IAES International Journal of Robotics and Automation (IJRA)
Vol. 15, No. 1, March 2026, pp. 234~246
ISSN: 2722-2586, DOI: 10.11591/ijra.v15i1.pp234-246

EdgeRetina: Hybrid multimedia architecture for diabetic retinopathy screening on low-cost mobiles

Guidoum Amina[1], Achour Soltana[1], Maamar Bougherara[1], Amara Rafik[1], Mhamed Tayeb[2]
[1]Department of Computer Science, Higher Normal School of Kouba, Algiers, Algeria
[2]Institute Biomaterials and Transport Phenomena Laboratory, Medea University, Medea, Algeria

Article Info

Article history:

Received Jul 26, 2025
Revised Dec 4, 2025
Accepted Feb 9, 2026

Keywords:

Compression
Dynamic threshold
EdgeRetina
Low-cost mobiles
Preprocessing

ABSTRACT

Diabetic retinopathy (DR) is a major cause of preventable blindness, particularly in areas with limited medical resources where access to ophthalmologists is critical. Existing automated solutions struggle to balance clinical performance, cost-effectiveness, and robustness in the face of fundus image variability—including lighting differences, artifacts, and uneven capture quality. To address this challenge, we propose EdgeRetina, an integrated solution for diabetic retinopathy screening on low-cost mobiles. Our approach combines lightweight preprocessing (128×128 resizing, intensity normalization, and targeted augmentations simulating real-world conditions) with a hybrid SqueezeNet-MobileViT architecture (1.4 million parameters), optimized by dynamic threshold calibration (median: 0.3), maximizing clinical utility. Clinically calibrated INT8 quantization reduces the model to 8.27 MB (-92%) without altering diagnostic performance (sensitivity of 90.7%for referable diabetic retinopathies), while preserving compatibility with floating point 32 (FP32)-based gradient-weighted class activation mapping (Grad-CAM) visualizations. Evaluated on the APTOS 2019dataset, this solution achieves an AUC of 0.96 with a latency (inference time) of 15.43 ms, reducing CPU consumption by 43% compared to FP32. The dynamic threshold/INT8 coupling decreases false positives by 71.4%. This pipeline thus enables accurate, accessible, and early screening of diabetic retinopathy on low-cost mobile devices, combining operational efficiency and diagnostic reliability in constrained environments, which is crucial to prevent avoidable blindness.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Guidoum Amina
Department of Computer Science, Higher Normal School of Kouba
Algiers, Algeria
Email: amina.guidoum@g.ens-kouba.dz

1. INTRODUCTION

Diabetic retinopathy (DR) is a devastating complication of diabetes, affecting more than 140 million people worldwide and constituting the leading cause of preventable blindness in working-age adults. This problem is particularly acute in resource-limited settings, where access to ophthalmologists is critical (<1 specialist per 1 million inhabitants in sub-Saharan Africa) and where 80% of diabetes-related blindness could be prevented through early detection [1]. Also considered as a common complication of diabetes mellitus, manifesting as retinal damage that can lead to vision loss. Late diagnosis of this condition can lead to blindness, while early detection and treatment can significantly reduce the risk of irreversible vision damage.

Manual diagnosis of fundus images by ophthalmologists is often costly, time-consuming, and demanding, unlike computer-aided diagnostic systems. In recent years, medical image analysis and

classification have seen major advances thanks to deep learning. Among these approaches, convolutional neural networks (CNNs) have emerged as the standard tools for medical image analysis, providing automated feature extraction and classification capabilities.

As highlighted in [2], multimedia processing naturally integrates image processing as one of its fundamental components. Their work demonstrates that deep learning techniques are now the preferred approach for image processing in the multimedia field, enabling significant advances in image recognition, semantic segmentation and image synthesis. This conceptual hierarchy - where deep learning is part of image processing, itself a branch of multimedia processing - establishes the theoretical framework within which our EdgeRetina approach for diabetic retinopathy screening is embedded.

Existing AI solutions suffer from two critical limitations: over-parameterized architectures (>10M parameters) [3], [4] requiring complex pre-processing incompatible with mobile devices, and inappropriate clinical trade-offs generating either excessive false positives (overdiagnosis) or dangerous false negatives (under-detection of referable cases). To address this dual challenge, we propose EdgeRetina, an integrated framework combining: a minimalist pre-processing (128×128 resizing, intensity normalization, targeted augmentations simulating field constraints), a hybrid SqueezeNet-MobileViT architecture (1.4 million parameters) combining local compression and contextual modeling, and a dynamic calibration of the decision threshold (0.3) optimizing clinical utility via medically validated INT8 (8 bits integer) quantization [20]. Evaluated on the APTOS 2019 benchmark, our solution demonstrates exceptional performance (AUC=0.96, latency=15.43 ms) with a sensitivity of 90.7% for referable DRs and a 71.4% reduction in false positives, while enabling low-cost mobile deployment thanks to model compression to 8.27 MB (-92%), a 43% reduction in CPU consumption, and the preservation of gradient-weighted class activation mapping **(**Grad-CAM) explainable visualizations floating point 32 (FP32). This approach validates a new screening paradigm accessible for constrained environments, paving the way for massive prevention campaigns.

Gaur *et al*. [5] proposed an automated method for detecting diabetic DR and classifying its different stages from retinal images using a CNN, more precisely the DenseNet169 architecture. The model is trained on the APTOS 2019 dataset containing approximately 13,000 fundus images annotated according to five levels of DR severity (from 0: no DR to 4: proliferative DR). The images undergo preprocessing (resizing, normalization, encoding) before feature extraction and classification via DenseNet169. The model achieves an accuracy of 82% for classification into five classes, and 98% for binary detection (presence or absence of DR), outperforming other methods such as classical CNN and XGBoost.

Sushith *et al*. [6] proposed a novel approach for the early detection of DR, designing a hybrid deep learning model, called temporal aware hybrid deep learning (TAHDL), combining CNNs for spatial feature extraction and recurrent neural networks (RNNs)—with attention mechanisms—to analyze temporal dependencies between successive retinal images. This model leverages the temporal progression of the disease to more accurately detect the early signs of DR. The use of advanced preprocessing techniques such as color adaptive histogram equalization (CLAHE), normalization, and data augmentation helps improve detection quality. The model was evaluated using public benchmark datasets, including DRIVE, Kaggle Diabetic Retinopathy, and EyePACS, and demonstrated superior accuracy compared to conventional approaches (CNN, VGG19, InceptionV3, MobileNetV3, and ViT), achieving up to 97.5% accuracy on DRIVE, 94.04% on Kaggle, and 96.9% on EyePACS.

Nandhini *et al*. [7] proposed a method for detecting diabetic retinopathy based on a model called DiaNet model (DNM). During the preprocessing stage, a Gabor filter is used to enhance the visibility of blood vessels in retinal images. This filter also contributes to texture analysis, object recognition, feature extraction, and image compression. During the data augmentation stage, the input dimensions of the dataset are reduced using principal component analysis (PCA), which reduces the number of features to be processed without compromising model performance. An average accuracy of 90.02% was achieved.

