Proceedings of the 6th International Conference on Biological Engineering and Medical Science
DOI: 10.54254/2753-8818/2026.38827

Lightweight AI Fundus Image Model for Early Screening of Diabetic Retinopathy

Chenyi Lin

Faculty of Life Sciences, University College London, London, UK
zczvcli@ucl.ac.uk

Abstract. Diabetic retinopathy (DR) is now the main cause of preventable blindness in working-age adults worldwide. With the rising prevalence of diabetes, the need for DR screening is becoming increasingly urgent. Traditional DR screening requires professional ophthalmologists; thus, there are problems of a shortage of expert resources, uneven distribution, and lengthy manual interpretation. Most of the current deep-learning models have too many parameters and are computationally expensive; therefore, they are not suitable for deployment on edge devices. To solve the problem of deployment bottlenecks, the development of lightweight artificial intelligence models for fundus images has become necessary. This paper systematically reviews the recent developments in lightweight artificial intelligence models for DR screening based on fundus photographs. Custom image pre-processing, a lightweight network design, advanced model compression techniques and interpretability mechanisms are employed to reduce both the number of parameters and computational cost substantially without losing the sensitivity to early retinopathy. The model's diagnostic reliability and computational resource consumption have been examined. Ablation studies and visual analyses show the respective contributions of each module to the model's performance and efficiency. Additionally, to improve the interpretability of the model and boost clinical trust, Gradient-weighted Class Activation Mapping (Grad-CAM) is introduced. Future research will enhance the adaptability of cross-device and cross-domain applications, expand the scope of clinical verification, and integrate multimodal clinical data for reliable and fair ophthalmic AI deployment.

Keywords: lightweight AI, image model, early screening, diabetic retinopathy

1. Introduction

Diabetic retinopathy (DR) is the leading cause of preventable blindness in working-age adults worldwide; hence, early screening and management of DR are among the most important emerging challenges in public health with the rising prevalence of diabetes. Diabetic retinopathy is insidious, as microaneurysms and small retinal hemorrhages may develop silently without causing visual disturbances. Early identification and treatment of DR can reduce the risk of severe vision impairment by more than 90% [1]. Therefore, a high-throughput and widely accessible early-screening service should be established clinically and economically.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.38827

DR screening is usually performed by direct ophthalmoscopy or color fundus photography, and trained specialists need to grade the photos or conduct fundus fluorescein angiography (FFA). The experts' ratings are relatively accurate; however, there are system-level operational bottlenecks in the current screening model that are blocking the progression of the DR screening process. First, the distribution of specialized ophthalmologists around the world is extremely uneven; most ophthalmologists are employed in tertiary hospitals. Therefore, there are fewer providers and a shortage of diagnoses in primary care or community health service areas that require such services. Second, there is a structural shortage that leads to delayed care and diagnostic errors. Manual grading of fundus photographs is very labor-intensive and may be subject to interobserver variability due to clinician fatigue or differences in their own expertise. Third, the provider-side bottleneck results in considerable barriers for patients, such as the burden of travel, long waiting times and high diagnostic costs; thus, follow-up is neglected and the progression of asymptomatic diseases to serious illness cannot be prevented.

In recent years, deep-learning-based artificial intelligence has achieved good results in the automated analysis of fundus images. Good-performing models are generally larger and more parameter-intensive. A network of this size is computationally expensive, has a relatively long inference delay, and is not suitable for edge devices, point-of-care portable fundus cameras, or hardware-constrained primary clinics.

This review addresses the problem of building a lightweight AI fundus image model. To keep the high sensitivity to small disease spots, structural pruning, quantization and other efficient feature-extraction modules have been used to create a lightweight model with fewer parameters and floating-point operations (FLOPs). Therefore, lightweight models can carry out fast and accurate AI diagnosis at the edge of a terminal to enable low-cost early screening and triage in community clinics or other environments with a shortage of medical staff.

This paper aims to collect the above-mentioned scattered information in one place and provide a general introduction to lightweight AI models for DR screening based on fundus photographs.

2. Methodology

Many factors need to be considered in the design of an appropriate architectural framework for AI [2]. In contrast, the high-computation-cost parameter-dense deep-learning model is not used; instead, customized image pre-processing, a lightweight network design, state-of-the-art model compression methods and explainability mechanisms are employed [3].

