Volume 3, Number 3, 2026 Pages 1-25

Bioscience Insights

ISSN: 3080-3489 (Print)
ISSN: 3080-3497 (Online)

BiosciInsights.com

EYENOVA: AN EXPLAINABLE DEEP LEARNING PROTOTYPE FOR DIABETIC RETINOPATHY SEVERITY GRADING AND LESION LOCALIZATION

Syed Zain Mir[*1], Syed Rizwan[2], Shahzad Karim Khawer[3], Syed Haider Abbas Naqvi[4], Ashraf Zia[*5], Shahida Batool[6]

*[*1] Faculty of Engineering Sciences & Technology, Iqra University, Karachi, Pakistan.*
[2] Faculty of Engineering Sciences & Technology, Iqra University, Karachi, Pakistan.
[3] Faculty of Engineering Sciences & Technology, Iqra University, Karachi, Pakistan.
[4] Faculty of Engineering Sciences & Technology, Iqra University, Karachi, Pakistan.
*[*5] Department of Computer Science, Abdul Wali Khan University Mardan, Mardan, KP, Pakistan.*
[6] Department of Zoology, University of Education, Lahore, Pakistan.

[*1]szainmir@iqra.edu.pk, [2]syed.rizwan@iqra.edu.pk, [3]shahzad.karim@iqra.edu.pk, [4]haider@iqra.edu.pk, [*5]ashrafzia@awkum.edu.pk, [6]shahidabatool349@gmail.com

https://doi.org/10.5281/zenodo.23411622

Article Information

Article History

Received:	June 19, 2026
Revised:	August 19, 2026
Accepted:	September 07, 2026
Available Online:	October 16, 2026

Keywords:

Diabetic retinopathy, fundus imaging, EfficientNetB0, YOLOv8n, lesion localisation, explainable artificial intelligence, Grad-CAM, screening-support system.

Abstract

Diabetic retinopathy is a major cause of preventable vision loss, and timely screening is essential because early disease stages may remain asymptomatic. This study presents EyeNova, a locally deployable web-based prototype for diabetic retinopathy screening support using retinal fundus images. The prototype combines image validation, preprocessing, image-quality assessment, five-class DR severity grading, exploratory lesion localization, Grad-CAM visualization, dashboard presentation, and structured report generation. For severity grading, an EfficientNetB0 model initialized with ImageNet weights was trained on the Kaggle Diabetic Retinopathy 224x224 Gaussian Filtered dataset using a class-folder-wise 80/20 random train–validation split with seed 42, resulting in 2,805 training images and 704 validation images across five DR severity classes. For lesion localisation, a YOLOv8n detector was trained on a YOLO-format reconstruction of the FGADR Seg-set pixel-level lesion masks, using an image-level 80/20 stratified train-validation split by DR grade with 1,476 training images, 366 validation images, and 103,354 lesion annotations across six lesion classes. On the available internal validation split, the classifier achieved 98.44% accuracy, 98.45% weighted precision, 98.44% weighted recall, 99.46% weighted specificity, and 98.43% weighted F1-score, with an average inference time of 24.61 ms per image. In contrast, lesion localisation remained the limiting subsystem, with 59.70% precision,

21.27% recall, 18.44% mAP@0.50, and 7.30% mAP@0.50:0.95. These findings indicate that EyeNova is promising as a preliminary explainable screening-support prototype for severity grading, but its evidence remains limited to internal validation, with no independent external test set currently available. Accordingly, EyeNova should not be interpreted as a clinically validated diagnostic system.

INTRODUCTION

Diabetic retinopathy continues to be the most prevalent retina disorder due to diabetes which can result in irreversible blindness when it is not detected early enough and treated. Fundus photographs can therefore be used in screening because they are non-invasive, cost effective and useful in screening multiple patients at once. The problem however is that manual interpretation of the images requires many resources and the expertise of retinal specialists which is not readily available. [1]-[3]

Recent works in the field of deep learning have provided reasonable image-classification tools, particularly due to the application of transfer learning and convolutional neural networks [13]. Moreover, there was a study devoted to lightweight models in the scope of limited computational capacities. Nevertheless, the surveys prove that merely a good image classification is not sufficient for clinicians in order to be able to trust these systems. They need not only good results but also easily comprehensible ones, image quality control, and the location of lesions. [2]

The system EyeNova is aimed at filling the niche of a certain lack. EyeNova does not provide just a severity grade because it is a locally deployable web-based screening tool for fundus images. In particular, the system EyeNova provides the validation and preprocessing of such images, their quality control, severity assessment of DR, attempts to identify the location of lesions, and provides a Grad-CAM heatmap. All the information is provided in a clinician dashboard and in a report. Thus, this paper focuses on EyeNova as an example of internal validation.

