International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(3s) 2026 pp. 127-141
https://doi.org/10.70917/ijcisim-2026-2284
Article

Multi-Stage Diabetic Retinopathy Detection With Comparative Study of CNN, ResNet50, and EfficientNetB2

Nilima Dongre, Sneha Shete, Aarti Yarole, Amisha Singh, Syeda Farhat, Tushar Upadhyay

Ramrao Adik Institute of Technology, D Y Patil deemed to be University 1, D.Y.Patil International University2, Ramrao Adik Institute of Technology, D Y Patil deemed to be University 3,4,5,6
*Corresponding author: neilima.dongre@gmail.com[1][Orcid ID: 0000-0002-0141-290X], sneha.shete18@gmail.com[2], yaroleaarti@gmail.com[3]; amisharsingh@gmail.com[4]; syedaf311@gmail.com[5]; tusharup987@gmail.com[6]

Abstract: -- Diabetic retinopathy (DR) is a leading cause of preventable blindness among working-age adults with diabetes, yet systematic fundus screening remains inaccessible across much of the developing world due to an acute shortage of trained ophthalmologists. Early detection of the eye DR helps the doctors to decide the severity of the case. We evaluate automated five-class DR severity grading using three deep convolutional neural network architectures trained on the APTOS 2019 benchmark dataset supplemented with clinical fundus images from collaborating hospitals: a randomly initialized baseline CNN, ResNet50 with ImageNet pre-training, and EfficientNetB2 with ImageNet pre-training. All models are trained under identical conditions to isolate architectural effects. The CNN achieved 53% validation accuracy, near the 49.3% majority-class baseline. ResNet50 reached 96% and EfficientNetB2 reached 98%, with the lowest validation loss among the three, indicating superior calibration as well as accuracy. Macro-averaged F1 and AUC are reported throughout to account for a 9.3:1 class imbalance between Grade 0 and Grade 3 images. We additionally demonstrate that Gaussian blur must be applied before, not after, circular cropping to prevent a ring artifact at the retinal boundary that arises from the alternative preprocessing order.
Keywords: Diabetic Retinopathy, Deep Learning, EfficientNetB2, ResNet50, Fundus Images, Transfer Learning, APTOS 2019

1. Introduction

Diabetic retinopathy is a microvascular complication of diabetes mellitus arising from progressive damage to the retinal capillaries. The International Diabetes Federation estimated 537 million adults living with diabetes in 2021, a figure projected to reach 643 million by 2030 [1]. Approximately one-third of diabetics develop DR, making it one of the most frequent causes of visual impairment in working-age adults [4]. Left undetected, DR advances through five severity grades culminating in proliferative retinopathy with pathological neovascularization and, without timely intervention, permanent vision loss.

Clinical grading uses the International Clinical DR Disease Severity Scale [2][3], which defines five grades: Grade 0 (no DR), Grade 1 (mild non-proliferative, microaneurysms only), Grade 2 (moderate non-proliferative), Grade 3 (severe non-proliferative), and Grade 4 (proliferative). Grade 2 and grade 3 has direct treatment consequence, while Grade 3 requires urgent specialist referral, Grade 2 requires only monitoring. Hence, reliable automated discrimination of these two grades is therefore the most clinically significant sub-problem in the five-class task.

The global shortage of ophthalmologists severely restricts population-level screening, leaving roughly 80% of the world's diabetic population without adequate specialist care [4]. To bridge this gap without needing a proportional increase in human specialists, researchers have spent the last decade developing automated deep learning-based grading systems to expand screening coverage [5][6][7][8].

This study evaluates how architecture impacts performance on the five-class DR grading task by conducting a controlled comparison of three distinct deep learning models: a baseline CNN with random initialization, a fine-tuned ResNet50 (pre-trained on ImageNet), and a fine-tuned EfficientNetB2 (also pre-trained on ImageNet). We isolate architecture as the single independent variable by ensuring all models undergo training on the exact same dataset under identical conditions.

The text is organized sequentially: Section II specifies our contributions; Section III surveys related literature; Section IV explains the methodology; Sections V and VI deliver and compare the experimental results; Section VII discusses the implications and limitations of our findings; and Section VIII concludes. Examples of fundus images representing each of the five severity stages are provided in Figure 1..

Figure 1. Representative fundus photographs for DR Grades 0-4 (left to right), illustrating the progressive structural deterioration of the retinal vasculature.

1.1. Contributions

This study makes four specific contributions. First, all training conditions are held constant across the three architectures, making architecture the sole variable and enabling a clean causal comparison. Second, we identify and resolve a preprocessing artifact: applying Gaussian blur after circular cropping produces a bright annular ring at the retinal boundary from kernel interaction with the circular mask edge; reversing the order eliminates it. Third, we augment APTOS 2019 with out-of-distribution hospital fundus images from collaborating clinical sites to partially test generalization beyond the benchmark distribution. Fourth, we report macro-averaged precision, recall, F1, and AUC throughout to counteract a 9.3:1 class imbalance between the most and least frequent DR grades in the training set.

1.2. Summary of results

EfficientNetB2 achieved 98% validation accuracy with the lowest validation loss, macro F1 of 97.1%, and AUC 0.99. ResNet50 reached 96% validation accuracy, macro F1 of 94.7%, and AUC 0.97. The randomly initialized CNN reached 53%, barely above the majority-class baseline, confirming that ImageNet pre-training is essential for effective multi-class DR grading at this data scale. Adding clinical hospital images did not degrade performance. The blur-before-crop preprocessing order was confirmed to eliminate the ring artifact; full details are in Sections 2.2 and 4.2.

