

Diabetic Retinopathy Lesion Classification using Machine Learning

By

Sibin Shibu

Submitted to

The University of Roehampton

In partial fulfilment of the requirements

for the degree of

Master of Science

in

Data Science

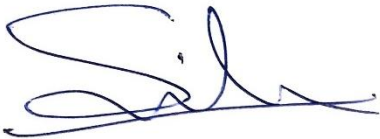
Declaration

I hereby certify that this report constitutes my own work, that where the language of others is used, quotation marks so indicate, and that appropriate credit is given where I have used the language, ideas, expressions, or writings of others.

I declare that this report describes the original work that has not been previously presented for the award of any other degree of any other institution.

Sibin Shibu

23/08/2025

A handwritten signature in blue ink, appearing to be 'Sibin Shibu', with a stylized, cursive script.

Acknowledgements

I would like to express my sincere appreciation to my supervisor, **Dr. Mohammed Farhan Khan**, for his invaluable guidance, constructive feedback, and continuous support throughout the development of my dissertation, *ADRGs: An Automatic Diabetic Retinopathy Grading System Through Machine Learning*. His expertise and encouragement have been instrumental in the successful completion of this research.

I am also grateful to the **University of Roehampton, London**, for providing the resources, facilities, and academic environment necessary to carry out this study.

Finally, I wish to extend my thanks to my family and friends for their patience, understanding, and unwavering support during my MSc journey.

Abstract

Early detection of diabetic retinopathy (DR) is vital but manual grading of fundus images is time-consuming and difficult to scale. This project develops and evaluates a lightweight, two-stage, feature-based pipeline for DR lesion classification on the DIARETDB1 dataset. After standard pre-processing and background suppression (vessels and optic disc), candidate regions are generated and described by a 30-dimensional hybrid feature vector capturing appearance (intensity statistics across colour/derived planes), morphology (e.g., circularity, solidity, extent, equivalent diameter, perimeter–area ratio), and simple anatomical context (distances to vessels/optic disc). Classification proceeds hierarchically: Stage-1 separates lesions from background within two families—bright (EX/CWS) and red (MA/HE). Stage-2 sub-types the positives (EX vs CWS and MA vs HM (HE mapped to HM)). To avoid leakage, an 80/20 split grouped by image is used so patches from the same eye never appear in both train and test. Class imbalance is addressed by under-sampling background in Stage-1 and train-only rebalancing in Stage-2; test sets retain natural class ratios. Three tabular baselines are compared: Logistic Regression (LR), Random Forest (RF), XGBoost (XGB) with metrics aligned to the source paper (Sensitivity, Specificity, Accuracy, Balanced Accuracy = (Sensitivity+Specificity)/2), plus ROC-AUC and PR-AUC.

Results show bright vs background is easier than red vs background. For bright, LR achieved high lesion sensitivity (0.854) with overall Accuracy 0.779 and ROC-AUC \approx 0.88; RF offered the highest specificity (fewest false positives), while XGB provided a balanced trade-off. For red, XGB delivered the best overall balance (Accuracy \approx 0.711, ROC-AUC \approx 0.81), while LR preserved the highest red-lesion recall. In Stage-2, EX vs CWS was separated best by LR (Balanced Accuracy \approx 0.83, ROC-AUC \approx 0.92), whereas MA vs HM was the hardest task; LR gave the best-balanced performance (Balanced Accuracy \approx 0.62, ROC-AUC \approx 0.68), RF the highest accuracy via conservatism on HM, and XGB lay between.

Table of Contents

Declaration	ii
Acknowledgements.....	iii
Abstract	iv
Table of Contents	v
List of Figures	vii
List of Tables.....	viii
List of Abbreviations	ix
Chapter 1 Introduction.....	1
1.1 Problem Description, Context and Motivation	1
1.2 Objectives.....	2
1.3 Methodology.....	3
1.4 Legal, Social, Ethical and Professional Considerations	4
1.5 Background	4
1.6 Structure of Report	5
Chapter 2 Literature – Technology Review.....	6
2.1 Literature Review	6
2.2 Technology Review	8
2.3 Summary	10
Chapter 3 Implementation.....	11
3.1 System Design and Architecture	11
3.2 Dataset and Data Preparation	12
3.3 Image Pre-Processing.....	17
3.4 Segmentation	19
3.5 Segmentation Stages	20
3.6 Major Vessel Extraction	21
3.7 Minor Vessel Extraction	22

3.8	Optic Disc Detection.....	23
3.9	Background Removal and Clean Segmented Image	24
3.10	Lesion Candidate Generation.....	25
3.11	Feature Extraction.....	27
3.12	Lesion Classification	29
3.13	Severity Grading.....	29
3.14	Technologies Used	30
3.15	Summary	31
Chapter 4 Evaluation and Results		32
4.1	Experimental Setup.....	32
4.2	Evaluation Metrics	33
4.3	Stage 1: Bright vs Background	34
4.4	Stage 1: Red vs Background	38
4.5	Stage 2: EX vs CWS	42
4.6	Stage 2: MA vs HM	46
4.7	Strengths, Limitations, Threats to Validity.....	50
4.8	Summary	51
Chapter 5 Conclusion		52
5.1	Future Work	54
5.2	Reflection	55
References.....		56
Appendices.....		I
Appendix A: Project Proposal		II
Appendix B: Project Management.....		III
Appendix C: Artefact/Dataset		IV
Appendix D: Screenshot		V

List of Figures

Figure 1: Grantt Chart	3
Figure 2: System Architecture.....	11
Figure 3: Annotation Data Frame Sample.....	12
Figure 4: Annotations Overlay on Raw Fundus Images	13
Figure 5: Lesion Type Distribution	14
Figure 6: Lesion Type vs Confidence	15
Figure 7: Annotations: Polygon vs Centroid	16
Figure 9: Image Pre-Processing Pipeline.....	17
Figure 10: Image Pre-Processing Stages	18
Figure 11: Segmentation Pipeline	19
Figure 12: Segmentation Stages	20
Figure 13: Major Vessels Sample	21
Figure 14: Minor Vessels Sample	22
Figure 15: Optic disc Sample.....	23
Figure 16: Background Removed Clean Segmented Sample.....	24
Figure 17: Lesion Candidate Generation Pipeline.....	25
Figure 18: Feature Extraction Pipeline Features.....	27
Figure 19: Lesion Classification Pipeline	29

List of Tables

Table 1: Intensity Features.....	27
Table 2: Structural Features.....	28
Table 3: Anatomical Context Features.....	28
Table 4: Bright vs Background Classification Report	34
Table 5: Bright vs Background Confusion Matrix.....	35
Table 6: Bright vs Background AUC-ROC Curve	36
Table 7: Bright vs Background Model Comparison.....	37
Table 8: Red vs Background Classification Report	38
Table 9: Red vs Background Confusion Matrix	39
Table 10: Red vs Background AUC-ROC Curve.....	40
Table 11: Red vs Background Model Comparison	41
Table 12: EX vs CWS Classification Report	42
Table 13: EX vs CWS Confusion Matrix	43
Table 14: EX vs CWS AUC-ROC Curve.....	44
Table 15: EX vs CWS Model Comparison	45
Table 16: MA vs HM Classification Report.....	46
Table 17: MA vs HM Confusion Matrix	47
Table 18: MA vs HM AUC-ROC Curve.....	48
Table 19: MA vs HM Model Comparison	49

List of Abbreviations

No.	Abbreviation	Meaning
1	MA	Microaneurysm
2	HE	Haemorrhages
3	EX	Hard Exudates
4	CWS	Cotton Wool Spots

Chapter 1 Introduction

1.1 Problem Description, Context and Motivation

Diabetic Retinopathy is a progressive eye disease caused by complications of diabetes and has become one of the leading causes of vision impairment worldwide. It is estimated that nearly one-third of individuals with diabetes will develop some degree of DR during their lifetime. If left untreated, the disease advances gradually and can result in irreversible blindness. [1][2][3]

One of the major challenges in managing DR is that it often progresses without noticeable symptoms during its early stages. By the time patients present with blurred vision, distortions, or severe sight loss, the disease has often reached an advanced stage. Traditional diagnostic approaches rely on ophthalmologists manually interpreting retinal fundus images to identify lesions such as MA, Haemorrhages, Hard Exudates, and Cotton Wool Spots. This process, while effective, is highly time-consuming, depends on the expertise of specialists, and is not feasible at scale.[4]

The global increase in diabetes prevalence intensifies the problem. Projections show that the number of patients affected by DR will continue to rise significantly in the coming decades, putting additional strain on healthcare systems. Early detection and treatment can reduce the risk of vision loss by as much as ninety percent, yet screening programs remain underutilised due to the lack of specialists and the complexity of manual diagnosis.[5]

In this context, the development of automated grading systems for DR has become a critical area of research. By combining fundus imaging with computer vision and machine learning techniques, automated systems have the potential to provide cost-effective, accurate, and scalable screening solutions. Such systems can reduce workloads for specialists, improve early diagnosis rates, and ultimately prevent blindness in millions of individuals. [1][2]

1.2 Objectives

The aim of this project is to design and evaluate a “**Diabetic Retinopathy Lesion Classification using Machine Learning**” that operates on retinal fundus images using machine learning techniques. The objectives are [1]:

1. To enhance the quality of low-contrast retinal images through advanced pre-processing techniques.
2. To segment background regions such as the optic disc and retinal vasculature in order to reduce misclassification errors.
3. To detect and classify retinal lesions through a hybrid feature extraction process.
4. To implement a hierarchical classification mechanism for grading DR severity.
5. To validate the performance of the system on benchmark dataset “DIARETDB1” using widely accepted metrics such as accuracy, sensitivity, specificity, and area under the ROC curve.

1.3 Methodology

Design

The project is designed in four distinct phases: pre-processing, segmentation, lesion classification, and severity grading. During pre-processing, a contrast enhancement method is applied to improve the clarity of retinal attributes. Segmentation of vessels and optic disc is achieved through a combination of mathematical morphology and composite filtering techniques. Lesion detection is then carried out using a set of intensity-based and structural features, and classification is performed in a hierarchical manner to improve accuracy.

Testing and Evaluation

The system is tested using benchmark datasets of retinal fundus images. Performance evaluation is conducted using accuracy, sensitivity, specificity, and area under the curve as primary metrics. Comparative analysis with existing methods is used to demonstrate the advantages of the proposed approach.

Project Management

The development follows an Agile framework, with milestones defined in a Gantt chart. Each sprint focuses on specific tasks, such as data preparation, algorithm design, segmentation validation, and system evaluation. This structure allows continuous refinement and responsiveness to challenges encountered during implementation. Fig 1. shows the Grant Chart sample.

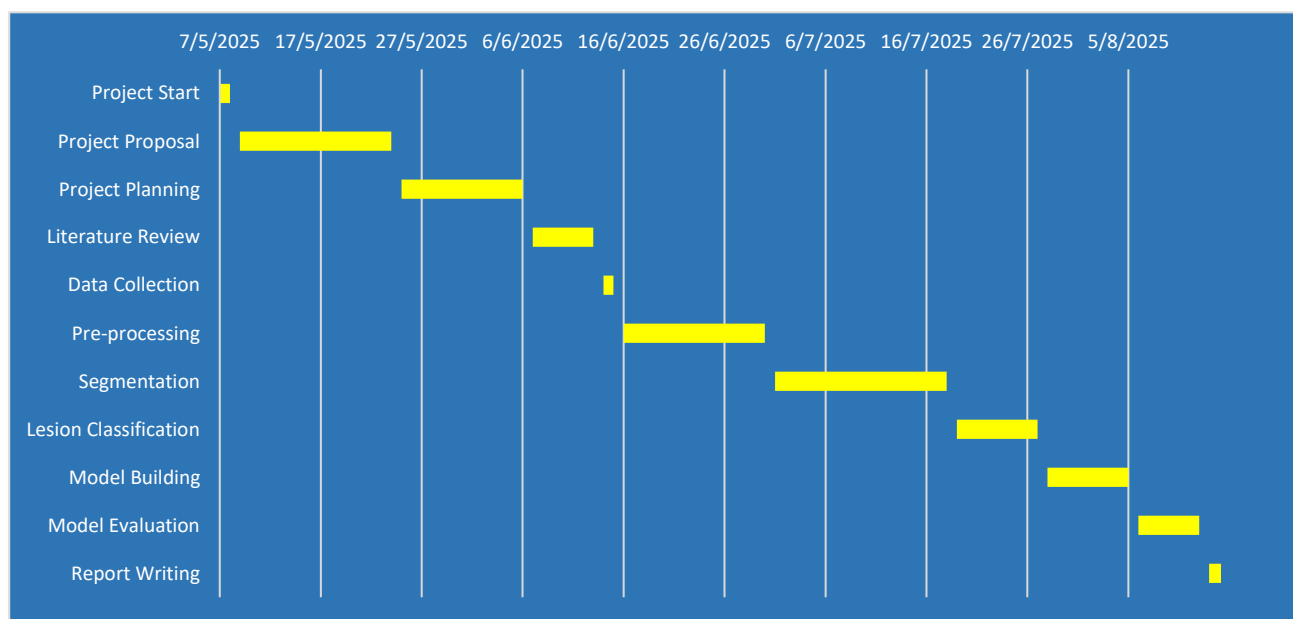


Figure 1: Grantt Chart

1.4 Legal, Social, Ethical and Professional Considerations

The project makes use of publicly available, annotated retinal image dataset. This ensures compliance with data protection standards and eliminates risks associated with patient confidentiality. From a social perspective, the system is designed to act as a diagnostic aid for clinicians rather than a replacement for professional medical judgment.

Ethical considerations include minimising potential bias in lesion detection and ensuring that results do not mislead clinical decision-making. Professional responsibility requires that the system be transparent, reliable, and accountable, with clear communication about its role as a decision-support tool. Legally, the use of datasets must respect licensing agreements, and any intellectual property generated from the project must be acknowledged appropriately.

1.5 Background

The rapid growth in global diabetes prevalence has resulted in DR becoming a major public health concern. Conventional diagnostic tools such as fluorescein angiography and optical coherence tomography are effective but either invasive, expensive, or limited in scope. Fundus photography, by contrast, offers a non-invasive imaging approach that can capture the retinal abnormalities associated with DR in an efficient manner. [14][11]

Research into automated DR detection has explored a variety of methods. Earlier systems relied on individual lesion segmentation using techniques such as mathematical morphology and texture analysis. While these methods achieved partial success, they lacked complete grading frameworks and were often limited by image quality issues. More recent advances have introduced machine learning and deep learning techniques, which offer strong performance but often require large annotated datasets and significant computational resources. [12], [15]

The present project builds upon these developments by creating an integrated system that combines pre-processing, segmentation, lesion classification, and severity grading into one coherent pipeline. By using both pixel intensity and structural features, the system aims to achieve high robustness in detecting and classifying lesions while maintaining computational efficiency. This balance makes the approach suitable for use in real-world clinical and community screening settings, particularly where resources are limited. [13], [14]

1.6 Structure of Report

This dissertation is structured as follows:

- **Chapter 1 – Introduction:** Provides the context, problem definition, objectives, and methodology.
- **Chapter 2 – Literature Review:** Reviews existing research on DR detection and grading, with emphasis on machine learning approaches.
- **Chapter 3 – Implementation:** Details the architecture, data pre-processing, and model training process.
- **Chapter 4 – Evaluation and Results:** Presents the evaluation metrics, performance analysis, and comparison with existing systems.
- **Chapter 5 – Discussion:** Discusses the findings, limitations, and implications for future research.
- **Chapter 6 – Conclusion:** Summarizes the work and suggests potential future directions.

Chapter 2 Literature – Technology Review

2.1 Literature Review

Diabetic Retinopathy (DR) has been widely recognised as a leading cause of preventable blindness, and its early detection has become a key area of medical research. Automated grading systems for DR have evolved considerably over the past two decades, moving from basic image enhancement techniques to sophisticated deep learning models. This section critically examines the major developments in the literature, highlighting both their strengths and limitations, while situating them within the context of the current project.

Early Approaches to DR Detection

The earliest automated approaches to DR diagnosis relied heavily on image pre-processing and segmentation of specific retinal abnormalities. Initial studies emphasised the importance of contrast enhancement and colour normalisation to improve image clarity. For example, contrast-limited adaptive histogram equalisation (CLAHE) was often employed to improve visibility of lesions in fundus images. While such methods improved detection of certain features, they were sensitive to noise and failed when images suffered from severe distortions. [8]

Lesion-based systems initially focused on identifying individual abnormalities. MA were frequently considered the earliest indicators of DR, and various clustering and thresholding methods were employed to detect them. Similarly, HE and EX were targeted using filters such as Gabor filters, Hough transforms, and texture-based descriptors. However, these approaches often failed to provide a complete grading system, since they detected lesions but did not integrate them into a severity classification. [5][10]

Emergence of Machine Learning Methods

Machine learning introduced more flexibility into DR detection by allowing algorithms to classify image pixels or regions based on learned features. Approaches such as Support Vector Machines (SVM), K-Nearest Neighbors (K-NN), and Adaptive Neuro-Fuzzy Inference Systems (ANFIS) were widely used to classify retinal features. SVM, in particular, demonstrated strong performance in separating lesion and non-lesion regions due to its capability of handling high-dimensional feature spaces.[3][4]

Feature extraction played a central role in these methods. Structural features such as area, eccentricity, and perimeter were used alongside intensity-based measures including mean pixel values and variance. While these hybrid feature-based systems improved detection accuracy, they remained dependent on hand-crafted features. As a result, performance varied significantly depending on dataset quality, and generalisation to unseen data was limited. [7][8][9]

Another limitation of traditional machine learning approaches was their reliance on accurate segmentation of background regions such as the optic disc and blood vessels. Misclassification often occurred when bright lesions overlapped with the optic disc or when vessel segmentation was incomplete. Some methods attempted to circumvent this by focusing only on vascular segmentation, but this excluded other critical features necessary for grading.

Rise of Deep Learning [1][2]

The introduction of deep learning marked a major shift in DR detection. Convolutional Neural Networks (CNNs) demonstrated the ability to learn features directly from raw images, eliminating the need for manual feature engineering. Models such as AlexNet, VGGNet, ResNet, and Inception networks were applied to large-scale retinal datasets, achieving state-of-the-art performance in many cases.

Deep learning approaches offered several advantages, including robustness to noise, the ability to capture complex lesion patterns, and scalability for large datasets. Systems such as IDx-DR and EfficientNet demonstrated that CNN-based approaches could achieve clinical-grade performance. However, deep learning also introduced challenges. These models required vast amounts of annotated data, which is often unavailable, and their computational demands made them less feasible for resource-constrained settings. Furthermore, deep learning models often functioned as “black boxes,” providing little interpretability for clinicians.