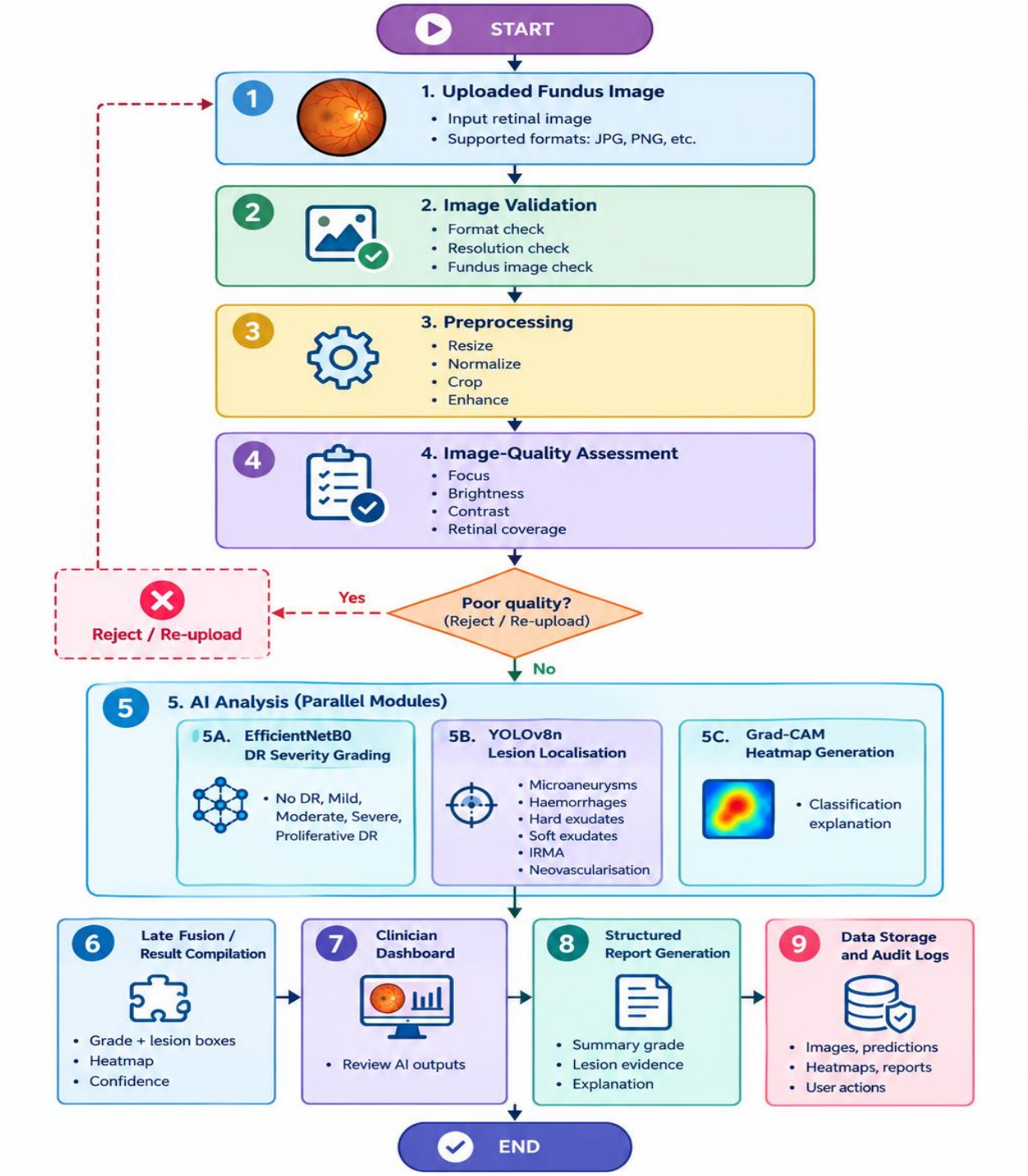


Figure 1. EyeNova Diabetic Retinopathy Screening-Support System Workflow

2. Related Work

Diabetic retinopathy diagnostic systems based on deep learning algorithms can be divided into three categories. The first category deals with the classification or grading of the disease severity through CNN or transfer learning-based techniques. Here, one can include such approaches as ensemble CNN grading, VGG-based grading, lightweight screening with MobileNet architectures, as well as recent custom deep models.

The second approach is aimed at the detection of lesions using detector-based architectures (e.g., YOLO, etc.).

Finally, the third approach attempts to integrate classification and lesion detection into one process. However, a key problem which constantly emerges in the review of this field is that despite the ability of

these algorithms to achieve extremely high benchmark accuracy, their integration into clinical practice still presents challenges. [17]

EfficientNet is suitable for medical image grading at a student level because of its compound scaling technique, which scales well with respect to depth, width, and resolution. Grad-CAM remains the most useful post hoc visual explanation tool for the CNN classifier. YOLOv8 becomes pertinent in the case of lesion localization as Ultralytics describes it as a family of anchor-free, flexible models including training, validation, inference, and export. Nevertheless, it is important to remember that YOLOv8 is to be cited as software/documentation, not as a research paper. [4][6]

The choice of a dataset greatly influences the results of a study. FGADR is a benchmark for fine-grained diabetic retinopathy research, which consists of 2,842 images. Out of them, 1,842 have been annotated by lesions, such as microaneurysms and hemorrhages. As for Messidor, it is a regular benchmark, containing 1,200 color photographs with expert retinopathy grading. According to surveys, EyePACS and APTOS 2019 are widely used in grading publicly. Lesion detection usually includes DDR and IDRiD datasets. [9], [11], [12]

Recent advancements in artificial intelligence, deep learning, and intelligent monitoring systems have demonstrated significant potential for improving automated disease detection, medical image interpretation, and clinical decision support. Khalil et al. [26] investigated advanced machine-learning techniques for early brain tumor detection using MRI images, highlighting the importance of computational approaches in medical image analysis. Similarly, Khalil et al. [27] explored deep neural networks for predictive analytics and real-time monitoring in heart transplantation, demonstrating the broader applicability of AI-driven clinical decision-support systems. Majeed et al. [28] examined mathematical optimization and high-dimensional data analysis techniques relevant to improving machine-learning model development.

Furthermore, Mir et al. [29] presented an IoT-driven intelligent energy monitoring framework, while Mir et al. [30] investigated the integration of artificial intelligence and IoT technologies for monitoring and predictive maintenance in solar power systems, illustrating the application of intelligent automated monitoring architectures. Siddiqui et al. [31], with Syed Zain Mir as a co-author, explored machine-learning-based maximum air temperature prediction for heatwave identification in Pakistan, highlighting the potential of predictive modelling for complex real-world datasets. Although these studies address different application domains, they provide relevant methodological perspectives on intelligent monitoring, predictive analytics, and machine-learning-based decision support. Building upon these developments and the established literature on retinal image analysis, the present study introduces EyeNova, an explainable deep-learning prototype integrating EfficientNetB0-based diabetic retinopathy severity grading, YOLOv8n-based lesion localization, Grad-CAM visual explanations, and structured clinical reporting. Unlike conventional classification-only approaches, EyeNova emphasizes the integration of multiple analytical components into a unified screening-support workflow while acknowledging the need for independent clinical validation.

3. Problem Statement

However, there are fewer studies dealing with integrated screening-support systems in deep learning. Normally, the studies show decent results in either one or another, but not both [14]. However, there are relatively few end-to-end solutions for performing severity grading, image quality assessment, giving lesion evidence and visual explanation, and generating structured report. EyeNova attempts to close this gap with all the existing limitations. The first is that they test classifiers only on the validation dataset, the origin of the classifier dataset remains undisclosed, the localization of lesions is inferior to the grading part, and it wasn't yet external evaluated.

4. Contributions

The main contributions of this study are as follows:

1. The solution combines image validation, pre-processing, image quality evaluation, DR grading, lesion localization, Grad-CAM explanation, dashboard, and report generation into one single local or web-based screening process.

2. Uses EfficientNetB0 transfer learning for five-class classification and provides the internal validation results in an open way.
3. Uses YOLOv8n lesion localization in an exploratory approach and provides transparency about its detection results, along with the known shortcomings of recall and mean average precision.
4. Provides a modular decision support system framework for further development through external validation, improved detector, calibration, and usability analysis by clinicians.

5. Methods and System Design

5.1 Datasets and Experimental Protocol

The dataset used for the classifier model was obtained from Kaggle Diabetic Retinopathy fundus-images dataset provided by Shajin R. P. This data can be found at https://www.kaggle.com/datasets/shajinrp/diabetic-retinopathy dated 5th June 2026 and it has an ODC Public Domain Dedication and License (PDDL). After local pre-processing, the dataset contained 3,509 images which included 2,805 training images and 704 validation images without any test set images. Before training, the images were resized into 224 × 224 pixels, normalized, and then augmented using methods such as image flips horizontally, a maximum rotation of 0.05 radians, and a random zoom of 0.1. On the other hand, the dataset used for lesion localization is the FGADR lesion dataset transformed to YOLO. The FGADR documentation states that there are 2,842 total images in their dataset consisting of 1,842 images from their Seg-set with lesions annotated and another 1,000 images from their Grade-set. The YOLO train-validation split, image counts, lesion counts, and class-wise annotation distribution must be documented before submission. The summary of available datasets shown in table 1.

Table 1: Dataset summary

Dataset	Role in EyeNova	Available size / split information	Labels / classes	Use in Present Study and Limitations
Kaggle Diabetic Retinopathy 224x224 Gaussian Filtered	Five-class DR severity grading	3,509 images total; 2,805 training images; 704 validation images; class-folder-wise random 80/20 split with seed 42; no independent test set	Mild, Moderate, No_DR, Proliferate_DR, Severe	Used to train and internally validate the EfficientNetB0 classifier. Dataset was accessed on 4 March 2026 and is distributed under CC0: Public Domain. A later audit found 0 filename duplicates but 12 SHA-256 hash overlaps across train/validation; therefore, results are reported as internal validation with duplication-risk limitation.
FGADR Seg-set	Exploratory YOLO lesion localisation	1,842 Seg-set images; 1,476 training images; 366 validation images; image-level 80/20 stratified split by FGADR DR grade with seed 42; no	Microaneurysm, hemorrhage, hard_exudate, soft_exudate, IRMA, neovascularization	Used to train and internally validate the YOLOv8n lesion-localisation model. YOLO labels were reconstructed from pixel-level masks using 8-connected foreground components, producing 103,354 lesion annotations across six classes. Duplicate checking found 0 duplicate basenames and 0 duplicate

		independent test set		SHA-256 hashes across train/validation.
EyePACS / APTOS 2019	Proposed external grading validation	Not used in the current EyeNova evaluation	DR severity grading labels	Not used in the current study; proposed for future external validation to assess classifier generalisation beyond the current internal Kaggle split.
Messidor	Proposed external grading validation	1,200 colour fundus images with expert retinopathy grading	Retinopathy grade and macular-oedema risk	Not used in the current study; proposed for future cross-dataset validation using expert retinopathy grades and macular-oedema risk labels.
DDR / IDRiD	Proposed lesion-detection benchmarking	Not used in the current EyeNova evaluation	DR grades and/or lesion annotations depending on dataset subset	Not used in the current study; proposed for future lesion-detection benchmarking and external validation of localisation performance.

