International Journal of Mathematical, Engineering and Management Sciences
Vol. 11, No. 1, 130-147, 2026
https://doi.org/10.33889/IJMEMS.2026.11.1.007

A Harmonized Multi-Source Dataset with Baseline Deep Learning Validation for Staging Diabetic Retinopathy

Mukesh Delu
Department of Mathematics,
Malaviya National Institute of Technology, Jaipur, 302017, Rajasthan, India.
E-mail: 2022rma9055@mnit.ac.in

Priyanka Harjule
Department of Mathematics,
Malaviya National Institute of Technology, Jaipur, 302017, Rajasthan, India.
Corresponding author: Priyanka.maths@mnit.ac.in

Rajesh Kumar
Department of Human Anatomy and Physiology,
University of Johannesburg, Johannesburg, 2006, South Africa.
&
Department of Electrical Engineering,
Malaviya National Institute of Technology, Jaipur, 302017, Rajasthan, India.
E-mail: rkumar@uj.ac.za

Kushal Gajjar
Department of Electronics and Communication Engineering,
Malaviya National Institute of Technology, Jaipur, 302017, Rajasthan, India.
E-mail: 2022uec1893@mnit.ac.in

(Received on August 31, 2025; Revised on November 15, 2025; Accepted on November 25, 2025)

Abstract
Accurate automated grading of diabetic retinopathy (DR) significantly depends on the quality of retinal fundus images. Inferior-quality pictures, resulting from inadequate lighting, motion blur, distortions, or incomplete retinal coverage, may obscure minor lesions and diminish the accuracy of model predictions. This study constructs a harmonized multi-source dataset using a multi-dimensional image quality assessment framework for multi-class DR staging. Retinal images are collected from IDRiD, Messidor-2, SUSTech-SYSU, APTOS 2019, DeepDRiD-v1.1, and Zenodo DR V03 datasets. The proposed pipeline includes preprocessing, image quality assessment using technical quality and medical relevance indicators, dataset-specific statistics, and adaptively thresholded using DR severity-aware percentiles derived from stratified samples with weighting to match diagnostic needs. Baseline deep learning models were trained for three hierarchical DR classification schemes to validate the dataset. Experimental results show that the quality-filtered merging of datasets improves model generalization accuracy by 3-7% compared to the normal merging of datasets. This work provides a benchmark dataset and baseline performance results to facilitate future research in DR staging and medical image classification.

Keywords- Diabetic retinopathy, Label harmonization, Image quality assessment, Retinal fundus images, Deep learning, Baseline validation, Hierarchical classification.

1. Introduction

Diabetes mellitus (DM) is a global public health crisis, affecting about 536.6 million people in the year 2021, and estimated that the number will increase to 783.2 million by 2045 (Sun et al., 2022). DR is a serious microvascular complication of DM. It causes vitreous hemorrhage, diabetic macular edema (DME),

and vision impairment. DR arises due to cytokine-mediated damage to capillaries. It leads to increased vascular permeability, ischemia, and irreversible vision loss (Zhang et al., 2024). It has a global prevalence of 34.6%. This includes proliferative DR (PDR) (6.96%), DME (6.81%), and vision-threatening complications (10.2%) (Yau et al., 2012).

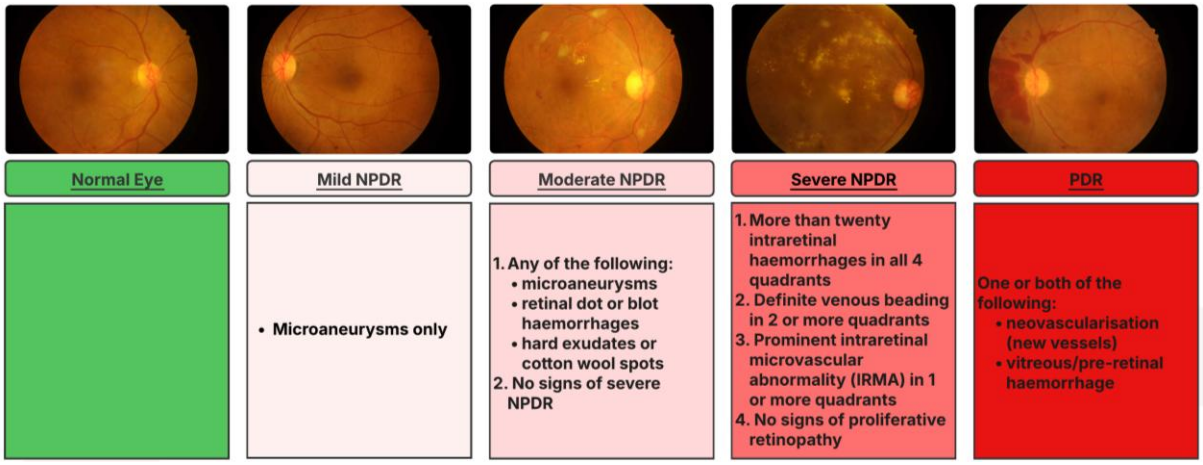

Figure 1. The International Clinical Diabetic Retinopathy (ICDR) severity scale and corresponding features. (Wilkinson et al., 2003).

As shown in **Figure 1,** DR begins with mild non-proliferative DR (NPDR). The mild NPDR is characterized by the presence of microaneurysms. It then progresses to moderate and severe NPDR. This progression is indicated by an increase in microaneurysms, intraretinal hemorrhages, hard exudates, and retinal venous irregularities. Finally, it tends to PDR. PDR is characterized by neovascularization. PDR can result in visual impairment (Rajesh et al., 2023). In the early stages of DR, vascular permeability may go up. Damage to the blood vessels in the retina can cause macular edema, or thickening of the retina. This can occur at any stage of the disease's progression.

Timely identification and diagnosis of DR can be helpful to prevent vision loss. Research by the National Eye Institute (2019) tells that timely intervention for DR decreases the possibility of blindness by 95%. The challenge lies in the fact that DR might be asymptomatic during its initial phases and can complicate diagnosis. Blurred vision and floaters indicate that the disease has progressed to an advanced stage. This significantly reduces the effectiveness of treatment. Current diagnosis methods can identify advanced DR based on clearly defined symptoms. Early-stage DR features are usually overlooked due to their small size and difficulty in separating them from normal variations (Abushawish et al., 2024; Almas et al., 2025).

People with diabetes should get an exam once a year so that they can get a correct diagnosis and find out their status quickly, which will help avoid problems linked to DR (Kim et al., 2025). Traditional methods of diagnosis of DR depend on the manual detection of indicators related to DR by trained ophthalmologists in retinal fundus images. This process takes significant time and work. Also, highly trained ophthalmologists are needed, which isn't always easy to find in developing countries (Taha et al., 2024). Treatment of DR and DME, such as vitrectomy, laser treatment, and intravitreal anti-VEGF injections, is continuously in demand and is highly expensive (Mansour et al., 2020). DR management is not easy to do effectively due to high tool costs, income disparities, and the need for close monitoring. Delays in diagnosis and treatment can also be caused by social, environmental, nutritional, and health issues (Hill-Briggs et al.,

2020). Medical image analysis through AI has become a powerful tool to solve these issues. AI enables the automated extraction of features. It also enhanced the accuracy of diagnosis across various clinical imaging modalities (Gulati et al., 2023, 2024, 2025).

Table 1. Detailed distribution of multistage DR datasets. Normalized Shannon Entropy (NSE) and Variance of Proportions (VP) are also included to check the dataset class balance.

