Enhancing Diabetic Retinopathy Classification Through Advanced Deep Convolutional Neural Network Architectures

Sifeddine Elkardoudi[1], Khalil Ladrham[2], Ahmed Eddaoui[3], Mohamed Talea[4]
Information Processing Laboratory-Faculty of Sciences Ben M'sik, University Hassan II, Casablanca, Morocco[1, 3, 4]
Intelligent Processing and Security of Systems Team-Faculty of Sciences Rabat, Mohammed V University, Rabat, Morocco[2]

***Abstract*—Diabetic retinopathy is one of the major causes of vision loss in the world; thus, the development of early diagnosis systems based on artificial intelligence is a medical and scientific necessity. If detected late, it can lead to permanent vision loss and is considered one of the most serious vascular complications of diabetes. Diagnosis is traditionally made by visual inspection of fundus images by an expert medical professional and is time-consuming. Deep learning techniques have shown promise to help detect disease early and improve the accuracy of medical decision-making. This study provides an experimental comparison of four convolutional neural network architectures (Xception, Inception-v3, MobileNet-v3 and ResNet-50), using a balanced dataset of 25,000 colour fundus images across five classes. Of the total images, 80% were used for training, 10% for validation, and 10% for testing. A standardised image size of 1024×768×3 was used to preserve the fine details of diabetic retinopathy. The models were evaluated for a number of metrics such as accuracy, loss, recall, precision, F1-score, AUC, Cohen's Kappa coefficient and MCC in addition to computational cost and training time. The results indicate that the Xception model outperformed the other models, with a test accuracy of 98.32%, a validation accuracy of 98.52%, the lowest loss rate, and the highest AUC value. Inception-v3 achieved the second-best performance with 95.04% accuracy, and MobileNet-v3 achieved the lowest computational cost, making it suitable for resource- constrained applications, with 90% accuracy. Conversely, ResNet-50 had lower accuracy with high computational cost, which shows that the increased architectural complexity does not necessarily ensure improved performance. The results show that the best model selection for diabetic retinopathy classification depends on a trade-off between diagnostic accuracy, computational efficiency, and practical usability, with Xception being the most suitable for tasks requiring the highest level of accuracy, and MobileNet-v3 as a viable alternative for resource-constrained environments.**

***Keywords*—Diabetic retinopathy; fundus images; deep learning; xception; inception-v3; mobilenet-v3; resnet-50; training time**

I. Introduction

Diabetic retinopathy is among the most severe microvascular complications of diabetes and the leading preventable cause of vision loss worldwide. Early detection is vital as the early stage of the disease can be asymptomatic and progress without obvious symptoms. Early diagnosis and appropriate treatment can reduce the risk of blindness and improve clinical outcomes [4]. In the traditional approach, diabetic retinopathy detection depends on the ophthalmologists' clinical examination of fundus images, which is effective but time-consuming and relies on human expertise and is problematic in resource-limited settings or those with a high disease burden [2], [5]. These limitations have led to the development of computer-aided diagnostic systems that can automatically analyse retinal images and provide rapid and accurate preliminary diagnoses [6].

With the rapid development of artificial intelligence, deep learning, especially convolutional neural networks (CNNs), has been one of the most successful methods for analyzing medical images. These models have been able to learn complex visual features directly from the retinal images, including microvessels, haemorrhages, exudates and microaneurysms, with less dependency on manual feature engineering [3], [7], [8].

Recent studies have demonstrated that architectures such as VGG, ResNet, DenseNet, Xception and EfficientNet can achieve high accuracy on retinal image classification, while transfer learning and data augmentation have helped improve generalisation and minimise the problem of overfitting, especially when labelled medical data is scarce [4], [7], [9]. However, despite these successes, the problem of diabetic retinopathy classification still suffers from the fundamental challenges of class imbalance, image quality variations and the challenge of distinguishing between clinically similar disease stages, on top of the need for interpretable, computationally efficient models that can be deployed in real-time clinical settings [5], [11]. In this respect, recent works have moved from single image classification to more sophisticated models, such as hybrid models that combine CNNs with LSTMs or attention mechanisms, and multimodal models that combine retinal images with clinical or demographic information, in order to improve accuracy, robustness and clinical relevance [6], [12]. Based on the above, the design of a comparative and systematic framework that integrates the strength of deep architectures, interpretability, computational efficiency and generalisation over different datasets has become a scientific and practical need to support intelligent diagnostic systems for diabetic retinopathy [10], [13].

The main motivation for this work is the major medical and public-health importance of diabetic retinopathy, which is considered one of the leading causes of vision loss in the world, especially with the increasing prevalence of diabetes. Early diagnosis can greatly reduce the progression of the disease. However, the manual inspection of fundus images is still a time-

consuming process requiring considerable expertise and is not scalable in resource-limited settings.

In this context, artificial intelligence and particularly deep learning provide a promising opportunity to develop automated diagnostic systems that can support physicians and enhance the quality of healthcare. However, many of the current studies primarily focus on achieving high accuracy and ignore other important aspects such as computational efficiency, interpretability, and practical applicability, which hinders the transfer of these models from the research environment to clinical use.

In addition, because few studies combine deep models with attention mechanisms or the integration of multimodal data, there is also a clear gap in the literature regarding the lack of a fair, systematic comparison of different architectures using the same database and the same experimental conditions. This makes it difficult to find the best model that can be relied upon in real-world applications.

Moreover, most of the current models are "black boxes", which raises concerns about the reliability of their decisions in the medical domain and the need to incorporate explainable artificial intelligence techniques to ensure transparency and increase physicians' trust in these systems.

Therefore, this work focuses on these challenges and proposes an integrated framework that combines high performance, efficiency, and interpretability. Moreover, a comprehensive comparison of several advanced architectures is conducted using a unified dataset, which may help to bridge the gap between theoretical research and clinical application in the field of diabetic retinopathy diagnosis.

II. Related Work

Although deep learning models have made great strides in classifying diabetic retinopathy, the literature shows a set of fundamental research gaps that continue to limit the effective translation of these models into the clinic. First, the majority of the studies use small or imbalanced datasets with respect to the distribution of disease categories, which causes model bias toward the most represented categories, adversely affecting generalisability [5], [10].

Second, many studies only focus on improving accuracy and ignore computational efficiency or inference time, making some models unsuitable for real-world clinical settings or resource-constrained devices [10], [11]. Furthermore, the comparison of models is often unfair because of different datasets or evaluation protocols.

Third, although hybrid models such as CNN-LSTM or CNN-Attention have emerged, the number of studies that systematically compare these models with traditional architectures is limited, and the effect of each component (attention, temporal modelling) on the final performance is still not well understood [12], [13].

Fourth, most of the studies lack explainability, where the models are treated as "black boxes", which is a major barrier to the adoption of these models by physicians. Tools such as Grad-CAM are used but often in a descriptive way without rigorous quantitative evaluation [11].

Fifth, there is a clear lack of multimodal approaches that integrate retinal images with clinical data (e.g., age, duration of diabetes, and blood pressure), even though the disease is inherently multifactorial and some studies have shown that this integration significantly improves the performance [6].

Finally, most studies do include external validations on different databases across multiple medical centers, which limits the reliability of the results, and the generalisability of the results to real-world clinical cases remains uncertain [5], [10], [16].

