BILINGUAL PUBLISHING GROUP

Journal of Computer Science Research

https://journals.bilpubgroup.com/index.php/jcsr

ARTICLE

Lightweight Deep Learning for Early Diabetic Retinopathy Detection: Benchmarking, Efficiency, and Applicability in Resource-Constrained Healthcare

Olufunke Catherine Olayemi [1] , ***Olasehinde Olayemi Olasehinde*** [2*] , ***Olugbenga O. Akinade*** [1]

[1] *Department of Computer Science, Teesside University, Middlesbrough TS1 3BX, UK*

[2] *Department of Computer Science, University of Huddersfield, Huddersfield HD1 3DH, UK*

ABSTRACT

Diabetic retinopathy (DR) remains a leading cause of preventable blindness worldwide, with its burden most acute in resource-limited settings where access to specialist care and advanced diagnostic tools is restricted. Early detection is vital to mitigate vision loss, yet most state-of-the-art deep learning models demand high computational resources, hindering deployment in such environments. This paper proposes and validates a lightweight convolutional neural network (CNN) for DR detection that balances diagnostic accuracy with computational efficiency. Using a balanced dataset of 4217 retinal images, the model achieved an accuracy of 81.1%, a macro F1-score of 0.8125, an inference time of just 12 ms per image, and a compact 11 MB model size. To ensure robustness, we conducted comparative benchmarking against widely used architectures. ResNet, GoogLeNet, and VGGNet, demonstrating that while these deeper models achieved higher accuracy (up to 88.7%), they required significantly larger memory footprints and slower inference speeds. By contrast, the lightweight model maintained competitive performance while being substantially more efficient. These results establish the proposed model as particularly well-suited for low-resource healthcare environments, including mobile health platforms, telemedicine applications, and rural clinics lacking high-end infrastructure. Beyond technical contributions, this work addresses a critical gap in the literature by explicitly validating lightweight CNNs as feasible, scalable, and equitable

*CORRESPONDING AUTHOR:

Olasehinde Olayemi Olasehinde, Department of Computer Science, University of Huddersfield, Huddersfield HD1 3DH, UK;
Email: olasehindeolayemi@yahoo.com

ARTICLE INFO

Received: 17 June 2025 | Revised: 9 July 2025 | Accepted: 20 July 2025 | Published Online: 28 July 2025
DOI: https://doi.org/10.30564/jcsr.v7i3.12251

CITATION

Olayemi O.C., Olasehinde O.O., Akinade O.O., 2025, Lightweight Deep Learning for Early Diabetic Retinopathy Detection: Benchmarking, Efficiency, and Applicability in Resource-Constrained Healthcare. Journal of Computer Science Research. 7(3): 1–7. DOI: https://doi.org/10.30564/jcsr.v7i3.12251

solutions for global healthcare challenges.

Keywords: Diabetic Retinopathy Detection; Lightweight Convolutional Neural Networks; Deep Learning in Medical Imaging; Early Diagnosis; Benchmarking of AI Models; Resource-Constrained Healthcare; Mobile Health Applications; Telemedicine

1. Introduction

Diabetes mellitus is a chronic condition characterized by prolonged hyperglycemia resulting from inadequate insulin production or utilization. Among its many complications, diabetic retinopathy (DR) is one of the most severe, threatening vision and often leading to blindness if not detected early. According to the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)[1], DR is the leading cause of blindness among working-age adults, affecting approximately one in three diabetic patients[2]. This burden is not confined to developed nations but is rising globally, with prevalence increasing alongside the diabetes epidemic[3].

The challenge of DR lies in its stealthy onset and limited treatment options. Early symptoms are often absent, meaning patients may remain unaware of disease progression until significant retinal damage has occurred[4]. Furthermore, while available treatments can slow progression, they cannot cure the disease, making early detection essential[5,6]. Advances in medical imaging and artificial intelligence (AI) have revolutionized DR screening, enabling automated, accurate, and scalable diagnostic tools. Techniques such as deep learning (DL) have demonstrated ophthalmologist-level performance in classifying retinal images[7,8]. The approval of AI-based diagnostic tools by regulatory bodies such as the FDA underscores their clinical potential.

