MELBA Journal Machine Learning for Biomedical Imaging

M-FIQ: Mobile Fundus Image Quality Assessment Dataset for Teleophthalmology Screening

João M. M. **Moreira** [1, 2], João D. S. **Almeida** [1], Elaine P. F. **Costa** [3], Jhones S. **Soares** [1], Luís F. R. **Pereira** [1, 2], Emily G. C. **Ribeiro** [1], Amanda S. **Almeida** [1], Iaze G. S. C. **Santos** [1], Aristófanes C. **Silva** [1], Darlan B. P. **Quintanilha** [1], Anselmo C. **Paiva** [1], Geraldo B. **Junior** [1], Marcos A. G. **Campos** [3]

1 Applied Computing Group (NCA), Federal University of Maranhão, P.O. Box 65.080-805, São Luís, MA, Brazil
2 Postgraduate Program in Computer Science (PPGCC), Federal University of Maranhão,P.O. Box 65.080-805, São Luís, MA, Brazil
3 School of Medicine Coordination, Federal University of Maranhão,P.O. Box 65.080-805, São Luís, MA, Brazil

Abstract

Retinal image quality assessment (RIQA) is a critical step in ophthalmic screening workflows, as low-quality fundus images can compromise both clinical diagnosis and the performance of automated artificial intelligence systems. Although several public datasets have been proposed for RIQA research, most were acquired using conventional tabletop fundus cameras, while datasets collected with portable devices remain scarce. This paper presents M-FIQ, a novel public dataset for retinal image quality assessment in portable fundus imaging. The dataset comprises 7,971 fundus photographs collected from 304 patients using the portable Eyer2 fundus camera in a community-based screening program in Brazil. M-FIQ includes both color (2,656) and red-free (5,315) fundus images, each annotated according to three quality levels: Good, Usable, and Reject. In addition, images labeled as Usable or Reject are further annotated with the specific degradation factors responsible for quality impairment, including low sharpness, underexposure, overexposure, incomplete field of view, peripheral shadowing, and motion artifacts. To facilitate reproducible research, we provide patient-level train/test splits and establish a benchmark using four deep learning architectures: DenseNet121, EfficientNet-B0, ResNet50, and ViT-B16. Experimental results demonstrate the suitability of the dataset for RIQA tasks, with ViT-B16 achieving the highest F1-score on the color fundus image test set (78.44%), whereas ResNet50 achieved the highest F1-score on the red-free fundus image test set (74.10%). By addressing the lack of publicly available RIQA datasets acquired with portable fundus cameras, M-FIQ aims to support the development and evaluation of robust image quality assessment methods for ophthalmic screening and teleophthalmology applications. The dataset is publicly available at https://www.synapse.org/Synapse:syn75868226.

Keywords

Retinal Image Quality Assessment, Portable Fundus Camera, Deep Learning, Dataset, Fundus Image

Article informations

https://doi.org/10.59275/j.melba.2026-b2dg
Volume 2026, Received: 2026-06-19, Published 2026-09-21
Corresponding author: joao.marcello@nca.ufma.br
Special issue: MICCAI Open Data 2026 x MELBA
Guest editors: AT, MS et al.

1. Background

Fundus imaging is the most widely used clinical examination for diagnosing ocular diseases such as diabetic retinopathy, glaucoma, and age-related macular degeneration (Shen et al., 2020), as it accurately captures important retinal anatomical structures, including the optic nerve, macula, and blood vessels, enabling effective assessment of the ocular fundus (Ju et al., 2020).

A medical image is considered to be of adequate quality when all structures, including pathological findings, are clearly visible (Raj et al., 2019). However, the final image quality may be compromised by a variety of factors, such as dirt on the camera lens, improper flash settings, blinking, eyelash occlusion, and lack of technical expertise (Abramovich et al., 2023), resulting in artifacts such as

uneven illumination, low contrast, blur, and glare.

Studies indicate that approximately 25% of fundus images require improved quality to enable accurate diagnosis, while other estimates suggest that 12% are of insufficient quality for professional interpretation (Gong et al., 2025). Low-quality images adversely affect both manual diagnosis by human specialists and automated analysis by computer-aided systems, making image reacquisition necessary, a process that is costly, time-consuming, and inconvenient for patients (Zago et al., 2018). Furthermore, poor-quality images hinder diagnosis by specialists in teleophthalmology systems.

In this context, automated retinal image quality assessment (RIQA) has received increasing attention from the scientific community. Automated quality assessment methods can be integrated into the image acquisition process, supporting real-time quality control (Gong et al., 2025). However, public RIQA datasets exhibit limited image diversity and often fail to encompass different acquisition modalities used in clinical practice, such as images acquired with handheld cameras (Jin et al., 2023).

Motivated by these limitations, we present the Mobile Fundus Image Quality Assessment (M-FIQ) dataset. It com-prises color and red-free fundus images acquired using a handheld fundus camera and annotated according to a three-level quality grading scheme (Good, Usable, and Reject), as well as additional labels describing the factors responsible for image quality degradation. Figure 1 illustrates an example of a color fundus image and its corresponding red-free version from the proposed dataset.

(a) Color fundus image. (b) Red-free fundus image.

Figure 1: Example of color and red-free fundus image.

