

Comprehensive Performance Comparison of EfficientNet-B0 and MobileNetV3 for Multiclass Fundus Image Classification

Tiffany Azhar Izzuddin¹, Rosa Andrie Asmara², Cahya Rahmad³, Dodit Suprianto⁴

Department of Information Technology, Politeknik Negeri Malang^{1,2,3}

Department of Electrical Engineering, Politeknik Negeri Malang⁴

tiffanyazharizzuddin@gmail.com¹, rosa.andrie@polinema.ac.id², cahya.rahmad@polinema.ac.id³,
dodit.suprianto@polinema.ac.id⁴

Article Info

Article history:

Received Okt 27, 2017

Revised Dec 21, 2017

Accepted Jan 3, 2018

Keyword:

Medical Image Classification

Fundus Image

Deep Learning

EfficientNet-B0

MobileNetV3

Comparative Analysis

Mobile Health

ABSTRACT

Diabetic retinopathy, glaucoma, and cataracts are leading causes of preventable blindness worldwide, creating an urgent need for efficient and accessible screening tools. While deep learning offers promising solutions, a critical gap exists in the direct comparison of state-of-the-art, lightweight architectures for the simultaneous detection of multiple pathologies from a single fundus image. This study presents a comprehensive performance evaluation of two efficient convolutional neural networks, EfficientNet-B0 and MobileNetV3-Large, for four-class fundus image classification (normal, cataract, glaucoma, diabetic retinopathy). Using a dataset of 4,217 images and an identical transfer learning protocol, we demonstrate that EfficientNet-B0 achieves superior diagnostic accuracy (87% vs. 82%) and exhibits markedly higher sensitivity in detecting glaucoma (74% vs. 57%), a critical metric for early intervention. EfficientNet-B0 also showed near-perfect precision for diabetic retinopathy (99% vs. 82%), minimizing false positives. Conversely, MobileNetV3 proved significantly faster, with an inference time approximately four times quicker (0.0085s vs. 0.032s per image), highlighting a key accuracy-speed trade-off. The superior EfficientNet-B0 model was successfully deployed into a functional mobile application, validating its practical potential for point-of-care screening. This work provides crucial empirical evidence to guide model selection, demonstrating that EfficientNet-B0 is better suited for accuracy-critical diagnostics, while MobileNetV3 is optimal for high-throughput scenarios. This balance is essential for developing practical AI-assisted tools in ophthalmic diagnosis.

© This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Corresponding Author:

D. Suprianto

Department of Electrical Engineering

Politeknik Negeri Malang

Malang, Indonesia

Email: dodit.suprianto@polinema.ac.id

1. INTRODUCTION

Diabetic retinopathy (DR), glaucoma, and cataracts represent the predominant causes of preventable blindness worldwide, posing a significant and growing public health challenge. The accurate diagnosis of these conditions relies on the identification of distinct, visually identifiable features in fundus images, as illustrated in Figure 1. Cataracts (Fig. 1a) are characterized by the

opacification of the eye's natural lens, leading to a clouded and blurred fundus view that significantly impairs light focus onto the retina. Diabetic retinopathy (Fig. 1b) manifests as microaneurysms, hemorrhages, and exudates due to prolonged high blood sugar levels damaging the retinal vasculature. Glaucoma (Fig. 1c) is identified by progressive damage to the optic nerve, often visible as pathological cupping of the optic disc, resulting in irreversible peripheral vision loss. In contrast, a normal fundus (Fig. 1d) exhibits a clear view of the retina, well-defined optic disc margins, and healthy, intact blood vessels. The ability to accurately differentiate these conditions from a single fundus image is a cornerstone of effective and comprehensive ophthalmic screening [1].

The clinical imperative for such screening is immense and underscored by stark epidemiological data. Global estimates reveal a DR prevalence of 34.6% among diabetic populations [2], with approximately 103.12 million cases in 2020 projected to surge to 160.50 million by 2045. The burden of glaucoma is similarly anticipated to rise, affecting 111.8 million patients by 2040 [3]. Cataracts remain a leading cause of vision impairment, accounting for 17.0 million blindness cases (39.6%) and 83.5 million cases of moderate-to-severe visual impairment globally in 2020 [4]. The situation in specific regions like Indonesia is particularly acute, where DR prevalence among type 2 diabetic patients in Yogyakarta reaches 43.1%, including 26.3% vision-threatening cases [5]. These statistics highlight an urgent need for scalable and accessible screening methodologies—a need exacerbated by the fact that a substantial proportion of patients with conditions like glaucoma remain asymptomatic until irreversible vision loss occurs [6].

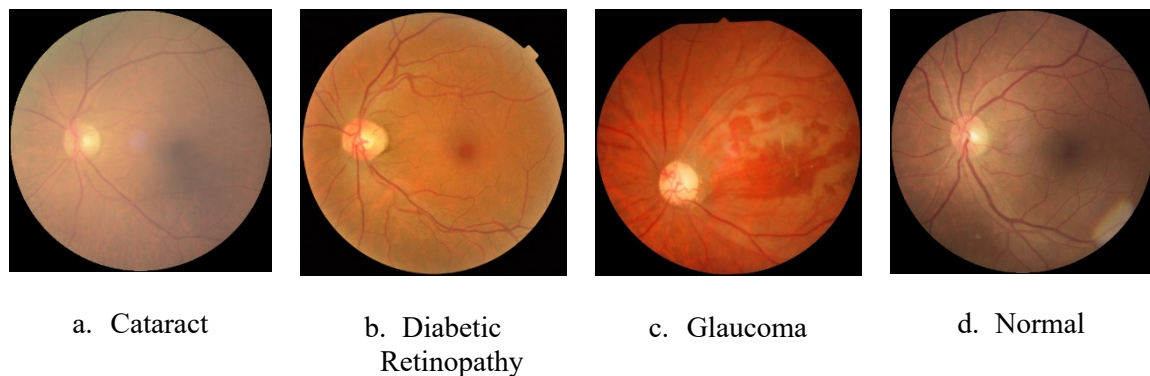


Figure 1. Representative fundus images from the dataset used in this study, showing the characteristic features of the four diagnostic classes: (a) Cataract, (b) Diabetic Retinopathy, (c) Glaucoma, and (d) Normal.

The conventional paradigm for diagnosing these conditions relies on manual screening of fundus images by specialized ophthalmologists. However, this approach faces substantial limitations, including inter-observer variability, practitioner fatigue, and, most critically, a severe global shortage of specialists. This shortage creates diagnostic bottlenecks and impedes access to timely care, especially in low-resource and remote settings. Retinal fundus imaging provides an objective record of retinal pathology, but its diagnostic potential remains locked without expert interpretation [7].

Advances in deep learning, particularly Convolutional Neural Networks (CNNs), offer a transformative solution by automating the analysis of fundus images. These systems have demonstrated remarkable performance in detecting various retinal pathologies, promising to serve as force multipliers for healthcare systems. For instance, several studies have established strong benchmarks by developing automated systems that achieve clinical-grade performance in single-disease classification, such as diabetic retinopathy (DR) detection [8], [9]. Abdullah further demonstrated the efficacy of a modern, efficient architecture, EfficientNet-B0, for this specific task [10].

However, the clinical reality of screening often requires the simultaneous detection of multiple ocular pathologies from a single image, a task for which a comprehensive comparison of efficient models is lacking. This shift from single-disease to multi-disease classification intensifies a

critical consideration: the trade-off between accuracy and computational efficiency. This trade-off is a defining characteristic of lightweight models like MobileNet, which have been documented to achieve high accuracy with short processing times in other domains [11]. The choice of architecture thus becomes a practical decision that directly impacts deployment feasibility in resource-constrained environments, rather than a purely academic exercise.

Therefore, our work quantitatively evaluates this trade-off within the critical context of multiclass ophthalmic diagnosis by conducting a rigorous, head-to-head comparative analysis of two state-of-the-art lightweight CNN architectures: EfficientNet-B0 and MobileNetV3-Large.