Another weakness was that deep learning approaches sometimes neglected pre-processing or segmentation, assuming the network would learn all relevant features. In practice, this resulted in reduced performance on low-quality images or when background features closely resembled lesions. Hybrid approaches, combining pre-processing, segmentation, and deep learning, have therefore been proposed to strike a balance between interpretability and performance.

Critical Evaluation

From this literature, several conclusions can be drawn. First, lesion-specific methods are inadequate for complete grading because they fail to account for the interplay between different lesion types. Second, while deep learning has achieved impressive results, its reliance on large datasets and high computational requirements make it less suitable for environments with limited resources. Third, hybrid approaches that combine pre-processing, segmentation, and classification offer a promising balance of accuracy, interpretability, and efficiency.

The present project is designed in light of these observations. By incorporating robust pre-processing, accurate segmentation, and a hybrid feature-based classification strategy, it aims to address the weaknesses of prior systems while leveraging their strengths.

2.2 Technology Review

A critical review of available technologies is necessary to determine the most appropriate tools and methods for this project. This includes both image processing techniques and machine learning algorithms, as well as the software platforms that support their implementation.

Image Pre-Processing Technologies

Image pre-processing is essential for enhancing the quality of retinal images. Techniques such as histogram equalisation, CLAHE, and spatial collaborative contrast enhancement (SCCE) are widely used. SCCE is particularly effective as it adjusts contrast based on spatial dependencies between pixels, enabling clearer visualisation of retinal features. Compared with CLAHE, SCCE offers improved handling of local contrast while reducing over-enhancement. [1]

Segmentation Techniques

Segmentation of the optic disc and blood vessels is a critical step to reduce misclassification. Mathematical morphology remains one of the most effective approaches for vessel segmentation, often combined with top-hat and bottom-hat transforms. For optic disc segmentation, composite filtering methods that combine Gaussian, Median, and Variance filters have demonstrated strong performance. These methods are computationally efficient and suitable for small datasets, making them appropriate for the project. [6][9]

Feature Extraction Methods

Hybrid feature extraction remains an important component of lesion classification. Structural features such as area, eccentricity, and solidity provide geometric information, while intensity-based features such as mean, variance, and standard deviation capture local image properties. Together, these features provide a comprehensive representation of lesion characteristics, helping to differentiate between bright and red lesions. [7][8]

Machine Learning Algorithms

Among machine learning methods, Support Vector Machines are particularly suitable for high-dimensional feature spaces and have consistently demonstrated strong classification performance. K-Nearest Neighbors provides simplicity and interpretability, though it can be sensitive to noise. Gaussian Mixture Models are effective for clustering-based classification and provide probabilistic outputs. The combination of these algorithms allows comparative evaluation and ensures robustness in classification.

While deep learning offers strong performance, its reliance on large datasets and computational resources makes it less suitable for this project, which aims to demonstrate a practical and efficient grading system using limited data. [1][2][3][4][5]

Software and Platforms

Python has been selected as the primary implementation platform, due to its powerful image processing toolbox and integrated support for machine learning algorithms. It allows rapid prototyping and testing of algorithms, with built-in support for visualisation and statistical analysis. Alternative platform such as MATLAB were considered, but python provides an easier code implementation and more toolkit for image processing.

Rationale for Choice

The choice of SCCE for pre-processing, mathematical morphology for segmentation, hybrid feature extraction, and machine learning classifiers such as SVM, K-NN, and GMM reflects a balance between efficiency, interpretability, and accuracy. These technologies are well-suited for limited datasets and provide clear, reproducible results, making them appropriate for both academic research and potential clinical applications.

2.3 Summary

This chapter has reviewed the literature and technologies relevant to automated diabetic retinopathy grading. Early lesion-specific methods, while pioneering, were incomplete and prone to misclassification. Machine learning introduced more flexibility but remained dependent on hand-crafted features. Deep learning achieved state-of-the-art results but required extensive datasets and resources, limiting its universal applicability. Integrated hybrid systems that combine pre-processing, segmentation, feature extraction, and classification offer the most balanced solution.

The technology review highlighted a range of available options. SCCE was identified as a superior pre-processing technique, while mathematical morphology and composite filtering were considered effective for segmentation. Hybrid feature extraction provides a comprehensive representation of lesions, and machine learning classifiers such as SVM, K-NN, and GMM were found to be robust choices for classification. MATLAB was chosen as the development platform for its integrated capabilities.

Overall, the analysis confirms that a hybrid feature-based machine learning approach offers a practical and effective solution for automated DR grading in this project. The insights gained from the literature and technology reviews directly inform the methodology by emphasising robust pre-processing, reliable segmentation, and efficient classification strategies as the foundation for the proposed system.

Chapter 3 Implementation

3.1 System Design and Architecture

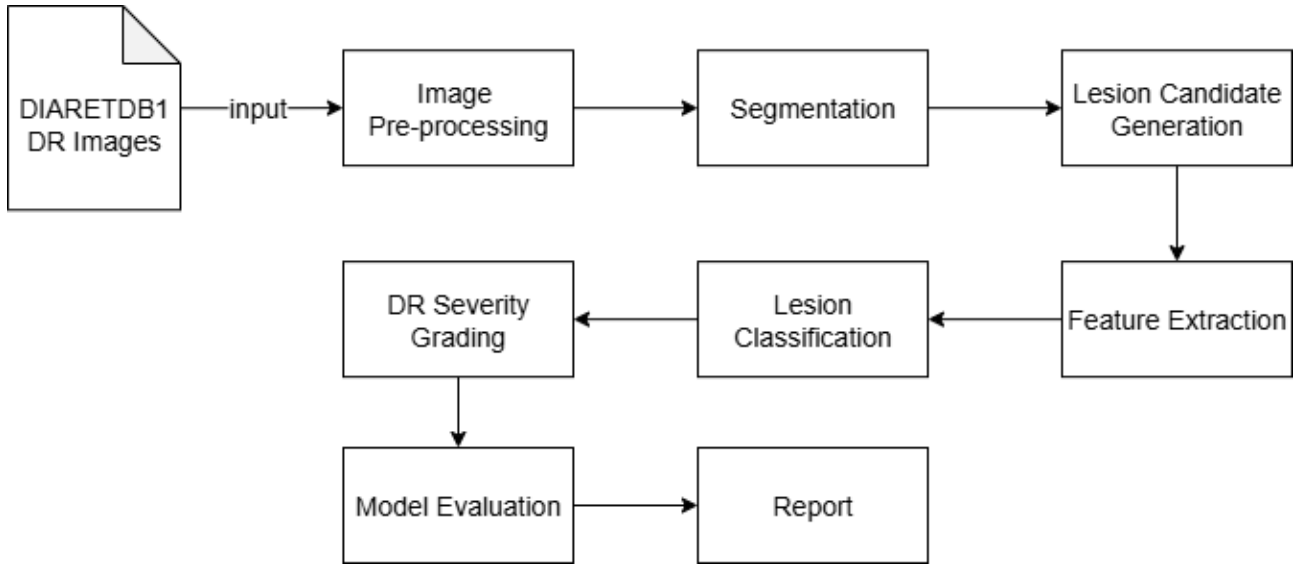


Figure 2: System Architecture

1. **Image Pre-processing** – standardise resolution, enhance contrast, denoise, and normalise the fundus images.
2. **Segmentation** – suppress normal anatomy by extracting **blood vessels** and the **optic disc** to reduce false positives.
3. **Lesion Detection & Feature Extraction** – generate lesion **candidates** on the cleaned retina and compute a **30-D hybrid feature vector** (appearance, shape, context) for each candidate.
4. **Lesion Classification** – apply a **hierarchical** scheme: Stage-1 (true vs. false) within the bright and red families, then Stage-2 subtype classification (**EX/CWS** or **MA/HM**).
5. **Severity Grading** – aggregate lesion counts and map them to an image-level DR grade (**No**, **Mild**, **Moderate**, **Severe**).

Fig 2 illustrates the data flow from DIARETDB1 inputs through pre-processing and segmentation to candidate detection, feature extraction, classification, and finally image-level grading. Each module is loosely coupled and executed in sequence, which simplifies debugging, testing, and iterative refinement. The **Model Evaluation** and **Report** blocks consume the grading outputs to produce metrics and visual summaries.

3.2 Dataset and Data Preparation

Description

The DIARETDB1 (Diabetic Retinopathy Database 1) is a publicly available benchmark dataset widely used in the development and evaluation of automated diabetic retinopathy (DR) detection systems. It was created to support research into image analysis methods for DR screening and is particularly valuable due to its expert annotations of retinal lesions. It has 89 color fundus images (84 images with signs of DR and 5 normal). The images were captured at Kuopio University Hospital (Finland) using a fundus camera with a 50° field of view.

The dataset provides image-level and lesion-level ground truth through expert annotations. Lesion information is supplied in XML files, each corresponding to a single image, with coordinate points marking the centers of suspected lesions. Each annotation also includes the lesion type label, covering four major categories: MA, HE, EX, and CWS.

The dataset was annotated independently by four ophthalmologists to ensure reliability and consensus. Due to its carefully curated annotations and expert validation, DIARETDB1 is regarded as a benchmark dataset for academic DR research. However, its relatively small size makes it more appropriate for algorithm prototyping, validation, and feature-based approaches than for training large-scale deep learning models.

	image_name	lesion_type	centroid_x	centroid_y	radius	polygon	rep_x	rep_y	confidence
0	diaretdb1_image001.png	EX	713	532	141.0	None	705	536	High
1	diaretdb1_image001.png	HE	493	647	5.0	None	493	647	High
2	diaretdb1_image001.png	HE	978	354	5.0	None	978	354	High
3	diaretdb1_image001.png	HE	618	94	5.0	None	618	94	High
4	diaretdb1_image001.png	HE	597	99	14.0	None	601	99	High
...
10325	diaretdb1_image089.png	MA	1237	542	5.0	None	1237	542	High
10326	diaretdb1_image089.png	MA	958	561	5.0	None	958	561	High
10327	diaretdb1_image089.png	MA	883	256	5.0	None	883	256	High
10328	diaretdb1_image089.png	MA	406	881	5.0	None	406	881	Medium
10335	diaretdb1_image089.png	MA	885	257	13.0	None	884	255	Medium

4875 rows × 9 columns

Figure 3: Annotation Data Frame Sample

Annotations Overlay

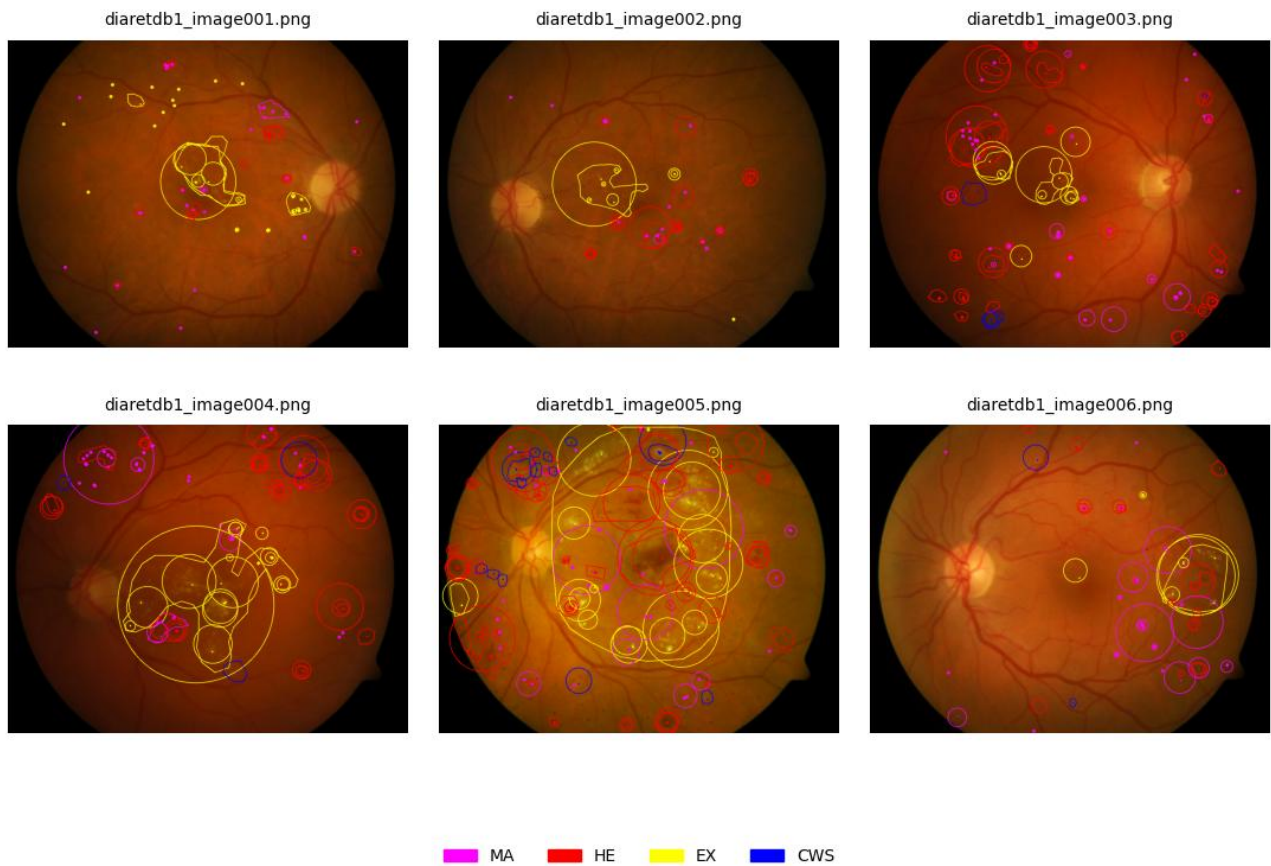


Figure 4: Annotations Overlay on Raw Fundus Images

Fig 4 illustrates the overlaid retinal image with expert-provided lesion annotations, where different colors represent different lesion types. Magenta markers denote *MA*, red circles indicate *HE*, yellow contours represent *EX*, and blue outlines correspond to *CWS*. The annotations were generated from XML coordinate points provided in the dataset and reviewed by four independent ophthalmologists to ensure accuracy. These overlays highlight the distribution and diversity of lesions across the retina, which form the ground truth for lesion detection, classification, and severity grading in the proposed system.

Exploratory Data Analysis (EDA)

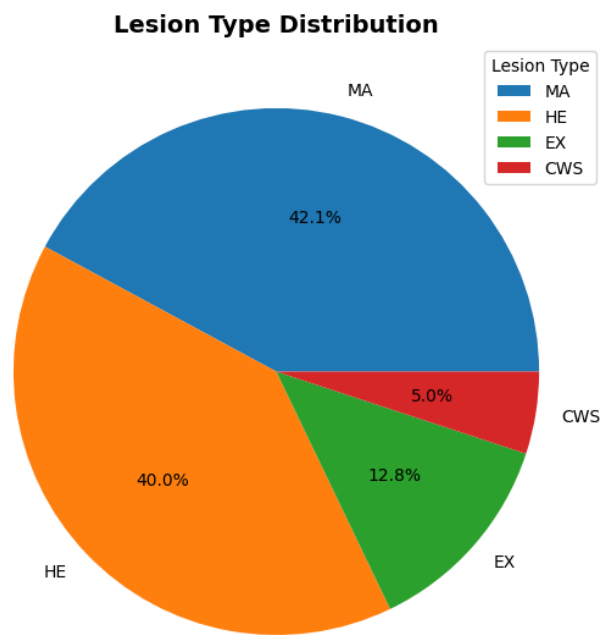


Figure 5: Lesion Type Distribution

The dataset contains four major lesion categories: **MA, (42.1%), HE (40.0%), EX (12.8%),** and **CWS (5.0%)**. Fig 5. illustrates that *MA* and *HE* are the most prevalent lesion types, together accounting for over 80% of all annotations. In contrast, *CWS* occur relatively infrequently. This imbalance is consistent with clinical observations, as *MA* are often the earliest indicators of DR, while *HE* is common in progressive stages. The skewed distribution has methodological implications: classifiers must be designed to handle class imbalance to ensure accurate detection across all lesion types, particularly the less frequent *CWS*.

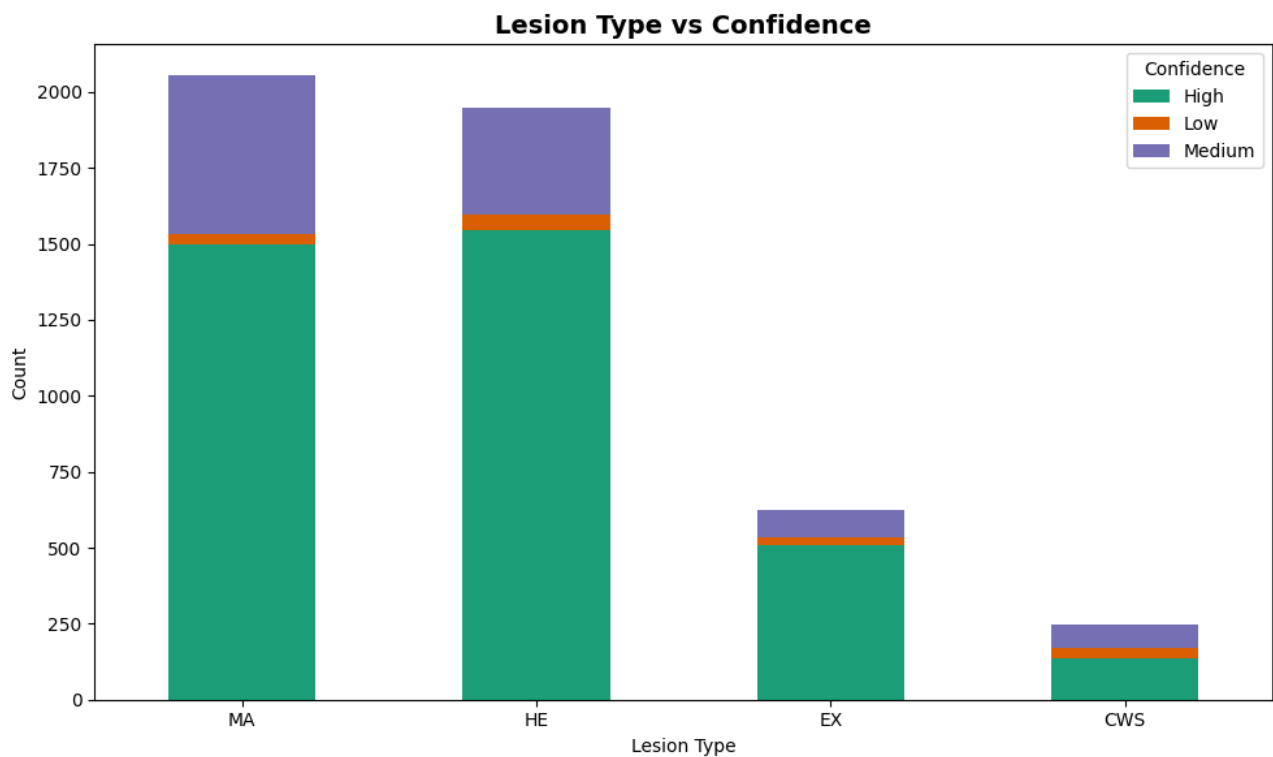


Figure 6: Lesion Type vs Confidence

Fig 6. shows lesion counts for the four annotated categories — **MA, HE, EX, and CWS** which is further divided by annotation confidence levels (high, medium, and low). Most annotations fall into the **high-confidence** category, particularly for MA and HE, which dominate the dataset. EX and CWS, although less frequent, also show a predominance of high-confidence annotations but with a proportionally larger share of medium-confidence cases. Low-confidence annotations represent only a small fraction across all lesion types. This confidence stratification is important for model development, as high-confidence annotations provide reliable ground truth for training, while medium and low-confidence points may require careful handling (e.g., weighting schemes or exclusion) to avoid introducing label noise during classifier training.