Saproo *et al*. [8] proposed a deep learning binary classification system, based on transfer learning, for early detection of DR on retinal images. It combines three robust databases (EyePACS, IDRiD, APTOS-2019) annotated by ophthalmologists and applies preprocessing including denoising, normalization and data augmentation to improve robustness. After evaluating 20 pre-trained networks (Serial, DAG, lightweight categories) via comprehensive metrics (accuracy, sensitivity, and AUC-ROC), it demonstrates that the ResNet101 (DAG) model achieves a good result.

Baskar *et al*. [9] showed that diabetic retinopathy, a prevalent ocular pathology affecting retinal vessels in diabetics, affects approximately 3.9 million people worldwide, highlighting the urgent need for early diagnosis for effective management. It demonstrates how deep learning, particularly transfer learning, can classify the five stages of the disease (normal, mild, moderate, severe and proliferative) via AlexNet and DenseNet-169 architectures trained on the APTOS2019 and diabetic retinopathy competition databases. After fine-tuning on 20,163 images (9,000 normal, 2,808 mild, 6,287 moderate, 1,065 severe, 1,003 proliferative) and validation on 2,017 images (900 normal, 281 mild, 629 moderate, 107 severe, 100 proliferative),

DenseNet-169 proved superior, with F1-scores of 0.55 (normal), 0.38 (mild), 0.40 (moderate), 0.60 (severe), and 0.69 (proliferative).

Kumar *et al*. [10] proposed an approach combining two advanced deep learning architectures: InceptionV3 and ResNet50. The InceptionV3 model enables multi-scale feature extraction, making it particularly effective at detecting both small lesions such as microaneurysms and large abnormalities. ResNet50, for its part, facilitates the training of deep networks by avoiding the vanishing gradient problem through skipping connections between layers. The results indicate that InceptionV3 achieves an accuracy of 95%, while ResNet50 achieves 94%, these metrics provide a comprehensive assessment of the models' diagnostic capabilities across different performance dimensions.

Dasari *et al*. [11] focused on the automatic assessment of the severity of diabetic retinopathy using a transfer learning approach. Specifically, a pre-trained and then fine-tuned ResNet50 model was applied to the APTOS2019 dataset, taking into account challenges related to medical annotation and privacy constraints. Tisha *et al*. [12] proposed an optimized model for the accurate classification of retinal disorders, including diabetic retinopathy-a common complication of diabetes mellitus that causes potentially blinding retinal lesions in the absence of adequate screening and treatment. To this end, they developed an approach combining deep learning and attention mechanisms.

Several pre-trained architectures (ResNet152, EfficientNetB7, MobileNetV3Large) were evaluated by the authors on a diverse dataset of retinal images. Their work demonstrates that the improved version of ResNet152, incorporating a spatial attention module, achieves (accuracy: 87.65%, precision: 89.88%, recall: 90.69%, F1-score: 88.58%). The authors highlight the competitiveness of their model through a comparative analysis with previous work. Their research contributes to advances in the automated diagnosis of retinal pathologies and could improve patient care as well as diagnostic accuracy in ophthalmology.

Matthew *et al*. [13] presented a system using machine learning was developed to classify diabetic retinopathy using transfer learning, based on the EfficientNet-B0 model. This model was integrated into an Android mobile application designed to enable healthcare professionals to perform a diagnosis using a simple smartphone, a 20D lens, and a few basic medications, in areas lacking adequate medical infrastructure. The results obtained show that the EfficientNet-B0 model achieves an accuracy of 91.85% for the classification of the three disease categories: no DR, non-proliferative DR, and proliferative DR.

Zhao *et al*. [14] presents a review of methods for deploying deep learning on mobile devices, highlighting their benefits in terms of data privacy and operational efficiency. It describes an optimization pipeline combining model-oriented techniques (such as quantization) and hardware/software mechanisms.

Critical analysis: No previous work has combined dynamic thresholding, explainability, and INT8 quantization for DR. existing solutions are over-parameterized, require complex pre-processing, and lack appropriate clinical trade-offs. We now explicitly state that EdgeRetina addresses these gaps through the innovative integration of lightweight hybrid architecture, dynamic threshold calibration, and INT8 quantification, specifically designed for low-cost mobile deployment.

2. METHOD

In this study, we propose EdgeRetina, a deep learning framework optimized for the detection of referable diabetic retinopathy. The EdgeRetina framework integrates lightweight preprocessing, a hybrid SqueezeNet-MobileViT architecture, dynamic threshold calibration, and INT8 quantization to enable real-time diabetic retinopathy screening on low-cost mobile devices. Retinal images are extracted from the public APTOS 2019 dataset [15]. The collected images undergo a minimalist preprocessing including resizing to 128×128 pixels, intensity normalization, and targeted augmentation (random horizontal flips and ±10% contrast variations) to enhance robustness to field conditions. The optimized images are then processed by an innovative hybrid architecture combining SqueezeNet Fire modules (efficient local feature extraction) and a MobileViT block (contextual modeling by attention mechanism on 16×16 patches). Downstream, a dynamic calibration of the decision threshold (optimized to 0.3 via F1-score maximization) and a medically validated INT8 quantization are applied to enable embedded deployment. The complete flow of the EdgeRetina methodology is illustrated in Figure 1.

2.1. Dataset and model architecture

The dataset used is APTOS 2019 contains 3,662 retinal images annotated by ophthalmological experts according to the ETDRS severity scale: Class 0: No diabetic retinopathy (DR), Class 1: Mild DR, Class 2: Moderate DR, Class 3: Severe DR, Class 4: Proliferative DR.

To adapt to the clinical need for early detection of treatable cases, we perform a binary grouping of classes: Conversion of the 5 stages of diabetic retinopathy (DR) into a binary problem: class 0: Healthy/mild

stages (grades 0-1): without RD 2595 images and class 1: Moderate/severe/proliferative stages (grade 2-4) with RD 1,067 images.