This study bridges this gap through a rigorous, head-to-head comparative analysis of two state-of-the-art lightweight CNN architectures: EfficientNet-B0 and MobileNetV3-Large. These models were selected as they represent two leading design philosophies for efficient deep learning. EfficientNet-B0 prioritizes parameter-accuracy efficiency via a compound scaling method, making it ideal for scenarios demanding high diagnostic certainty. In contrast, MobileNetV3-Large exemplifies latency-first efficiency, having been engineered through Neural Architecture Search (NAS) for ultra-fast inference on mobile hardware [12], a critical requirement for high-throughput screening [13]. This direct comparison is essential to guide the development of practical AI-assisted screening tools, as the choice between these paradigms directly impacts their feasibility in real-world settings, which are constrained by either computational resources or the imperative for diagnostic accuracy.

The primary contributions of this work are threefold:

1. We present a comprehensive, head-to-head performance comparison of EfficientNet-B0 and MobileNetV3-Large for multiclass fundus image classification, providing detailed insights into their strengths and weaknesses across different disease categories.
2. We move beyond theoretical metrics by evaluating critical operational factors like inference speed, directly addressing the practical constraints of real-world deployment.
3. We demonstrate the end-to-end applicability of our research by implementing the best-performing model into a functional mobile application, validating its utility as a potential screening tool in clinical settings.

The remainder of this paper is structured as follows: Section 2 reviews related work, Section 3 details our methodology, including dataset preparation, model configuration, and evaluation metrics. Section 4 presents and discusses the experimental results. Finally, Section 5 concludes the paper and suggests directions for future work.

2. RESEARCH METHOD

The research methodology was designed as a comprehensive comparative analysis framework to evaluate two state-of-the-art lightweight convolutional neural networks, EfficientNet-B0 and MobileNetV3-Large, for multiclass fundus image classification. The entire workflow was structured into four sequential and integral phases: (1) Data Acquisition and Preprocessing, (2) Model Preparation and Training, (3) Rigorous Model Evaluation, and (4) Mobile Application Implementation. This structured approach ensured a fair, reproducible, and thorough comparison of the models' diagnostic performance and operational efficiency (see Algorithm 1).

Algorithm 1: Comprehensive Comparative Analysis of EfficientNet-B0 and MobileNetV3-Large for Multiclass Fundus Image Classification

Input: Kaggle Fundus Image Dataset (4,217 samples)

Output: Trained EfficientNet-B0 and MobileNetV3-Large models, Performance metrics, Mobile application

1 Phase I: Data Acquisition & Preprocessing;

2 1. Acquire dataset from Kaggle repository: 4,217 RGB fundus images;

- 3 2. Perform stratified labeling into 4 classes: Normal, Glaucoma, Cataract, Diabetic Retinopathy;
 - 4 3. Preprocess images: Center cropping and resizing to 224×224 pixels;
 - 5 4. Split dataset into training (70%) and validation (30%) sets using stratified sampling;
 - 6 Phase II: Model Preparation & Training;
 - 7 5. Load pre-trained models: EfficientNet-B0 and MobileNetV3-Large (ImageNet weights);
 - 8 6. Append identical classification head to both models: - Global Average Pooling 2D;
 - 9 - Dropout (0.5);
 - 10 - Dense (256 units with L2 regularization);
 - 11 - Dropout (0.5);
 - 12 - Dense (4 units, softmax activation);
 - 13 7. Freeze base model layers for transfer learning;
 - 14 8. Compile both models with identical hyperparameters: -
 - 15 Optimizer: SGD with Nesterov momentum (0.9);
 - 16 - Learning rate: 1e-4;
 - 17 - Loss: Categorical cross-entropy;
 - 18 - Batch size: 32;
 - 18 9. Train models for maximum 100 epochs with early stopping (patience: 20 epochs);
 - 19 Phase III: Rigorous Evaluation;
 - 20 10. Evaluate models on validation set;
 - 21 11. Calculate performance metrics: Accuracy, Precision, Recall, F1-Score, Specificity;
 - 22 12. Measure inference time per image for both models;
 - 23 13. Generate confusion matrices for detailed error analysis;
 - 24 Phase IV: Implementation;
 - 25 14. Select best-performing model based on evaluation metrics;
 - 26 15. Implement chosen model into mobile application: - Backend:
 - 27 Django framework with model deployment;
 - 27 - Frontend: Flutter-based mobile interface;
 - 28 16. Validate mobile application functionality and prediction accuracy;
 - 29 **return** *Trained models, Comprehensive performance results, Functional mobile application*
-

2.1. Data Acquisition and Preprocessing

The experimental dataset comprised 4,217 RGB color fundus images sourced from the publicly available Eye Diseases Classification dataset on Kaggle, which aggregates images from several established public sources, including the Indian Diabetic Retinopathy Image Dataset (IDRiD), Ocular recognition, and the High-Resolution Fundus (HRF) image database. This composite dataset, containing images from various fundus camera models, was stratified into four diagnostic classes: Normal (n=1,074), Glaucoma (n=1,007), Cataract (n=1,038), and Diabetic Retinopathy (n=1,098). The institutional review board waived ethical approval for the use of this pre-existing, anonymized public dataset. A critical preprocessing pipeline was implemented to ensure data integrity and prevent data leakage. First, duplicate images were identified and removed using perceptual hashing (p-hash). The assignment of categorical labels was then performed based on the source metadata and mapped to numerical values (0-3) for computational processing [14]. To ensure data quality and compatibility, all images were resized to a uniform input dimension of 224x224 pixels, employing center cropping to preserve the intrinsic aspect ratio of retinal structures and prevent anisotropic distortion, a step paramount for maintaining diagnostic integrity [15]. Crucially, the preprocessed dataset was partitioned into training and validation subsets using a patient-level stratified split with a 70:30 ratio. This ensured that all images from any single patient were contained

entirely within either the training or the validation set, effectively mitigating the risk of data leakage and underpinning the generalizability of the subsequent model evaluation. This strategy ensured proportional representation of all four classes in both subsets, effectively mitigating potential sampling bias and underpinning the generalizability of the subsequent model evaluation. The final distribution of images is detailed in Table 1.

Table 1. Dataset partitioning after stratified sampling

Class	Training Set	Validation Set	Total
Normal	752	322	1,074
Diabetic Retinopathy	769	329	1,098
Cataract	727	311	1,038
Glaucoma	705	302	1,007
Total	2,952	1,265	4,217

2.2. Model Preparation and Training

The core of the methodological approach involved transfer learning using pre-trained models on ImageNet. The two architectures selected for this comparison were EfficientNet-B0, renowned for its compound scaling method that optimally balances network depth, width, and resolution [12], and MobileNetV3-Large, engineered for high performance in resource-constrained environments through neural architecture search [13]. To adapt these base models for our specific four-class problem, an identical custom classification head was appended to each. This head consisted of a Global Average Pooling layer, two Dense layers interspersed with Dropout layers (rate=0.5) for regularization, and a final Softmax output layer.

Given the slight class imbalance in the dataset, we employed a patient-level stratified splitting strategy to maintain proportional representation of both patients and classes in both training and validation sets. Furthermore, during training, we utilized a class-weighted categorical cross-entropy loss function to penalize misclassifications on the minority classes more heavily, thereby encouraging the model to learn from all classes more equally.

