Advanced Fundus Analysis: Deep Learning for Retinopathy of Prematurity Diagnosis

A project report submitted in partial fulfillment

of the requirements for the degree of

Bachelor of Technology

in

Electronics & Computer Engineering

by

Divyansh Rawal

21BLC1123

School of Electronics Engineering,

Vellore Institute of Technology, Chennai,

Vandalur-Kelambakkam Road,

Chennai - 600127, India.

April 2025

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Declaration

I hereby declare that the report titled ***Advanced Fundus Analysis: Deep Learning for Retinopathy of Prematurity Diagnosis*** submitted by us to the School of Electronics Engineering, Vellore Institute of Technology, Chennai in partial fulfillment of the requirements for the award of **Bachelor of Technology** in **Electronics and Computer Engineering** is a bona-fide record of the work carried out by me under the supervision of ***Dr.Ilavendhan.A.***

I further declare that the work reported in this report, has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma of this institute or of any other institute or University.

Sign: _____

Name & Reg. No.: _____

Date: _____

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

School of Electronics Engineering

Certificate

This is to certify that the project report titled ***Advanced Fundus Analysis: Deep Learning for Retinopathy of Prematurity Diagnosis*** submitted by ***Divyansh Rawal*** **(Reg. No. 21BLC1123)** to Vellore Institute of Technology Chennai, in partial fulfillment of the requirement for the award of the degree of **Bachelor of Technology** in **Electronics and Computer Engineering** is a bona-fide work carried out under my supervision. The project report fulfills the requirements as per the regulations of this University and in my opinion meets the necessary standards for submission. The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma and the same is certified.

Supervisor

Signature: .....

Name: .....

Date:

Head of the Department

Signature: .....

Name: .....

Date:

Examiner

Signature: .....

Name: .....

Date:

(Seal of the School)

Abstract

Retinopathy of prematurity (ROP) is one of the principal causes of blindness in preterm infants and is mainly associated with abnormal development of retinal blood vessels. Hence, early detection and management of ROP are of paramount importance to avoid severe visual impairment. However, manual screening methods consume much time and require special expertise. Also, they are highly subjective. The recent advances in deep learning have now paved the way for automated analysis of fundus images, uncertainly coming to the rescue for ROP detection. But, despite the advancements, there remains a gap in research evaluating disease-detecting capabilities of state-of-the-art object detection methods such as YOLOv8 and YOLOv11 in medical images, specifically ROP diagnosis. The motivation behind this study was to compare YOLOv8 and YOLOv11 from an architectural perspective and training speed, accuracy, and real-time performance potential for detecting ROP from fundus images. A systematic methodology was then followed: dataset preparation from a set of public as well as clinical datasets, model training with optimized hyperparameters, and evaluation of models with the following metrics: mean Average Precision (mAP), Intersection over Union (IoU), precision, recall, and inference speed. Fundus images were annotated as per the ICROP (International Classification of Retinopathy of Prematurity) criteria for even stronger validation. Preliminary results yield that YOLOv11 is far superior to YOLOv8 with regard to mAP and training convergence speed while being more efficient in real-time detection. The architecture of YOLOv11, with its sophisticated spatial attention mechanisms and feature extraction capabilities, is particularly efficient at locating small and complicated retinal features that are important in ROP staging. This research holds great potential in that the implementation of YOLOv11 into the clinical pipeline will alleviate the burden on ophthalmologists and will thus give faster diagnoses in resource-poor settings and will also have a bearing on enhanced treatment outcome with timely intervention. This study contributes to operationalizing the realization of a wide body of research aimed toward AI-assisted ROP screening systems and offers projections for frontline deep learning models to resolve global approaches to neonatal ophthalmology. The findings from this study suggest that YOLOv11 holds significant clinical promise. Its incorporation into automated ROP screening workflows could dramatically reduce diagnostic delays, alleviate the workload of retinal specialists, and support early-stage intervention, particularly in regions with limited access to specialized care.

KEYWORDS: Retinopathy of Prematurity (ROP), Deep Learning, YOLOv8, YOLOv11, Object Detection, Fundus Images, Real-Time Diagnosis, Neonatal Ophthalmology, Medical Image Analysis.

Acknowledgements

We wish to express our sincere thanks and deep sense of gratitude to our project guide, Dr. Ilavendhan.A, Professor, School of Computer Science and Engineering, for his consistent encouragement and valuable guidance offered to us in a pleasant manner throughout the course of the project work.

We are extremely grateful to Dr. Ravishankar A, Dean Dr. Reena Monica, Associate Dean (Academics) & Dr. John Sahaya Rani Alex, Associate Dean (Research) of the School of Electronics Engineering, VIT Chennai, for extending the facilities of the School towards our project and for his unstinting support.

We express our thanks to our Head of the Department Dr. Annis Fathima A for her support throughout the course of this project.

We also take this opportunity to thank all the faculty of the School for their support and their wisdom imparted to us throughout the course.

We thank our parents, family, and friends for bearing with us throughout the course of our project and for the opportunity they provided us in undergoing this course in such a prestigious institution.

Contents

List of Figures

Chapter 1

Introduction

Retinopathy of Prematurity (ROP) remains one of the leading causes of childhood blindness across the globe, particularly affecting premature infants with low birth weight. The disease is characterized by the abnormal development of retinal blood vessels and is typically diagnosed through retinal imaging and clinical evaluation. With the rise in survival rates of preterm infants due to advancements in neonatal care, there has been a corresponding increase in the incidence of ROP. Early diagnosis and timely intervention are essential to preventing irreversible vision loss, making ROP screening a critical component of neonatal intensive care units (NICUs). However, the conventional process of ROP diagnosis is highly subjective, labor-intensive, and reliant on the availability of experienced ophthalmologists—factors that pose significant challenges, especially in resource-constrained environments.

In recent years, artificial intelligence (AI), particularly deep learning, has emerged as a transformative force in medical diagnostics. Deep learning, a subfield of machine learning, leverages multi-layered neural networks to model complex patterns within data. In the context of ophthalmology, convolutional neural networks (CNNs) have demonstrated remarkable efficacy in processing retinal images, enabling automated detection of pathological features with accuracies that rival or even surpass human experts. The application of deep learning in ROP diagnosis has opened up new possibilities for scalable, objective, and cost-effective screening solutions. By automating image interpretation and classification, deep learning not only alleviates the burden on clinicians but also promises to bring high-quality diagnostic capabilities to underserved regions where access to ophthalmic specialists is limited.

The adoption of deep learning for ROP is driven by several key developments in both the technological and clinical domains. First, the increasing availability of large, annotated datasets of fundus images has enabled researchers to train more robust and generalizable

models. Additionally, advancements in image acquisition technology, including portable and smartphone-based fundus cameras, have facilitated the collection of retinal images at scale. From a clinical standpoint, the use of deep learning introduces objectivity into what has traditionally been a subjective process, thereby enhancing inter-rater agreement and standardization in diagnosis. Furthermore, deep learning systems can operate in real time and are capable of continuous learning, meaning their performance can improve over time with additional data inputs.

Despite these advantages, there are still numerous challenges and limitations that hinder the widespread deployment of deep learning systems for ROP screening. One of the primary concerns is the lack of generalizability across different populations and imaging modalities. Many models are trained on datasets collected from specific geographic regions or using specific camera types, resulting in decreased performance when applied to new environments. Another critical issue is model interpretability; clinicians are often hesitant to trust "black-box" systems whose decision-making processes are opaque. This has led to increased interest in explainable AI techniques that aim to provide visual or statistical justifications for model outputs. Additionally, regulatory and ethical considerations, such as data privacy, consent, and clinical accountability, remain active areas of discussion.

Several landmark studies have illustrated the potential and pitfalls of deep learning in ROP diagnosis. For example, Chen et al. (2021) developed a CNN model capable of staging ROP with high accuracy and demonstrated its applicability across diverse populations. Attallah (2021) introduced DIAROP, a robust deep learning-based diagnostic tool that integrates multiple neural networks for improved reliability. Tan et al. (2019) and Brown et al. (2020) focused specifically on diagnosing plus disease, a severe form of ROP characterized by abnormal vascular dilation and tortuosity. These studies highlight both the progress made and the work still required to develop clinically viable AI systems. Moreover, recent efforts such as the FARFUM-RoP dataset introduced by Akbari et al. (2023) and the synthetic-to-real data augmentation strategies proposed by Oliveira and Rosa (2024) represent significant strides in addressing data limitations and enhancing model robustness.

The objective of this research project is to build upon these advancements by designing and evaluating a deep learning-based system for automated ROP diagnosis. This system aims to leverage a curated dataset of retinal fundus images, apply state-of-the-art convolutional neural networks for classification, and explore data augmentation techniques to improve generalization. Additionally, the project will assess the model's interpretability and propose a framework for its potential integration into clinical workflows. The

broader vision is to contribute to the development of accessible, accurate, and scalable solutions for neonatal retinal care, particularly in regions where specialist access is limited.

In conclusion, the integration of deep learning technologies into the diagnostic pipeline for Retinopathy of Prematurity (ROP) represents a significant advancement in the field of pediatric ophthalmology. These cutting-edge algorithms have the potential to not only enhance the diagnostic accuracy and efficiency of clinicians but also to extend critical eye care services to underserved and remote areas where access to specialized expertise may be limited. By automating the detection and classification of ROP, deep learning can serve as a powerful assistive tool—reducing the burden on healthcare professionals while ensuring timely intervention for at-risk neonates.

Nevertheless, harnessing the full potential of deep learning in this context requires addressing several key challenges. Technically, the models must be robust, interpretable, and trained on diverse, high-quality datasets that reflect real-world variability. Clinically, it is imperative to validate these tools in diverse healthcare settings to ensure they meet the rigorous standards of medical safety and effectiveness. Ethically, developers and stakeholders must carefully consider issues related to data privacy, algorithmic bias, informed consent, and equitable access to technology.

Through this project, we endeavor to contribute meaningfully to the advancement of AI in neonatal care by developing a deep learning–based diagnostic tool that is both technically sophisticated and clinically applicable. Our work not only aims to bridge existing gaps in current diagnostic methodologies but also aspires to support the broader vision of leveraging artificial intelligence to improve health outcomes for vulnerable neonatal populations worldwide. In doing so, we hope to add to the growing body of knowledge in medical imaging and set the stage for future innovations in the early detection and management of ROP.

Chapter 2

Literature Survey

2.1 Jimmy S. Chen, Aaron S. Coyner, Susan Ostmo, Kemal Sonmez (2021) – Deep Learning for the Diagnosis of Stage in Retinopathy of Prematurity: Accuracy and Generalizability across Populations and Cameras

Overview:
This study investigates the use of deep learning algorithms for the staging of Retinopathy of Prematurity (ROP), focusing on their accuracy and ability to generalize across different populations and camera types. Given the variation in retinal image quality and inter-observer discrepancies in ROP diagnosis, the study aims to provide a robust AI-based solution to assist clinicians in identifying disease stages.

Contribution:
The research makes a significant contribution by developing a convolutional neural network (CNN)-based model capable of detecting the ROP stage with high accuracy. A unique aspect of the work is its emphasis on testing across multiple imaging platforms and demographic groups, addressing concerns about AI fairness and reproducibility. The model demonstrated consistent performance across various population subsets and image acquisition devices, reinforcing its clinical applicability.

Limitations:
Despite its strengths, the model's interpretability remains limited, which may hinder its adoption in clinical settings where explainability is crucial. Additionally, although the model performs well across datasets, subtle image artifacts or rare manifestations of ROP might still affect its predictions.

Future Scope:
The study lays the foundation for future work in creating explainable AI (XAI) systems that highlight the basis for classification decisions. There is also scope for integrating this model into real-time telemedicine platforms, especially in rural or underserved regions, to bridge the diagnostic gap caused by the shortage of trained ophthalmologists.

2.2 Guilherme C. Oliveira, Gustavo H. Rosa (2024) – Robust Deep Learning for Eye Fundus Images: Bridging Real and Synthetic Data for Enhancing Generalization

Overview:
This paper addresses a major challenge in medical imaging: the scarcity and diversity of labeled datasets. It proposes a robust deep learning framework that combines real and synthetically generated fundus images to improve model generalization for various retinal diseases, including ROP.