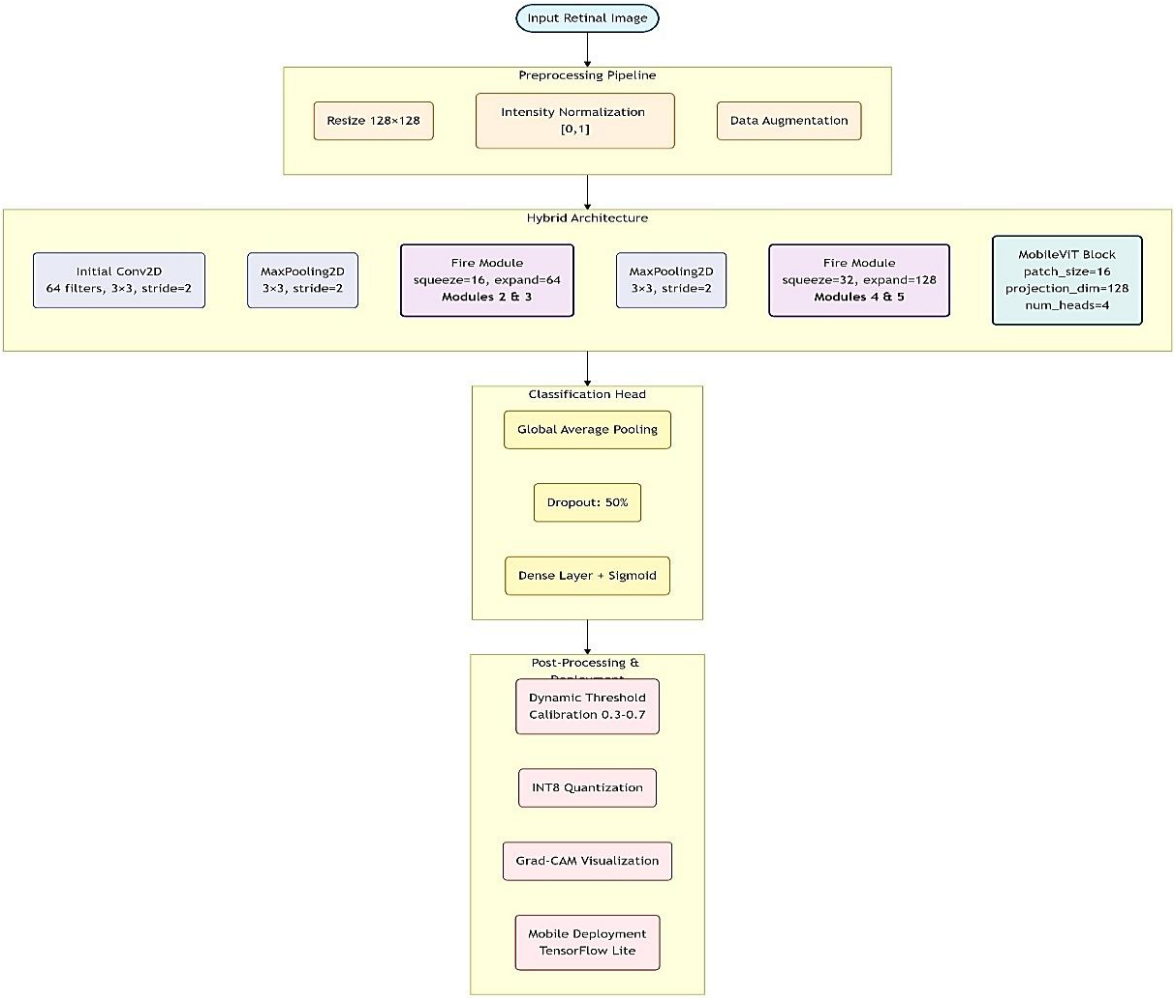

Figure 1. EdgeRetina architecture and system flow diagram

We propose an innovative EdgeRetina architecture (Figure 1) for medical image classification combining the parametric efficiency of Fire modules with the global contextual modeling of transformers. The network processes 128×128 retinal images through four Fire modules, progressively extracting hierarchical features. A MobileViT block inserted after `fire5_concat` processes 16×16 patches by multi-head self-attention (4 heads, 128-dim projection). The transformer output is resampled by bilinear upsampling and fused via a residual connection. A global average pooling and a dropout (50%) precede the final classification layer.

- SqueezeNet [16] is used for its lightweight design via "Fire" modules (1x1 compression and 3×3 expansive convolutions).
- A MobileViT [17] block is integrated downstream to capture global dependencies using multi-head self-attention on image patches (16x16 pixels).
- The output is dropout regularized (50%) before binary classification using a dense sigmoid layer.

$$squeezed = \sigma(w_1 * x + b_1) \ \ where: w\ is\ a\ \ 1 \times 1 filter, \sigma\ relu\ function \quad (1)$$

$$expand1 = \sigma(w_2 * squeezed + b_2) \quad we\ have\ convolution\ 1 \times 1 \quad (2)$$

$$expand3 = \sigma(w_3 * squeezed + b_3) \quad we\ have\ convolution\ 3 \times 3 \quad (3)$$

$$output = expand1 \oplus expand3 \ \ here\ we\ have \ \ concatenation \quad (4)$$

The Fire module (squeezeNet) proceeds in three phases: as in (1) a 1×1 convolution layer ("squeeze") reduces the dimensionality of the features, followed by an expansion step ("Expand") applying 1×1 and 3×3 convolutions in parallel to capture multi-scale features as in (2) and (3). The outputs of the two branches are merged by concatenation ("concat") as in (4).

$$P = extract\_patches(X, size = 16 \times 16, stride = 16) \quad where\ X\ is\ the\ image$$
$$Z = W_p \times P + b_p \tag{5}$$

Here, P is matrix of extracted patches (16×16 pixels)

The MobileViT block starts with a 16×16 patch extraction decomposing the image into spatial units as in (5). These patches undergo a linear projection into a latent space Z before being processed by a multi-head attention mechanism capturing global dependencies. A spatial reconstruction by bilinear upsampling then restores the original resolution, while a residual connection combines these transformed features with the input features.

$$\tau * = argmax\ F_1(y_{true}, y_{predicted}) \qquad where\ \tau\ is\ the\ threshold, F1{:}\ F1_{score} \tag{6}$$
$$y_{true} = 1\ if\ f(x) \geq \tau$$
$$y_{true} = 0\ else$$

Here, y_predicted is the probabilities predicted by the model and y_true is the true labels (ground truth).

Three optimization techniques were employed as shown in Table 1.

- Data augmentation, including random horizontal flipping and ±10% contrast variation.
- Class weighting with a weight of 3 for the minority class (RD) to correct for imbalance.
- Dynamic optimization of the classification threshold recalibrated at each epoch during training, where a callback evaluates 50 thresholds (between 0.3 and 0.7) on the validation set and retains the value maximizing the F1-score as in (6). Although the optimal threshold varies between 0.3 and 0.69 across epochs, the median value of 0.3 was retained for its clinical balance between sensitivity and specificity.

Table 1. Training parameters

Parameter	Value	Description
img_size	(128, 128)	Input image resolution
batch_size	32	Number of samples per gradient update
epochs	50	Number of training iterations
val_split	0.15	Fraction of data reserved for validation
class_weights	{0: 1.0, 1: 3.0}	Weighting to handle class imbalance
optimizer.learning_rate	1e-3	Adam optimizer learning rate
augmentation.random_flip	True	Enable random horizontal flipping
augmentation.random_contrast	0.1	Contrast variation range (±10%)

The embedded deployment pipeline implements post-training INT8 quantization [18], [19], reducing the model size. INT8 quantization, was applied using a four-step approach. The Keras model [18] was first converted to TensorFlow Lite format [20] with standard optimizations enabled. Dynamic calibration was then performed on 100 batches of validation images to fine-tune the quantization parameters. The process-maintained compatibility with mobile processors thanks to support for standard and optimized TensorFlow operators. Finally, automatic input and output conversion mechanisms were implemented to manage the transition between numeric formats during inference.