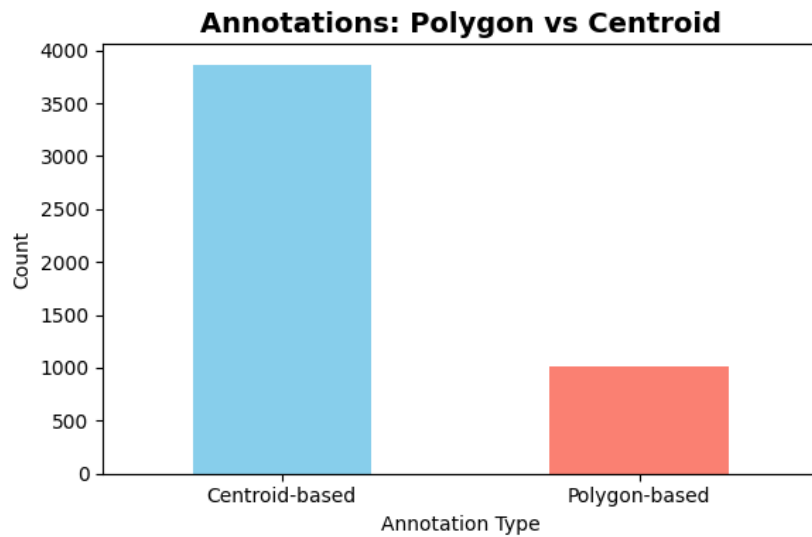


Figure 7: Annotations: Polygon vs Centroid

Fig 7. Shows the two main forms of expert annotations: **centroid-based** and **polygon-based**. Centroid-based annotations, where lesions are marked with a single $(x,y)(x,y)(x,y)$ coordinate representing the lesion center, are the most common, with nearly 3,900 instances. In contrast, polygon-based annotations, which delineate lesion boundaries more precisely, account for just over 1,000 cases. The predominance of centroid annotations reflects the dataset's emphasis on lesion localization rather than detailed boundary tracing. For model development, centroid-based annotations are typically sufficient for lesion detection and classification tasks, whereas polygon-based annotations are particularly useful for evaluating segmentation performance or when lesion size and shape need to be explicitly quantified.

3.3 Image Pre-Processing

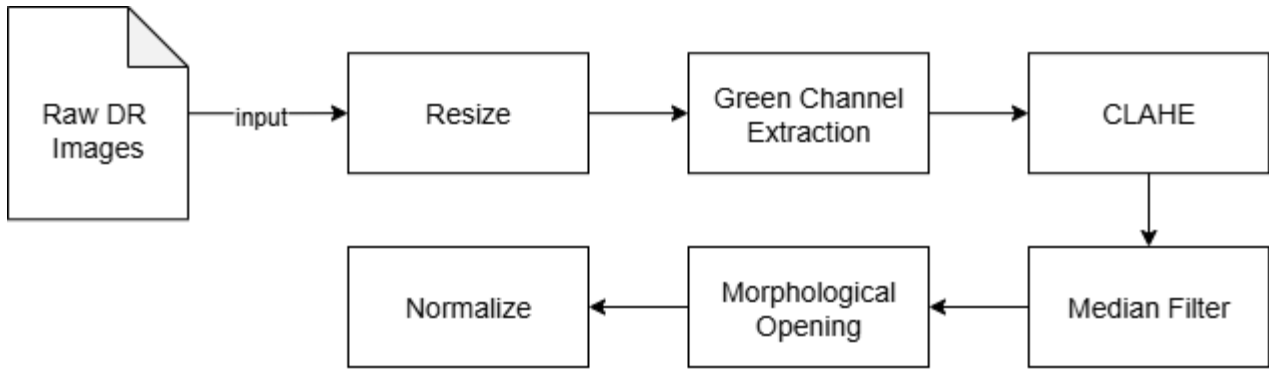


Figure 8: Image Pre-Processing Pipeline

The image pre-processing pipeline was designed to standardize the raw fundus images, suppress artefacts, and enhance lesion visibility prior to segmentation and classification. The pipeline consists of six sequential stages as illustrated in Fig 9. Each stage is described below.

Pipeline Description:

1. **Resize** – All images were resized to **512 × 512 pixels** to ensure uniformity and reduce computational cost.
2. **Green Channel Extraction** – Fundus images are captured in RGB format, but the **green channel** provides the highest contrast between retinal structures and background. By isolating channel Green, noise from red and blue channels is suppressed.
3. **Contrast Enhancement (CLAHE)** – Applied Contrast Limited Adaptive Histogram Equalization (CLAHE) with a clip limit of 2.0 and tile grid size of 8×8 to improve local contrast and make lesions more visible.
4. **Median Filtering** – Used a median filter (kernel size 3) to suppress salt-and-pepper noise while preserving edges.
5. **Morphological Opening** – Applied with a 3×3 elliptical kernel to remove small artefacts and smooth the background.
6. **Normalization** – Pixel intensity values were normalized to the range [0,1] for numerical stability in subsequent processing.

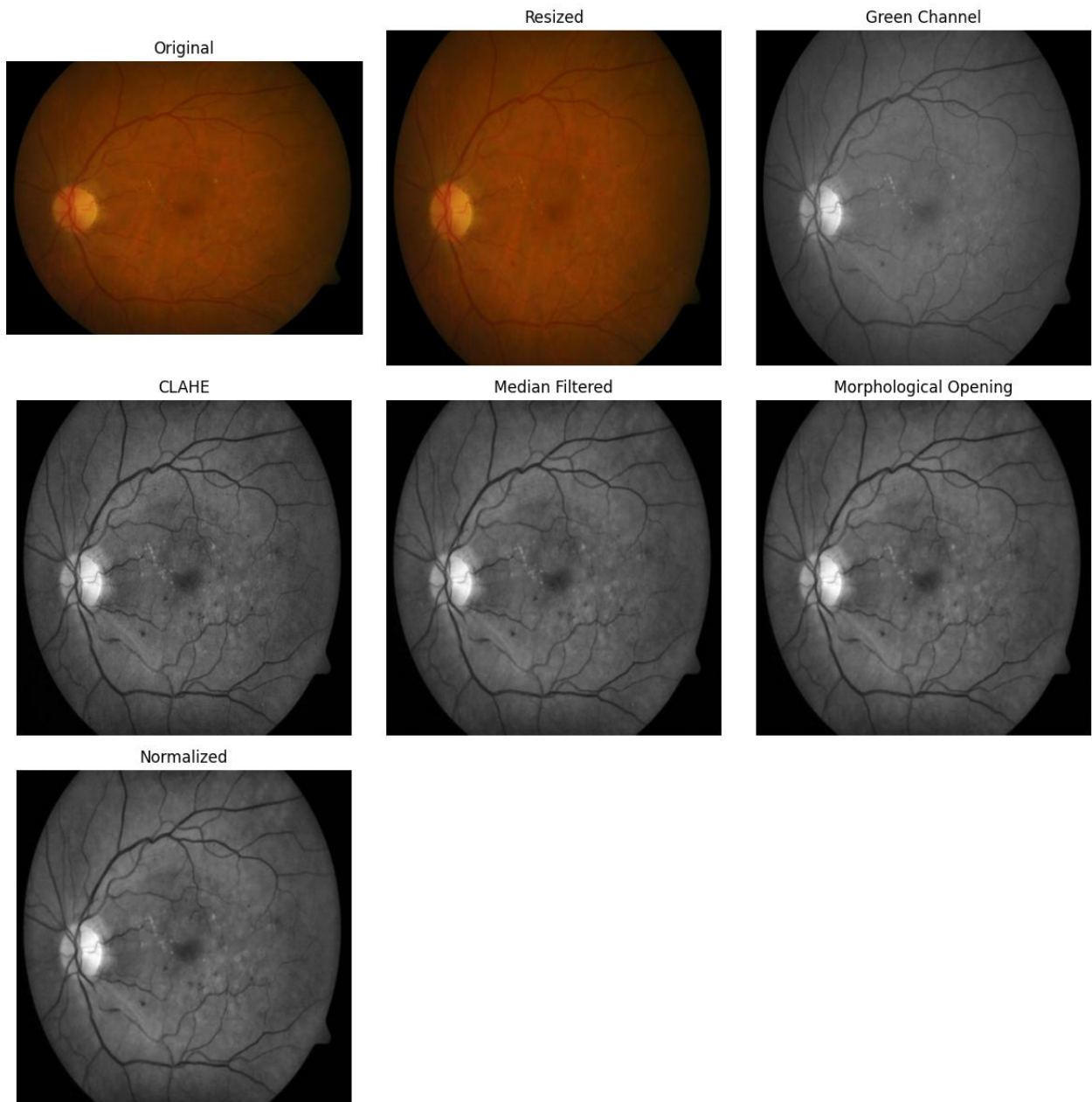


Figure 9: Image Pre-Processing Stages

Fig 10. illustrates the sequential steps in the pre-processing pipeline: the original fundus image is first resized to a fixed resolution (512×512 pixels), after which the **green channel** is extracted to enhance vessel and lesion contrast. Contrast Limited Adaptive Histogram Equalization (**CLAHE**) is then applied to improve local contrast, making MA and EX more visible. A **median filter** (3×3 kernel) reduces salt-and-pepper noise while preserving edges, followed by **morphological opening** with an elliptical structuring element to suppress small artefacts and smooth background variations. Finally, the image is **normalized** to the [0,1] range to standardize pixel intensities for subsequent processing.

3.4 Segmentation

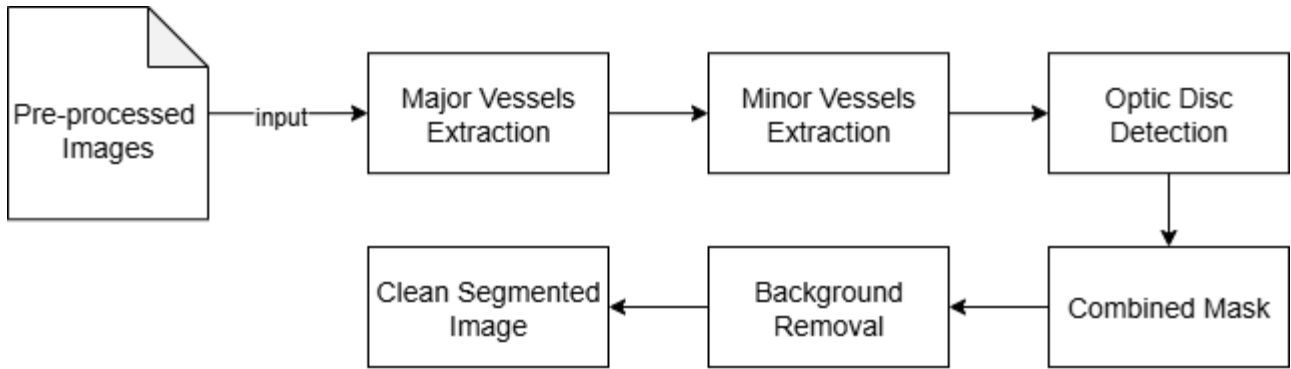


Figure 10: Segmentation Pipeline

The **segmentation pipeline** was developed to suppress background structures such as retinal vessels and the optic disc, which often resemble lesions and contribute to false positives in detection. The pipeline consists of a sequence of stages, illustrated in **Fig 11**, that progressively extract vessel masks, detect the optic disc, and generate a combined background mask. Each stage is described below.

Pipeline Description:

1. Major and Minor Vessels:

The large blood vessels in the retina are detected first, since they appear as strong, tubular structures. Smaller vessel branches are extracted separately using a lower threshold. This ensures that thin vessels, which can resemble small lesions, are also suppressed.

2. Optic Disc Detection:

The optic disc is identified as the brightest circular region in the retina. It is masked out because its brightness is similar to EX and can cause false positives.

3. Combine Mask:

The masks of major vessels, minor vessels, and the optic disc are merged into a single background mask. Morphological operations are used to refine the mask and close small gaps.

4. Background Removal:

The combined mask is applied to the pre-processed image to remove vessels and the optic disc, leaving behind only lesion-relevant regions.

5. Clean Segmented Image:

The final output is a cleaned image where the background structures are suppressed, making lesions easier to detect and classify.

3.5 Segmentation Stages

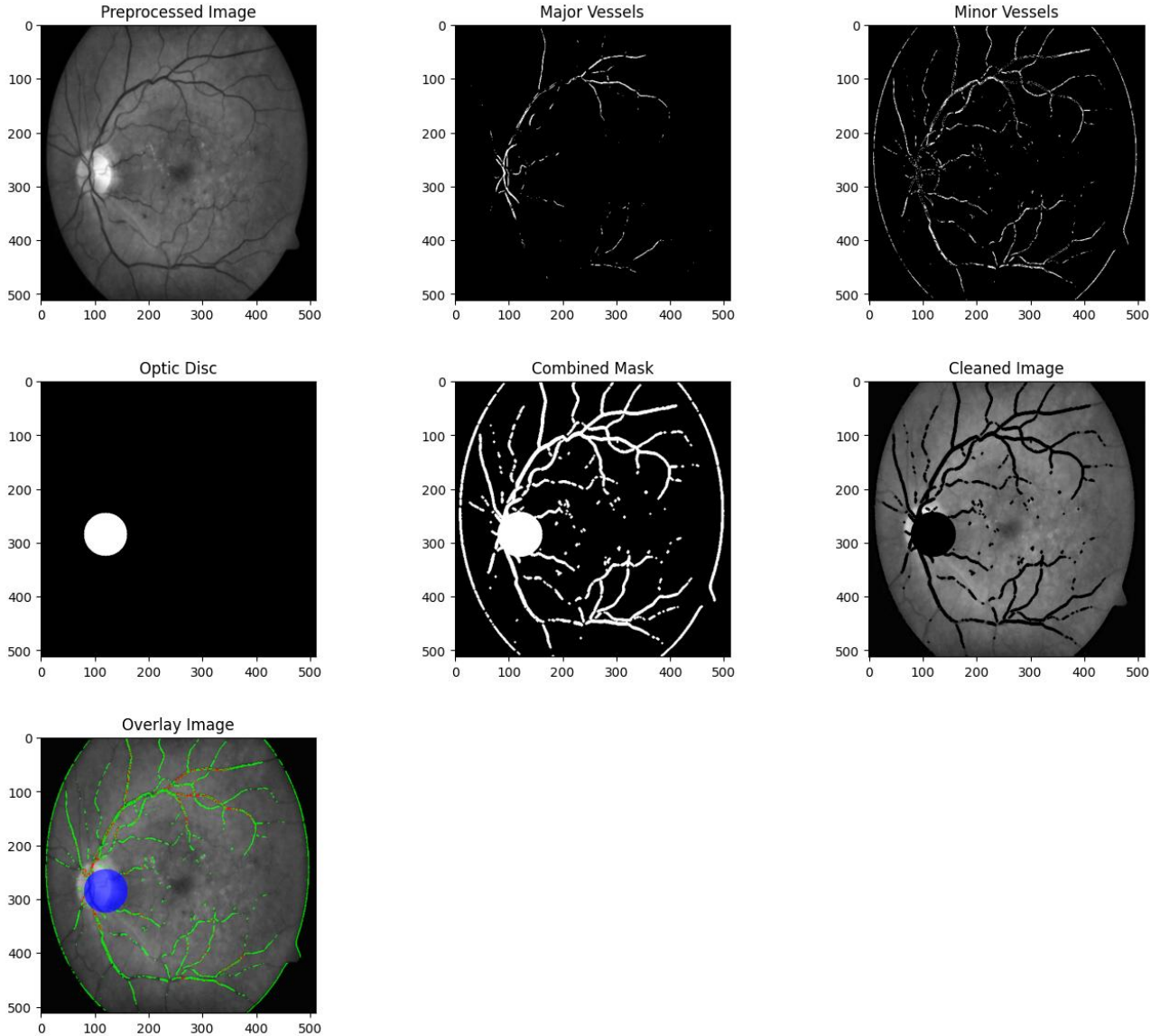


Figure 11: Segmentation Stages

Fig 12. illustrates the sequential steps of the segmentation pipeline. Starting with the **pre-processed image**, the **Frangi filter** was applied to enhance vascular structures. High and low thresholding produced separate masks for **major vessels** and **minor vessels**. The **optic disc** was detected by Gaussian blurring and intensity maxima localization, and represented by a circular mask. These masks were merged into a **combined mask**, which was refined using morphological dilation. The mask was then applied to the input to generate the **cleaned image**, where vessels and the optic disc were suppressed, preserving only lesion-relevant regions. The **overlay image** demonstrates the alignment of vessel and optic disc masks on the original fundus image, confirming accurate background suppression.