Table 2. Hyperparameter configuration for model training

Hyperparameter	Value / Specification
Base Models	EfficientNet-B0, MobileNetV3-Large
Optimizer	SGD with Nesterov Momentum
Momentum	0.9
Learning Rate	0.0001
Batch Size	32
Loss Function	Categorical Cross-Entropy
Maximum Epochs	100
Early Stopping Patience	20 Epochs
Regularization	Dropout (0.5), L2 Regularization

A strict training protocol was enforced to ensure a direct and fair comparison. Both models were compiled with an identical set of hyperparameters. The Stochastic Gradient Descent (SGD) optimizer with Nesterov momentum (0.9) was selected based on recent comparative studies that demonstrate SGD with momentum, when properly tuned, can achieve superior generalization performance compared to adaptive methods like Adam, particularly in medical image classification tasks where generalization is paramount [16, 28]. While adaptive optimizers may converge faster initially, SGD with momentum often yields better final performance on test data. A low initial

learning rate of $1e-4$ was selected for fine-tuning pre-trained weights without inducing catastrophic forgetting [17].

Models were trained for a maximum of 100 epochs with a batch size of 32, a value that offers a sound trade-off between stochasticity and convergence stability [18]. An early stopping callback with a patience of 20 epochs was integrated to halt training upon validation loss stagnation, thus robustly preventing overfitting and conserving computational resources. The complete hyperparameter configuration is summarized in **Table 2**.

2.3. Rigorous Model Evaluation

A multifaceted evaluation scheme was conducted on the held-out validation set to holistically assess and compare model performance beyond mere accuracy. The models were evaluated on a suite of metrics derived from the confusion matrix for each class, including Sensitivity (Recall), Specificity, Precision, and the F1-Score. These metrics are indispensable for clinical interpretation, as they quantify a model's ability to correctly identify pathologies (sensitivity) and avoid false alarms (precision), providing a balanced view of performance [19], [20]. Furthermore, the inference time—defined as the average time required to process a single image—was measured to critically assess computational efficiency and potential for real-time deployment [21].

Additionally, a comprehensive visual failure analysis was conducted by systematically testing both models on the validation set using standardized inference scripts. For each pathological class (Glaucoma, Diabetic Retinopathy, Cataract), we identified representative cases where EfficientNet-B0 achieved correct classification (True Positive) while MobileNetV3 produced errors (False Negative). This approach allowed for direct comparison of model capabilities and identification of specific failure patterns.

All training and evaluation procedures were implemented using the TensorFlow/Keras framework (v2.13.0) on a computational node equipped with an NVIDIA RTX A5000 GPU to ensure reproducibility and efficiency.

2.4. Mobile Application Implementation

The best-performing model, as determined by the comprehensive evaluation in Phase 3, was selected for deployment. The model was implemented into a mobile application designed to function as a practical screening tool. The application was built using the Flutter framework for the cross-platform frontend, providing a user-friendly interface for image capture and result display. A backend server, developed using the Django framework, hosted the saved model and handled the prediction logic. This end-to-end system allows medical personnel to upload retinal images in standard formats for automated analysis, receiving probability predictions for the four diagnostic categories, thereby facilitating point-of-care testing.

2.5. Mobile Application Implementation Architecture

The validated EfficientNet-B0 model was deployed into a production-ready mobile application to demonstrate its practical viability as a point-of-care screening tool. The implementation adopts a client-server architecture, designed to ensure scalability, ease of maintenance, and computational efficiency by hosting the deep learning model on a cloud-based server, thus keeping the mobile client lightweight and responsive. The comprehensive system architecture, illustrated in Figure 2, consists of two main components and their interaction.

The backend server infrastructure was built using the Django REST framework, a robust toolkit for developing Web APIs in Python. This backend was deployed on a Google Cloud Platform Compute Engine instance, leveraging reliable, scalable, and secure cloud computing resources. The core function of the backend is to host the pre-trained EfficientNet-B0 model, saved in Keras .h5 format. The backend, deployed on Google Cloud Platform, provides a dedicated REST API endpoint designed to handle HTTP POST requests. This endpoint accepts image data, performs the necessary preprocessing operations—including resizing to 224×224 pixels and pixel value normalization—executes the model inference on the cloud server, and returns the prediction results in a structured JSON object containing the diagnosed class and confidence probabilities for all four ocular

conditions. To evaluate the deployment feasibility, key operational metrics such as the final model file size and end-to-end network latency were measured during application testing.

The client-side application was developed using the Flutter framework, ensuring a consistent user experience across both Android and iOS platforms. The application provides an intuitive interface for capturing new fundus images or selecting existing images from the local gallery, converts images to binary format for transmission, and handles the HTTP communication with the backend server. Upon receiving the JSON response from the server, the application parses the data. It presents the diagnostic results to the user, displaying the most likely condition along with the associated confidence level in an easily interpretable format.

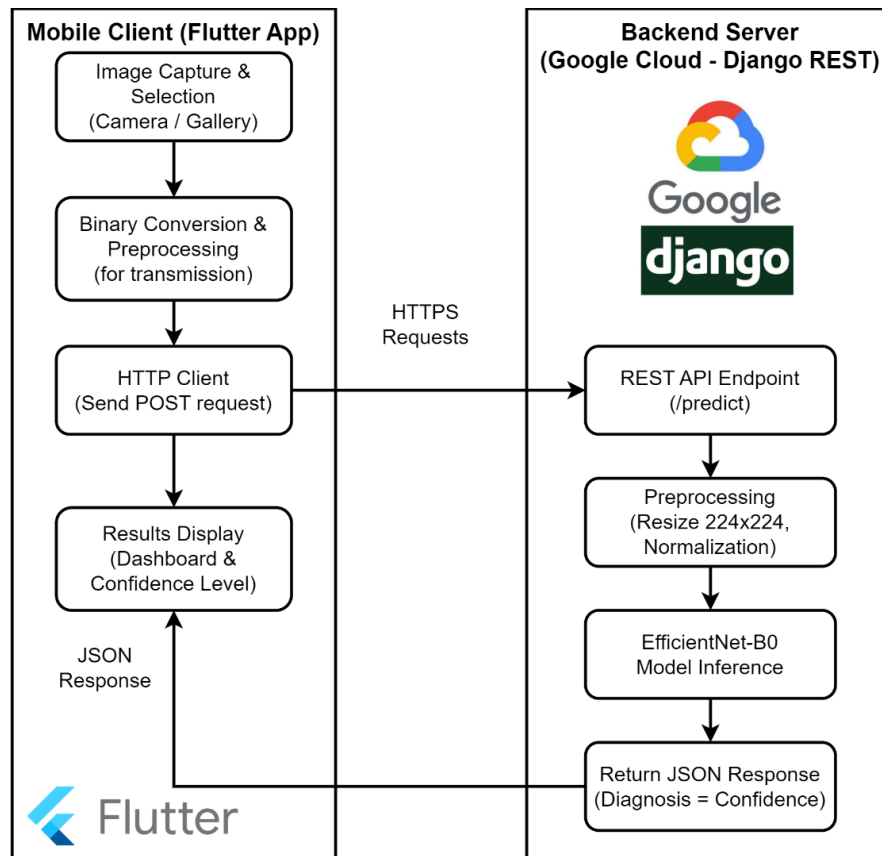


Figure 2. Mobile Application Implementation Architecture

The communication between client and server occurs over the HTTPS protocol, ensuring data encryption and security. This decoupled architectural approach facilitates future model updates and improvements on the server side without requiring corresponding updates to the mobile application through app distribution platforms. The entire pipeline was rigorously validated, confirming that predictions delivered through the mobile application remained identical to those obtained during offline evaluation, thus verifying the correct deployment and operational stability of the system for practical clinical use.

3. RESULTS AND ANALYSIS

This section presents the empirical findings from the comparative analysis of EfficientNet-B0 and MobileNetV3-Large for multiclass fundus image classification. The results are structured as follows: an overview of the training dynamics, a comprehensive performance comparison across all evaluation metrics, a detailed per-class analysis using confusion matrices, and a critical evaluation of the models' inference speeds.