Dataset name	No DR	Mild NPDR	Moderate NPDR	Severe NPDR	PDR	Total	(NSE, VP)	Camera	Location
IDRiD	168	25	168	93	62	516	(0.90, 0.012)	Kowa VX-10 Alpha	India
Messidor-2	1017	270	347	75	53	1744	(0.72, 0.40)	Topcon TRC NW6 + 3CCD	France
Zenodo DR V03	711	6	110	349	261	1437	(0.76, 0.029)	Zeiss Visucam 500	Paraguay
SUSTech-SYSU	631	24	401	87	76	1219	(0.71, 0.037)	Topcon TRC-50DX	China
APTOS 2019	1805	370	999	193	295	3662	(0.80, 0.027)	Standard fundus cameras	India
DeepDRiD-v1.1	914	222	398	354	112	2000	(0.86, 0.019)	Non-mydriatic fundus cameras	China

Deep learning-based methods have shown strong potential for automated DR detection and staging, which reduces dependence on manual screening and subjective clinical evaluation (Butt et al., 2022; Kotiyal and Pathak, 2022; Mutawa et al., 2023; Raghad and Hamad, 2024; Saproo et al., 2024; Shakibania et al., 2024; Almas et al., 2025; Refat et al., 2025; Zafar et al., 2025; Zhang et al., 2025). However, their clinical performance depends a lot on the quality, diversity, and balance of the dataset (Rajesh et al., 2023; Abushawish et al., 2024; Taha et al., 2024). As shown in **Table 1**, publicly available dataset classes vary in distribution, geographical origin, and imaging devices, causing domain shift and acquisition bias that affect generalization in different population groups and clinical conditions (Men et al., 2025). Serious class imbalance, especially in advanced stages such as insufficient representation of PDR, deforms the adaptation of the model and limits sensitivity to visually impaired cases (Rajesh et al., 2023; Abushawish et al., 2024).

The use of different grading systems, inconsistencies in annotations, and inter-observer variability in the dataset decrease the reliability of the model (Riotto et al., 2025). The variety in ages, races, and comorbidities makes it harder for models to be valid on the external dataset. Changes in resolution, lighting, and field of view of the image can also affect model performance. When models use a single source of data, they are more likely to overfit (Kim et al., 2025; Refat et al., 2025). These issues suggest the urgent need for standardized, big, multi-center, and well-annotated datasets (Taha et al., 2024). This type of data enables the development of clinically robust and generalizable DR classification systems. If the challenges mentioned in **Figure 2** are not resolved, the deep learning framework will prove weak when transferred from research settings to real-world applications.

To resolve these challenges, this study presents a multi-dimensional image quality assessment framework, which integrates basic, technical, and medical relevance quality indicators of image through a three-component scoring structure. Each metric is normalized using dataset-specific statistics and adaptively thresholded using DR severity-specific percentiles derived from stratified samples, with severity-specific weighting to match diagnostic needs. Strict thresholds for early DR detection and more relaxed criteria for advanced stages. For detecting and staging DR, retinal images must be technically, clinically, and anatomically easy to understand. The proposed quality filtering pipeline examines retinal images from multiple directions. This approach differs from conventional techniques that consider basic quality parameters. It shows that medically relevant quality metrics are essential for optimal dataset curation. Enhanced imaging is necessary for the identification of early-stage conditions (microaneurysms and minor vascular abnormalities).

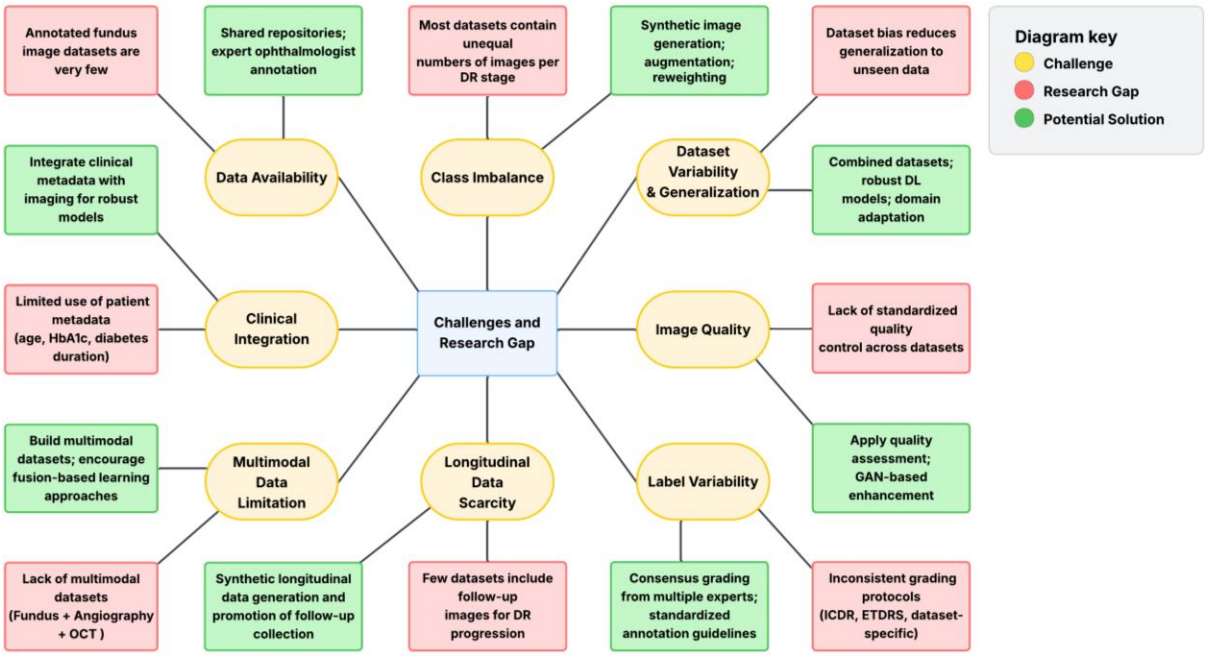

Figure 2. Overview of key challenges, research gaps, and potential solutions in DR fundus image datasets (Butt et al., 2022; Kotiyal and Pathak, 2022; Mutawa et al., 2023; Rajesh et al., 2023; Abushawish et al., 2024; Raghad and Hamad, 2024; Saproo et al., 2024; Shakibania et al., 2024; Taha et al., 2024; Almas et al., 2025; Kim et al., 2025; Refat et al., 2025; Riotto et al., 2025; Zafar et al., 2025; Zhang et al., 2025).

The framework is deployed in constructing the Diabetic Retinopathy Enhanced, Adapted, and Merged Retinal Fundus Image (DREAM-RFI) Dataset. This is a harmonized and curated fundus image dataset for DR staging. It has low bias and balanced representation of all classes. The DREAM-RFI merges images from IDRiD (Porwal et al., 2018), Messidor-2 (Decencière et al., 2014), SUSTech-SYSU (Lin et al., 2020), APTOS 2019 (Karthik et al., 2019), DeepDRiD-v1.1 (Liu et al., 2022), and Zenodo DR V03 (Benítez et al., 2021) datasets. The DREAM-RFI assigns DR grades according to the ICDR severity scale. The DREAM-RFI dataset immediately solves the fundamental problems of the existing DR dataset. It does this by providing high-quality, equally distributed, and diverse data. This dataset can improve the accuracy of deep learning models and be useful in more situations.

The key contributions of this study are outlined as follows:

i. Proposed a multi-dimensional image quality assessment methodology that includes basic, technical, and medical relevance indicators to identify low-quality samples and enhance dataset reliability.
ii. Executed integration of DR datasets based on image quality from several imaging devices and populations to augment dataset diversity and cross-population generalizability. Also, the robustness was maintained by removing low-quality images and minimizing excessive overfitting on the same source dataset.
iii. Developed and openly released a complete dataset creation pipeline via GitHub, assuring the reproducibility, transparency, and accessibility of the DREAM-RFI dataset for future research.
iv. Established baseline performance benchmarks to allow comparison studies and direct future model development by training and evaluating a number of deep learning models on DREAM-RFI.

The remaining structure of this article is as follows: Section 2 presents the functioning of the multi-dimensional image quality assessment framework. Section 3 describes the experimental setup and validation results. Finally, in Section 4, the summary of our findings and possible directions for future functions are given.