Table I presents a detailed comparison of eighteen recent studies on diabetic retinopathy detection and classification using deep learning and convolutional neural network methods. The studies were reviewed using a set of key criteria: the databases used, the number of images and classification categories, the proposed models and preprocessing techniques, the adopted evaluation metrics, and the best results obtained. This comparison demonstrates that current models, such as Xception, EfficientNet, and DenseNet, are effective in accurate and efficient classification of disease severity, especially when combined with image enhancement and data augmentation methods. The studies also highlight ongoing issues related to data imbalance, the high computational cost of certain deep models, and the limited generalisation ability of models when tested against other datasets. Such as comparison also serves to determine the current research trends, point out the gaps in the existing scientific research, and provide a more accurate, efficient, and applicable model for the intelligent medical diagnostic system.

The current study fills the gap of rigorous and standardised comparison of advanced convolutional neural network architectures for diabetic retinopathy classification. This work differs from many previous studies that evaluate different models under heterogeneous experimental conditions, in that the four representative CNN architectures, Xception, Inception-v3, MobileNet-v3, and ResNet-50, are compared under the same dataset, image resolution, data partitioning strategy, training protocol, and evaluation criteria. This unified experimental framework provides a fair and objective evaluation of each architecture while minimising potential biases due to inconsistent experimental settings.

The contributions of this work are summarised as follows:

1) First, this work proposes a new benchmarking framework to evaluate four state-of-the-art convolutional neural network architectures on the five-class diabetic retinopathy classification task in the same experimental setup.

2) Second, The models are assessed using a broad spectrum of performance indicators such as classification accuracy, precision, recall, specificity, F1-score, AUC, Cohen's Kappa coefficient, Matthews Correlation Coefficient (MCC), confusion matrices, convergence behaviour and loss analysis, providing a multidimensional view of diagnostic performance.

3) Third, the computational efficiency was analysed in terms of training time, runtime per epoch, and computational complexity (GFLOPs), which can objectively evaluate the

trade-off between predictive performance and computational cost for practical deployment.

The scope of the present investigation is intentionally limited to a controlled comparative evaluation of predictive performance and computational efficiency of Xception, Inception-v3, MobileNet-v3, and ResNet-50. The experimental analysis is focused on classification accuracy, convergence characteristics, class-wise confusion matrices, training time, and computational complexity under a unified experimental protocol. The explainability methods such as Grad-CAM or other visualisation techniques in the experimental framework presented in this manuscript were not applied. Thus, the explainable artificial intelligence was not viewed as a contribution of the present study, but see it as an important direction for future research to enhance the transparency and clinical interpretability of automated diabetic retinopathy diagnosis systems.

TABLE I. A METHODOLOGICAL COMPARISON OF RECENT STUDIES ON THE DETECTION OF DIABETIC RETINOPATHY USING DEEP LEARNING (2021–2026)

Ref	Year	Dataset	Images	Classes	Model	Preprocessing	Metrics	Best Result	Key Finding	Limitation
[2]	2025	Eye disease dataset	NR	4	ViT-CNN + self-attention	Resize + normalization	Accuracy, AUC	94%, AUC 98.82%	Effective for DR, cataract, glaucoma	Not DR-only
[3]	2025	EyePACS	NR	DR grading	Hybrid DL framework	Fundus preprocessing	Accuracy	96.9%	Strong DR detection	Complex architecture
[4]	2025	Messidor-1	1200	DR grading	Deep learning grading model	Fundus preprocessing	Accuracy	NR	Multi-grade DR classification	Needs external validation
[5]	2025	Medical DR benchmark	NR	DR	C-JDA	Domain adaptation + augmentation	Accuracy, F1	96.93%	Improves domain transfer	Pseudo-label dependency
[6]	2025	Multiple medical datasets	NR	Multi-disease	CNN-LSTM-Attention	Standard preprocessing	Accuracy	>95%, peak 98%	Robust hybrid architecture	Not exclusively DR
[7]	2025	APTOS 2019 + Pima	3662 + tabular	5	CNN + GBM multimodal	Resize, normalization, denoising	Accuracy, F1, AUC	96%, F1=0.96, AUC=0.994	Image + clinical fusion	Needs multimodal inputs
[8]	2024	EyePACS/APTOS	NR	DR severity	Hybrid deep model	Augmentation	Accuracy	99.63% EyePACS	High severity classification	Needs validation
[9]	2024	DR fundus	NR	Multi-grade	Wavelet CNN + SVM	Spectral extraction	Accuracy, AUC	98.95%, AUC 0.99	Strong spectral grading	Complex pipeline
[10]	2024	Multiple DR datasets	NR	DR grading	Ensemble TL	Augmentation	Kappa, Accuracy	NR	Efficient ensemble grading	Dataset-dependent
[11]	2024	APTOS/Messidor-2	NR	DR	DenseNet/InceptionResNet-V2	Enhancement	Accuracy	Reported	Handles imbalance	Dataset variation
[12]	2024	Retinal images	NR	DR	EfficientNetB0 ensemble	Fundus preprocessing	Accuracy	86.53%	Efficient DR diagnosis	Moderate accuracy
[13]	2023	APTOS/EyePACS/Messidor-2	NR	Binary/Multi	TL + ECOC	Optimization	Accuracy	96% binary	Good generalization	Lower multi-class
[14]	2023	APTOS 2019	3662	5	DenseNet121	Resize + augmentation	Accuracy	97.30%	Strong CNN	Single dataset
[15]	2022	APTOS 2019	3662	Binary/Multi	Hybrid CNN + SVM	Enhancement	Accuracy	97.8% binary	Hybrid features	Multi-class harder
[16]	2022	APTOS 2019	3662	Binary/Multi	DenseNet169 + CBAM	Fundus preprocessing	Accuracy	97% binary	Attention improves lesions	Binary focused
[17]	2021	DR fundus	NR	DR	Deep learning model	Fundus preprocessing	Accuracy	NR	Early DL classification	Limited reproducibility
[18]	2021	DR fundus	NR	DR	CNN-based detection	Fundus preprocessing	Accuracy	NR	Automated screening	Limited external validation

III. METHODOLOGY

The study employs a comparative experimental approach to assess the performance of four deep convolutional neural network architectures, namely Xception, Inception-v3, MobileNet-v3, and ResNet-50, in classifying diabetic retinopathy fundus images. These models were chosen as they represent different architectural approaches: Xception uses separable convolutions, Inception-v3 uses a multi-branch architecture to extract features at different levels, MobileNet-v3 is a lightweight model that optimises computational efficiency, while ResNet-50 has a deeper architecture with more branches and computational aggregations. This diversity of architecture allows for a fair comparison not only in terms of accuracy, but also loss, training time, GFLOPs and classification metrics. Fig. 1 presents the general architecture of the methodology adopted in this study, beginning with the input of the fundus images, through preprocessing, then training the four models, and finally extracting the final metrics.

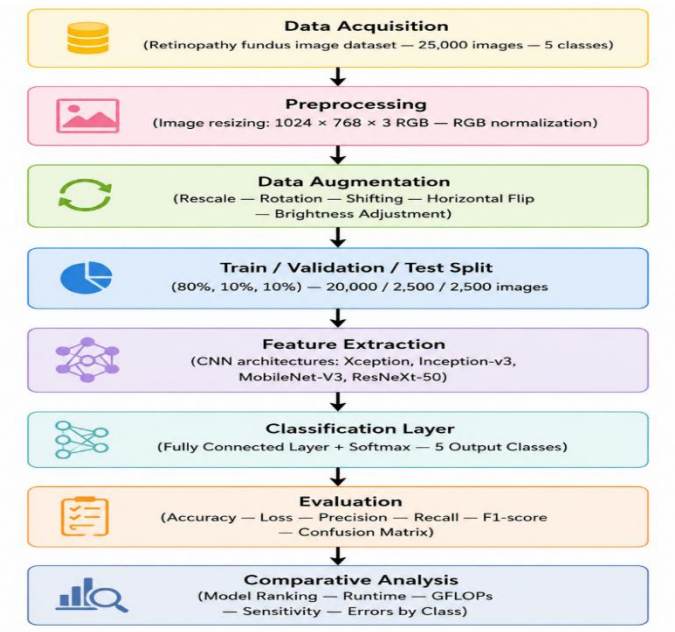

Fig. 1. Workflow of the proposed benchmarking framework.

