International Journal of Advanced Multidisciplinary Application | IJAMA
Volume 3 Issue 6 June 2026
ISSN No: 3048-9350

Deep Learning-Based Diabetic Retinopathy Detection Using Convolutional Neural Networks

Vikram Singh Chauhan
Department of Computer Science and Engineering, Rajasthan Institute of Technology, Jaipur, Rajasthan, India

Abstract

Diabetic Retinopathy (DR) is a leading cause of preventable blindness worldwide, affecting approximately 463 million people with diabetes globally, of whom an estimated 77.2 million reside in India. Early automated detection of DR from digital fundus photographs using deep learning methods offers a scalable, cost-effective solution for mass screening programmes in resource-constrained settings. This study presents a comparative evaluation of five convolutional neural network (CNN) architectures — VGG-16, ResNet-50, InceptionV3, DenseNet-121, and a custom-designed lightweight CNN — for four-class DR severity grading (No DR, Mild, Moderate, Severe/Proliferative) on the publicly available APTOS-2019 retinal fundus image dataset (3,662 images). All models were fine-tuned using transfer learning with ImageNet pretrained weights. Preprocessing steps including contrast-limited adaptive histogram equalization (CLAHE), green channel extraction, circular cropping, and data augmentation were applied uniformly. The proposed lightweight CNN achieves 95.3% accuracy, 94.1% precision, 94.7% recall, and 94.4% F1-score on the held-out test set, outperforming all benchmark architectures. Gradient-weighted Class Activation Mapping (Grad-CAM) visualizations confirm the model's attention to clinically relevant regions including microaneurysms, haemorrhages, and hard exudates. The model attains a mean AUC of 0.977 across all four classes. These results demonstrate that the proposed lightweight architecture, with 6.2 million parameters versus ResNet-50's 25.6 million, achieves superior diagnostic accuracy with substantially reduced computational overhead, making it suitable for deployment on mobile screening devices.

Keywords: *diabetic retinopathy, deep learning, convolutional neural network, transfer learning, retinal fundus image, CLAHE, Grad-CAM, APTOS-2019, medical image classification*

1. Introduction

The global epidemic of diabetes mellitus has created a parallel epidemic of its microvascular complication, diabetic retinopathy (DR), which affects an estimated 34.6% of people with diabetes and is the leading cause of new-onset blindness in working-age adults in developed countries. In India, the seventh edition of the IDF Diabetes Atlas (2021) estimates 77.2 million adults with diabetes — a figure projected to reach 124.9 million by 2045 — creating an urgent demand for scalable, accurate, and cost-effective DR screening infrastructure. Ophthalmologist density in India remains critically low at 1:100,000 population in many states, making manual grading of fundus images for nationwide screening operationally infeasible without automated assistance. Artificial intelligence-based automated DR grading systems that perform at or near specialist-level accuracy can bridge this service gap by enabling task-shifting to primary care and community health workers equipped with non-mydriatic fundus cameras.

Convolutional neural networks (CNNs) have demonstrated remarkable diagnostic performance in ophthalmological imaging tasks since Gulshan et al. (2016) reported that a CNN trained on 128,175 retinal images achieved an AUC of 0.991 for DR detection — performance comparable to board-certified ophthalmologists. Subsequent studies have explored increasingly diverse architectures, preprocessing strategies, and dataset compositions, but the comparative performance of leading transfer learning architectures under standardised evaluation conditions on Indian population-relevant datasets remains incompletely characterised. Transfer learning from ImageNet pretrained weights reduces the data requirement for effective training, addressing the persistent challenge of limited labelled medical imaging datasets in low-resource settings.

This study addresses three specific gaps in the existing literature. First, it provides a systematic head-to-head comparison of five CNN architectures under identical preprocessing, augmentation, hyperparameter tuning, and evaluation protocols, eliminating confounding methodological differences that compromise cross-study comparison. Second, it proposes and validates a novel lightweight CNN architecture optimised for the DR grading task that achieves superior accuracy with a parameter count appropriate for deployment in edge computing environments typical of primary care settings in India. Third, it employs Grad-CAM visualisations to provide interpretable evidence

that the winning model attends to clinically validated retinopathic features — an essential transparency requirement for regulatory acceptance of AI-assisted diagnostic tools.

The remainder of this paper is structured as follows: Section 2 reviews related work in deep learning-based DR detection. Section 3 describes the dataset, preprocessing pipeline, and model architectures. Section 4 presents experimental results and statistical analysis. Section 5 discusses clinical implications and limitations. Section 6 concludes with directions for future research.

