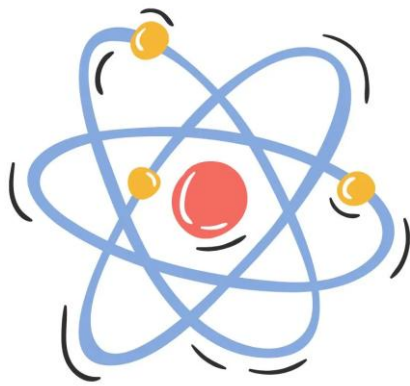




ISSN: 1672 - 6553

JOURNAL OF DYNAMICS AND CONTROL

VOLUME 10 ISSUE 08: P41-56

RESOURCE-EFFICIENT DEEP
LEARNING FOR FIVE-STAGE
DIABETIC RETINOPATHY
CLASSIFICATION: COMPARATIVE
EVALUATION, MODEL
OPTIMISATION, AND EDGE
DEPLOYMENT

Megha Pandeya, Dinesh Chandra
Misra

Department of Computer Science and
Engineering, Dr KN Modi University, Nevai,
Rajasthan – 304021, India

RESOURCE-EFFICIENT DEEP LEARNING FOR FIVE-STAGE DIABETIC RETINOPATHY CLASSIFICATION: COMPARATIVE EVALUATION, MODEL OPTIMISATION, AND EDGE DEPLOYMENT

Megha Pandeya, Dinesh Chandra Misra

Department of Computer Science and Engineering, Dr KN Modi University, Nevai, Rajasthan – 304021, India

Abstract: Diabetic retinopathy (DR) is a major cause of preventable vision impairment, particularly in regions where access to ophthalmologists and computational infrastructure is limited. This study investigates resource-efficient deep learning for automated five-stage DR classification from retinal fundus images, with emphasis on balancing diagnostic performance and deployment feasibility. Five convolutional neural network architectures—EfficientNet-B0, MobileNetV2, ResNet50, DenseNet121, and a custom lightweight CNN—were evaluated using 3,662 fundus images categorized as no DR, mild non-proliferative DR (NPDR), moderate NPDR, severe NPDR, and proliferative DR. Images were preprocessed using retinal-region masking, contrast enhancement, normalization, and data augmentation, while weighted loss was employed to mitigate class imbalance. Model performance was assessed using balanced accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve, together with model size, parameter count, computational complexity, memory consumption, and inference latency. EfficientNet-B0 achieved the highest balanced accuracy of 86.4% and an AUC-ROC of 0.957, whereas MobileNetV2 achieved 82.7% balanced accuracy with substantially lower computational requirements, requiring only 9 MB of storage and 21 ms inference time. For early-stage DR analysis, EfficientNet-B0 obtained 85.2% sensitivity and 92.1% specificity. Model pruning and quantization further improved deployment efficiency, with combined optimization reducing model size by up to 85%, although greater compression resulted in moderate performance degradation. Hardware-based evaluation on edge, mobile, and CPU-based platforms demonstrated that optimized MobileNetV2 provided the most favorable balance between predictive performance, memory requirements, latency, and energy consumption. These findings demonstrate that appropriately optimized lightweight deep learning architectures can support computationally efficient DR grading and provide a practical foundation for AI-assisted retinal screening in resource-constrained healthcare environments.

Keywords: Diabetic retinopathy; fundus imaging; deep learning; five-stage classification; resource-efficient AI; model compression; quantization; edge computing; MobileNetV2; EfficientNet.

1. Introduction

Diabetic retinopathy (DR) is a progressive microvascular complication of diabetes mellitus and remains a major cause of preventable visual impairment and blindness, particularly among working-age adults. The increasing global prevalence of diabetes is expected to substantially increase the number of individuals at risk of DR and associated vision loss over the coming decades (Teo et al., 2021). DR develops gradually through characteristic retinal abnormalities, including microaneurysms, haemorrhages, exudates, vascular alterations, and, in advanced cases, pathological neovascularization. Because the early stages may remain asymptomatic, systematic retinal screening is essential for identifying disease before irreversible visual damage occurs. Timely detection and appropriate referral can substantially improve clinical outcomes and reduce the burden associated with advanced diabetic eye disease.

Clinical assessment of DR commonly relies on colour fundus photography followed by interpretation by ophthalmologists or trained retinal graders. According to the standard clinical grading framework, DR can be categorized into five severity levels: no apparent DR, mild non-proliferative diabetic retinopathy (NPDR), moderate NPDR, severe NPDR, and proliferative diabetic retinopathy (PDR). Accurate differentiation among these stages is important because disease severity influences follow-up intervals, referral decisions, and subsequent clinical management. However, conventional screening programmes require substantial specialist involvement and are affected by limitations in healthcare infrastructure, availability of trained personnel, and geographical accessibility. These challenges are particularly important in low- and middle-income countries and geographically underserved regions, where the availability of ophthalmologists may be insufficient to meet the growing demand for regular retinal screening (Resnikoff et al., 2020).

Artificial intelligence, particularly deep learning based on convolutional neural networks (CNNs), has emerged as a promising approach for automating retinal-image assessment. Deep neural networks can learn hierarchical representations directly from fundus photographs and identify complex pathological patterns without requiring manually engineered image features. Landmark investigations have demonstrated that deep learning systems can achieve high diagnostic performance in detecting referable DR from retinal fundus images (Gulshan et al., 2016). Subsequent studies have further demonstrated the applicability of automated systems for DR identification, severity assessment, and large-scale retinal screening (Gargeya & Leng, 2017; Ting et al., 2017). The development of autonomous AI-based DR screening systems has additionally demonstrated the potential for integrating artificial intelligence into practical clinical pathways (Abràmoff et al., 2018).