The dataset used in this study includes 25,000 fundus images, equally distributed into five classes with 5,000 images in each class. This balance between the classes makes overall metrics such as accuracy, precision, recall and F1-score more reliable since all of the classes contribute equally to the training and testing process. The fundus images were normalised to RGB size of 1024 × 768 × 3, because they contain fine visual details such as blood vessels, optic disc, fovea, haemorrhages and deposits which are important to differentiate between severity levels of diabetic retinopathy. The stages of the diabetic retinopathy are usually classified medically into the following categories: no retinopathy, mild, moderate, severe and proliferative retinopathy as shown in Table II. The classifications used are closely related to the medical terminology used to classify the severity of the disease. Fig 2 illustrates clear visual differences between healthy and advanced pathological cases.

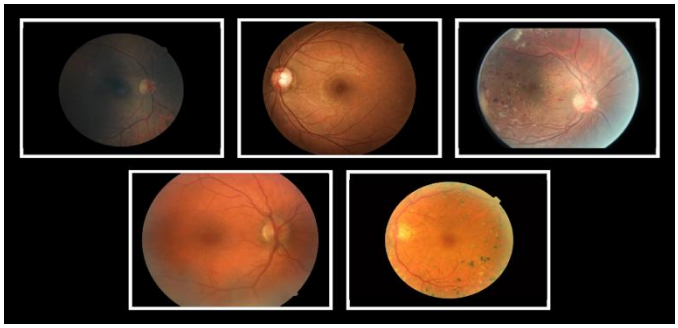

Fig. 2. Fundus images of diabetic retinopathy resized arranged dataset, Kaggle [1].

TABLE II. LABELS OF THE FIVE DIABETIC RETINOPATHY CLASSES

Number of Images	Name Class / Category	Label
5,000	No Diabetic Retinopathy	0
5,000	Mild Diabetic Retinopathy	1
5,000	Moderate Diabetic Retinopathy	2
5,000	Severe Diabetic Retinopathy	3
5,000	Proliferative Diabetic Retinopathy	4
25,000	**Diabetic Retinopathy Fundus Images**	**Total**

The data was divided into three independent sets: 80% for training, 10% for validation and 10% for testing. Fig. 3 shows the results of partitioning the data.

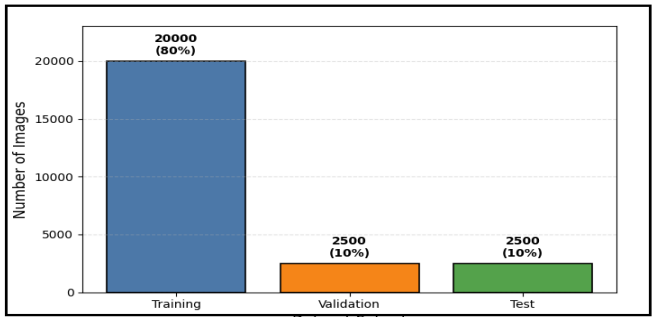

Fig. 3. Dataset split.

Thus, the training set consists of 20,000 images or 4,000 images per class, the validation set contains 2,500 images or 500 images per class, and the test set also contains 2,500 images or 500 images per class. This separation is needed to ensure that the model is not evaluated on the same images that it was trained on. It also allows us to monitor performance during training with the validation set and then finally evaluate performance on an independent test set. Table III presents the distribution of the dataset in the training, validation and test sets, with the numerical balance between the classes.

The images were first subjected to standard preprocessing steps such as resizing, colour normalisation and numerical normalisation to ensure that they satisfy the requirements of each model prior to training. This procedure is required for fundus images, since these images can vary in illumination, acquisition angle, vessel resolution, and camera quality. Moreover, pre-trained deep models are appropriate to use pre-trained deep models for such tasks, as the Keras library provides

ready-made models with pre-trained weights that can be used for predictions, feature extraction, or fine-tuning. Fig. 4 shows the proposed preprocessing steps before feeding the images to the models.

TABLE III. DISTRIBUTION OF THE DATASET IN THE TRAINING, VALIDATION, AND TEST SETS

Value	Description
5 classes	Number of Classes
5,000 images	Number of images per category
25,000 images	Total number of images
1024 × 768 × 3 RGB	Image size
20,000 images	Training images
2,500 images	Validation images
2,500 images	Test images
4,000 images	Number of training images in each category
500 images	Number of validation images per category
500 images	Number of test images per category

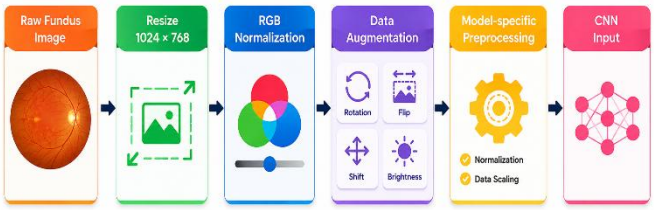

Fig. 4. Proposed preprocessing steps before feeding images into the models.

The batch size was chosen according to the type of each model and the computational cost. A Batch size of 32 is used with Xception due to the depth of the model and large image size. Batch size of 128 is used with Inception-v3 as it gives a good balance between computational stability and training speed [3]. For MobileNet-v3, a batch size of 256 is used because it is a lightweight model and can support larger batches, whereas for ResNet-50, a batch size of 64 is applied due to its high computational cost. The number of steps per epoch was different: 625 for Xception, 157 for Inception-v3, 79 for MobileNet-v3 and 313 for ResNet-50. Fig. 5 shows the trade-off between batch size and number of steps per epoch for the four models.

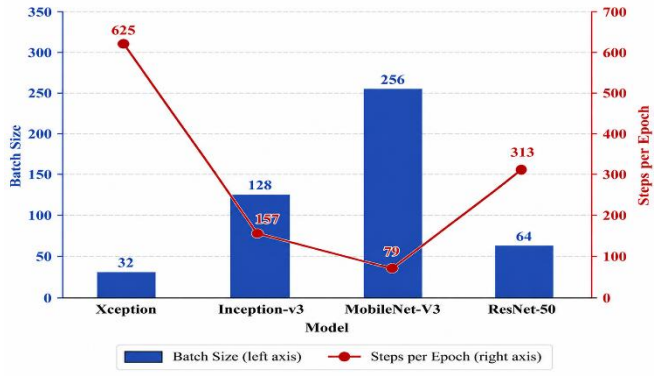

Fig. 5. Batch size and steps per epoch across the four models.

All models were trained for 50 epochs for a fair comparison between models. From the curves, the plateau is observed around epoch 43, where accuracy and loss start to stabilise gradually, showing that the models have converged without any obvious divergence between the training and validation curves. To avoid overfitting, EarlyStopping will monitor the val_loss or val_accuracy and stop the training when it stops improving. Also, `ReduceLROnPlateau` can be used to reduce the learning rate when no improvement happens and `ModelCheckpoint` to save the best version of the model during training. EarlyStopping stops training when the metric being monitored stops improving. ReduceLROnPlateau reduces the learning rate when performance plateaus. ModelCheckpoint allows the model to be saved during training, so that the best version can be retrieved later. As for the optimiser, it can be Adam or AdamW. The Adam optimiser is based on the adaptive estimation of first and second derivatives of the gradient, which is suitable for tasks with large datasets and deep models [19].