2. Methodology

This study presents a multi-dimensional framework for assessing image quality in retinal fundus images that considers technical excellence and medical significance, as seen in **Figure 3**. The framework is intended to identify and systematically eliminate low-quality retinal pictures that may have a negative impact on model training while maintaining clinical diversity for effective DR staging. Each fundus image is represented by a set of 10 scalar quality metrics:

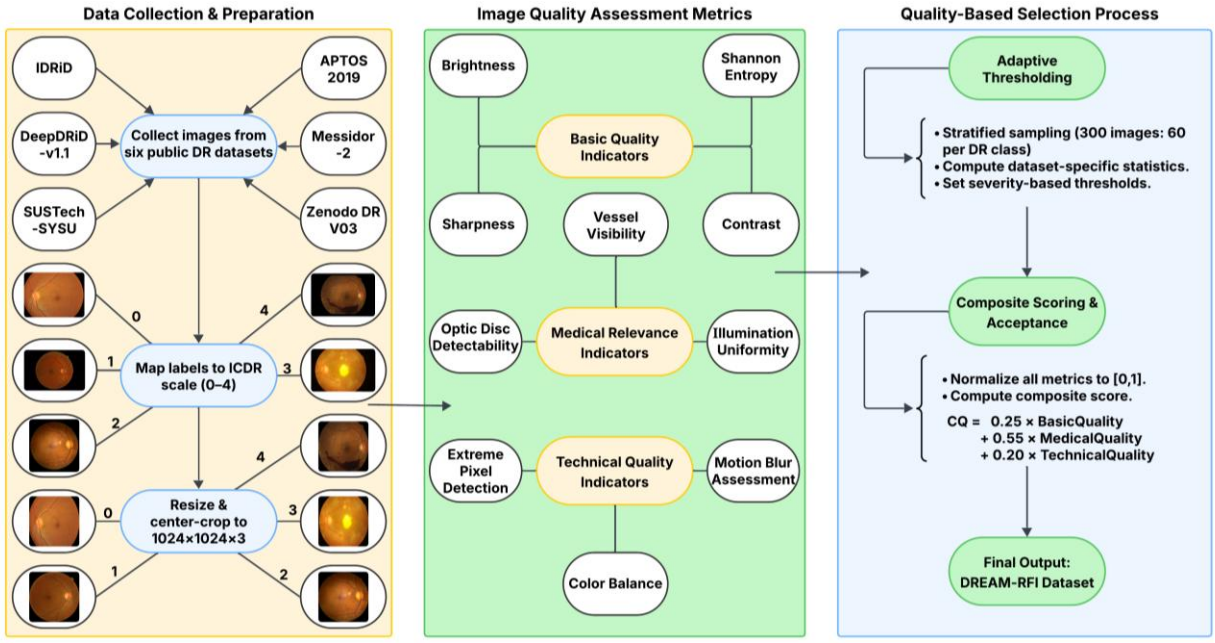

Figure 3. Detailed pipeline for retinal image quality assessment and dataset preparation.

Table 2. Components of the retinal image quality assessment vector Q.

Category	Metrics	Interpretation
Basic quality metrics	Brightness (B)	Measures overall illumination uniformity across the retinal field.
	Contrast (C)	Captures intensity spread and distinguishes vessel vs. background regions.
	Sharpness (S)	Reflects clarity of structural boundaries, such as vessels and lesions.
	Entropy (H)	Quantifies the richness of visual information and textural variability.
Medical quality metrics	Illumination Uniformity (U)	Assesses smoothness and absence of artifacts in the fundus background.
	Vessel Visibility (V)	Reflects clarity and delineation of vascular structures critical for diagnosis.
	Optic Disc Visibility (OD)	Ensures optic disc is well-defined for reliable anatomical localization.
Technical quality metrics	Extreme Pixel Detection (EP)	Validates proper exposure of retinal image without under/over saturation.
	Motion Blur Assessment (MB)	Indicates presence of motion-related degradation.
	Color Balance (CB)	Ensures natural color tone necessary for clinical interpretation.

$$\boldsymbol{Q} = \left[\underbrace{B, C, S, H}_{\text{Basic}}, \underbrace{U, V, OD}_{\text{Medical}}, \underbrace{EP, MB, CB}_{\text{Technical}}\right] \quad (1)$$

where, each component corresponds to a scalar quality metric described in **Table 2**.

2.1. Data Preparation

Retinal fundus images were gathered from six public DR datasets (IDRiD (Porwal et al., 2018), Messidor-2 (Decencière et al., 2014), SUSTech-SYSU (Lin et al., 2020), APTOS 2019 (Karthik et al., 2019), DeepDRiD-v1.1 (Liu et al., 2022), and Zenodo DR V03 (Benítez et al., 2021)) and mapped to the standard ICDR scale (0 - 4). Non-standard labels were converted accordingly. All images were uniformly resized and center-cropped to $1024 \times 1024 \times 3$ as seen in **Figure 4**. Image cropping is the initial preprocessing step to eliminate the irrelevant background surrounding the fundus images, which typically includes a large black border. This black area does not give any clinical information and can make learning algorithms work worse. Therefore, only the region containing the eye is preserved using a bounding-box-based cropping method, followed by resizing the result to a standard resolution.

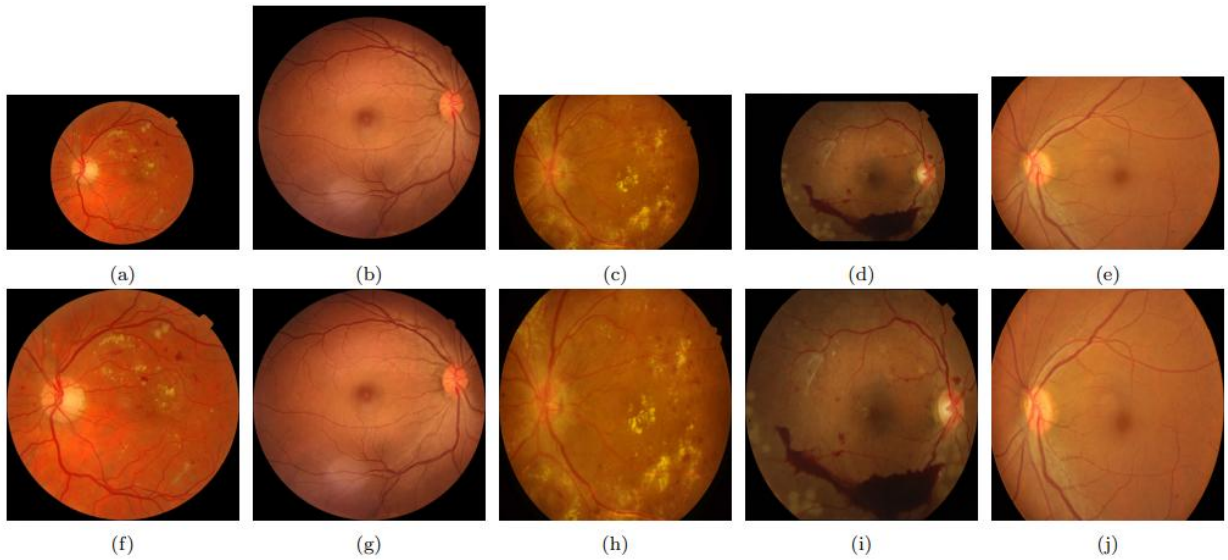

Figure 4. Visual results for data preparation step: (a - e) original images and (f - j) center-cropped and resized versions.

Each image is first converted to grayscale, followed by applying a low threshold of 10 to separate the dark background from the eye region, as seen in Equation (2).

$$T(i,j) = \begin{cases} 2, & \text{if } I(i,j) > 10 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where, $I(i,j)$ is the grayscale intensity at pixel (i,j), and $T(i,j)$ is the binarized image. This transformation shows the eye as a white region on a dark background. The binary mask is used to identify related regions, and the largest contour is assumed to include the eye. To calculate the area (A) of each contour, use the shoelace formula:

$$A = \frac{1}{2}\sum_{l=0}^{n-1} (i_l\, j_{l+1} - i_{l+1}\, j_l),\ where\ (i_n\ ,\ j_n) = (i_0,\ j_0) \quad (3)$$

Once the largest contour is identified, a bounding box is drawn around it, which returns the bounding box coordinates (i, j) and dimensions (w, h). i goes from top to bottom and j goes from left to right:

$$(i, j, w, h) = (\text{row}, \text{column}, \text{width}, \text{height}) \quad (4)$$

By getting rid of the black background, this method successfully separates the Region of Interest in the retinal image. This method gets the input ready for jobs like quality filtering and diagnosis.

2.2 Image Quality Indicators

The quality of the retinal image plays a crucial role in accurate DR diagnosis. This section describes the image quality parameters used.