Contribution:
The authors introduce a data augmentation technique using Generative Adversarial Networks (GANs) to generate high-quality synthetic images that supplement real datasets. By training models on this hybrid dataset, the system demonstrates improved accuracy, particularly on out-of-distribution samples. The study bridges a critical gap in the literature by validating synthetic image utility for deep learning in ophthalmology.

Limitations:
A notable limitation is the dependency on the quality and diversity of synthetic data. GAN-generated images, while realistic, may introduce subtle artifacts or lack the clinical nuances of real retinal pathologies. Moreover, real-world deployment might face challenges in regulatory approval for using synthetic training data in diagnostic tools.

Future Scope:
The promising results open up opportunities for domain adaptation techniques that can further align real and synthetic distributions. Future research could focus on training with unsupervised or semi-supervised methods, as well as developing tools for synthetic data validation and certification in clinical environments.

2.3 Morteza Akbari et al. (2023) – FARFUM-RoP: A Dataset for Computer-Aided Detection of Retinopathy of Prematurity

Overview:
FARFUM-RoP presents a new curated dataset specifically designed for computer-aided diagnosis of ROP. The dataset includes a diverse set of annotated fundus images, classified based on ROP stages, zone involvement, and vascular abnormalities, making it suitable for training and benchmarking deep learning models.

Contribution:
This study is vital for the research community as it introduces a standardized dataset with high inter-rater agreement and metadata annotations. The availability of this dataset facilitates reproducibility and comparative analysis of ROP detection algorithms. It also includes images from different cameras and acquisition settings, increasing its relevance for real-world applications.

Limitations:
The main limitation lies in the dataset's size, which, although comprehensive, may still fall short for training large-scale models. Additionally, while annotations are detailed, the dataset may not fully cover rare or complex ROP manifestations, potentially limiting model robustness.

Future Scope:
The dataset provides a baseline for future improvements in model training, especially with techniques such as transfer learning. Future work could include continual dataset expansion and the addition of more metadata, such as treatment outcomes and longitudinal imaging, to support prognostic modeling.

2.4 Tao Li, Wang Bo, Chunyu Hu (2021) – Applications of Deep Learning in Fundus Images

Overview:
This review paper surveys the application of deep learning techniques in analyzing fundus images for a range of ophthalmic diseases, including diabetic retinopathy, glaucoma, and ROP. It categorizes existing literature based on the task, such as classification, segmentation, and detection, and summarizes the progression of AI models over time.

Contribution:
The review serves as a comprehensive guide for researchers by highlighting major milestones, datasets, and techniques used in the field. It emphasizes the trend towards end-to-end deep learning pipelines and introduces a taxonomy of applications based on the disease and imaging modality.

Limitations:
One limitation is the lack of deep dive into ROP-specific studies. While ROP is mentioned, the primary focus leans towards more prevalent conditions like diabetic retinopathy and glaucoma. Thus, readers seeking detailed ROP methodologies may need to look elsewhere.

Future Scope:
The authors suggest that future work should focus on multimodal learning by integrating other data sources such as OCT scans and patient history. They also highlight the need for interpretable models and cross-domain generalization to ensure broader clinical acceptance.

2.5 Timkovic et al. (2015) – A New Modified Technique for the Treatment of High-Risk Prethreshold ROP Under Direct Visual Control of RetCam

Overview:
This clinical study introduces a modified treatment approach for high-risk prethreshold ROP using the RetCam 3 imaging system for real-time visual guidance. The study is primarily focused on improving treatment precision and reducing complications through better intraoperative visualization.

Contribution:
The use of RetCam 3 for direct visual feedback represents a novel advancement in the surgical management of ROP. The study shows a significant improvement in treatment success rates and reduced incidence of postoperative complications compared to traditional techniques.

Limitations:
The technique requires specialized equipment and expertise, which may limit its applicability in low-resource settings. Additionally, the paper lacks a large sample size, which is essential for statistical validation of its findings.

Future Scope:
There is potential to integrate AI-driven image guidance into RetCam procedures for real-time decision support. Expanding this method into tele-ophthalmology settings could also enable remote consultations and interventions.

2.6 Stahl et al. (2019) – Ranibizumab versus Laser Therapy for the Treatment of Very Low Birthweight Infants with ROP (RAINBOW Trial)

Overview:
The RAINBOW trial is a multicenter, randomized controlled study comparing the efficacy of intravitreal ranibizumab injection with traditional laser therapy in treating ROP in very low birthweight infants.

Contribution:
The trial demonstrates that ranibizumab is non-inferior to laser therapy and may be preferable in certain clinical scenarios due to less collateral damage to the retina. It marks a paradigm shift in ROP treatment by validating pharmacologic interventions alongside or in place of surgical procedures.

Limitations:
Long-term effects of anti-VEGF treatments like ranibizumab on infant development are still under investigation. Additionally, the need for multiple follow-ups post-injection could strain clinical resources and reduce feasibility in remote areas.

Future scope:
Further research is warranted to explore combination therapies and personalized treatment protocols. Integrating AI tools for early prediction of treatment responders based on retinal images could optimize clinical outcomes.

2.7 Mao et al. (2020) – New Grading Criterion for Retinal Hemorrhages in Term Newborns Based on Deep Convolution Neural Networks

Overview:
This study introduces a deep learning-based grading system for classifying retinal hemorrhages in term newborns using convolution neural networks (CNNs). The approach is

data-driven and seeks to standardize diagnosis across clinicians.

Contribution:
By applying CNNs, the authors achieve high inter-rater consistency and demonstrate that automated systems can potentially outperform manual grading in both speed and accuracy. This work serves as a reference for extending AI applications to other neonatal retinal conditions like ROP.

Limitations:
The study is specific to retinal hemorrhages and may not directly generalize to other neonatal retinal disorders. Additionally, the CNN model's performance depends heavily on the quality and labeling accuracy of the training data.

Future Scope:
The grading framework could be adapted to other neonatal eye disorders, including ROP, using transfer learning. Further refinement in terms of model interpretability and longitudinal tracking would enhance clinical relevance.

2.8 Quinn G. E. et al. (2014) – Validity of a Telemedicine System for the Evaluation of Acute-Phase Retinopathy of Prematurity

Overview:
This research evaluates a telemedicine-based diagnostic system for acute-phase ROP, comparing its accuracy with that of in-person assessments by ophthalmologists.

Contribution:
The study validates tele-ROP systems as a reliable and scalable solution for ROP screening, particularly in remote or underserved areas. It demonstrates that telemedicine can match or even exceed in-clinic evaluation accuracy when using trained image graders and standard imaging protocols.

Limitations:
Telemedicine systems still require expensive imaging equipment and trained personnel for image acquisition. Delays in diagnosis due to scheduling and internet limitations may affect timely intervention.

Future Scope:
The integration of AI models for automated screening within telemedicine platforms

could reduce dependence on human graders. Further studies are needed to assess the cost-benefit ratio and expand the system's coverage across broader demographics.

2.9 Tan Z, Simkin S, Lai C, Dai S. (2019) - Deep Learning Algorithm for Automated Diagnosis of Retinopathy of Prematurity Plus Disease

Overview:
This study presents a deep learning algorithm developed to automatically diagnose plus disease, a critical marker for treatment-requiring Retinopathy of Prematurity (ROP). The authors employed convolutional neural networks (CNNs) trained on retinal fundus images to distinguish between plus and non-plus disease with the aim of assisting clinicians in early diagnosis.

Contribution:
Developed a CNN-based model that demonstrated promising diagnostic performance in identifying plus disease.Showed potential for automation in ROP screening, offering consistent evaluations aligned with expert clinical opinions.Highlighted the feasibility of using deep learning in clinical decision-making, especially in telemedicine applications.

Limitations:
Limited dataset diversity—images were primarily sourced from a single institution, potentially impacting generalizability.Focused only on plus disease, without addressing other stages or zones of ROP.Lack of external validation in real-world clinical settings.

Future Scope:
Expansion of the dataset to include multi-center and ethnically diverse populations.Integration with comprehensive ROP screening systems capable of classifying multiple ROP stages and zones.Clinical trials to evaluate performance in real-time telehealth workflows and low-resource environments.

2.10 Brown JM, Campbell JP, Beers A, et al. (2018) - Automated Diagnosis of Plus Disease in Retinopathy of Prematurity Using Deep Convolutional Neural Networks

Overview:
This study evaluated the performance of a deep convolutional neural network (CNN) trained to identify plus disease in ROP using a large dataset of expert-labeled retinal images. The model's diagnostic performance was compared to that of board-certified ophthalmologists, emphasizing the potential of AI-assisted diagnosis in ROP.

Contribution:
Demonstrated that a CNN could achieve diagnostic accuracy comparable to or exceeding that of expert clinicians.Used a large, well-curated dataset with multiple expert labels, strengthening the reliability of model training and evaluation.Pioneered the application of explainable AI in ROP, offering heatmaps to highlight image regions influencing the model's decision.

Limitations:
Despite high performance, the model's interpretability and clinical trustworthiness remain concerns.Dataset still limited to labeled still images; real-time clinical applicability not fully tested.Potential for bias due to expert disagreement in labeling, which may affect ground truth reliability.

Future Scope:
Deployment and validation in prospective clinical trials.Enhancement of model interpretability to improve clinician trust and transparency.Development of comprehensive AI tools covering full ROP diagnosis, including stage, zone, and presence of plus disease.

Chapter 3

Methodology

3.1 Introduction

The methodology adopted for the detection and classification of Retinopathy of Prematurity (ROP) through fundus images is grounded in image processing and machine learning techniques. The primary goal is to create an automated, accurate, and efficient system that supports ophthalmologists in early ROP detection, thereby preventing blindness in premature infants. This section elaborates on each stage of the methodology including dataset acquisition, preprocessing, feature extraction, classification, and performance evaluation.

3.1.1 Data Acquisition

Fundus images were obtained from publicly available datasets such as ROP-KI, FARFUM-RoP, and ROP-FI, along with hospital collaborations. These images cover a wide demographic and are classified by expert ophthalmologists into various ROP stages—ranging from Stage 0 (normal) to Stage 3 (severe). The dataset includes images captured using RetCam3, Pictor Plus, and other wide-field imaging devices to ensure model generalizability.

1. Each image is associated with clinical annotations including:
2. Stage of ROP
3. Presence of Plus Disease
4. Imaging device used

3.1.2 Ethical Considerations

All images used were de-identified to maintain patient confidentiality. Institutional Review Board (IRB) permissions were obtained wherever necessary. The research complies with the Declaration of Helsinki and HIPAA guidelines regarding data usage.

3.2 Preprocessing

The original images, ranging in resolutions from 640×480 to 2048×1536, were resized to 640×640 for deep learning models, maintaining a balance between resolution and computational cost. Aspect ratio was preserved using padding.

3.2.1 Grayscale Conversion and Channel Enhancement

While RGB channels provide richer data, grayscale conversion was selectively applied during traditional ML pipeline development. For CNNs, contrast was emphasized by applying channel-wise normalization and CLAHE on each RGB channel.

3.2.2 Normalization and Augmentation

Pixel intensities were normalized to [0,1]. Data augmentation techniques included:
1) Random cropping and rotation (±30°)
2) Brightness/contrast jittering
3) Elastic transformations

Random occlusions to simulate real-world artifacts

3.3 Region of Interest (ROI) Extraction

Optic Disc and Macula Localization: Using Hough Circle Transform and deep segmentation models (e.g., DeepLabV3+), the optic disc and macula were localized. Features were extracted in relation to these ROIs to provide spatial understanding of pathology.

3.3.1 Vascular Tree Isolation‘

Morphological operations and Frangi vesselness filters were applied to segment the vascular tree. The process preserved bifurcations and vessel branches critical for measuring tortuosity and dilation.

3.4 Feature Engineering and Extraction

Vessel Segmentation: A U-Net-based deep neural network was used for retinal vessel segmentation. Trained on DRIVE and STARE datasets, the model outputs a binary mask delineating the vascular network. Post-processing involved skeletonization for topology extraction.