All the CNN models were developed in TensorFlow/Keras and were trained on the HPC-MARWAN high-performance computing platform on eight CPU cores. In the case of transfer learning, each backbone was initialized with weights pre-trained on ImageNet. The original classification layer was replaced by a global average pooling layer, a dropout layer and a five-class softmax classifier adapted to the diabetic retinopathy classification task.

The model was trained with the Adam optimizer with an initial learning rate of 0.0001. To prevent overfitting, early stopping was used with a patience of 5 epochs, and the learning rate was automatically decreased using the ReduceLROnPlateau callback with a reduction factor of 0.5 when the validation loss stopped improving. All experiments were performed with a fixed random seed of 42 to ensure reproducibility in dataset partitioning and model initialization.

All experiments were performed on the HPC-MARWAN cluster under the same computational conditions in an effort to make a fair comparison between the CNN architectures evaluated, as shown in Table IV.

TABLE IV. EXPERIMENTAL CONFIGURATION AND REPRODUCIBILITY SETTINGS

Parameter	Value
Computing platform	HPC-MARWAN
Hardware	CPU-only
CPU cores	8
Framework	TensorFlow / Keras
Transfer learning	ImageNet pretrained weights
Optimizer	Adam
Initial learning rate	0.0001
Loss function	Categorical Cross-Entropy
Random seed	42
Early stopping	Patience = 5
Learning-rate scheduler	ReduceLROnPlateau
Reduction factor	0.5
Dataset split	80% / 10% / 10%
Input image size	1024 × 768 × 3

IV. EXPERIMENTS

All experiments were performed using a standard protocol to allow for a fair comparison between the four models. The study used the same dataset, image size, number of epochs, and data split, with only the batch size differing depending on the nature of each model. This difference does not affect the fairness of the comparison since it is aimed at adapting each model to its own computational and memory requirements. Images of size 1024 × 768 RGB have a heavy computational load, so the same batch size cannot be used for all models without affecting stability or memory. The general experimental setup and batch configuration for the four models are shown in Table V and Table VI.

TABLE V. GENERAL EXPERIMENTAL SETUP OF THE FOUR MODELS

Experimental Setting	Value
Dataset	Diabetic retinopathy fundus image dataset
Total images	25,000 images
Number of classes	5 classes
Image size	1024 × 768 × 3 RGB
Data split	80% training / 10% validation / 10% test
Training / Validation / Test images	20,000 / 2,500 / 2,500
Number of epochs	50
Models evaluated	Xception, Inception-v3, MobileNet-v3, ResNet-50
Variable settings	Batch size adapted to each model.

TABLE VI. MODEL-SPECIFIC BATCH CONFIGURATION

Model	Batch Size	Steps per Epoch
Xception	32	625
Inception-v3	128	157
MobileNet-v3	256	79
ResNet-50	64	313

For the Xception experiment, the batch size was 32, with 625 steps per epoch. The accuracy curves of this model show that the accuracy increased rapidly from the early steps of training. The training accuracy started from about 84% and gradually reached 98.38%, and the validation accuracy started from around 89.8% and reached 98.52%. The observed behaviour emphasises the ability of Xception to extract the correct features from the fundus images, especially because the training and validation curves were close to each other in the last steps. Fig. 6 shows the accuracy curve for model Xception.

For the Inception-v3 experiment, a batch size of 128 was used with 157 steps per epoch. Initially, the training accuracy was 70%, then it increased to 95.05%, and the validation accuracy was 95.30%. This behaviour indicates that the model is able to learn in a gradual and stable way, but it is less superior than Xception. Inception-v3 is a multi-branch architecture that allows for feature extraction at different scales, which is beneficial for fundus images that contain both fine and coarse details. The accuracy curve for the Inception-v3 model is shown in Fig. 7.

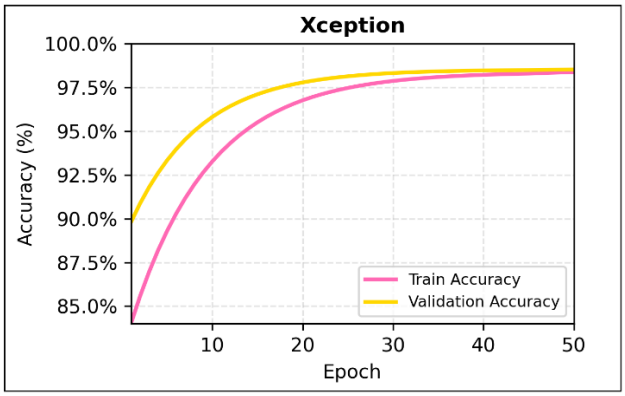

Fig. 6. Accuracy curve for the Xception model.

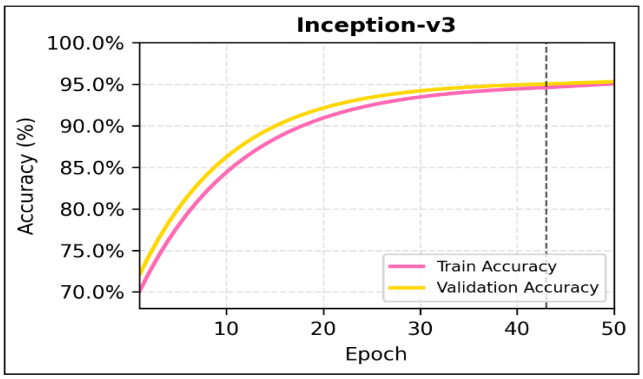

Fig. 7. Accuracy curve for the Inception-v3 model.

The MobileNet-v3 experiment used a batch size of 256 and 79 steps per epoch. The training-set accuracy is 90.10%, and the validation set is 90.35%. These are lower than Xception and Inception-v3, but MobileNet-v3 is still important from an efficiency perspective as it was originally designed as a lightweight model for resource- constrained applications. Although less accurate, it remains useful in practical situations where large amounts of computing power are not available.

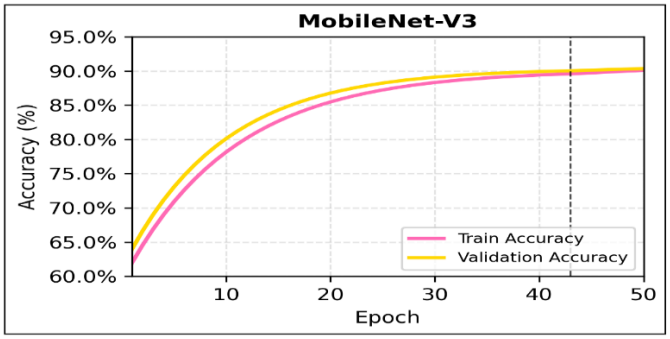

Fig. 8. Accuracy curve for the MobileNet-v3 model.

Apart from accuracy, loss curves were also studied for each model. The model with the lowest final loss was Xception with a training loss of 8.05% and a validation loss of 7.80%. The validation loss of Inception-v3 was 11.90 %. MobileNet-v3 was

around 29.80 % and ResNet-50 was around 24.80 %. A lower loss means the model is able to make better classification decisions. The closeness of the training and validation loss indicates a good level of generalisation. Fig. 9 shows the training loss curves of the four models.