3.6 Major Vessel Extraction

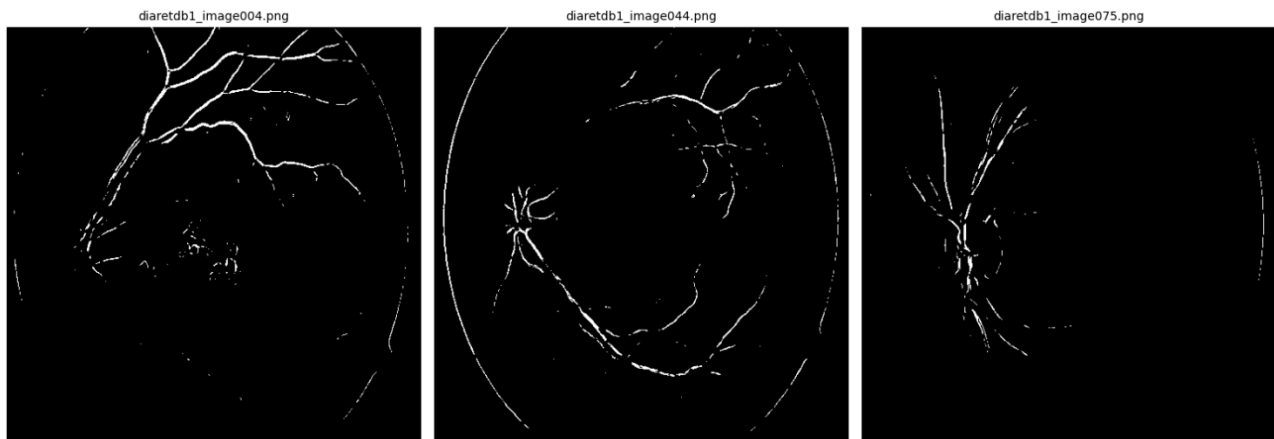


Figure 12: Major Vessels Sample

Fig 13. Illustrates sample major vessels extracted for Background Segmentation. Large retinal vessels often appear as dark, elongated tubular structures that run across the fundus. These vessels can closely resemble pathological lesions, particularly HE, when only intensity or shape-based features are considered. To address this, the Frangi vesselness filter was applied to enhance and isolate strong vascular structures. By setting a high threshold, only the major vessels with the strongest vesselness response were retained in a binary mask. Removing these large vessels from the analysis is essential because their intensity variations and linear morphology can easily be mistaken for elongated HE. Suppressing them ensures that the lesion detection stage focuses on true pathological features rather than normal anatomical structures.

3.7 Minor Vessel Extraction

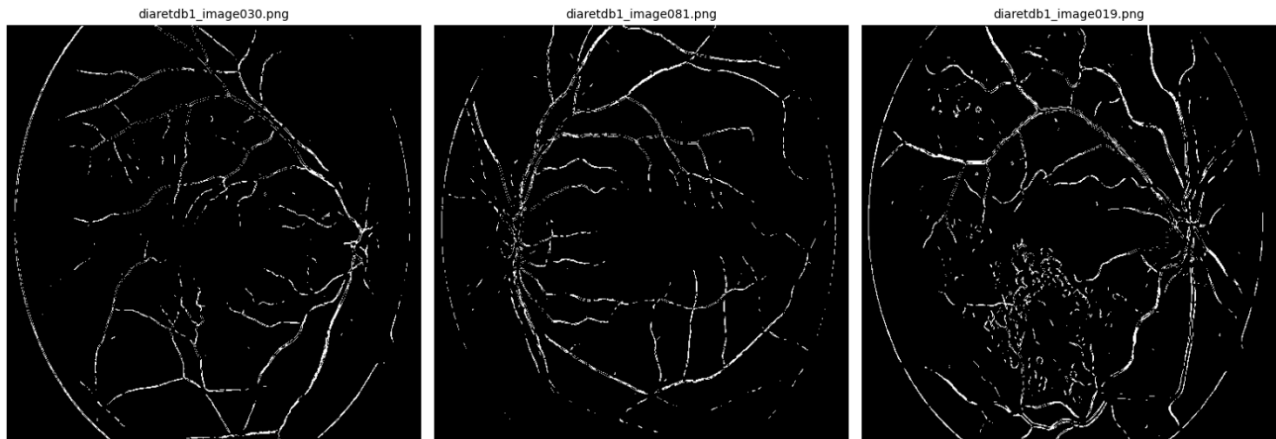


Figure 13: Minor Vessels Sample

In addition to major vessels, the retina also contains thin capillaries that form a dense vascular network. These small vessels, while physiologically normal, can resemble early-stage lesions such as MA due to their small, rounded appearance when fragmented or partially visible. To minimize such false positives, a lower threshold was applied to the Frangi response, generating a mask of **minor vessels** illustrated in Fig 14. By suppressing these finer vascular branches, the system avoids misclassifying them as MA or small HE, which are key indicators of diabetic retinopathy progression. This step improves the specificity of lesion detection by ensuring that only genuine abnormal structures remain visible.

3.8 Optic Disc Detection

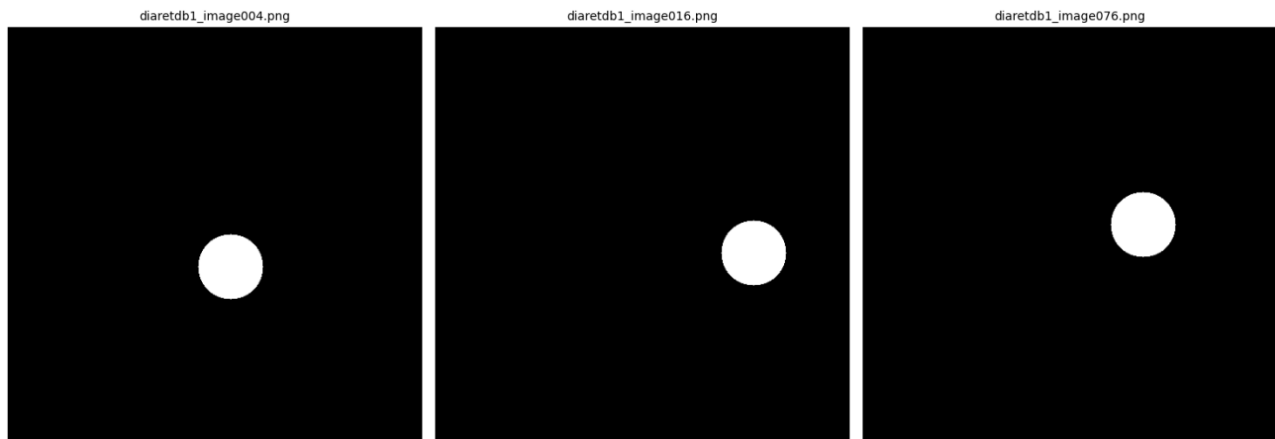


Figure 14: Optic disc Sample

The optic disc (OD) is one of the brightest structures in a retinal fundus image, characterized by a circular region where blood vessels converge. Its brightness and shape often overlap with the visual characteristics of bright lesions such as EX or CWS. Without proper masking, the optic disc could significantly increase false positives in lesion classification, as algorithms may interpret it as a large exudate cluster. To prevent this, Gaussian blurring was first applied to smooth local intensity variations, followed by identification of the brightest point in the image. A circular mask illustrated in Fig 15 was then created around this point to cover the optic disc region. By removing the optic disc, the system eliminates a major source of false positives, improving both sensitivity and specificity in detecting actual pathological bright lesions.

3.9 Background Removal and Clean Segmented Image

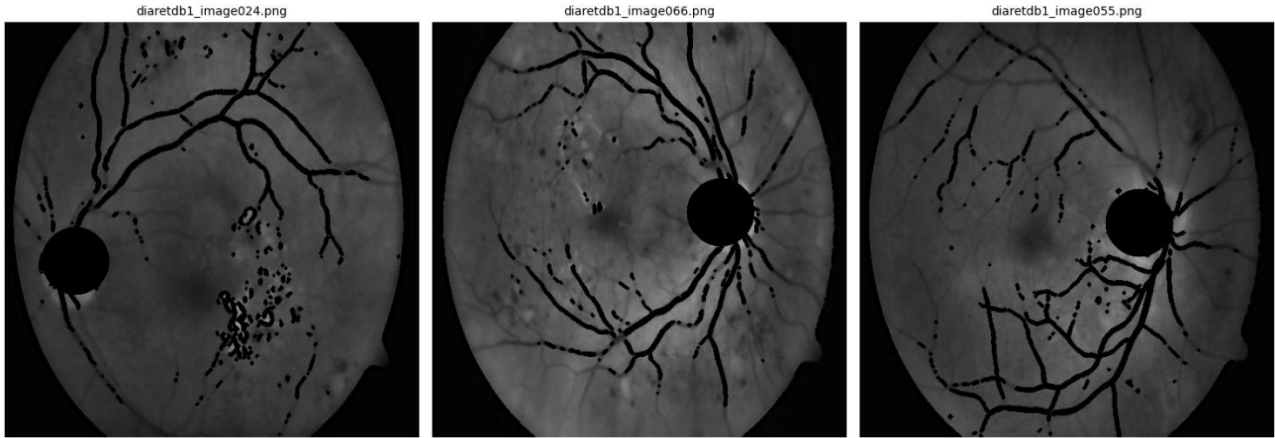


Figure 15: Background Removed Clean Segmented Sample

The final step of segmentation involved applying the combined mask directly to the pre-processed image illustrated in Fig 17, setting all pixels corresponding to vessels and the optic disc to zero. This effectively removed normal anatomical structures, leaving only potential lesion regions visible. Since diabetic retinopathy lesions such as MA, HE, and EX typically occur on the retinal background rather than within vessel walls or on the optic disc, masking these non-relevant areas significantly reduced the search space for lesion detection. The resulting **clean segmented image** provided a focused and accurate representation of the retina, minimizing the risk of confusing normal structures with pathological ones. By ensuring that only lesion-relevant regions remained, this step improved computational efficiency, reduced false positives, and enhanced the robustness, forming a reliable foundation for subsequent lesion detection and severity classification.

3.10 Lesion Candidate Generation

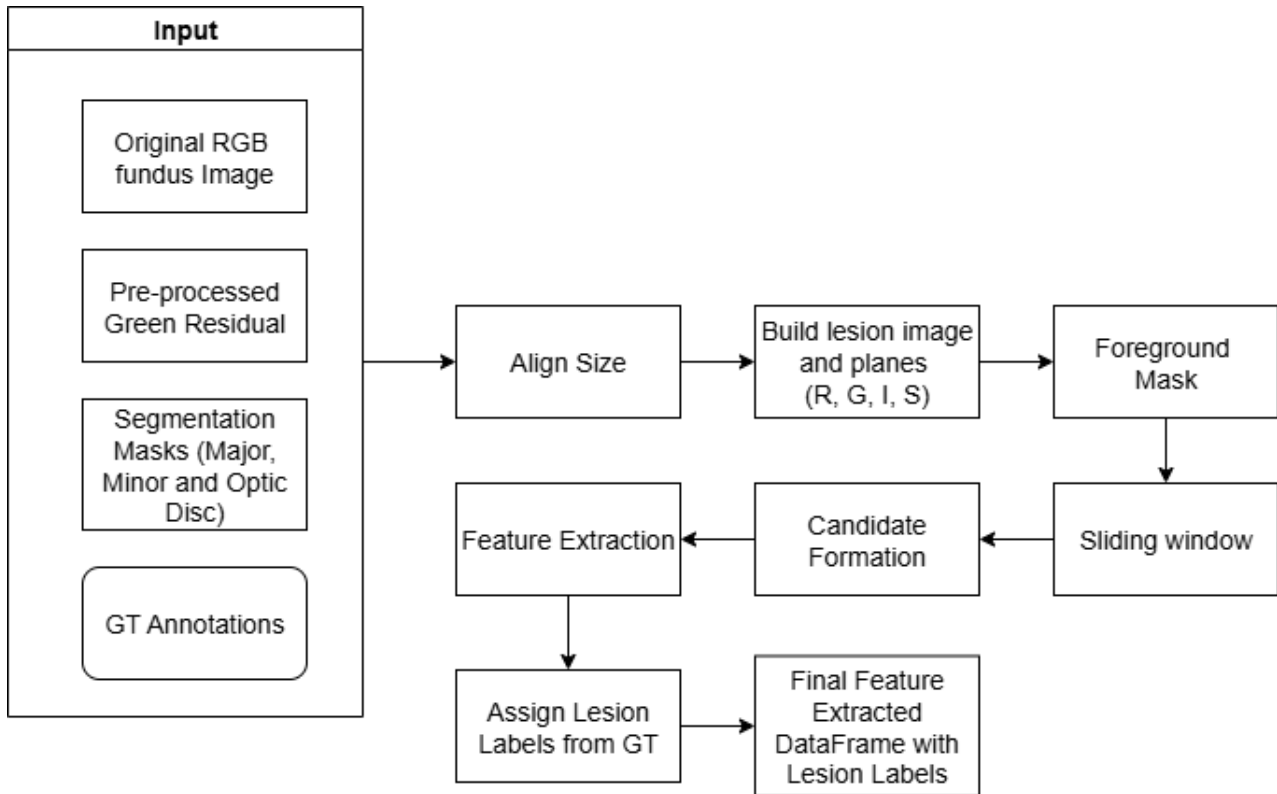


Figure 16: Lesion Candidate Generation Pipeline

Pipeline:

1. Original RGB fundus image

This is the raw colour photograph captured by the fundus camera. It provides the red and blue channels used later for appearance cues and for computing auxiliary planes (intensity and saturation). Keeping the original RGB alongside the processed green channel preserves clinically useful colour information.

2. Pre-processed Green Residual

The green residual is the pre-processed green channel after contrast enhancement and background suppression. It offers the highest lesion-to-background contrast and is the primary signal used for proposing candidates in the sliding-window stage.

3. Segmentation masks (major, minor, optic disc)

Binary masks flagging large/small vessels and the optic disc are used to exclude normal anatomy from analysis. Removing these regions up front reduces systematic false positives.

4. Ground-truth (GT) annotations

Expert annotations (polygons, centroid, radius) are used only for supervision and evaluation. They do not drive detection; instead, they assign labels to the candidates generated.

5. Align Size

The RGB fundus image is resized to exactly match the height and width of GrG_rGr. This pixel-level alignment ensures that all planes (R, GrG_rGr, B) and all masks index the same spatial locations, preventing feature/label mismatches downstream.

6. Build lesion image and planes (R, G, I, S)

Derive two auxiliary planes: mean intensity I and HSV saturation S. Using these four planes [R, G, I, S] captures complementary cues colour, brightness and local contrast which help distinguish bright lesions (EX/CWS) from red lesions (MA/HE).

7. Foreground mask

A foreground gate is created by excluding all pixels marked by vessel and optic-disc masks. Only the remaining retinal background is considered for candidate search, which improves efficiency and reduces false alarms on normal anatomy.

8. Sliding window

The image is scanned with overlapping 17x17 windows at a stride of 5 pixels. Windows with insufficient foreground are skipped to avoid unstable measurements at boundaries and to keep computation focused.

9. Candidate formation

For each kept window we produce a single, stable candidate region centred on the window location. Practically, we form a small binary region around the centre (via local adaptive thresholding) and keep the connected component that contains the centre pixel. This enforces one candidate per window and yields a coherent blob for shape analysis.

10. Feature extraction

From every candidate/window we compute a 30-dimensional hybrid descriptor. Appearance (16 features): mean, variance, minimum and maximum over R, G, S and I. Together this capture how the region looks, how it is shaped and where it sits anatomically.

11. Assign lesion labels from GT

Each candidate's centre is intersected with the ground-truth for the same image: point-in-polygon if a polygon exists, otherwise an inside-circle test using centroid and radius. Candidates are labelled as lesion or not-lesion, with lesion group and type (MA/HE/EX/CWS).

12. Final feature-extracted DataFrame with lesion labels

All candidates are consolidated into a table containing image ID, coordinates, the 30 features and GT-derived labels. This dataset is the direct input to the Lesion Classification stage (e.g., SVM/KNN/GMM) and later aggregation for Severity Grading.

3.11 Feature Extraction

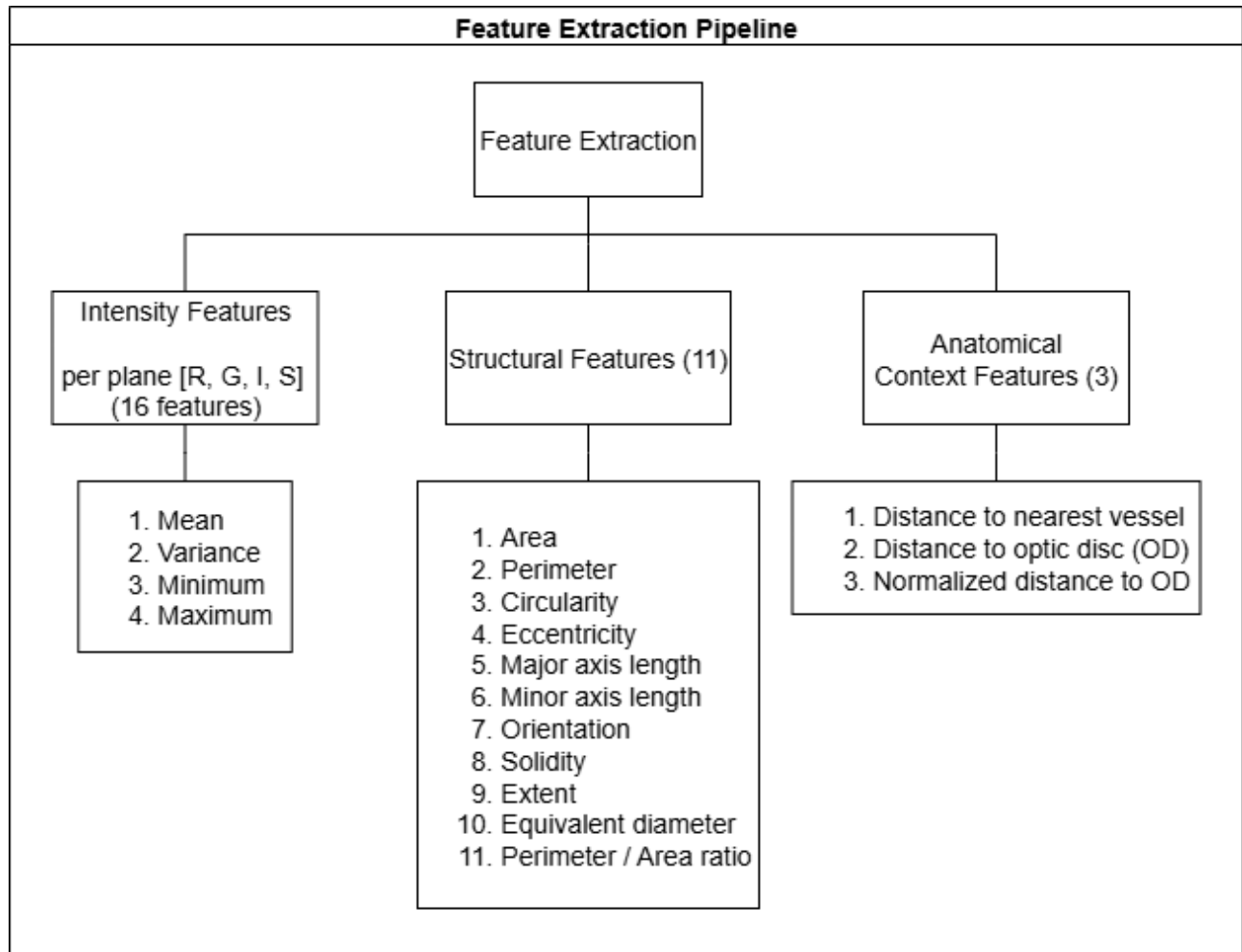


Figure 17: Feature Extraction Pipeline Features

Fig 20 illustrates the features created per candidate. A total of 30 features were computed.