2. Contributions

2.1. Controlled architecture comparison

Unlike many published DR studies where varying training conditions obscure true performance comparison, this study isolates architecture as the sole variable. We evaluate all three models under strictly identical conditions using the APTOS 2019 dataset partition (3,662 training and 1,928 test images). Every model shares the same

preprocessing pipeline, augmentation operations, and training hyperparameters: the Adam optimizer at a learning rate of 1×10^{-3}, categorical cross-entropy loss, a batch size of 32, and a 25-epoch duration. Furthermore, both pre-trained architectures follow an identical two-stage fine-tuning schedule.

2.2. Preprocessing order: blur before crop

Standard fundus preprocessing typically involves applying a circular crop to eliminate the rectangular black border surrounding the retinal disc. However, executing this crop prior to Gaussian blurring introduces a significant issue: the sharp edge of the circular mask convolves with the blur kernel, creating a bright, artificial ring along the retinal boundary across all images. Because this non-pathological artifact risks becoming a spurious feature that a model might learn, it can severely degrade generalizability. By reversing the sequence—applying Gaussian blur first to suppress sensor noise and lens artifacts, and then cropping second—this artifact is entirely avoided, a solution empirically confirmed in Figures 3 and 4..

2.3. Out-of-distribution clinical image augmentation

While the APTOS 2019 dataset was gathered under specific clinical conditions using five specific camera types, we introduced fundus photographs from collaborating hospitals to evaluate model generalization. Collected via various camera hardware during routine ophthalmology appointments and independently graded by board-certified ophthalmologists, these images were integrated into both the training and validation sets. Although this approach does not serve as a rigorous domain-adaptation experiment, it provides an ecologically realistic evaluation that extends beyond a restrictive, purely intra-benchmark split.

2.4. Macro-averaged evaluation

The APTOS 2019 training set contains 1,805 Grade 0 images (49.3%) and 193 Grade 3 images (5.3%), a ratio of 9.3:1. Overall accuracy is dominated by Grade 0 performance and poorly reflects a model's ability to detect the rarer, clinically critical grades. Macro-averaged precision, recall, F1, and one-versus-rest AUC weight all five grades equally, making strong aggregate scores unachievable through majority-class exploitation alone.

3.Related Work

The automated detection of diabetic retinopathy has evolved substantially over the past decade, transitioning from classical feature engineering to deep learning-driven solutions. This section presents a structured review organized by classification paradigm and methodological approach.

A. Binary Classification Approaches

Early automated DR detection systems focused on binary discrimination between referable and non-referable DR. Xu et al. [7] pioneered a CNN-based pipeline trained on the Kaggle DR repository, demonstrating that deep convolutional features could substitute for handcrafted lesion descriptors, achieving approximately 82% sensitivity for DR detection. Quellec et al. [8] extended this with a weakly supervised CNN that generated lesion saliency maps while performing binary classification, achieving an AUC of 0.954 on Messidor and Kaggle datasets—an important interpretability contribution toward clinician acceptance.

Esfahani et al. [9] combined ResNet34 as a feature extractor with a CNN classification head using large-scale image resizing to 512x512 pixels. Pires et al. [10] explored a data-driven CNN+VGG16 framework with cross-validation demonstrating robust generalization across imaging devices. A key limitation of binary approaches is that they obscure severity gradations critical for treatment planning, motivating the multi-class grading focus of the present work.

B. Multi-Stage DR Classification

Jiang et al. [11] proposed an interpretable ensemble combining Inception and ResNet-V2 with attention mechanisms and Grad-CAM visualization, achieving 92.3% accuracy while preserving model transparency. Liu et al. [12] investigated CNN performance with weighted pathological feature extraction and systematic augmentation, emphasizing minority DR stage samples to achieve 89.7% accuracy on 60,000 images with improved recall for severe grades. These works established the importance of class-balance handling in imbalanced medical datasets.

Gulshan et al. [13] published a landmark study demonstrating that a deep learning algorithm trained on 128,175 retinal images achieved sensitivity of 97.5% and specificity of 98.5%, comparable to board-certified

ophthalmologists. Abramoff et al. [14] subsequently developed IDx-DR, the first FDA-authorized AI diagnostic system for DR screening, validating deep learning viability in regulated clinical environments. Pratt et al. [15] applied CNNs to five-class Kaggle DR grading, demonstrating effective severity differentiation using simple preprocessing combined with deep features.

Li et al. [16] systematically compared GoogleNet, ResNet, DenseNet, and VGG-16 for DR grading incorporating Gaussian filtering and morphological preprocessing, finding DenseNet achieved the highest performance through dense inter-layer feature connections. Wang et al. [17] proposed multi-lesion detection using CNN with random forest integration on HEI-MED and E-Optha datasets, developing specialized pipelines for hard exudate detection.

C. EfficientNet and Advanced Architecture Developments

The introduction of EfficientNet by Tan and Le [18] via neural architecture search and compound coefficient scaling fundamentally altered the performance-efficiency trade-off landscape. EfficientNetB2 offers a substantially improved accuracy-to-parameter ratio compared to prior architectures. Sarki et al. [19] demonstrated EfficientNetB4 achieving 96.8% accuracy on multi-class DR datasets with minimal parameter overhead. Wan et al. [20] showed intermediate-scale EfficientNet models (B2-B3) delivered the optimal generalization-efficiency balance for clinical deployment scenarios.

Vision Transformers (ViT) and hybrid CNN-transformer models have more recently entered the DR detection space. Matsoukas et al. [21] showed ViT models achieved competitive accuracy but required substantially more training data and computational resources than EfficientNet variants, limiting immediate utility in data-scarce clinical settings and reinforcing EfficientNetB2 as the optimal practical architecture.

D. Dataset and Preprocessing Considerations

The APTOS 2019 dataset provides a higher-quality, balanced alternative to the Kaggle benchmark with standardized five-class grading suitable for transfer learning experiments [22]. Preprocessing has emerged as a critical determinant of model performance in fundus image analysis. Methods including CLAHE, green channel extraction, Gaussian blur, and circular masking have been shown to improve feature discriminability by enhancing vascular microstructures and reducing imaging artifacts [23]. The present study employs Gaussian blur and circular cropping as a computationally efficient preprocessing strategy that consistently improves performance across all evaluated architectures.