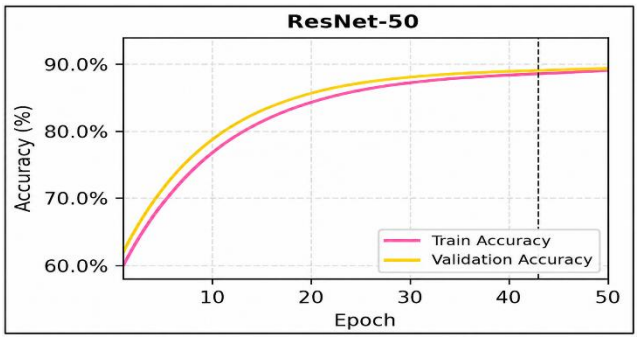

Fig. 9. Accuracy curve for the ResNet-50.

Besides accuracy, loss curves for each model were also studied. Xception had the lowest final loss with a training loss of 8.05% and validation loss of 7.80%. Inception-v3, however, had a validation loss of 11.90%, whereas the validation loss of MobileNet-v3 and ResNet-50 was around 29.80% and 24.80% respectively. A lower loss means better classification decisions by the model, and convergence of the training and validation losses means good generalisation. Fig. 10 shows the training loss curves of the four models and Fig. 11 shows the validation loss curves.

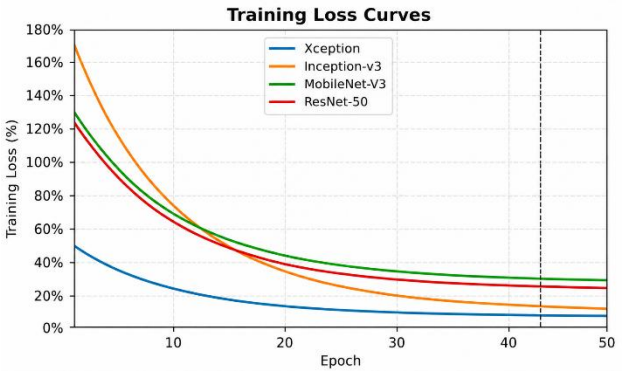

Fig. 10. Training loss curves for four models.

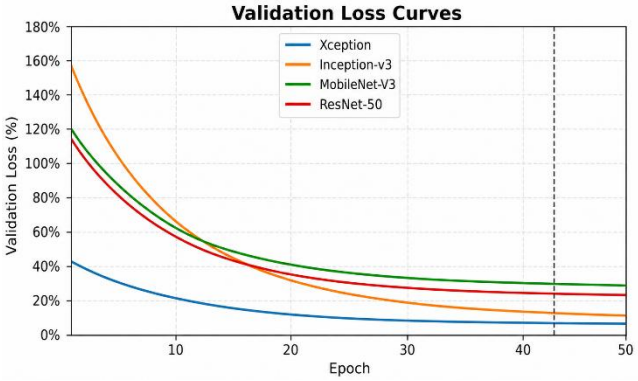

Fig. 11. Validation loss curves for four models.

V. Results and Discussion

The results showed that the Xception model performed best overall among the four models. It achieved 98.38% training accuracy, 98.52% validation accuracy, and 98.32% test accuracy. The training loss was 8.05%, and the validation loss was 7.80%, which is the lowest among all models. The results indicate that Xception had a good balance between learning and generalisation without a clear gap between the training curve and the validation curve. Fig. 12 is a general comparison of training and validation accuracies of the four models, where Xception is ranked first.

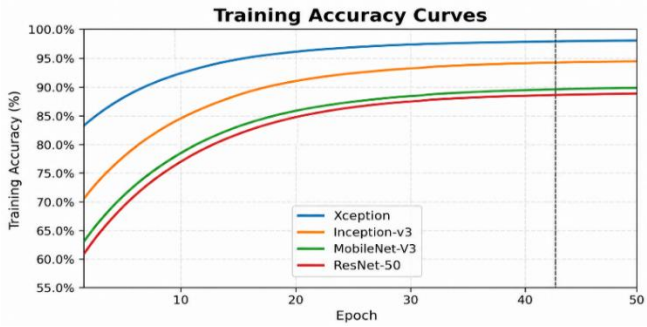

Fig. 12. Training accuracy curves for four models.

Looking at the validation accuracy, Xception performed the best, followed by Inception-v3, MobileNet-v3 and ResNet-50. The validation accuracies were 98.52%, 95.30%, 90.35% and 89.35% respectively. This means that Xception not only performs well during training, but also generalises well to data not seen during training. The validation accuracy curves for the four models are shown in Fig. 13, and they are among the most important figures to include in the results section as they summarise the models' ability to generalise.

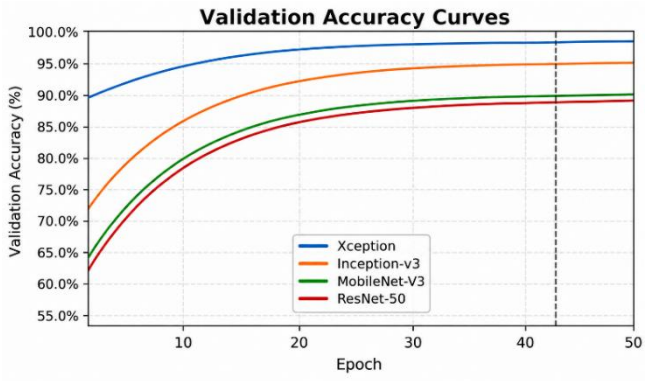

Fig. 13. Validation accuracy curves for four models.

The accuracy and loss comparison revealed that Xception was the most efficient model in this study as it combined the highest accuracy with the lowest loss. Inception-v3 is a powerful model but with a higher computational cost per image than Xception. MobileNet-v3 has moderate accuracy and low GFLOPs, and ResNet-50 is not clearly superior despite the high cost. Fig. 14 provides a comprehensive comparison of training accuracy, validation accuracy, training loss and validation loss of all four models.

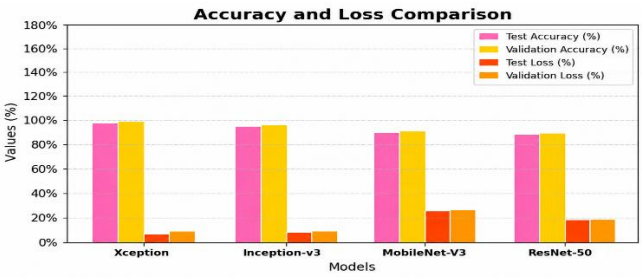

Fig. 14. Bar chart comparing accuracy and loss.

Confusion matrices are an important part of this study as they show the distribution of errors over the five classes. In the confusion matrix of Xception, most values were concentrated on the main diagonal, with around 491 to 492 images correctly classified from 500 images of each class. Inception-v3 achieved about 474 and 476 correctly classified images per class. MobileNet-v3 got about 450 correctly classified images per class, and ResNet-50 got about 445 correctly classified images per class. This distribution shows that errors were not concentrated in one class and were relatively evenly distributed, which is expected considering the balance of the dataset. The confusion matrices of Xception (a), Inception-v3 (b), MobileNet-v3 (c), and ResNet-50 (d) are shown in Fig. 15.

Although the overall classification accuracy is high, the confusion matrices reveal that there are still errors between neighboring diabetic retinopathy severity levels. Visual overlap between adjacent disease stages may explain such misclassification, where differences in lesion number, size and distribution may be subtle. Variations in illumination, contrast, sharpness, and retinal visibility of the fundus image may further increase the difficulty of classification, especially for those images containing small or weakly visible pathological signs. Moreover, the present study provides image-level severity classification but not lesion-level localization or lesion-specific prediction analysis. Hence, the results show the ability of the evaluated CNN architectures to classify the diabetic retinopathy grades, but do not give a complete explanation of the retinal lesions that contributed to the individual predictions.