1.1 Related Works

In recent years, several datasets have been proposed to support the development and evaluation of automated RIQA methods.

EyeQ is one of the most widely used datasets for fundus image quality assessment. Proposed by Fu et al. (2019), the dataset contains 28,792 images selected from EyePACS and reannotated according to a three-level quality grading protocol: Good, Usable, and Reject. This strategy aims to represent intermediate scenarios between images of high and poor quality for clinical analysis, making EyeQ a widely adopted benchmark in RIQA research.

The FQS (Fundus Quality Score) dataset was proposed by Gong et al. (2025) and contains 2,246 fundus images annotated using the same three quality levels adopted in EyeQ, along with a continuous quality score (*Mean Opinion Score* – MOS) ranging from 0 to 100. This dual annotation strategy allows the problem to be addressed as both a classification and a regression task, enabling more precise assessments of image quality.

The FIVES (Fundus Image Vessel Segmentation) dataset (Jin et al., 2022) was primarily developed for retinal vessel segmentation tasks and contains 800 high-resolution color fundus images with annotations of retinal vessels and various ocular pathologies. In addition to the segmentation annotations, the dataset provides image quality assessments across three aspects: illumination or color distortion, blur, and low contrast.

The MSHF (Multi-Source Heterogeneous Fundus) dataset (Jin et al., 2023) was the only public RIQA dataset identified in this study that includes images acquired using a handheld device. The dataset comprises 1,302 images from different acquisition modalities, including conventional fundus images, ultra-widefield images, and 302 images captured with the portable DEC200 camera. This heterogeneity was intentionally introduced to represent diverse clinical scenarios and imaging conditions, contributing to the evaluation of the robustness of image quality assessment methods.

Despite the significant advances represented by these studies, there remains a lack of large-scale, publicly available RIQA datasets specifically designed for handheld fundus cameras.

1.2 Contributions

Our main contributions are as follows:

- To the best of our knowledge, this is the largest public fundus image dataset acquired using a handheld fundus camera for retinal image quality assessment;
- Our dataset contains both color and red-free fundus images and is the first to include both modalities annotated using the same quality grading scheme adopted by EyeQ (Fu et al., 2019) (Good, Usable, and Reject). In addition, it provides labels for the factors responsible for image quality degradation, including low sharpness, underexposure, overexposure, incomplete field of view, peripheral shadowing, and motion blur or camera shake. As a result, the dataset can be used for both multiclass

classification tasks (quality assessment) and multi-label classification tasks (identification of image degradation factors);

- We provide a dedicated test set and establish a benchmark for retinal image quality assessment to facilitate fair comparison among methods.

2. Summary

The proposed M-FIQ dataset contains 7,971 fundus images acquired using an Eyer2 handheld fundus camera, each annotated with quality and degradation-factor labels. Among these images, 2,656 are color fundus images and 5,315 are red-free grayscale fundus images. The original images are provided in JPEG format with a resolution of 1900×1600 pixels. We also provide a preprocessed version in which mask-related artifacts have been removed and the images have been resized to 768×768 pixels.

The objective of this dataset is to provide a collection of images that serves as a benchmark for the development and evaluation of machine learning methods for fundus image quality assessment, with a particular focus on images acquired using handheld devices.

To demonstrate the utility of the dataset, we evaluated the performance of four backbone architectures widely used in the literature—DenseNet121, EfficientNet, ResNet50, and ViT-B16—for quality assessment of both color and red-free fundus images. Among them, ViT-B16 achieved the best performance on the color fundus image test set, reaching an F1-score of 78.44%, whereas ResNet50 achieved the best results on the red-free fundus image test set, reaching an F1-score of 74.10%.

This work aims to fill a gap in publicly available datasets for fundus image quality assessment, as annotated images acquired using handheld devices remain scarce.

3. Resource availability

3.1 Potencial Use Cases

The M-FIQ dataset offers numerous opportunities for research in fundus image quality assessment and the development of artificial intelligence methods for handheld devices. By combining color and red-free fundus images acquired using a handheld fundus camera, along with annotations of both overall image quality and the specific factors responsible for image degradation, the dataset can be explored in a variety of research and clinical application scenarios.

3.1.1 Multiclass Image Quality Classification

The M-FIQ dataset can be used to train and evaluate fundus image quality classification models based on three quality levels (Good, Usable, and Reject), enabling the identification of whether an image is of sufficient quality for clinical assessment or for subsequent processing by computer-aided diagnosis systems.

3.1.2 Multi-Label Classification of Image Degradation Factors

Beyond overall quality classification, M-FIQ enables the development of models capable of simultaneously identifying multiple factors responsible for image degradation, such as low sharpness, underexposure, overexposure, incomplete field of view, peripheral shadowing, and motion blur or camera shake. This approach provides more detailed information about the causes of poor image quality, supporting operator feedback systems during the image acquisition process.

3.1.3 Quality Assessment for Handheld Fundus Cameras and Screening Programs

As it consists exclusively of images acquired using a handheld device, the M-FIQ dataset can be used to train and validate image quality assessment models specifically designed for this type of equipment. The acquisition characteristics represented in the dataset, including variations in illumination, field of view, and artifacts commonly observed in handheld devices, enable the investigation and development of quality assessment methods tailored to the unique challenges of portable fundus imaging.