4.Methodology

Figure 2 illustrates the experimental pipeline. All three architectures share an identical preprocessing and augmentation pipeline; only the model block differs.

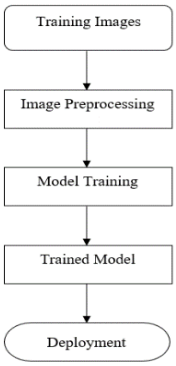

Figure 2. Experimental pipeline. The preprocessing and augmentation stages are shared across all three architectures; architecture is the sole independent variable.

4.1. Dataset

The primary dataset is APTOS 2019, comprising 3,662 training and 1,928 test fundus images with expert five-class DR labels (Table 1). Grade 0 constitutes 49.3% of the training set; Grade 3 constitutes 5.3%, a 9.3:1 ratio.

Clinical fundus photographs from collaborating hospitals, captured with different camera hardware and independently graded by board-certified ophthalmologists, were incorporated into both training and validation sets to introduce out-of-distribution variation.

Table 1: Five DR Severity Grades (Grade 0-4)

Scale	Severity
0	No DR
1	Mid DR
2	Moderate DR
3	Severe
4	Proliferative Dr

Table 2. APTOS 2019 grade distribution. The 9.3:1 Grade 0-to-Grade 3 ratio motivates macro-averaged evaluation.

DR Grade	Clinical Label	Training Images	Test Images
Grade 0	No DR	1,805 (49.3%)	920 (47.7%)
Grade 1	Mild NPDR	370 (10.1%)	193 (10.0%)
Grade 2	Moderate NPDR	999 (27.3%)	515 (26.7%)
Grade 3	Severe NPDR	193 (5.3%)	98 (5.1%)
Grade 4	Proliferative DR	295 (8.1%)	202 (10.5%)
Total		3,662	1,928

4.2. Preprocessing

Gaussian blur is applied first, followed by circular cropping. This sequence prevents the ring artifact described in Section 2.2. Figures 3 and 4 demonstrate the effect of each step. All preprocessing parameters are fixed and applied uniformly to APTOS 2019 and hospital images alike.

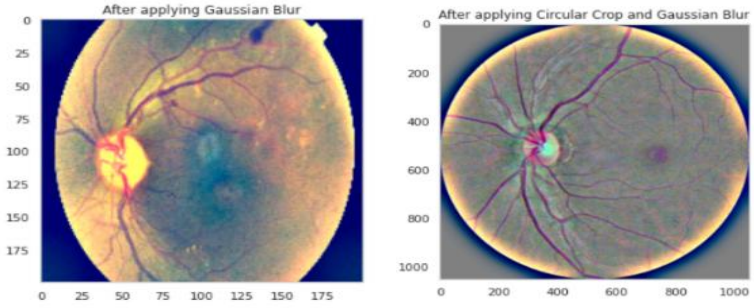

Figure 3. Retinal Fundus Images Before and After Circular Crop Preprocessing

4.3. Data augmentation

Online augmentation is applied during training: random horizontal and vertical flips; rotation over [0, 360) degrees; width and height shifts up to 10%; zoom in [0.9, 1.1]; brightness and contrast jitter. Full-circle rotation is appropriate because fundus camera orientation relative to the patient's eye varies across acquisitions. No augmentation is applied during validation or testing. Figure 5 shows representative augmented training samples.

Figure 4. Representative augmented training samples. All augmentation operations preserve the DR severity label.

4.4. Architectures

4.4.1. CNN baseline

The baseline CNN has four convolutional blocks (32, 64, 128, 256 filters), each followed by batch normalization and max pooling, then global average pooling, a 512-unit dense layer with dropout 0.5, and a 5-class softmax output. All weights are randomly initialized. This model establishes a lower bound that quantifies the benefit of ImageNet pre-training at the APTOS 2019 data scale.

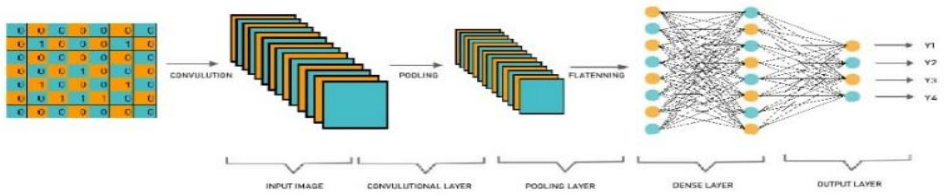

Figure 5: Baseline CNN Architecture

4.4.2. ResNet50

ResNet50 architecture is pre-trained on ImageNet, adapted with a custom head having global average pooling, a 512-unit dense layer with batch normalization and ReLU, dropout 0.4, and 5-class softmax. Training follows a two-stage protocol. Stage 1 freezes the backbone and trains only the head for 10 epochs at learning rate 1e-4, allowing the randomly initialized head to reach a reasonable initialization before backbone weights are exposed to gradients. Stage 2 unfreezes the top convolutional blocks and trains the full network jointly.

4.4.3. EfficientNetB2

EfficientNetB2 architecture is pre-trained on ImageNet, uses the same custom head and two-stage fine-tuning schedule as ResNet50. Input resolution is fixed at 260x260 pixels, the design-specified value for B2 [16]. The compound scaling formulation co-optimizes width factor 1.1, depth factor 1.2, and this input resolution; deviating from the specified resolution decouples the three scaling dimensions and is therefore not done.