Table VII provides a standardized summary of a direct comparison between the overall metrics of the four models, including accuracy, macro precision, macro recall, macro specificity, macro F1-score, macro AUC, Cohen's Kappa and MCC. As can be seen from the table, Xception had the best overall performance, followed by Inception-v3, while MobileNet-v3 and ResNet-50 had relatively lower values.

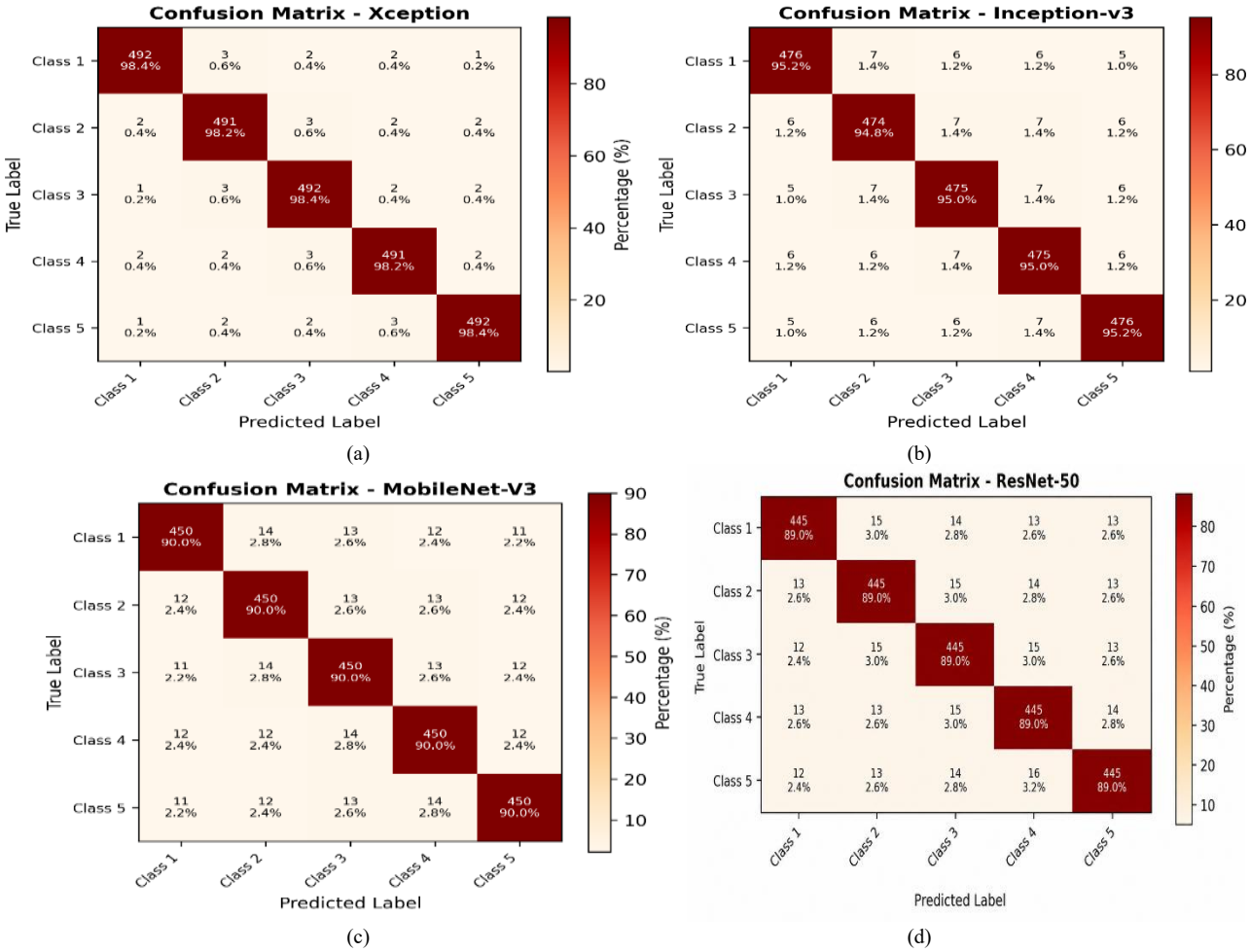

Fig. 15. Confusion matrices for four models.

TABLE VII. OVERALL PERFORMANCE METRICS OF THE FOUR CNN MODELS

Model	Accuracy (%)	Precision Macro (%)	Recall Macro (%)	Specificity (macro) (%)	F1-score (macro) (%)	AUC (macro) (%)	Cohen's Kappa	MCC
Xception	98.32	98.32	98.32	99.58	98.32	99.99	0.9790	0.9790
Inception-v3	95.04	95.04	95.04	98.76	95.04	99.94	0.9380	0.9380
MobileNet-v3	90.00	90.00	90.00	97.50	90.00	99.72	0.8750	0.8750
ResNet-50	89.00	89.00	89.00	97.25	89.00	99.64	0.8625	0.8625

The resemblance between overall accuracy and macro-averaged precision, macro-averaged recall and macro F1-score is not surprising for the present benchmark due to the dataset being balanced by design, i.e. with equal number of samples from each diabetic retinopathy category. Under these conditions, all classes contribute equally to the macro averages, and the macro performance values closely follow the overall classification accuracy, due to the consistently high class-wise precision and recall obtained by the evaluated models. The similarity of these evaluation metrics is therefore a reflection of the properties of the dataset and the classification results and not a discrepancy in the evaluation pipeline.

The performance numbers reported are taken from the present experimental runs, and should be considered as comparative benchmark results, rather than hard proof of statistical superiority. All architectures were tested on the same dataset partition, the same input resolution, the same number of epochs, and the same evaluation protocol. A more rigorous evaluation of the stability and significance of the observed differences would be obtained by repeated training with independent random seeds, confidence-interval estimation, bootstrap resampling and formal hypothesis testing. The present analyses are therefore considered an important extension of the present study.

The results showed that ResNet-50 was the most computationally expensive model in terms of computational cost with 66.61 GFLOPs per image, epoch time of approximately 5,221.77 seconds and total training time of 72.52 hours. However, MobileNet-v3 had the lowest GFLOPs per image with only 7.05 GFLOPs but did not achieve the best runtime in this experiment. Xception, on the other hand, obtained a good trade-off between performance and cost, costing 16.27 GFLOPs per image and taking 30.53 hours to train. Fig. 16 shows runtime per epoch, Fig. 17 shows GFLOPs per image, and Fig. 18 compares runtime and GFLOPs.

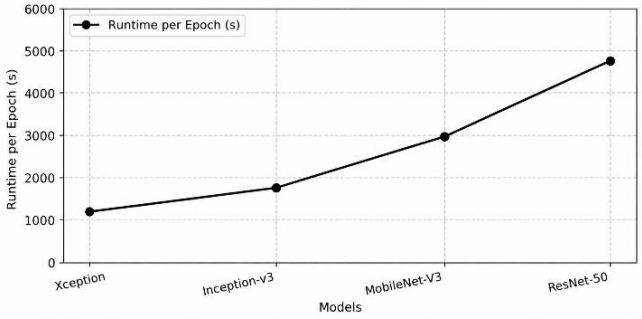

Fig. 16. Runtime per epoch for the four models.