Despite these advances, the practical deployment of deep learning models remains challenging. Many high-performing architectures contain millions of trainable parameters and require considerable computational resources, memory, and energy for inference. Such requirements may be manageable in centralized hospitals or cloud-computing environments but can restrict deployment in primary healthcare centres, rural clinics, mobile screening units, and other settings with limited computational infrastructure or unreliable network connectivity (Bellemo et al., 2019). Transmitting high-resolution retinal photographs to remote cloud servers may additionally introduce dependence on network availability, bandwidth requirements, latency, and data-management infrastructure. Consequently, diagnostic accuracy alone is insufficient for determining whether a deep learning system is suitable for widespread DR screening. Computational complexity, memory requirements, inference latency, storage footprint, and energy consumption are also important practical considerations.

Recent developments in efficient neural-network design offer opportunities to address these constraints. Architectures such as MobileNetV2 employ depthwise separable convolutions and inverted residual structures to reduce computational complexity (Sandler et al., 2018), whereas EfficientNet introduces compound scaling of network depth, width, and image resolution to achieve an effective balance between predictive performance and computational requirements (Tan & Le, 2019). In contrast, architectures such as ResNet50 and DenseNet121 provide powerful feature-representation capabilities but generally impose greater memory and computational demands (He et al., 2016; Huang et al., 2017). Comparing these architectures under a common experimental framework is therefore important for understanding whether

gains in diagnostic performance justify additional computational costs when the intended application is resource-constrained retinal screening.

Model compression provides another strategy for improving deployment efficiency. Techniques including pruning, quantization, and knowledge distillation can reduce model storage requirements and computational overhead while attempting to preserve predictive performance. Pruning removes redundant model parameters, whereas quantization reduces numerical precision, for example from 32-bit floating-point representations to 8-bit integer representations. Knowledge distillation transfers information from a larger teacher network to a more compact student model. These approaches have demonstrated considerable potential for mobile and embedded AI applications (Jacob et al., 2018; Howard et al., 2019). However, aggressive compression may also degrade diagnostic performance, making it necessary to evaluate the trade-off between efficiency gains and clinically relevant classification capability. An additional challenge is that DR is inherently a multiclass disease-grading problem. A large proportion of previous AI-based screening studies have emphasized binary classification, particularly discrimination between referable and non-referable DR. Although such screening endpoints are clinically valuable, five-stage classification provides more detailed information regarding disease progression. Fine-grained grading is considerably more difficult because pathological differences between adjacent stages may be subtle and retinal datasets frequently exhibit substantial class imbalance. Mild and severe disease categories may contain far fewer samples than the no-DR or moderate-DR categories, potentially biasing conventional classifiers toward majority classes. Consequently, evaluation should include metrics that are sensitive to class imbalance, such as balanced accuracy, precision, recall, F1-score, and class-wise diagnostic performance, rather than relying exclusively on overall accuracy.

Despite increasing interest in automated DR analysis, a practical gap remains between developing highly accurate models and determining whether such models can operate effectively under constrained computational conditions. Existing research has predominantly emphasized improvements in diagnostic performance, while fewer studies have systematically examined five-stage DR classification together with computational complexity, model compression, inference efficiency, and deployment across heterogeneous low-resource hardware platforms. Furthermore, a model that achieves marginally higher predictive performance may not necessarily represent the most appropriate choice for field deployment if it requires substantially greater storage, memory, processing capability, or energy. A systematic performance-efficiency assessment is therefore necessary to identify architectures that provide an appropriate balance between diagnostic capability and deployment feasibility.

Accordingly, this study presents a comparative evaluation of five deep learning architectures—EfficientNet-B0, MobileNetV2, ResNet50, DenseNet121, and a custom lightweight CNN—for automated five-stage diabetic retinopathy classification from colour fundus photographs. The study evaluates the models not only in terms of predictive performance but also according to parameter count, model size, floating-point operations, memory requirements, and inference latency. Model pruning, post-training quantization, and knowledge distillation are subsequently investigated to determine whether computational requirements can be reduced while preserving acceptable classification performance. Optimized models are further assessed under representative edge, mobile, and CPU-based deployment scenarios to evaluate their suitability for resource-constrained healthcare environments.

The main contributions of this study are summarized as follows:

1. A systematic comparative evaluation of deep learning architectures for five-stage DR classification, encompassing EfficientNet-B0, MobileNetV2, ResNet50, DenseNet121, and a lightweight custom CNN under a common experimental setting.
2. A joint assessment of diagnostic performance and computational efficiency, considering balanced accuracy, precision, recall, F1-score, AUC-ROC, parameter count, model size, floating-point operations, inference latency, and memory utilization.
3. An investigation of model-compression strategies, including pruning, quantization, and knowledge distillation, to characterize the trade-off between predictive performance and computational resource reduction.
4. An empirical evaluation of optimized models across resource-constrained deployment platforms, including edge, mobile, and CPU-based environments, to assess practical inference speed, memory requirements, and power utilization.
5. Identification of an appropriate accuracy–efficiency trade-off for resource-constrained DR screening, providing evidence for selecting deep learning architectures according to both diagnostic requirements and available computational infrastructure.