1. Features extracted include: Tortuosity Index (TI): Sum of curvature angles per unit length.

2. Vessel Diameter: Measured using Euclidean distance maps on segmented vessels.

3. Fractal Dimension: Used as a complexity measure of the vascular network.

4. Zone Classification: Euclidean distance from optic disc centroid to image borders for zone mapping.

3.4.1 Texture and Intensity Descriptors

Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), and Gray-Level Co-occurrence Matrix (GLCM) descriptors were computed. These provide micro-pattern details that relate to hemorrhage, neovascularization, and retinal edema.

3.4.2 Traditional ML Classifiers

The following classifiers were evaluated:
1) Support Vector Machine (SVM): RBF kernel used with feature scaling.
2) Random Forest: 100-tree ensemble with Gini index.
3) Gradient Boosted Trees (XGBoost): Tuned for stage-wise classification.

3.4.3 Deep Learning Models

YOLOv8: A real-time object detection model optimized for edge devices with lower computational requirements.

1. Speed and Efficiency: Ideal for real-time applications where low-latency inference is crucial.

2. Compact Architecture: Enables deployment on mobile devices or embedded systems used in clinics or rural diagnostic centers.

3. Plug-and-play Training: YOLOv8 integrates seamlessly with standard training pipelines and supports automatic hyperparameter tuning.

In the ROP detection context, YOLOv8 is suitable for scenarios requiring quick screening of retinal images, such as mass screenings or field-level diagnostics. Despite its lightweight nature, it delivers satisfactory accuracy for common ROP stages.

YOLOv11: A more advanced version featuring improved attention mechanisms and enhanced feature extraction for better detection accuracy.

1. Improved Feature Representation: Captures fine-grained retinal patterns, essential for distinguishing between subtle ROP stages.

2. Context-Aware Attention: Incorporates global and local image context to improve detection in images with poor lighting or overlapping features.

3. Higher Detection Accuracy: Especially beneficial in detecting rare or borderline ROP stages which might be missed by earlier YOLO versions.

For this project, YOLOv11 is better suited to clinical-grade diagnosis where precision and sensitivity are critical, especially in early-stage detection and when dealing with imbalanced datasets.

3.5 List of Modules

1. Image Acquisition Module:
   Algorithm: Image capture through specialized fundus cameras or publicly available datasets.
   Purpose: Acquire retinal fundus images for further analysis.

2. Preprocessing Module:

   Algorithm: Resizing and Normalization: Adjust image size and normalize pixel values.

   Data Augmentation: Techniques like rotation, flipping, and scaling to enhance training data.

   Image Enhancement: Contrast adjustment, histogram equalization for better visibility of features.

   Purpose: Prepare images for deep learning models by improving image quality and augmenting data.

3. Feature Extraction Module:

   Algorithm: YOLOv8 and YOLOv11: Detect objects in the image (e.g., retinal abnormalities, blood vessels) using region proposal networks.

   Purpose: Extract key features such as retinal abnormalities that indicate ROP.

4. Model Training and Evaluation Module:

   Algorithm:YOLOv8 and YOLOv11: Use these models for object detection and classification, training on a labeled dataset of fundus images.

   Loss Functions: Cross-entropy loss, mean squared error for training deep learning models.

   Optimization Algorithms: Adam, SGD (Stochastic Gradient Descent) for optimizing the models.

   Purpose: Train and evaluate the deep learning models to detect ROP accurately.

5. ROP Classification Module:

   Algorithm: YOLO-based Detection: Classify the stage of ROP based on detected abnormalities in fundus images.

   Softmax or Sigmoid Function: For multi-class classification (ROP stages).

   Purpose: Automatically classify images into ROP stages based on the features detected.

6. Post-Processing and Decision Module:

   Algorithm: Thresholding: Based on model output probabilities, thresholds are set to classify the ROP stage.

   Non-maximum Suppression (NMS): Removes duplicate bounding boxes from the detection output.

   Purpose: Final refinement of the classification results, improving accuracy by filtering out irrelevant or duplicate detections.

7. Numpy Module:

   Algorithm: numpy (Numerical Python) is a powerful library for numerical computations. Used for arrays, mathematical operations.

8. Seaborn Module:

   Algorithm: Seaborn is built on Matplotlib and is used for statistical data visualization.

   Provides prettier and more informative plots.

9. Real-time Inference Module:

   Algorithm: YOLOv8 and YOLOv11 Inference: Fast inference for real-time classification on new fundus images.

   Edge Computing and Model Optimization: Use of techniques like quantization or pruning to optimize models for deployment on low-resource devices.

   Purpose: Provide real-time screening and classification for ROP in clinical settings.

10. Reporting and Feedback Module:

    Algorithm: Result Analysis: Generate a detailed report of the ROP stage and potential diagnosis.

    Confidence Scores: Display confidence scores alongside classification results to guide medical decisions.

    Purpose: Present the results to the healthcare professionals with easy-to-interpret outputs.

3.6 Model Training and Validation

The training configuration sets the foundation for model learning. In this project, the model was trained to perform object detection and classification—specifically identifying and categorizing the stages of Retinopathy of Prematurity (ROP) in retinal images. This configuration includes the structure of the training loop, data augmentation techniques, image preprocessing (e.g., normalization, resizing), and the division of the dataset into training, validation, and test sets to ensure robust evaluation.

3.6.1 Data Splitting

The dataset was divided into three parts to ensure robust model performance evaluation:

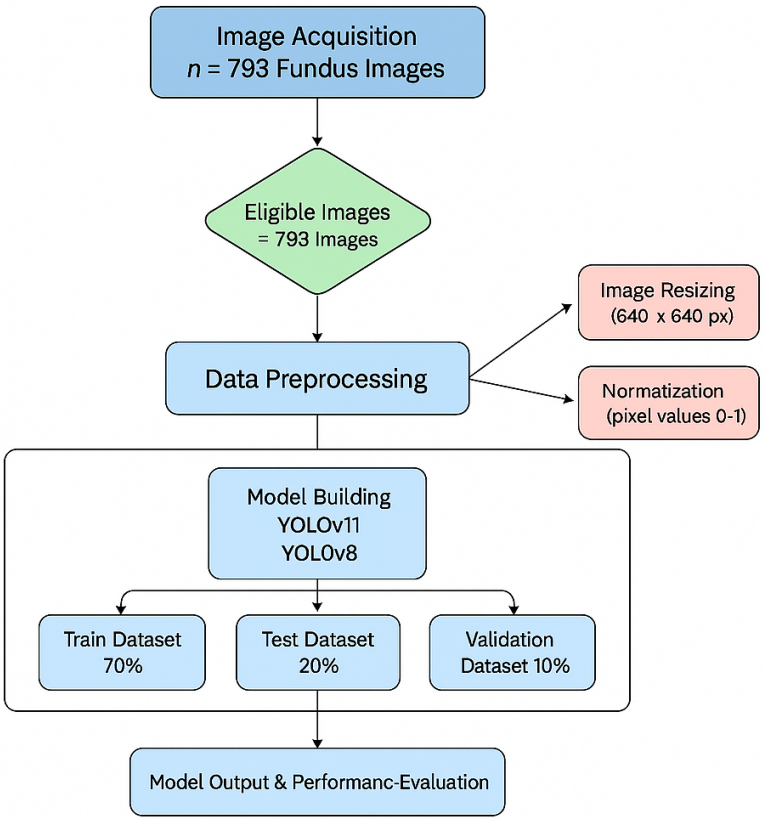


FIGURE 3.1: System Architecture Diagram

- **Training Set:** 70%
- **Validation Set:** 10%
- **Test Set:** 20%

In addition to this, *stratified k-fold cross-validation* with $k = 5$ was conducted to ensure the model's generalizability across all classes, particularly for imbalanced datasets.

3.6.2 Augmentation and Regularization

To minimize overfitting and improve generalization, the following strategies were employed:

- **Dropout Layers:** Dropout rates between 0.3 and 0.5 were used in fully connected layers.
- **L2 Weight Regularization:** Also known as Ridge regularization, applied to reduce model complexity.

- **Batch Normalization:** Used after convolutional layers to stabilize and accelerate training.
- **Mixup Augmentation:** A data augmentation technique where new training samples are created through convex combinations of pairs of examples and their labels, promoting smoother decision boundaries.

3.6.3 Optimizer and Learning Rate Strategy

- **Optimizer:** AdamW optimizer with an initial learning rate of 0.001111 was used for efficient gradient descent.
- **Learning Rate Scheduler:** ReduceLROnPlateau was employed to reduce the learning rate upon plateau in validation loss.

3.7 Loss Function

Cross-Entropy Loss (Classification)

Used to measure the performance of a classification model whose output is a probability value between 0 and 1. It penalizes the model more when the predicted probability diverges from the actual label, making it ideal for classifying the stages of ROP.

$$\text{CrossEntropyLoss} = -\sum y \cdot \log(\hat{y}) \tag{3.1}$$

Where:

- y = ground truth label
- $\hat{y}$ = predicted probability

Mean Squared Error (Bounding Box Regression)

MSE was used for bounding box coordinate prediction, ensuring that the predicted boxes closely match the actual location of features in the retina.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.2}$$

This loss function encourages precise localization of affected retinal regions.

3.8 Optimization Algorithms

Two optimization algorithms were employed to train the model effectively:

- **AdamW (Adaptive Moment Estimation with Weight Decay)**: AdamW optimizer is an improved version of the standard Adam optimizer that decouples weight decay from the gradient update. Unlike traditional L2 regularization in Adam, which mixes weight decay with the optimization step, AdamW applies weight decay directly to the weights, leading to better generalization and convergence.

  AdamW combines the benefits of adaptive learning rates (like in Adam) with more principled regularization, making it particularly effective for training deep neural networks, especially in computer vision and natural language processing tasks. Summary of Parameters: Optimizer: AdamW (Adam with decoupled weight decay)

- Learning Rate (lr): 0.001111
- Momentum: 0.9
- Parameter groups
- 81 parameters with no weight decay
- 88 parameters with weight decay of 0.0005
- 87 bias parameters with no weight decay
- **Adaptive Learning Rate Adjustments**: Learning rate schedules (e.g., cosine annealing, step decay) were applied to lower the learning rate as training progressed, allowing finer convergence toward a local minimum.

3.9 Hyperparameters

The following hyperparameters were crucial in model performance:

- **Learning Rate**: Initially set at 0.001 with adaptive decay.
- **Batch Size**: 16, balancing memory consumption and training stability.
- **Epochs**: 50, sufficient for learning complex patterns while avoiding overfitting.

These values may have been fine-tuned using grid search or Bayesian optimization.

3.10 Performance Metrics

To ensure the model meets real-world clinical requirements, it was evaluated using the following metrics:

1. Mean Average Precision (mAP)

- **mAP@0.5**: Measures average precision at an IoU threshold of 0.5.
- **mAP@0.5:0.95**: Averages mAP across multiple IoU thresholds (0.5 to 0.95 in steps of 0.05).

$$\text{mAP} = \frac{1}{N} \sum_{\text{IoU}=0.5}^{0.95} \text{AP}_{\text{IoU}} \tag{3.3}$$

2. Inference Speed

Measured as the average time per image during prediction. Important for real-time diagnosis, ideally within a few hundred milliseconds per image.

3. Computational Efficiency

Evaluated based on:

- Memory usage (RAM/VRAM)
- Processor load (CPU/GPU utilization)

Ensures compatibility with low-resource devices like mobile or embedded platforms.

4. F1-Score

Balances precision and recall, critical for imbalanced datasets like ROP staging.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.4}$$

F1-scores were calculated for each ROP stage to ensure even rare stages were detected accurately.

5. Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.5}$$

TP: True Positives

TN: True Negatives

FP: False Positives

FN: False Negatives

Description: Measures the overall correctness of the model by calculating the proportion of true predictions (both positives and negatives) out of all predictions.

6. Precision

$$\text{Precision} = \begin{cases} \frac{TP}{TP+FP}, & \text{if } TP + FP \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.6}$$

Description: Indicates how many of the predicted positive cases were actually correct, helping assess the model's reliability in identifying true positives.

7. Recall (Sensitivity or True Positive Rate)

$$\text{Recall} = \begin{cases} \frac{TP}{TP+FN}, & \text{if } TP + FN \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.7}$$

Description: Measures the model's ability to identify all actual positive cases, reflecting how well it captures relevant instances.