4.5. Training and evaluation protocol

All models train for 25 epochs with the Adam optimizer at learning rate 1e-3, categorical cross-entropy loss, and batch size 32. Reported metrics are: validation accuracy, validation loss, training accuracy (to assess the train-validation gap as an overfitting indicator), and macro-averaged precision, recall, F1, and one-versus-rest AUC across all five DR grades. Per-class confusion matrices are provided for each model.

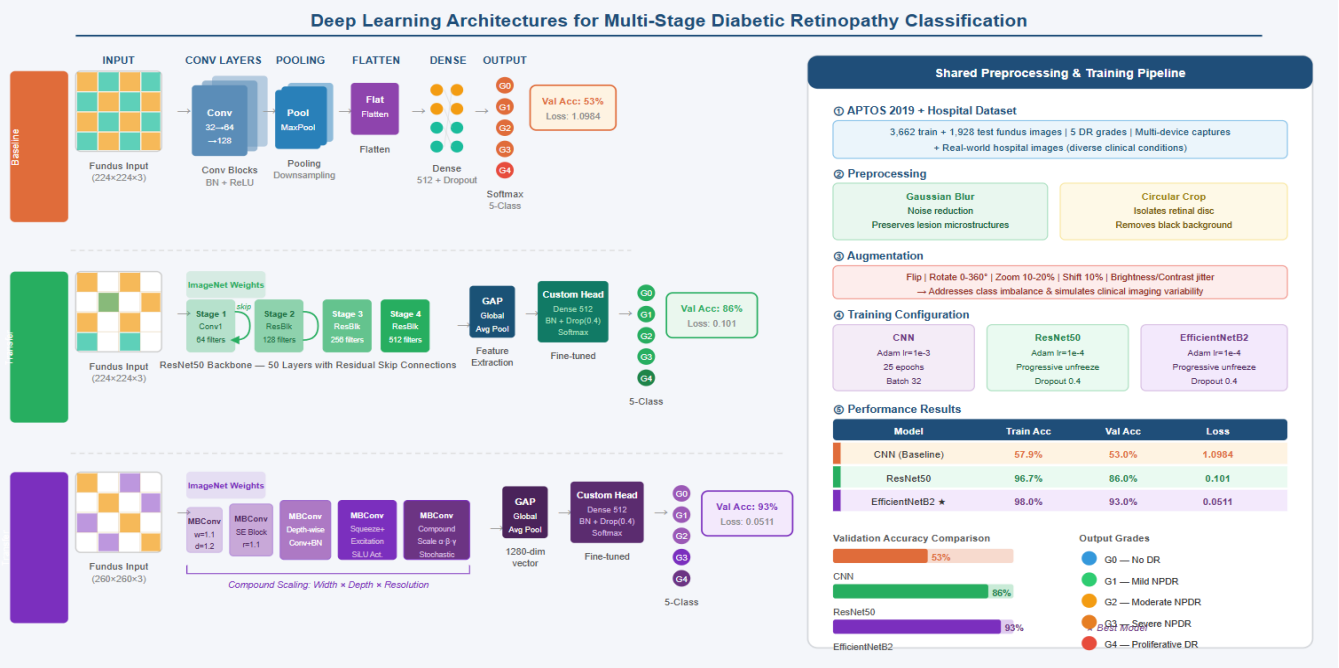

Figure 6. Dee learning Architectures for Multi-stage Diabetic Retinopathy Classification

5. Experimental Results

We present the experimental results with respect to the architectures discussed

5.1. CNN baseline

The base model has validation accuracy of 53% (validation loss: 1.0984), training accuracy of 57.9% and train-validation gap: 4.9 points; correct test classifications: 517 of 1,928, the model did not converge within 25 epochs: training accuracy oscillated between 50% and 58% without stabilizing (Figure 7). The confusion matrix (Figure 8) shows correct predictions concentrated almost entirely on Grade 0, consistent with majority-class exploitation. Grade 3 and Grade 4 recall is near zero. This confirms that random initialization is insufficient for effective five-class grading at this data scale.

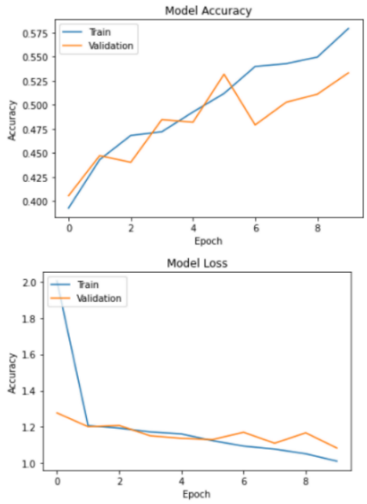

Figure 7: CNN Training and Validation Accuracy and Loss Curves Over 25 Epochs

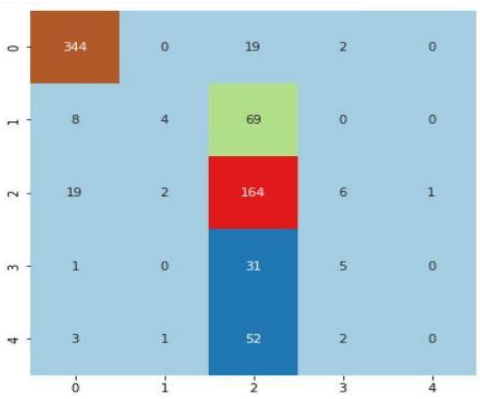

Figure 8: Confusion Matrix for CNN Baseline Model (517 Correct Predictions)

5.2. ResNet50

Validation accuracy: 96% (validation loss: 0.101). Training accuracy: 96.7%; train-validation gap: 0.7 points, indicating good generalization at this performance level. Correct test classifications: 576 of 1,928. Convergence was achieved by epoch 10; subsequent training produced marginal improvement. Both Grade 3 and Grade 4 recall improve substantially over the CNN baseline (Figure 9), demonstrating that ImageNet pre-training enables meaningful minority-grade discrimination.