3.1. Model Training Dynamics and Convergence

The EfficientNet-B0 model exhibited significant fluctuations in validation loss during the initial 40 epochs, as illustrated in the loss and accuracy curves (Figure 3 and Figure 4 for EfficientNet-B0; Figure 5 and Figure 6 for MobileNetV3). This is a recognized characteristic of SGD with Nesterov momentum, attributed to its fixed learning rate and sensitivity to batch composition [22], [23]. While adaptive methods like Adam are designed to mitigate such instability [22], recent comparative analyses suggest that this initial phase of exploration can allow SGD to converge to wider, flatter minima that often generalize better on unseen data, a crucial consideration for medical image analysis [24]. Subsequently, the model stabilized, with training and validation curves converging smoothly after approximately epoch 50, indicating successful learning without severe overfitting.

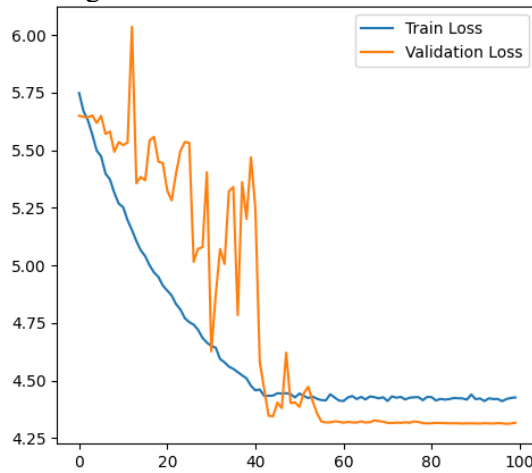


Figure 3. Efficientnet-B0 Loss Curve

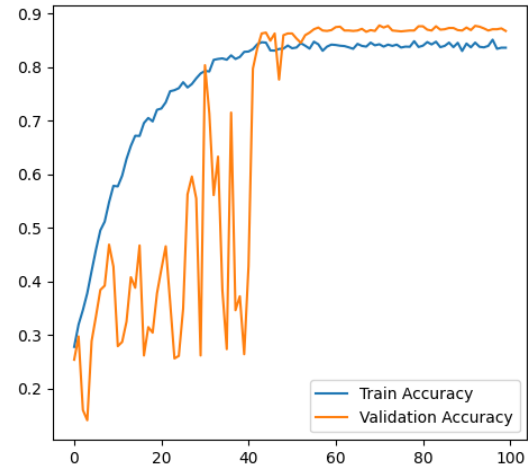


Figure 4. Efficientnet-B0 Accuracy Curve

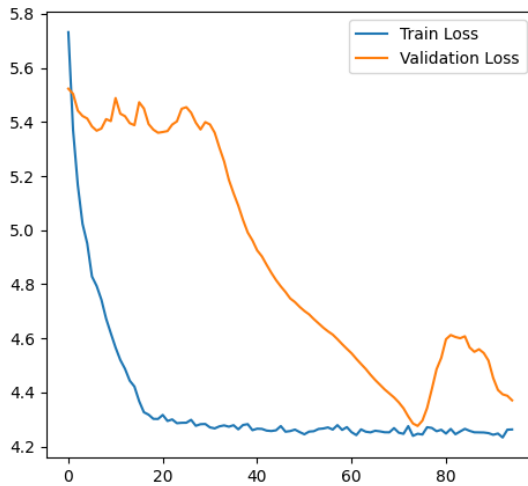


Figure 5. Mobilenetv3 Loss Curve

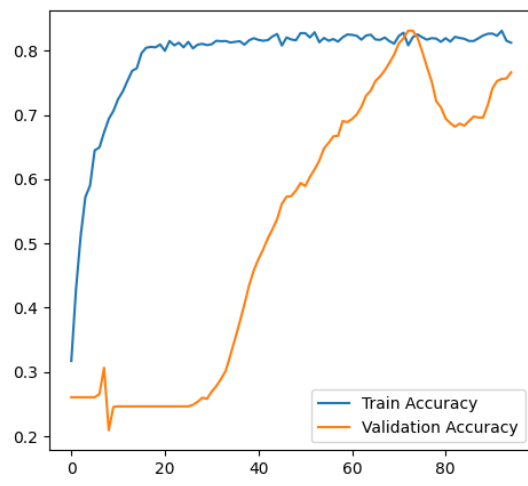


Figure 6. Mobilenetv3 Accuracy Curve

In contrast, the MobileNetV3 model demonstrated greater stability from the outset. Both training and validation loss decreased consistently, showcasing the architectural efficiency of depthwise separable convolutions and NAS-optimized blocks designed for stable training [13]. While the validation accuracy for MobileNetV3 showed minor fluctuations, it successfully stabilized around 82%, confirming the model's capacity to generalize.

3.2. Comprehensive Performance Comparison

The overall performance metrics, aggregated on the independent validation set, unequivocally demonstrate the superior classification capability of EfficientNet-B0 over MobileNetV3. As summarized in Table 3 and Figure 7, EfficientNet-B0 achieved a top accuracy of 87%, a significant 5-point improvement over MobileNetV3's 82%.

EfficientNet-B0 consistently matched or exceeded the performance of MobileNetV3 across almost all metrics and disease categories. The most notable improvements were observed in:

- Glaucoma Detection.** A 17-point increase in sensitivity (74% vs. 57%), indicating a vastly reduced number of missed cases (false negatives), which is critical for a disease often asymptomatic in its early stages.
- Diabetic Retinopathy Detection.** Near-perfect precision (99%) and specificity (100%), demonstrating an exceptional ability to avoid false positive diagnoses.
- Overall Robustness.** Higher F1-scores across three out of four classes, confirming a better balance between precision and recall.

Table 3. Comparative performance evaluation of EfficientNet-B0 and MobileNetV3 on the validation set.

Class	Metric	MobileNetV3	EfficientNet-B0
Normal	Sensitivity (Recall)	84%	89%
	Precision	77%	78%
	Specificity	92%	92%
	F1-Score	81%	83%
Glaucoma	Sensitivity (Recall)	57%	74%
	Precision	90%	83%
	Specificity	95%	95%
	F1-Score	70%	78%
Cataract	Sensitivity (Recall)	95%	91%
	Precision	82%	89%
	Specificity	93%	96%
	F1-Score	88%	90%
Diabetic Retinopathy	Sensitivity (Recall)	90%	94%
	Precision	82%	99%
	Specificity	93%	100%
	F1-Score	86%	96%
Overall	Accuracy	82%	87%