2.2.1 Basic Quality Indicators

This study used four basic quality parameters that have a direct effect on how easy it is to read an image:

i. **Brightness** checks for images that are too or too little exposed by looking at their global exposure (Gonzalez and Woods, 2018). It gives a global intensity indicator to define the visual differences seen in fundus pictures. Brightness imbalance can conceal small DR lesions by either washing them out or hiding them in the background. It is defined by

$$B = \frac{1}{1024 \times 1024} \sum_{i=1}^{1024} \sum_{j=1}^{1024} I(i, j) \quad (5)$$

ii. **Contrast** locates local variations in image intensity. It is not based on edge gradients, which are not stable in low-light or noisy areas. A high value of contrast indicates better separation of retinal structures (Gonzalez and Woods, 2018). It is determined as the standard deviation of pixel intensities:

$$C = \sqrt{\frac{1}{1024 \times 1024} \sum_{i=1}^{1024} \sum_{j=1}^{1024} (I(i, j) - B)^2} \quad (6)$$

iii. **Sharpness** is used in measuring the structural clarity of vessels and micro-lesions. This is a good measure to maintain fine vascular boundaries (Gonzalez and Woods, 2018). The Laplacian operator is equal to:

$$L(i, j) = \nabla^2 I(i, j) = \frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2} \quad (7)$$

The sharpness is estimated by:

$$S = \frac{1}{1024 \times 1024} \sum_{i=1}^{1024} \sum_{j=1}^{1024} \left(L(i, j) - \overline{L}\right)^2 \quad (8)$$

where, $\overline{L}$ denotes the mean value of $L(i, j)$ across the entire image. Sharper images result from higher S values. Blurrier images result from lower S values. Sharpness is crucial in fundus image quality filtering, as blurry pictures might mask small vascular structures and minor DR lesions, which can lead to misdiagnosis.

iv. **Shannon Entropy** is a quantification of visual information in the image. It reflects textural and structural complexity essential for DR stage differentiation. It is defined by

$$H = -\sum_{q=0}^{255} p_q \log_2 p_q \quad (9)$$

where, p_q denotes the normalized histogram probability of the intensity level q. Lower value of H reduces intensity variation (Gonzalez and Woods, 2018) and causes DR signs (microaneurysms and fine vascular

changes) to remain hidden.

2.2.2 Medical Relevance Indicators

Three medical relevance indicators essential in the diagnosis of DR have been included in this work:

i. **Illumination Uniformity** measures the uniformity of the lighting throughout the image (Gonzalez and Woods, 2018). The image is divided into a 3×3 grid of 9 regions $\{R_m\}_{m=1}^{9}$. The mean brightness in each region is:

$$B_m = \frac{1}{|R_m|}\sum_{(i,j)\in R_m} I(i,j) \tag{10}$$

where, $|R_m|$ is the number of pixels in R_m. Illumination uniformity is then defined as

$$U = 1 - \frac{\sigma(B_1,\ldots,B_9)}{\overline{B}} \tag{11}$$

where, $\sigma(.)$ the standard deviation and $\overline{B}$ the mean of $B_1, B_2, \ldots,$ and B_9. A higher value of U indicates evenly distributed illumination. A lower value of U shows uneven illumination, which can hide DR signs or create false pathological patterns.

ii. **Vessel Visibility** quantifies how clearly blood vessels appear in the retinal image. The green channel I_{green} is enhanced using multi-directional filters (K_h, K_v, K_{d1}, K_{d2}) to capture horizontal, vertical, and diagonal vessel structures (Gonzalez and Woods, 2018):

$$VS(i,j) = \max_{k\in\{h,v,d1,d2\}} \text{filter2D}(I_{\text{green}}, K_k) \tag{12}$$

The vessel visibility score is defined by:

$$V = \frac{\#\{V(i,j) > P_{95}(VS)\}}{1024\times1024} \tag{13}$$

where, # denotes the number of pixels satisfying the condition. $P_{95}(VS)$ is the 95th percentile of VS. A high value of V means that the vessel is easier to see. A low value of V value indicates poor picture quality. Images that aren't clear can hide vessel narrowing, tortuosity, and new blood vessel growth.

iii. **Optic Disc Detectability** evaluates the visibility of the optic disc, typically the brightest region in a retinal image. It is defined as the fraction of pixels with intensity above the 95th percentile:

$$OD = \frac{\#\{I(i,j) > P_{95}(I)\}}{1024\times1024} \tag{14}$$

A high value of OD shows a clearly visible optic disc. A low value of OD gives a poor image quality, which hinders the distinction between normal bright regions and exudates.

2.2.3 Technical Quality Indicators

In this study, three technical quality parameters that determine the clarity of the retinal image are assessed:

i. **Extreme Pixel Detection** is used to determine the percentage of very dark ($I(i,j) < 10$) or very bright ($I(i,j) > 245$) pixels, which can reflect improper exposure (Gonzalez and Woods, 2018):

$$EP = \frac{\#\{I(i,j) < 10\} + \#\{I(i,j) > 245\}}{1024\times1024} \tag{15}$$

A large EP value means that brightness has high variations. These variations can hide small retinal lesions. They may also produce false artifacts that may reduce the interpretation of an image.

ii. **Motion Blur Assessment** uses the gradient magnitude to find image sharpness (Gonzalez and Woods, 2018):

$$MB = \frac{1}{1024\times1024}\sum_{i=1}^{1024}\sum_{j=1}^{1024}\sqrt{\left(\frac{\partial I}{\partial x}(i,j)\right)^2+\left(\frac{\partial I}{\partial y}(i,j)\right)^2} \quad (16)$$

High values of MB will imply stronger edges and less motion blur. Low values of MB reflect blurring caused by eye movement or camera shake. Images with Low MB value may mask fine vascular structures and small lesions in DR.

iii. **Color Balance** helps in evaluating the continuity of color channels. It is defined by:

$$CB = \sigma([R_{\text{mean}}, G_{\text{mean}}, B_{\text{mean}}]) \quad (17)$$

where, R_{mean}, G_{mean}, and B_{mean} are the mean intensities of the red, green, and blue channels, respectively. A low CB shows balanced colors. A high CB is associated with color cast issues, such as overly reddish or bluish tones. These color issues may alter the severity of the lesions.

2.3 Adaptive Thresholding Strategy

Different imaging protocols, equipment, and demographics produce varying baseline quality characteristics. To avoid bias from uniform quality standards, the proposed method generates dataset-specific thresholds. A stratified sample of approximately 300 images proportionally sampled across available DR severity classes is analyzed to compute mean, standard deviation, and percentile distributions for each metric. These statistics define normalization constants and initial thresholds. Different DR severities require different quality standards for reliable diagnosis. The proposed adaptive approach uses a severity-specific percentile threshold obtained from dataset calibration, as shown in **Table 3**. This shows that clinical reality that high-quality images are required to detect micro lesions in early DR stages, while in advanced stages, where major lesions are present, comparatively low technical quality is also acceptable.

Table 3. Adaptive percentile thresholds for image quality across DR severity levels.

DR severity	Class label	Percentile threshold
No DR	0	15th percentile (strictest)
Mild NPDR	1	12th percentile
Moderate NPDR	2	10th percentile
Severe NPDR	3	8th percentile
PDR	4	5th percentile (most relaxed)

2.4 Adaptive Threshold Customization

This framework allows users to adapt the quality threshold according to their requirements. After the preliminary adaptive threshold assessment, this system prepares the flagging image sample and all quality metrics, so that users can review the filtering results and make the necessary amendments. This manual customization capacity allows researchers to adjust the individual metric threshold up or down, so that the characteristics and quality requirements of the dataset are consistent. Users can assess whether the thresholds automatically match their expectations and clinical requirements by looking at the flagged samples. The manual threshold section provides an interface that allows the threshold to be replaced for any of the ten quality metrics. Their impact on the removal image can be reviewed, and the norms can be sophisticated sequentially before creating the final quality-filtered dataset. This method keeps the structural rigor of a multidimensional evaluation framework while making it adaptable to different research needs.

2.5 Dataset-Adaptive Normalization and Composite Quality Scoring

This system uses a two-tier normalization method to make sure that datasets with different imaging properties can be fairly compared. To account for differences in equipment and acquisition, basic quality metrics are normalized with a z-score:

$$Q_{k,norm} = \max\left(0, \min\left(1, \frac{z_{score}+3}{6}\right)\right) \quad (18)$$

where, $z_{score} = \frac{Q_k - \mu_k}{\sigma_k}$, with μ_k and σ_k are the mean and standard deviation for the metric k. This z-score transformation with clipping ensures that normalized values to $[0,1]$. It preserves relative quality differences within each dataset.

