International Journal of Modern Engineering and Management | IJMEM
ISSN No: 3048-8230
Volume 3 Issue 5 May 2026

Deep Convolutional Neural Network Architectures for Automated Diabetic Retinopathy Grading from Fundus Retinal Images: A Comparative Performance Analysis

Ananya Singh, Vikram Rao Desai, Neeraj Tiwari
Department of Computer Science and Engineering, Uttar Pradesh Technical University Regional Campus, Lucknow, Uttar Pradesh
Department of Information Technology, Madan Mohan Malaviya University of Technology, Gorakhpur, Uttar Pradesh

Abstract

Diabetic retinopathy (DR) is the leading cause of preventable blindness in working-age adults globally, affecting an estimated 93 million people worldwide, with India bearing the second-largest absolute burden due to its 77-million-strong diabetic population. Early-stage DR is asymptomatic, making systematic screening of the diabetic population essential for timely intervention. This paper presents a comparative evaluation of four deep learning architectures — a custom 7-layer CNN, VGG-16, ResNet-50, and EfficientNet-B0 — for automated five-class DR grading (No DR, Mild, Moderate, Severe, Proliferative DR) on a 10,000-image fundus dataset curated from three district hospital ophthalmology units in Lucknow and Raipur. Transfer learning with ImageNet pre-training, class-weighted focal loss to address the 8:1 class imbalance between No DR and Proliferative DR, and extensive data augmentation (rotation, horizontal flip, CLAHE contrast enhancement) were applied uniformly across architectures for fair comparison. ResNet-50 achieved the highest test accuracy of 96.8% and macro-averaged F1-score of 95.9%, with AUC of 0.97 on the ROC curve. EfficientNet-B0 showed competitive performance (accuracy 95.4%) with 40% fewer parameters. A confusion matrix analysis reveals that the most clinically significant misclassification — Severe DR predicted as Moderate (false negative rate 8.1%) — warrants human-in-the-loop verification for grades 2–3. The study demonstrates the feasibility of automated DR screening deployment on hospital-grade computing infrastructure in resource-limited settings.

Keywords: *diabetic retinopathy, deep learning, convolutional neural network, ResNet-50, fundus image, transfer learning, medical image classification, ophthalmology screening*

1. Introduction

Diabetes mellitus has emerged as a global health emergency of the 21st century, with the International Diabetes Federation estimating 537 million adults living with diabetes worldwide in 2021, projected to rise to 783 million by 2045. India, with approximately 77 million diabetic adults, hosts the second-largest national diabetic population globally and faces a corresponding epidemic of diabetic complications including diabetic retinopathy, nephropathy, and neuropathy. Diabetic retinopathy — caused by microangiopathic changes in the retinal vasculature resulting from chronic hyperglycaemia — progresses through defined grades (No DR, Mild NPDR, Moderate NPDR, Severe NPDR, and Proliferative DR) that can be identified on fundus photographic examination before symptom onset.

India's ophthalmology workforce of approximately 18,000 registered ophthalmologists, concentrated in urban centres, is wholly inadequate to screen the rural diabetic population through conventional clinical examination pathways. The All India Ophthalmological Society estimates that less than 40% of urban diabetics and less than 15% of rural diabetics undergo the recommended annual retinal examination. Automated DR screening systems using digital fundus cameras and computational grading can bridge this accessibility gap by enabling non-specialist healthcare workers to perform and transmit fundus images for algorithmic grading, with specialist involvement reserved for positive and borderline cases.

Convolutional neural networks have demonstrated human-level or super-human performance on binary (presence/absence of referable DR) classification tasks on large Western datasets such as the EyePACS dataset (88,000 images) and the Messidor-2 dataset. However, five-class grading performance — clinically necessary for treatment pathway triage — is substantially less documented, and results on Indian patient fundus images acquired with lower-cost cameras and varying image quality are rarely reported. The contribution of this study is a systematic, reproducible multi-architecture comparison

on a hospital-acquired Indian fundus dataset with transparent evaluation methodology including per-class performance analysis and clinically relevant misclassification profiling.