Models trained on M-FIQ can also be integrated into ophthalmic screening workflows to perform automatic image quality verification at the time of acquisition. Such systems may reduce the number of inadequate examinations forwarded for analysis, decrease the need for image reacquisition, and improve the efficiency of large-scale ocular disease screening programs.

3.1.4 Domain Adaptation Between Clinical and Handheld Fundus Cameras

M-FIQ can also be used in domain adaptation studies, enabling the investigation of how models trained on datasets acquired with clinical fundus cameras can be adapted to images acquired using handheld devices. Likewise, M-FIQ can serve as a target domain for transfer learning techniques, contributing to the development of models that generalize more effectively across different image acquisition conditions.

3.2 Data Location

The M-FIQ dataset is publicly available on an open-source platform to support research in the field and is hosted at https://www.synapse.org//Synapse:syn75868226.

3.3 Licensing

This dataset is released under the Creative Commons Attribution - onCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license. This permits users to share and adapt the dataset for non commercial purposes, provided appropriate credit is given, any changes are indicated, and derivative works are distributed under the same license.

3.4 Ethical Considerations

Ethical approval for the present study was granted by the Brazilian Research Ethics Committee under protocol number CAAE 89498825.7.0000.5086, entitled "New Technologies Applied to Blindness Prevention in the State of Maranhão".

4. Methods

The images were acquired from patients in a quilombola (Afro-Brazilian descendant) community using the Eyer2 handheld fundus camera. Following a filtering stage in which outliers were removed, the quality of each image was assessed by two specialists. The entire process is described in detail in the following sections.

4.1 Data Collection

The images were collected by technicians who had previously received online and hands-on training between September and November 2024 in the municipality of Itapecuru, Maranhão, Brazil. These images were acquired from 304 patients belonging to a quilombola community using the Eyer2 handheld fundus camera (Figure 2-A), developed by Phelcom, configured with a 45° field of view (FOV), following pupil dilation with 1% tropicamide. Figure 2-B illustrates one of the image acquisition sessions conducted in a quilombola community in the state of Maranhão, Brazil.

The original images have a resolution of 1900×1600 pixels and include metadata indicating whether the image corresponds to the left eye (OE) or right eye (OD). In addition, a colored circular label located in the upper-right corner automatically indicates images showing evidence of retinal abnormalities: red denotes a high probability, yellow a moderate probability, and green a low probability. This label is generated by EyerMaps, an artificial intelligence feature integrated into Phelcom's handheld devices, which analyzes the images and produces attention maps highlighting regions associated with these abnormalities. Figure 3 presents some examples of images from the proposed dataset.

After the acquisition of color fundus images, the device automatically generates grayscale versions, referred to as red-free images. Red-free imaging enhances the contrast

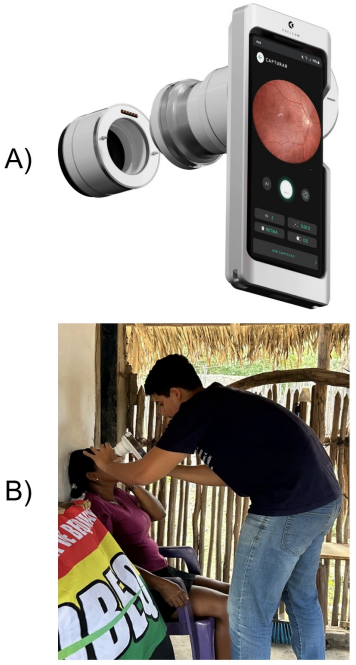

Figure 2: A) Eyer2 Portable Fundus Camera and B) Fundus Image Acquisition.

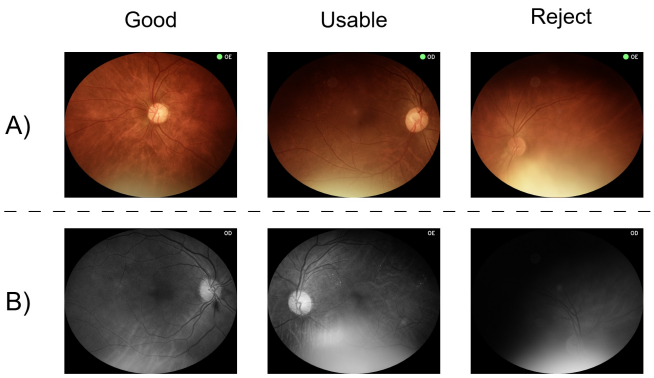

Figure 3: Examples of images included in the proposed dataset along with their corresponding quality labels. A) Color fundus images. B) Red-free fundus images.

of certain retinal structures, such as the retinal nerve fiber layer, facilitating the detection of localized or diffuse defects, particularly in patients with glaucoma. In addition, small vascular lesions may become more visible due to the increased contrast. Retinal folds and surface distortions are also more easily observed in this imaging modality. Red-free images further facilitate the identification of microaneurysms, hemorrhages, and neovascularization in patients with diabetic retinopathy. Because color and red-free fundus images highlight different retinal features, their combined use provides a more comprehensive assessment of the fundus. Consequently, the combined interpretation of both imaging modalities can improve diagnostic accuracy.