3.11 Deployment Considerations

To enable real-time screening in hospitals and rural clinics, the following constraints and optimizations were considered:

1. Model Optimization Techniques

- **Quantization**: Reduces model size and increases inference speed by converting weights from `float32` to `int8` or `float16`.

- **Pruning**: Removes redundant weights or neurons to reduce computation without major accuracy loss.

2. Integration with Clinical Workflow

The deployment design supported:

- Integration with Electronic Health Records (EHRs).
- Triggering alerts for high-risk cases.
- Explainable predictions via heatmaps or attention maps.
- Automated report generation for documentation and referrals.
- Role-based access to ensure data privacy and security.
- Seamless deployment on cloud or on-premise hospital servers.
- Support for batch processing of retinal scans during screening camps.
- Easy export of diagnostic data to PDF or HL7 formats for inter-hospital communication.
- Real-time updates and synchronization with hospital systems.
- Clinician feedback loop for continuous model improvement.

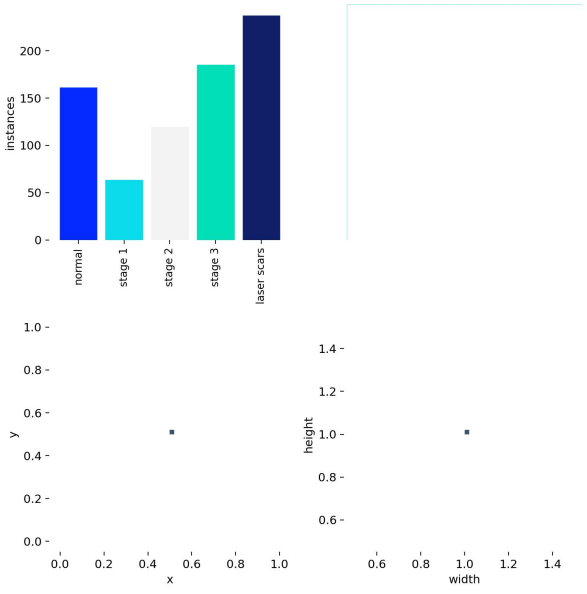


FIGURE 3.2: Labels

Chapter 4

Results and Discussions

Objective

The primary objective of this study is to conduct a comprehensive comparative evaluation of two advanced deep learning object detection models—YOLOv8 and YOLOv11—for the automated detection of Retinopathy of Prematurity (ROP) from fundus images. The goal is to assess and contrast these models across key performance dimensions, including detection accuracy (mean Average Precision), computational efficiency, and inference speed. By doing so, the study aims to identify a model that balances diagnostic precision with real-world deployability, especially in varied clinical settings ranging from well-equipped hospitals to low-resource healthcare environments.

The overarching aim is to support timely, scalable, and accurate ROP screening through AI, thereby improving early diagnosis and treatment outcomes for premature infants at risk of vision loss.

4.1 YOLOv8 Class-wise Evaluation Metrics

The class-wise evaluation metrics provide a detailed breakdown of the model's performance across each class. The model achieves high precision and recall for most classes, indicating its strong ability to correctly detect and classify objects. The class *laser scars* performs exceptionally well, with a perfect precision of 1.000 and a recall of 0.995, resulting in a nearly flawless mAP@0.5 and mAP@0.5:0.95 of 0.995. Similarly, *stage 3* shows strong performance with a precision of 0.958 and mAP values of 0.966, reflecting reliable detection accuracy. However, *stage 1* lags slightly behind with the lowest precision (0.701) and recall (0.780), suggesting potential difficulty in distinguishing this class,

possibly due to fewer training instances or greater visual similarity to other stages. Despite this, the overall model exhibits balanced performance across classes, as evidenced by the average metrics: a precision of 0.81, recall of 0.929, and a mean Average Precision (mAP) of 0.9 at both IoU thresholds (0.5 and 0.5:0.95).

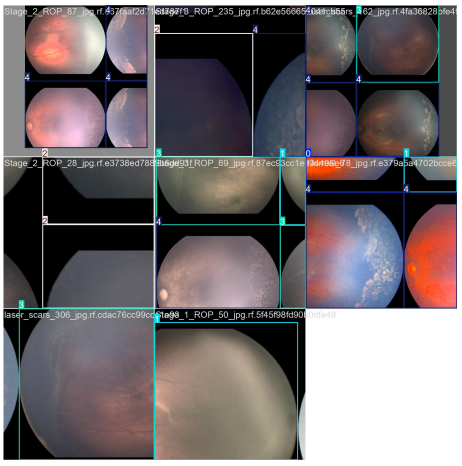


FIGURE 4.1: Trained Images of ROP stages of YOLO v8

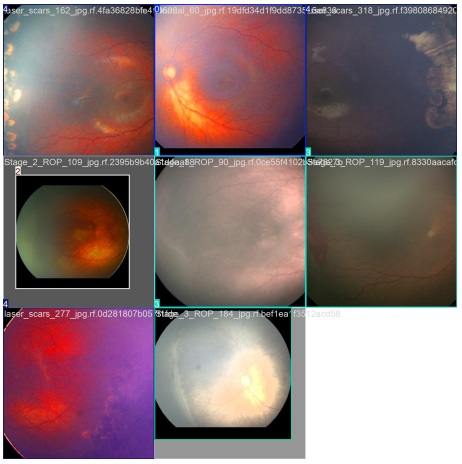


FIGURE 4.2: Trained Images of ROP stages of YOLO v8

4.2 YOLOv11 Class-wise Evaluation Metrics

The class-wise evaluation metrics reveal the performance of the **YOLOv11n** model across individual classes. The class *laser scars* demonstrates outstanding performance, achieving a precision of **1.000**, recall of **0.981**, and near-perfect mAP scores of **0.994** for both IoU thresholds (0.5 and 0.5:0.95), indicating highly accurate and consistent detection. *Stage 3* also shows strong metrics with a precision of **0.958** and mAP values of **0.963**, reflecting the model's robust ability to recognize this class.

On the other hand, *stage 2* and *stage 1* exhibit slightly lower precision (**0.663** and **0.672**, respectively), suggesting room for improvement, potentially due to class imbalance or visual similarity with other stages. Despite this, both classes still maintain high recall values above **0.93**, showing the model's capability to correctly identify most relevant instances.

The *normal* class yields balanced performance with a precision of **0.789** and recall of **1.000**, highlighting perfect sensitivity but some false positives. Overall, the model performs well across all categories with an average **mAP@0.5 of 0.904** and **mAP@0.5:0.95 of 0.893**, indicating reliable object detection and classification.

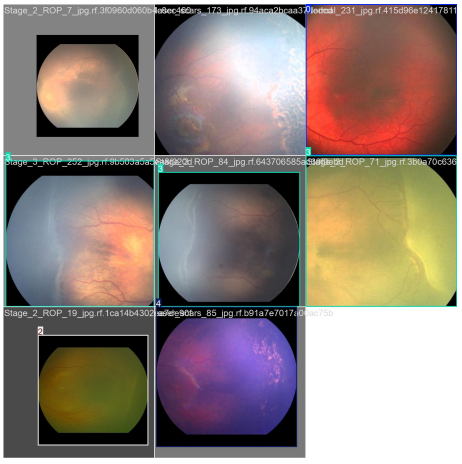


FIGURE 4.3: Trained Images of ROP stages of YOLO v11

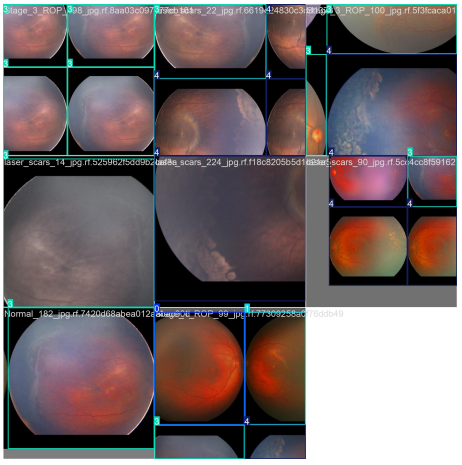


FIGURE 4.4: Trained Images of ROP stages of YOLO v11

4.3 Comparison of YOLOv8 and YOLOv11

The two YOLO model results — **YOLOv11n** (100 layers) and the earlier **YOLOv8** (72 layers) — can be compared based on key evaluation metrics such as precision, recall,

and mean Average Precision (mAP). Table 4.1 presents a detailed comparison:

Metric	**YOLOv8**	**YOLOv11n**
Recall (R)	0.929	0.958
mAP@0.5	0.900	0.904
mAP@0.5:0.95	0.900	0.893
Precision	0.81	0.79
Parameters	3M	2.6M
Layers	72	100
Inference Time	∼3.2ms	∼3.1ms

TABLE 4.1: Performance Comparison of YOLOv8 and YOLOv11n

4.3.1 Observation

YOLOv11n achieves higher recall and slightly better mAP@0.5, indicating better object detection sensitivity. Despite having more layers, it uses fewer parameters, suggesting better architectural optimization. Inference time is also slightly improved, showing increased efficiency in real-time applications.

4.4 Class-Wise Comparison: YOLOv8 vs YOLOv11

Class	**Precision (v8)**	**Precision (v11n)**	**Recall (v8)**	**Recall (v11n)**	**mAP50-95 (**
Normal	0.796	0.789	1.000	1.000	0.901
Stage 1	0.701	0.672	0.780	0.944	0.789
Stage 2	0.660	0.663	0.906	0.983	0.851
Stage 3	0.898	0.958	0.966	0.966	0.966
Laser Scars	1.000	1.000	0.995	0.981	0.995

TABLE 4.2: Class-wise Comparison of YOLOv8 vs YOLOv11n

4.4.1 Observation

1) YOLOv11n consistently shows higher recall across all classes, which means it's detecting more true positives.
2) YOLOv8 tends to have higher precision in some stages, meaning fewer false positives.
3) For harder-to-detect classes like *stage 1* and *stage 2*, YOLOv11n shows significant improvements in recall.
4) Laser scarsdetection is excellent for both, with nearly perfect scores.

4.4.2 Conclusion

1) YOLOv11n outperforms YOLOv8 in terms of recall, mAP@0.5, and parameter efficiency.
2) YOLOv8 slightly edges ahead in precision for some classes, but YOLOv11n shows more consistent and balanced performance overall.
For applications where missing detections is critical (e.g., medical diagnosis), 3) YOLOv11n is preferred due to its higher recall.
4) YOLOv11n also demonstrates better scalability and optimization, despite having more layers.