2.1. Pre-processing and image standardization

Artifacts appear in images from the primary care setting using different camera specifications due to nonuniform illumination and low contrast [2]. Therefore, images are standardized from different imaging devices during pre-processing to extract similar features. Contrast-limited adaptive histogram equalization (CLAHE) enhances subtle vascular structures and focal pathological lesions, such as microaneurysms and soft exudates. The green-channel signals provide a high signal-to-noise ratio concerning the retinal vasculature. These are then cropped and mean-variance normalized to remove inter-site noise and thus ensure better downstream model convergence.

DOI: 10.54254/2753-8818/2026.38827

2.2. Lightweight architectural core

Lightweight convolutional networks provide the core feature representation by maximizing the gain from features with minimal mathematical operations [3]. Therefore, the backbone employs depthwise separable convolutions, inverted residuals, and channel-shuffle operations rather than typical parameter-heavy convolutions morphologically adapted to early DR. Clinically, the initial changes in the retina are small in extent, have a low-intensity edge gradient and poor local contrast, and may present as microaneurysms, small dot hemorrhages, and very small hard exudates in the adjacent retinal tissue. Depthwise Separable Convolution divides the processes of spatial filtering and channel-wise transformation to prevent cross-channel noise and achieve fine-grained separation of spatial features. Therefore, pointwise convolution can extract relevant low-level lesion features. First, the input features are expanded to a high-dimensional space for the inverted residual structure prior to depthwise filtering. At the same time, linear projection bottlenecks do not enhance weak high-frequency signals in small exudates via non-linear activation functions. Channel-shuffle operations can be employed to enhance the cross-channel exchange of information, and thus the strong background retinal structures, such as large blood vessels and optic disc boundaries, will not obscure weak focal pathological responses. Many factors reduce the number of parameters and computational cost, thereby improving the diagnostic high-frequency response required for early diabetic retinopathy (DR) screening.

2.3. Model compression methodologies

Structural pruning, quantization, and knowledge distillation, three complementary compression strategies, have been integrated within a lesion-aware optimization pipeline for seamless deployment on point-of-care resource-constrained equipment and portable edge devices [2]. Rather than considering compression as an abstract parameter reduction, the three techniques were designed to preserve subtle features of early diabetic retinopathy, such as small microaneurysms, low-contrast hemorrhages, and tiny exudate margins. Structural pruning determines each filter's importance according to activation responsiveness toward the focal pathological region; thus, redundant background noise-sensitive channels are eliminated, while the feature maps representing micro-lesions are retained [3]. Subsequently, weights and activation maps are converted from FP32 floating-point precision to INT8 precision using Quantization-Aware Training (QAT). During QAT, lesion-weighted loss functions let the model adjust quantization step sizes so that subtle gradients of pixel-level intensity are preserved, and the boundaries of low-contrast exudates remain visible in a low-precision representation [2]. Finally, knowledge distillation uses the high-capacity teacher network to transfer soft-label probability distributions and inter-class dark knowledge to the compressed student backbone. Thus, the lightweight model inherits subtle decision boundaries that separate real early pathology from non-pathological retinal artifacts or noise [3]. The three methods drastically reduce parameter count, memory footprint, and inference latency on edge terminals, allowing real-time diagnostic speed without losing sensitivity to early retinal disease.

2.4. Model interpretability integration

Transparent reasoning will be a clinical acceptance hurdle for AI diagnostics in primary care [4]. To this end, Gradient-weighted Class Activation Mapping (Grad-CAM) has been included in the diagnostic pipeline to deal with the "black box" dilemma of deep neural networks. Gradient-weighted Class Activation Mapping (Grad-CAM) generates a visual localization heatmap by

DOI: 10.54254/2753-8818/2026.38827

calculating the gradient of the class score concerning the feature maps in the last convolutional layer, indicating the spatial location from which the network derives the prediction. Thus, such a map localizes the distinctive spatial patterns of early DR lesions: microaneurysms, as focal, punctate highlight spots; dot or flame-shaped hemorrhages, as dispersed high-response areas, often tracking the retinal microvasculature; and hard exudates, as localized hotspots with irregular, feathered edges. In this respect, such explicit visualization of the pathological spatial distributions allows primary care physicians to cross-reference rapidly the highlighted visual cues against the raw color fundus image. This will decrease the chances of false-negative diagnostic cases due to non-specialist oversight and thus substantially improve the clinical credibility of automated referral recommendations.