Subsequently, all images underwent a manual filtering process in which records that did not correspond to fundus

images, as well as images corresponding to activation maps automatically generated by the device, were removed.

4.2 Annotation Process

A software tool was developed for image quality annotation, enabling the independent labeling of both image quality and the type of degradation present in the image. The quality labels follow the three-level grading scheme first introduced by the EyeQ dataset (Fu et al., 2019): 1) Good, for sharp images in which all important fundus structures (optic disc, macula, and retinal blood vessels) are clearly visible; 2) Usable, for images presenting some irregularities, such as artifacts, blur, or illumination issues, but that still allow a reliable diagnosis; and 3) Reject, for images with severe quality issues that prevent reliable clinical assessment. For images labeled as Usable or Reject, the factors responsible for image degradation were also annotated, including low sharpness, underexposure, overexposure, incomplete field of view, peripheral shadowing, and motion blur or camera shake.

The annotation process was conducted by two ophthalmology specialists. To streamline the workflow, the dataset was partitioned at the patient level, and each patient was assigned to a single specialist. During annotation, the software displayed all fundus images belonging to one patient at a time. After the specialist completed the annotation for that patient, the images of the next assigned patient were loaded automatically. Consequently, every image was annotated by exactly one specialist, while ensuring that all images from the same patient were consistently evaluated by the same annotator. After the annotation phase, all images underwent a review process in which the specialists discussed and established the final label for cases where inconsistencies were identified.

4.3 Data Organization

The dataset was organized into two main directories: `data` for storing the fundus images and `labels` for storing the corresponding annotations.

Within the `data` directory, there are two subdirectories:

- **original**: contains the original fundus images at a resolution of 1900 × 1600 pixels, including both color and red-free grayscale images.
- **processed**: contains preprocessed fundus images with a resolution of 768 × 768 pixels, in which mask-related artifacts have been removed.

The `labels` directory includes three CSV files:

- **train_multiclass_labels.csv** and **test_multiclass_labels.csv**: define the training and test partitions, respectively. Each file contains the following columns:
  - `id`: anonymized patient identifier
  - `image`: directory to the corresponding fundus image
  - `eye`: indicates the eye laterality, specifying whether the image corresponds to the right eye (`OD`) or the left eye (`OE`).
  - `quality`: quality assessment (0 = good, 1 = usable, 2 = reject)
  - `category`: specifies whether the image is color (*color*) or red-free grayscale (*gray*)
- **quality_degradation_labels.csv**: provides detailed annotations regarding image degradation factors. In addition to the `id` and `image` columns, the file includes binary indicators (1 = present, 0 = absent) for the following factors: Low Sharpness, Underexposure, Overexposure, Incomplete Field of View, Peripheral Shadowing, and Motion Blur or Camera Shake.

The M-FIQ dataset contains a total of 7,971 fundus images, divided into 6,239 images for training and 1,732 images for testing. The images were categorized into three quality levels (Good, Usable, and Reject) and two modalities (Color and Red-free grayscale). The training–test split was performed using patient-level stratification based on quality labels to prevent data leakage between partitions while preserving a consistent class distribution across the datasets. The distribution of images across partitions and quality categories is summarized in Table 1.

Table 1: Image distribution by quality and category.

Partition	Quality	Color	Gray	Total
Train	Good	302	146	448
	Usable	772	516	1288
	Reject	1002	3501	4503
	Total	2076	4163	6239
Test	Good	62	42	104
	Usable	189	126	315
	Reject	329	984	1313
	Total	580	1152	1732
Overall	Good	364	188	552
	Usable	961	642	1603
	Reject	1331	4485	5816
	Total	2656	5315	7971

As shown in Table 1, the M-FIQ dataset exhibits a pronounced class imbalance, with the Good category representing the minority class. Out of the 7,971 images, only 552 (6.92%) were labeled as Good, whereas 1,603 (20.11%) were classified as Usable and 5,816 (72.96%) as Reject.

This distribution reflects the real-world conditions under which the dataset was acquired using portable fundus cameras in large-scale screening campaigns, where poor image quality is considerably more frequent than high-quality acquisitions. Consequently, the dataset presents a challenging task for retinal image quality assessment methods, requiring models to effectively learn from highly imbalanced class distributions while accurately distinguishing clinically acceptable images from poor-quality ones.

In addition to the overall quality labels, images classified as Usable or Reject were independently annotated for the presence of specific quality degradation factors. Table 2 summarizes the distribution of the annotated degradation factors across the M-FIQ dataset.

Table 2: Distribution of annotated quality degradation factors in the M-FIQ dataset.

Quality degradation	Images
Low sharpness	3,331
Underexposure	699
Overexposure	17
Incomplete field of view	2,171
Peripheral shadowing	3,241
Motion artifacts	544

5. Technical Validation

For the demonstration of the proposed dataset's applicability, four neural network architectures widely recognized in the literature were employed: DenseNet121 (Huang et al., 2018), EfficientNet-B0 (Tan and Le, 2020), ResNet50 (He et al., 2015), and ViT-B16 (Dosovitskiy et al., 2021). The task consisted of performing fundus image quality classification on color and red-free fundus images using three quality categories: Good, Usable, and Reject.