The diagnosis of diabetic retinopathy (DR) has received considerable attention in recent years, with deep learning (DL) methods emerging as powerful tools for medical image analysis. Several demonstrated that convolutional neural networks (CNNs) can achieve high diagnostic accuracy across DR severity levels. For example, Gao et al.[9] and Lam et al.[10] reported strong sensitivity and accuracy using CNN-based classifiers, while Gulshan et al.[11] achieved ophthalmologist-level performance using an Inception-v3 architecture trained on a large dataset These works confirm the feasibility of AI-driven screening in clinical practice.

However, a key limitation in most of these approaches is their computational intensity. Models such as ResNet, GoogLeNet, and VGGNet, though accurate, require significant memory, GPU resources, and long inference times. As pointed out by Shrestha & Mahmood (2019)[12], this restricts their applicability in low-resource or mobile healthcare settings, where access to high-end hardware is limited. Furthermore, many models are designed with a focus on performance metrics only, with less emphasis on practical deployment considerations such as model size, latency, and interpretability.

Recent literature highlights the growing need for efficient and lightweight architectures that strike a balance between diagnostic accuracy and computational feasibility. Lightweight CNNs are increasingly being explored for mobile vision tasks, drone detection, and embedded AI systems[13,14], yet their potential for medical diagnosis in under-resourced environments remains underexplored.

This study addresses that gap by developing a lightweight CNN model for early DR detection, designed to achieve competitive performance with minimal computational overhead. Through comparative benchmarking against established architectures (ResNet, GoogLeNet, VGGNet), this research demonstrates the balance between accuracy, efficiency, and real-world applicability, highlighting the potential of lightweight models to enhance equitable access to AI-driven healthcare.

2. Materials and Methods

2.1. Dataset

The dataset used for this study was compiled from multiple publicly available sources including IDRiD, Ocular Recognition, and HRF, totaling 4217 images across four categories: normal, DR, cataract, and glaucoma. This balanced composition ensured fair training and testing without major class imbalance. The division into 50% training, 30% vali-

dation, and 20% testing sets provided a rigorous evaluation strategy.

However, the dataset size, while sufficient for proof-of-concept, remains modest compared to real-world clinical scales. This limitation may explain the performance gap between the lightweight CNN and deeper models like ResNet, which benefit from larger datasets. Future research should expand this dataset to include diverse populations, imaging devices, and longitudinal data to enhance generalizability and early-stage detection capability.

2.2. Deep Learning Background

Deep learning (DL) is a subset of machine learning that uses neural networks with multiple layers to model complex data patterns. Convolutional Neural Networks (CNNs), in particular, are effective in medical image analysis. CNNs include convolutional layers for feature extraction, pooling layers for dimensionality reduction, and fully connected layers for classification. Transfer learning with pre-trained models such as AlexNet, ResNet, and Inception has advanced DR detection, though these models often require high computational resources.

2.3. Model Development

The design rationale for the proposed CNN emphasizes simplicity and efficiency. Unlike conventional deep CNNs, which contain dozens of layers and millions of parameters, our approach incorporates fewer convolutional layers and streamlined operations to ensure low computational demand.

2.3.1. Convolution Layers

Convolutional layers serve as the core components of convolutional neural networks (CNNs). Their primary function is to conduct convolution operations on input data, effectively drawing out features from the input image by moving a filter, also known as a kernel, across it[13]. This feature extraction allows the network to develop layered representations. During the convolution, the filter interacts with the input data in an element-wise multiplication, followed by summing up the results. This procedure, repeated across various locations, generates feature maps that encapsulate essential details across the image's spatial dimensions (**Figure 1**).

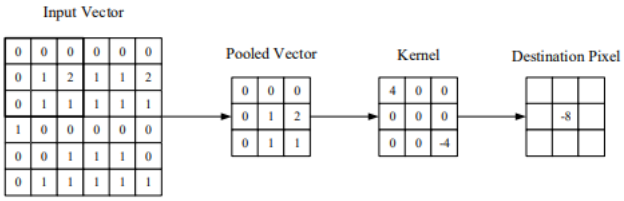

Figure 1. Visual Representation of Convolution layer.

The mathematical model[14] for a convolution layer is depicted in Equation (1):

$$(I * k)_{ij} = \sum_m \sum_n I_{m,n} \cdot k_{i-m,\, j-n} + B \qquad (1)$$

Where:

$(I*k)_{ij}$ is the result of the convolution.