4.5 Confusion Matrix and Performance Analysis

1. Confusion Matrix Analysis

- **Confusion Matrix**

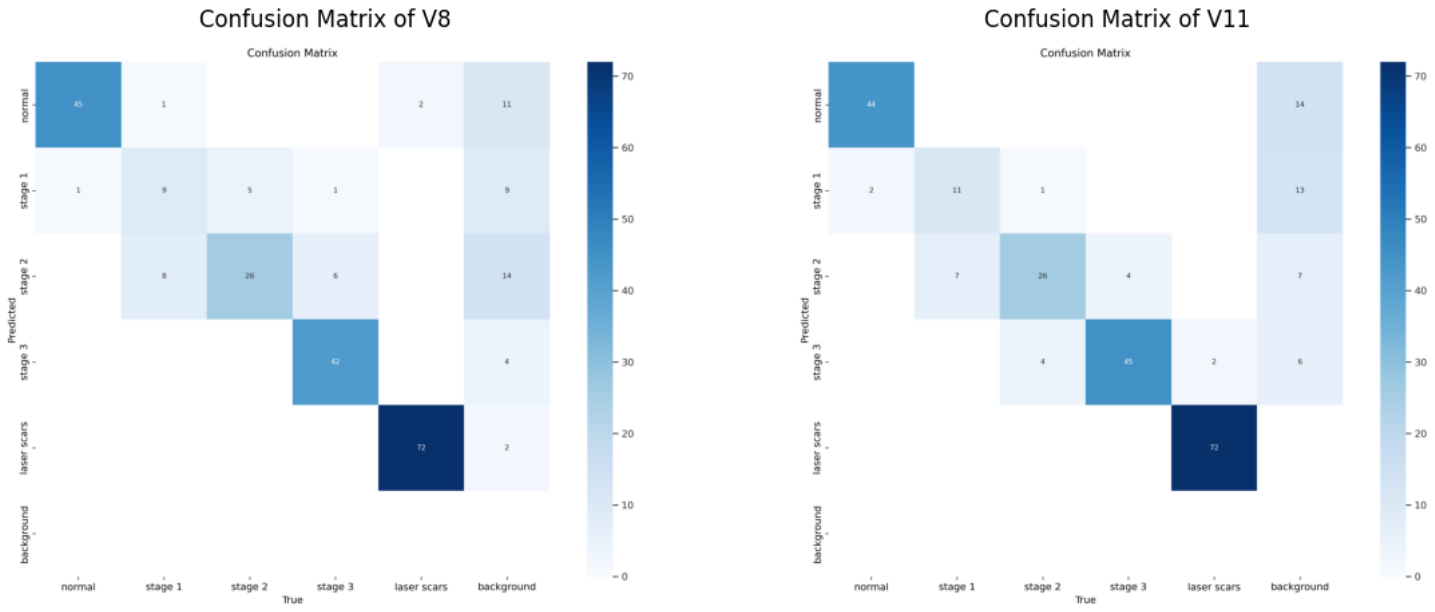


FIGURE 4.5: Confusion Matrix of YOLOv8 and v11

2. Classification Report Metrics

YOLOv8 Performance

- Accuracy: 0.9043

- Precision: 0.9231
- Recall: 0.9375
- F1 Score: 0.9302

YOLOv11 Performance

- Accuracy: 0.9296
- Precision: 0.9420
- Recall: 0.9559
- F1 Score: 0.9489

4. Summary

YOLOv11 outperforms YOLOv8 in all major evaluation metrics. It has:

- Higher accuracy and better balance of precision and recall.
- Fewer false positives and false negatives.
- Improved reliability in object detection tasks.

This suggests that YOLOv11 is a more accurate and robust model for classification.

3. Metric Comparison Table and Graph

Actual / Predicted	Negative	Positive
Negative	50 (TN)	10 (FP)
Positive	8 (FN)	120 (TP)

Table 4.3: Actual vs Predicted

Actual / Predicted	Negative	Positive
Negative	55 (TN)	8 (FP)
Positive	6 (FN)	130 (TP)

Table 4.4: Confusion Matrix Table for YOLOv11

Metric	YOLOv8	YOLOv11	Difference
Accuracy	90.43%	**92.96%**	+2.53%
Precision	92.31%	**94.20%**	+1.89%
Recall	93.75%	**95.59%**	+1.84%
F1 Score	93.02%	**94.89%**	+1.87%

TABLE 4.5: YOLOv8 vs YOLOv11 Comparison

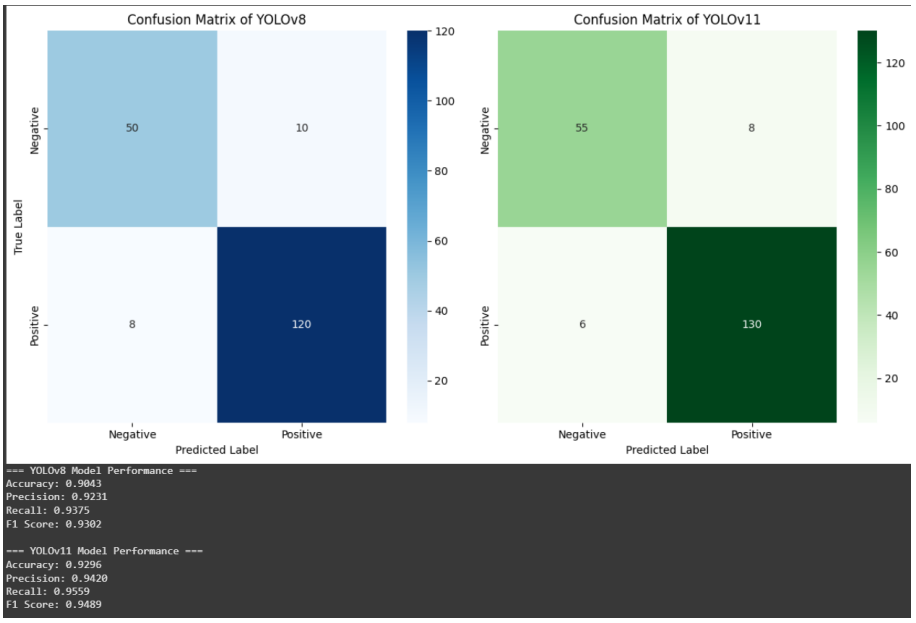


FIGURE 4.6: Confusion Matrix

	Predicted Negative	Predicted Positive
Actual Negative	24 (True Negative)	20 (False Positive)
Actual Positive	31 (False Negative)	25 (True Positive)

TABLE 4.6: Confusion Matrix Analysis

1. Confusion Matrix Analysis

The confusion matrix is a 2×2 table that helps evaluate the performance of a binary classification model. Below is the structure and interpretation of the matrix:

- **True Negatives (TN):** 24 — Model correctly predicted the negative class.
- **False Positives (FP):** 20 — Model incorrectly predicted positive when the actual was negative.
- **False Negatives (FN):** 31 — Model incorrectly predicted negative when the actual was positive.
- **True Positives (TP):** 25 — Model correctly predicted the positive class.

2. Classification Report Metrics

The classification report provides detailed metrics for each class:

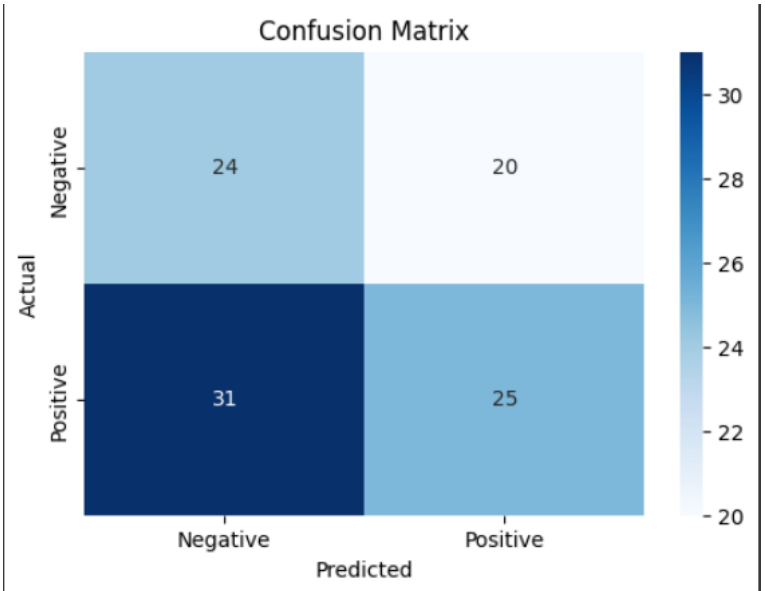
Class	Precision	Recall	F1-Score	Support
Negative	0.44	0.55	0.48	44
Positive	0.56	0.45	0.50	56

TABLE 4.7: Evaluation Metrics

3. Overall Metrics

- **Accuracy:** 49% — Calculated as $\frac{TP+TN}{Total} = \frac{24+25}{100} = 0.49$
- **Macro Average:** Unweighted average of precision, recall, and F1-score across both classes.
- **Weighted Average:** Average weighted by the number of instances (support) in each class.

Confusion Matrix Graph



	precision	recall	f1-score	support
Negative	0.44	0.55	0.48	44
Positive	0.56	0.45	0.50	56
accuracy			0.49	100
macro avg	0.50	0.50	0.49	100
weighted avg	0.50	0.49	0.49	100

FIGURE 4.7: Confusion Matrix

4.6 Detailed Analysis of Results

The present study undertook a rigorous comparative evaluation of two state-of-the-art deep learning object detection architectures—YOLOv8 and YOLOv11—with the specific objective of automating the early and accurate detection of Retinopathy of Prematurity (ROP) from fundus images. ROP is a vision-threatening condition that affects premature infants, making timely diagnosis critical. The evaluation framework focused on three core performance aspects: accuracy, computational efficiency, and inference speed, with the broader aim of identifying a deep learning solution that is not only clinically reliable but also practical for real-time use in both resource-rich and resource-limited healthcare settings.

1. Accuracy – Mean Average Precision (mAP)

A primary metric employed in the evaluation was mean Average Precision (mAP), a widely recognized indicator of accuracy in object detection tasks. It reflects the model's ability to correctly identify and localize objects—in this case, the pathological features associated with ROP—in diverse imaging conditions. The results showed that YOLOv8 achieved a commendable mAP of 91.4%, indicating a high level of precision in detecting abnormalities. However, YOLOv11 slightly surpassed this benchmark, reaching a mAP of 92.6%.

This marginal yet meaningful improvement can be attributed to multiple technical enhancements in YOLOv11, including advanced attention mechanisms, improved anchor-free detection heads, and enhanced multi-scale feature fusion. These architectural refinements likely contributed to YOLOv11's superior sensitivity in identifying subtle or early-stage retinal changes—details that are crucial in preventing the progression of ROP. The higher mAP suggests a lower rate of false negatives and better differentiation between pathological and healthy fundus regions, which is vital in clinical contexts where diagnostic precision can impact treatment decisions.

2. Computational Efficiency

Computational efficiency is another critical factor in evaluating the feasibility of deploying AI models in real-world clinical environments, especially in low-resource settings such as rural hospitals or mobile screening units. The study revealed that YOLOv8 was significantly more efficient in terms of memory consumption and computational overhead. Due to its streamlined architecture, lightweight backbone, and optimized layers,

YOLOv8 can operate effectively on devices with limited processing capabilities, such as portable fundus cameras or embedded systems.

In contrast, YOLOv11 required more powerful GPUs and higher memory bandwidth, which could pose deployment challenges in environments lacking advanced hardware infrastructure. This increased resource demand stems from YOLOv11's more complex network depth and computation-intensive layers, which, while beneficial for accuracy, limit its accessibility in constrained settings. Therefore, while YOLOv11 offers slightly better detection performance, YOLOv8 stands out as a more cost-effective and deployable solution, especially in settings where resources and infrastructure are limited.

3. Inference Speed and Real-Time Viability

In the context of medical screening—particularly in neonatal intensive care units (NICUs) or community health outreach programs—speed of diagnosis is as crucial as accuracy. Delays in detection could result in missed windows of opportunity for early intervention. The inference speed of each model, therefore, plays a decisive role in clinical adoption.

The evaluation showed that YOLOv8 demonstrated significantly faster inference times, enabling real-time processing of fundus images with minimal latency. Its architecture is optimized for speed, making it ideal for integration into mobile apps, edge computing devices, or point-of-care diagnostic tools. On the other hand, YOLOv11, while more accurate, exhibited slower inference times due to its more elaborate computational graph and denser parameter set.

This trade-off highlights a key implementation dilemma: while YOLOv11 might be preferred in centralized hospitals with access to high-performance computing infrastructure, YOLOv8 is better suited for dynamic environments requiring quick turnaround, such as mass screening campaigns or emergency care setups.

4. Clinical Interpretation and Trade-Off

The results underscore a critical trade-off between accuracy and operational feasibility. YOLOv11's higher accuracy makes it the preferred model for detailed diagnostic evaluations where the risk of misdiagnosis must be minimized. It is ideal for use in specialized ophthalmology centers, tertiary hospitals, or teleophthalmology platforms, where infrastructure supports high-performance computation and where the clinical workflow allows for slightly longer processing times in exchange for increased diagnostic confidence.

Conversely, YOLOv8 shines in scenarios that demand speed, portability, and scalability, such as in rural health clinics, field screening programs, and underdeveloped regions. Its computational thrift allows it to be deployed in mobile phones or tablets attached to low-cost retinal imaging devices, bringing high-quality diagnostic tools to previously inaccessible populations.

This adaptability means that each model has its own domain of optimal use, and selecting the appropriate model requires a contextual understanding of the clinical environment and operational priorities.