Grading data was unbalanced across the split that could be used, having Proliferative DR and No DR as the biggest classes while Moderate DR represented the smallest class. Thus, the evaluation based solely on overall accuracy was deemed unsuitable and per-class metrics of precision, recall, specificity, and F1 score were also calculated. Since there was no documentation of the process that generated the split to verify whether this was a random, stratified, patient-based or pre-specified split, there is an inherent possibility of sampling error and data leakage.

5.2 Preprocessing and image-quality assessment

The project documentation describes a preprocessing pipeline in figure 2. that resizes retinal images, normalises pixel values, improves contrast when needed, Noise reduction and background cropping were performed before training using Gaussian blur, fixed binary thresholding, morphological closing and opening, and largest-connected-component analysis. A grayscale threshold of 15 was used to generate the retinal/background mask. Both the Gaussian and morphological kernels were set to 9 × 9, and the Gaussian blur parameter was $\sigma = 0.$, and converts the image into a model-compatible tensor. Before inference, each uploaded fundus image was evaluated using a rule-based image quality assessment module. The module checked image focus using blur estimation, illumination using brightness-level analysis, contrast using intensity variation, and retinal coverage using fundus-region visibility. Images failing the predefined quality checks were not passed to the classification and lesion-localization models and were instead flagged for re-upload. Image-quality assessment was included because poor focus, illumination artefacts, and incomplete retinal coverage can degrade downstream DR grading. [18]

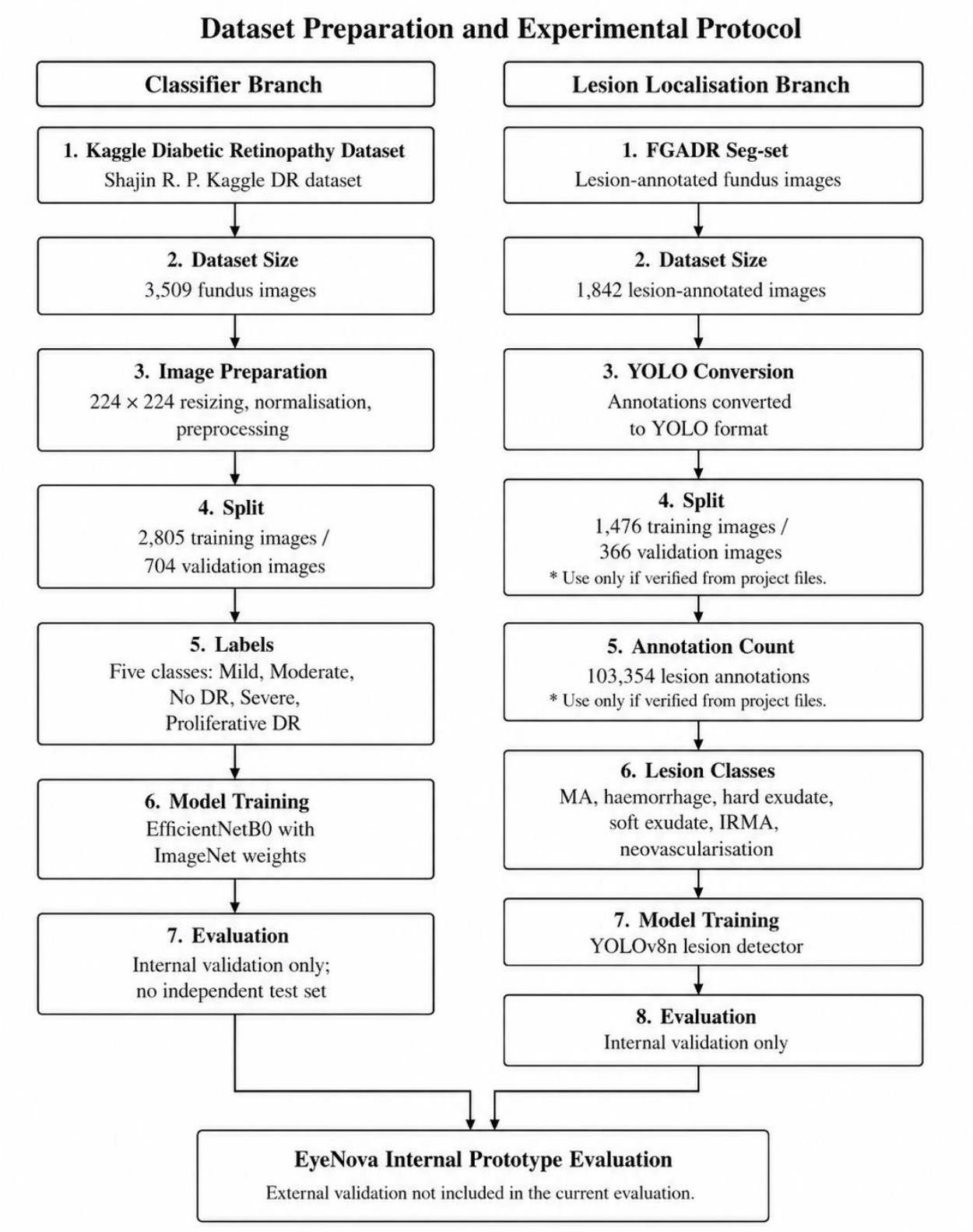


Figure 2. Dataset preparation and experimental protocol for the classifier and lesion-localisation branches.

5.3 EfficientNetB0 severity classifier

The grading model is an **EfficientNetB0** classifier initialised with ImageNet-pretrained weights. The classifier was trained with the following settings: input size **224 × 224**, **10 epochs**, **batch size 8**, **Adam** optimiser, **learning rate 0.001**, **categorical cross-entropy** loss, and augmentation consisting of **horizontal flips**, **random rotation 0.05**, and **random zoom 0.1**. EfficientNetB0 was selected because its compound