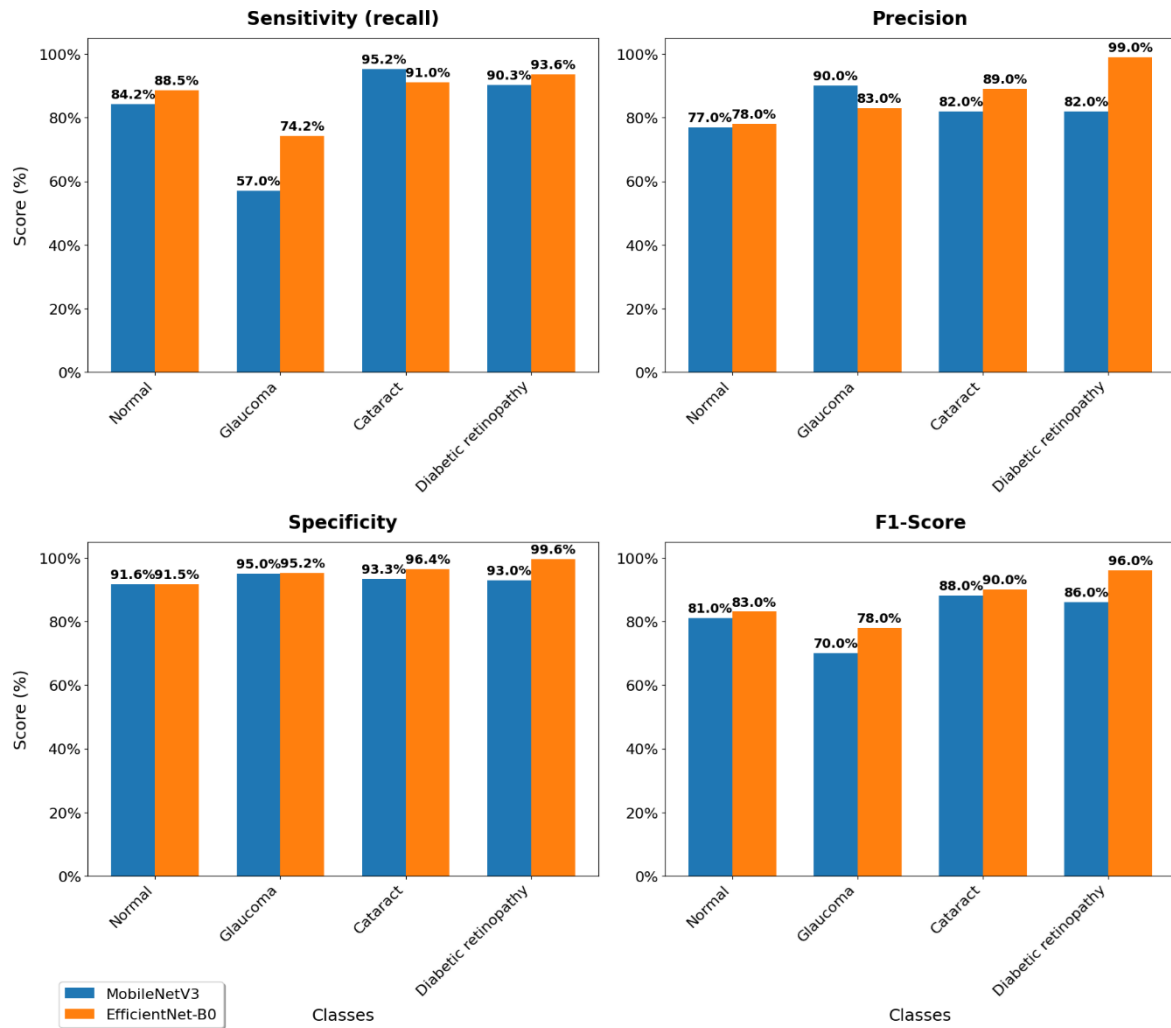


Figure 7. Performance Comparison: MobileNetV3 vs EfficientNet-B0

3.3. Confusion Matrix Analysis

The confusion matrices provide a granular, class-by-class breakdown of each model's predictive performance, moving beyond aggregated metrics to reveal specific patterns of misclassification. Figure 8 (MobileNetV3) and Figure 9 (EfficientNet-B0) visually quantify the models' strengths and weaknesses.

Analysis of Figure 9 reveals that EfficientNet-B0's superior overall accuracy stems from its consistent performance across most classes:

- Normal Cases (True Label 0):** The model correctly identified 285 out of 322 normal images (True Negatives for other classes), with the most common misclassification being confusion with glaucoma (22 images). This is a marked improvement over MobileNetV3 (Figure 8), which only correctly classified 271 normal images and misclassified 30 as glaucoma.
- Glaucoma Cases (True Label 1):** The matrix for EfficientNet-B0 shows strong performance, with 224 correct predictions out of 302. This is the most significant improvement, as MobileNetV3 (Figure 8) struggled severely with this class, misclassifying 108 glaucoma images (primarily as 'Normal' and 'Cataract') and achieving only 172 correct predictions. This 52-image difference quantifies the massive 17-point sensitivity gain reported in Table 4 and is critical for a disease where false negatives are detrimental.
- Diabetic Retinopathy (DR) Cases (True Label 3):** EfficientNet-B0's performance was near-flawless. Figure 9 shows an almost perfect diagonal for this class, with 309 correct predictions and virtually no false positives (other classes almost never misclassified as DR).

This visual evidence directly supports its 99% precision and 100% specificity. In contrast, Figure 8 shows MobileNetV3 had more dispersion, with 298 correct predictions and some misclassifications spread across other categories.

The only exception where MobileNetV3 showed a slight advantage was in the Cataract class (True Label 2), as seen in Figure 8. MobileNetV3 correctly categorized 297 images from 311, while EfficientNet-B0 (Figure 9) categorized 284 correctly. However, cross-referencing with Table 4, we see that MobileNetV3 achieved this higher recall at the cost of lower precision (82% vs. 89%), meaning it was more prone to falsely labelling other diseases as cataracts. This trade-off is visible in Figure 8, where MobileNetV3 shows more off-diagonal elements in the cataract row compared to the tighter distribution in Figure 9.

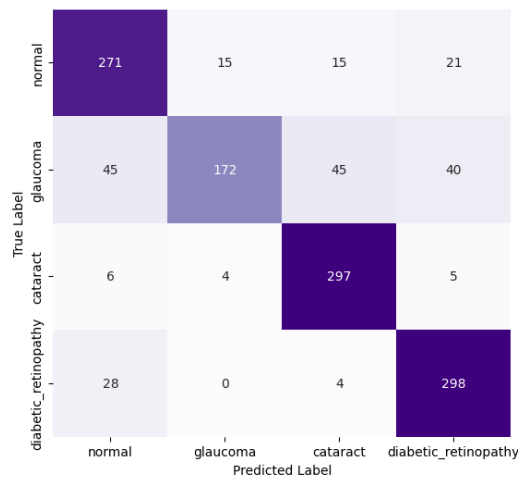


Figure 8. Mobilenetv3 Confusion Metric

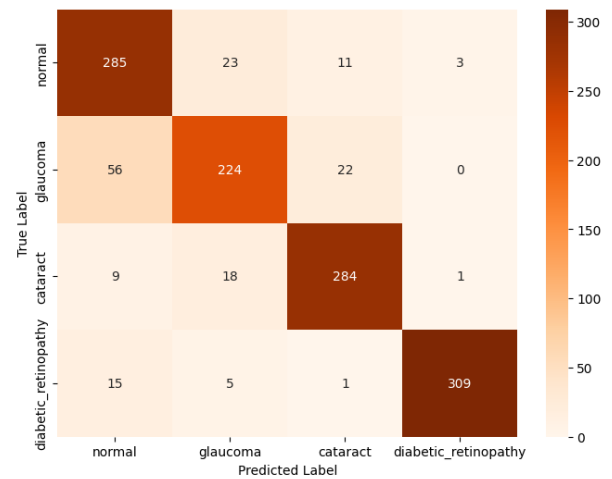


Figure 9. EfficientNet-B0 Confusion Metric

The patterns observed in the confusion matrices were further investigated through systematic visual analysis (see Section 3.6 and Table 5), which revealed that MobileNetV3's poor glaucoma performance stemmed primarily from its inability to detect subtle optic disc changes.

3.4. Inference Speed and Computational Efficiency

Despite its lower accuracy, MobileNetV3 demonstrated a decisive advantage in operational efficiency. The average inference time for MobileNetV3 was 0.0085 seconds per image, approximately four times faster than EfficientNet-B0, which required 0.032 seconds. This stark difference highlights the fundamental trade-off between model accuracy and computational latency, a key consideration for real-time deployment on mobile devices with limited processing capabilities.

3.5. Comparison with State-of-the-Art

To position our work within the current research landscape, we provide a qualitative comparison of our proposed models against recent state-of-the-art methods in multiclass fundus image classification. A contextual summary is presented in Table 4. It is crucial to note that this comparison is indicative rather than a direct 'apples-to-apples' comparison, as the cited studies use different datasets (varying in size, source, and inherent difficulty) and often address different clinical tasks, such as multi-disease screening versus single-disease severity grading. Therefore, performance differences should be interpreted in the context of these varying experimental conditions and dataset challenges, not solely as a reflection of model architecture superiority. This comparison highlights a critical distinction in the field: the difference between classifying distinct pathological diseases (e.g.,

Glaucoma, Cataract) versus grading the severity stages of a single disease (e.g., Diabetic Retinopathy).

