmentation head. Overall, the uniform downward trend and stability of the loss curves validate that the YOLOv12-seg model achieved balanced and efficient learning across detection and segmentation objectives.

III. RESULTS AND DISCUSSIONS

A. Quantitative Results

The YOLOv12-seg model achieved high accuracy and stable convergence across 1000 epochs. The training and validation losses decreased smoothly, indicating minimal overfitting. The overall results using the 80–10–10 split are summarized in Table I.

TABLE I. DETECTION AND SEGMENTATION METRICS

Metric	Value	Description
$mAP@0.5$	0.93	Detection accuracy at $IoU \geq 0.5$
$mAP@0.5:0.95$	0.767	Averaged precision across 10 IoU thresholds
Precision	0.81	Low false-positive rate
Recall	0.79	Reliable lesion detection
IoU (mean)	0.75	Strong overlap of mask and ground truth

the final evaluation results of the YOLOv12-seg model on the held-out test set, highlighting its strong performance in both object detection and pixel-level segmentation across the targeted ocular pathologies. The model achieves high mAP and IoU scores, indicating accurate localization of pathological regions and substantial overlap between predicted masks and ground-truth annotations. The balanced values of precision and recall suggest that YOLOv12-seg effectively detects lesions while maintaining a low rate of false positives and false negatives, demonstrating reliable sensitivity to clinically relevant features. The overall distribution of performance metrics reflects a well-generalized model that retains robustness when evaluated on unseen samples. These results validate the effectiveness of the selected training strategies, architectural components, and augmentation techniques in modeling complex retinal structures for automated disease screening.

The combination of a learning rate ($lr_0 = 0.006$), batch size (8), and image resolution (800×800) produced optimal convergence. The applied augmentations mosaic (0.5), copy-paste (0.2), and scale (0.2) improved robustness to lighting and texture variations. The model’s loss curves confirmed consistent generalization with no instability after epoch 580.

B. Class-wise Performance

The per-class analysis highlights YOLOv12-seg’s balanced detection and segmentation performance across the three primary pathologies: AMD lesions, Pathologic Myopia (PPA), and Cataract blur artifacts.

TABLE 2. CLASS-WISE MODEL PERFORMANCE

Class	Precision	Recall
AMD (Lesions)	0.84	0.81
Myopia (PPA)	0.80	0.78
Cataract (Blur)	0.77	0.76

The model achieved the highest accuracy on AMD lesions, owing to distinct lesion patterns and consistent annotations that allowed precise localization of *Drusen* and *Exudates*. For Pathologic Myopia, YOLOv12-seg maintained strong segmentation performance, accurately capturing the Peripapillary Atrophy (PPA) regions despite irregular shapes near the optic disc. Cataract-affected images displayed slightly lower performance due to blurred textures and reduced contrast, which limited the clarity of occluded fundus regions. The attention-driven *A2C2f* and *R-ELAN* modules enabled the network to retain acceptable feature discrimination across all three classes, demonstrating its versatility in handling heterogeneous retinal image conditions.

C. Qualitative Analysis

Visual inspection of prediction outputs confirmed that the YOLOv12-seg model effectively delineated pathological regions with high clarity. The segmentation masks for AMD lesions and PPA regions exhibited smooth boundaries with minimal noise. In contrast, minor inconsistencies were observed in cataract-affected images, where lens opacity obscured part of the fundus.