3. Evaluation metrics

Lightweight AI models for point-of-care screening have two perspectives: diagnostic reliability and computational resource constraints. The primary care workflow maximizes diagnostic sensitivity first to avoid false negatives and not miss subtle early-stage lesions, balanced against high specificity to prevent overloading tertiary referral networks. The F1 score and area under the receiver operating characteristic curve (AUC-ROC) consider class imbalance, which is typical of population screening [5, 6]. Meanwhile, parameters, FLOPs, memory footprint, and inference latency represent local memory occupancy, device power consumption, and the feasibility of feedback within a single consultation. Table 1 summarizes the main metrics, their analysis focus, and their clinical or engineering implications.

Table 1. Comprehensive summary of evaluation metrics for lightweight AI screening models

Metric	Category	Calculation Formula	Significance Description
Sensitivity (Recall)	Clinical	true positive (TP) / (TP + false negative [FN])	Minimizes false negatives; essential for early detection of micro-lesions.
Specificity	Clinical	TN / (TN + FP)	Minimizes false positives; prevents overwhelming tertiary eye hospitals.
F1 Score	Clinical	2 × (Precision × Recall) / (Precision + Recall)	Evaluates overall balanced diagnostic accuracy under severe class imbalance.
AUC-ROC	Clinical	Integral of ROC Curve	Measures global discrimination capability across varying decision thresholds.
Parameter Count	Computational	Total trainable weights (M)	Determines model file size and local random-access memory (RAM)/Flash memory storage overhead.
FLOPs	Computational	Floating-point operations	Measures algorithmic complexity and hardware compute requirements.
Inference Latency	Computational	Execution time per image (ms)	Dictates real-time feasibility and immediate feedback in primary care consultations.

4. Analysis

The diagnostic capacity of lightweight AI models must thus be understood at the intersection of accuracy, precision, and recall concerning the subtle morphological spectrum of early-stage diabetic

DOI: 10.54254/2753-8818/2026.38827

retinopathy (DR) [7]. Early DR pathology is difficult to diagnose clinically because, in most cases, the microaneurysms, punctate hemorrhages and focal hard exudates are small in size, have poor contrast with the background fundus microvasculature and exhibit high inter-rater variability. In general, a relatively large model can achieve good overall accuracy by using background recognition in class-imbalanced screening datasets by default. Lightweight models are pruned and quantized in a lesion-aware way to extract features of interest more effectively. By preserving the spatial gradients of high-frequency information, a lightweight architecture can retain high sensitivity (recall) for very small microaneurysms and thus avoid harmful false negatives due to delayed early intervention [7]. At the same time, high-precision filtering of features can reduce false positives caused by non-pathological artifacts, such as changes in retinal pigment or dust on the camera lens, and thus improve accuracy. Thus, the diagnostic trade-offs can help the network detect early pathological changes well before there is any reduction in visual acuity; therefore, robust screening can be performed at the primary care level, and early diagnosis is necessary to prevent irreversible vision loss [7].

5. Ablation studies and visualization

5.1. Architectural ablation analysis

Microaneurysms, punctate hemorrhages, and faint hard exudates are relatively small, low-contrast lesions easily occluded by background retinal structures. Therefore, the architectural dilemma for lightweight models on edge devices is balancing the need for extreme diagnostic sensitivity to these early micro-lesions against the need to aggressively compress computational overhead. Accordingly, systematic architectural ablation analyses have been performed at the network level by adding or removing lightweight convolutional modules, compression strategies, and lesion-oriented attention modules [8-10]. Hence, such an ablation provides empirical evidence of each module's contribution to clinical diagnostic performance and computational efficiency and a sound empirical basis for optimizing screening models for edge deployment [8, 10].