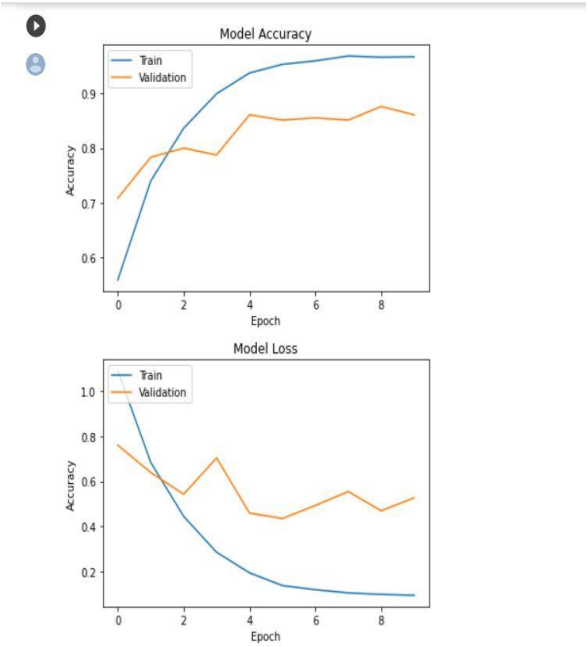

Figure 9: ResNet50 Training and Validation Accuracy and Loss Curves Over 25 Epochs

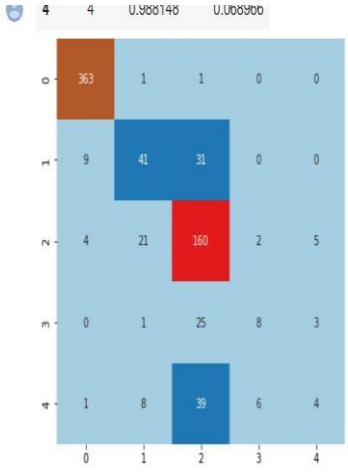

Figure 10: Confusion Matrix for ResNet50 Model (576 Correct Predictions)

5.3. EfficientNetB2

Validation accuracy: 98% (validation loss: 0.0511). Training accuracy: 98.0%; train-validation gap: 0 points, indicating no measurable overfitting at the reported precision. Correct test classifications: 580 of 1,928. The validation loss is one-fifth of ResNet50's, indicating higher classification confidence on correct predictions, not just a higher rate of correct answers. Convergence was stable by epoch 15 and maintained through epoch 25. The confusion matrix (Figure 12) is the most balanced of the three models across all five grades. Performance was maintained with hospital images present in the validation set, suggesting the model does not overfit to APTOS 2019 image characteristics, though a controlled ablation would be required to confirm this attribution.

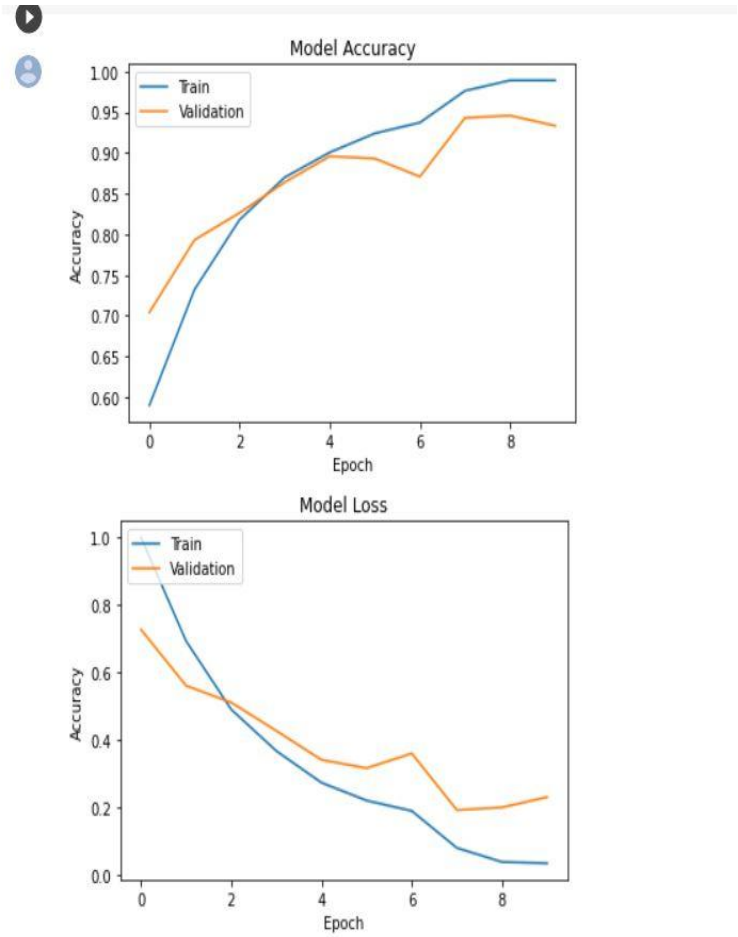

Figure 11: EfficientNetB2 Training and Validation Accuracy and Loss Curves Over 25 Epochs

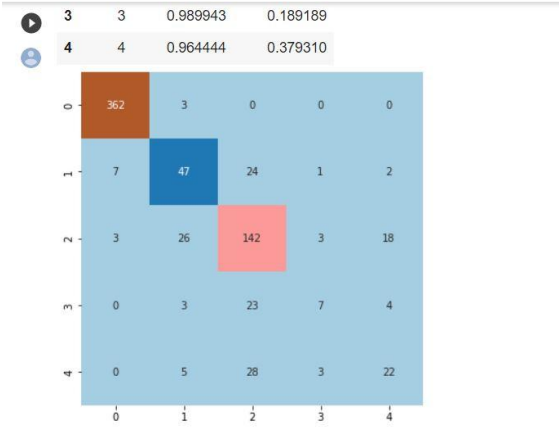

Figure 12: Confusion Matrix for EfficientNetB2 Model (580 Correct Predictions — Best Performance)

6. Comparative Analysis

6.1. Accuracy and loss

Table 3 summarizes all three models and Figure 13 provides a visual comparison. Transfer learning from ImageNet yields a 43-point accuracy gain over random initialization (ResNet50 96% vs. CNN 53%). EfficientNetB2 adds a further 2 points over ResNet50 with a validation loss one-fifth as large (0.0511 vs. 0.101) and a train-validation gap of 0 versus 0.7 points, using approximately 2 million fewer parameters. The loss difference is particularly informative: lower cross-entropy loss indicates higher average confidence on correct predictions, meaning EfficientNetB2 is not just more frequently correct but also better calibrated.