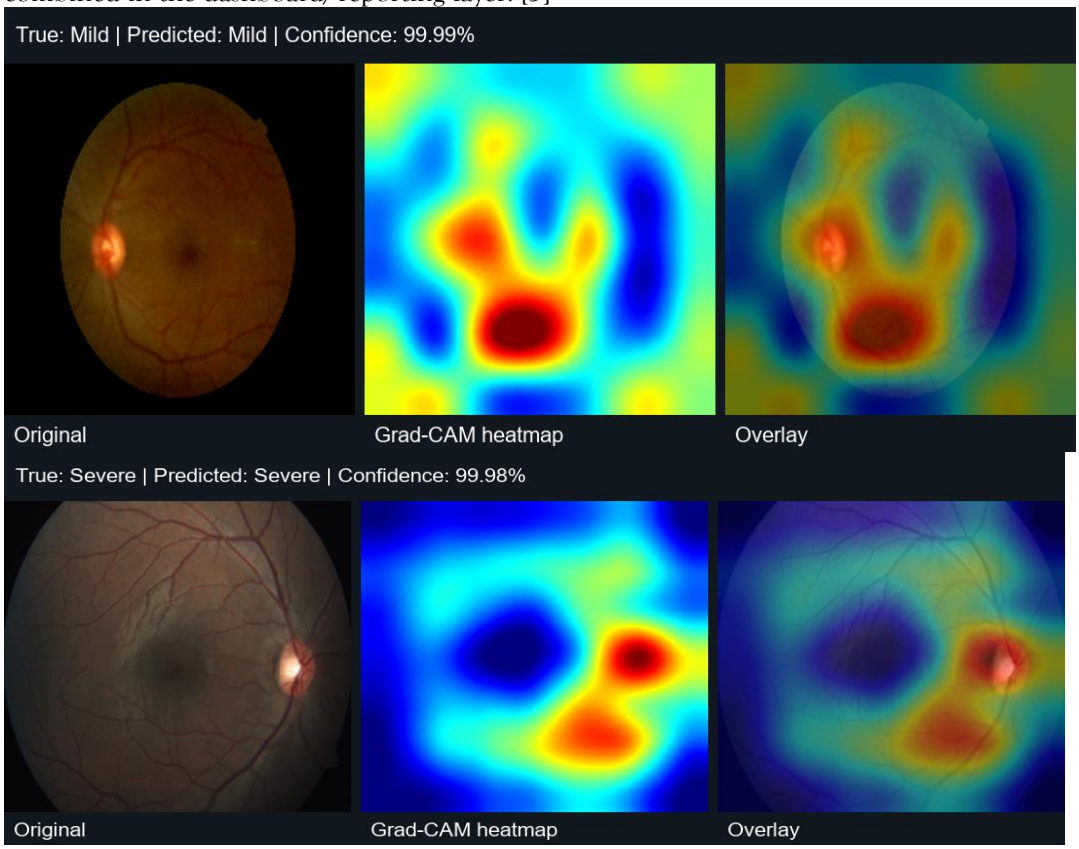
scaling strategy was explicitly designed to improve the accuracy efficiency trade-off across depth, width, and input resolution.

5.4 YOLOv8n lesion detector

The detector branch uses **YOLOv8n** trained on FGADR converted to YOLO format, with confirmed settings of **50 epochs**, **batch size 4**, and **image size 768 × 768**. Ultralytics documents YOLOv8 as an anchor-free family supporting detection, segmentation, classification, training, validation, inference, and export, and it explicitly recommends citing YOLOv8 as software/documentation because no formal YOLOv8 research paper was published. The Ultralytics software version and documentation access date should be cited.

5.5 Grad-CAM explainability and decision logic

The explainability module uses **Grad-CAM** to generate a heatmap from the classifier's final convolutional representation and overlays it on the original fundus image shown in figure 3. In the documented workflow, the system validates the uploaded image, preprocesses it, applies image-quality assessment, performs severity grading, runs lesion localisation, generates Grad-CAM, compiles the outputs into a clinician-facing dashboard, and optionally generates a structured report that stores the image, quality result, grade, lesion output, heatmap, and relevant metadata. EyeNova uses application-level late fusion: classifier predictions, YOLO lesion outputs, and Grad-CAM heatmaps are generated independently and combined in the dashboard/reporting layer. [5]



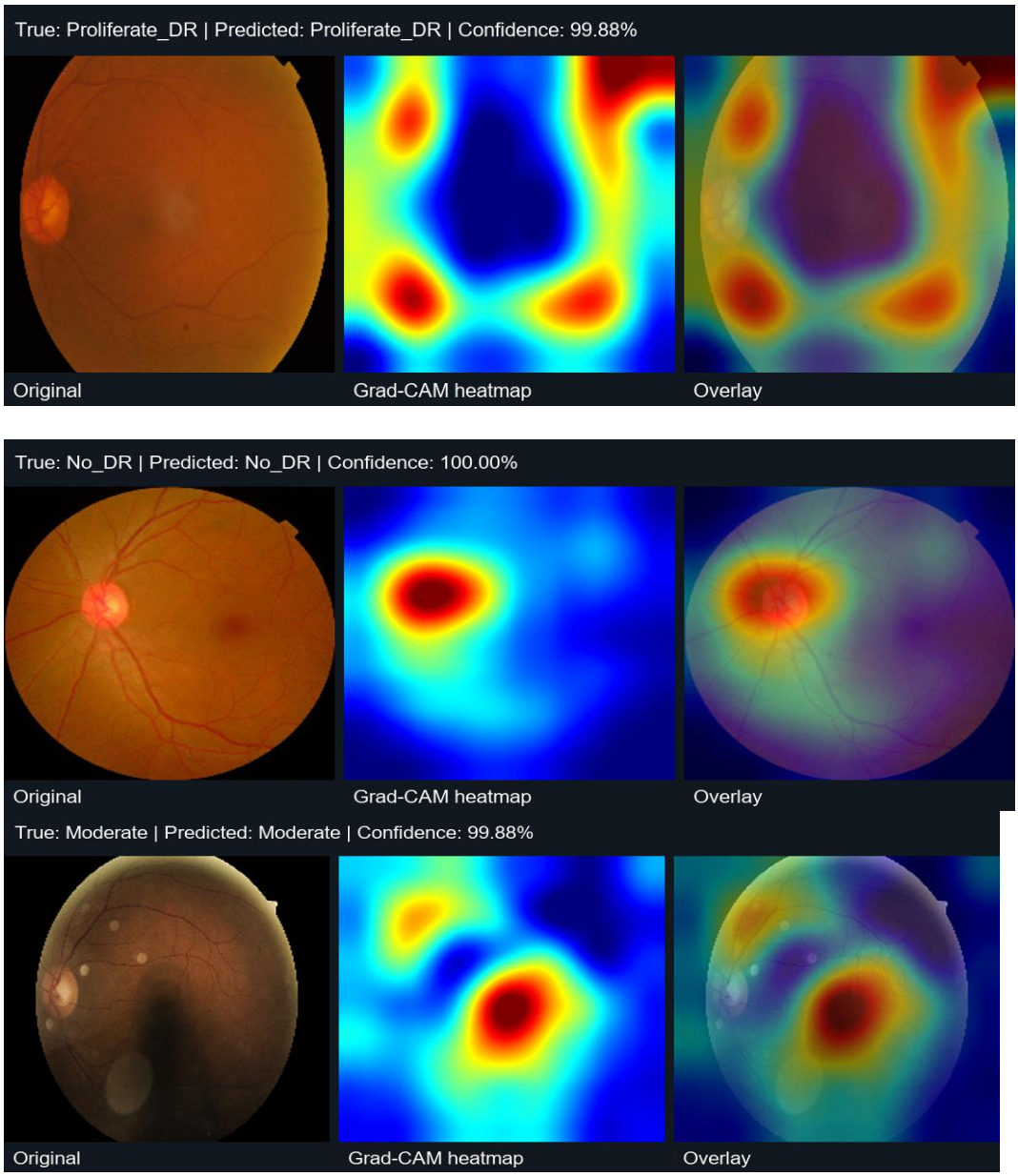


Figure 3. Explainable AI (XAI) workflow for retinal fundus image analysis using Grad-CAM heatmaps, YOLO lesion localization, and severity grading via application-level late fusion.