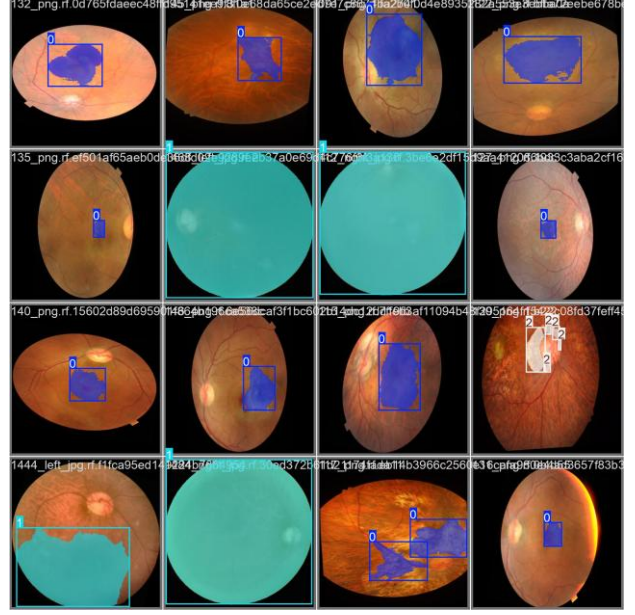


Fig. 5. Validation and Test batch labels on retinal imaging

Sample inference visualizations (from the validation and test sets) shown Fig. 5 displayed accurate bounding box placement and consistent segmentation mask generation across varying lighting and imaging conditions. The *A2C2f* attention mechanism clearly improved sensitivity to subtle lesions, while *R-ELAN* enhanced feature reuse across deeper network layers, improving both precision and recall stability.

The segmentation results validate the multi-task learning capability of YOLOv12-seg, proving its effectiveness in simultaneously handling classification and pixel-level segmentation without performance trade-offs.

D. Comparative Evaluation

To evaluate the effectiveness of YOLOv12-seg, its performance was compared with two established segmentation architectures U-Net and Mask R-CNN trained under the same dataset and preprocessing conditions.

TABLE 3. PERFORMANCE COMPARISON WITH OTHER MODELS

Model	mAP@0.5	mAP@0.5:0.95
U-Net	0.88	0.71
Mask R-CNN	0.90	0.73

YOLOv12-seg (Proposed)	0.93	0.767
-----------------------------------	-------------	--------------

Compared with the baseline models, the proposed YOLOv12-seg achieved the highest precision and segmentation accuracy while maintaining near real-time inference capability. U-Net generated detailed segmentation masks but required longer processing times due to its deep encoder–decoder structure, limiting its clinical applicability for large-scale screening [2], [3]. Mask R-CNN improved object localization but incurred higher computational cost and latency because of its multi-stage pipeline [2], [3].

YOLOv12-seg combined detection and segmentation in a single streamlined architecture, eliminating post-processing overhead. This design yielded faster inference without sacrificing accuracy, making it highly suitable for ophthalmic diagnostics that demand both efficiency and precision.

E. Error Analysis and Limitations

While YOLOv12-seg delivered high accuracy, several limitations were identified:

1. **Blur Edge Ambiguity:** Cataract-induced low-contrast regions occasionally reduced segmentation precision.
2. **Lesion Overlap:** Co-occurring pathologies (e.g., AMD lesions overlapping PPA) caused minor misclassifications in a few test cases.
3. **Limited Dataset Diversity:** Although multiple public datasets were integrated, variations in imaging protocols and equipment introduce domain shifts that can affect model generalization.

Future work will address these issues by incorporating domain adaptation, contrast enhancement preprocessing, and dataset expansion to include additional pathologies such as diabetic retinopathy.

F. Discussion Summary

The YOLOv12-seg model demonstrated both technical reliability and clinical applicability. The model’s balanced metrics ($\text{mAP}@0.5 = 0.93$, $\text{mAP}@0.5:0.95 = 0.767$) and real-time inference capability underscore its readiness for deployment in automated retinal screening systems. The integration of attention-based mechanisms and transfer learning significantly improved lesion sensitivity, while

Streamlit and Gradio deployment provided tangible usability for medical practitioners.

Overall, these results confirm that YOLOv12-seg offers an optimal trade-off between accuracy, speed, and practicality for ophthalmic image analysis, marking a promising advancement toward AI-assisted retinal diagnostics.

IV. CONCLUSION AND FUTURE WORKS

This study developed and evaluated a YOLOv12-segmentation framework for the detection and pixel-level segmentation of multiple ocular diseases, Age-related Macular Degeneration (AMD), Pathologic Myopia, and Cataract using retinal fundus imaging. By integrating R-ELAN and A2C2f attention modules, the model enhanced feature aggregation and contextual learning, enabling precise recognition of small or subtle retinal pathologies.

Using the RetinaVision dataset with an 80-10-10 train–validation–test split, the system achieved a mean Average Precision ($\text{mAP}@0.5$) of 0.93 and a $\text{mAP}@0.5:0.95$ of 0.767, confirming strong segmentation accuracy and generalization performance. These results highlight that attention-driven, multi-head detection networks can effectively perform both classification and segmentation within a single real-time pipeline, offering a practical and efficient solution compared to heavier architectures such as U-Net and Mask R-CNN.