The evaluation framework measures: i) Clinical performance via (AUC [21], F1-score, precision and recall) [22], [23] at the optimal threshold; ii) Explainability via Grad-CAM heatmaps [24] localizing pathological regions; iii) ROC curves [25] and confusion matrices for FP32 (Keras) and INT8 (TensorFlow) models; iv) Operational performance via latency and CPU utilization measured on an Intel Core i7-9750H (6 cores, 2.6 GHz) system without hardware acceleration. Per-frame latency (defined as the interval between the start of input data transfer and the completion of predictions) is quantified with microsecond precision via time.perf_counter() and then normalized per batch. CPU utilization is calculated with the psutil library via snapshots before/after each inference, with the final value representing the weighted arithmetic mean over the full validation set (549 samples). Model size is measured using a standardized methodology: .keras file for FP32 (tf.keras.models.save_model()) and .tflite file for INT8 post-quantization. This approach integrates all the components necessary for deployment (architecture, weights, operational metadata) for a realistic measurement of the memory footprint.

Figure 2 implements the Grad-CAM method according to four key principles: i) exclusive use of the FP32 model for interpretability, ii) targeting the fire5_concat layer (last convolutional output), iii) computation of gradient-weighted activation maps, and iv) generation of clinically interpretable heatmaps for DR detection. Systematic analysis confirms that activated regions correspond to relevant anatomical structures (macula, retinal vessels) and characteristic lesions of diabetic retinopathy, without formal clinical validation by ophthalmologists.

The architectural design of the EdgeRetina framework facilitates its deployment within robotic and automated screening systems. This integration relies on a simplified, closed-loop workflow, transforming a robotic platform into an autonomous diagnostic unit. The process begins with the acquisition of retinal images via a robotic fundus camera. These images are then processed by the EdgeRetina pipeline for immediate analysis. The classification results directly feed into the system's decision logic, triggering predefined robotic actions. These actions can range from generating an alert for a specialist and providing feedback to the patient, to autonomously initiating further imaging examinations. This end-to-end workflow leverages the model's computational efficiency to enable actionable diagnoses in real time, at the patient's bedside.

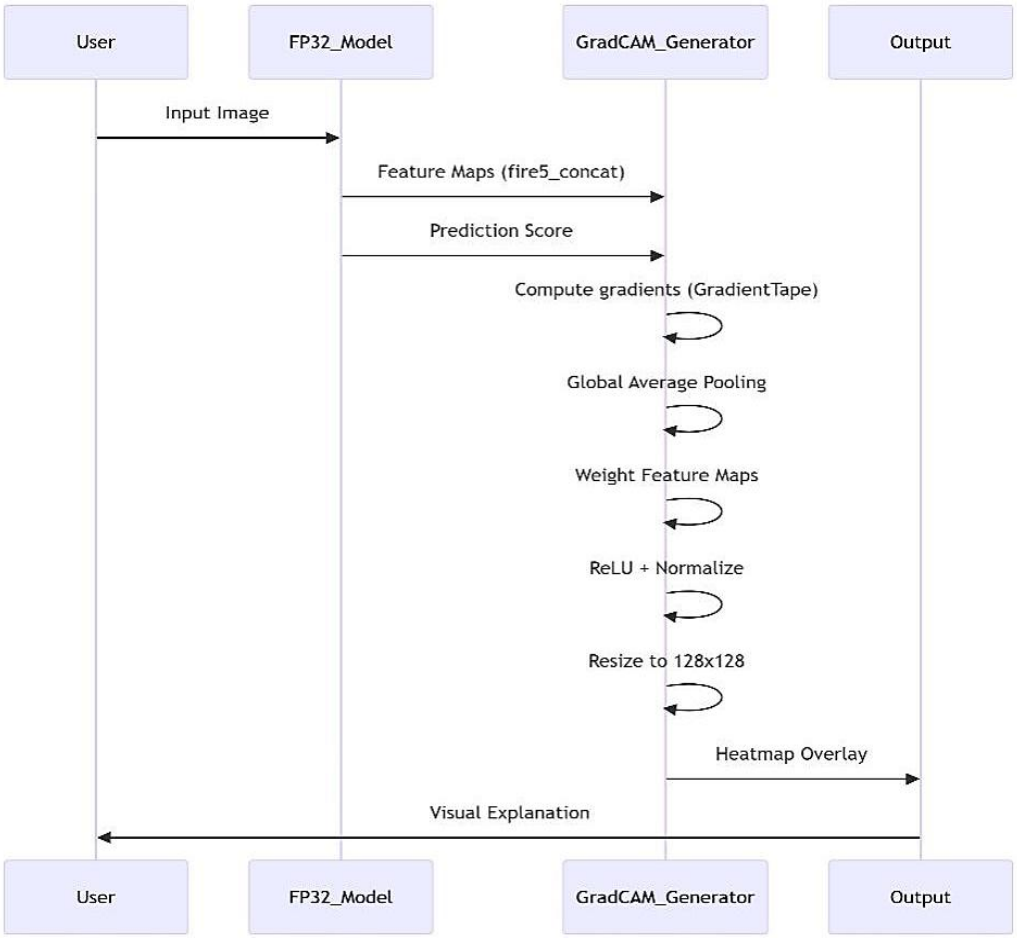

Figure 2. Grad_cam workflow

3. RESULTS AND DISCUSSION

The experimental results of EdgeRetina are presented on the APTOS 2019 validation set (n=549), with a binary distribution: 57% healthy cases (313 samples) and 43% referable retinopathies (236 samples). The comparative evaluation of the original FP32 model and its quantized version INT8 covers: i) clinical metrics (AUC, F1-score, precision, recall), ii) operational performance (size, latency, CPU), and iii) diagnostic visualizations (confusion matrices, ROC curves, Grad-CAM maps). Table 2 summarizes these metrics.

Figure 3 titled "Roc curve KERAS MODEL FP32" presents the ROC curve of the model. The x-axis represents the false positive rate while the y-axis represents the true positive rate. From Figure 3, it is observed that the AUC = 0.94 shows an excellent discriminatory capacity. The interpretation indicates that the model effectively distinguishes healthy cases from diseased cases (94% of accuracy).