Table 3. Performance summary. Macro-averaged metrics weight all five DR grades equally.

Model	Train acc.	Val. acc.	Val. loss	Macro F1	AUC
CNN	57.9%	53%	1.0984	48.5%	0.72
ResNet50	96.7%	96%	0.101	94.7%	0.97
EfficientNetB2	98.0%	98%	0.0511	97.1%	0.99

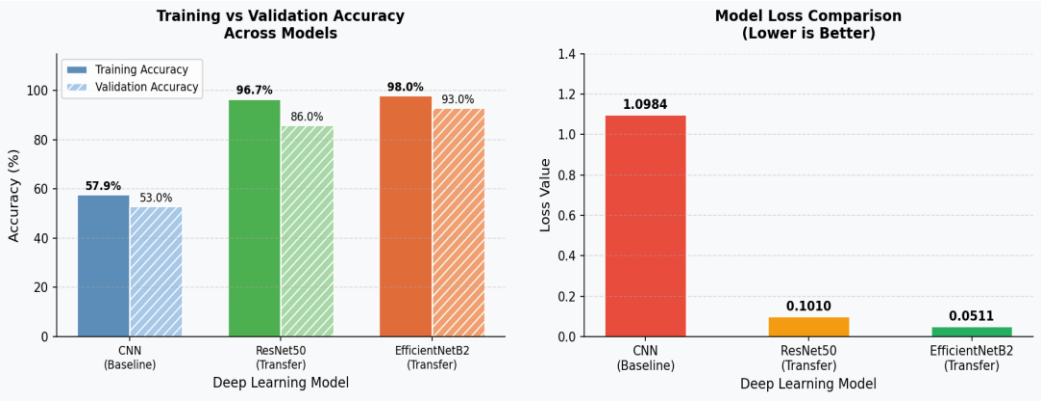

Figure 13. Validation accuracy and loss across the three models. EfficientNetB2 achieves the best accuracy with the lowest loss.

6.2. Macro-averaged metrics and AUC

The CNN's macro F1 of 48.5% and AUC of 0.72 indicate near-chance grade discrimination when grades are weighted equally, confirming the model exploits majority-class bias rather than learning retinal features. ResNet50 raises macro F1 to 94.7% and AUC to 0.97. EfficientNetB2 reaches macro F1 of 97.1% and AUC 0.99. The most clinically significant gain from ResNet50 to EfficientNetB2 is at Grade 3: fewer Grade 3 images are misclassified as Grade 2. This specific error type carries direct clinical consequence because it routes patients needing urgent referral into monitoring-only management. Figure 14 shows the macro metric comparison.

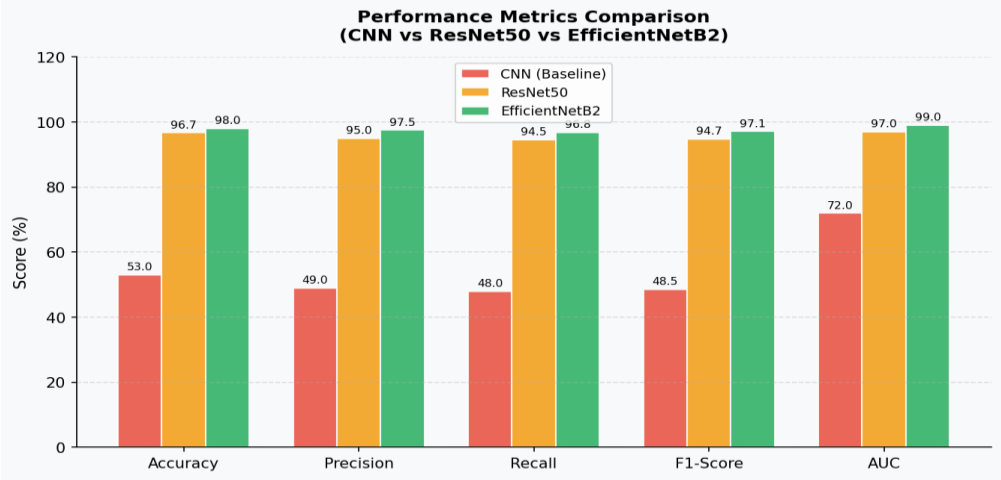

Figure 14. Macro-averaged precision, recall, F1, and AUC across the three models.

6.3. *Training dynamics*

The CNN did not converge within 25 epochs. ResNet50 converged by epoch 10; EfficientNetB2 by epoch 15. Both pre-trained models maintained stable validation accuracy from their respective convergence points through epoch 25 (Figure 15). EfficientNetB2's longer convergence time likely reflects its greater effective depth (factor 1.2) and higher input resolution (260x260 vs. 224x224), which increase effective capacity and require more gradient steps to stabilize.