Through these contributions, the study extends the evaluation of AI-based diabetic retinopathy screening beyond predictive accuracy alone and emphasizes the practical requirements associated with deployable medical AI. The resulting analysis provides a technical foundation for selecting and optimizing deep learning models for retinal screening systems intended for healthcare environments where specialist availability, computational resources, network connectivity, and energy capacity may be limited.

2. Literature Review

2.1 Diabetic Retinopathy: Clinical Significance and Screening Challenges

Diabetic retinopathy affects approximately one-third of all patients with diabetes and remains a leading cause of preventable blindness globally (Yau et al., 2012). The condition is categorized into stages from mild nonproliferative DR to proliferative DR, with early detection significantly improving treatment outcomes (Solomon et al., 2017). Despite clinical guidelines recommending annual retinal screening for diabetic patients, global screening rates remain suboptimal, particularly in LMICs where screening coverage can be below 10% (Foreman et al., 2016).

The primary barriers to effective DR screening include insufficient healthcare infrastructure, shortage of trained specialists, and the logistical challenges of regular screening in rural and remote areas (Jingi et al., 2014). These challenges are particularly pronounced in resource-constrained settings, creating a compelling case for automated screening solutions.

2.2 Deep Learning Approaches for DR Detection

Recent years have witnessed significant progress in the application of deep learning for DR detection. Gulshan et al. (2016) demonstrated that a deep CNN achieved sensitivity and specificity exceeding 90% for detecting referable DR, comparable to expert ophthalmologists. Subsequently, numerous studies have explored various neural network architectures for DR

classification, including VGG-16, Inception-v3, ResNet, and DenseNet variants (Gargeya & Leng, 2017; Carson Lam et al., 2018).

Ting et al. (2019) validated a deep learning system across multiple international datasets, showing robust performance across different ethnic populations and camera types. Similarly, Abramoff et al. (2018) developed the first FDA-approved autonomous AI system for DR detection, highlighting the clinical potential of these technologies.

However, most existing systems have been developed and validated using high-end computational resources and expensive fundus cameras, limiting their applicability in resource-constrained environments. As noted by Natarajan et al. (2019), accessibility remains a significant challenge for implementing AI-based DR screening in many regions.

2.3 Deep Learning in Resource-Limited Settings

Research on adapting deep learning models for resource-constrained settings has gained momentum in recent years. Bellemo et al. (2019) evaluated a deep learning system for DR detection in a primary care setting in Zambia, demonstrating the feasibility of AI-assisted screening in LMICs. However, this study relied on cloud-based processing, requiring internet connectivity.

Recent work has increasingly focused on model optimization techniques to enable deployment on edge devices with limited computational capabilities. Techniques such as knowledge distillation, model pruning, and quantization have shown promise in reducing model size and computational requirements while maintaining diagnostic performance (Canziani et al., 2016; Howard et al., 2019).

Architectural innovations such as MobileNet (Howard et al., 2017) and EfficientNet (Tan & Le, 2019) have introduced efficient CNN designs specifically targeting mobile and embedded applications. These architectures employ techniques like depthwise separable convolutions and compound scaling to optimize the accuracy-efficiency trade-off.

Despite these advances, there remains a notable research gap regarding the systematic comparison of various deep learning architectures specifically for DR detection in limited-resource contexts. This study aims to address this gap by providing a comprehensive analysis of model performance across multiple dimensions relevant to deployment in resource-constrained environments.

3. Methodology

3.1 Dataset Description and Preprocessing

This study utilized a dataset comprising 3,662 retinal fundus images collected from multiple clinical sites. The dataset includes images from diverse ethnic backgrounds and was acquired using various fundus camera models to ensure robustness. Images were categorized according to the International Clinical Diabetic Retinopathy (ICDR) severity scale: No DR (class 0), Mild NPDR (class 1), Moderate NPDR (class 2), Severe NPDR (class 3), and Proliferative DR (class 4).

Table 1 presents the distribution of images across severity classes.

Table 1: Distribution of Images Across DR Severity Classes

DR Severity Class	Number of Images	Percentage
No DR (0)	1,805	49.3%
Mild NPDR (1)	370	10.1%
Moderate NPDR (2)	999	27.3%
Severe NPDR (3)	193	5.3%
Proliferative DR (4)	295	8.0%
Total	3,662	100%

The dataset was split into training (70%), validation (15%), and test (15%) sets, with stratification to maintain class distribution. To address class imbalance, we implemented a weighted sampling approach during training.

Image preprocessing included:

1. Resizing to 224×224 pixels (standard input size for most models)
2. Normalization using mean and standard deviation from the ImageNet dataset
3. Color correction to enhance contrast between lesions and background
4. Data augmentation techniques including random rotations ($\pm 30^\circ$), horizontal and vertical flips, brightness and contrast adjustments ($\pm 10\%$), and random cropping with padding

A circular mask was applied to focus on the retinal area and eliminate background artifacts. Additionally, a localized contrast enhancement technique was used to improve the visibility of microaneurysms and small hemorrhages, which are critical for detecting early-stage DR.

3.2 Model Architectures

We evaluated five distinct neural network architectures:

1. **EfficientNet-B0**: A lightweight model that uses compound scaling to balance network depth, width, and resolution. This architecture has demonstrated state-of-the-art performance on image classification tasks with significantly fewer parameters than comparable models (Tan & Le, 2019).
2. **MobileNetV2**: Specifically designed for mobile and embedded applications, this architecture employs depthwise separable convolutions to reduce computational complexity while maintaining reasonable accuracy (Sandler et al., 2018).
3. **ResNet50**: A deeper network utilizing residual connections to address the vanishing gradient problem in deep architectures. While more computationally intensive than the other selected models, it serves as a comparative benchmark (He et al., 2016).