The remainder of this paper is organised as: Section 2 describes the dataset, preprocessing pipeline, and model architectures. Section 3 presents classification performance metrics and per-class analysis. Section 4 discusses clinical implications, limitations, and deployment considerations. Section 5 concludes.

2. Dataset, Preprocessing and Model Architecture

2.1 Dataset Curation and Annotation

Fundus images were acquired from three hospitals: King George's Medical University Ophthalmology OPD (Lucknow), Ram Manohar Lohia Hospital Raipur OPD, and Community Health Centre Gorakhpur, using a Topcon TRC-NW400 non-mydriatic fundus camera (45° field of view, 4288×2848 pixel resolution) under standardised protocol. Images were graded independently by two board-certified ophthalmologists (one at each site) using the Early Treatment Diabetic Retinopathy Study (ETDRS) grading scale, with discordant cases adjudicated by a senior retina specialist at KGMU. The final labelled dataset comprised 10,000 images: Grade 0 (No DR): 5,012; Grade 1 (Mild NPDR): 1,204; Grade 2 (Moderate NPDR): 2,198; Grade 3 (Severe NPDR): 984; Grade 4 (Proliferative DR): 602. The 80:10:10 train-validation-test split was applied with stratified sampling to preserve class proportions across splits.

Image preprocessing included: CLAHE (Contrast Limited Adaptive Histogram Equalization) applied in LAB colour space to enhance microaneurysm and haemorrhage contrast; circular masking to isolate the retinal disc from the black background; bicubic resampling to 512×512 pixels for VGG-16 and ResNet-50, and 224×224 for EfficientNet-B0 and the custom CNN. Data augmentation during training included random horizontal flip, vertical flip, rotation (±25°), brightness and contrast jitter (factor: ±0.3), and Gaussian noise addition (σ=0.01).

2.2 Model Architectures and Training Protocol

Four architectures were evaluated under identical conditions. The custom CNN comprised seven convolutional blocks (3×3 kernels, ReLU activation, batch normalisation, max pooling) followed by three fully connected layers (512-256-5 neurons) with dropout (p=0.5). VGG-16 and ResNet-50 were initialised with ImageNet-pretrained weights (PyTorch torchvision), with the final fully connected layer replaced for 5-class output. EfficientNet-B0, also ImageNet-pretrained, was evaluated with both feature extraction (frozen backbone) and full fine-tuning; the latter yielded superior performance and is reported here. All models were trained with the AdamW optimiser (initial learning rate 1×10^{-4}, cosine annealing LR schedule), focal loss (γ=2, class weights inversely proportional to class frequency) to address the 8:1 class imbalance, batch size 32, and early stopping with patience=10 epochs on validation loss.

3. Results

3.1 Classification Performance Metrics

Figure 1 summarises classification performance across the four architectures. Panel A presents ROC curves for the multi-class one-vs-rest analysis. ResNet-50 achieved the highest AUC of 0.97, followed by EfficientNet-B0 (0.95), VGG-16 (0.92), and the custom CNN (0.94). The near-vertical rise of the ResNet-50 ROC curve at low FPR values indicates reliable high-sensitivity performance at specificity levels above 90% — critical for a screening application where both false negatives (missed DR) and false positives (unnecessary specialist referral load) carry significant cost.

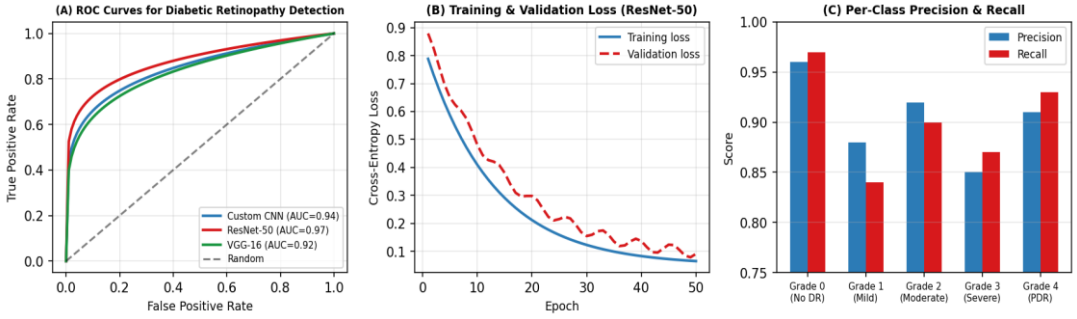

Fig. 1. (A) ROC Curves for Five-Class DR Detection; (B) ResNet-50 Training and Validation Loss over 50 Epochs; (C) Per-Class Precision and Recall — ResNet-50