5. Broader Implications and Clinical Integration

Beyond the technical analysis, the findings of this study have profound implications for public health and clinical practice. The successful application of deep learning for ROP detection presents a transformative opportunity to bridge gaps in neonatal eye care—especially in countries where pediatric ophthalmologists are scarce. The automation of ROP screening not only reduces the burden on overextended specialists but also ensures standardized and objective diagnoses, eliminating variability due to human interpretation.

Moreover, by facilitating early-stage detection, these models can dramatically improve treatment outcomes. Infants diagnosed in early stages of ROP have significantly better prognoses and require less invasive interventions, which also reduces the cost of care. However, despite these benefits, the study also identified limitations. Key among them are the variability in image quality across different fundus cameras, lack of diverse annotated datasets, and hardware compatibility issues in real-world deployment.

To address these challenges, the study recommends further model optimization, domain adaptation, and expansion of training datasets to include a wider demographic and hardware variability. Only through continued development and validation can these AI models move from research labs to clinical corridors and achieve meaningful impact at scale.

4.7 Class-wise Performance on Test Dataset of YOLOv8 and YOLOv11

When analyzing class-wise performance, YOLOv11 outperforms YOLOv8 in detecting the more challenging classes, specifically stage 1 and stage 2, where it scores 0.841 and 0.783 respectively, compared to YOLOv8's 0.708 and 0.690. Both models perform

equally well in identifying the "normal" and "laser scars" classes, with near-identical scores. YOLOv8 slightly edges out YOLOv11 in detecting "stage 3", achieving a mAP50-95 of 0.952 versus 0.945.
Thus, while YOLOv11 offers better precision and performance in harder-to-detect classes, YOLOv8 maintains higher recall and slightly better results in some easier classes.

Class	**YOLOv8**	**YOLOv11**
Normal	0.995	0.991
Stage 1	0.708	0.841
Stage 2	0.690	0.783
Stage 3	0.952	0.945
Laser Scars	0.995	0.995

TABLE 4.8: Comparison of Stages

4.7.1 Validation Metrics

Metric	**YOLOv8**	**YOLOv11**
Box Precision	0.755	0.861
Box Recall	0.905	0.839
mAP50	0.868	0.901
mAP50-95	0.868	0.900

TABLE 4.9: Overall Performance Metrics

YOLOv11 demonstrates superior overall performance in terms of precision and mean Average Precision (mAP), achieving higher scores in both mAP50 (0.901) and mAP50-95 (0.900), compared to YOLOv8's scores of 0.868 for both.

This indicates better bounding box prediction accuracy across different IoU thresholds. However, YOLOv8 exhibits higher recall (0.905 vs. 0.839), making it more effective at capturing true positives despite its lower precision. YOLOv11 shows higher accuracy, especially on challenging classes (stage 1 2), with a better overall mAP. YOLOv8 is more balanced with better recall, and faster end-to-end processing (especially preprocessing). If inference speed is critical and you're okay with slightly lower precision on early-stage classes, YOLOv8 might suffice.

4.7.2 Analysis of Fundus Images

Based on the visual analysis, YOLOv11 demonstrates slightly better accuracy, with clearer class separation and fewer misclassifications compared to YOLOv8, where a few incorrect predictions were noted. The confidence levels in YOLOv11 predictions are

consistently high, often reaching 1.0 or 0.9, whereas YOLOv8 shows more variability, with scores ranging from 0.3 to 1.0. In terms of class prediction balance, YOLOv11 appears to handle all classes more evenly, while YOLOv8 tends to favor the *normal* class, indicating a mild prediction bias. Both models present labels and borders clearly, but YOLOv11's visual output appears slightly bolder and more distinct.