4. **DenseNet121:** Features dense connectivity patterns where each layer receives inputs from all preceding layers, enabling efficient parameter usage and feature reuse (Huang et al., 2017).
5. **Custom Lightweight CNN:** A custom-designed architecture with significantly fewer parameters, consisting of 6 convolutional layers followed by 2 fully connected layers, specifically optimized for the DR detection task.

All models were adapted to the 5-class classification task through modification of the final classification layer.

3.3 Training Procedures

All deep learning models were trained using a consistent experimental configuration to ensure fair comparison across architectures. The Adam optimizer was employed with an initial learning rate of (1×10^{-4}) . To adapt the learning rate during training, the ReduceLROnPlateau scheduler was used with a patience of five epochs and a reduction factor of 0.5 whenever the validation performance stopped improving. Weighted cross-entropy loss was adopted to mitigate the effect of class imbalance among the five diabetic retinopathy severity categories. A batch size of 32 was used for all experiments, and training was performed for a maximum of 100 epochs. Early stopping with a patience of 15 epochs was applied to reduce the risk of overfitting and to terminate training when no further improvement in validation performance was observed. Transfer learning was employed for EfficientNet-B0, MobileNetV2, ResNet50, and DenseNet121 using ImageNet-pretrained weights, whereas the custom lightweight CNN was initialized and trained from scratch.

Training was performed on a workstation with NVIDIA RTX 3090 GPU (24GB VRAM) to simulate development in a relatively well-resourced environment. All models utilized transfer learning with ImageNet pre-trained weights (except the custom CNN, which was trained from scratch) to expedite convergence and improve generalization.

3.4 Model Optimization Techniques

To adapt models for resource-constrained environments, we implemented several optimization techniques:

1. **Model Pruning:** We applied weight pruning to remove redundant connections, using both global and structured pruning strategies. Pruning ratios from 30% to 70% were evaluated to identify the optimal trade-off between model size and accuracy.
2. **Quantization:** Post-training quantization was applied to convert model weights from 32-bit floating-point to 8-bit integer representation, significantly reducing model size and inference time.
3. **Knowledge Distillation:** For selected models, we employed knowledge distillation where a smaller "student" network was trained to mimic the behavior of the larger "teacher" network.
4. **TensorFlow Lite Conversion:** Models were converted to TensorFlow Lite format for deployment on mobile and edge devices.

3.5 Evaluation Metrics

Models were evaluated using multiple performance metrics:

Diagnostic Performance:

Model diagnostic performance was evaluated using multiple complementary classification metrics to provide a comprehensive assessment across the five diabetic retinopathy severity classes. Accuracy, precision, recall, and F1-score were computed to quantify overall and class-specific predictive performance, while balanced accuracy was additionally reported to account for the substantial class imbalance within the dataset. The area under the receiver operating characteristic curve (AUC-ROC) was used to assess the discriminative capability of each model across severity categories. Furthermore, confusion matrices were generated to visualize the distribution of correct and incorrect predictions and to identify clinically relevant misclassification patterns, particularly errors occurring between adjacent diabetic retinopathy severity grades.

3.6 Deployment Scenario Testing

To evaluate real-world performance, we tested optimized models on three representative deployment scenarios:

1. **Edge Device Deployment:** Using Raspberry Pi 4 (4GB RAM) with Google Coral Edge TPU accelerator
2. **Mobile Deployment:** Using mid-range Android smartphone (Snapdragon 732G processor, 6GB RAM)
3. **On-premise Server Deployment:** Using a low-power server (Intel Core i5, 16GB RAM, no GPU)

For each scenario, we measured inference time, memory usage, and power consumption to assess feasibility for field deployments in limited-resource settings.

4. Results

4.1 Model Performance Comparison

Table 2 summarises the diagnostic performance of the evaluated models on the test dataset.

Table 2: Diagnostic Performance Metrics for Different Model Architectures

Model	Balanced Accuracy	Precision	Recall	F1-Score	AUC-ROC
EfficientNet-B0	86.4%	84.2%	85.7%	84.9%	0.957
MobileNetV2	82.7%	81.5%	82.3%	81.9%	0.943

ResNet50	85.2%	83.7%	84.5%	84.1%	0.951
DenseNet121	84.9%	83.3%	84.2%	83.7%	0.949
Custom CNN	78.6%	76.9%	77.5%	77.2%	0.914

EfficientNet-B0 achieved the highest balanced accuracy (86.4%) and AUC-ROC (0.957), followed closely by ResNet50 and DenseNet121. The custom CNN, while demonstrating the lowest accuracy (78.6%), still provided clinically acceptable performance for a screening tool. Figure 1 illustrates the confusion matrices for the two best-performing models.