5.6 System architecture and implementation environment

EyeNova follows a layered architecture in figure 4. consisting of a presentation layer, application/service layer, AI processing layer, and data persistence layer. The presentation layer provides login, dashboard access, image upload, image-quality feedback, clinical decision-support outputs, explainable visualisation, report generation, and administrative functions. The application/service layer manages authentication, request validation, workflow coordination, result compilation, and report generation. The AI processing layer performs image preprocessing, image-quality assessment, five-class DR severity grading, YOLO-based lesion localisation, and Grad-CAM heatmap generation. The data persistence layer stores user records, uploaded image metadata, quality-assessment outputs, classification results, lesion-localisation outputs, heatmaps, generated reports, and audit logs. Security controls are implemented through authenticated access, role-based access control, controlled storage of uploaded data, and audit logging. EyeNova uses application-level late fusion, where classifier predictions, YOLO lesion-localisation outputs, and Grad-CAM heatmaps are generated independently and then combined in the dashboard and structured report to support clinician interpretation.

The local EyeNova implementation was evaluated on an HP ProBook 445 G8 Notebook PC running Windows 11 Pro Education, version 25H2, build 26200.8246, with an AMD Ryzen 5 5600U processor, integrated AMD Radeon Graphics, 8 GB RAM, and a 64-bit x64-based system. The software environment used Python 3.11.9, TensorFlow 2.16.1, PyTorch 2.11.0, Ultralytics 8.3.0, and OpenCV 4.11.0.86. The EfficientNetB0 classifier was trained with ImageNet initialisation, an input size of 224 × 224 pixels, batch size 8, Adam optimisation, an initial learning rate of 0.001, categorical cross-entropy loss, early stopping monitored on validation accuracy with patience 4, and ReduceLROnPlateau monitored on validation loss with factor 0.5 and patience 2. The classifier training time was not recorded in the training logs or repository files; available file metadata only indicates that the classifier model and train_summary.txt were last written around 09:56 local time on 16 May 2026, so the total classifier training duration cannot be reported. The YOLOv8n detector was trained for 50 epochs with batch size 4 and input size 768 × 768; checkpoint metadata indicated a Kaggle training path and a training time of 2758.84 seconds, approximately 45 minutes and 59 seconds. The FGADR YOLO documentation reports an image-level 80/20 stratified split with 1,476 training images, 366 validation images, 103,354 lesion annotations, and 0 duplicate basename or SHA-256 leakage across train and validation splits. No independent external test set was available for either the classifier or detector; therefore, the reported results are interpreted as internal prototype validation rather than external clinical validation.

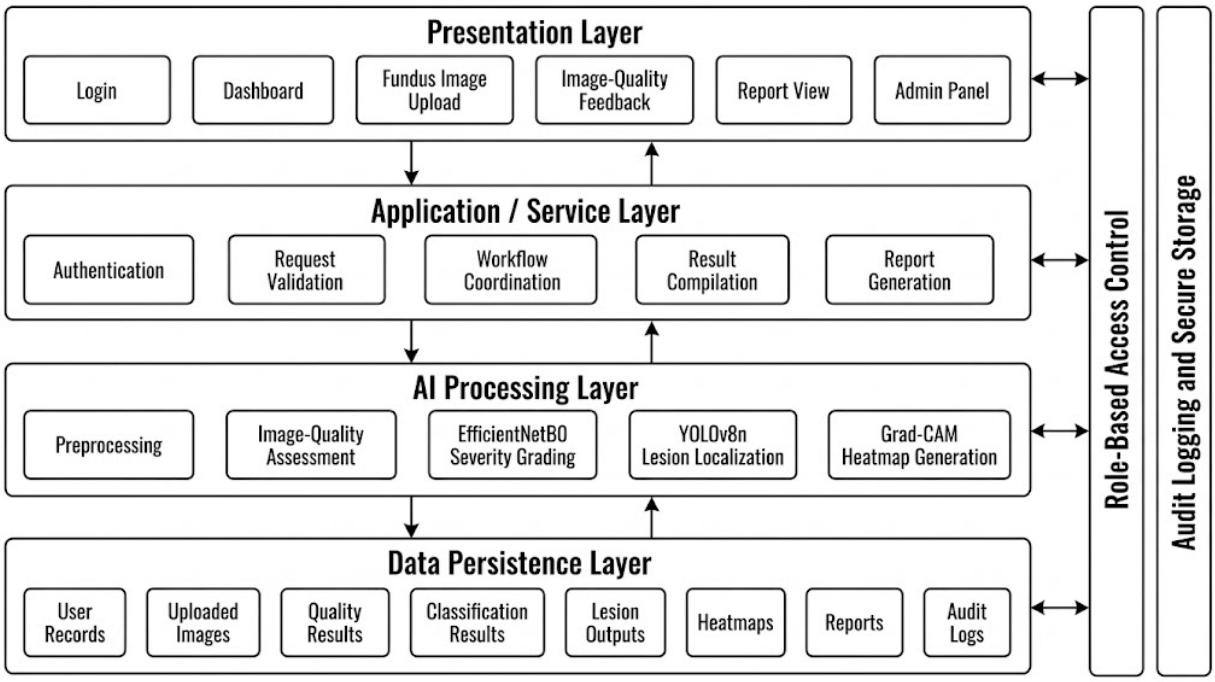


Figure 4. Layered architecture of EyeNova showing presentation, application/service, AI processing, and data-persistence components

The modeling phase consists of three tree-based models: XGBoost, In table 2, all implementation parameters, training settings, and deployment specifications for the EyeNova screening aid prototype are provided. Severity of the disease is assessed based on an EfficientNetB0 model, which is trained locally with early stopping and learning rate schedule for 10 epochs (224 × 224) in five DR severity classes. Multi-class lesion detection is done with YOLOv8n model, which is trained on Kaggle with 103,354 annotations in six lesions classes for 50 epochs (768 × 768) Local deployment is done with a Python-based deep learning framework with application-level late-fusion design, which includes classification, lesion detection, and Grad-CAM explanations.

Table 2: Confirmed model hyperparameters

Component	Setting	Confirmed value
EfficientNetB0 classifier	Initial weights	ImageNet
	Input size	224 × 224
	Epochs	10
	Batch size	8
	Optimiser	Adam
	Learning rate	0.001
	Loss	Categorical cross-entropy
	Classes	Mild, Moderate, No_DR, Proliferate_DR, Severe
	Augmentation	Horizontal flip, random rotation 0.05, random zoom 0.1
	Early stopping	Yes; monitor="val_accuracy", mode="max", patience=4, restore_best_weights=True
	Learning-rate schedule	ReduceLROnPlateau(monitor="val_loss", factor=0.5, patience=2)
	Dataset split	2,805 training images and 704 validation images
	Test set	No independent test set available
	Training platform	Local laptop
	Training time	Not recorded in training logs/files; available timestamps only show completion around 09:56 local time on 16 May 2026
	Completion evidence	Classifier model was last written at 2026-05-16 09:54:52; train_summary.txt was last written at 2026-05-16 09:56:18, indicating completion around 09:56 local time on 16 May 2026, but not total training duration
YOLOv8n detector	Model	YOLOv8n
	Epochs	50
	Batch size	4, based on recovered checkpoint metadata
	Input size	768 × 768
	Classes	Microaneurysm, haemorrhage, hard exudate, soft exudate, IRMA, neovascularisation
	Dataset split	1,476 training images and 366 validation images
	Annotation count	103,354 lesion annotations across six classes
	Split method	Image-level 80/20 stratified split by FGADR DR grade, seed 42
	Test set	No independent test set documented
	Training platform	Kaggle; checkpoint path points to /kaggle/working/...
	Training time	2758.84 s, approximately 45 min 59 s