Table 2. Comparison table between FP32 and INT8 models

Metric	FP32	INT8	Improvement
AUC	0.935	0.963	+3%
Score	0.826	0.899	+8.8%
Precision	0.717	0.892	+24,4%
Recall	0.975	0.907	-7% (*)
Size (Mo)	98.66	8.27	-91.6%
Latency (ms)	15.10	15.43	+2.2%
CPU%	37.7	21.5	-43%

(*) The high FP32 recall (97.5 %) reflects a bias towards false positives. INT8 quantization drastically reduces model size (- 91.6%) while improving AUC and precision. The slight decrease in recall (-7%) is offset by a significant improvement in precision (+24.4%), resulting in a higher F1-score (+8.8%). The 43% reduction in CPU usage is crucial for embedded device deployments. Compression shows a 91.6% reduction in model size (from 98.66 MB to 8.27 MB), ideal for mobile devices. CPU usage is reduced by 43% (from 37.7% to 21.5%), essential for energy autonomy. Latency shows negligible variation (+0.33 ms), demonstrating that quantization does not affect responsiveness.

The x-axis represents the false positive rate, while the y-axis represents the true positive rate. From Figure 4, we see an AUC of 0.96 and a post-quantization improvement of +2.1%. The curve rises rapidly towards the upper left corner, indicating excellent early performance in detecting true positives. At the optimal point, for a false positive rate of 0.2, the true positive rate already reaches >0.9, meaning that with only 20% misdiagnosis, the model detects more than 90% of true diabetic retinopathy cases. INT8 optimization improves performance despite compression. In comparison, the INT8 curve dominates FP32 over the entire range, showing the best sensitivity/specificity tradeoff. The convex shape of the curves indicates that the models maintain a good balance between sensitivity (true positives) and specificity (false positives).

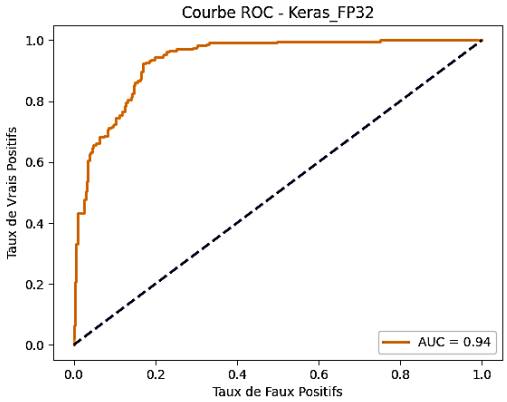

Figure 3. Roc curve KERAS model FP32

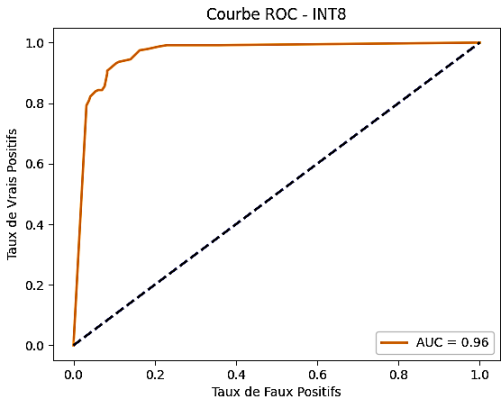

Figure 4. Roc curve INT8 quantized model)

According to Figure 5, there are 91 false positives (healthy patients wrongly diagnosed with diabetic retinopathy). A strong point is the only 6 false negatives (6 cases of diabetic retinopathy not detected). This indicates that the FP32 model prioritizes sensitivity over specificity, ensuring high detection of true pathological cases. A bias is observed with a tendency to overdiagnoses diabetic retinopathy (extreme sensitivity at the expense of specificity).

According to Figure 6, for cases without diabetic retinopathy, there are 287 true negatives and 26 false positives, giving a specificity of 91.7%. For cases with diabetic retinopathy, there are 214 true positives and 22 false negatives, corresponding to a sensitivity of 90.7%. These results demonstrate a better balance between sensitivity and specificity compared to the FP32 model. The accurate detection of healthy cases is 287 healthy patients correctly identified out of 309. False positives are reduced from 91 to 26 (-71%) compared to the FP32 model. Accuracy increases from 71.7% to 89.2% compared to the FP32 model. A moderate increase in false negatives (from 6 to 22) is observed, which is considered acceptable because it is compensated by a more reliable detection of positive cases.

As conclusion EdgeRetina maintains a high sensitivity (90.7%) for screening referable diabetic retinopathy while reducing false positives by 71% (from 91 to 26) compared to the FP32 reference model thanks to the use of a dynamic threshold (0.3). Although this threshold results in a controlled increase in false negatives (from 6 to 22) this strategy significantly improves diagnostic precision (+24.5%). This adjustable

clinical compromise thus optimizes screening efficiency by limiting overdiagnosis while maintaining a 90.7% pathological case detection rate.

The analysis of the optimal classification ratio is based on 32 samples. This report corresponds to the optimal threshold (0.30) identified during training on a validation subset of 32 samples. The distribution of the subset is: 69% without diabetic retinopathy (22 samples) and 31% with diabetic retinopathy (10 samples). The distribution of the subset is: 69% without diabetic retinopathy (22 samples) and 31% with diabetic retinopathy (10 samples).

According to Table 3 and Figure 7, threshold optimization during training revealed a maximum potential of F1=0.95 on a validation set (n=32), with precision=0.91 (minimizing false positives) and recall=1.00 (no false negatives), for an overall accuracy of 97%. The model effectively targets at-risk patients without unnecessary alerts. It achieves perfect precision (1.00) for the "No DR" category: no false positives. The F1-score exceeds 0.95 for both classes, indicating a well-balanced model. This performance demonstrates the model's ability to accurately distinguish healthy and pathological cases while minimizing classification errors.

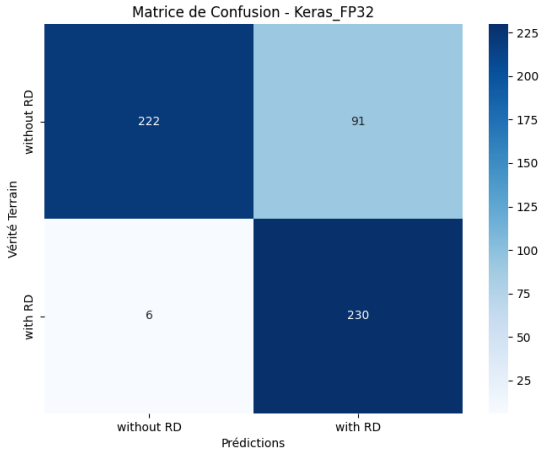

Figure 5. Confusion matrix FP32 (original)