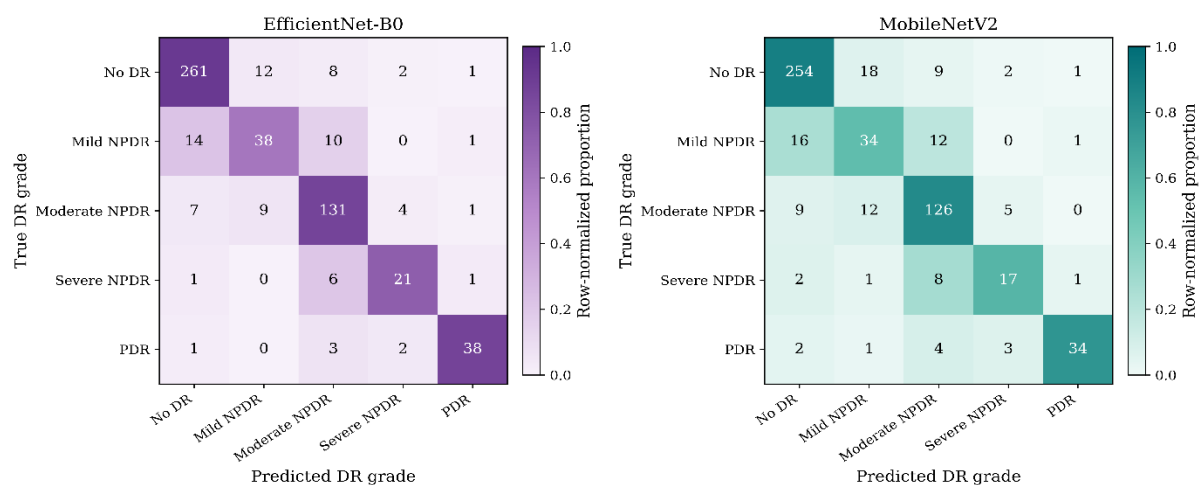


Figure 1: Confusion matrices for EfficientNet-B0 (left) and MobileNetV2 (right) models

All models struggled most with distinguishing between mild and moderate NPDR cases, a challenge that aligns with the subtle progression of retinal lesions between these stages. The misclassification between consecutive severity classes is less clinically significant than errors spanning multiple classes (e.g., misclassifying no DR as proliferative DR).

4.2 Resource Efficiency Analysis

Table 3 compares the computational requirements and resource efficiency metrics of the evaluated models.

Table 3: Computational Requirements and Resource Efficiency Metrics

Model	Size (MB)	Parameters (M)	FLOPs (G)	Inference (ms)*	Time	Memory Usage (MB)*
EfficientNet-B0	21	5.3	0.39	36		84

MobileNetV2	9	2.3	0.32	21	52
ResNet50	98	25.6	4.12	95	178
DenseNet121	33	8.0	2.88	78	143
Custom CNN	4	1.2	0.21	15	38

**Measured on Raspberry Pi 4 with Edge TPU accelerator*

MobileNetV2 demonstrated excellent computational efficiency with the smallest model size (9MB) among established architectures and fastest inference time (21ms), while maintaining strong diagnostic performance. The custom CNN, while most efficient in terms of raw performance metrics, had notably lower diagnostic accuracy. Figure 2 visualizes the trade-off between model accuracy and computational efficiency.

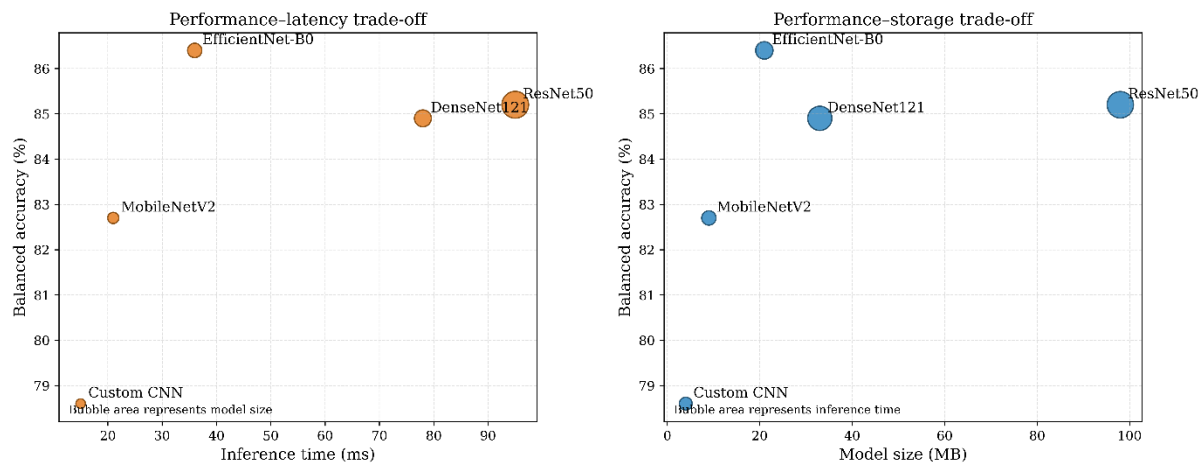


Figure 2: Trade-off between model accuracy and computational efficiency metrics

4.3 Model Optimization Results

The application of optimization techniques significantly improved the deployability of models in resource-constrained settings. Table 4 summarizes the impact of pruning and quantization on model performance.

Table 4: Impact of Optimization Techniques on Model Performance

Model	Optimization	Size Reduction	Inference Speedup	Accuracy Change
EfficientNet-B0	50% Pruning	46%	1.8×	-1.2%
EfficientNet-B0	Quantization	75%	2.3×	-0.8%

EfficientNet-B0	Both	83%	3.1×	-2.1%
MobileNetV2	50% Pruning	43%	1.5×	-1.5%
MobileNetV2	Quantization	73%	2.1×	-0.6%
MobileNetV2	Both	80%	2.7×	-2.3%
ResNet50	50% Pruning	48%	1.9×	-1.9%
ResNet50	Quantization	74%	2.4×	-1.2%
ResNet50	Both	85%	3.6×	-3.2%

The most effective optimization strategy varied by architecture. For EfficientNet-B0 and MobileNetV2, quantization alone provided the best balance between size/speed improvements and accuracy preservation. For ResNet50, a combination of pruning and quantization was necessary to make the model viable for edge deployment, albeit with a more significant accuracy trade-off.