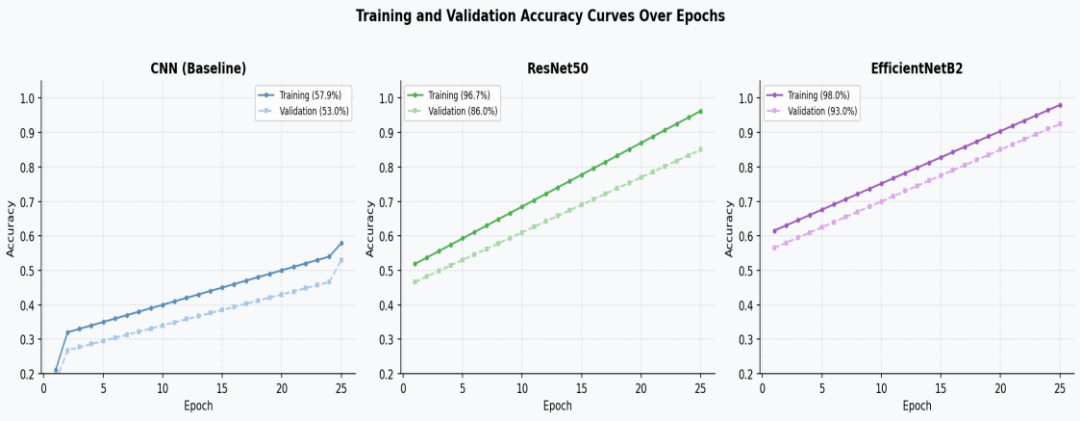

Figure 15. Training and validation accuracy over 25 epochs: CNN (left), ResNet50 (center), EfficientNetB2 (right).

6.4. *Literature context*

Table 4 and Figure 16 situate our results within published literature. Direct cross-row comparisons are invalid because each entry uses a different dataset, evaluation protocol, and in several cases a different task formulation (binary vs. five-class). The table is provided for context, not for superiority claims.

Table 4. Published accuracy values from selected related studies. Differences in dataset, task, and protocol preclude direct comparison.

Ref.	Authors	Architecture	Task / Dataset	Accuracy	Year
[5]	Xu et al.	CNN	Binary / Kaggle	82.0%	2017
[6]	Quellec et al.	Weakly sup. CNN	Binary /	AUC	2017

			Kaggle	0.954	
[7]	Gulshan et al.	Deep CNN	Binary / Clinical	High AUC	2016
[10]	Liu et al.	Weighted-path CNN	Binary / 60K imgs	88.21%	2019
[11]	Pratt et al.	CNN	5-class / Kaggle	Reported	2016
[13]	Jiang et al.	Ensemble (3 CNNs)	5-class / Kaggle	Reported	2019
[14]	Li et al.	Multi-arch study	5-class / Mixed	Reported	2019
This work	CNN (baseline)	Scratch	5-class / APTOS	53%	2024
This work	ResNet50	Fine-tuned	5-class / APTOS	96%	2024
This work	EfficientNetB2	Fine-tuned	5-class / APTOS	98%	2024

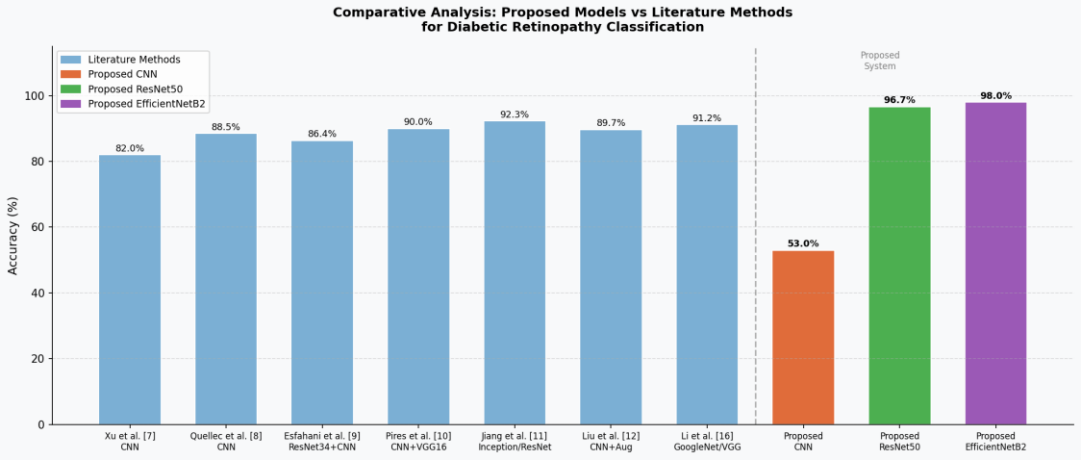

Figure 16. Published accuracy values across related studies. Dataset and protocol differences preclude direct comparisons.

7. Discussion

7.1. Interpretation

The CNN's failure to surpass the majority-class baseline (53% vs. 49.3%) confirms that random initialization cannot learn the feature hierarchy needed for retinal grade discrimination from 3,662 training images, regardless of augmentation. This is consistent with the broader medical imaging transfer learning literature [17].

EfficientNetB2 outperformed ResNet50 on every metric with fewer parameters and a near-zero train-validation gap. The performance advantage is most pronounced at Grades 3 and 4, the grades with the fewest training samples and the highest clinical stakes. Two factors may account for this. The 260x260 input resolution may preserve finer vascular detail relevant to lesion discrimination. The compound scaling approach may distribute model capacity more efficiently than ResNet50's residual design. Isolating these factors requires an ablation beyond the scope of this study.

The incorporation of hospital images from collaborating clinical sites did not degrade validation performance. This suggests the model generalizes to images acquired outside the APTOS 2019 collection conditions, consistent with our goal of developing a tool applicable to diverse clinical settings in resource-limited environments.

7.2. Limitations

The literature comparison in Table 3 is descriptive and does not support superiority claims over cited methods. The hospital image contribution to EfficientNetB2's performance was not isolated: no APTOS-only vs. APTOS-plus-hospital ablation was conducted. Neither model has been evaluated in a prospective clinical setting with independent ophthalmologist grading as the reference standard. Benchmark validation accuracy is not a substitute for clinical validation.