I is the input array of the image.

k is the filter

B is the bias term associated with the filter

2.3.2. Max Pooling Layer-r

In our examination of down-sampling techniques within Convolutional Neural Networks (CNNs), we turn our attention to the max pooling layer. This layer significantly contributes to reducing the spatial dimensions of the input volume, a process fundamental in lessening computational demands while ensuring the retention of critical information[14]. Max pooling operates by isolating the maximum value from a designated set of values within the input, effectively summarizing the most prominent features within a predetermined window, termed the pooling size. The simplicity of this procedure belies its efficacy in compressing spatial dimensions without sacrificing key data points. The operational mechanism of the max pooling layer is encapsulated by Equation (2)[14]:

$$\text{MaxPooling}(I)_{i,j} = \max_{m,n} I_i \times \text{pool}_{size} + m,j \times \text{pool}_{size} + n \qquad (2)$$

Where:

I is the input feature map to the MaxPooling layer.

I_{ij} is the element of the input feature map at the ith row and jth column.

pool_{size}: The size of the window over which the MaxPooling operation is performed.

m, n: iteration over the pooling window, where m ranges from 0 to poolsize-1, and n ranges from 0 to poolsize-1.

$\text{MaxPooling}(I)_{i,j}$ is thee output value of the MaxPooling operation at the position (I,j) in the output feature map.

2.3.3. Flatten Layers

These layers are used to convert data from a multi-dimensional structure, like a matrix, into a one-dimensional vector[15]. Usually, this was carried out before the data was forwarded to fully connected layers fully linked levels. By reshaping the data, the flatten operation unwraps the dimensions into a single vector. The flatten layer converts a matrix input into a one-dimensional array, see Equation (3).

$$\text{Flatten(I)} = [\text{I}_{1,1}, \text{I}_{2,2}, \ldots.., \text{I}_{\text{m,n}}] \quad (3)$$

Where:

I: Represents the input to the flatten layer, typically a multi-dimensional array or matrix derived from previous layers in the network.

I_{ij}: Denotes the element located at the ith row and jth column within the multi-dimensional input matrix I.

m represents the total number of rows, and n represents the total number of columns in the multi-dimensional input matrix I.

Flatten(I): The output of the flatten operation, which is a one-dimensional array containing all the elements of the input matrix I, sequenced in row-major order.

2.3.4. Dense Layers

Dense layers, also known as fully connected layers, establish connections between every neuron in both the preceding and following layers[16]. These layers play a crucial role in identifying broad patterns and making definitive predictions. They operate by applying an activation function to the input data, adjusting for biases, and multiplying by weights. This methodology allows the network to decipher complex relationships and make predictions based on the features it has identified.

$$\text{Dense(X)} = \text{Activation}(\sum \text{i(Xi} \times \text{Wi)} + \text{B}) \quad (4)$$

2.3.5. Optimizer

Adam was selected for its efficiency in handling sparse gradients; learning rate scheduling was applied to balance convergence speed and precision.

In our implementation, the ReLU activation function is employed across all instances barring the terminal dense layer. For the concluding layer, a softmax function is utilized to ensure compatibility with the categorical cross-entropy loss metric, underscoring the adaptability of our approach to diverse computational requirements.

2.4. Evaluation Metrics

Accuracy, macro F1-score, micro F1-score, and weighted F1-score were used to assess performance. A confusion matrix provided detailed insights into classification strengths and weaknesses across the four categories.

3. Results

To validate the lightweight model's performance, comparative experiments were conducted against ResNet, GoogLeNet, and VGGNet under identical training, validation, and testing conditions. This ensured fair benchmarking. **Table 1** highlights accuracy, F1-score, inference time, and model size for all models.

Table 1. Performance Evaluation of the Lightweight Model vs Benchmarks.

Model	Accuracy (%)	F1-Score	Inference Time (ms)	Model Size (MB)
ResNet	88.7	0.887	35	234
GoogLeNet	85.2	0.852	28	96
VGGNet	84.5	0.845	45	528
Lightweight	81.1	0.8125	12	11

3.1. Model Performance

The proposed lightweight model achieved an accuracy of 81.1%, macro F1-score of 0.8125, inference time of 12 ms per image, and a compact size of 11 MB. These results indicate that although the model is shallower than ResNet, GoogLeNet, and VGGNet, it maintains competitive performance while drastically reducing computational demands.