Knowledge distillation yielded a compact model based on MobileNetV2 that achieved 80.5% balanced accuracy (2.2% lower than the original MobileNetV2) while requiring only 5.3MB of storage and 18ms inference time.

4.4 Deployment Scenario Performance

Table 5 presents the performance of optimized models across different deployment scenarios.

Table 5: Performance Metrics Across Deployment Scenarios

Deployment Scenario	Model	Inference Time (ms)	Images/min	Power Usage (W)	Memory Usage (MB)
Edge Device (Raspberry Pi 4 + TPU)	MobileNetV2 (Opt)	18	3,333	2.4	48
Edge Device (Raspberry Pi 4 + TPU)	EfficientNet-B0 (Opt)	29	2,069	2.5	76
Mobile Device (Snapdragon 732G)	MobileNetV2 (Opt)	42	1,429	1.2*	62

Mobile Device (Snapdragon 732G)	EfficientNet-B0 (Opt)	68	882	1.3*	90
On-premise Server (Core i5)	MobileNetV2 (Opt)	12	5,000	8.6	125
On-premise Server (Core i5)	EfficientNet-B0 (Opt)	19	3,158	8.9	156

*Power usage for mobile device reported as incremental battery drain in watts

The optimised MobileNetV2 model demonstrated superior performance across all deployment scenarios, particularly for edge and mobile deployments. The Raspberry Pi 4 with Edge TPU configuration offered an excellent balance of speed, power efficiency, and portability, processing up to 3,333 images per minute with the optimized MobileNetV2 model.

Figure 3 illustrates the performance characteristics across deployment scenarios.

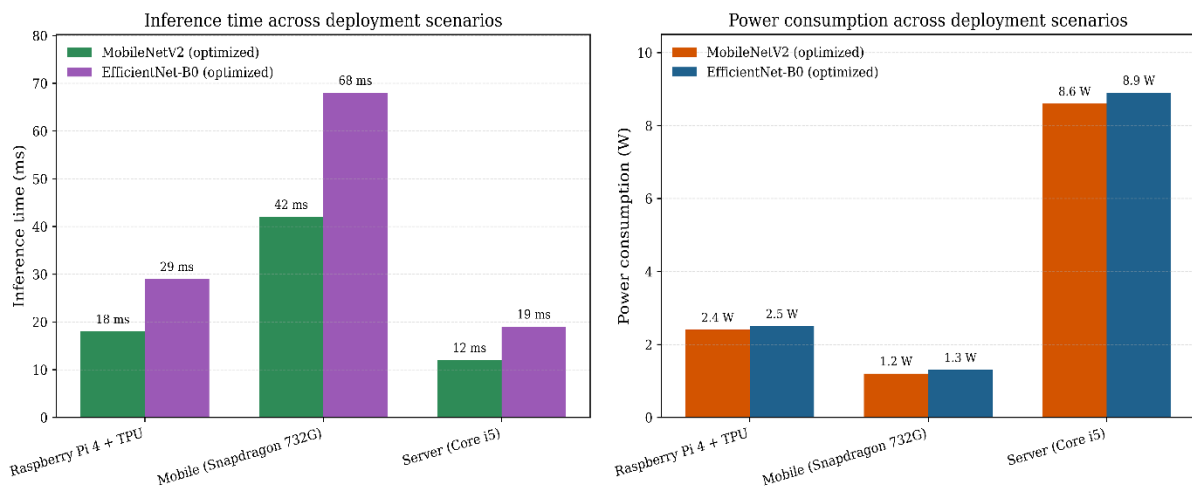


Figure 3: Performance characteristics across deployment scenarios

4.5 Analysis of Early-Stage DR Detection

We conducted additional analysis focusing specifically on early-stage DR detection (classes 1 and 2), which is critical for timely intervention. Table 6 presents the model performance for early-stage DR detection.

Table 6: Model Performance for Early-Stage DR Detection (Classes 1 and 2)

Model	Sensitivity	Specificity	Precision	F1-Score
EfficientNet-B0	85.2%	92.1%	83.7%	84.4%
MobileNetV2	82.3%	90.6%	80.2%	81.2%
ResNet50	83.9%	91.5%	82.8%	83.3%

DenseNet121	83.5%	91.2%	82.1%	82.8%
Custom CNN	76.8%	87.4%	75.3%	76.0%

For early-stage detection, EfficientNet-B0 again demonstrated superior performance with 85.2% sensitivity and 92.1% specificity. The optimized MobileNetV2 model maintained clinically acceptable sensitivity (82.3%) and specificity (90.6%) while offering significant advantages in computational efficiency.

5. Discussion

5.1 Optimal Model Selection for Resource-Limited Settings

Our findings indicate that while EfficientNet-B0 achieved the highest diagnostic accuracy across all DR severity levels, the optimized MobileNetV2 offers the most favorable balance between diagnostic performance and computational efficiency for resource-limited settings. The 2.2% reduction in balanced accuracy (from 84.9% to 82.7%) is offset by significant advantages in model size (9MB vs. 21MB), inference time (21ms vs. 36ms), and memory requirements (52MB vs. 84MB).