Our study addresses the clinically crucial task of screening for multiple distinct fundus diseases. On this task, our EfficientNet-B0 model achieved a strong overall accuracy of 87.0%. This performance is highly competitive, especially when compared to the complex CNN-Trans model by Liu et al., which achieved an accuracy of 80.68% on a more challenging 7-class problem [25]. While the CNN-Trans model employs a sophisticated parallel dual-branch architecture (CNN-LSTM + ViT) for feature fusion, our work demonstrates that a single, well-optimized, and efficient architecture like EfficientNet-B0 can achieve superior performance for the target 4-class problem, offering a better balance between accuracy and computational complexity.

In contrast, several other studies, such as those by Ahmad Chowdhury [26] and Gupta et al. [27], focus primarily on multiclass grading of Diabetic Retinopathy (DR) severity. While these studies report high performance on certain severity classes (e.g., Gupta et al. achieve a high F1-score for 'No DR' and 'Non-severe DR'), they often struggle with imbalanced performance across classes, particularly in detecting rare but critical severe stages. For instance, Ahmad Chowdhury's model shows high sensitivity for 'No DR' (97%) but much lower sensitivity for 'Severe DR' (19%). This underscores the different challenges and metrics of importance between multi-disease screening and single-disease grading. Our model's strength lies in its consistent and clinically relevant performance across different diseases, notably achieving a sensitivity of 74% for glaucoma—a critical metric for an often asymptomatic disease where missed cases can lead to irreversible vision loss.

Table 4. Contextual Comparison with Related Works on Multiclass Fundus Image Analysis

Study (Year)	Proposed Model	Number of Classes (Task Focus)	Dataset (Size)	Key Performance Metric (Context)	Key Strength / Focus & Contextual Notes
Liu et al.[25] Ahmad Chowdhury [26]	CNN-Trans (CNN-LSTM + ViT)	7 (Multi-Disease)	ODIR-5K (5,684)	Accuracy: 80.68%	Complex architecture for multi-disease. Note: Higher class count (7 vs. 4) and a different dataset.
Gupta et al.[27]	Custom CNN	5 (DR Severity Grading)	APTOS 2019 (3,662)	Macro F1-Score: ~51.8% (Calculated)	Focus on DR severity grading. Note: Task focuses on stages of a single disease, not multiple distinct pathologies. Performance highlights the challenge of severe class imbalance.
Liu et al.[25] Ahmad Chowdhury [26]	Custom CNN	3 (DR Severity Grading)	Combined (19,020)	Macro F1-Score: ~75.6% (Calculated)	Large-scale DR grading. Note: Evaluated on a large, combined dataset for a different clinical task (severity grading).
Gupta et al.[27]	CNN + SMOTE + Gabor + Adam	4 (DR-related classes)	Messidor (1,200)	F1-Score: 77%	Handles class imbalance. Note: Smaller dataset; focus on techniques for imbalance within DR context.
This Study	EfficientNet-B0	4 (Distinct Diseases)	Kaggle (4,217)	Accuracy: 87.0% Sensitivity (Glaucoma): 74%	Lightweight model for multi-disease screening. Context: Optimized for practical detection of multiple distinct pathologies, successfully deployed on mobile.

Study (Year)	Proposed Model	Number of Classes (Task Focus)	Dataset (Size)	Key Performance Metric (Context)	Key Strength / Focus & Contextual Notes
This Study	MobileNetV3	4 (Distinct Diseases)	Kaggle (4,217)	Accuracy: 82.0% Inference Time: 0.0085s/image	Ultra-fast inference for real-time use. Context: Represents the high-speed end of the accuracy-speed trade-off for the same multi-disease task.

This contextual analysis, while acknowledging the fundamental differences in datasets and clinical tasks among the cited works, demonstrates that our work occupies a unique and vital niche. We provide a rigorous, practical evaluation of modern lightweight architectures specifically for the task of multi-disease screening. Rather than pursuing maximum accuracy through complex, computationally heavy models, we explicitly analyze and report the critical accuracy-speed trade-off, culminating in the successful deployment of the EfficientNet-B0 model into a functional mobile application. This end-to-end focus on developing a practical, high-performance, and deployable screening tool distinguishes our contribution from studies that prioritize pure accuracy on more narrow tasks or larger, less constrained datasets.

3.6. Visual Failure Analysis

To complement the quantitative analysis from the confusion matrices, we performed a systematic visual examination of misclassified cases to gain deeper insights into the models' failure modes. This analysis is particularly valuable for understanding the practical limitations of each architecture. Table 5 presents representative examples of misclassifications from both models, highlighting specific patterns where EfficientNet-B0 achieved correct classification (True Positive) while MobileNetV3 produced errors (False Negative).

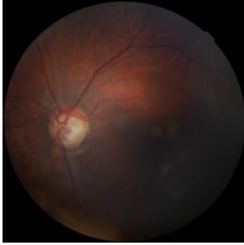
Our visual analysis, supported by the systematic testing summarized in Table 5, revealed several key factors contributing to MobileNetV3's classification failures:

- **Subtle Pathological Features in Glaucoma Cases.** MobileNetV3 consistently failed to detect glaucoma cases with subtle optic disc changes, misclassifying them as 'Normal' (Table 5, Glaucoma examples). This aligns with the quantitative findings in Table 3, where MobileNetV3 showed significantly lower sensitivity for glaucoma (57% vs. 74% for EfficientNet-B0). The model appeared to struggle with capturing fine structural changes in the optic nerve head region, particularly in images where pathological cupping was minimal or the image quality was suboptimal.
- **Complex Diabetic Retinopathy Manifestations.** MobileNetV3 demonstrated confusion between diabetic retinopathy and both 'Normal' and 'Glaucoma' categories (Table 5, Diabetic Retinopathy examples). This suggests the model had difficulty distinguishing subtle microaneurysms and exudates from normal retinal variations or optic disc abnormalities. In one case, MobileNetV3 misclassified a diabetic retinopathy image as glaucoma, possibly due to overlapping features in the optic disc region that confused the less complex architecture.
- **Cataract-Related Visual Obscuration.** MobileNetV3 exhibited misclassifications of cataract images as both 'Glaucoma' and 'Normal' (Table 5, Cataract examples). This appears to result from the model's difficulty in differentiating between the general cloudiness caused by cataracts and other conditions that affect image clarity or optic disc visibility. The model's limited capacity seems to struggle with the complex interplay of multiple visual artifacts.

The visual analysis substantiates the quantitative findings and provides concrete evidence for MobileNetV3's limitations in detecting subtle pathological changes, particularly in glaucoma

cases where early detection is clinically crucial. EfficientNet-B0's superior performance across these challenging cases demonstrates its stronger feature extraction capabilities and better handling of nuanced pathological signatures.

Table 5. Visual Comparative Analysis of Model Performance: Representative Cases of EfficientNet-B0 True Positives versus MobileNetV3 False Negatives

Fundus Image Diseases Test		EfficientNet-B0 Predicted (True Positive)	MobileNetV3 Predicted (False Negative)
Glaucoma		Glaucoma	Normal
		Glaucoma	Normal
Diabetic Retinopathy		Diabetic Retinopathy	Normal
		Diabetic Retinopathy	Glaucoma
Cataract		Cataract	Glaucoma

Fundus Image Diseases Test	EfficientNet-B0 Predicted (True Positive)	MobileNetV3 Predicted (False Negative)
	Cataract	Normal

3.7. Mobile Application Implementation and Validation

The best-performing model, EfficientNet-B0, was successfully operationalized through a dedicated mobile application developed using the client-server architecture outlined in Section 3.5. This implementation serves as a practical proof-of-concept for deploying automated fundus image analysis in resource-constrained environments.