5.1 Implementation Details

Initially, the images were separated into two modalities: color and red-free fundus imagens. Although both were annotated using the same quality criteria, the two modalities exhibit distinct visual characteristics, which could affect the experimental results if analyzed together. Therefore, all experiments were conducted separately for each modality.

The image preprocessing stage aimed to remove undesired artifacts present in the original image masks, specifically the labels indicating the eye laterality (OD and OE) and the colored circle used to indicate the presence of retinal abnormalities. The removal of these elements was intended to prevent the models from making predictions based on external markers rather than analyzing the relevant retinal structures. Additionally, preprocessing included resizing the images from 1900×1600 pixels to 224×224 pixels with padding as the original images did not have square aspect ratios, which is a requirement for properly feeding the backbone architectures employed in this study.

To mitigate the severe class imbalance in the M-FIQ dataset, a balanced mini-batch sampling strategy was adopted during training. Instead of sampling images according to the original class distribution, each training iteration constructs a mini-batch containing an equal number of samples from each quality class (Good, Usable, and Reject). Specifically, if K samples are selected per class, each mini-batch contains $3K$ images. Since batches are constructed with a fixed number of samples per class, images are reused with replacement whenever necessary, with repetitions occurring more frequently in minority classes due to their limited number of samples. Consequently, training epochs are defined by a fixed number of iterations rather than by a single pass through the dataset, making sample repetition an inherent component of the sampling strategy. This approach ensures that every parameter update is computed from a class-balanced mini-batch, preventing the optimization process from being dominated by the majority class. To further reduce the risk of overfitting caused by this sampling strategy, online data augmentation was employed as a regularization technique. Random rotations and translations were applied during training to increase sample diversity while preserving the original color characteristics of the fundus images. Color-based augmentations were intentionally avoided, as color information is an important cue for retinal image quality assessment and modifying it could alter the quality attributes that the model is intended to learn.

All backbone architectures were trained using the same configuration: the AdamW optimizer, a learning rate of 3×10^{-4}, a weight decay of 0.005, a batch size of 36 (12 per class), and an early stopping strategy configured with a patience of 10 epochs, a minimum of 15 training epochs, and a tolerance of 0.001. Each model was trained for 40 epochs with 580 iterations using 5-fold cross-validation. Due to computational constraints, the batch size was reduced to 24 (8 per class) for the ViT-B16 experiment. The backbone architectures were initialized with weights pre-trained on ImageNet-21k when available, and on the standard ImageNet dataset otherwise.

The experiments were conducted on a workstation equipped with an Nvidia GeForce GTX 1660 GPU with 6 GB of VRAM, an Intel Core i7 processor, and 16 GB of RAM. The algorithms were implemented using the *PyTorch* library (Paszke et al., 2019).

Table 3: Macro-averaged cross-validation results across different backbones using color fundus images from the M-FIQ dataset.

Backbone	F1-Score	Precision	Recall	Specificity
DenseNet121	$\mathbf{0.7947 \pm 0.0132}$	$\mathbf{0.7848 \pm 0.0160}$	$\mathbf{0.8146 \pm 0.0151}$	$\mathbf{0.8989 \pm 0.0075}$
EfficientNet-B0	0.7891 ± 0.0156	0.7806 ± 0.0165	0.8092 ± 0.0154	0.8950 ± 0.0079
ResNet50	0.7856 ± 0.0146	0.7781 ± 0.0137	0.8001 ± 0.0172	0.8923 ± 0.0068
ViT-B16	0.7797 ± 0.0137	0.7684 ± 0.0105	0.8021 ± 0.0174	0.8888 ± 0.0119

5.2 Evaluation

For evaluation, metrics appropriate for the multiclass classification task were computed to capture different aspects of model performance. The considered metrics were F1-score, precision, recall, and specificity, as they provide a comprehensive assessment of the models' ability to correctly distinguish among the three quality levels. Equations 1, 2, 3, and 4 present the formulations of precision, recall, specificity, and F1-score, respectively.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3}$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

To ensure a balanced evaluation across the three classes, all metrics were reported using the macro-averaged approach, that is, each metric was computed independently for each class and subsequently aggregated using the arithmetic mean.

5.3 Results and Discussion

5.3.1 Color fundus image quality assessment results

The macro-averaged cross-validation results for the evaluated backbone architectures using color fundus images from the M-FIQ dataset are summarized in Table 3. DenseNet121 achieved the best performance across all evaluation metrics, particularly in terms of F1-score (0.7947) and recall (0.8146). This behavior suggests that the architecture was particularly effective at capturing relevant patterns for the task, providing a favorable balance between precision and recall. The remaining architectures achieved comparable but consistently lower results, with ViT-B16 exhibiting the weakest performance across nearly all metrics.

For the test set evaluation, after the cross-validation stage, an ensemble was constructed by selecting the three best-performing folds for each backbone based on the F1-score. The final quality classification was obtained by averaging the outputs of the three selected folds.

On the independent test set (Table 4), ViT-B16 achieved the best performance across all evaluation metrics, obtaining an F1-score of 0.7844 and a specificity of 0.8983. ResNet50 ranked second in terms of F1-score and precision, whereas DenseNet121, which achieved the best cross-validation performance, ranked third on the test set. Although this change in ranking differs from the cross-validation results, the performance differences among the evaluated backbones during cross-validation were relatively small and accompanied by overlapping standard deviations. Moreover, the test evaluation was conducted on a single independent partition containing 580 images, which may introduce additional variability in the observed ranking. Overall, these results suggest that the evaluated architectures exhibit comparable performance, while ViT-B16, despite its lower average cross-validation results, demonstrated the strongest generalization on the selected test set.