Medical relevance indicators are normalized using direct scaling with validated bounds:

$$U_{norm} = \max(0, \min(1, U)) \quad (19)$$

$$V_{norm} = \max(0, \min(1, V \times 100)) \quad (20)$$

$$OD_{norm} = \max(0, \min(1, OD \times 50)) \quad (21)$$

Technical quality metrics are normalized considering their specific characteristics:

$$EP_{norm} = \max(0, \min(1{,}1 - (EP \times 2))) \quad (22)$$

$$MB_{norm} = \max(0, \min(1, MB/50)) \quad (23)$$

$$CB_{norm} = \max(0, \min(1{,}1 - (CB/50))) \quad (24)$$

A composite score integrates normalized quality metrics using a weighted combination of three quality components:

$$CQ = 0.25 \times \text{BasicQuality} + 0.55 \times \text{MedicalQuality} + 0.20 \times \text{TechnicalQuality} \quad (25)$$

where:

$$\text{BasicQuality} = \text{mean}([B_{\text{norm}}, C_{\text{norm}}, S_{\text{norm}}, H_{\text{norm}}]) \quad (26)$$

$$\text{MedicalQuality} = \text{mean}([U_{\text{norm}}, V_{\text{norm}}, OD_{\text{norm}}]) \quad (27)$$

$$\text{TechnicalQuality} = \text{mean}([EP_{\text{norm}}, MB_{\text{norm}}, CB_{\text{norm}}]) \quad (28)$$

An image is accepted if its composite quality score exceeds the severity-specific threshold determined during dataset calibration.

2.6 Quality Assessment Results

For each dataset, about 300 stratified samples were looked at to find the normalization factors and percentile distributions that were unique to that dataset. There was an equal number of these samples in each of the DR classes. The computed percentile values for each DR stage were utilized to establish quality benchmarks. This ensured that adaptive filtering was applied appropriately to the unique characteristics of each dataset. To make the most use of the computer's resources, progress reports had to be generated every 500 photographs, and trash had to be removed every 100 photos for large-scale processing.

2.7 DREAM-RFI Dataset

This process produces the DREAM-RFI dataset. It contains 3461 retinal images of five DR stages: No DR (1256 images), Mild NPDR (493 images), Moderate NPDR (779 images), Severe NPDR (396 images), and

PDR (537 images). The DREAM-RFI dataset provides a more balanced class distribution than other datasets mentioned in **Table 1**. It has a high normalized Shannon entropy (0.94) and a very low variance of proportions (0.008), indicating a well-distributed representation of all five DR severity levels. All these images were obtained from several public datasets and standardized through preprocessing, quality evaluation, and label harmonization. This ensures balanced class distribution and representation of various populations and reduces dataset bias. The DREAM-RFI dataset supports the image quality assessment, domain adaptation, explainable AI, and preprosecuting research. This pipeline allows reuse, expansion, and application in other medical imaging collections. This technique provides reproducibility of the dataset, transparency, and traceability. Any researcher having the required source material can recreate the DREAM-RFI dataset using the code available on https://doi.org/10.5281/zenodo.17587015.

Table 4. Description of the Hierarchical classification schemes for DR staging used in this study.

Type	Description	Clinical purpose	Classes – image count
Two-class scheme	Offers a clinically relevant screening framework that enables rapid disease detection and is well-suited for large-scale population screening.	**Screening:** Identify the presence vs. absence of DR for diagnosis.	i.No DR (0) – 1256 DR (1 to 4) – 2205
Three-class scheme	Balance diagnostic precision with robustness; address the need to identify patients requiring closer follow-up or referral without overburdening healthcare resources; mitigate class imbalance.	**Referral Prioritization:** Identify patients needing routine monitoring vs. urgent ophthalmic evaluation.	i.No DR (0) – 1256 ii.Early DR (1 and 2) – 1272 Advanced DR (3 and 4) – 933
Five-class scheme (ICDR Severity Scale)	Follows the ICDR severity scale; provides detailed stratification of DR; captures disease progression; supports clinical decision-making, enables comparison with expert annotations; benchmark for advanced diagnostic models.	**Treatment Planning:** Enables stage-wise assessment and comparison with expert diagnosis.	i.No DR (0) – 1256 ii.Mild NPDR (1) – 493 iii.Moderate NPDR (2) – 779 iv.Severe NPDR (3) – 396 PDR (4) – 537

3. Experimental Validation

A comprehensive experimental research was conducted to create reliable baselines and validate the utility of the DREAM-RFI dataset.

3.1 Classification Schemes

As indicated in **Table 4**, three classification schemes were used for experimental verification and evaluation of the DREAM-RFI dataset. Multiple classification methods can serve unique and complementary functions in DR staging. The hierarchical structure of these schemes makes them adaptable to different clinical and technical settings, facilitating rapid automated screening and sophisticated decision support in specialized care.

3.2 Experimental Setup

Figure 5 illustrates the pipeline for the experimental authentication of the DREAM-RFI dataset. For the backbone model, VGG-16, ResNet-50, Inception-V3, DenseNet-121, and MobileNet-V2 were chosen for their strong representational capability and complementary architectural qualities to construct accurate benchmarks on the DREAM-RFI dataset. All models were trained under standardized conditions using the AdamW optimizer with an initial learning rate of 1×10^{-4}, a batch size of 128, and 50 epochs with early stopping based on validation loss. Identical data partitioning criteria preserving a 70:10:20 train-validation-test data split ratio were utilized for facilitating a fair comparison with prior DR classification methodologies. The fundus photos were resized to $384 \times 384 \times 3$ pixels. To preserve anatomical validity of the augmented retinal image, medically relevant image augmentation techniques: rescaling, rotation of

small angles (maximum ± 12), width and height shifts, slight zooms, brightness adjustment, and the green channel enhancement through the CLAHE were used. Horizontal flips were excluded as they generate anatomically impossible retinal orientation, which is not found in clinical behavior and can affect the capacity of the model's generalization on real clinical data. All experiments were run on an NVIDIA GeForce RTX 3090 GPU using Keras with a TensorFlow backend.

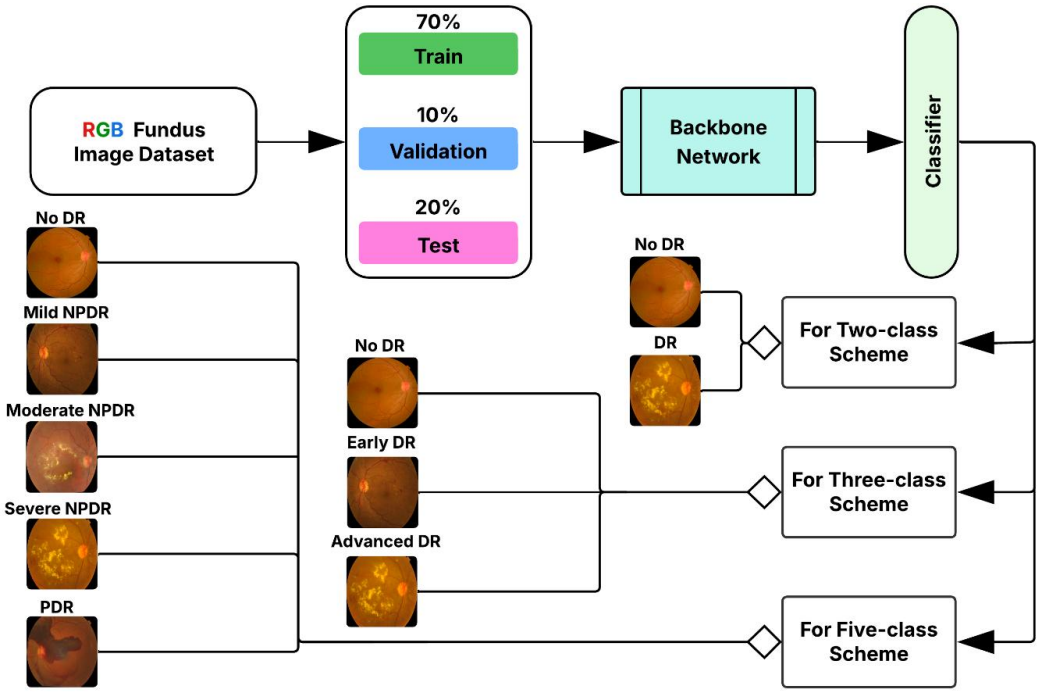

Figure 5. The pipeline for experimental validation of the DREAM-RFI dataset.