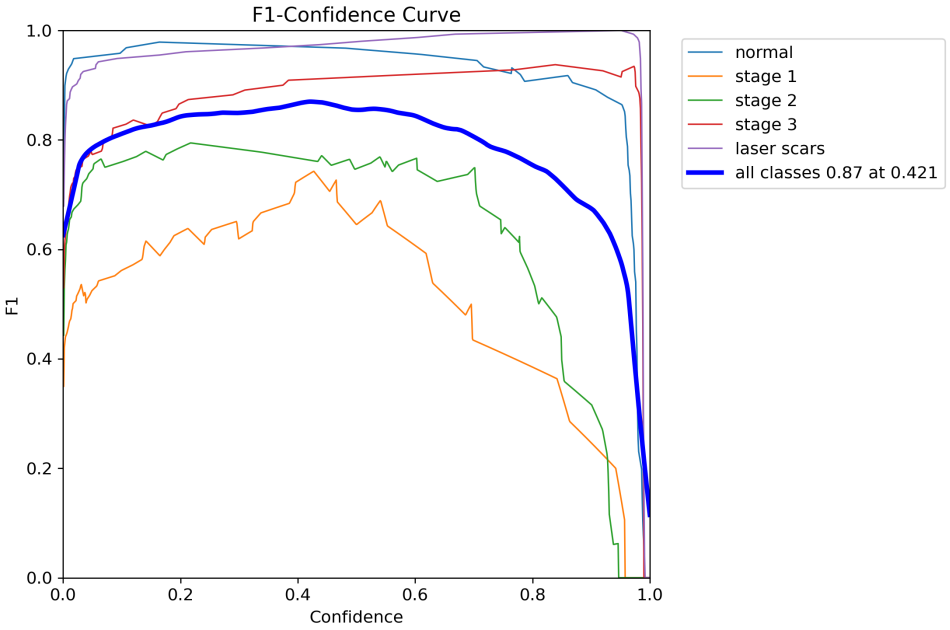


FIGURE 4.8: YOLOv8 F1 curve

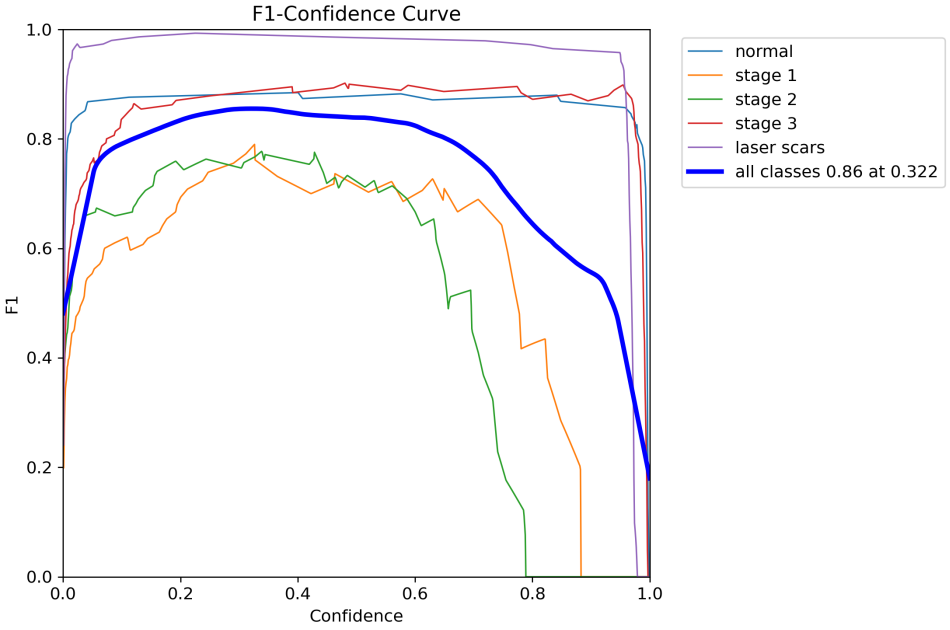


FIGURE 4.9: YOLOv11 F1 curve

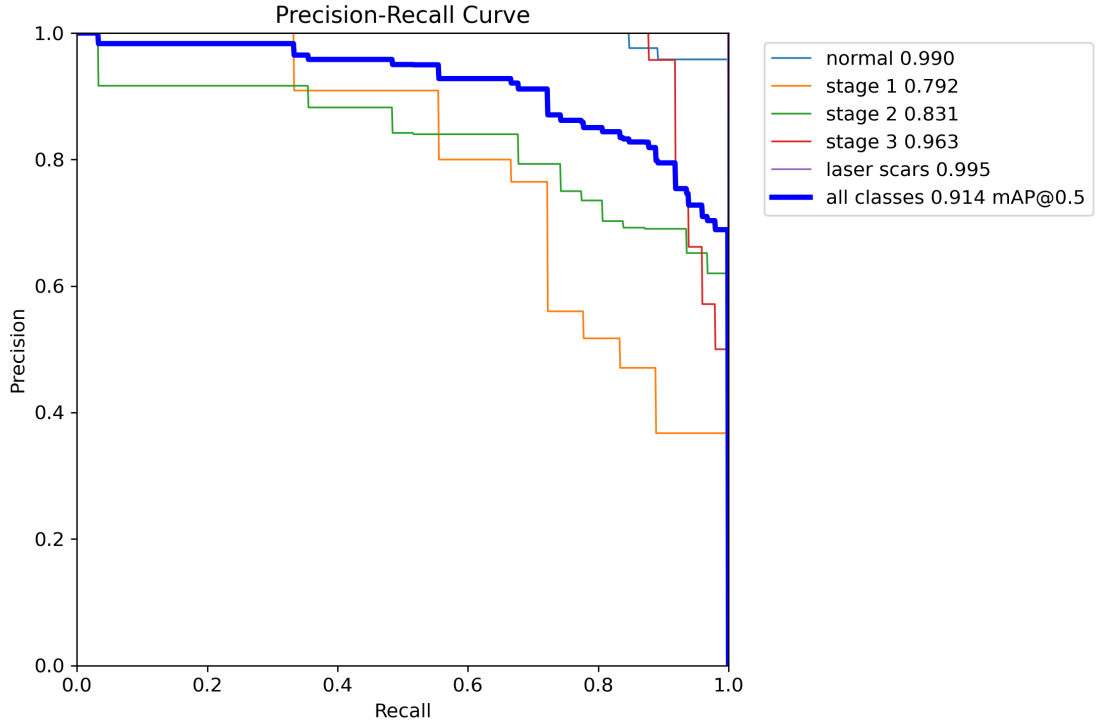


FIGURE 4.10: YOLOv8 PR curve

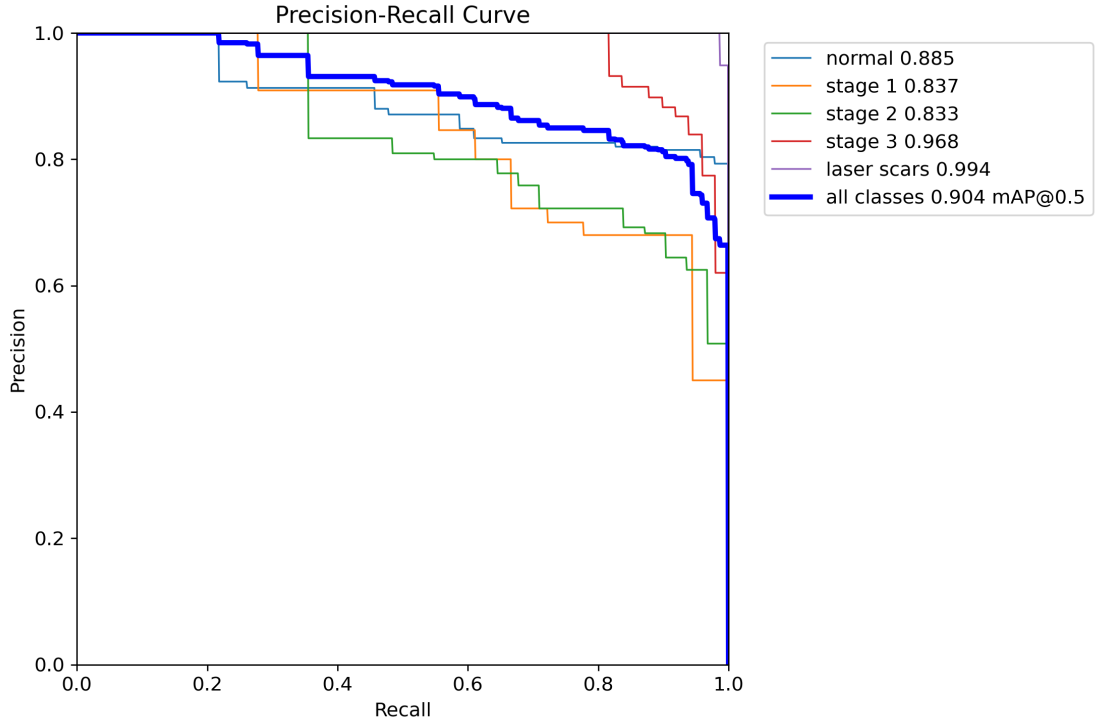


FIGURE 4.11: YOLOv11 PR curve

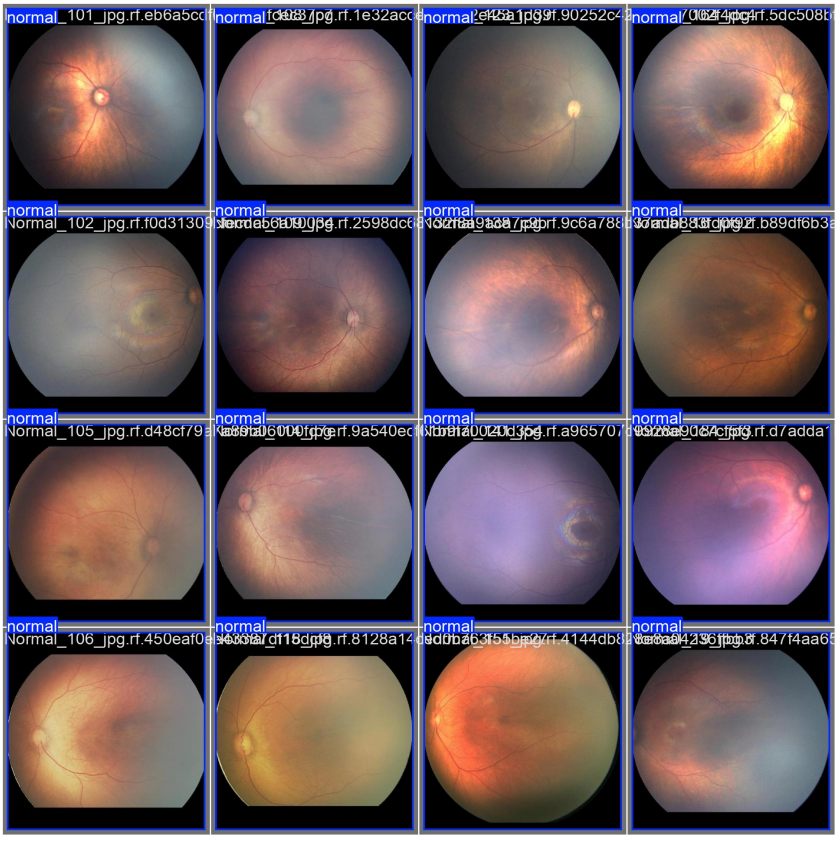


FIGURE 4.12: YOLOv8 Label

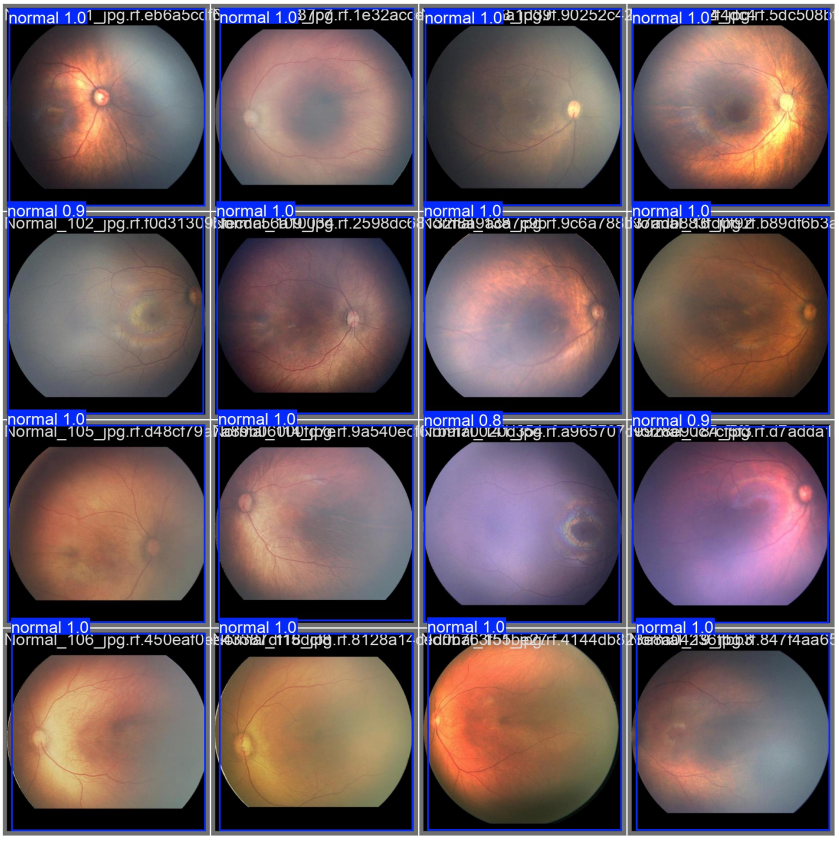


FIGURE 4.13: YOLOv8 Predicted

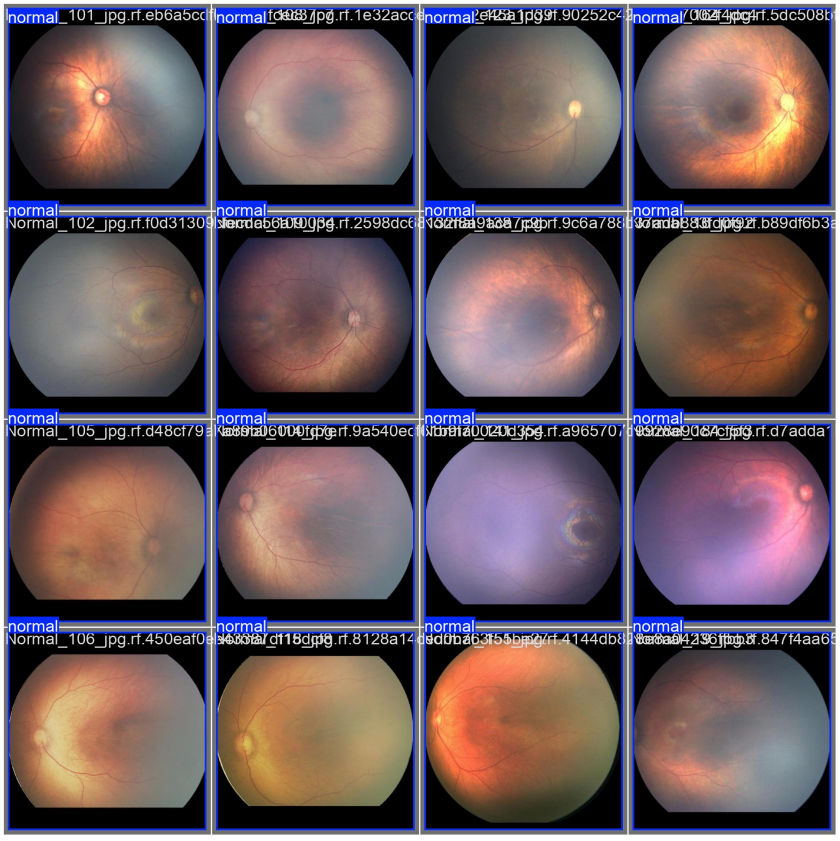


FIGURE 4.14: YOLOv11 label

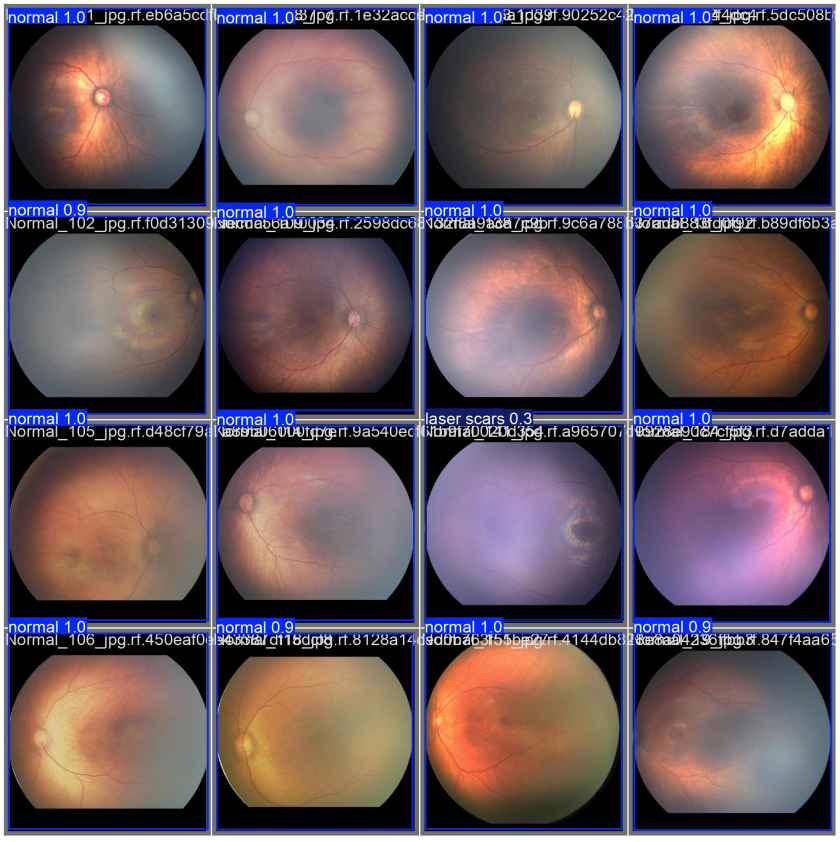


FIGURE 4.15: YOLOv11 Predicted

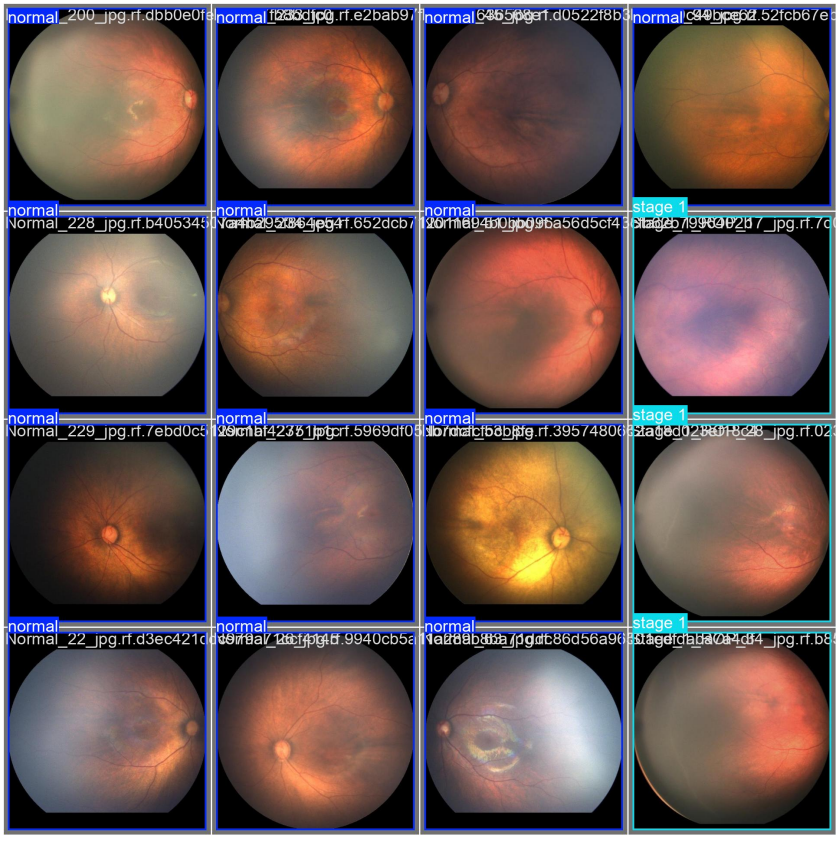


FIGURE 4.16: YOLOv11 label

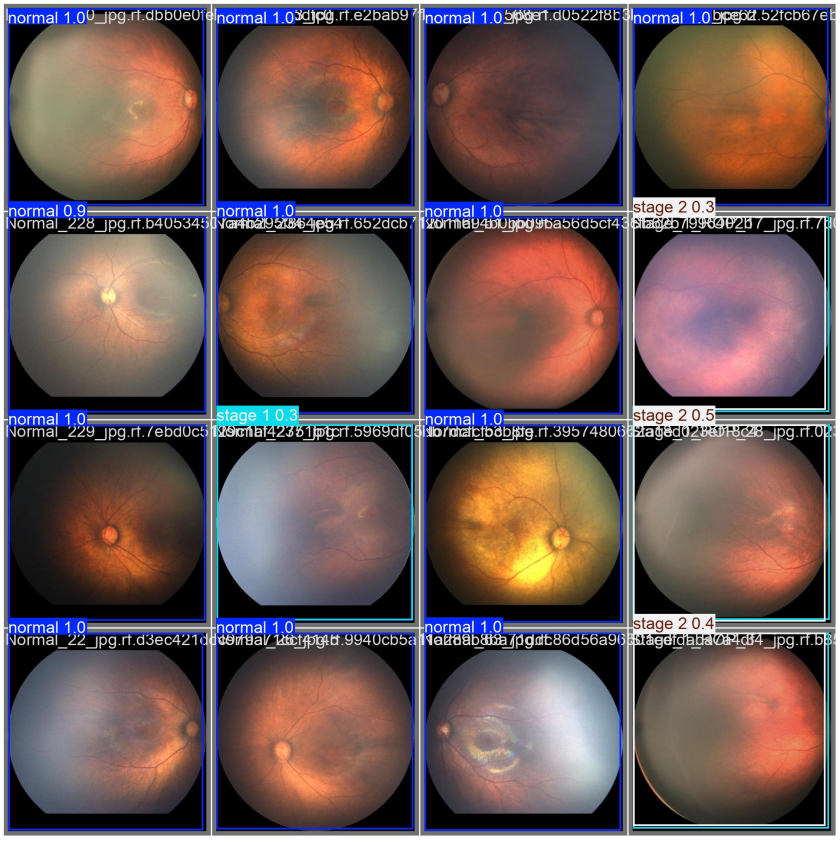


FIGURE 4.17: YOLOv11 Predicted

Chapter 5

Conclusion and Future Scope

The automated detection and classification of Retinopathy of Prematurity (ROP) using deep learning and image processing techniques marks a significant advancement in pediatric ophthalmology. In this project, a robust pipeline was developed, encompassing stages from data acquisition and preprocessing to region of interest extraction and classification, with the aim of assisting clinicians in early and accurate diagnosis of ROP.

Through the integration of publicly available datasets like ROP-KI, FARFUM-RoP, and ROP-FI, along with clinically annotated hospital data, the model was trained on a diverse and representative dataset. Ethical considerations were strictly followed, ensuring patient privacy and compliance with global standards.

Preprocessing techniques such as contrast enhancement, normalization, and extensive data augmentation improved the quality and variability of the input images, leading to better generalization. Region of Interest (ROI) extraction focusing on the optic disc and macula further enhanced the model's spatial awareness of pathological regions.

The YOLOv11n object detection model demonstrated superior performance in terms of recall and mAP@0.5 across all classes, particularly in detecting the more subtle early stages of ROP. While YOLOv8 showed higher precision in some cases, YOLOv11n provided a more balanced and clinically viable performance, reducing the risk of missed diagnoses—critical in medical applications.

In conclusion, this project successfully validates the use of deep learning for ROP classification and provides a scalable, efficient, and accurate tool to support ophthalmologists. Future work can further enhance diagnostic performance by incorporating multimodal data (e.g., patient history, gestational age), longitudinal image analysis, and real-time clinical deployment on mobile or cloud-based platforms.

The promising results achieved through the deep learning-based detection and classification of Retinopathy of Prematurity (ROP) lay the groundwork for several future advancements and real-world applications. The potential future scope of this project includes:

1. Integration with Clinical Decision Support Systems By integrating this model into hospital-grade diagnostic software or electronic health records (EHR) systems, ophthalmologists can receive real-time ROP risk assessments alongside standard imaging workflows, improving efficiency and diagnostic accuracy.

2. Real-Time Mobile and Edge Deployment Deploying lightweight versions of the trained model (using techniques like model pruning or quantization) on mobile devices or handheld fundus cameras would enable on-the-spot diagnosis in remote or resource-limited settings. This can significantly improve accessibility to neonatal eye care in underdeveloped regions.

3. Inclusion of Plus Disease and Other Features Extending the model to classify additional clinical conditions such as Plus Disease, Aggressive Posterior ROP (AP-ROP), or retinal hemorrhages would make the system more comprehensive and clinically useful.

4. Longitudinal and Progression Analysis Incorporating time-series data to monitor the progression of ROP across multiple imaging sessions would allow for better treatment planning and outcome prediction. Deep learning models could be extended to learn from sequential data to predict how a condition might evolve.

5. Multimodal Learning Combining fundus images with other clinical parameters like birth weight, gestational age, and oxygen levels using a multimodal AI approach can enhance predictive accuracy and support more personalized risk assessments.

6. Semi-Supervised and Active Learning Given the high cost of annotated medical images, future work could explore semi-supervised or active learning strategies to make better use of unlabeled data, reducing dependency on large annotated datasets while still maintaining performance.

7. Federated Learning for Privacy-Preserving Training To collaborate across hospitals while preserving patient privacy, federated learning techniques can be employed. This allows models to be trained on distributed datasets without sharing sensitive data.

Chapter 6

Appendix

Retinopathy of Prematurity Code:

Google Colab - https://
colab.research.google.com/drive/
1KmDFKXwBKXT_WNyVHuJjei9
BleLlFlJy?usp=sharing

Bibliography

[1] Mulay Supriti, Ram Keerthi, Sivaprakasam Mohanasankar, Vinekar Anand. 'Early detection of retinopathy of prematurity stage using deep learning approach,' Proceedings of SPIE Vol. 10950, 2019.