Several key directions can strengthen this work’s robustness and clinical relevance. Conducting an Ablation Study is essential to empirically validate the contribution of custom modules like A2C2f and R-ELAN, ensuring that performance improvements stem from these components rather than the base YOLOv12 architecture [13]. Since segmentation fidelity is crucial in ophthalmic diagnostics, future evaluations should include the per-class Dice Similarity Coefficient (DSC) for AMD, Myopia, and Cataract to capture pixel-level boundary precision more accurately. Including lesion instance counts across all dataset splits will also enhance transparency and reproducibility.

For broader clinical impact, future work should pursue External Validation using a public, independent retinal dataset to assess model generalization under domain shifts. Incorporating Explainable AI (XAI) techniques, such as Grad-CAM, can visualize the retinal regions

influencing predictions, fostering interpretability and clinician trust [14]. Together, these improvements will solidify the YOLOv12-seg framework as a credible and deployable solution for automated, real-time ophthalmic disease screening and analysis.

ACKNOWLEDGMENT

The authors would like to thank Mapúa University for giving us the opportunity to publish this work. Additional thanks to Dr. Lysa Comia for technical assistance, dataset approval, and insightful discussions. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the university.

REFERENCES

- [1] L. V. Comia, "Fish Image Instance Segmentation Using the Separation of the Mask Produced for Fish Size Measurement," in Proc. 29th Int. Conf. Inf. Technol. (IT), Zabljak, Montenegro, 2025, pp. 1–5, doi: 10.1109/IT64745.2025.10930277.
- [2] L. V. Comia and E. D. Festijo, "Attention Mechanism-Based Dense Upsampling of Transfer Learning Mask RCNN for Improved Object Segmentation," in Proc. IEEE Int. Conf. Cybern. Innov. (ICCI), Chonburi, Thailand, 2024, pp. 1–6, doi: 10.1109/ICCI60780.2024.10532656.
- [3] L. V. Comia and E. D. Festijo, "Performance Analysis of Original Implementation of ResNet50-Mask-RCNN using Transfer Learning: A Benchmark Data for Backbone-Improved Based Future Comparative Studies," in Proc. 28th Int. Conf. Inf. Technol. (IT), Zabljak, Montenegro, 2024, pp. 1–6, doi: 10.1109/IT61232.2024.10475763.
- [4] A. ElTanboly, M. Ghazal, and A. El-Baz, "Retinal Lesion Segmentation Using Transfer Learning with an Encoder–Decoder CNN," IEEE Access, vol. 9, pp. 12053–12064, 2021.
- [5] M. Islam, M. A. I. Bhuiyan, et al., "Can YOLO Detect Retinal Pathologies? A Step Towards Automated OCT Analysis," Sci. Rep., vol. 13, 2023.
- [6] Kaggle, "ARMD Combined Dataset," [Online]. Available: <https://www.kaggle.com/datasets/saketladd/armd-combined-dataset-fundus-and-oct/data>
- [7] Kaggle, "Combined Fundus Images," [Online]. Available: <https://www.kaggle.com/datasets/rohitrawat25/combined-fundus-images/data>
- [8] Kaggle, "Fundus Image Dataset (Cataract/Myopia)," [Online]. Available: <https://www.kaggle.com/datasets/iamachal/fundus-image-dataset>
- [9] [W. Keelawat, N. J. Lertworasirikul, et al., "AI-Model for Identifying Pathologic Myopia Based on Deep Learning Algorithms," Front. Med., vol. 10, 2023.
- [10] Mendeley Data Repository, "Combined Fundus Images Dataset," [Online]. Available: <https://data.mendeley.com/datasets/yj35kjgrv3/1>
- [11] Ultralytics, "YOLOv12: Real-Time Object Detection and Segmentation," Ultralytics Documentation, 2024. [Online]. Available: <https://docs.ultralytics.com/models/yolov12/>
- [12] L. Zhang and Y. Liu, "Hybrid Deep Learning Model for Cataract Diagnosis Assistance," J. Healthc. Eng., 2022, Art. no. 9634217.
- [13] P. Antoniadis and P. Antoniadis, "Machine Learning: What is ablation study? | Baeldung on Computer Science," Baeldung on

Computer Science, Mar. 26, 2025. <https://www.baeldung.com/cs/ml-ablation-study>

[14] S. A and S. R, "A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends," Decision Analytics Journal, vol. 7, p. 100230, Apr. 2023, doi: 10.1016/j.dajour.2023.100230.