No.	Intensity Features	Description
1	Mean	average brightness
2	Variance	local contrast/texture
3	Minimum	darkest value
4	Maximum	brightest value

Table 1: Intensity Features

No.	Structural Features	Description
1	Area	size of the region
2	Perimeter	boundary length
3	Circularity	roundness (MAs \approx high)
4	Eccentricity	elongation (HEs \approx higher)
5	Major axis length	long radius of fitted ellipse
6	Minor axis length	short radius of fitted ellipse
7	Orientation	angle of the fitted ellipse
8	Solidity	convex-hull area (compactness)
9	Extent	bounding-box area (fill factor)
10	Equivalent diameter	diameter of a circle with same area
11	Perimeter / Area ratio	boundary complexity

Table 2: Structural Features

No.	Anatomical Context Features	Description
1	Distance to nearest vessel	Pixels from candidate center to the closest vessel
2	Distance to optic disc (OD)	Euclidean distance from candidate center to OD center
3	Normalized distance to OD	OD distance divided by OD diameter

Table 3: Anatomical Context Features

3.12 Lesion Classification

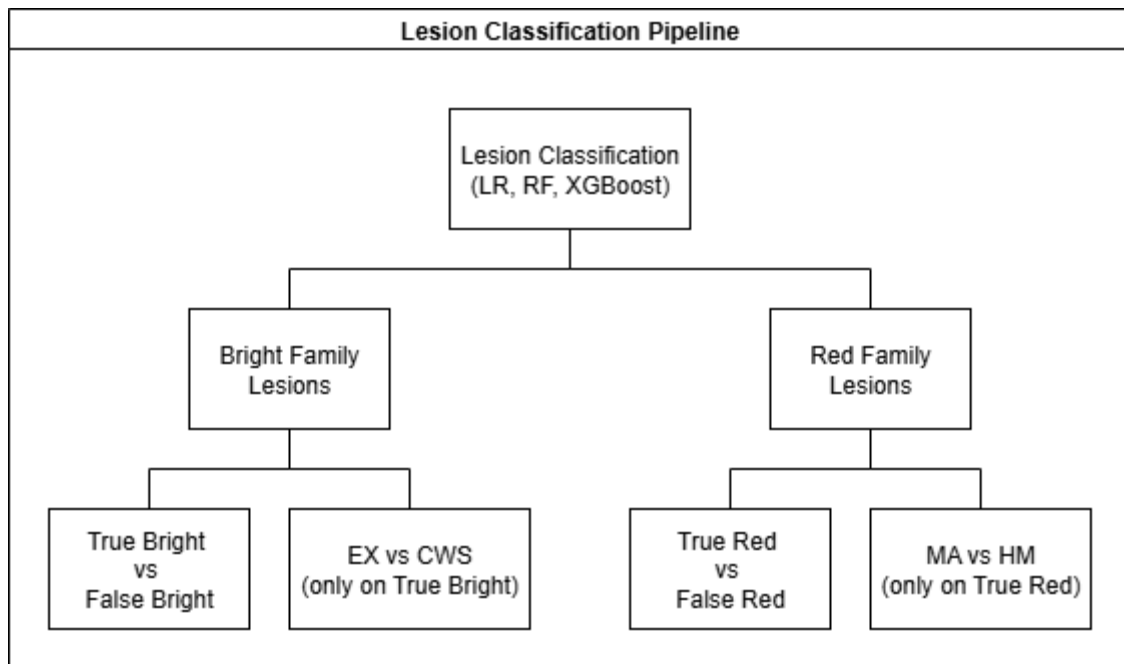


Figure 18: Lesion Classification Pipeline

The lesion classification module takes the 30-dimensional feature vector for each candidate and routes it through a two-branch, two-stage hierarchy illustrated in Fig 21. Candidates are first separated into bright and red families; each family runs a Stage-1 filter that distinguishes *true lesions* from *false look-alikes* using a supervised classifier. For training, positives are the family’s true lesions (EX/CWS for bright, MA/HM for red) and negatives are not_lesion samples; lesions from the other family are excluded. Only candidates that pass Stage-1 proceed to Stage-2 subtype classification—EX vs CWS in the bright branch and MA vs HM in the red branch. The outputs of both branches are then aggregated by counting {EX, CWS, MA, HM} per image, which forms the input to the subsequent severity grading step.

3.13 Severity Grading

Severity grading works at the image level. After classification, we count each lesion type that is EX, CWS, MA, HM and form two totals: bright and red. Once the system has tagged the little spots in a photo, it counts how many of each kind it can see. We group them into “bright” changes (yellow-white patches like EX or CWS) and “red” changes (tiny red dots or bleeds such as MA and HE). As a rule of thumb, more red or bright changes = more serious disease. Using a small set of straightforward rules, the system turns those counts into an overall stage for that eye: No DR, Mild, Moderate, or Severe.

3.14 Technologies Used

- **Python** was the primary implementation language.

Core Scientific Stack

- **NumPy** — array operations and numerical routines.
- **Pandas** — tabular data handling (feature tables, labels, experiment logs).
- **SciPy (ndimage)** — distance transforms, morphology helpers (e.g., `distance_transform_edt`).

Image Processing

- **OpenCV (cv2)** — I/O, resizing, color-space conversion (RGB \leftrightarrow HSV), filtering, morphology.
- **scikit-image** — feature-level ops: `filters.frangi` (vesselness), `measure.regionprops`, `morphology.convex_hull_image`, labeling.

Machine Learning

- **scikit-learn** — model training/evaluation:
 - Utilities: train/test splits, cross-validation, scalers (min-max), metrics (accuracy, sensitivity, specificity, ROC-AUC).

Visualization & Reporting

- **Matplotlib** — plots for ROC curves, confusion matrices, and qualitative overlays.
- **Seaborn** — for EDA visualizations

Data Preparation

- **XML parsing** — convert DIARETDB1 annotation XML (polygons/centroids) into a structured dataframe used for supervision and evaluation.

Utilities & Tooling

- **tqdm** — progress bars for dataset loops.

3.15 Summary

This chapter detailed the end-to-end implementation of the Automatic Diabetic Retinopathy Grading System (ADRGs) as a modular, data-driven pipeline. Starting from the DIARETDB1 dataset (with expert XML annotations), images are standardised via a compact pre-processing pipeline. A segmentation stage removes normal anatomy—major/minor vessels and the optic disc producing a clean retinal background and lowering false positives downstream. On this background, the system performs lesion candidate generation with masked sliding windows, forms a single centred blob per window, and computes a hybrid 30-D descriptor per candidate that combines appearance (R, G, I, S), shape/structure, and simple anatomical context (distances to vessels/OD). Ground-truth overlays provide labels for supervised learning.

Lesion classification follows a hierarchical design in two parallel streams: bright (Stage-1 True vs False, then EX vs CWS) and red (Stage-1 True vs False, then MA vs HM). The per-candidate outputs are aggregated at the image level for severity grading, where counts of bright and red lesions are mapped via simple rules to No/Mild/Moderate/Severe DR. The system is implemented in Python with NumPy/Pandas/SciPy, OpenCV and scikit-image for vision, scikit-learn for modelling, and Matplotlib/Seaborn for analysis and reporting. This modular design enables clear diagnostics, easy iteration, and a clean handoff to Chapter 4 for Model Evaluation.

Chapter 4 Evaluation and Results

This chapter evaluates the System in terms of technical performance, strengths/weaknesses, and practical considerations.

4.1 Experimental Setup

This study evaluates a simple two-step classifier on DIARETDB1 using the features previously extracted. Those numeric features (e.g., brightness/variance per channel, shape cues like solidity/extent/equivalent diameter, perimeter-area ratio, and distances to vessels/optic disc) are the model inputs; only ID/label columns (image name, x/y, lesion labels) are excluded. Step 1 decides if a region is a lesion or background, done separately for bright and red families. Step 2 then sub-types the positives (EX vs CWS for bright; MA vs HM for red, with HE mapped to HM). To keep testing fair, data are split 80% train / 20% test by image, so patches from the same image never appear in both sets. Because background greatly outnumbers lesions, background is trimmed in Step 1, and only the training split is rebalanced in Step 2 so the test stays realistic. Pre-processing is light: missing values are filled; only Logistic Regression uses feature scaling. Three common tabular models are compared—Logistic Regression, Random Forest, and XGBoost—using a default 0.5 cutoff (with a quick check of an alternative cutoff that balances catching lesions vs avoiding false alarms). Results are reported with sensitivity, specificity, accuracy, and their average (the paper’s AUC), plus ROC-AUC and PR-AUC for extra context. A fixed random seed is used so runs are repeatable.

4.2 Evaluation Metrics

To judge how well the models work, this study reports a small set of clear, class-imbalance-aware metrics. The reference is from [1]:

- **Sensitivity (Recall of the positive class)**

The proportion of true positives that the model correctly finds.

Interpretation: “Of all real lesions of the chosen type, how many were caught?”

Why it matters: screening benefits from high sensitivity.

- **Specificity (True-Negative Rate)**

The proportion of true negatives that the model correctly rejects. *Interpretation:* “Of all background (or the non-positive subtype), how many were correctly ignored?”

Why it matters: high specificity reduces false alarms and reviewer workload.

- **Accuracy.**

The overall fraction of correct predictions (both classes). *Caveat:* with class imbalance, accuracy can be misleading on its own, so it is always read alongside Sensitivity and Specificity.

- **AUC (paper definition).**

Following the source paper, AUC is reported as the average of Sensitivity and Specificity:

Note: this is equivalent to Balanced Accuracy and makes results directly comparable to the paper.

- **ROC-AUC (modern)**

The area under the Receiver Operating Characteristic curve across all probability thresholds.

Interpretation: overall ability to separate positives from negatives, independent of a single cutoff.

- **Confusion Matrix (counts and row-normalized).**

Shows true negatives (TN), false positives (FP), false negatives (FN), and true positives (TP).

Use: makes error types visible (e.g., missed lesions vs false alarms) and supports the three metrics above.

4.3 Stage 1: Bright vs Background

Classification report:					Classification report:				
	precision	recall	f1-score	support		precision	recall	f1-score	support
background	0.8396	0.7101	0.7694	9848	background	0.6461	0.9202	0.7592	9848
bright	0.7318	0.8536	0.7880	9126	bright	0.8412	0.4562	0.5915	9126
accuracy			0.7791	18974	accuracy			0.6970	18974
macro avg	0.7857	0.7818	0.7787	18974	macro avg	0.7437	0.6882	0.6754	18974
weighted avg	0.7877	0.7791	0.7784	18974	weighted avg	0.7399	0.6970	0.6786	18974
Logistic Regression					Random Forest				
Classification report:					Classification report:				
	precision	recall	f1-score	support		precision	recall	f1-score	support
background	0.7101	0.8702	0.7821	9848	background	0.6461	0.9202	0.7592	9848
bright	0.8149	0.6167	0.7021	9126	bright	0.8412	0.4562	0.5915	9126
accuracy			0.7483	18974	accuracy			0.6970	18974
macro avg	0.7625	0.7435	0.7421	18974	macro avg	0.7437	0.6882	0.6754	18974
weighted avg	0.7605	0.7483	0.7436	18974	weighted avg	0.7399	0.6970	0.6786	18974
XGBoost									

Table 4: Bright vs Background Classification Report

Table 4 compares the three models on Stage-1 (Bright vs Background) and shows a clear trade-off. Logistic Regression is the most *recall-oriented* for bright lesions: it catches 85.4% of bright patches (recall), with bright precision 0.732, background recall 0.710, and the highest accuracy (0.779) and balanced accuracy/AUC (paper) (0.782) of the three—good for screening but with more false alarms on background. Random Forest flips the trade-off: background recall is very high (0.920), so false positives are low, and bright precision is strong (0.841), but bright recall drops to 0.456 (it misses many lesions) and overall accuracy falls to 0.697. XGBoost sits in between: background recall 0.870, bright recall 0.617, bright precision 0.815, and accuracy 0.748, giving a more balanced profile. In short, LR is best if *finding* bright lesions is the priority; RF is best if *avoiding false positives* matters most; XGB provides a middle ground—its threshold can be nudged to trade a little specificity for extra sensitivity depending on the deployment goal.

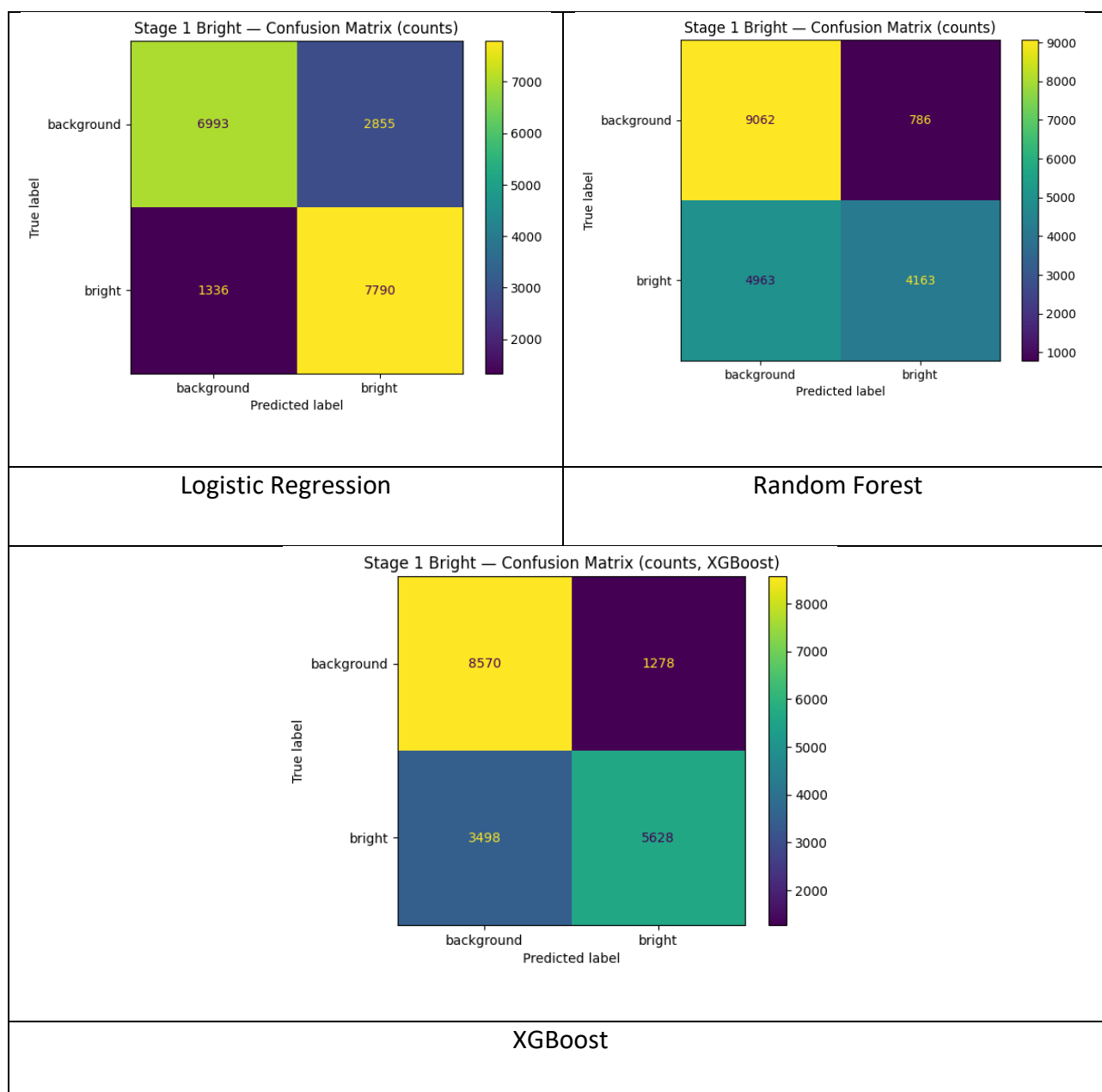


Table 5: Bright vs Background Confusion Matrix

The three confusion matrices show the trade-off clearly. Logistic Regression finds most bright lesions (TP 7,790; FN 1,336) but raises more false alarms on background (FP 2,855; TN 6,993)—that’s sensitivity ≈ 0.854 and specificity ≈ 0.710 . Random Forest is very conservative: it has the fewest false positives (FP 786; TN 9,062, specificity ≈ 0.920) but misses many lesions (FN 4,963; TP 4,163, sensitivity ≈ 0.456). XGBoost sits in between (TP 5,628; FN 3,498; FP 1,278; TN 8,570) with sensitivity ≈ 0.617 and specificity ≈ 0.870 . In short, choose LR if catching as many bright lesions as possible is the priority, RF if minimizing false positives matters most, and XGB as a balanced compromise (and you can nudge any model’s threshold to shift this trade-off).