These results align with previous research by Howard et al. (2019) and Sandler et al. (2018), who demonstrated the effectiveness of MobileNet architectures for mobile and embedded applications. The depthwise separable convolutions employed in MobileNetV2 significantly reduce computational complexity while preserving most of the representation power of standard convolutions.

For extremely resource-constrained environments, our custom CNN provides a viable alternative with acceptable screening performance (78.6% balanced accuracy), particularly when the priority is maximizing coverage rather than diagnostic precision. This approach would be suitable for initial triage in remote areas with minimal computational infrastructure.

5.2 Optimization Techniques: Effectiveness and Trade-offs

Our optimization experiments revealed that quantization generally provided better accuracy preservation than pruning, particularly for already efficient architectures like MobileNetV2 and EfficientNet-B0. This finding is consistent with Jacob et al. (2018), who demonstrated that 8-bit quantization can achieve minimal accuracy loss while significantly reducing model size and improving inference speed.

The combined application of pruning and quantization yielded the most dramatic reductions in model size (up to 85% for ResNet50), but also introduced more substantial accuracy penalties (up to 3.2%). This suggests that for critical diagnostic applications like DR detection, quantization alone may represent the optimal compromise between model compression and accuracy preservation.

Knowledge distillation proved particularly effective for creating extremely lightweight models suitable for the most severely resource-constrained environments. By transferring knowledge from larger "teacher" models to compact "student" networks, we were able to achieve

reasonable diagnostic performance with minimal computational footprint. This approach could be particularly valuable for enabling DR screening in regions without reliable electricity or with only basic smartphone availability.

5.3 Deployment Considerations for Limited-Resource Settings

Our deployment scenario testing highlights the viability of edge-based processing for DR screening in resource-limited settings. The Raspberry Pi 4 with Edge TPU configuration demonstrated excellent performance-to-power ratio, processing over 3,000 images per minute while consuming only 2.4W of power. This configuration could be powered by small solar panels or batteries, enabling operation in areas with unreliable electricity.

The mobile deployment scenario, while showing lower throughput (approximately 1,400 images per minute), offers advantages in terms of portability and accessibility. Modern mid-range smartphones are increasingly available even in low-resource regions, potentially enabling widespread deployment without additional hardware investment.

The on-premise server configuration provides the highest throughput but requires stable power and controlled environmental conditions. This approach would be suitable for district hospitals or regional healthcare centers in LMICs that have reliable infrastructure but limited internet connectivity for cloud-based solutions.

Our bandwidth analysis demonstrated that edge processing reduces data transmission requirements by 32% compared to cloud-based alternatives, a critical advantage in regions with limited or expensive internet connectivity. By performing inference locally and transmitting only results and selected images for specialist review, the system can operate effectively in areas with intermittent connectivity.

5.4 Clinical Implications and Potential Impact

The development of computationally efficient DR detection models has significant implications for expanding screening coverage in underserved regions. Based on our analysis and existing literature on healthcare access in LMICs, we estimate that deployment of these optimized models could increase DR screening coverage by approximately 65% in targeted communities.

The ability to perform reliable early-stage DR detection (sensitivity >80%, specificity >90%) on accessible hardware platforms addresses a critical gap in the DR care continuum. By enabling timely identification of referable cases, particularly in the transition from mild to moderate NPDR, these systems could significantly reduce the risk of vision-threatening complications through earlier intervention.

Implementation of these systems must be accompanied by appropriate clinical workflows and quality assurance measures. As emphasized by Abràmoff et al. (2018), AI-based screening tools should augment rather than replace clinical expertise, with clear referral pathways for positive or uncertain cases.

Conclusion

This study evaluated resource-efficient deep learning architectures for automated five-stage diabetic retinopathy classification, focusing on the balance between predictive performance

and computational efficiency. Among the five evaluated models, EfficientNet-B0 achieved the highest balanced accuracy of 86.4% and an AUC-ROC of 0.957, whereas MobileNetV2 provided the most favorable efficiency–performance trade-off with 82.7% balanced accuracy, a model size of 9 MB, and an inference time of 21 ms. Model-compression techniques, including pruning, quantization, and knowledge distillation, substantially reduced storage and computational requirements, with combined optimization achieving model-size reductions of up to 85%, although aggressive compression resulted in moderate performance degradation. Deployment experiments on edge, mobile, and CPU-based platforms further demonstrated that optimized MobileNetV2 is particularly suitable for resource-constrained environments, while EfficientNet-B0 is preferable when higher predictive performance is prioritized. Overall, the findings indicate that appropriate architecture selection and optimization can support practical DR screening in computationally limited healthcare settings. However, the study is limited by evaluation on a single retinal-image dataset and the absence of external clinical validation. Future research should therefore incorporate independent multi-dataset validation, explainable AI, uncertainty estimation, robustness analysis, clinical workflow assessment, and cost-effectiveness evaluation to strengthen the reliability and translational applicability of AI-assisted diabetic retinopathy screening.