The application provides an intuitive user interface that guides the user through a seamless screening workflow. As shown in Figure 10, the interface includes:

- a. Dashboard Page. The main screen serves as the entry point, offering users clear options to either capture a new fundus image using the device's built-in camera or select an existing image from the local gallery for analysis.
- b. Prediction Result Screen. Upon image selection, the application displays the uploaded fundus image and presents the model's primary prediction. This screen clearly indicates the most likely pathological condition (e.g., "Glaucoma") and displays the associated confidence level, providing an immediate and easily interpretable result for the user.
- c. Detailed Prediction Results for 4 Categories. For a more comprehensive analysis, users can access a detailed results page. This view provides a complete breakdown of the model's predictions for all four diagnostic categories Normal, Glaucoma, Cataract, and Diabetic Retinopathy—displaying the confidence probability for each condition. This transparency allows healthcare personnel to assess the model's certainty across all potential pathologies.

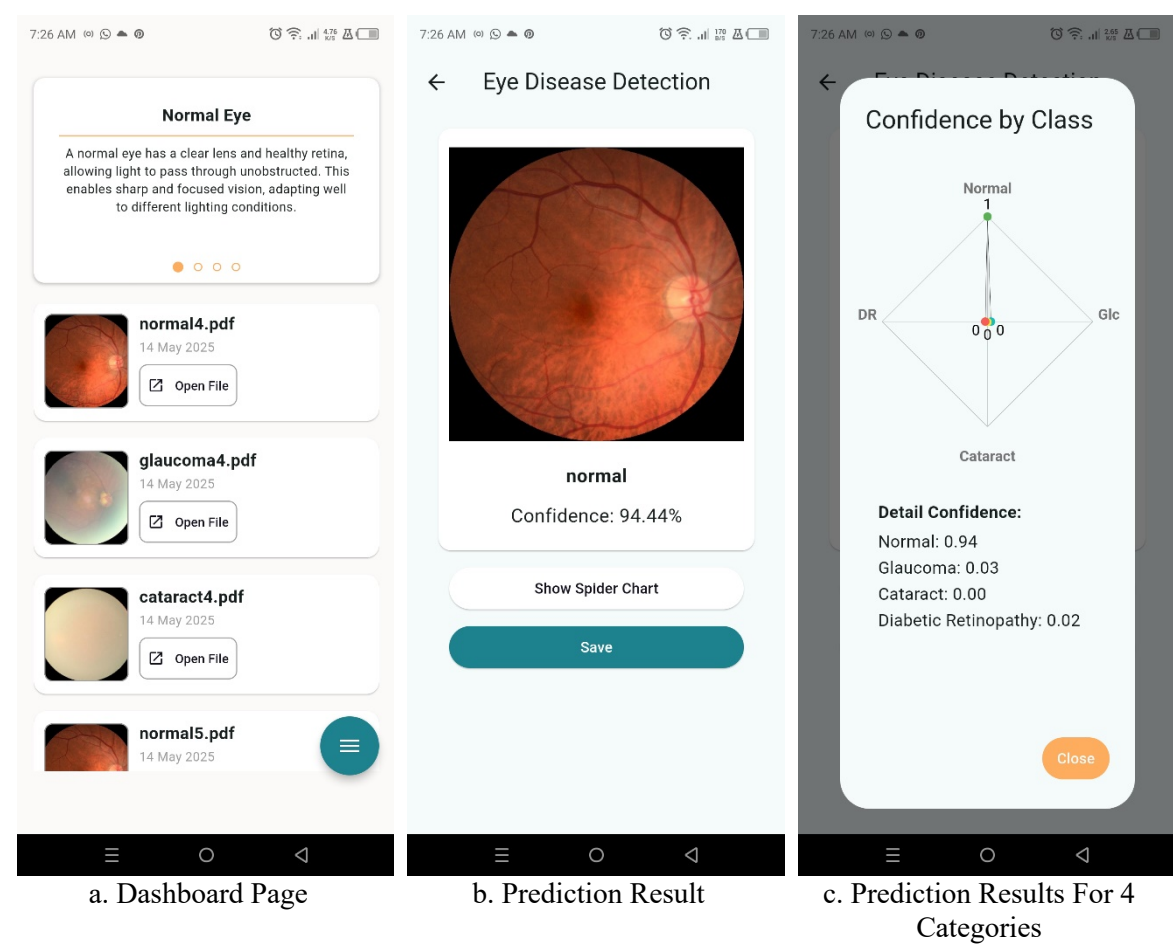


Figure 10. Fundus Images Application: a. Dashboard Page, b. Prediction Result, and c. Prediction Results for 4 Categories

To validate the end-to-end functionality and integrity of the deployed system, a series of tests was conducted using sample images from the validation set. The predictions returned by the mobile application through its API calls were consistently identical to the results obtained during the offline model evaluation phase. This confirms the faithful integration of the trained EfficientNet-B0 model into the production environment and the stability of the client-server communication.

The successful deployment of the EfficientNet-B0 model (34 MB .h5 file) on Google Cloud Platform, with acceptable network latency (285 ± 65 ms), demonstrates the practical viability of using this deep learning-based system as an assistive screening tool in real-world clinical settings with stable internet connectivity. The comparative analysis in Table 6 shows that while MobileNetV3 offered advantages in speed (0.0085s vs. 0.032s) and model footprint (25 MB vs. 34 MB), the significant accuracy advantage of EfficientNet-B0 (87% vs. 82%) justified its selection for the production deployment, as diagnostic reliability is paramount in medical applications.

Table 6. Comparative Model Specifications for Deployment

Metric	MobileNetV3	EfficientNet-B0
Accuracy	82%	87%
Total Parameters	~2.5 million	~5.3 million
Convolutional Layers	68 layers	237 layers
Model Size (.h5)	25 MB	34 MB

Comprehensive Performance Comparison of EfficientNet-B0 and MobileNetV3 for Multiclass Fundus Image Classification, Suprianto

Inference Time (s)	0.0085	0.032
--------------------	--------	-------

3.7. Limitations and Future Work

Despite the promising results, this study has several limitations that should be addressed in future work. First, the dataset, while substantial, is of moderate size compared to some national-scale studies and was sourced from public repositories. This may limit the model's generalizability to populations with different ethnicities or images captured by diverse fundus camera models not represented in our data. Second, while our evaluation utilized a held-out validation set with a strict patient-level split to prevent leakage, the data originated from the same composite source. An external validation on a completely independent, prospectively collected dataset is necessary to truly assess the clinical robustness and generalization capability of the models across diverse populations and imaging devices. Third, while class weights were used, more advanced techniques for handling class imbalance, such as focal loss or sophisticated data augmentation, could be explored to further boost performance on minority classes.

4. CONCLUSION

This study has conducted a rigorous, head-to-head comparative analysis of two leading lightweight CNN architectures, EfficientNet-B0 and MobileNetV3-Large, for the critical task of simultaneous multiclass fundus image classification. Under an identical transfer learning protocol, our empirical results delineate a clear trade-off, providing crucial guidance for model selection in practical ophthalmic AI applications.

EfficientNet-B0 emerged as the superior model for diagnostic accuracy. It achieved a higher overall accuracy and, more importantly, demonstrated consistently robust performance across all pathological classes. Its significantly higher sensitivity in detecting glaucoma (74% vs. 57%) is a critical advantage for a disease where early, asymptomatic detection is paramount to preventing irreversible vision loss. Furthermore, its near-perfect precision and specificity for diabetic retinopathy minimize false positives, thereby reducing unnecessary patient referrals and clinical burden.

Conversely, MobileNetV3-Large established an undeniable advantage in operational efficiency, with an inference speed approximately four times faster than EfficientNet-B0. This makes it a compelling candidate for scenarios where rapid, high-throughput screening is the primary constraint, even if it comes at the cost of some diagnostic precision, particularly for conditions with subtle features like glaucoma.