In the present experiment, Xception showed the best trade-off between classification performance and computational demand overall. It achieved the highest accuracy and the lowest loss with a significantly lower computational cost than ResNet-50. Inception-v3 showed a competitive performance, but was less efficient than Xception, while MobileNet-v3 was still the most appropriate lightweight choice with a lower number of operations. It is deeper in architecture and computationally more expensive. However, ResNet-50 did not provide a similar improvement in forecasting performance.

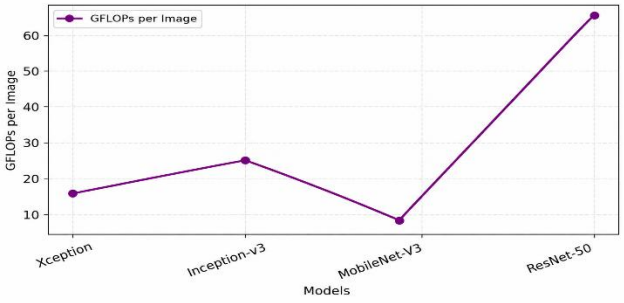

Fig. 17. GFLOPs per image for the four models.

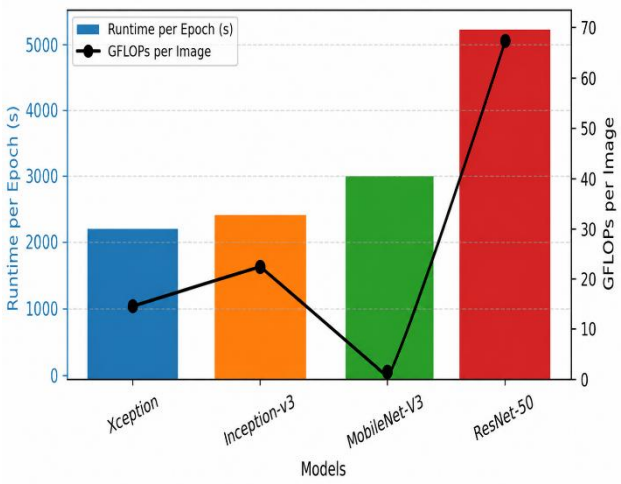

Fig. 18. Computational cost: runtime and GFLOPs.

VI. CONCLUSION AND FUTURE PROSPECTS

In this study, a comparative experimental framework was presented to evaluate the performance of four deep convolutional neural network architectures on the task of classifying fundus images associated with diabetic retinopathy: Xception, Inception-v3, MobileNet-v3 and ResNet-50. The study used a balanced dataset comprising 25,000 colour fundus images, equally divided among five classes, with 5,000 images per class. The dataset is divided to 80% for training, 10% for validation and 10% for testing, which corresponds to 20,000 images for training, 2,500 images for validation and 2,500

images for testing. This division is necessary for a more objective assessment of the performance of the models, since the models are trained on a large set of images, their generalization capacity is monitored during the training process, and finally, they are tested on independent data that were not part of the learning process. Class balance also helps the models avoid being biased towards a particular class and makes metrics such as accuracy, recall, precision, and F1-score more interpretable and comparable.

The results showed that the Xception model performed the best among the four models in general. The proposed model achieved 98.38% training accuracy, 98.52% validation accuracy, and 98.32% test accuracy. It also boasted the lowest loss values with a training loss of 8.05% and a validation loss of 7.80%. The results obtained show that Xception obtained a high accuracy on the training data, as well as a good generalization ability to the validation and test data. This is important in the medical field as a model that only learns the characteristics of the training data without generalizing to new data is not suitable for use in real-world environments. Moreover, the training, validation, and test accuracies converge, which indicates that the model does not suffer from obvious overfitting and is able to learn stable and effective visual representations for the fundus images.

The reason for Xception's better performance in this study is that it is designed with depthwise separable convolutions. This It also enables the extraction fine visual features and reduces the computational load with the heavier models. This property seems useful in the context of fundus images, as it allows the model to capture small local patterns such as hemorrhagic spots or subtle changes in the retina, while still maintaining a good ability to represent the overall structure of the image. The confusion matrix for this model shows that most of the correct classifications are near the main diagonal, with a small spread of errors across the classes, indicating a good ability to discriminate between the five classes.

The Inception-v3 model also performed well with a validation accuracy of 95.30% and a test accuracy of 95.04%. The performance indicates that the multi-branch structure of the model can extract visual features from different levels, which is relatively well-suited for fundus images with both fine and coarse details simultaneously. However, Inception-v3 still had worse performance than Xception in terms of accuracy, loss, and computational cost per image. Inception-v3 needs more GFLOPs per image than Xception and does not improve the accuracy. Thus, it can be regarded as a robust and stable model but not the most efficient one among the models tested in this study.

For MobileNet-v3, the test accuracy was 90.00%, which is lower than those of Xception and Inception-v3, but still an important result given the computational efficiency. MobileNet-v3 was mainly designed to be a lightweight model that can be used in resource-limited environments such as mobile devices, embedded systems, or low-cost medical screening devices. The model has the lowest GFLOPs per image, and thus is suitable when the goal is to minimize the computational load rather than to achieve the highest possible accuracy. However, its lower accuracy compared to more powerful models emphasizes the need to be cautious when using it in sensitive medical applications, especially when the aim is to support a diagnostic decision that requires a high degree of confidence.

ResNet-50 provided the lowest accuracy of the four models, with a test accuracy of 89.00%, but the highest computational cost in this experiment. The result indicates that the increase in the depth or in the architectural complexity does not always lead to performance improvement, especially when the architecture is not well matched to the nature of the data, or when the increase in the computational costs is not compensated by a clear gain in accuracy. ResNet-50 showed a long training time and high GFLOPs without outperforming the other models. Thus, the present results show that this model does not seem to be the best choice for fundus image classification in this experimental setup, especially if the study aims to balance performance and efficiency.

The overall results confirm the compromise between accuracy and computational cost as a key factor in the choice of a suitable model. In this study, when the priority is higher diagnostic performance, Xception is the best choice because it has high accuracy, low loss, and clear generalization ability. However, if the objective is to create a lightweight system, MobileNet-v3 is a better option to consider, despite having a lower accuracy than powerful models. Inception-v3 is a better middle ground in performance but less efficient than Xception in these results. On the other hand, ResNet-50 does not provide any clear practical advantage in this setting, given the high computational cost relative to its accuracy.

The value of this study is that it does not just rank models by their accuracy, but provides a multidimensional view of their performance by combining accuracy, loss, classification metrics, confusion matrices, training time, and computational cost. This type of evaluation is critical in medical AI applications, as a model ready for clinical deployment must be accurate, stable, interpretable, generalizable, and able to run within realistic time and compute constraints. Based on the reported results, Xception can be regarded as the most balanced architecture within the present experimental setting in this study and MobileNet-v3 as a promising option for lightweight applications; Inception-v3 and ResNet-50 need further optimization or fine-tuning before being used in a practical diagnostic assistance system. However, the findings of this study need to be interpreted in view of the limitations of its experimental setting. While the dataset is sufficiently large and balanced to compare the models, in order to generalize the results to real-world clinical settings, the models need to be tested on independent external datasets obtained across different imaging devices and multiple medical institutions. Moreover, high performance in a controlled experiment does not ensure that the model will perform at the same level with low-quality images, images under different lighting conditions, or images from patients with different demographic backgrounds. The present results thus represent an important step toward the establishment of a preliminary benchmark but do not exclude the need for further validation studies before any direct clinical application.

This study has some limitations; the experimental evaluation was performed on one publicly available diabetic retinopathy

dataset. This allows a fair comparison of the evaluated CNN architectures in the same experimental conditions, but further validation with independent datasets from different clinical settings and imaging devices is needed to further evaluate the robustness and the generalizability of the proposed framework.