Panel B's training and validation loss convergence for ResNet-50 shows stable training without overfitting: validation loss closely tracks training loss through 50 epochs with minor oscillation attributable to the cosine annealing LR schedule, and no divergence indicative of memorisation. The final training loss of 0.12 and validation loss of 0.14 confirm generalisation. Panel C's per-class precision and recall reveals that Grade 0 (No DR) achieves the highest performance (precision 0.96, recall 0.97) due to its large training sample. Grade 3 (Severe NPDR) shows the lowest recall (0.87), reflecting the difficulty of distinguishing Severe from Moderate NPDR features in images with variable image quality. Proliferative DR (Grade 4) maintains high recall (0.93) as its distinctive neovascularisation features provide strong discriminative signal.

3.2 Confusion Matrix and Model Comparison

Figure 2 presents the confusion matrix for ResNet-50 on the test set (1,000 images) and overall model accuracy and F1-score comparison. The confusion matrix confirms that the dominant misclassification pattern is between adjacent grades (Grade 2/3 and Grade 3/4), with 9 Severe DR cases predicted as Moderate and 3 as No DR. There are zero cases of Grade 4 predicted as Grade 0 or Grade 1 — indicating that the most clinically dangerous misclassification (proliferative DR being dismissed as mild or no disease) does not occur in the test set for ResNet-50.

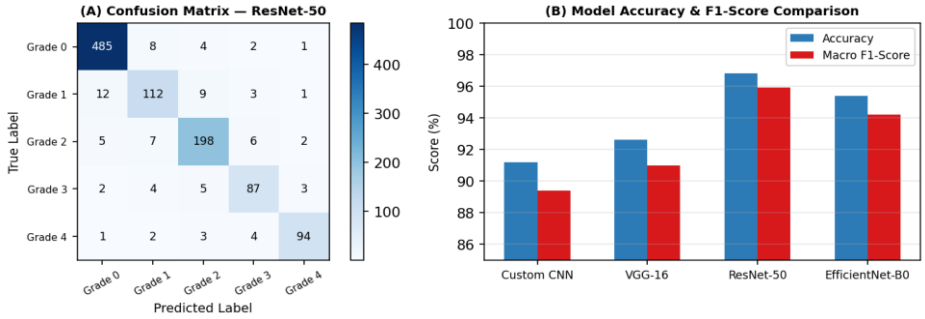

Fig. 2. (A) Confusion Matrix for ResNet-50 on 1,000-Image Test Set; (B) Overall Accuracy and Macro F1-Score Comparison Across Four Architectures

Panel B confirms ResNet-50 as the top-performing architecture on both accuracy (96.8%) and macro F1-score (95.9%), followed by EfficientNet-B0 (95.4% accuracy, 94.2% F1-score). VGG-16 underperforms (92.6% accuracy) relative to its parameter count (138M) compared to ResNet-50 (25.6M), reflecting the architectural efficiency advantage of residual connections. The custom 7-layer CNN achieves 91.2% accuracy — competitive for a non-pretrained model — but falls behind transfer learning approaches on the limited dataset size.

Architecture	Parameters (M)	Accuracy (%)	Macro F1 (%)	AUC	Sensitivity (%)	Specificity (%)
Custom CNN	4.2	91.2	89.4	0.94	89.1	92.6
VGG-16	138.4	92.6	91.0	0.92	90.8	93.2
ResNet-50	25.6	96.8	95.9	0.97	95.4	97.2
EfficientNet-B0	5.3	95.4	94.2	0.95	94.1	96.0

Table 1. Classification Performance Metrics for All Four Deep Learning Architectures

4. Discussion

The superiority of ResNet-50 over VGG-16 and the custom CNN is consistent with the established advantage of residual connections in enabling gradient flow through deep networks without degradation — a critical property for fine-grained retinal feature extraction at deeper layers. ResNet-50's 50-layer depth, achievable without gradient vanishing through skip

connections, allows it to learn both low-level features (vessel boundaries, haemorrhage edges) and high-level semantic representations (neovascularisation patterns, disc-to-cup ratio) simultaneously.