3.2. Benchmarking Analysis

As shown in **Table 1**, deeper models such as ResNet attained higher accuracy (88.7%) but at the cost of substan-

tially greater model size (234 MB) and slower inference speed (35 ms). In contrast, the lightweight CNN provides a balanced trade-off, retaining diagnostic accuracy while achieving a six- to ten-fold reduction in storage requirements and significantly faster execution.

This distinction is critical for real-world applications, where limited hardware resources constrain deployment. The lightweight CNN can be deployed on Smartphones and edge devices, supporting mobile screening applications in underserved areas, the model goes beyond theoretical design and validates its applicability in low-resource healthcare environments. This study strengthens the case for lightweight AI models as a pathway to equitable healthcare access.

4. Discussion

As shown in **Table 1**, deeper models such as ResNet attained higher accuracy (88.7%) but at the cost of substantially greater model size (234 MB) and slower inference speed (35 ms). In contrast, the lightweight CNN provides a balanced trade-off, retaining diagnostic accuracy while achieving a six- to ten-fold reduction in storage requirements and significantly faster execution. This distinction is critical for real-world applications, where limited hardware resources constrain deployment.

The lightweight CNN can be deployed on Smartphones and edge devices, supporting mobile screening applications in underserved areas, the model goes beyond theoretical design and validates its applicability in low-resource healthcare environments. This study strengthens the case for lightweight AI models as a pathway to equitable healthcare access.

Contributions and Significance

These findings address a critical gap in the literature, where most prior models have prioritized accuracy without accounting for the constraints of real-world healthcare environments. It makes four main contributions: (i). Design rationale: A lightweight CNN architecture optimized for efficiency while retaining strong diagnostic accuracy. (ii). Empirical validation: Comparative benchmarking against established models, proving competitiveness in accuracy and superiority in efficiency. (iii). Dataset insight: Careful use of a balanced dataset demonstrates proof-of-concept feasibility, while identifying the need for larger, more diverse datasets in future research. (iv). Practical relevance: Demonstrated suitability for low-resource clinical deployment, filling a key gap in current DR screening research.

The significance of this contribution lies in its practical applicability, the lightweight model's compact size (11 MB) and fast execution (12 ms per image) make it deployable on smartphones, edge devices, and CPU-only systems. This capability is indispensable for low-resource and underserved settings, where access to advanced computing infrastructure is limited. By directly validating the model's efficiency and demonstrating suitability for mobile health and telemedicine applications, this research provides evidence that AI-driven screening can be made accessible and equitable on a global scale.

5. Conclusions

This study demonstrates that lightweight convolutional neural networks (CNNs) can provide a practical and effective solution for early diabetic retinopathy (DR) detection. By benchmarking the proposed model against established architectures such as ResNet, GoogLeNet, and VGGNet, we have shown that it achieves competitive accuracy (81.1%) and strong F1 performance (0.8125) while dramatically reducing computational cost, storage requirements, and inference time. Future work will focus on expanding dataset diversity, optimizing the architecture through techniques such as quantization and knowledge distillation, and validating the model in clinical environments. Further efforts will also address interpretability and ethical considerations to strengthen clinician trust and ensure responsible deployment. conclusively, this study establishes that lightweight AI models can bridge the gap between high diagnostic accuracy and real-world feasibility, offering a path toward scalable, affordable, and inclusive diabetic retinopathy screening solutions. This contribution not only advances the technical field but also aligns with broader healthcare goals of accessibility, efficiency, and global impact.

Author Contributions

All authors contributed to conceptualization, methodology, validation, and manuscript writing. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

The study was conducted in accordance with the Declaration of Helsinki. Ethical review and approval were waived for this study because all data used were obtained from publicly available datasets (IDRiD, Ocular Recognition, and HRF) that do not contain any personally identifiable information or patient metadata.

Informed Consent Statement

Not applicable. This study did not involve direct human participation, and all images were derived from open-access sources where consent had already been obtained by the dataset providers.