Accuracy (acc), precision (prec), recall, F1-score (F1), and AUC were used to evaluate the DREAM-RFI dataset. The accuracy of the model reflects the proportion of correctly identified samples of the total samples, which indicates the effectiveness of the model. The precision measures the percentage of correct predictions in positive predictions and helps in reducing false positives. The recall evaluates the ability to identify the real positives of the model and aids in reducing false negatives. F1-score is the harmonic mean of precision and recall, and balances both standards and is useful for uneven data. Additionally, AUC assesses the discriminatory ability of the model on various classification thresholds, where high values reflect better performance. This method assesses the effect of each class and reduces bias towards large or small classes. This makes the model assessment of DR and grading more accurate. These criteria ensure intensive and rigorous evaluation of the effectiveness of the model in important applications such as DR diagnosis.

3.3 Results

Table 5 compares five widely used CNN backbones, including VGG-16, ResNet-50, Inception-V3, MobileNet-V2, and DenseNet-121, trained on (i) the combined dataset without quality filtering and (ii) the proposed DREAM-RFI dataset. DREAM-RFI consistently outperforms all architectures across all three DR classification schemes (two-class, three-class, and five-class), proving the efficacy of the quality-filtered merging procedure. The results indicate that DenseNet-121 consistently outperforms other architectures in all tasks. DenseNet-121 gained the highest classification accuracy: 95.20% for two-class, 92.84% for three-

class, and 85.12% for five-class classification. VGG-16 and MobileNet-V2 provided relatively low accuracies. ResNet-50 and Inception-V3 performed moderately.

Technically, the improvement in accuracy is consistent rather than model-specific. There are gains of 3–5% in two-class classification, 4–7% in three-class classification, and 4–6% in five-class classification. This shows that DREAM-RFI increases inter-class separability, especially in more difficult multi-class cases where label noise and low-quality images regularly propagate errors into deeper network layers. The merged dataset without quality filtering provides lower and more variable performance. This happens due to irregular illumination, motion blur, occlusions, and device-specific abnormalities. These distortions significantly reduce the performance in three-class and five-class settings. This impairs the ability to differentiate between small lesions (microaneurysms, IRMA, and neovascularization). Severity-aware thresholding and adaptable quality filtering of the DREAM-RFI reduce noise within classes. This helps models to identify patterns that are diagnostically important.

Table 5. Comparison of baseline backbone models on the DREAM-RFI dataset and the merged dataset without quality filtering.

Backbone	Merged dataset without quality filtering			DREAM-RFI Dataset		
	Two-class	Three-class	Five-class	Two-class	Three-class	Five-class
VGG-16	85.24%	80.12%	72.30%	89.10%	84.32%	76.45%
ResNet-50	87.60%	82.40%	74.85%	91.05%	87.22%	79.20%
Inception-V3	89.45%	84.65%	76.95%	92.80%	89.35%	81.60%
MobileNet-V2	86.15%	81.05%	73.10%	90.20%	85.10%	77.80%
DenseNet-121	91.42%	86.56%	80.24%	95.20%	92.82%	85.12%

Additional assessments were conducted using precision, recall, F1-score, and AUC metrics. **Table 6** details the performance of DenseNet-121. Nearly complete discrimination was achieved in the two-class setting by yielding an AUC of 0.9973. The three-class configuration achieved an AUC of 0.9947. DenseNet-121 achieved an AUC of 0.9409 in the challenging five-class scenarios. To ensure class balance and reliability of the results, experiments were verified by using a number of data divisions through many training and evaluation stages, ensuring class balance and reliability of the results. **Figure 6** displays the confusion matrix and ROC curves. This verification demonstrates the capabilities of DenseNet-121 and the DREAM-RFI dataset at different classification levels. This can serve as a benchmark dataset for assessing DR severity.

Table 6. Performance of DenseNet-121 for different classification settings on the DREAM-RFI dataset.

Classification Type	Accuracy	Precision	Recall	F1-score	AUC
Two-class	95.20%	95.95%	94.53%	95.23%	0.9973
Three-class	92.82%	94.96%	91.61%	93.29%	0.9947
Five-class	85.12%	85.20%	85.12%	85.14%	0.9409

Table 7 provides a comparative review of the recent studies. This table shows the datasets, functions, and different classification strategies that are adopted in the literature. The suggested DREAM-RFI dataset with DenseNet-121 achieved competitive performance. It recivied 95.20%, 92.82%, and 85.12% accuracy in two-class, three-class, and five-class settings, respectively. These findings confirm that the DREAM-RFI data is a competitive multi-level DR classification benchmark. Find out also that simple architectures can also perform reliably on this dataset. At this point, we are concerned with basic certification with CNN backbones. In the future, it will be upgraded to incorporate more complex architectures. Special training techniques will continue to facilitate performance improvements.

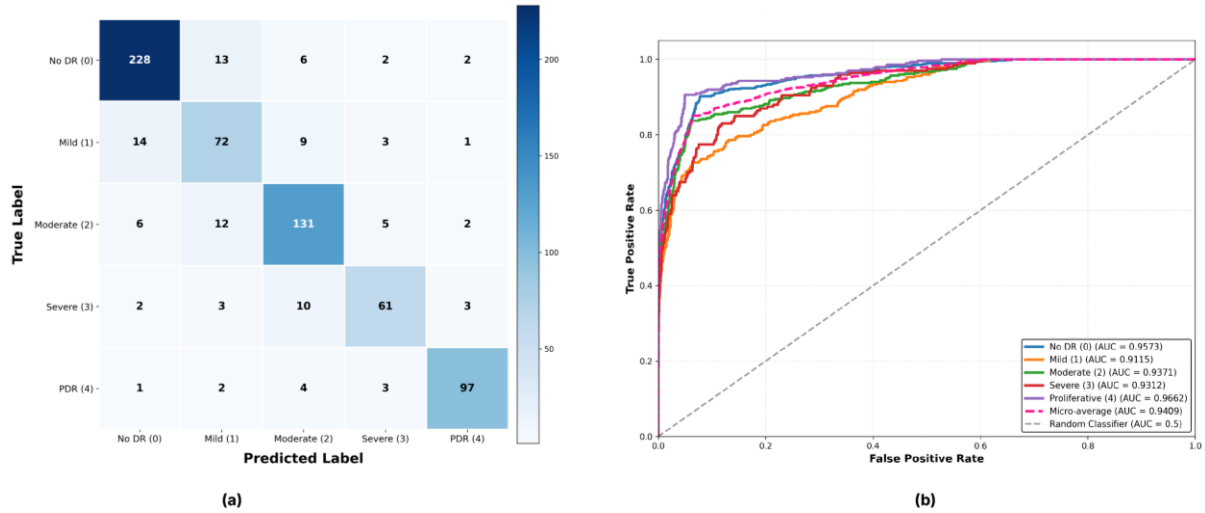

Figure 6. Visual results of DenseNet-121 for five-class settings on the DREAM-RFI dataset. (a) Confusion matrix. (b) ROC curves.

Table 7. A summary of recent studies on DR detection.

Study	Dataset combinations	Methodology	Classification type	Performance
Kotiyal and Pathak (2022)	IDRiD	Inception-V3, Xception, and VGG-19	Two-class	Acc: Inception-V3 (95%), Xception (92.50%), and VGG-19 (89.94%).
Butt et al. (2022)	APTOS 2019	Hybrid GoogleNet + ResNet-18 with SVM	Two-class and Three-class	Acc: Two-class (97.80%) and Three-class (89.29%).
Mutawa et al. (2023)	EyePACS + APTOS 2019 + ODIR	DenseNet-121	Two-class 8059663067	Acc: Two-class (98.97%).
Raghad and Hamad (2024)	IDRiD + Messidor-2	VGG-19, DenseNet-121, and EfficientNet B6	Five-class	Acc: VGG-19 (80%), DenseNet-121 (88%), and EfficientNet B6 (90%).
Saproo et al. (2024)	DRD-EyePACS + IDRiD + APTOS 2019	VGG-19, ResNet-101, and ShuffleNet	Two-class	Acc: VGG-19 (96.22%), ResNet-101 (97.33%), and ShuffleNet (96.66%).
Shakibania et al. (2024)	APTOS 2019 + IDRiD + Messidor-2	Dual Branch Model (ResNet-50 + EfficientNetB0)	Two-class and Five-class	Acc: Two-class (98.50%) and Five-class (89.60%).
Almas et al. (2025)	EyePACS and Kaggle	Enhanced Stacked auto-encoders	Five-class	Acc: EyePACS (88%) and Kaggle (79.50%).
Zafar et al. (2025)	APTOS 2019 + DDR + FairVision	lightweight 37-layer CNN model	Two-class and Four-class	Acc: Two-class (99.06%) and Four-class (90.75%).
Refat et al. (2025)	APTOS 2019 + DDR + IDRiD + Messidor-2 + RETINO	VR-FuseNet (VGG19 + ResNet50 fusion)	Five-class	Acc:91.82%, Prec: 92.61%, Recall: 92.23%, F1: 92.39%.
Zhang et al. (2025)	EyePACS + APTOS 2019	CNN, ViT, Hybrid Models	Five-class	Acc:72.93%, AUC: 0.93.
This Study	DREAM-RFI (**Proposed**)	DenseNet-121	Two-class, Three-class, and Five-class	Acc: Two-class (95.20%), Three-class (92.82%), and Five-class (85.12%). AUC: Two-class (0.9973), Three-class (0.9947), and Five-class (0.9409).