Deployment description	Architecture	Locally deployable web-based screening-support prototype
	Fusion strategy	Application-level late fusion of classifier output, YOLO lesion output, and Grad-CAM heatmaps
Environment details	Hardware	HP ProBook 445 G8 Notebook PC
	Processor	AMD Ryzen 5 5600U with Radeon Graphics, 2.30 GHz
	RAM	8 GB
	GPU	Integrated AMD Radeon Graphics
	System type	64-bit operating system, x64-based processor
	Operating system	Windows 11 Pro Education
	OS version	25H2
	OS build	26200.8246
	Python version	3.11.9
	TensorFlow version	2.16.1
	Keras version	3.13.2 in backend/classification environment; 3.14.0 in root environment
	PyTorch version	2.11.0
	Ultralytics version	8.3.0
	OpenCV version	4.11.0.86

6. Results

6.1 Reported grading results

Classifier performance in table 3 is reported only on the supplied internal validation split and is not interpreted as test-set or clinical performance. On **704 validation images**, the confirmed classifier metrics are **98.44% accuracy**, **98.45% weighted precision**, **98.44% weighted recall/sensitivity**, **99.46% weighted specificity**, **98.43% weighted F1-score**, and **24.61 ms/image** average inference time. These results indicate high internal-validation performance on the available split; however, the absence of an independent test set limits generalisability.

Table 3: Overall classification results

Metric	**Value**
Validation accuracy	98.44%
Weighted precision	98.45%
Weighted recall / sensitivity	98.44%
Weighted specificity	99.46%
Weighted F1-score	98.43%
Average inference time	24.61 ms/image

The confusion matrix in figure 5 shows 693 correct predictions among 704 validation images. Misclassifications are limited, but they are not evenly distributed. Moderate DR is the weakest class on recall, and a small number of Mild and Severe cases are predicted as Proliferative DR. This error pattern indicates that aggregate accuracy should be interpreted alongside per-class recall and F1-score.

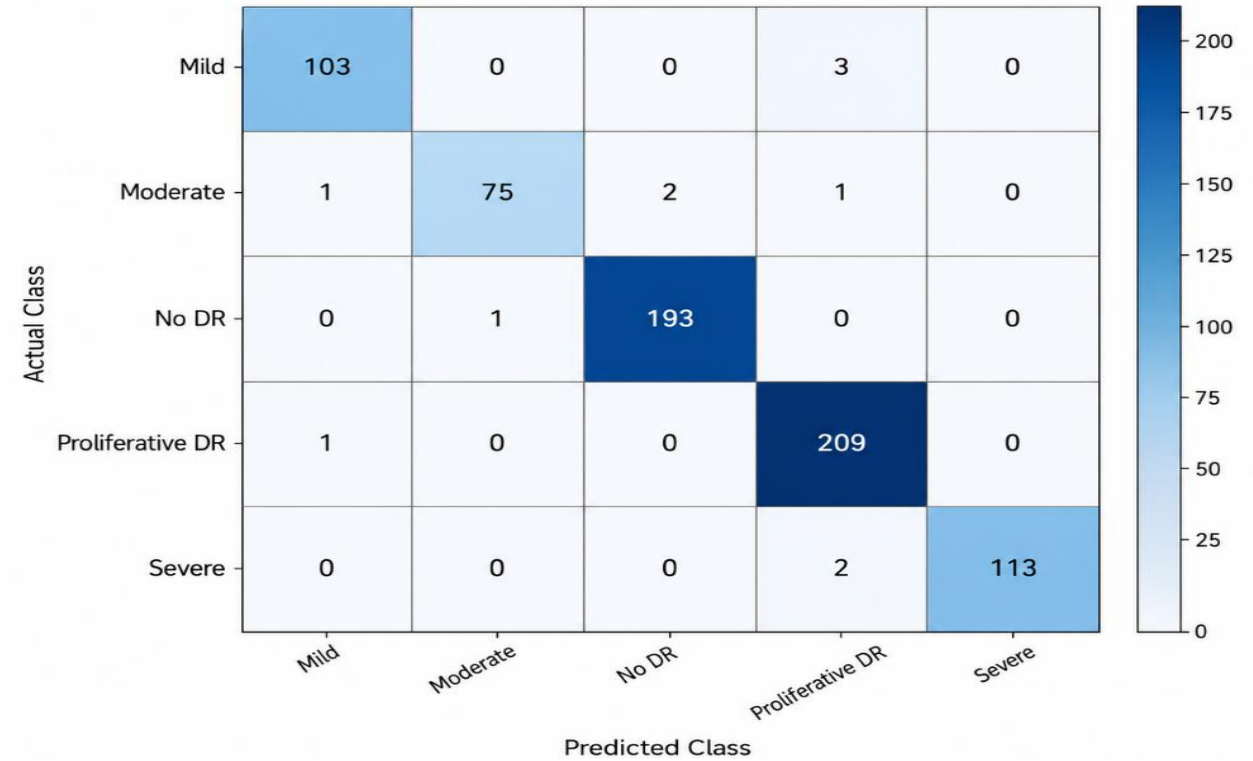


Figure 5. Confusion matrix of the EfficientNetB0 classifier on the 704-image internal validation split.

Table 4: Per-class classification metrics

6.2 Per-class grading results

Per-class results in table 4 identify class-specific weaknesses that are obscured by aggregate accuracy. Moderate DR has the lowest reported recall (**94.94%**) and F1-score (**96.77%**), while Severe achieves **100.00% precision** and **100.00% specificity** on the current validation split. No DR and Proliferative DR achieved high recall on the internal validation split. These findings should be interpreted as evidence of strong internal-validation performance, not as evidence that DR grading is clinically solved.

Class	Precision	Recall / Sensitivity	Specificity	F1-score	Support
Mild	98.10%	97.17%	99.67%	97.63%	106
Moderate	98.68%	94.94%	99.84%	96.77%	79
No DR	98.97%	99.48%	99.61%	99.23%	194
Proliferative DR	97.21%	99.52%	98.79%	98.35%	210
Severe	100.00%	98.26%	100.00%	99.12%	115

Table 5: YOLOv8n lesion-localisation results

6.3 Reported lesion-localisation results

Lesion localisation was the weakest subsystem. The confirmed YOLOv8n validation metrics in table 5 are **59.70% precision**, **21.27% recall**, **18.44% mAP@0.50**, and **7.30% mAP@0.50:0.95**. The stricter COCO-style summary metric is mAP@0.50:0.95 Because mAP@0.50:0.95 is the stricter summary metric, the detector's 7.30% mAP@0.50:0.95 is emphasised when interpreting localisation performance. Given its

low recall and mAP, the detector should be interpreted only as an exploratory visualisation component rather than a reliable lesion-localisation module.