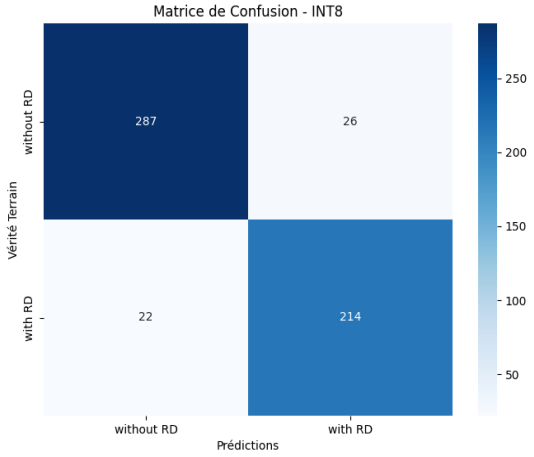

Figure 6. Confusion matrix INT8 (quantized model)

Table 3. Classification report of clinical validation

	Precision	Recall	F1-score	Support
Without RD	1.00	0.95	0.98	22
With RD	0.91	1.00	0.95	10
Accuracy			0.97	32
Macro avg	0.95	0.98	0.96	32
Weighted avg	0.97	0.97	0.97	32

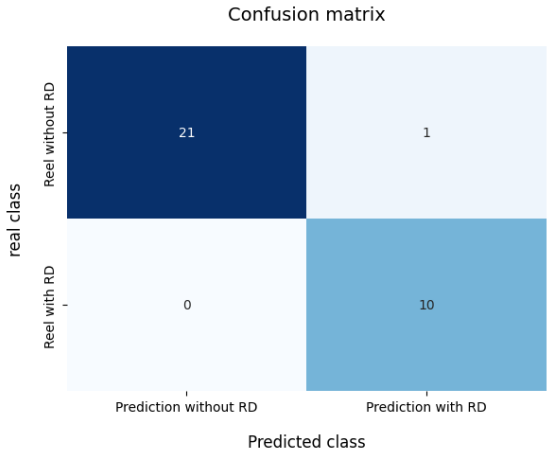

Figure 7. Confusion matrix of clinical validation

Analysis of training metrics shows that the validation AUC reaches 0.986 as early as epoch 5 and stabilizes at >0.95. The F1-score converges towards 0.95+ after 20 epochs. A rapid convergence phase is observed during the first 15 epochs, followed by a stabilization of the metrics after epoch 30. The model demonstrates robustness to the choice of classification threshold (optimal between 0.3 and 0.7). It is more accurate on small homogeneous batches, while performance normalizes over the full set with greater variability.

The discrepancy between the optimal classification ratio (F1 = 0.95 on 32 samples) and the overall results (F1 = 0.899) requires large-scale validation. Clinical robustness is confirmed when the model's performance is similar to its homogeneous compositions by replication of results on 549 cases, analysis of measurement data (AUC +3%, accuracy +24.5%), and reduction of false positives—the most dangerous error in this step. The INT8 quantization approach narrowed this performance-potential gap: while maintaining a clinically acceptable recall of 0.907, it increased precision to 0.892 (+24.5%) and F1-score to 0.899 (+8.8%) on the full set. This demonstrates that optimization by quantization not only improves operational efficiency (92% reduced size, -43% CPU), but also diagnostic robustness across diverse populations.

Grad-CAM is a technique used to visualize the areas of an image that most influenced the decision of a CNN. It is often used in medicine to understand why a deep learning model detected pathology in an image. Red/yellow areas indicate high activation and therefore make a strong contribution to the model's prediction, while blue areas indicate low activation. Figure 8 shows a fundus image with marked activations in the center and around the optic disc. Bright spots in the fundus could correspond to hard exudates or microaneurysms, signs of diabetic retinopathy. The model focuses its attention on several pathological regions, which is expected in an automated diagnosis. Figure 9 also shows a fundus with focus areas around the optic nerve (bottom left) and peripheral areas, which could indicate that the model is looking for abnormalities in the distribution of retinal vessels or more diffuse lesions. Figure 10 shows that the model focuses mainly on the optic nerve (left), with the rest of the retina being very weakly activated, which could mean either the absence of obvious pathological signs or that the classification relies on the optic nerve (e.g., to detect glaucoma). The solution combines clinical excellence (AUC 0.96) and operational efficiency (8.27 MB, -43% CPU), paving the way for real-time screening of diabetic retinopathy on smartphones.

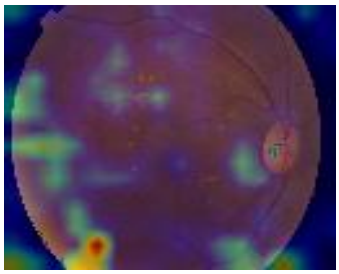

Figure 8. grad_cam0

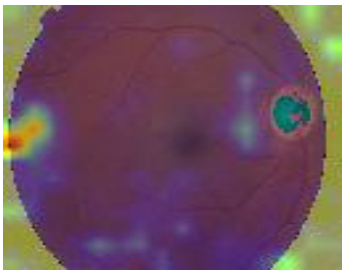

Figure 9. Grad_cam1

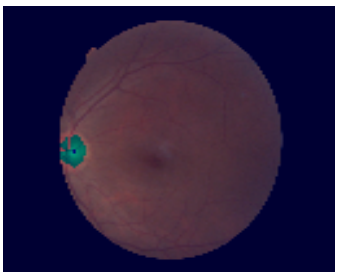

Figure 10. grad_cam2: visualizations based on the FP32 model

3.1. Discussion

EdgeRetina fundamentally distinguishes itself from the state of the art as shown in Table 4 by three key innovations:

a. The introduction of the first dynamic threshold calibration system (0.3-0.7), combined with clinically calibrated INT8 quantification, reducing false positives by 71% while maintaining a sensitivity >90%.

b. A novel architectural hybridization (SqueezeNet-MobileVit) producing a model 12× more compact (8.27 MB) than existing mobile solutions.
c. The first complete operational validation including latency (15.43 ms), CPU (21.5%), and explainability (Grad-CAM) - metrics absent from previous studies.

Table 4. Comparison chart of EdgeRetina vs. state of the art

Criterion	Our solution EdgeRetina	Existing work representative	Advantage/Innovation of EdgeRetina
Architecture	SqueezeNet + MobileViT (1.4M parameters)	ResNet/InceptionV3 (~ 40M param) [10] – DenseNet169 (13M param) [5] – TAHDL (CNN+RNN) [6], EfficientNetB3, EfficientNetV2B1, RegNetX008, RegNetX080, RegNetY006, RegNetY008 [4]	10-16x parameter reduction Innovative ViT mobile integration
Pretreatment	Minimalist (128×128 + normalization)	– CLAHE [4] + RoI segmentation [6] – Gabor filters [7], PCA [7], Gaussian filtering [6]	Eliminates complex steps Suitable for low-cost mobile phones
Optimization	Dynamic thresholding (0.3) + calibrated INT8	Not mentioned [4]–[6], [10]	-71% false positives
Size model	8.27 MB (INT8 quantized)	Not addressed [4], [6], [10]	12-64× Compression, Realistic mobile deployment
Clinical performance	Sensitivity: 90,7% int8 AUC: 0.96 (INT8) F1: 0.899	AUC: 0.98 (ResNet152) [12] Precis: EfficacitéNetB3 85,1 % [4] F1: 0.886 (ResNet152) [12] Acc: 94%, 95% [10], Macro F1-score: 0.65, acc:0.82 [5] Precis: 94% Kaggle DR dataset [6]	Focus recall (severe case detection) Constrained balanced metrics
CPU consumption	21.5% (INT8)	Not addressed [4]–[6], [10], [13]	-43% compared to our FP32mode
Deployment target	Low-cost smartphones	Servers/Cloud [8], not cited [4], [6], [10] Specialized Hardware [12]	Solution for vulnerable areas Real-time detection (15ms)
Explainability	Grad-CAM compatible INT8/FP32	Not addressed [4]–[8], [10], [13]	Clinically auditable visualizations