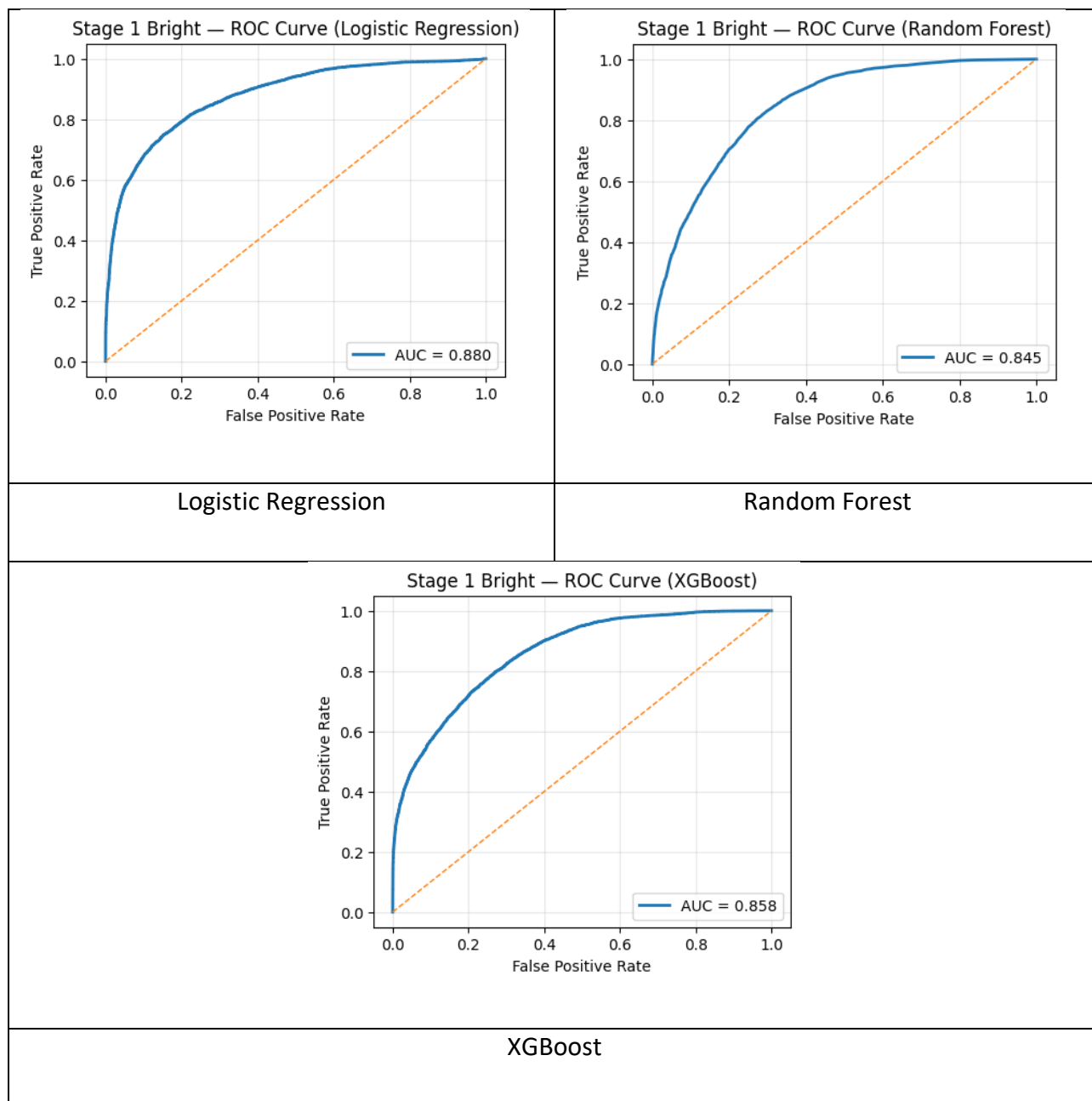


Table 6: Bright vs Background AUC-ROC Curve

All three curves sit well above the diagonal, so the models separate bright lesions from background reliably. Logistic Regression has the best ranking ability ($AUC \approx 0.88$), XGBoost is close behind (≈ 0.86), and Random Forest is slightly lower (≈ 0.85). The steep rise near the origin means each model can achieve high true-positive rates while keeping false positives modest if the threshold is chosen appropriately. This also explains the earlier trade-offs: RF's low recall at the 0.5 cutoff is a threshold choice, not a lack of separability—the ROC shows it could recover sensitivity by moving the cutoff. In short, LR provides the strongest overall ranking, XGB is a balanced second, and RF trails slightly but remains usable with threshold tuning.

Model	Sensitivity (bright)	Specificity (background)	Balanced Accuracy	Accuracy	ROC AUC	PR AUC (bright)	TP	FP	TN	FN
LR	0.8536	0.7101	0.7818	0.7791	0.8802	0.8817	7790	2855	6993	1336
XGB	0.6167	0.8702	0.7435	0.7483	0.8584	0.855	5628	1278	8570	3498
RF	0.4562	0.9202	0.6882	0.697	0.8448	0.8169	4163	786	9062	4963

Table 7: Bright vs Background Model Comparison

Table 7 (Bright vs Background) in one line: Logistic Regression is the best at finding bright lesions (Sensitivity 0.8536, TP=7,790) but let's more background slip through as false alarms (FP=2,855; Specificity 0.7101). Random Forest is the strictest on background (Specificity 0.9202, FP=786; TN=9,062) but misses many lesions (Sensitivity 0.4562, FN=4,963). XGBoost sits in between (Sensitivity 0.6167, Specificity 0.8702) and delivers a balanced Accuracy (0.7483) with solid ranking ability (ROC-AUC 0.858; PR-AUC 0.855). Overall: choose LR if screening and recall are the priority, choose RF if minimizing false positives is critical, and use XGB for a balanced trade-off—and note that small threshold shifts can move any model toward more sensitivity or more specificity without retraining.

4.4 Stage 1: Red vs Background

Classification report:					Classification report:				
	precision	recall	f1-score	support		precision	recall	f1-score	support
background	0.7386	0.6485	0.6906	9848	background	0.7102	0.7717	0.7397	9848
red	0.6266	0.7199	0.6700	8069	red	0.6885	0.6157	0.6500	8069
accuracy			0.6806	17917	accuracy			0.7015	17917
macro avg	0.6826	0.6842	0.6803	17917	macro avg	0.6993	0.6937	0.6949	17917
weighted avg	0.6882	0.6806	0.6813	17917	weighted avg	0.7004	0.7015	0.6993	17917
Logistic Regression					Random Forest				
Classification report:					Classification report:				
	precision	recall	f1-score	support		precision	recall	f1-score	support
background	0.7326	0.7465	0.7395	9848	background	0.7326	0.7465	0.7395	9848
red	0.6833	0.6674	0.6752	8069	red	0.6833	0.6674	0.6752	8069
accuracy			0.7109	17917	accuracy			0.7109	17917
macro avg	0.7079	0.7070	0.7074	17917	macro avg	0.7079	0.7070	0.7074	17917
weighted avg	0.7104	0.7109	0.7106	17917	weighted avg	0.7104	0.7109	0.7106	17917
XGBoost									

Table 8: Red vs Background Classification Report

The class-wise reports show that red-lesion detection is harder than bright: scores are lower across all models. Logistic Regression prioritizes catching red lesions (red recall ≈ 0.72) but is weaker at rejecting background (background recall ≈ 0.65). Random Forest does the opposite—strong background control (background recall ≈ 0.77) but it misses more red lesions (red recall ≈ 0.62). XGBoost is the most balanced, with background recall ≈ 0.75 , red recall ≈ 0.67 , and the best overall accuracy (~~0.71~~) and ~~macro/weighted F1~~ (0.70–0.71). In practice: choose LR (or XGB with a lower threshold) when not missing red lesions is the priority; choose RF when minimizing false positives on background matters more; otherwise XGB is a sensible default.

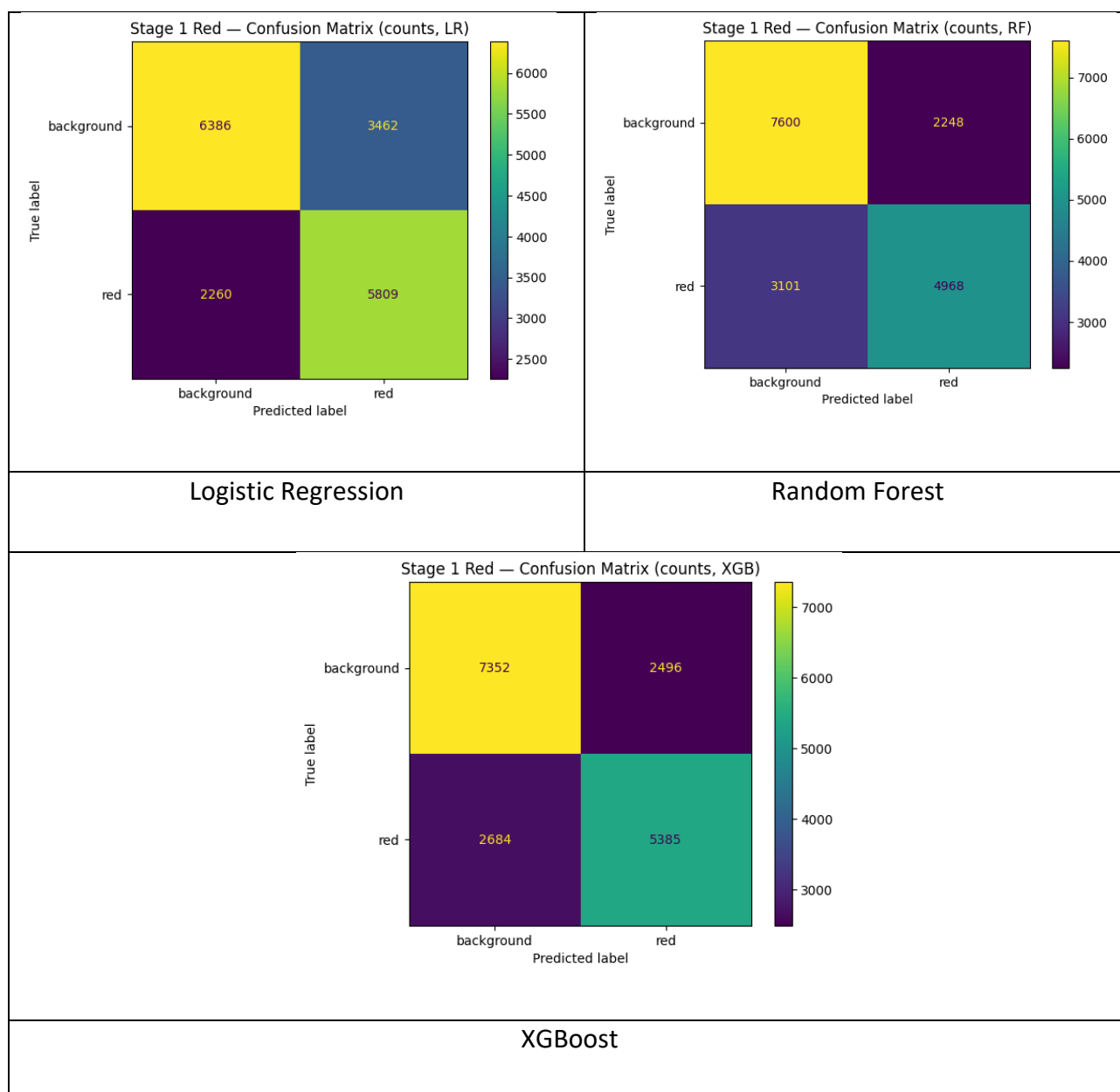


Table 9: Red vs Background Confusion Matrix

The three confusion matrices for Stage-1 Red vs Background tell the same story as the reports. Logistic Regression is recall-oriented for red lesions: it correctly detects 5,809 reds and misses 2,260 (sensitivity ≈ 0.72), but it also raises many false alarms on background (3,462 FPs; specificity ≈ 0.65). Random Forest is much stricter on background (7,600 TNs, 2,248 FPs; specificity ≈ 0.77), yet it misses more red lesions (3,101 FNs; sensitivity ≈ 0.62). XGBoost sits between the two, with 5,385 TP, 2,684 FN, 7,352 TN and 2,496 FP (sensitivity ≈ 0.67 , specificity ≈ 0.75), giving the best overall accuracy of the three. In short: LR catches the reddest lesions but with more false positives; RF minimizes false positives but misses more lesions; XGB offers the most balanced trade-off. Adjusting the decision threshold can nudge any model toward higher sensitivity or higher specificity depending on the deployment goal.

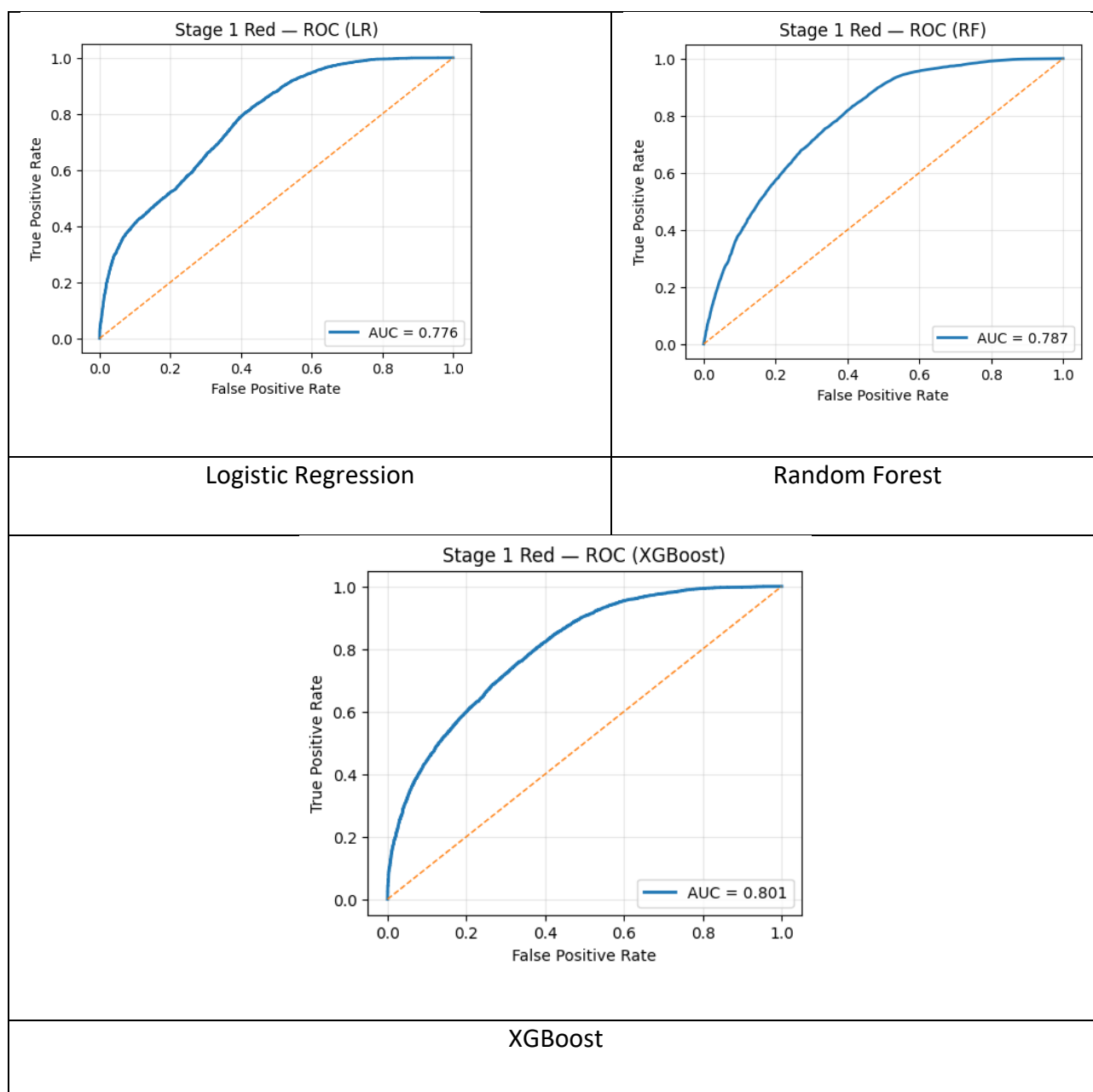


Table 10: Red vs Background AUC-ROC Curve

The ROC curves for Stage-1 Red vs Background sit above the random diagonal, confirming all three models can separate red lesions from background to a useful degree, but less cleanly than for bright lesions. XGBoost gives the best ranking (AUC ≈ 0.801), Random Forest is close (≈ 0.787), and Logistic Regression trails slightly (≈ 0.776). The curves rise steeply near the origin, so each model can achieve reasonable sensitivity at low false-positive rates if the threshold is chosen carefully; however, pushing to very high sensitivity requires accepting a faster increase in false positives—evidence of stronger overlap between red lesions and background/vessels. In short, red detection is harder than bright (lower AUCs overall), XGBoost is the safest default for ranking, and the operating point can be shifted toward sensitivity or specificity via threshold tuning depending on screening vs. precision priorities.

Model	Sensitivity (red)	Specificity (background)	Balanced Accuracy	Accuracy	ROC AUC	PR AUC (red)	TP	FP	TN	FN	AUC (paper def)
LR	0.7199	0.6485	0.6842	0.6806	0.7763	0.7359	5809	3462	6386	2260	0.6842
RF	0.6157	0.7717	0.6937	0.7015	0.7866	0.7167	4968	2248	7600	3101	0.6937
XGB	0.6674	0.7465	0.707	0.7109	0.8009	0.7622	5385	2496	7352	2684	0.707

Table 11: Red vs Background Model Comparison

Performance is lower than in the bright task, confirming red lesions are harder. Logistic Regression is most sensitivity-oriented (0.7199; TP=5,809) but has the lowest specificity (0.6485; FP=3,462), yielding the lowest accuracy (0.6806) and AUC (paper)=0.684. Random Forest flips the trade-off with the highest specificity (0.7717; FP=2,248) but the lowest sensitivity (0.6157; FN=3,101), giving accuracy 0.7015 and AUC (paper)=0.694. XGBoost is the most balanced: sensitivity 0.6674 and specificity 0.7465 lead to the best accuracy (0.7109), the highest ROC-AUC (0.8090), the strongest PR-AUC (0.7622), and the top AUC (paper)=0.707. In practice: pick LR (or a lower XGB threshold) when missing red lesions is unacceptable, choose RF when minimizing false positives matters most, and use XGB as the default balanced model—its threshold can be nudged toward either goal without retraining.

4.5 Stage 2: EX vs CWS

Classification report:					Classification report:				
	precision	recall	f1-score	support		precision	recall	f1-score	support
CWS	0.2660	0.8799	0.4085	783	CWS	0.1578	0.3691	0.2210	783
EX	0.9863	0.7801	0.8712	8646	EX	0.9350	0.8215	0.8746	8646
accuracy			0.7884	9429	accuracy			0.7840	9429
macro avg	0.6261	0.8300	0.6399	9429	macro avg	0.5464	0.5953	0.5478	9429
weighted avg	0.9264	0.7884	0.8327	9429	weighted avg	0.8704	0.7840	0.8203	9429
Logistic Regression					Random Forest				
Classification report:					Classification report:				
	precision	recall	f1-score	support		precision	recall	f1-score	support
CWS	0.1732	0.4534	0.2506	783	CWS	0.1732	0.4534	0.2506	783
EX	0.9420	0.8040	0.8675	8646	EX	0.9420	0.8040	0.8675	8646
accuracy			0.7748	9429	accuracy			0.7748	9429
macro avg	0.5576	0.6287	0.5591	9429	macro avg	0.5576	0.6287	0.5591	9429
weighted avg	0.8782	0.7748	0.8163	9429	weighted avg	0.8782	0.7748	0.8163	9429
XGBoost									

Table 12: EX vs CWS Classification Report

All three models land around 0.77–0.79 accuracy, but they trade CWS vs EX performance differently. Logistic Regression catches most CWS (recall ≈ 0.88) yet its CWS precision is low (~ 0.27), meaning many EX are mistakenly called CWS; at the same time, it keeps very high EX precision (~ 0.99) with moderate EX recall (~ 0.78). Random Forest is more conservative on CWS (recall ≈ 0.37) while staying strong on EX (precision ≈ 0.94 , recall ≈ 0.82). XGBoost offers the best balance: it raises CWS recall (≈ 0.45) compared to RF while keeping EX performance high (precision ≈ 0.94 , recall ≈ 0.84). The very small CWS class ($\sim 8\%$ of test) explains the low CWS precision across models; if the goal is to not miss CWS, LR (or lowering the XGB threshold) is preferable, whereas if preventing false CWS alarms is key, RF/XGB are safer.