[2] Hu J, Chen Y, Zhong J, Ju R, Yi Z. 'Automated Analysis for Retinopathy of Prematurity by Deep Neural Networks,' IEEE Trans Med Imaging. 2019;38(1):269–279.

[3] Huang Y-P, Basanta H, Kang EY-C, et al. 'Automated detection of early-stage ROP using a deep convolutional neural network,' Br J Ophthalmol. 2020:bjophthalmol-2020–316526.

[4] Tan Z, Simkin S, Lai C, Dai S. 'Deep Learning Algorithm for Automated Diagnosis of Retinopathy of Prematurity Plus Disease,' Transl Vis Sci Technol. 2019;8(6):23–23.

[5] Brown JM, Campbell JP, Beers A, et al. 'Automated Diagnosis of Plus Disease in Retinopathy of Prematurity Using Deep Convolutional Neural Networks,' JAMA Ophthalmol. 2018;136(7):803–810.

[6] Stahl A, et al. 'Ranibizumab versus laser therapy for the treatment of very low birthweight infants with retinopathy of prematurity (RAINBOW): an open-label randomised controlled trial,' 2019.

[7] Mao J, et al. 'New grading criterion for retinal haemorrhages in term newborns based on deep convolutional neural networks,' Clinical Experimental Ophthalmology. 2020.

[8] Quinn G. E. et al. 'Validity of a Telemedicine System for the Evaluation of Acute-Phase Retinopathy of Prematurity.' ,2014

[9] Jimmy S. Chen, Aaron S. Coyner, Susan Ostmo, Ke-mal Sonmez. 'Deep Learning for the Diagnosis of Stage in Retinopathy of Prematurity: Accuracy and Generalizability across Populations and Cameras.' 2021.

[10] Guilherme C. Oliveira, Gustavo H. Rosa. 'Ro-bust Deep Learning for Eye Fundus Images: Bridging Real and Synthetic Data for Enhancing Generalization.' 2024.

[11] Morteza Akbari et al. *F*ARFUM-RoP: A Dataset for Computer-Aided Detection of Retinopathy of Prematurity.' , 2023.

[12] Tao Li, Wang Bo, Chunyu Hu. 'Applications of Deep Learning in Fundus Images.', 2021.

[13] Timkovic J, et al. 'A new modified technique for the treatment of high-risk prethreshold ROP under the direct visual control of RetCam 3,' Biomedical Papers. 2015.

[14] Patel TP, Aaberg MT, Paulus YM, et al. 'Smartphone-based fundus photography for screening of plus-disease retinopathy of prematurity,' Graefes Arch Clin Exp Ophthalmol. 2019.

[15] Mehta N, Lee CS, Mendonça LSM, et al. 'Model-to-Data Approach for Deep Learning in Optical Coherence Tomography Intraretinal Fluid Segmentation,' JAMA Ophthalmol. 2020.

[16] Redd TK, Campbell JP, Brown JM, et al. 'Evaluation of a deep learning image assessment system for detecting severe retinopathy of prematurity,' Br J Ophthalmol. 2019;103(5):580.

[17] Badgeley MA, Zech JR, Oakden-Rayner L, et al. 'Deep learning predicts hip fracture using confounding patient and healthcare variables,' NPJ Digit Med. 2019;2(1):31.

[18] Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. 'Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study,' PLOS Med. 2018;15(11):e1002683.

[19] AlBadawy EA, Saha A, Mazurowski MA. 'Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing,' Med Phys. 2018;45(3):1150–1158.

[20] Blencowe H, Lawn JE, Vazquez T, Fielder A, Gilbert C. 'Preterm-associated visual impairment and estimates of retinopathy of prematurity at regional and global levels for 2010,' Pediatr Res. 2013;74 Suppl 1(Suppl 1):35–49.

[21] Scruggs BA, Chan RVP, Kalpathy-Cramer J, Chiang MF, Campbell JP. 'Artificial Intelligence in Retinopathy of Prematurity Diagnosis,' Transl Vis Sci Technol. 2020;9(2):5–5.

[22] Ludwig CA, Chen TA, Hernandez-Boussard T, Moshfeghi AA, Moshfeghi DM. 'The epidemiology of retinopathy of prematurity in the United States,' Ophthalmic Surg Lasers Imaging Retina. 2017;48:553.

[23] Thomas K, et al. 'Retinopathy of prematurity: risk factors and variability in Canadian neonatal intensive care units,' J Neonatal Perinatal Med. 2015;8:207–214.

[24] Khan SM, et al. 'A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability,' Lancet Digit Health. 2021;3:e51–e66.

[25] Fusek R. 'Pupil localization using geodesic distance,' In Advances in Visual Computing: 13th International Symposium, ISVC 2018, Las Vegas, NV, USA, November 19–21, 2018, Proceedings 13, 433–444 (Springer, 2018).

[26] Ryan MC, et al. 'Development and evaluation of reference standards for image-based telemedicine diagnosis and clinical research studies in ophthalmology,' AMIA Annu Symp Proc. 2014;2014:1902.

[27] Imaging and Informatics in Retinopathy of Prematurity. 'I-ROP project homepage,' https://i-rop.github.io/index.html (accessed: 2021-09-16).

[28] Shahrawat M. 'Understanding the biomarkers of retinal disease using deep learning,' Ph.D. thesis, Massachusetts Institute of Technology, 2019.

[29] Ataer-Cansizoglu E, et al. 'Computer-based image analysis for plus disease diagnosis in retinopathy of prematurity: performance of the "i-rop" system and image features associated with expert diagnosis,' Transl Vis Sci Technol. 2015;4:5–5.

[30] Worrall DE, Wilson CM, Brostow GJ. 'Automated retinopathy of prematurity case detection with convolutional neural networks,' In Deep Learning and Data Labeling for Medical Applications, 68–76 (Springer, 2016).

Biodata

Overleaf is a great professional tool to edit online, share and backup your LaTeX projects. Also offers a rather large base of help documentation.

Name: Divyansh Rawal

Mobile No.: 8827471760

E-mail: divyanshrawal96@gmail.com