Table 4: Performance of ensembles on the M-FIQ color fundus image test set using macro-averaged metrics. Values in bold indicate the best performance per metric, and underlined values correspond to the second-best.

Ensemble	F1	Prec	Recall	Spec
DenseNet121	0.7477	0.7330	<u>0.7790</u>	0.8863
EfficientNet-B0	0.7454	0.7300	0.7787	<u>0.8929</u>
ResNet50	<u>0.7554</u>	<u>0.7473</u>	0.7707	0.8908
ViT-B16	**0.7844**	**0.7724**	**0.8021**	**0.8983**

Table 5 presents the per-class performance of the evaluated backbone ensembles on the M-FIQ color fundus image test set. ViT-B16 achieved the highest F1-score across all quality classes, demonstrating the most consistent performance among the evaluated backbones. ResNet50 generally ranked second, while DenseNet121 and EfficientNet-B0 exhibited less balanced results.

Figure 4 presents the confusion matrix of ViT-B16 ensemble on the color fundus image test set of M-FIQ dataset. It can be observed that the primary challenge for the model lies in distinguishing between the Usable and Reject classes. Despite the lower representation of the Good class in the training set, the model demonstrated strong learning capa-

bility for this category, making only 13 misclassifications among the 62 available images. Notably, no Good quality image was incorrectly classified as Reject, which would be particularly critical in a real-world deployment scenario.

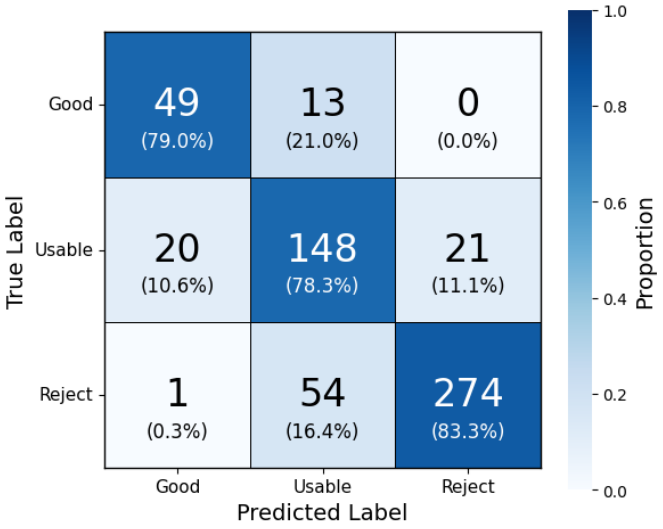

Figure 4: Confusion matrix of the ViT-B16 ensemble on the M-FIQ color fundus image test set.

5.3.2 *Red-free fundus image quality assessment results*

Table 6 summarizes the macro-averaged cross-validation results obtained by the evaluated backbones using red-free fundus images from the M-FIQ dataset. EfficientNet-B0 achieved the best overall performance, obtaining the highest F1-score (0.7794), while also achieving the highest precision (0.7522). ResNet50 yielded the highest recall (0.8279) and specificity (0.9349), indicating a better ability to correctly identify each quality class. In contrast, ViT-B16 presented the lowest overall performance, with an F1-score of 0.7410, suggesting that its transformer-based architecture was less effective than the CNN-based backbones for the red-free fundus image quality assessment task.

Table 5: Per-class performance comparison of ensembles on the M-FIQ color fundus image test set. Values in bold indicate the best performance per metric, and underlined values correspond to the second-best.

Class	Ensemble	F1	Prec	Recall	Spec
Good	DenseNet121	0.7000	0.6282	**0.7903**	0.9440
	EfficientNet-B0	0.6713	0.5926	0.7742	0.9363
	ResNet50	0.6923	0.6618	0.7258	0.9556
	ViT-B16	**0.7424**	**0.7000**	**0.7903**	**0.9595**
Usable	DenseNet121	0.6983	0.6336	0.7778	0.7826
	EfficientNet-B0	0.7033	0.6419	0.7778	0.7903
	ResNet50	0.7101	0.6533	0.7778	0.8005
	ViT-B16	**0.7327**	**0.6884**	**0.7831**	**0.8286**
Reject	DenseNet121	0.8447	0.9370	0.7690	0.9323
	EfficientNet-B0	0.8614	**0.9556**	0.7842	**0.9522**
	ResNet50	0.8636	0.9268	0.8085	0.9163
	ViT-B16	**0.8782**	0.9288	**0.8328**	0.9163

As in the color fundus image experiments, the test set evaluation was performed using an ensemble strategy. After the cross-validation stage, the three best-performing folds for each backbone, based on the F1-score, were selected to form the ensemble. The final quality prediction was obtained by averaging the output probabilities of the selected models.