3.4 Discussion

The Dream-RFI dataset improves the applicability of deep learning models. It does this by unifying heterogeneous sources into a standardized format with quality control. Such a combination can be used to

neutralize biases in datasets. It also enables the trained models to be more transferable to unknown clinical settings so that they can be substantially used in real DR screening as well. This baseline evaluation provides the necessary standard for future research by using different backbone architectures. These standards validate the DREAM-RFI dataset as an appropriate testing platform. It also has performance benchmarks for comparison with advanced models.

Research experiments on the DREAM-RFI dataset were performed in three classification systems: (i) two-class scheme, (ii) three-class scheme, and (iii) Five-Class scheme. In this method, the model was tested based on the levels of different clinical complications, and it showed the adaptability of the backbone. The best performance was recorded with a two-class scheme based on the highest accuracy of 95.20% and AUC of 0.9973. This makes it effective in the first screening tasks. The three-class system had good results with an accuracy of 92.82% and an AUC of 0.9947. This shows that this data can be used to do intermediate grading. It effectively indicates the worsening of the sickness. The five-class classification is more difficult and useful clinically because it gives 85.12% accuracy and 0.9409 AUC. The findings indicate that DenseNet-121 is effective in all levels of classification. It can be utilized as a benchmark for the DREAM-RFI multi-level DR classification workload.

This paper demonstrates that the use of quality-filtered merging of datasets improves the accuracy of model generalization by 3-7% over the standard merging of datasets. This is empirical evidence that the systematic quality evaluation of the data set is the key to improving the clinical utility of automated systems of DR screening. Extensions to the future will be based on a hierarchical three-step classification framework, which is sequential in nature. Such a hierarchical approach is likely to enhance interpretability. It would be more compatible with clinical decision-making procedures. This also offers an in-depth understanding of the reinforcement of the DREAM-RFI data in clinical environments with stages of clinical challenges. Also, the installation of fractional activation beans in lightweight models will enable the preservation of comparative efficiency due to enhanced performance. The use of other public datasets to expand the DREAM-RFI dataset will add more diversity as well as capacity to the dataset and increase its clinical relevance.

4. Conclusion

This paper has presented a DREAM-RFI dataset and gives a full pipeline for constructing, assessing, and benchmarking this dataset. The proposed pipeline was created due to a deep examination where both technical and medically applicable indicators were taken into consideration. The DREAM-RFI data set solves an issue of low-quality images, which interferes with the proper identification and staging of DRS. It was found that the simple evaluation of the three-class and five-class configurations based on two-class and DenseNet-121 demonstrated that they reached competitive accuracies of 95.20%, 92.82%, and 85.12%, respectively. The findings can offer a strong background to the usefulness of the DREAM-RFI dataset. To provide a better structure of improvement in the explanation and better coordination with clinical decision-making, the three-step hierarchical classification structure will be used in the future. Besides, the light models will be aimed at incorporating the functions of fractional activation and extension of DREAM-RFI to other open datasets in such a way that their variety, growth, and clinical applicability can be further reinforced.

Conflicts of Interest

The authors declare that they have no financial or personal conflicts of interest that may have influenced this paper.

Acknowledgements
This research acknowledges with appreciation the contributions of the original dataset sources (IDRiD (Porwal et al., 2018), Messidor-2 (Decencière et al., 2014), SUSTech-SYSU (Lin et al., 2020), APTOS 2019 (Karthik et al., 2019), DeepDRiD v1.1 (Liu et al., 2022), and Zenodo DR V03 (Benítez et al., 2021)) and extends our thanks to the creators and maintainers of the open-source tools and libraries that facilitated this research.
The code for producing the DREAM-RFI dataset is accessible at https://doi.org/10.5281/zenodo.17587015.
All data are from publicly available datasets that comply with institutional review requirements; no new human subjects data were collected.

AI Disclosure
During the preparation of this work the author(s) used generative AI in order to improve the language of the article. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Appendix: Metrics for evaluating dataset balance

Medical diagnosis require a balanced dataset, as misclassifying rare illnesses can lead to delayed therapy. An imbalanced dataset may cause models to favor the majority class, resulting in a missing diagnosis for rare but clinically relevant illnesses. Balance enables the model to learn from all classes equally, enhancing its sensitivity and accuracy for reliable and prompt medical decision-making. We have utilized Normalized Shannon Entropy and Variance of Proportions to evaluate the class balance of datasets. These measures are detailed in **Table 8**.

Table 8. Metrics for evaluating dataset balance in an n-class dataset. Here p_i is the proportion of samples in the i^{th} class and $\overline{p}$ is the mean class proportion.

Metrics	Formula	Best value
Normalized Shannon Entropy (NSE)	$\text{NSE} = -\frac{\sum_{i=0}^{n} p_i \log_2(p_i)}{\log_2(n)}$	1
Variance of Proportions (VP)	$\text{VP} = \frac{1}{n}\sum_{i=1}^{n}(p_i - \overline{p})^2$	0

References

Abushawish, I.Y., Modak, S., Abdel-Raheem, E., Mahmoud, S.A., & Hussain, A.J. (2024). Deep learning in automatic diabetic retinopathy detection and grading systems: a comprehensive survey and comparison of methods. *IEEE Access*, *12*, 84785-84802. https://doi.org/10.1109/access.2024.3415617.

Almas, S., Wahid, F., Ali, S., Alkhyyat, A., Ullah, K., Khan, J., & Lee, Y. (2025). Visual impairment prevention by early detection of diabetic retinopathy based on stacked auto-encoder. *Scientific Reports*, *15*(1), 2554. https://doi.org/10.1038/s41598-025-85752-2.

Benítez, V.E.C., Matto, I.C., Román, J.C.M., Noguera, J.L.V., García-Torres, M., Ayala, J., Pinto-Roa, D.P., Gardel-Sotomayor, P.E., Facon, J., & Grillo, S.A. (2021). Dataset from fundus images for the study of diabetic retinopathy. *Data in Brief*, *36*, 107068. https://doi.org/10.1016/j.dib.2021.107068.

Butt, M.M., Iskandar, D.N.F.A., Abdelhamid, S.E., Latif, G., & Alghazo, R. (2022). Diabetic retinopathy detection from fundus images of the eye using hybrid deep learning features. *Diagnostics*, *12*(7), 1607. https://doi.org/10.3390/diagnostics12071607.

Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., Gain, P., Ordonez, R., Massin, P., Erginay, A., Charton, B., & Klein, J. (2014). Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, *33*(3), 231. https://doi.org/10.5566/ias.1155.

Gonzalez, R.C., & Woods, R.E. (2018). *Digital image processing* (4th ed.). Pearson, UK.

Gulati, S., Guleria, K., & Goyal, N. (2023). Classification of diabetic retinopathy using pre-trained deep learning model- DenseNet 121. In 14*th International Conference on Computing Communication and Networking Technologies* (pp. 1-6), Delhi, India. https://doi.org/10.1109/icccnt56998.2023.10308181.

Gulati, S., Guleria, K., & Goyal, N. (2025). Privacy-preserving and collaborative federated learning model for the detection of ocular diseases. *International Journal of Mathematical Engineering and Management Sciences*, *10*(1), 218-248. https://doi.org/10.33889/ijmems.2025.10.1.013.