The ablation series characterizes the feature extraction, artifact suppression, and resource consumption using standard fundus benchmarks based on this uncompressed baseline model. The parameters and FLOPs skyrocketed when depthwise separable convolution and inverted residual structures were not applied [8]. Still, lesion sensitivity did not increase statistically significantly, proving that standard dense convolution is a severe case of mathematical redundancy and will not facilitate the fine-grained feature extraction [10]. Furthermore, ablation studies of the lesion-oriented attention module, such as convolutional block attention module (CBAM) or lesion-weighted loss, significantly reduced the recall of faint, low-contrast exudates during quantization; thus, adaptive step-size scaling is required to prevent gradient clipping near subtle lesion boundaries when reducing precision [8]. Without knowledge distillation, the network is unable to suppress background noise and non-pathological retinal artifacts; thus, there will be many false positives and a relatively low overall F1 score [9]. Therefore, ablation studies have shown that the synergistic combination of structural compression, attention guidance and knowledge distillation significantly reduces the parameters and latency of the lightweight backbone by an order of magnitude without sacrificing sensitivity to subtle early-stage DR pathology [8, 10].

DOI: 10.54254/2753-8818/2026.38827

5.2. Interpretability and Grad-CAM heatmap visualization

One of the main problems in translating deep-learning models to clinical use, especially for safety-sensitive applications in primary care, is their black-box nature. Early DR lesions are small, of low contrast, and not easily distinguishable from background retinal structures; thus, clinical acceptance will rest on the assurance that model prediction is based on actual pathological changes rather than spurious noise or artifacts from the images. In this respect, Grad-CAM overcomes such a problem by providing attention heatmaps that localize the spatial region predominantly driving the AI diagnostic decision-making [9]. At this level of explainability, outputs from lightweight models will likely undergo clinical validation and be trusted for edge deployment in primary DR screening workflows [8, 10].

Visual inspection of these heatmaps shows very different spatial activation patterns for early clinical DR pathology. Microaneurysms are sharply focal punctate highlight spots. Dot- or flame-shaped hemorrhages cause highly responsive activation areas strictly confined within the retinal microvasculature. Hard exudates present concentrated hotspots with irregular feathery edges [8, 9]. Comparative visualization shows that conventional compressed networks typically generate diffuse, scattered, or even misaligned attention maps due to the loss of spatial detail by parameter compression. Conversely, lesion-focused attention modules like CBAM and knowledge distillation restore the spatial alignment of activation hotspots at subtle lesion edges [9]. This microscopic-level visual feedback will provide a straightforward audit trail for primary care physicians, so that nonspecialist operators may cross-check model-generated heatmaps against raw color fundus images [10]. This will significantly reduce the risks of missed and false negatives in triaging early disease stage patients and increase clinical confidence in lightweight AI screening systems [10].

6. Conclusion

Diabetic Retinopathy (DR) is among the top causes of preventable blindness worldwide. However, operational bottlenecks regarding specialist availability, diagnostic lag time, and poor patient adherence in underprivileged socioeconomic settings are fundamental to the late diagnostic paradigms based on conventional screening. The state-of-the-art architectures are parameter-heavy and computationally demanding; therefore, they are not locally deployable on resource-scarce point-of-care devices or handheld fundus cameras despite the excellent diagnostic performance of deep-learning models. This review will synthesize the state of the art regarding the architecture, methodology, and evaluation of lightweight AI models for fundus imaging and finally cover the last mile between high-accuracy deep learning and edge screening accessibility. Lightweight backbones, therefore, decouple spatial filtering from channel-wise transformation by systematically combining depthwise separable convolutions, inverted residual structures, and channel-shuffle operations. Such a new network structure allows decoupling fine-grained spatial texture and subtle boundary gradient from the network without heavy mathematical overhead. From this point forward, the lesion-aware optimization pipeline compresses models for edge terminals by structural pruning, QAT, and knowledge distillation. This joint approach is enabled not by generic parameter reduction but by selective pruning of noise-sensitive channels, adjustment of quantization step size to retain low-contrast margins of exudates, and transfer of dark knowledge from heavyweight teacher models. Therefore, this lightweight framework can realize the gigantic compression of parameters and FLOPs without sacrificing diagnostic sensitivity to minor early-stage retinal pathologies, such as isolated microaneurysms and punctate hemorrhages.

DOI: 10.54254/2753-8818/2026.38827

Therefore, the evaluation of lightweight models for point-of-care use can only be carried out by considering both clinical diagnostic accuracy and computational efficiency. Sensitivity is used in the primary care workflow to reduce false negatives and the omission of subtle lesions, while high specificity prevents an overload of the tertiary referral network. F1 score and AUC-ROC address the class imbalance characteristic of population screening, while parameters, FLOPs, and inference latency reflect memory usage, device power consumption, and feasibility of real-time feedback during a local-level consultation. Grad-CAM is used to show which areas of the image the model focuses on, and these areas are related to the disease. It provides a simple and easy-to-understand explanation for non-specialist primary care physicians on how the AI's recommendations are based on the original fundus photographs, reduces diagnostic errors, and increases clinical trust.