Metric	Value
Precision	59.70%
Recall	21.27%
mAP@0.50	18.44%
mAP@0.50:0.95	7.30%

Lesion-localisation outputs are assistive visual cues and must not be interpreted as definitive evidence of lesion presence or absence. This limitation is consistent with prior lesion-detection studies reporting difficulty in detecting small retinal lesions, particularly microaneurysms. [10], [15],

7. Technical Analysis and Literature Comparison

EyeNova's main contribution is workflow integration rather than algorithmic novelty. Many published studies report either grading performance or lesion detection, but EyeNova combines upload validation, fundus preprocessing, image-quality control, five-class grading, lesion localisation, Grad-CAM, dashboarding, report generation, and secure record handling in one documented workflow. This workflow-oriented contribution addresses the gap between model accuracy and clinically understandable decision-support design by integrating diabetic retinopathy severity grading, lesion localisation, and explainability into a unified prototype framework. However, the system should be interpreted as an internally validated research prototype rather than a clinically deployable diagnostic tool. [20] [24]

Several methodological limitations reduce the strength of the evidence. First, the classification branch is evaluated only on an internal validation split. Second, the exact dataset provenance and split-generation logic for the classification dataset remain under documented. Third, the detector's 21.27% recall and 7.30% mAP@0.50:0.95 are insufficient to support strong lesion-level conclusions. Fourth, the study does not include ablation experiments, calibration analysis, confidence intervals, or clinician usability evaluation. Collectively, those issues reduce both reproducibility and external validity. The replication literature in DR deep learning has already shown how sensitive conclusions can be to incomplete reporting of data sources, grading protocols, and hyperparameters. [19] [21]

A further methodological point concerns **fusion**. Published multi-task or fusion systems often attempt tighter coupling between grading and lesion evidence. EyeNova uses application-level late fusion, in which classifier predictions, detector outputs, and Grad-CAM maps are combined after separate inference. That is acceptable for a prototype study and This design is appropriate for a prototype study but should not be described as a jointly optimised fusion architecture. [7]

Table 6: Comparison with representative literature

Study	Main task	Dataset(s)	Core approach	Reported performance in original study	Interpretation relative to EyeNova
Alyoubi et al., 2021 [10]	Five-class grading + lesion localisation	DDR, APTOS	CNN512 + YOLOv3 + fusion	88.6% accuracy on DDR; 84.1% on APTOS; lesion mAP 0.216; fused accuracy 89%	Relevant precedent for combining DR grading with lesion evidence.

Santos et al., 2022 [11]	Lesion detection	DDR, IDRiD	YOLOv5 + preprocessing, tiling, augmentation	DDR validation mAP@0.5 = 0.2630; test mAP@0.5 = 0.1540	Illustrates the difficulty of lesion detection even in dedicated pipelines.
Wahab Sait, 2023 [12]	Lightweight DR grading	APTOS, EyePACS	YOLOv7 feature extraction + QMPA + MobileNetV3	98.0% and 98.4% accuracy; F1-scores of 93.7 and 93.1	Relevant lightweight grading reference with deployment emphasis.
Lv et al., 2026 [16]	Small-lesion detection	Fundus lesion dataset	Improved YOLOv8 with ECA, RepNCSPELAN, and GIoU	+12.3% mAP over baseline YOLOv8; 16% fewer parameters	Suggests possible detector-improvement directions for EyeNova.
Rahim et al., 2026 [17]	Microaneurysm detection	Colour fundus image set	YOLOv9 + image-processing comparison	91% accuracy for best YOLOv9 model	Supports future exploration of a microaneurysm-focused detector branch, but accuracy is not directly comparable with mAP.
EyeNova	DR grading + exploratory lesion localisation + explanation	Shajin R. P. Kaggle dataset; FGADR YOLO	EfficientNetB0 + YOLOv8n + Grad-CAM + dashboard/reporting	98.44% internal-validation grading accuracy; YOLOv8n mAP@0.50:0.95 = 7.30%	High internal grading performance, weak lesion localisation, and strongest value as an integrated explainable prototype.

The summarized studies are not directly comparable because they differ in dataset composition, label schemas, image quality, split protocols, and evaluation designs; therefore, the comparison in table 6 is presented as a qualitative overview rather than a direct performance leaderboard. [16],

8. Discussion, Limitations, and Future Work

8.1 Clinical interpretation and deployment language

EyeNova is presented as a **screening-support tool** that organises AI outputs for review rather than replacing ophthalmologists. That wording is already compatible with the internal system design, which describes a clinician-facing dashboard, structured reports, role-based access, audit logging, and secure storage. For deployment language, The deployment description is limited to a locally deployable web-based prototype unless cloud deployment is empirically demonstrated.

8.2 Explainability and privacy considerations

Grad-CAM was included to provide coarse visualisation of image regions contributing to the classifier prediction here because it can make CNN predictions visually inspectable, but the paper should avoid overstating what a heatmap means. Grad-CAM provides a coarse localisation of discriminative regions; it

does **not** establish causal explanation, clinical sufficiency, or lesion-level correctness by itself Future work may compare Grad-CAM with Grad-CAM++ to evaluate whether alternative explanation methods provide more localised visual evidence. [22]
The prototype includes authenticated access, role-based permissions, secure storage, audit logging, and HTTPS where applicable. However, explicit statements on patient-image anonymisation, retention policy, encryption at rest, consent, and local data-governance procedures are currently unspecifiedPatient-image anonymisation, consent procedures, retention policy, encryption at rest, and local data-governance procedures remain future implementation requirements.

8.3 Limitations
This study has several limitations that restrict the strength and generalisability of the findings. The most important limitations are: no separate held-out test set for the classifier; no external validation; incomplete provenance details for the grading dataset; unspecified split-generation method and therefore possible leakage risk; no detailed duplicate-control statement; weak detector recall and mAP; unspecified YOLO train/validation split and per-lesion AP values; absent training curves in the current confirmed file set; absent calibration analysis and confidence intervals; and no clinician usability evaluation. These limitations materially restrict the reproducibility, external validity, and clinical interpretability of the current findings.

8.4 Future Work
A future version should retain EfficientNetB0 as a reproducible baseline and improve the detection branch in a targeted way. The literature suggests several defensible extensions: high-resolution tiling and background handling, as in recent YOLO-based fundus lesion work; attention-enhanced or improved-loss YOLOv8 variants, as reported by Lv et al.; and microaneurysm-focused configurations, as explored by Rahim et al. A more ambitious next step would be a **lesion-guided grading model** in which detector-derived lesion maps or counts are fused with the classifier's global fundus representation. That direction is consistent with both FGADR's benchmark motivation and prior multi-task DR designs. 17]
Future evaluation should include a documented split protocol; an independent held-out test set; external grading validation on at least one of APTOS, EyePACS, or Messidor; external lesion benchmarking on DDR or IDRiD if label compatibility permits; ablation comparisons for augmentation, quality gating, and detector design; per-lesion AP reporting; calibration metrics; and clinician review of dashboard/report usability. Because Moderate DR is the weakest EyeNova grading class, targeted analysis of borderline or ambiguous cases should also be added. Recommended evaluation matrix for future research is shown in table 7. [23][25]

Table 7: Recommended evaluation metrics for the revised paper

Evaluation layer	Metrics to report	Why they matter
Grading	Accuracy, weighted precision, weighted recall, weighted specificity, weighted F1, per-class metrics, confusion matrix, AUC where available, quadratic weighted kappa	Standard DR grading evaluation and ordinal-awareness
Localisation	Precision, recall, mAP@0.50, mAP@0.50:0.95, per-class AP, PR curves, false positives per image	Necessary to understand lesion sensitivity, not just one summary number
Efficiency	Inference time per image, memory footprint, model size	Important for local screening workflows

Reliability	Confidence calibration, confidence intervals, repeatability checks	Needed before any stronger deployment claim
Workflow	Rejection rate from IQA, time-to-report, clinician acceptance/usability results	Connects model output to actual screening utility

9. Conclusion

This study presented EyeNova, a locally deployable web-based prototype for explainable diabetic retinopathy screening support. The EfficientNetB0 classifier achieved high internal-validation performance on the available validation split, while YOLOv8n lesion localisation remained exploratory because of low recall and mAP. The results support EyeNova as an internally validated workflow-integration prototype, not as a clinically validated diagnostic system. Future work should prioritise a leakage-free and preferably patient-level split, independent test-set evaluation, external validation, detector improvement, per-lesion AP reporting, calibration analysis, confidence intervals, and clinician usability assessment.