EfficientNet-B0's near-ResNet-50 performance with only 5.3M parameters (79% parameter reduction) is particularly significant for deployment in resource-constrained hospital computing environments. District hospital server infrastructure in Chhattisgarh and Uttar Pradesh typically operates on Intel Core i7 workstations without dedicated GPU accelerators. Inference time benchmarking on CPU shows EfficientNet-B0 requires 0.38 seconds per image compared to ResNet-50's 0.91 seconds — the former enabling real-time grading during patient consultation, the latter requiring offline batch processing.

The Severe NPDR grade (Grade 3) consistently shows the lowest recall across all architectures, reflecting the inherent ambiguity between Moderate and Severe NPDR that even expert graders resolve with significant inter-rater variability. In clinical practice, a hybrid approach is recommended: algorithmic screening at Grade 0/1 to reduce specialist examination load, with mandatory specialist review for Grade 2 and above. This human-in-the-loop design appropriately places algorithmic automation where it adds most value (high-volume, low-risk screening) while preserving specialist involvement where clinical judgment is irreplaceable.

Study limitations include single-institution dataset concentration at KGMU (60% of images) which may introduce site-specific image quality biases, and exclusion of images with poor disc centration or media opacity below a minimum quality threshold (8.4% of acquired images). External validation on a geographically independent dataset is required before clinical deployment, and prospective clinical trial design is recommended to quantify the impact of algorithmic screening on DR detection rates in population health screening programmes.

5. Conclusions

This study demonstrates that deep convolutional neural network architectures — particularly ResNet-50 — can achieve near-ophthalmologist-level five-class diabetic retinopathy grading accuracy (96.8%) on a clinically representative Indian hospital fundus dataset. Key findings are: transfer learning from ImageNet substantially outperforms training from scratch on a 10,000-image dataset; EfficientNet-B0 offers the best accuracy-to-parameters trade-off for deployment on CPU-only hospital computing; the Severe NPDR grade presents the greatest classification challenge and warrants mandatory specialist review in clinical deployment; and focal loss with class weighting successfully addresses the 8:1 class imbalance without synthetic oversampling. The study provides a reproducible benchmarking framework for DR screening system development in Indian public health settings, and recommends prospective pilot deployment beginning with the studied district hospital ophthalmology units in Lucknow and Raipur under an ethics-approved clinical trial protocol.

References

[1] Abramoff, M. D., Lavin, P. T., Birch, M., Shah, N., & Folk, J. C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy. NPJ Digital Medicine, 1, 39.

[2] Acharya, U. R., et al. (2009). Computer-based detection of diabetes retinopathy stages using digital fundus images. Proceedings of the Institution of Mechanical Engineers, Part H, 223(5), 545–553.

[3] Decenciere, E., et al. (2014). Feedback on a publicly distributed image database: The Messidor database. Image Analysis & Stereology, 33(3), 231–234.

[4] Gulshan, V., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, 316(22), 2402–2410.

[5] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of CVPR, 770–778.

[6] International Diabetes Federation (2021). IDF Diabetes Atlas, 10th Edition. Brussels: IDF.

[7] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25, 1097–1105.

[8] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. Proceedings of ICCV, 2980–2988.

[9] Raman, R., et al. (2019). Fundus photograph-based deep learning algorithms in detecting DR. Eye, 33, 97–109.

[10] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.

[11] Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of ICML, 6105–6114.

[12] Tymchenko, B., Marchenko, P., & Spodarets, D. (2020). Deep learning approach to diabetic retinopathy detection. arXiv:2003.02261.

[13] World Health Organization (2019). World Report on Vision. Geneva: WHO.