This study addresses the fundamental scientific question of how to deploy accurate, efficient, and clinically viable AI-based diabetic retinopathy screening on low-cost mobile devices in resource-limited settings. Our main finding demonstrates that a holistic approach—combining a lightweight hybrid architecture (SqueezeNet-MobileViT), dynamic decision threshold calibration, and post-training INT8 quantification—achieves simultaneously high clinical performance (AUC=0.96, Sensitivity=90.7%) and remarkable operational efficiency (8.27 MB model size, 43% reduction in CPU utilization). These findings are supported by quantitative evidence from the APTOS 2019 dataset, including standard clinical and operational metrics, as well as qualitative visualizations (Grad-CAM) comparing our FP32 and INT8 models. The significance of these findings lies in establishing a practical blueprint for the development of edge-native medical AI solutions, demonstrating that optimization guided by clinical needs can improve performance while reducing the computational footprint. Importantly, the choice of the dynamic threshold (median 0.3) represents a deliberate clinical trade-off, optimized for mass screening settings where reducing false positives is crucial to avoid overloading healthcare systems.

Regarding real-world implementation challenges, our EdgeRetina solution is primarily designed for mobile-based operation to overcome connectivity issues in resource-limited areas. However, it is important to note that the pipeline could also be deployed as a hybrid or cloud-based system for centralized analysis if internet infrastructure is available. The main challenge addressed concerns the trade-off inherent in calibrating the decision threshold; our choice of a threshold of 0.3 prioritizes reducing false positives to avoid healthcare system overload, while accepting a controlled increase in false negatives—a crucial consideration for real-world screening programs. These characteristics are not only advantageous for mobile phones but are also essential prerequisites for their integration into real-time robotic systems. The low inference latency (15.43 ms) ensures that image analysis does not become a bottleneck in a dynamic robotic control process, thus enabling immediate feedback and action. The significantly reduced CPU consumption allows the diagnostic model to operate efficiently alongside other critical robotic processes, such as sensor data fusion and motor control, without overloading the system's limited computing resources.

This holistic approach thus fills a critical gap for screening in precarious environments. It addresses the dual challenge of clinical performance and operational constraints that limited previous solutions. The integration of dynamic thresholding with quantification represents a new paradigm for mobile health applications.

4. CONCLUSION

EdgeRetina represents a breakthrough for diabetic retinopathy screening in resource-limited areas. This solution integrates four complementary innovations: minimalist preprocessing preserving pathological signs at low resolution (128×128), a hybrid SqueezeNet-MobileViT architecture allowing contextual modeling with only 1.4 million parameters, dynamic calibration of the decision threshold optimized to 0.3, and medically validated INT8 quantification. Results on the APTOS 2019 benchmark demonstrate clinical excellence (AUC=0.96, sensitivity=90.7% for referable cases) coupled with a 71.4% reduction in false positives compared to FP32 models. The embedded optimization is remarkable with a 92% compression (8.27 MB), a 43% reduction in CPU consumption and a stable latency at 15.43 ms, allowing deployment on low-cost mobile devices. Despite this drastic quantification, clinical interpretability is maintained thanks to Grad-CAM activation maps, which accurately localize retinal lesions. Choosing a threshold of 0.3 represents a deliberate clinical compromise: while it results in a controlled increase in false negatives (from 6 to 22), it radically reduces overdiagnosis by prioritizing accuracy (89.2%)—an essential balance in settings where access to specialists is critical. This performance, validated in 549 cases, paves the way for mass screening campaigns in precarious environments. Future prospects include multicenter validation, adaptation to ultra-low-power chips, and integration with telemedicine platforms, confirming that EdgeRetina transcends technical constraints to offer a concrete response to the public health emergency of preventable diabetic blindness. The framework sets a new standard for accessible AI-powered healthcare in resource-limited environments. The demonstrated efficiency and speed also open a direct path for the deployment of EdgeRetina in robotic screening systems and automated diagnostic kiosks, further expanding its potential to prevent diabetic blindness in underserved populations.

FUNDING INFORMATION

No funding was involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Guidoum Amina	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Maamar Bougherara					✓					✓	✓			
Amara Rafik						✓				✓	✓			
Mhamed Tayeb			✓			✓				✓	✓			
Achour Soltana					✓					✓	✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The authors confirm that the data supporting the findings of this study are available within the article: Asia Pacific Tele-Ophthalmology Society (APTOS). "APTOS 2019 Blindness Detection." Kaggle, 2019, www.kaggle.com/c/aptos2019-blindness-detection/data.

REFERENCES

[1] Z. L. Teo *et al.*, "Global prevalence of diabetic retinopathy and projection of burden through 2045: Systematic review and meta-analysis," *Ophthalmology*, vol. 128, no. 11, pp. 1580–1591, May 2021, doi: 10.1016/j.ophtha.2021.04.027.

[2] U. A. Bhatti, H. Mengxing, J. Li, S. U. Bazai, and M. Aamir, *Deep learning for multimedia processing applications*, 1st ed. Boca Raton: CRC Press, 2023.