Gulati, S., Guleria, K., Goyal, N., AlZubi, A.A., & Castilla, Á.K. (2024). A privacy-preserving collaborative federated learning framework for detecting retinal diseases. *IEEE Access*, *12*, 170176-170203. https://doi.org/10.1109/access.2024.3493946.

Hill-Briggs, F., Adler, N.E., Berkowitz, S.A., Chin, M.H., Gary-Webb, T.L., Navas-Acien, A., Thornton, P.L., & Haire-Joshu, D. (2020). Social determinants of health and diabetes: a scientific review. *Diabetes Care*, *44*(1), 258-279. https://doi.org/10.2337/dci20-0053.

Karthik, Maggie, & Dane, S. (2019). *APTOS 2019 blindness detection*. Kaggle. https://www.kaggle.com/c/aptos2019-blindness-detection.

Kim, S.J., Lim, J.I., Bailey, S.T., Kovach, J.L., Vemulakonda, G.A., Ying, G.S, & Flaxel, C.J. (2025). Idiopathic macular hole preferred practice pattern®. *Ophthalmology, 132*(4), 234-269. https://doi.org/10.1016/j.ophtha.2024.12.021.

Kotiyal, B., & Pathak, H. (2022). Diabetic retinopathy binary image classification using Pyspark. *International Journal of Mathematical Engineering and Management Sciences*, *7*(5), 624-642. https://doi.org/10.33889/ijmems.2022.7.5.041.

Lin, L., Li, M., Huang, Y., Cheng, P., Xia, H., Wang, K., Yuan, J., & Tang, X. (2020). The SUSTech-SYSU dataset for automated exudate detection and diabetic retinopathy grading. *Scientific Data*, *7*(1), 409. https://doi.org/10.1038/s41597-020-00755-0.

Liu, R., Wang, X., Wu, Q., Dai, L., Fang, X., Yan, T., Son, J., Tang, S., Li, J., Gao, Z., Galdran, A., Poorneshwaran, J.M., Liu, H., Wang, J., Chen, Y., Porwal, P., Tan, G.S.W., Yang, X., Dai, C., Song, H., Chen, M., Li, H., Jia, W., Shen, D., Sheng, B., & Zhang, P. (2022). DeepDRID: diabetic retinopathy—grading and image quality estimation challenge. *Patterns*, *3*(6), 100512. https://doi.org/10.1016/j.patter.2022.100512.

Mansour, S.E., Browning, D.J., Wong, K., Flynn, H.W., Jr, & Bhavsar, A.R. (2020). The evolving treatment of diabetic retinopathy. *Clinical Ophthalmology*, *14*, 653-678. https://doi.org/10.2147/opth.s236637.

Men, Y., Fhima, J., Celi, L.A., Ribeiro, L.Z., Nakayama, L.F., & Behar, J.A. (2025). Deep learning generalization for diabetic retinopathy staging from fundus images. *Physiological Measurement, 46*, 015001. https://doi.org/10.1088/1361-6579/ada86a.

Mutawa, A.M., Alnajdi, S., & Sruthi, S. (2023). Transfer learning for diabetic retinopathy detection: a study of dataset combination and model performance. *Applied Sciences*, *13*(9), 5685. https://doi.org/10.3390/app13095685.

National Eye Institute. (2019). *People with diabetes can prevent vision loss.* National Institutes of Health. https://www.nei.nih.gov/sites/default/files/2019-06/diabetes-prevent-vision-loss.pdf.

Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., & Meriaudeau, F. (2018). Indian Diabetic Retinopathy Image Dataset (IDRID): a database for diabetic retinopathy screening research. *Data*, *3*(3), 25. https://doi.org/10.3390/data3030025.

Raghad, R., & Hamad, A.H. (2024). Multi-Label diabetic retinopathy detection using transfer learning based convolutional neural network. *Fusion Practice and Applications*, *17*(2), 279-293. https://doi.org/10.54216/fpa.17022.

Rajesh, A.E., Davidson, O.Q., Lee, C.S., & Lee, A.Y. (2023). Artificial intelligence and diabetic retinopathy: AI framework, prospective studies, head-to-head validation, and cost-effectiveness. *Diabetes Care*, *46*(10), 1728-1739. https://doi.org/10.2337/dci23-0032.

Refat, S.R., Raha, Z.S., Sarker, S., Preotee, F.F., Rahman, M.M., Muhammad, T., & Alam, M.S. (2025). *VR-FuseNet: A fusion of heterogeneous fundus data and explainable deep network for diabetic retinopathy classification*. arXiv. https://arxiv.org/abs/2504.21464.

Riotto, E., Tsai, W.S., Khalid, H., Lamanna, F., Roch, L., Manoj, M., & Sivaprasad, S. (2025). intergrader agreement on qualitative and quantitative assessment of diabetic retinopathy severity using ultra-widefield imaging: INSPIRED study report 1. *Diagnostics*, *15*(14), 1831. https://doi.org/10.3390/diagnostics15141831.

Saproo, D., Mahajan, A.N., & Narwal, S. (2024). Deep learning based binary classification of diabetic retinopathy images using transfer learning approach. *Journal of Diabetes & Metabolic Disorders*, *23*(2), 2289-2314. https://doi.org/10.1007/s40200-024-01497-1.

Shakibania, H., Raoufi, S., Pourafkham, B., Khotanlou, H., & Mansoorizadeh, M. (2024). Dual branch deep learning network for detection and stage grading of diabetic retinopathy. *Biomedical Signal Processing and Control*, *93*, 106168. https://doi.org/10.1016/j.bspc.2024.106168.

Sun, H., Saeedi, P., Karuranga, S., Pinkepank, M., Ogurtsova, K., Duncan, B.B., Stein, C., Basit, A., Chan, J.C.N., Mbanya, J.C., Pavkov, M.E., Ramachandaran, A., Wild, S.H., James, S., Herman, W.H., Zhang, P., Bommer, C., Kuo, S., Boyko, E.J., & Magliano, D.J. (2022). IDF diabetes atlas: global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Research and Clinical Practice*, *183*, 109119. https://doi.org/10.1016/j.diabres.2021.109119.

Taha, A.A., Dinesen, S., Vergmann, A.S., & Grauslund, J. (2024). Present and future screening programs for diabetic retinopathy: a narrative review. *International Journal of Retina and Vitreous*, *10*(1), 14. https://doi.org/10.1186/s40942-024-00534-8.

Wilkinson, C.P., Ferris, F.L., Klein, R.E., Lee, P.P., Agardh, C.D., Davis, M., Dills, D., Kampik, A., Pararajasegaram, R., & Verdaguer, J.T. (2003). Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*, *110*(9), 1677-1682. https://doi.org/10.1016/s0161-6420(03)00475-5.

Yau, J.W.Y., Rogers, S.L., Kawasaki, R., Lamoureux, E.L., Kowalski, J.W., Bek, T., Chen, S.J., Dekker, J.M., Fletcher, A., Grauslund, J., Haffner, S., Hamman, R.F., Ikram, M.K., Kayama, T., Klein, B.E.K., Klein, R., Krishnaiah, S., Mayurasakorn, K., O'Hare, J.P., Orchard, T.J., Porta, M., Rema, M., Roy, M.S., Sharma, T., Shaw, J., Taylor, H., Tielsch, J.M., Varma, R., Wang, J.J., Wang, N., West, S., Xu, L., Yasuda, M., Zhang, X., Mitchell, P., & Wong, T.Y. (2012). Global prevalence and major risk factors of diabetic retinopathy. *Diabetes Care*, *35*(3), 556-564. https://doi.org/10.2337/dc11-1909.

Zafar, A., Kim, K.S., Ali, M.U., Byun, J.H., & Kim, S.H. (2025). A lightweight multi-deep learning framework for accurate diabetic retinopathy detection and multi-level severity identification. *Frontiers in Medicine*, *12*. https://doi.org/10.3389/fmed.2025.1551315.

Zhang, D., Zhang, Y., Kang, J., & Li, X. (2024). Nonlinear relationship between diabetes mellitus duration and diabetic retinopathy. *Scientific Reports*, *14*(1), 30223. https://doi.org/10.1038/s41598-024-82068-5.

Zhang, W., Belcheva, V., & Ermakova, T. (2025). Interpretable deep learning for diabetic retinopathy: a comparative study of CNN, VIT, and hybrid architectures. *Computers*, *14*(5), 187. https://doi.org/10.3390/computers14050187.