REFERENCES

[1] B. Goutam, M. F. Hashmi, Z. W. Geem, and N. D. Bokde, "A Comprehensive Review of Deep Learning Strategies in Retinal Disease Diagnosis Using Fundus Images," IEEE Access, vol. 10, 2022, doi: 10.1109/ACCESS.2022.3178372.

[2] S. Zhu, C. Xiong, Q. Zhong, and Y. Yao, "Diabetic Retinopathy Classification With Deep Learning via Fundus Images: A Short Survey," IEEE Access, vol. 12, 2024, doi: 10.1109/ACCESS.2024.3361944.

[3] W. L. Alyoubi, W. M. Shalash, and M. F. Abulkhair, "Diabetic retinopathy detection through deep learning techniques: A review," Informatics in Medicine Unlocked, vol. 20, Art. no. 100377, 2020, doi: 10.1016/j.imu.2020.100377.

[4] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proc. 36th Int. Conf. Machine Learning, vol. 97, Proc. Machine Learning Research, pp. 6105–6114, 2019.

[5] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization," in Proc. IEEE Int. Conf. Computer Vision, pp. 618–626, 2017.

[6] G. Jocher, A. Chaurasia, and J. Qiu, Ultralytics YOLOv8, ver. 8.0.0, software/documentation, 2023, accessed Jun. 3, 2026.

[7] Y. Zhou, B. Wang, L. Huang, S. Cui, and L. Shao, "A Benchmark for Studying Diabetic Retinopathy: Segmentation, Grading, and Transferability," IEEE Trans. Medical Imaging, vol. 40, no. 3, pp. 818–828, 2021, doi: 10.1109/TMI.2020.3037771.

[8] FGADR Dataset Official Website, "FGADR: Fine-Grained Annotated Diabetic Retinopathy Dataset," official dataset site, accessed Jun. 3, 2026.

[9] E. Decencière et al., "Feedback on a publicly distributed database: the Messidor database," Image Analysis & Stereology, vol. 33, no. 3, pp. 231–234, 2014, doi: 10.5566/ias.1155.

[10] W. L. Alyoubi, M. F. Abulkhair, and W. M. Shalash, "Diabetic Retinopathy Fundus Image Classification and Lesions Localization System Using Deep Learning," Sensors, vol. 21, Art. no. 3704, 2021, doi: 10.3390/s21113704.

[11] C. Santos, M. Aguiar, D. Welfer, and B. Belloni, "A New Approach for Detecting Fundus Lesions Using Image Processing and Deep Neural Network Architecture Based on YOLO Model," Sensors, vol. 22, Art. no. 6441, 2022, doi: 10.3390/s22176441.

[12] A. R. Wahab Sait, "A Lightweight Diabetic Retinopathy Detection Model Using a Deep-Learning Technique," Diagnostics, vol. 13, Art. no. 3120, 2023, doi: 10.3390/diagnostics13193120.

[13] Q. H. Nguyen, R. Muthuraman, L. Singh, G. Sen, A. C. Tran, B. P. Nguyen, and M. Chua, "Diabetic Retinopathy Detection using Deep Learning," in Proc. 4th Int. Conf. Machine Learning and Soft Computing, 2020, doi: 10.1145/3380688.3380709.
[14] S. Qummar et al., "A Deep Learning Ensemble Approach for Diabetic Retinopathy Detection," IEEE Access, vol. 7, 2019, doi: 10.1109/ACCESS.2019.2947484.
[15] S. Akhtar et al., "A deep learning based model for diabetic retinopathy grading," Scientific Reports, vol. 15, Art. no. 3763, 2025, doi: 10.1038/s41598-025-87171-9.
[16] H. Lv, X. Zhang, F. Zhao, and P. Li, "Diabetic Retinopathy Detection Algorithm Based on Improved YOLOv8," Cognitive Computation, vol. 18, Art. no. 50, 2026, doi: 10.1007/s12559-026-10589-4.
[17] S. S. Rahim, A. Deo, R. Mumtaz, and V. Palade, "An Improved Microaneurysms Detection for Diabetic Retinopathy Screening Using YOLO," Biomedicines, vol. 14, Art. no. 359, 2026, doi: 10.3390/biomedicines14020359.
[18] S. K. Saha, B. Fernando, J. Cuadros, D. Xiao, and Y. Kanagasingam, "Deep Learning for Automated Quality Assessment of Color Fundus Images in Diabetic Retinopathy Screening," arXiv:1703.02511, 2017.
[19] M. Voets, K. Møllersen, and L. A. Bongo, "Replication study: Development and validation of deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," arXiv:1803.04337, 2018.
[20] S. M. S. Islam, M. M. Hasan, and S. Abdullah, "Deep Learning based Early Detection and Grading of Diabetic Retinopathy Using Retinal Fundus Images," arXiv:1812.10595, 2018.
[21] B. Tymchenko, P. Marchenko, and D. Spodarets, "Deep Learning Approach to Diabetic Retinopathy Detection," arXiv:2003.02261, 2020.
[22] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks," arXiv:1710.11063, 2017.
[23] G. Lepetit-Aimon, C. Playout, M. C. Boucher, R. Duval, M. H. Brent, and F. Cheriet, "MAPLES-DR: MESSIDOR Anatomical and Pathological Labels for Explainable Screening of Diabetic Retinopathy," arXiv:2402.04258, 2024.
[24] J. Hou et al., "Cross-Field Transformer for Diabetic Retinopathy Grading on Two-field Fundus Images," arXiv:2211.14552, 2022.
[25] O. Abramovich et al., "FundusQ-Net: a Regression Quality Assessment Deep Learning Algorithm for Fundus Images Quality Grading," arXiv:2205.01676, 2022.
[26] Khalil, A., Hussain, M., Majeed, M. K., Hamza, A., Ali, A., Ajaz, K., et al. (2025). Artificial Intelligence in Neuro-Oncology: Integrating Advanced Machine Learning Techniques for Accurate and Early Detection of Brain Tumors Through MRI Imaging. Spectrum of Engineering Sciences, 413–435.
[27] Khalil, A., Hussain, A. W., Khan, A. H., Majeed, M. K., Mir, S. Z., Abbasi, M. D., et al. (2025). Transforming Heart Transplantation with AI: Deep Neural Networks for Predictive Analytics and Real-Time Monitoring in Clinical Decision Support Systems. Pakistan Journal of Medical & Cardiological Review, 4(3), 390–411.
[28] Majeed, M. K., Dilshad, W., Essa, R., Ali, I., Majid, M., & Zia, A. (2026). Advanced Linear Algebra and Mathematical Optimization Techniques for High-Dimensional Data Analysis and Machine Learning Applications. Spectrum of Engineering Sciences, 4(5), 2225–2249.
[29] Mir, S. Z. M., Majeed, M. K., Khawer, S. K., Younus, I., & Shakil, M. H. (2025). IoT-Driven Smart Net Energy Meter for Optimized Energy Management in Grid-Connected Solar Systems. Sukkur IBA Journal of Emerging Technologies, 8(2), 23–36.
[30] Mir, S. Z., Majeed, M. K., Siddiqui, M. H. S., Khawer, S. K., Younus, I., Shah, S. R., et al. (2025). Integrating AI and IoT in Solar Power Systems for Improved Monitoring and Predictive Maintenance. Spectrum of Engineering Sciences, 971–982.
[31] Siddiqui, M. H. S., Abbasi, M. D., Mir, S. Z., Nadeem, G., Majeed, M. K., Khawer, S. K., et al. (2025). Prediction of Maximum Air Temperature for Defining Heat Wave in Various States of Pakistan Using Machine Learning Algorithm. Spectrum of Engineering Sciences, 942–960.