Lightweight AI models will be used for scalable, low-cost, and decentralized early disease detection. This class of lightweight architectures has good structural efficiency and is specifically designed for model compression to achieve rapid, accurate point-of-care triage at the primary healthcare level. Other future directions of research include cross-device domain adaptation, further prospective clinical validation in more diverse populations, and multimodal clinical data integration to achieve the goal of trustworthy and equitable ophthalmic AI.

References

[1] Chong, D. D., Das, N., & Singh, R. P. (2024). Diabetic retinopathy: Screening, prevention, and treatment. *Cleveland Clinic Journal of Medicine, 91*(8), 503–510. https://doi.org/10.3949/ccjm.91a.24028

[2] Popescu Patoni, S. I., Mușat, A. A. M., Patoni, C., Popescu, M. N., Munteanu, M., Costache, I. B., Pîrvulescu, R. A., & Mușat, O. (2023). Artificial intelligence in ophthalmology. *Romanian Journal of Ophthalmology, 67*(3), 207–213. https://doi.org/10.22336/rjo.2023.37

[3] Waisberg, E., Ong, J., Kamran, S. A., Masalkhi, M., Paladugu, P., Zaman, N., Lee, A. G., & Tavakkoli, A. (2025). Generative artificial intelligence in ophthalmology. *Survey of Ophthalmology, 70*(1), 1–11. https://doi.org/10.1016/j.survophthal.2024.04.009

[4] Fayyaz, A. M., Abdulkadir, S. J., Talpur, N., Al-Selwi, S. M., Hassan, S. U., & Sumiea, E. H. (2025). Grad-CAM (gradient-weighted class activation mapping): A systematic literature review. *Computers in Biology and Medicine, 198*(Pt B), 111200. https://doi.org/10.1016/j.compbiomed.2025.111200

[5] Sinclair, S. H., & Schwartz, S. (2024). Diabetic retinopathy: New concepts of screening, monitoring, and interventions. *Survey of Ophthalmology, 69*(6), 882–892. https://doi.org/10.1016/j.survophthal.2024.07.001

[6] Cioana, M., Deng, J., Nadarajah, A., Hou, M., Qiu, Y., Chen, S. S. J., Rivas, A., Toor, P. P., Banfield, L., Thabane, L., Chaudhary, V., & Samaan, M. C. (2023). Global prevalence of diabetic retinopathy in pediatric type 2 diabetes: A systematic review and meta-analysis. *JAMA Network Open, 6*(3), e231887. https://doi.org/10.1001/jamanetworkopen.2023.1887

[7] Vujosevic, S., Aldington, S. J., Silva, P., Hernández, C., Scanlon, P., Peto, T., & Simó, R. (2020). Screening for diabetic retinopathy: New perspectives and challenges. *The Lancet Diabetes & Endocrinology, 8*(4), 337–347. https://doi.org/10.1016/S2213-8587(19)30411-5

[8] Al-Dolat, W., Alhatamleh, S., Alqudah, N., Alhazimi, A., Amin, M., Daamseh, A., Madain, R., Malkawi, R., Al-Omari, R., Almarek, F., & Aljefri, S. H. (2026). Enhancing fundus image analysis for diabetic retinopathy using CheXNet with CBAM and Grad-CAM visualization. *Frontiers in Medicine, 13*, 1732109. https://doi.org/10.3389/fmed.2026.1732109

[9] Zhang, Z., Zhao, H., Dong, L., Luo, L., & Wang, H. (2025). A study on the interpretability of diabetic retinopathy diagnostic models. *Bioengineering, 12*(11), 1231. https://doi.org/10.3390/bioengineering12111231

[10] Long, X., Gan, F., Fan, H., Qin, W., Li, X., Ma, R., Wang, L., Hu, R., Xie, Y., Chen, L., Cao, J., Shao, Y., Liu, K., & You, Z. (2025). EfficientNetB0-based end-to-end diagnostic system for diabetic retinopathy grading and macular edema detection. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy, 18*, 1311–1321. https://doi.org/10.2147/DMSO.S506494