Table 7 presents the performance of the ensemble models on the M-FIQ red-free fundus image test set. ResNet50 achieved the best overall performance, obtaining the highest F1-score (0.7410), while also exhibiting competitive precision (0.7429) and specificity (0.9025). ViT-B16 achieved the highest recall (0.7408) and specificity (0.9250), whereas EfficientNet-B0 obtained the highest precision (0.7508) and the second-best F1-score (0.7367). In contrast, DenseNet121 yielded the lowest overall performance, with an F1-score of 0.6836. Although EfficientNet-B0 achieved the highest F1-score during cross-validation, ResNet50 demonstrated better generalization on the independent test set. Conversely, ViT-B16, which showed the lowest cross-validation performance, improved considerably on the test set, whereas DenseNet121 experienced the largest performance degradation.

Table 7: Performance of ensembles on the M-FIQ red-free fundus image test set using macro-averaged metrics. Values in bold indicate the best performance per metric, and underlined values correspond to the second-best.

Ensemble	F1	Prec	Recall	Spec
DenseNet121	0.6836	0.6893	0.6813	0.8778
EfficientNet-B0	0.7367	**0.7508**	0.7243	0.9034
ResNet50	**0.7410**	0.7429	0.7391	0.9025
ViT-B16	0.7283	0.7438	**0.7408**	**0.9250**

Table 8 presents the per-class performance of the evaluated backbone ensembles on the M-FIQ red-free fundus image test set. Unlike the color fundus image experiments, the results were more evenly distributed across the evaluated backbones, with no single model consistently outperforming the others in all quality classes. ResNet50 achieved the highest F1-score for the Good class, ViT-B16 for the Usable class, and EfficientNet-B0 for the Reject class, whereas DenseNet121 generally yielded the lowest F1-scores across the evaluated quality classes.

Figure 5 presents the confusion matrix of the ResNet50 ensemble on the M-FIQ red-free fundus image test set. The results indicate that the primary source of misclassification occurred in the Usable class, with a considerable number of images being confused with both Good and Reject. In contrast, the model exhibited excellent discrimination be-

Table 6: Macro-averaged cross-validation results across different backbones using red-free fundus images from the M-FIQ dataset.

Backbone	F1-Score	Precision	Recall	Specificity
DenseNet121	0.7766 ± 0.0179	0.7501 ± 0.0288	0.8199 ± 0.0254	0.9292 ± 0.0062
EfficientNet-B0	$\mathbf{0.7794 \pm 0.0192}$	$\mathbf{0.7522 \pm 0.0348}$	0.8179 ± 0.0051	0.9342 ± 0.0097
ResNet50	0.7755 ± 0.0319	0.7419 ± 0.0374	$\mathbf{0.8279 \pm 0.0233}$	$\mathbf{0.9349 \pm 0.0058}$
ViT-B16	0.7410 ± 0.0331	0.7094 ± 0.0439	0.8060 ± 0.0197	0.9314 ± 0.0107

Table 8: Per-class performance comparison of ensembles on the M-FIQ red-free fundus image test set. Values in bold indicate the best performance per metric, and underlined values correspond to the second-best.

Class	Ensemble	F1	Prec	Recall	Spec
Good	DenseNet121	0.5682	0.5435	0.5952	0.9811
	EfficientNet-B0	0.6329	0.6757	0.5952	0.9892
	ResNet50	**0.6667**	0.6667	**0.6667**	0.9874
	ViT-B16	0.6111	**0.7333**	0.5238	**0.9928**
Usable	DenseNet121	0.5259	0.5755	0.4841	**0.9561**
	EfficientNet-B0	0.6135	**0.6160**	0.6111	0.9532
	ResNet50	0.5944	0.6016	0.5873	0.9522
	ViT-B16	**0.6218**	0.5215	**0.7698**	0.9133
Reject	DenseNet121	0.9567	0.9490	0.9644	0.6964
	EfficientNet-B0	**0.9635**	0.9606	**0.9665**	0.7679
	ResNet50	0.9619	0.9605	0.9634	0.7679
	ViT-B16	0.9521	**0.9765**	0.9289	**0.8690**

tween the extreme quality classes, as no Reject images were misclassified as Good, and only one out of 42 Good images were classified as Reject, indicating that confusion between the extreme quality classes was minimal.

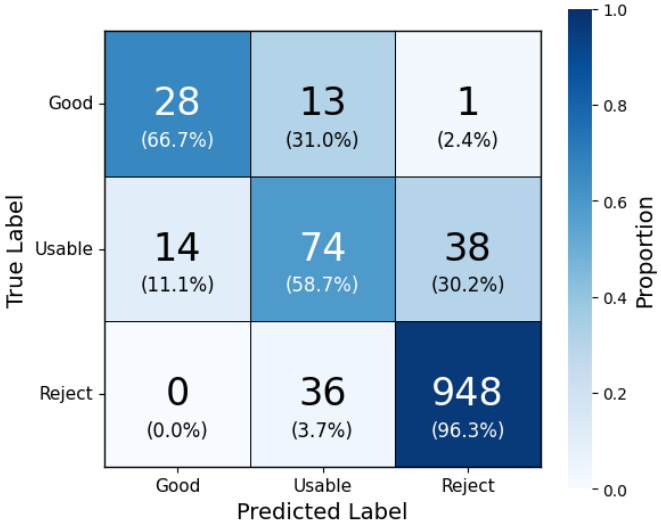

Figure 5: Confusion matrix of the ResNet50 ensemble on the M-FIQ red-free fundus image test set.