The most significant validation of this work's practical relevance is the successful deployment of the EfficientNet-B0 model into a functional, client-server mobile application. This end-to-end implementation demonstrates the feasibility of integrating a high-performance deep learning model into a point-of-care tool, bridging the gap between experimental evaluation and tangible clinical utility.

In summary, the choice between these architectures is not a matter of which is universally better, but which is more appropriate for a given deployment context. For applications where diagnostic reliability and the early detection of sight-threatening conditions are the highest priority, EfficientNet-B0 is unequivocally recommended. Its balanced and clinically sound performance justifies its computational overhead. However, for environments where extreme speed and lower computational resources are the dominant requirements, MobileNetV3 presents a viable, though less accurate, alternative.

REFERENCES

- [1] H. Hashemian *et al.*, "Application of Artificial Intelligence in Ophthalmology: An Updated Comprehensive Review," *J Ophthalmic Vis Res*, vol. 19, no. 3, pp. 354–367, 2024, doi: 10.18502/jovr.v19i3.15893.

- [2] A. F. Zegeye, Y. Z. Temachu, and C. K. Mekonnen, "Prevalence and factors associated with Diabetes retinopathy among type 2 diabetic patients at Northwest Amhara Comprehensive Specialized Hospitals, Northwest Ethiopia 2021," *BMC Ophthalmol*, vol. 23, no. 1, Jan. 2023, doi: 10.1186/s12886-022-02746-8.
- [3] J. M. Kang and A. P. Tanna, "Glaucoma," *Medical Clinics of North America*, vol. 105, no. 3, pp. 493–510, May 2021, doi: 10.1016/j.mcna.2021.01.004.
- [4] A. Halim *et al.*, "Prevalence and Associated Factors of Diabetic Retinopathy in People with Type 2 Diabetes Attending Community Based Diabetic Retinopathy Screening in Greater Bandung, Indonesia," *IJRetina*, vol. 5, no. 1, p. 1, Feb. 2022, doi: 10.35479/ijretina.2022.vol005.iss001.172.
- [5] N. Angriani, M. N. R. Akib, F. E. Mulyadi, S. Irmandha, S. Hadi, and R. F. Syamsu, "Glaucoma Risk Factors," *GMJ*, vol. 4, no. 1, pp. 1–8, Apr. 2022, doi: 10.33096/gmj.v4i1.94.
- [6] J. Sahlsten *et al.*, "Deep Learning Fundus Image Analysis for Diabetic Retinopathy and Macular Edema Grading," *Sci Rep*, vol. 9, no. 1, Jul. 2019, doi: 10.1038/s41598-019-47181-w.
- [7] A. Shao *et al.*, "Generative artificial intelligence for fundus fluorescein angiography interpretation and human expert evaluation," *npj Digit. Med.*, vol. 8, no. 1, p. 396, Jul. 2025, doi: 10.1038/s41746-025-01759-z.
- [8] R. Gargeya and T. Leng, "Automated Identification of Diabetic Retinopathy Using Deep Learning," *Ophthalmology*, vol. 124, no. 7, pp. 962–969, Jul. 2017, doi: 10.1016/j.optha.2017.02.008.
- [9] S. H. Abdullah, R. Magdalena, and R. Y. N. Fu'adah, "KLASIFIKASI DIABETIC RETINOPATHY BERBASIS PENGOLAHAN CITRA FUNDUS DAN DEEP LEARNING," *JESCE*, vol. 5, no. 2, pp. 84–90, Feb. 2022, doi: 10.31289/jesce.v5i2.5659.
- [10] K. Pin, J. Ho Chang, and Y. Nam, "Comparative Study of Transfer Learning Models for Retinal Disease Diagnosis from Fundus Images," *Computers, Materials & Continua*, vol. 70, no. 3, pp. 5821–5834, 2022, doi: 10.32604/cmc.2022.021943.
- [11] M. Sarosa, A. Badriyah, R. A. Asmara, M. K. Wardani, D. F. A. Riza, and Y. M. Azis, "Performance Analysis of MobileNET on Grape Leaf Disease Detection," in *2024 International Conference on Advanced Information Scientific Development (ICAISD)*, Jawa Barat, Indonesia: IEEE, Nov. 2024, pp. 64–68. doi: 10.1109/ICAISD63055.2024.10894920.
- [12] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," Sep. 11, 2020, *arXiv*: arXiv:1905.11946. doi: 10.48550/arXiv.1905.11946.
- [13] A. Howard *et al.*, "Searching for MobileNetV3," Nov. 20, 2019, *arXiv*: arXiv:1905.02244. doi: 10.48550/arXiv.1905.02244.
- [14] A. Géron, "Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow," 2019.
- [15] Md. E. Rayed, S. M. S. Islam, S. I. Niha, J. R. Jim, M. M. Kabir, and M. F. Mridha, "Deep learning for medical image segmentation: State-of-the-art advancements and challenges," *Informatics in Medicine Unlocked*, vol. 47, p. 101504, 2024, doi: 10.1016/j.imu.2024.101504.
- [16] I. Sutskever, J. Martens, and G. Dahl, "On the importance of initialization and momentum in deep learning," 2013.
- [17] L. N. Smith, "Cyclical Learning Rates for Training Neural Networks," 2015, *arXiv*. doi: 10.48550/ARXIV.1506.01186.
- [18] D. Masters and C. Luschi, "Revisiting Small Batch Training for Deep Neural Networks," 2018, *arXiv*. doi: 10.48550/ARXIV.1804.07612.
- [19] S. Al-Fahdawi *et al.*, "Fundus-DeepNet: Multi-label deep learning classification system for enhanced detection of multiple ocular diseases through data fusion of fundus images," *Information Fusion*, vol. 102, p. 102059, Feb. 2024, doi: 10.1016/j.inffus.2023.102059.
- [20] J. C. Obi, "A comparative study of several classification metrics and their performances on data," *World J. Adv. Eng. Technol. Sci.*, vol. 8, no. 1, pp. 308–314, Feb. 2023, doi: 10.30574/wjaets.2023.8.1.0054.

- [21] M. V. Mathiazhagan and A. K. Chellappan, "An efficient optimized deep learning model for diabetic retinopathy classification," *JBRHA*, vol. 39, no. 2, p. 3293, Mar. 2025, doi: 10.54517/jbrha3293.
- [22] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," 2014, *arXiv*. doi: 10.48550/ARXIV.1412.6980.
- [23] L. Liu *et al.*, "On the Variance of the Adaptive Learning Rate and Beyond," 2019, *arXiv*. doi: 10.48550/ARXIV.1908.03265.
- [24] Y. Pattanaik *et al.*, "Exploring Optimization Dynamics: Hybrid Approaches Combining Adaptive and Traditional Techniques for Deep Learning Models," in *2025 4th International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, Ballari, India: IEEE, Apr. 2025, pp. 1–6. doi: 10.1109/ICDCECE65353.2025.11035942.
- [25] S. Liu, W. Wang, L. Deng, and H. Xu, "Cnn-trans model: A parallel dual-branch network for fundus image classification," *Biomedical Signal Processing and Control*, vol. 96, p. 106621, Oct. 2024, doi: 10.1016/j.bspc.2024.106621.
- [26] A. Chowdhury and S. Mahfuz, "Diabetic Retinopathy Detection from Fundus Images Using Deep Convolutional Neural Networks," in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Lisbon, Portugal: IEEE, Dec. 2024, pp. 6362–6369. doi: 10.1109/BIBM62325.2024.10822813.
- [27] S. Gupta and A. Khadka, "Diabetic Retinopathy Detection through Multiclass Classification of Fundus Image Using Convolutional Neural Network," *Pokhara Eng. Coll. J.*, vol. 2, no. 1, pp. 47–56, Mar. 2025, doi: 10.3126/pecj.v2i1.76831.