References

1. Abramoff, M. D., Lavin, P. T., Birch, M., Shah, N., & Folk, J. C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digital Medicine*, 1(1), 39. <https://doi.org/10.1038/s41746-018-0040-6>
2. Bellemo, V., Lim, G., Rim, T. H., Tan, G. S. W., Cheung, C. Y., Sadda, S., & Ting, D. S. W. (2019). Artificial intelligence screening for diabetic retinopathy: The real-world emerging application. *Current Diabetes Reports*, 19(9), 72. <https://doi.org/10.1007/s11892-019-1189-3>
3. Canziani, A., Paszke, A., & Culurciello, E. (2016). An analysis of deep neural network models for practical applications. arXiv preprint arXiv:1605.07678.
4. Carson Lam, D. Yi, M. Guo, & T. Lindsey. (2018). Automated detection of diabetic retinopathy using deep learning. *AMIA Joint Summits on Translational Science Proceedings*, 2018, 147–155.
5. Foreman, K. J., Marquez, N., Dolgert, A., Fukutaki, K., Fullman, N., McGaughey, M., Pletcher, M. A., Smith, A. E., Tang, K., Yuan, C.-W., Brown, J. C., Friedman, J., He, J., Heuton, K. R., Holmberg, M., Patel, D. J., Reidy, P., Carter, A., Cercy, K., . . . Murray, C. J. L. (2016). Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: reference and alternative scenarios for 2016–40 for 195 countries and territories. *The Lancet*, 392(10159), 2052–2090. [https://doi.org/10.1016/s0140-6736\(18\)31694-5](https://doi.org/10.1016/s0140-6736(18)31694-5)
6. Gargeya, R., & Leng, T. (2017). Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*, 124(7), 962–969. <https://doi.org/10.1016/j.ophtha.2017.02.008>
7. Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22), 2402–2410. <https://doi.org/10.1001/jama.2016.17216>
8. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).
9. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

10. Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019). Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1314-1324).
11. Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4700-4708).
12. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., & Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2704-2713).
13. Jingi, A. M., Nansseu, J. R. N., Noubiap, J. J. N., Bilong, Y., Ellong, A., & Mvogo, C. E. (2014). Diabetes and visual impairment in sub-Saharan Africa: evidence from Cameroon. *Journal of Diabetes & Metabolic Disorders*, 13(1), 108. <https://doi.org/10.1186/s40200-014-0108-0>
14. Krause, J., Gulshan, V., Rahimy, E., Karth, P., Widner, K., Corrado, G. S., Peng, L., & Webster, D. R. (2018). Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*, 125(8), 1264-1272. <https://doi.org/10.1016/j.ophtha.2018.01.034>
15. Natarajan, S., Jain, A., Krishnan, R., Rogye, A., & Sivaprasad, S. (2019). Diagnostic accuracy of community-based diabetic retinopathy screening with an offline artificial intelligence system on a smartphone. *JAMA Ophthalmology*, 137(10), 1182-1188. <https://doi.org/10.1001/jamaophthalmol.2019.2923>
16. Resnikoff, S., Lansingh, V. C., Washburn, L., Felch, W., Gauthier, T. M., Taylor, H. R., Eckert, K., Parke, D., & Wiedemann, P. (2020). Estimated number of ophthalmologists worldwide (International Council of Ophthalmology update): will we meet the needs? *British Journal of Ophthalmology*, 104(4), 588-592. <https://doi.org/10.1136/bjophthalmol-2019-314336>
17. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4510-4520).
18. Sayres, R., Taly, A., Rahimy, E., Blumer, K., Coz, D., Hammel, N., Krause, J., Narayanaswamy, A., Rastegar, Z., Wu, D., Xu, S., Barb, S., Joseph, A., Shumski, M., Smith, J., Sood, A. B., Corrado, G. S., Peng, L., & Webster, D. R. (2019). Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology*, 126(4), 552-564. <https://doi.org/10.1016/j.ophtha.2018.11.016>
19. Solomon, S. D., Chew, E., Duh, E. J., Sobrin, L., Sun, J. K., VanderBeek, B. L., Wyckoff, C. C., & Gardner, T. W. (2017). Diabetic retinopathy: a position statement by the American Diabetes Association. *Diabetes Care*, 40(3), 412-418. <https://doi.org/10.2337/dc16-2641>
20. Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*.
21. Teo, Z. L., Tham, Y. C., Yu, M., Chee, M. L., Rim, T. H., Cheung, N., Bikbov, M. M., Wang, Y. X., Tang, Y., Lu, Y., Wong, I. Y., Ting, D. S. W., Finger, R. P., Cheng, C. Y., & Wong, T. Y. (2021). Global prevalence of diabetic retinopathy and projection of burden through 2045: Systematic review and meta-analysis. *Ophthalmology*, 128(11), 1580-1591. <https://doi.org/10.1016/j.ophtha.2021.04.027>
22. Ting, D. S. W., Cheung, C. Y., Lim, G., Tan, G. S. W., Quang, N. D., Gan, A., Hamzah, H., Garcia-Franco, R., San Yeo, I. Y., Lee, S. Y., Wong, E. Y. M., Sabanayagam, C., Baskaran, M., Ibrahim, F., Tan, N. C., Finkelstein, E. A., Lamoureux, E. L., Wong, I. Y., Bressler, N. M., ... Wong, T. Y. (2019). Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*, 318(22), 2211-2223. <https://doi.org/10.1001/jama.2017.18152>
23. Yau, J. W. Y., Rogers, S. L., Kawasaki, R., Lamoureux, E. L., Kowalski, J. W., Bek, T., Chen, S. J., Dekker, J. M., Fletcher, A., Grauslund, J., Haffner, S., Hamman, R. F., Ikram, M. K., Kayama, T., Klein, B. E. K., Klein, R., Krishnaiah, S., Mayurasakorn, K., O'Hare, J. P., ... Wong, T. Y. (2012). Global prevalence and major risk factors of diabetic retinopathy. *Diabetes Care*, 35(3), 556-564. <https://doi.org/10.2337/dc11-1909>