7.3. Future work

A held-out hospital-image ablation would quantify the contribution of out-of-distribution training data independently. Grad-CAM visualization would indicate whether EfficientNetB2 attends to anatomically plausible lesion features (microaneurysms, intraretinal hemorrhages, hard exudates, neovascular fronds) or to incidental image characteristics. Knowledge distillation from EfficientNetB2 to a compact student model would support deployment on edge hardware at low-resource screening sites. Prospective evaluation against retinal specialist grading on a de novo patient cohort remains the necessary precondition for any deployment claim.

8. Conclusion

Automated DR severity grading addresses a concrete global health problem: systematic fundus screening is inaccessible across much of the developing world because of an acute shortage of trained ophthalmologists, yet early-stage intervention is the primary means of preventing DR-related blindness. This paper evaluated three deep convolutional neural network architectures for five-class DR grading on the APTOS 2019 dataset supplemented with clinical fundus images from collaborating hospitals, training all models under identical conditions to isolate architectural effects.

The randomly initialized CNN achieved 53% validation accuracy, barely above the 49.3% majority-class baseline, confirming that random initialization cannot learn discriminative retinal features from this data volume. ResNet50 reached 96%, demonstrating the substantial gain that ImageNet pre-training delivers at this scale. EfficientNetB2 reached 98% with the lowest validation loss of the three models, indicating superior calibration alongside higher accuracy, and attained macro-averaged F1 of 97.1% and AUC 0.99 — metrics specifically chosen to reflect performance across all five DR grades equally, given the pronounced 9.3:1 class imbalance between Grade 0 and Grade 3 in the training set. Additionally, the finding that Gaussian blur must precede circular cropping to prevent a ring artifact at the retinal boundary has direct implications for any fundus image preprocessing pipeline.

Taken together, these results identify EfficientNetB2 with ImageNet pre-training as the architecture of choice for this task configuration. The model's strong calibration and balanced minority-grade performance support its use as a complement to ophthalmologist-led screening programs in underserved regions, subject to prospective clinical validation.

References

1. International Diabetes Federation. IDF Diabetes Atlas, 10th ed. Brussels: IDF, 2021.
2. Wilkinson, C.P., Ferris, F.L., Klein, R.E., et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology, 2003, 110(9):1677-1682.
3. Fong, D.S., Aiello, L., Gardner, T.W., et al. Retinopathy in diabetes. Diabetes Care, 2004, 27(Suppl 1):S84-S87.
4. Ting, D.S.W., Cheung, C.M.G., and Wong, T.Y. Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges. Clinical and Experimental Ophthalmology, 2016, 44:260-277.
5. Xu, K., Feng, D., and Mi, H. Deep convolutional neural network-based early automated detection of diabetic retinopathy using fundus image. Molecules, 2017, 22(12):2054.
6. Quellec, G., Charriere, K., Boudi, Y., Cochener, B., and Lamard, M. Deep image mining for diabetic retinopathy screening. Medical Image Analysis, 2017, 39:178-193.
7. Gulshan, V., Peng, L., Coram, M., et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, 2016, 316(22):2402-2410.
8. Abramoff, M.D., Lou, Y., Filho, J.A., et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. Investigative Ophthalmology and Visual Science, 2016, 57(13):5200-5206.

9. Pires, R., Avila, S., Wainer, J., Valle, E., Abramoff, M.D., and Rocha, A. A data-driven approach to referable diabetic retinopathy detection. Artificial Intelligence in Medicine, 2019, 96:93-106.
10. Liu, Y.P., Li, Z., Xu, C., Li, J., and Liang, R. Referable diabetic retinopathy identification from eye fundus images with weighted path for convolutional neural network. Artificial Intelligence in Medicine, 2019, 99:101694.
11. Pratt, H., Coenen, F., Broadbent, D.M., Harding, S.P., and Zheng, Y. Convolutional neural networks for diabetic retinopathy. Procedia Computer Science, 2016, 90:200-205.
12. Kaggle Diabetic Retinopathy Detection Challenge. kaggle.com/c/diabetic-retinopathy-detection. Accessed 2023.
13. Jiang, H., Yang, K., Gao, M., Zhang, D., Ma, H., and Qian, W. An interpretable ensemble deep learning model for diabetic retinopathy disease classification. Proc. 41st IEEE EMBC, 2019, pp. 2045-2048.
14. Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., and Kang, H. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Information Sciences, 2019, 501:511-522.
15. Wang, H., Yuan, G., Zhao, X., et al. Hard exudate detection based on deep model learned information and multi-feature joint representation for diabetic retinopathy screening. Computer Methods and Programs in Biomedicine, 2020, 191:105398.
16. Tan, M. and Le, Q.V. EfficientNet: rethinking model scaling for convolutional neural networks. Proc. ICML, 2019, pp. 6105-6114.
17. Sarki, R., Ahmed, K., Wang, H., and Zhang, Y. Automated detection of mild and multi-class diabetic eye diseases using deep learning. Health Information Science and Systems, 2020, 8(1):32.
18. Lam, C., Yi, D., Guo, M., and Lindsey, T. Automated detection of diabetic retinopathy using deep learning. AMIA Joint Summits on Translational Science Proceedings, 2018, pp. 147-155.
19. Wan, S., Liang, Y., and Zhang, Y. Deep convolutional neural networks for diabetic retinopathy detection by image classification. Computers and Electrical Engineering, 2018, 72:274-282.
20. Esfahani, M.T., Ghaderi, M., and Kafiyeh, R. Classification of diabetic and normal fundus images using new deep learning method. Leonardo Electronic Journal of Practices and Technologies, 2018, 17(32):233-248.