Although this study provided a clear empirical comparison of four deep learning architectures to classify fundus images related to diabetic retinopathy, future work could focus first on expanding the database used, both in terms of the number of images and the variety of their sources. Models were evaluated on images obtained from different imaging devices, different medical institutions and diverse clinical cases will give an insight for the generalizability of the models in conditions closer to clinical settings of real-world. Moreover, the utilization of independent external databases will enhance the robustness of the obtained results, as good performance in a single internal split does not, on its own, prove the validity of the model to be used in different clinical settings.

The study can be further extended by using newer and more diverse models such as EfficientNet, ConvNeXt, Vision Transformers, and hybrid models combining convolutional networks and attention mechanisms. This enables us to investigate whether modern architectures can improve the trade-off between accuracy and computational efficiency. In addition, it is possible to improve the image preprocessing step by using fundus image-specific techniques such as contrast enhancement, light adjustment, removal of irrelevant areas, and automatic image quality detection before the images are introduced into the model.

In future work, each training experiment should be repeated with a number of different independent random seeds. Additionally, the mean, standard deviation and confidence interval for the main performance indicators should be reported. Also, before reaching some conclusions about the superiority of the model, it is required to perform paired bootstrap comparisons or other statistical tests considered relevant.

In the clinical context, the proposed framework aims to support ophthalmologists by providing quick and consistent initial assessments of diabetic retinopathy, not to replace expert clinical judgment. In real-world screening programs, the reduction of false negative predictions is particularly important to avoid missed diagnoses and delayed treatment. Therefore, additional clinical validation, workflow integration, and prospective evaluation are required before the proposed framework can be considered for routine clinical use.

A crucial aspect for future development is the integration of visual interpretation techniques such as Grad-CAM and Grad-CAM++ to identify the regions on which the model relies to make its decision. This is especially true in the medical domain where doctors need not only the final classification but also an explanation that can help them understand the reasoning behind the decision. Moreover, a more fine-grained error analysis can be performed, in particular between neighboring grades of diabetic retinopathy, which are often of higher clinical relevance. Finally, future work can be focused on building a medical decision support system with accuracy, efficiency, and interpretability. The system may include image classification, a confidence score and an interpretive map indicating the regions that influence the decision. In the future, the model could be extended to a multimodal one by combining fundus images with clinical data such as duration of diabetes, HbA1c level, blood pressure, and age of the patient. This development would bring the model more in line with the real world of medical practice, while emphasizing that its role must be that of an aid to the physician and not a substitute for specialized clinical evaluation.

REFERENCES

[1] A. Arora, Diabetic Retinopathy Resized Arranged Dataset, Kaggle, 2019. Available: https://www.kaggle.com/datasets/amanneo/diabetic-retinopathy-resized-arranged

[2] N. Ayesha, "A Vision Transformer-Based Convolutional Neural Network for the Automated Diagnosis of Eye Diseases Using Self-Attention Mechanisms," Engineering, Technology & Applied Science Research, vol. 15, no. 4, pp. 24493–24497, 2025, doi: 10.48084/etasr.10649.

[3] M. Sushith et al., "A hybrid deep learning framework for early detection of diabetic retinopathy using retinal fundus images," Scientific Reports, vol. 15, Art. no. 15166, 2025, doi: 10.1038/s41598-025-99309-w.

[4] Akhtar, S., Aftab, S., Ali, O. et al. A deep learning based model for diabetic retinopathy grading. Sci Rep 15, 3763 (2025). https://doi.org/10.1038/s41598-025-87171-9

[5] J. Zhang, R. Li, C. Liu, and X. Ji, "Improving Domain Transfer with Consistency-Regularized Joint Distribution Alignment for Medical Image Classification," Symmetry, vol. 17, no. 4, Art. no. 515, 2025, doi: 10.3390/sym17040515.

[6] M. T. Hayat et al., "A Hybrid Convolutional Neural Network–Long Short-Term Memory (CNN–LSTM)–Attention Model Architecture for Precise Medical Image Analysis and Disease Diagnosis," Diagnostics, vol. 15, no. 21, Art. no. 2673, 2025, doi: 10.3390/diagnostics15212673.

[7] A. S. Abdalrada, A. F. Neamah, H. L. Majeed, and A. R. Al-Sudani, "A Hybrid Machine Learning Approach for Enhanced Diabetes Prediction: Integrating Image and Numerical Data," Mesopotamian Journal of Big Data, pp. 211–221, 2025.

[8] A. Shazia et al., "Automated Early Diabetic Retinopathy Detection Using a Deep Hybrid Model," Trends in Emerging Technologies and Artificial Intelligence, 2024.

[9] S. Sundar, S. Subramanian, and M. Mahmud, "Classification of Diabetic Retinopathy Disease Levels by Extracting Spectral Features Using Wavelet CNN," Diagnostics, vol. 14, no. 11, Art. no. 1093, 2024, doi: 10.3390/diagnostics14111093.

[10] S. V. Chilukoti et al., "A reliable diabetic retinopathy grading via transfer learning and ensemble learning with quadratic weighted kappa metric," BMC Medical Informatics and Decision Making, vol. 24, no. 1, Art. no. 37, 2024, doi: 10.1186/s12911-024-02446-x.

[11] K. Sriporn, C.-F. Tsai, L.-J. Rong, P. Wang, T.-Y. Tsai, and C.-W. Chen, "Optimizing Deep Learning for Diabetic Retinopathy Diagnosis," International Journal of Advanced Computer Science and Applications, vol. 15, no. 11, 2024, doi: 10.14569/IJACSA.2024.0151135.

[12] L. Arora et al., "Ensemble deep learning and EfficientNet for accurate diabetic retinopathy diagnosis," Scientific Reports, vol. 14, 2024.

[13] W. K. Wong, F. Juwono, and C. Capriono, "Diabetic Retinopathy Detection and Grading: A Transfer Learning Approach Using Simultaneous Parameter Optimization and Feature-Weighted ECOC Ensemble," IEEE Access, 2023, doi: 10.1109/ACCESS.2023.3301618.

[14] C. Mohanty et al., "Using Deep Learning Architectures for Detection and Classification of Diabetic Retinopathy," Sensors, vol. 23, no. 12, Art. no. 5726, 2023, doi: 10.3390/s23125726.

[15] M. M. Butt, D. A. Iskandar, S. E. Abdelhamid, G. Latif, and R. Alghazo, "Diabetic Retinopathy Detection from Fundus Images of the Eye Using Hybrid Deep Learning Features," Diagnostics, vol. 12, no. 7, Art. no. 1607, 2022, doi: 10.3390/diagnostics12071607.

[16] Farag, Mohamed & Fouad, Mariam & Abdel-Hamid, Amr. (2022). Automatic Severity Classification Of Diabetic Retinopathy Based On DenseNet And Convolutional Block Attention Module. IEEE Access. 10. 1-1. 10.1109/ACCESS.2022.3165193.

[17] Mushtaq, Gazala & Siddiqui, Farheen. (2021). Detection of diabetic retinopathy using deep learning methodology. IOP Conference Series: Materials Science and Engineering. 1070. 012049. 10.1088/1757-899X/1070/1/012049.

[18] A. A. Ayala et al., "Diabetic Retinopathy Classification with Deep Learning via Fundus Images: A Short Survey," Procedia Computer Science, 2021.

[19] K. Ladrham and H. Gueddah, "Optimizing Dermatological Image Classification Using an Efficient Convolutional Neural Network Architecture," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 16, no. 12, pp. 577–587, 2025.