6. Conclusion

This study presents the M-FIQ, a new public database for retinal image quality assessment acquired with the Eyer2 portable fundus camera. The dataset contains 7,971 fundus images, including both color and red-free images, annotated at three quality levels (Good, Usable, and Rejected), as well as labels for the factors responsible for image degradation.

As part of our technical validation, we evaluated four deep learning architectures widely used in the literature, establishing an initial benchmark for fundus image quality assessment in the M-FIQ dataset. The results demonstrate that M-FIQ represents a challenging and suitable scenario for the development of methods for the automatic evaluation of image quality.

By making available one of the largest public datasets dedicated to evaluating the quality of retinal images obtained with portable devices, including those from the quilombola population, we hope to contribute to research on artificial intelligence applied to ophthalmology, particularly in the areas of screening, teleophthalmology, and eye disease detection.

Acknowledgments

This work was supported by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)* —Finance Code 001, the *Fundação de Amparo à Pesquisa e ao Desenvolvimento Científico e Tecnológico do Maranhão (FAPEMA)*(Gra number APP-02680/25), and the *Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)*(Grant number 406497/2024-9 and 305253/2025-5)

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, complying with all applicable laws and regulations regarding the treatment of human subjects. Ethical approval for the present study was granted by the Brazilian Research Ethics Committee under protocol number CAAE 89498825.7.0000.5086, entitled "New Technologies Applied to Blindness Prevention

in the State of Maranhão".

Conflicts of Interest

The authors declare they don't have conflicts of interest.

Data availability

The dataset is publicly available through our dedicated repository at

Synapse: https://www.synapse.org/Synapse:syn75868226.

DOI: 10.7303/SYN75868226

References

Or Abramovich, Hadas Pizem, Jan Van Eijgen, Ilan Oren, Joshua Melamed, Ingeborg Stalmans, Eytan Z. Blumenthal, and Joachim A. Behar. Fundusq-net: A regression quality assessment deep learning algorithm for fundus images quality grading. *Computer Methods and Programs in Biomedicine*, 239:107522, 2023. ISSN 0169-2607. . URL https://www.sciencedirect.com/science/article/pii/S0169260723001876.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Huazhu Fu, Boyang Wang, Jianbing Shen, Shanshan Cui, Yanwu Xu, Jiang Liu, and Ling Shao. *Evaluation of Retinal Image Quality Assessment Networks in Different Color-Spaces*, page 48–56. Springer International Publishing, 2019. ISBN 9783030322397. . URL http://dx.doi.org/10.1007/978-3-030-32239-7_6.

Zheng Gong, Zhuo Deng, Run Gan, Zhiyuan Niu, Lu Chen, Canfeng Huang, Jia Liang, Weihao Gao, Fang Li, Shaochong Zhang, and Lan Ma. Acquire continuous and precise score for fundus image quality assessment: Fthnet and fqs dataset. *Scientific Reports*, 15(1):40524, November 2025. ISSN 2045-2322. . URL https://doi.org/10.1038/s41598-025-24423-8.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018. URL https://arxiv.org/abs/1608.06993.

Kai Jin, Xingru Huang, Jingxing Zhou, Yunxiang Li, Yan Yan, Yibao Sun, Qianni Zhang, Yaqi Wang, and Juan Ye. Fives: A fundus image dataset for artificial intelligence based vessel segmentation. *Scientific Data*, 9(1):475, August 2022. ISSN 2052-4463. . URL https://doi.org/10.1038/s41597-022-01564-3.

Kai Jin, Zhiyuan Gao, Xiaoyu Jiang, Yaqi Wang, Xiaoyu Ma, Yunxiang Li, and Juan Ye. Mshf: A multi-source heterogeneous fundus (mshf) dataset for image quality assessment. *Scientific Data*, 10(1):286, May 2023. ISSN 2052-4463. . URL https://doi.org/10.1038/s41597-023-02188-x.

Lie Ju, Xin Wang, Quan Zhou, Hu Zhu, Mehrtash Harandi, Paul Bonnington, Tom Drummond, and Zongyuan Ge. Bridge the domain gap between ultra-wide-field and traditional fundus images via adversarial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. arXiv:2003.10042v2.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. URL https://arxiv.org/abs/1912.01703.

Aditya Raj, Anil Kumar Tiwari, and Maria G. Martini. Fundus image quality assessment: survey, challenges, and future scope. *IET Image Processing*, 13(7):1211–1224, 2019. . URL https://doi.org/10.1049/iet-ipr.2018.6212.

Ugur Sevik. Drimdb (diabetic retinopathy images database) database for quality testing of retinal images, 04 2014.

Yaxin Shen, Bin Sheng, Ruogu Fang, Huating Li, Ling Dai, Skylar Stolte, Jing Qin, Weiping Jia, and Dinggang Shen. Domain-invariant interpretable fundus image quality assessment. *Medical Image Analysis*, 61:101654, 2020. .

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020. URL https://arxiv.org/abs/1905.11946.

Gabriel Tozatto Zago, Rodrigo Varejão Andreão, Bernadette Dorizzi, and Evandro Ottoni Teatini

Salles. Retinal image quality assessment using deep learning. *Computers in Biology and Medicine*, 103: 64–70, 2018. .