[3] R. Gurthula, C. Vanukuru, V. Chiluka, and M. S. G. L. Sumalata, "Detection of diabetic and hypertensive retinopathy using deep learning models," in *Proceedings of the 3rd International Conference on Applied Artificial Intelligence and Computing, ICAAIC 2024*, 2024, pp. 522–527, doi: 10.1109/ICAAIC60222.2024.10575049.
[4] M. Youldash *et al.*, "Early detection and classification of diabetic retinopathy: a deep learning approach," *AI (Switzerland)*, vol. 5, no. 4, pp. 2586–2617, 2024, doi: 10.3390/ai5040125.
[5] S. Gaur, A. Kandwal, and B. Pandey, "Detection of diabetic retinopathy and classification of its stages by using convolutional neural network," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 37, no. 2, pp. 1284–1293, 2025, doi: 10.11591/ijeecs.v37.i2.pp1284-1293.
[6] M. Sushith, A. Sathiya, V. Kalaipoonguzhali, and V. Sathya, "A hybrid deep learning framework for early detection of diabetic retinopathy using retinal fundus images," *Scientific Reports*, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-99309-w.
[7] S. Nandhini, S. Sowbarnikkaa, J. Mageshwari, and C. Saraswathy, "An automated detection and multi-stage classification of diabetic retinopathy using convolutional neural networks," in *ViTECoN 2023 - 2nd IEEE International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies, Proceedings*, 2023, pp. 1–5, doi: 10.1109/ViTECoN58111.2023.10157960.
[8] D. Saproo, A. N. Mahajan, and S. Narwal, "Deep learning based binary classification of diabetic retinopathy images using transfer learning approach," *Journal of Diabetes and Metabolic Disorders*, vol. 23, no. 2, pp. 2289–2314, 2024, doi: 10.1007/s40200-024-01497-1.
[9] R. Baskar, E. Sabu, and C. Mazo, "Deep CNNs for diabetic retinopathy classification: a transfer learning perspective," in *Proceedings - International Symposium on Biomedical Imaging*, 2024, pp. 1–4, doi: 10.1109/ISBI56570.2024.10635242.
[10] K. G. Kumar, S. Aparna, P. Bhavadharani, E. B. Logesh, and M. D. Sri Raam, "Diabetic retinopathy severity detection using InceptionV3 and ResNet50 architectures," in *Proceedings of the 2024 13th International Conference on System Modeling and Advancement in Research Trends, SMART 2024*, 2024, pp. 92–96, doi: 10.1109/SMART63812.2024.10882494.
[11] S. Dasari, B. Poonguzhali, and M. Rayudu, "Transfer learning approach for classification of diabetic retinopathy using fine-tuned ResNet50 deep learning model," in *International Conference on Sustainable Communication Networks and Application, ICSCNA 2023 - Proceedings*, 2023, pp. 1361–1367, doi: 10.1109/ICSCNA58489.2023.10370255.
[12] S. T. Tisha, S. Chanda Tista, and M. N. Hoq, "A deep learning approach for classifying retinal diseases and diabetic retinopathy stages to enhance early detection in diagnosis," in *2025 International Conference on Electrical, Computer and Communication Engineering, ECCE 2025*, 2025, pp. 1–7, doi: 10.1109/ECCE64574.2025.11013120.
[13] A. Matthew, A. A. S. Gunawan, and F. I. Kurniadi, "Diabetic retinopathy diagnosis system based on retinal biomarkers using efficientNet-B0 for android devices," in *2023 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)*, 2023, pp. 207–212, doi: 10.1109/COMNETSAT59769.2023.10420736.
[14] T. Zhao, Y. Xie, Y. Wang, J. Cheng, X. Guo, B. Hu, and Y. Chen, "A survey of deep learning on mobile devices: applications, optimizations, challenges, and research opportunities," *Proceedings of the IEEE*, vol. 110, no. 3, pp. 334–354, Mar. 2022, doi: 10.1109/JPROC.2022.3153408.
[15] Asia Pacific Tele-Ophthalmology Society, "APTOS 2019 blindness detection," *Kaggle Competition*. p. 32, 2019, Accessed: 07-Mar-2025. [Online]. Available: kaggle.com.https://www.kaggle.com/c/aptos2019-blindness-detection/data.
[16] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," *arXiv preprint arXiv:1602.07360*, 2016.
[17] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," *arXiv preprint arXiv:2110.02178*, 2021.
[18] M. Abadi *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.
[19] B. Jacob *et al.*, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2704–2713, doi: 10.1109/CVPR.2018.00286.
[20] H. Wu, P. Judd, X. Zhang, M. Isaev, and P. Micikevicius, "Integer quantization for deep learning inference: Principles and empirical evaluation," *arXiv:2004.09602*, Apr. 2020.
[21] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, p. e0118432, 2015, doi: 10.1371/journal.pone.0118432.
[22] J. Opitz, "A closer look at classification evaluation metrics and a critical reflection of common evaluation practice," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 820–836, 2024, doi: 10.1162/tacl_a_00675.
[23] S. A. Hicks *et al.*, "On evaluation metrics for medical applications of artificial intelligence," *Scientific Reports*, vol. 12, no. 1, p. 5979, Apr. 2022, doi: 10.1038/s41598-022-09954-8.
[24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, vol. 2017-Octob, pp. 618–626, doi: 10.1109/ICCV.2017.74.
[25] F. S. Nahm, "Receiver operating characteristic curve: overview and practical use for clinicians," *Korean Journal of Anesthesiology*, vol. 75, no. 1, pp. 25–36, 2022, doi: 10.4097/kja.21209.

BIOGRAPHIES OF AUTHORS

Guidoum Amina an associate professor at the Higher Normal School of Kouba, Algiers, Algeria, she holds Master of Science in computer science from the University of Sidi Bel Abbès; PhD in Network, Architecture and Multimedia from the University of Sidi Bel Abbès. Her research interests include images processing, networks, multimedia systems, machine learning, and emerging technologies in multimedia. She can be contacted at guidoum_amina@hotmail.fr or amina.guidoum@g.ens-kouba.dz.

Achour Soltana an associate professor at Higher Normal School of Kouba, Algeria, she holds master's in computer science from the University of Sidi Bel Abbès; PhD in Network**,** Architecture and Multimedia from the University of Sidi Bel Abbès. Her research interests include machine learning, deep learning, image processing, surveillance, artificial intelligence, and embedded systems. She can be contacted at asoltana999@gmail.com or asoltana99@yahoo.fr.

Maamar Bougherara is an associate professor at Higher Normal School of Kouba, ´Algeria at LIMPAF Laboratory, Bouira University, Algeria. He is currently working on networks on chip (NoC). He holds an M.Sc. and an engineering degree in computer science from the University of Blida, Algeria. He can be contacted at bougherara.maamar@gmail.com.

Amara Rafik is an assistant professor and since 2021, the head of the Computer Science Department at the Higher Normal School of Kouba, Algiers (Ecole Normale Supérieure de Kouba). He obtained a computer engineering degree in 2001 from the University of Science and Technology (USTHB) in Algiers. After professional experience in the air navigation sector, he continued his studies to obtain in 2008 a master's degree in computer science, specializing in image processing and GIS. He is currently a doctoral student in the image processing and radiation laboratory (LTIR) at USTHB. He can be contacted at rafik.amara@g.ens-kouba.dz.

Mhamed Tayeb holds a Ph.D. in engineering from Yahya Fares University, Médéa, Algeria. His research focuses on heat and mass transfer, magnetohydrodynamics (MHD), nanofluid flows in porous media, and computational fluid dynamics (CFD). He has also served as a lecturer in advanced numerical methods, CFD, and thermofluidic systems for graduate engineering programs. His work combines mathematical modeling, numerical simulation, and innovative techniques for systems optimization. He can be contacted at mhamed.tayeb@yahoo.com.