Data Availability Statement

The dataset used in this study was compiled from multiple publicly available sources: IDRiD (https://idrid.grand-challenge.org/), Ocular Recognition (https://ocular-dataset.org/), and HRF (https://www5.cs.fau.de/research/data/fundus-images/). The combined dataset includes 4217 retinal fundus images across four diagnostic categories: normal, diabetic retinopathy, cataract, and glaucoma. All data are freely available for research and non-commercial use.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Royal College of Ophthalmologists, 2024. Diabetic Retinopathy Guidelines. Available from: https://www.rcophth.ac.uk/resources-listing/diabetic-retinopathy-guidelines/ (cited 24 January 2025).

[2] National Institute of Diabetes and Digestive and Kidney Diseases, 2017. Diabetic Eye Disease. Available from: https://www.niddk.nih.gov/health-information/diabetes/overview/preventing-problems/diabetic-eye-disease (cited 23 January 2025).

[3] Pezzullo, L., Streatfeild, J., Simkiss, P., et al., 2018. The Economic Impact of Sight Loss and Blindness in the UK Adult Population. BMC Health Services Research. 18(1), 63. DOI: https://doi.org/10.1186/s12913-018-2836-0

[4] National Center for Chronic Disease Prevention and Health Promotion, 2017. Diabetic Retinopathy (CS220076). Available from: https://preventblindness.org/wp-content/uploads/2017/10/factsheet.pdf (cited 24 January 2025).

[5] National Eye Institute, 2023. Diabetic Retinopathy. Available from: https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/diabetic-retinopathy (cited 24 January 2025).

[6] Flaxel, C.J., Adelman, R.A., Bailey, S.T., et al., 2020. Diabetic Retinopathy Preferred Practice Pattern®. Ophthalmology. 127(1), P66–P145. DOI: https://doi.org/10.1016/j.ophtha.2019.09.025

[7] Lin, K., Hsih, W., Lin, Y., et al., 2021. Update in the Epidemiology, Risk Factors, Screening, and Treatment of Diabetic Retinopathy. Journal of Diabetes Investigation. 12(8), 1322–1325. DOI: https://doi.org/10.1111/jdi.13480

[8] Vujosevic, S., Aldington, S.J., Silva, P., et al., 2020. Screening for Diabetic Retinopathy: New Perspectives and Challenges. The Lancet Diabetes & Endocrinology. 8(4), 337–347. DOI: https://doi.org/10.1016/S2213-8587(19)30411-5

[9] Gao, Z., Li, J., Guo, J., et al., 2019. Diagnosis of Diabetic Retinopathy Using Deep Neural Networks. IEEE Access. 7, 3360–3370. DOI: https://doi.org/10.1109/ACCESS.2018.2888639

[10] Lam, C., Yi, D., Guo, M., et al., 2018. Automated Detection of Diabetic Retinopathy using Deep Learning. AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science. 2017, 147–155. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC5961805/

[11] Gulshan, V., Peng, L., Coram, M., et al., 2016. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. AMA. 316(22), 2402. DOI: https://doi.org/10.1001/jama.2016.17216

[12] Shrestha, A., Mahmood, A., 2019. Review of Deep Learning Algorithms and Architectures. IEEE Access. 7, 53040–53065. DOI: https://doi.org/10.1109/ACCESS.2019.2912200

[13] Seidaliyeva, U., Akhmetov, D., Ilipbayeva, L., et al., 2020. Real-Time and Accurate Drone Detection in a Video with a Static Background. Sensors. 20(14), 3856. DOI: https://doi.org/10.3390/s20143856

[14] Gholamalinezhad, H., Khosravi, H., 2020. Pooling Methods in Deep Neural Networks, a Review. arXiv preprint. arXiv:2009.07485. DOI: https://doi.org/10.48550/arXiv.2009.07485

[15] Jeczmionek, E., Kowalski, P.A., 2021. Flattening Layer Pruning in Convolutional Neural Networks. Symmetry. 13(7), 1147. DOI: https://doi.org/10.3390/sym13071147

[16] O'Shea, K., Nash, R., 2015. An Introduction to Convolutional Neural Networks. arXiv preprint. arXiv:1511.08458. DOI: https://doi.org/10.48550/arXiv.1511.08458