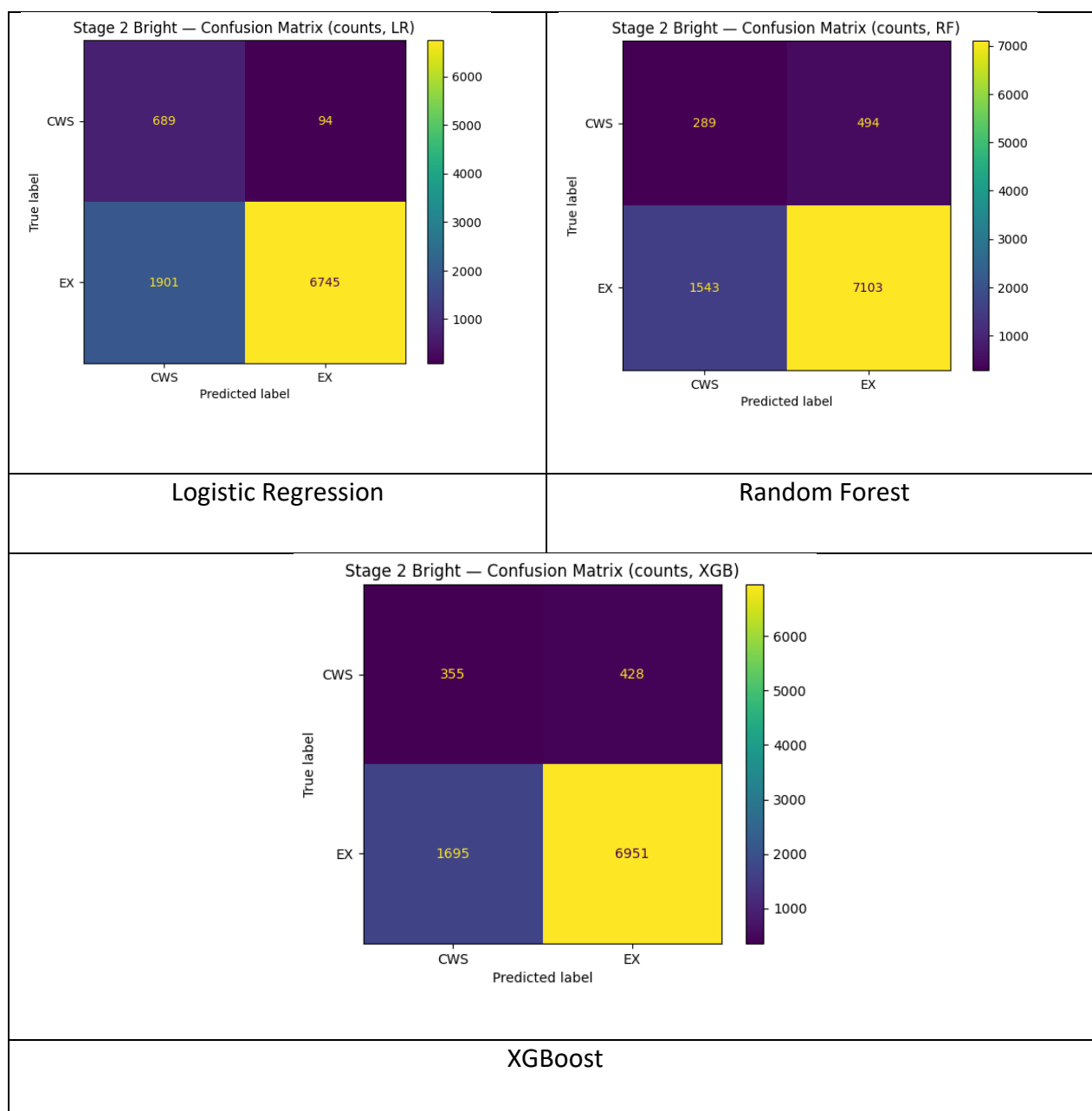


Table 13: EX vs CWS Confusion Matrix

The EX vs CWS confusion matrices show how each model trades CWS vs EX performance. Logistic Regression catches most CWS (689 CWS→CWS, 94 CWS→EX; CWS recall ≈ 0.88) but often over-calls CWS on EX patches (1,901 EX→CWS), which explains its low CWS precision and strong EX precision (EX: 6,745 EX→EX, 1,901 EX→CWS; recall ≈ 0.78). Random Forest is much stricter: it labels far fewer EX as CWS (1,543 EX→CWS) and gets very high EX recall (7,103 EX→EX; ≈ 0.82), but it misses many CWS (289 CWS→CWS, 494 CWS→EX; CWS recall ≈ 0.37). XGBoost lands in the middle—compared to RF it recovers more CWS (355 CWS→CWS, 428 CWS→EX; recall ≈ 0.45) while keeping EX strong (6,951 EX→EX, 1,695 EX→CWS; recall ≈ 0.80). In short: LR favors not missing CWS (at the cost of extra false CWS on EX), RF favors protecting EX from false CWS, and XGB offers the best balance between the two.

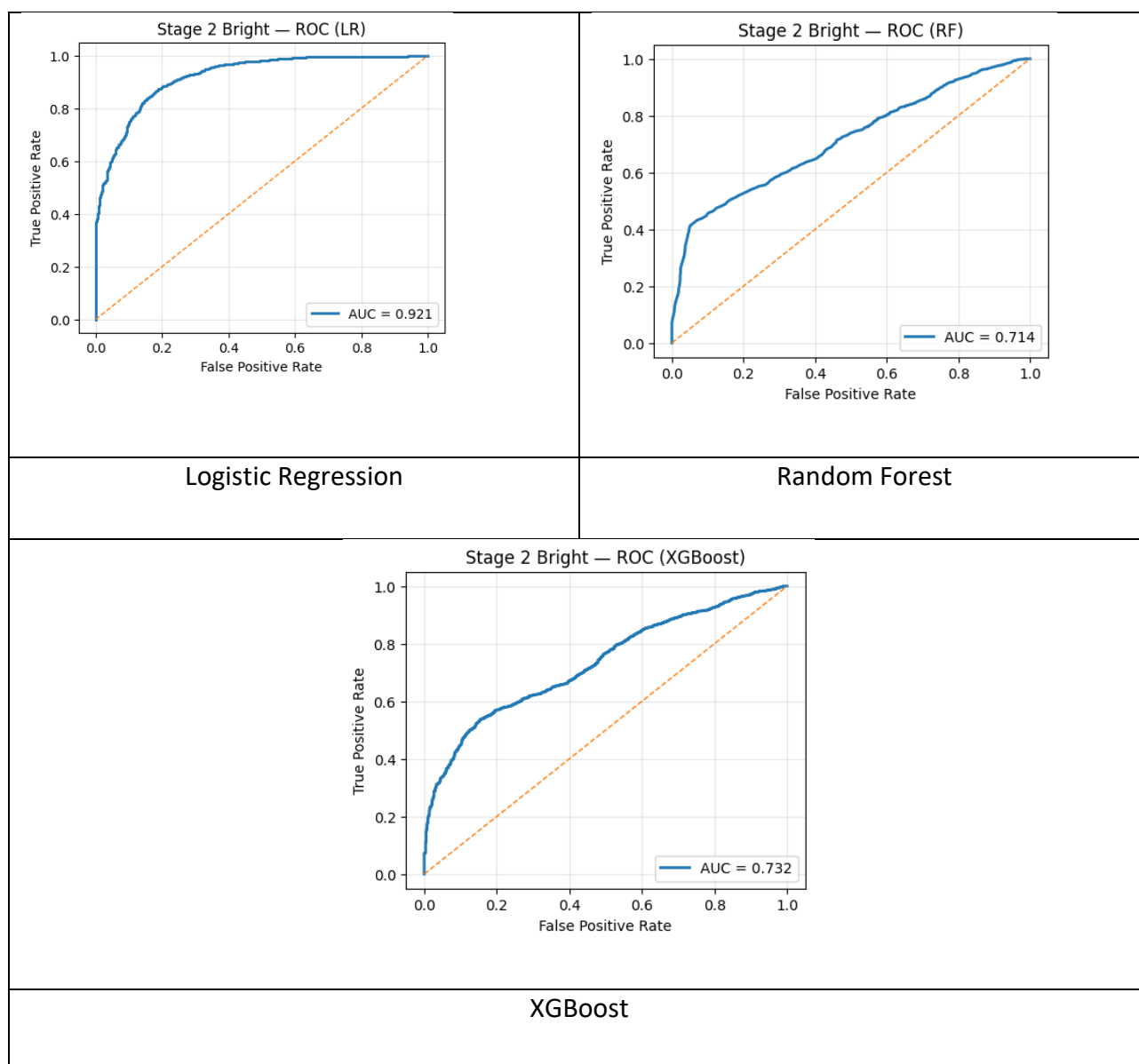


Table 14: EX vs CWS AUC-ROC Curve

The curves show Logistic Regression clearly separates EX from CWS the best (AUC ≈ 0.92), with a steep rise near the origin—so it can reach high true-positive rates for EX while keeping false positives relatively low by choosing a suitable threshold. XGBoost (AUC ≈ 0.73) and Random Forest (AUC ≈ 0.71) separate the classes less cleanly, which matches their weaker trade-offs seen in the confusion matrices. The gap suggests that, for these hand-crafted features, a simple linear boundary (LR) captures the EX–CWS differences better than the tree models. Given the strong class imbalance (CWS is small), the operating threshold still matters: moving it can recover more CWS or protect EX, but LR remains the best overall ranking model for this stage.

Model	Sensitivity	Specificity	Balanced Accuracy	Accuracy	ROC AUC	PR AUC	TP	FP	TN	FN
LR	0.7801	0.8799	0.83	0.7884	0.9212	0.9916	6745	94	689	1901
XGB	0.804	0.4534	0.6287	0.7748	0.7319	0.9685	6951	428	355	1695
RF	0.8215	0.3691	0.5953	0.784	0.7138	0.9661	7103	494	289	1543

Table 15: EX vs CWS Model Comparison

All three models are strong on this stage, but Logistic Regression (LR) is the most balanced: Sensitivity (EX) = 0.780, Specificity (CWS) = 0.880, Balanced Acc = 0.830, Accuracy = 0.788, and the best ROC-AUC = 0.921. Its errors are few on CWS (FP = 94; TN = 689), so it protects the minority class well while still retrieving most EX (TP = 6,745; FN = 1,901). Random Forest (RF) and XGBoost (XGB) push sensitivity a bit higher (0.822 and 0.804) but at a large cost to specificity (0.369 and 0.453), meaning many CWS are mislabelled as EX (FPs: 494 RF, 428 XGB). That's why their Balanced Accuracy drops (0.596 RF, 0.629 XGB) even though overall accuracy stays similar (majority class is EX). The very high PR-AUC values (≈ 0.97 – 0.99) reflect that EX is the positive and dominant class. Bottom line: choose LR for the best-balanced performance and CWS protection; choose RF/XGB only if maximizing EX recall is the priority, and consider threshold tuning to recover some CWS specificity.

4.6 Stage 2: MA vs HM

Classification report: precision recall f1-score support					Classification report: precision recall f1-score support				
HM	0.8858	0.7282	0.7993	7480	HM	0.8551	0.9378	0.8945	7480
MA	0.2621	0.5070	0.3455	1424	MA	0.3357	0.1650	0.2213	1424
accuracy			0.6928	8904	accuracy			0.8142	8904
macro avg	0.5740	0.6176	0.5724	8904	macro avg	0.5954	0.5514	0.5579	8904
weighted avg	0.7861	0.6928	0.7268	8904	weighted avg	0.7720	0.8142	0.7869	8904
Logistic Regression					Random Forest				
Classification report: precision recall f1-score support									
HM	0.8591	0.8422	0.8506	7480					
MA	0.2489	0.2746	0.2611	1424					
accuracy			0.7515	8904					
macro avg	0.5540	0.5584	0.5559	8904					
weighted avg	0.7615	0.7515	0.7563	8904					
XGBoost									

Table 16: MA vs HM Classification Report

These reports reflect the strong class imbalance (HM \gg MA). Random Forest achieves the highest overall accuracy (≈ 0.81) by being very conservative: it classifies HM extremely well (HM recall ≈ 0.94 ; F1 ≈ 0.89) but misses most MAs (MA recall ≈ 0.17 ; F1 ≈ 0.22). Logistic Regression shifts the trade-off toward finding MAs: it more than doubles MA recall (≈ 0.51)—the best of the three—but with low MA precision (≈ 0.26) and lower accuracy (≈ 0.69) because many HMs are over-called as MA. XGBoost sits between them (HM recall ≈ 0.84 ; MA recall ≈ 0.27 ; accuracy ≈ 0.76). In short: pick LR (or lower the threshold on XGB) if not missing microaneurysms is the priority; pick RF when minimizing false MA alarms matters most; XGB provides a balanced middle ground. Further gains for MA could come from adjusting the train rebalance ratio/thresholds or stronger class-weighted learning.

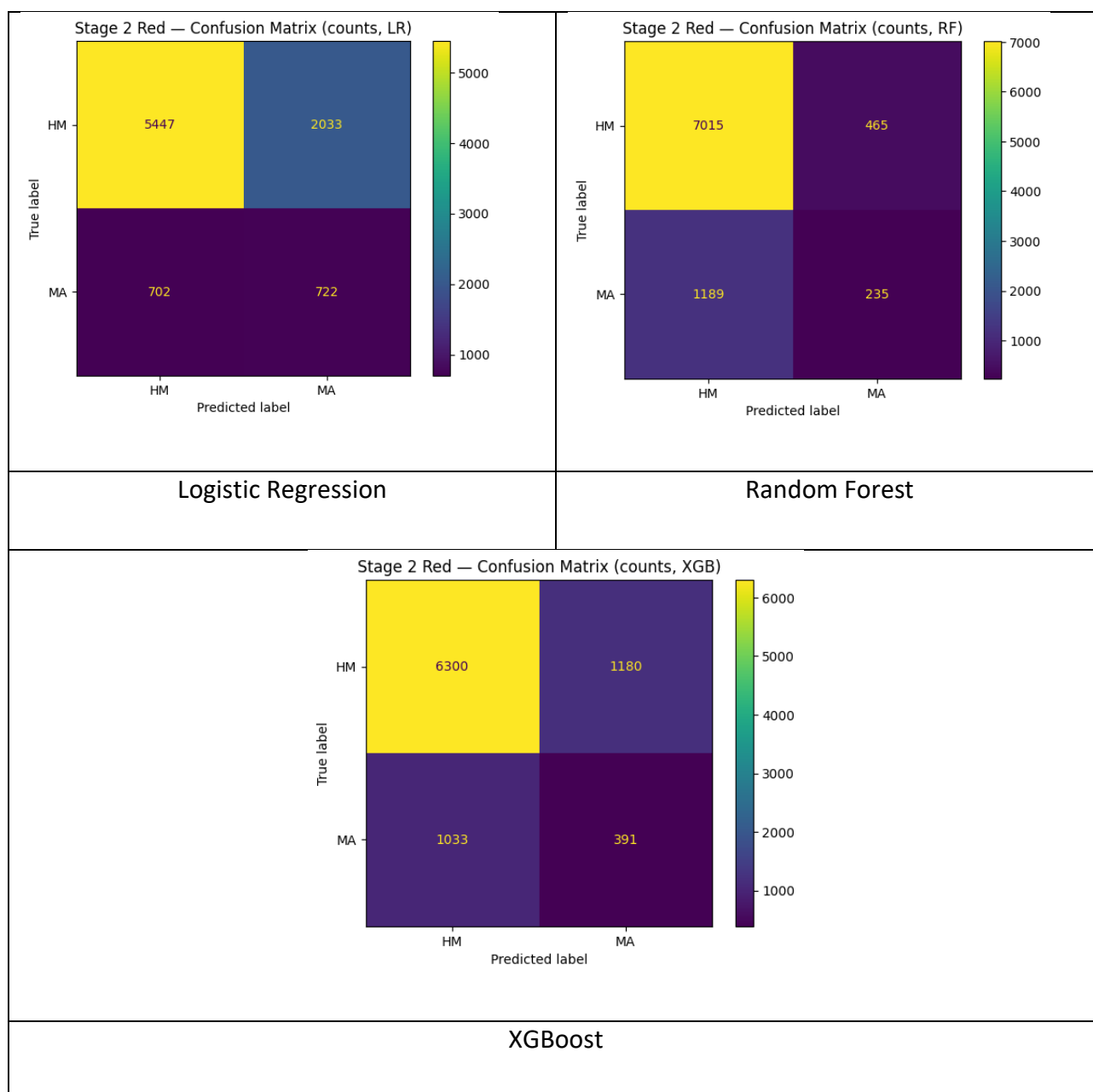


Table 17: MA vs HM Confusion Matrix

The counts show very different trade-offs. Logistic Regression finds the most microaneurysms: TP=722, FN=702 → MA recall ≈ 0.51 , but it also over-calls MA on HM (FP=2,033, HM specificity ≈ 0.73). Random Forest is the strictest: TN=7,015, FP=465 → HM specificity ≈ 0.94 , yet it misses most MAs (TP=235, FN=1,189, MA recall ≈ 0.17). XGBoost sits between them (TP=391, FN=1,033, TN=6,300, FP=1,180), giving MA recall ≈ 0.27 and HM specificity ≈ 0.84 . Because HM greatly outnumbers MA, RF's conservatism yields higher overall accuracy but poorer MA detection. If the clinical goal is to not miss MAs, prefer LR (or lower XGB's threshold / increase MA weight); if the goal is to minimize false MA alarms, RF is safer.

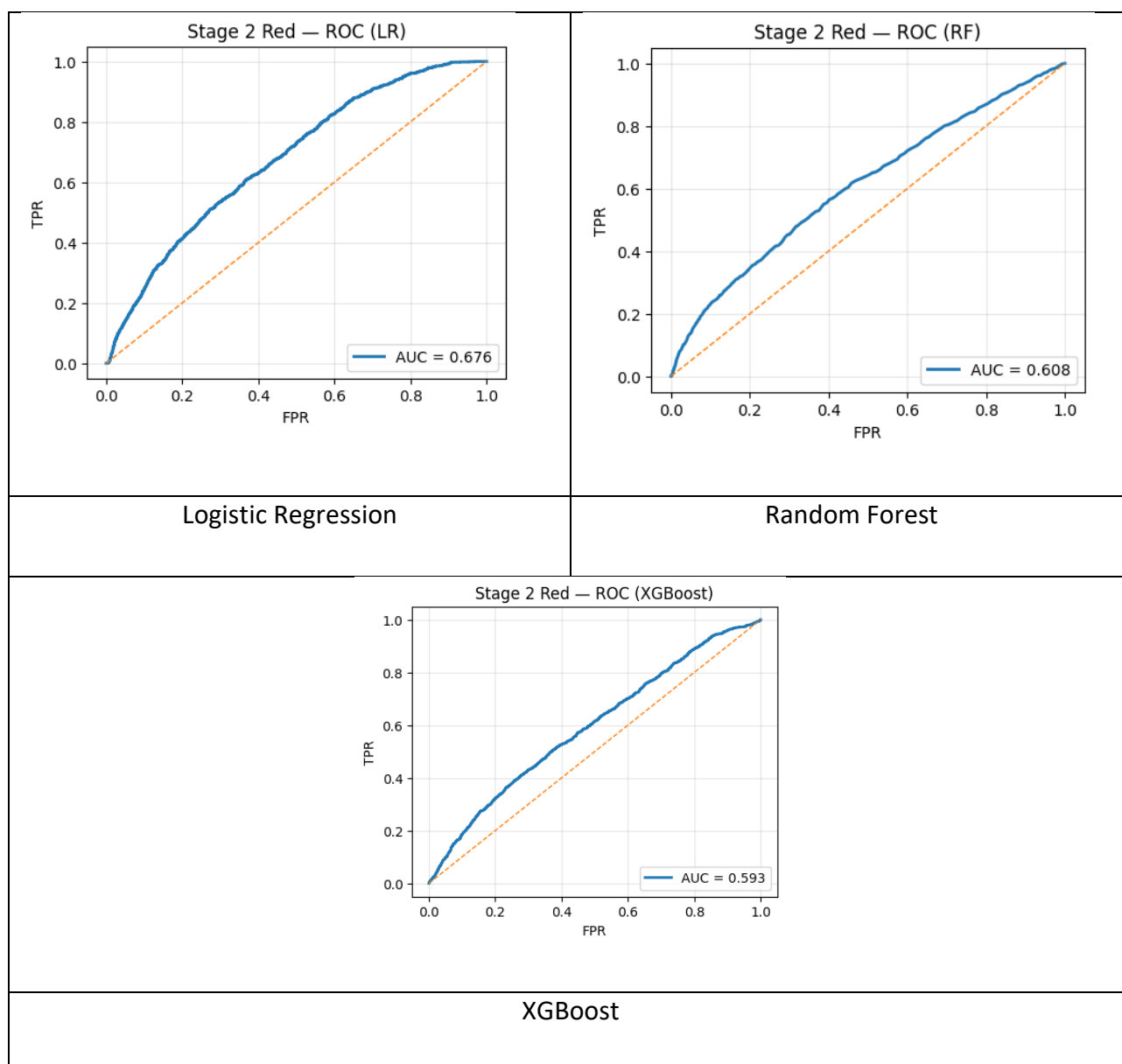


Table 18: MA vs HM AUC-ROC Curve

All three curves lie only modestly above the diagonal, confirming that MA vs HM is the hardest task. Logistic Regression ranks positives best (AUC ≈ 0.676), while Random Forest (≈ 0.608) and XGBoost (≈ 0.593) show weaker separation. The gradual rise of the curves means that achieving high MA recall quickly increases false positives—evidence of substantial feature overlap between MA and HM. In practice, LR is the safest default for ranking MAs ahead of HMs; moving the threshold downward can lift MA sensitivity, whereas RF/XGB behave more conservatively and keep more HMs correctly negative at the cost of missing MAs. Overall, the ROC shapes match the confusion matrices: detecting MAs is challenging, and performance is dominated by the sensitivity–specificity trade-off.

Model	Sensitivity (MA)	Specificity (HM)	Balanced Accuracy	Accuracy	ROC AUC	PR AUC (MA)	TP	FP	TN	FN	AUC (paper def)
LR	0.507	0.7282	0.6176	0.6928	0.6764	0.263	722	2033	5447	702	0.6176
RF	0.165	0.9378	0.5514	0.8142	0.6084	0.2421	235	465	7015	1189	0.5514
XGB	0.2746	0.8422	0.5584	0.7515	0.5926	0.2134	391	1180	6300	1033	0.5584

Table 19: MA vs HM Model Comparison

This is the toughest stage, and the numbers show it. Logistic Regression gives the best-balanced performance: Sensitivity (MA)=0.507, Specificity (HM)=0.728, Balanced Acc/AUC(paper)=0.618, and the top ROC-AUC=0.676—so it finds about half of the MAs while keeping HM errors moderate (TP=722, FN=702). Random Forest reports the highest overall accuracy (0.814), but that’s driven by the majority HM class: Specificity=0.938 with very low MA sensitivity=0.165 (TP=235, FN=1,189), which is risky if missing MAs is unacceptable. XGBoost sits between them (Sensitivity=0.275, Specificity=0.842, Balanced Acc=0.558, ROC-AUC=0.593). The uniformly low PR-AUC (0.21–0.26) reflects how rare MAs are—precision drops quickly when chasing recall. In practice: choose LR (or a lower threshold on XGB) when MA detection is the priority; choose RF when minimizing false MA alarms matters most; otherwise XGB offers a compromise but still trails LR on balanced accuracy.

4.7 Strengths, Limitations, Threats to Validity

Strengths

Leakage-safe evaluation: splits are grouped by *image*, so patches from the same eye never appear in both train and test.

- **Clear, comparable metrics:** reports the paper's AUC alongside Accuracy, Sensitivity, Specificity, ROC-AUC, and PR-AUC.
- **Transparent imbalance handling:** background under-sampling for Stage-1 and **train-only** rebalance for Stage-2; test sets kept natural.
- **Two-stage design:** mirrors clinical reasoning and improves separability in Stage-2.
- **Reproducibility:** fixed random seed, minimal preprocessing, and results tables saved per model/stage.
- **Use of extracted features:** the same engineered intensity/shape/distance features drive all experiments, ensuring consistency.

Limitations

- **No full hyper-parameter optimisation:** only sensible defaults were used; performance might improve with systematic tuning or calibration.
- **Single dataset:** trained/evaluated on DIARETDB1 only; no external validation.
- **Under-sampling cost:** trimming background may discard useful variability; Stage-2 remains imbalanced for rare classes (e.g., MA, CWS).
- **Cascade error propagation:** Stage-2 performance in deployment depends on Stage-1; errors in Stage-1 would reduce end-to-end recall.
- **Per-patch evaluation:** metrics are patch-level; per-image or per-patient utility is not measured.
- **Model scope:** compared LR/RF/XGB only; not an exhaustive comparison (e.g., SVM/k-NN/GMM from the paper or modern CNNs).

- **No user/clinical study:** no usability testing or reader study to assess practical workflow impact.

Threats to validity

- **Sampling bias:** manual background under-sampling and train-only rebalancing change class priors vs. real clinics; thresholds may need retuning in the wild.
- **Measurement bias / feature design:** engineered features may miss texture cues; results depend on the chosen feature set and preprocessing.
- **Label issues:** mapping HE→HM and any annotation noise could affect Stage-2 red results.
- **Domain shift:** performance may drop on other cameras, populations, or acquisition settings.
- **Statistical uncertainty:** single 80/20 split (no group k-fold); confidence intervals not reported.

4.8 Summary

This work set out to test a two-stage, feature-based pipeline for diabetic-retinopathy lesions and to compare three tabular models (Logistic Regression, Random Forest, XGBoost). Using leakage-safe splits and the extracted intensity/shape/distance features, Stage-1 showed that bright vs background is easier than red vs background: Logistic Regression maximized bright-lesion sensitivity, Random Forest maximized background specificity, and XGBoost offered the most balanced trade-off. For red vs background, XGBoost provided the best overall balance and accuracy, while LR kept the highest red recall. In Stage-2 subtyping, EX vs CWS was separated best by Logistic Regression (highest ROC-AUC and balanced accuracy), whereas MA vs HM was the hardest task; LR achieved the best-balanced accuracy by raising MA recall, RF achieved the highest overall accuracy by being conservative on the majority HM class, and XGB lay between them. Across stages, simple threshold tuning shifted models toward higher sensitivity (screening) or higher specificity (fewer false alarms) without retraining. Overall, the results confirm the viability of a lightweight, two-stage classical approach on engineered features, while highlighting remaining challenges for rare classes (e.g., MA, CWS) and motivating future work on calibration, grouped cross-validation, and external validation.

Chapter 5 Conclusion

This project set out to design and evaluate a **two-stage, feature-based Automatic Diabetic Retinopathy Grading System (ADRGs)** on DIARETDB1. The pipeline follows a clinically intuitive flow: pre-processing, background suppression (vessels/optic disc), candidate generation, **30-D hybrid feature extraction, hierarchical lesion classification** (Stage-1: lesion vs background; Stage-2: subtype), and rule-based image-level grading. Evaluation used leakage-safe splits by image and the paper's metrics (Sensitivity, Specificity, Accuracy, **AUC = (Sens+Spec)/2**), alongside ROC-AUC/PR-AUC.

Extent to which objectives were met.

1. **Image enhancement** was achieved with a compact pipeline (resize, green channel, CLAHE, denoise, morphology, normalization) that improved lesion visibility.
2. **Background suppression** successfully masked major/minor vessels and the optic disc, reducing false positives and narrowing the search space.
3. **Hybrid feature extraction** produced a consistent 30-feature descriptor per candidate (appearance, shape, and anatomical distances).
4. **Hierarchical classification** was implemented and validated: Stage-1 (bright vs background, red vs background) and Stage-2 (EX vs CWS; MA vs HM).
5. **Benchmark validation** on DIARETDB1 reported all required metrics with clear tables, ROC curves, and confusion matrices.

Key findings.

- **Stage-1:** *Bright vs background* is easier than *red vs background*. Logistic Regression (LR) achieved the highest bright-lesion sensitivity (screening-friendly), Random Forest (RF) maximized background specificity, and XGBoost (XGB) provided a balanced trade-off. For *red vs background*, overall scores were lower; **XGB** offered the best balance/accuracy while **LR** preserved higher red recall.
- **Stage-2:** *EX vs CWS* was separated best by **LR** (highest balanced accuracy and ROC-AUC), indicating that a linear boundary aligns well with the engineered features for these classes.

MA vs HM was the hardest task: **LR** achieved the best-balanced accuracy by increasing MA recall, **RF** attained the highest overall accuracy by being conservative on the majority HM class, and **XGB** lay between them.

- **Operating point matters.** Threshold choice strongly influenced Sensitivity/Specificity. Simple tuning (e.g., Youden's J) shifted models toward screening (higher sensitivity) or precision (fewer false alarms) **without retraining**.

Contributions and significance:

The project delivers a reproducible, leakage-safe classical baseline for DR lesion analysis; a clean feature table and code for all stages; transparent handling of imbalance; and clear, clinician-interpretable outputs (confusion matrices, ROC/PR curves). The results confirm the viability of a **lightweight, explainable alternative** to data-hungry deep models—useful where datasets or compute are limited—and highlight where future gains are most needed (rare classes, calibration, external validation).

5.1 Future Work

External validity and robustness

- **Test on additional datasets** (e.g., Messidor, IDRiD) to quantify domain shift across cameras/populations; report confidence intervals via grouped k-fold or bootstrap.
- **Per-image aggregation**: move from patch metrics to clinically meaningful **image- or eye-level** decisions (e.g., multiple-instance learning or spatial clustering of patch outputs).

Learning under imbalance

- Replace undersampling with **class-balanced or focal losses**, or cost-sensitive training; couple this with **thresholds tied to clinical costs** (missed lesion vs false alarm).
- Explore **rebalancing strategies per stage** and **calibration** (Platt/isotonic) so probability scores map reliably to operating points.

Model and feature improvements

- Run **lightweight HPO** (e.g., Optuna) for LR/RF/XGB; include **SVM/k-NN/GMM** baselines to mirror the source paper and round out comparisons.
- Add richer **texture/GLCM/gradient** features and re-check importance; use **feature selection** (L1-LR, mutual information) to reduce redundancy and improve generalisation.
- Trial **modern representation learning** (CNN/ViT or self-supervised patch embeddings) with a simple linear head, using the current pipeline as a strong classical baseline.

End-to-end behaviour and usability

- Measure **cascade effects** explicitly (Stage-1 → Stage-2) and study **gating strategies** (confidence filters) to control error propagation.
- Build a **reviewer UI** and run a small **think-aloud** or reader study with graders to determine preferred thresholds and workload savings.

5.2 Reflection

What worked well

The project's biggest strength was **evaluation discipline**: grouping by image prevented leakage, yielding credible numbers. The modular pipeline made debugging and iteration straightforward. A second strength was **clarity in metrics and trade-offs**—reporting both the paper's AUC and modern ROC/PR views made the imbalance effects and threshold choices transparent. Finally, the results highlighted an important lesson: **simple models can be powerful on good features**—LR outperformed the tree models for EX vs CWS.

What was challenging

Time and scope constraints limited **hyperparameter optimisation** and **external validation**. Red-family tasks—especially **MA vs HM**—exposed the limits of the current hand-crafted features and undersampling strategy. Tooling hiccups (e.g., XGBoost early-stopping incompatibilities) required fallback to simpler training loops. And while the pipeline outputs are clear, the work stopped short of a **per-image** decision layer and **usability** testing with clinicians.

What was learned

Three practical insights stand out. First, **leakage control** is non-negotiable in medical imaging; splitting by image meaningfully changes results. Second, **class imbalance** influences not only training but also storytelling—accuracy alone can mislead, and **operating thresholds** are often the real lever for clinical utility. Third, the **feature–model fit** matters: when features encode the right cues (e.g., solidity/extent/distance for bright lesions), linear models can generalise better than expected.

Closing remark

Within the constraints of a single-dataset MSc project, the work delivers a robust, explainable baseline and a clear map of where effort should go next. With calibrated thresholds, imbalance-aware learning, and external validation, the system can evolve from a solid research prototype into a practical screening aid.

References

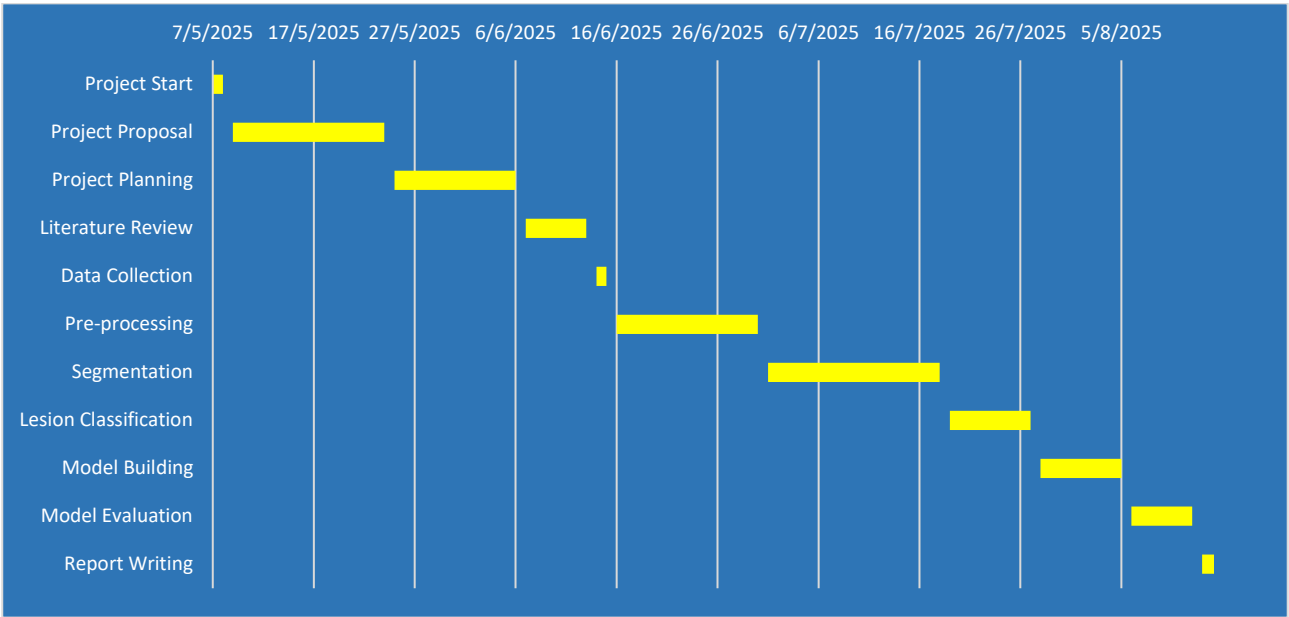
1. Y. A. S. Devi and K. M. Chari, "ADRGs: an automatic diabetic retinopathy grading system through machine learning," 2024.
2. S. Vujosevic et al., "Novel artificial intelligence for diabetic retinopathy and other retinal vascular diseases," *Acta Diabetologica*, 2024.
3. S. Prathibha and R. Prabhavathi, "Advancing diabetic retinopathy diagnosis with fundus photography: A survey," *Intelligent Medicine*, 2024.
4. D. Bhulakshmi et al., "A systematic review on diabetic retinopathy detection and classification using deep learning," *Bioengineering*, 2024.
5. B. Qian et al., "DRAC 2022: A public benchmark for diabetic retinopathy analysis," *Patterns*, 2024.
6. M. Monemian et al., "Red-lesion extraction in retinal fundus images by dynamic iterative separation," *Scientific Reports*, 2021.
7. M. Canayaz, "Classification of diabetic retinopathy with feature selection and optimized hyper-parameters," *Applied Soft Computing*, 2022.
8. W. L. Alyoubi et al., "Diabetic retinopathy fundus image classification and lesion detection," *Sensors*, 2021.
9. M. Driban et al., "Artificial intelligence in chorioretinal pathology through fundus imaging," *Eye and Vision*, 2024.
10. G. Lepetit-Aimon et al., "MAPLES-DR: MESSIDOR Anatomical and Pathological Labels for Explainable Screening," *Scientific Data*, 2024.
11. L. Zago Ribeiro et al., "Automated machine learning for diabetic retinopathy classification from multi-field handheld images," *Scientific Reports*, 2024.
12. [A. Senapati, "Artificial intelligence for diabetic retinopathy detection," *Artificial Intelligence in Medicine*, 2024. (state-of-the-art review)
13. American Academy of Ophthalmology, *Diabetic Retinopathy Preferred Practice Pattern®*, 2024. (screening & treatment guidance)
14. World Health Organization (WHO) Europe, "Promoting diabetic retinopathy screening," 2023. (screening as effective intervention)
15. L. Teo et al., "Global prevalence of diabetic retinopathy and projections," *Ophthalmology*, 2021. (meta-analysis of DR prevalence in diabetics)

Appendices

Appendix A: Project Proposal

https://github.com/sibinshibu/Diabetic-Retinopathy-MSc-Dissertation/tree/main/project_proposal

Appendix B: Project Management



Appendix C: Artefact/Dataset

- Dataset: <https://github.com/sibinshibu/Diabetic-Retinopathy-MSc-Dissertation/tree/main/dataset>
- Code: <https://github.com/sibinshibu/Diabetic-Retinopathy-MSc-Dissertation/tree/main/code>

Appendix D: Screencast

[MSc Project Video](#)