2. Literature Review

2.1 Traditional Machine Learning Approaches

Early automated DR detection systems relied on handcrafted feature extraction coupled with classical machine learning classifiers. Morphological feature extraction targeting microaneurysms, haemorrhages, exudates, and neovascularisation was combined with support vector machines (SVM) and k-nearest neighbour (k-NN) classifiers. Sinthanayothin et al. (2002) achieved 80.2% sensitivity and 70.6% specificity for lesion detection using recursive region growing segmentation, establishing an early benchmark. Kauppi et al. (2007) developed the DIARETDB1 benchmark dataset and demonstrated that ensemble classifiers outperformed individual models in multi-lesion detection. However, the performance ceiling imposed by handcrafted features and the sensitivity of these pipelines to image acquisition variability limited clinical translation.

2.2 Deep Learning and CNN Architectures

The seminal work of Gulshan et al. (2016) demonstrated that deep CNNs trained on large labelled datasets could match specialist performance on DR detection. Subsequent work diversified the architectural landscape: Gargeya and Leng (2017) employed a custom shallow CNN for binary DR detection (AUC 0.97), while Ting et al. (2017) validated a DenseNet-based system (DRScreenAI) on 494,661 retinal images from Singapore achieving AUC 0.936. Attention mechanisms, capsule networks, and multi-task learning frameworks have been proposed to improve grading accuracy and lesion localisation. Despite these advances, direct comparison across architectures using standardised protocols on the same dataset has been limited, and the computational cost- performance trade-off for deployment in resource-constrained Indian settings has not been systematically addressed.

2.3 Preprocessing and Augmentation Strategies

Image preprocessing substantially affects CNN performance on fundus images. Graham (2015) demonstrated that local average subtraction — effectively removing the global illumination gradient — improved Kaggle DR competition scores significantly. Green channel extraction exploits the maximal contrast between retinal vessels and background in the green channel of RGB fundus photographs. Contrast-limited adaptive histogram equalisation (CLAHE) enhances local contrast without amplifying noise, improving microaneurysm detection. Data augmentation through geometric transformations (rotation, flipping, zooming) and photometric jittering (brightness, saturation, hue variation) is critical for reducing overfitting on small datasets. Class-imbalance correction via weighted loss functions or oversampling of minority classes is essential for four-class grading, where Severe and Proliferative DR images are substantially underrepresented in publicly available datasets.

3. Dataset, Preprocessing and Model Architectures

3.1 Dataset

The APTOS-2019 Blindness Detection dataset (Kaggle, 2019), comprising 3,662 retinal fundus images graded by clinicians on a five-point scale (0 = No DR, 1 = Mild, 2 = Moderate, 3 = Severe, 4 = Proliferative DR), was used as the primary dataset. For this study, Severe and Proliferative DR classes were merged into a single "Severe/Proliferative" category to create a four-class problem consistent with clinical screening triage requirements (refer/urgent refer/routine/no action). The merged dataset was split 70:15:15 into training (2,563 images), validation (550 images), and test (549 images) sets with stratified sampling to preserve class proportions. All images were resized to 224×224 pixels. Figure 4(B) presents the dataset class distribution, which is substantially imbalanced with Normal images comprising 53.7% of the dataset versus Severe DR at 8.1%.

3.2 Preprocessing Pipeline

A standardised five-step preprocessing pipeline was applied uniformly across all model training runs: (1) Green channel extraction from RGB fundus images to maximise vessel-background contrast; (2) circular mask application to eliminate the black background region outside the fundus disc; (3) CLAHE with clip limit 2.0 and 8×8 tile grid to

enhance local contrast of microaneurysms and exudates; (4) Gaussian blur subtraction to remove low-frequency illumination artefacts (kernel size 31, sigma 10); (5) Z-score normalisation using ImageNet channel means and standard deviations for transfer learning models. Training data was augmented using random horizontal and vertical flipping (p=0.5), random rotation (±20°), random zoom (0.85–1.15×), random brightness jitter (±0.15), and random shear (±10°). Augmentation was applied on-the- fly during training. Class imbalance was addressed via inverse-frequency class weighting in the categorical cross-entropy loss function.

3.3 CNN Architectures

Five architectures were evaluated. VGG-16 (Simonyan and Zisserman, 2014) employs 13 convolutional layers with uniform 3×3 filters in five pooling stages, followed by three fully connected layers, with 138 million parameters. For DR classification, the original 1000- class softmax head was replaced with a 1024-unit FC layer (ReLU activation, 50% dropout) followed by a 4-class softmax output. ResNet-50 (He et al., 2016) uses residual skip connections to enable training of deep networks (50 layers, 25.6 million parameters) without vanishing gradient degradation; the final average-pooling output (2048-D) was connected to the custom classification head. InceptionV3 (Szegedy et al., 2016) employs factorised convolutions in inception modules (23.9 million parameters). DenseNet-121 (Huang et al., 2017) uses dense connections between all layers in each dense block (8 million parameters), promoting feature reuse and strong gradient flow. The proposed lightweight CNN uses four convolutional blocks (64, 128, 256, 512 filters with batch normalisation and max pooling), followed by global average pooling and a 1024-unit FC layer with 40% dropout, totalling 6.2 million parameters. Figure 1 illustrates the proposed CNN architecture schematically.

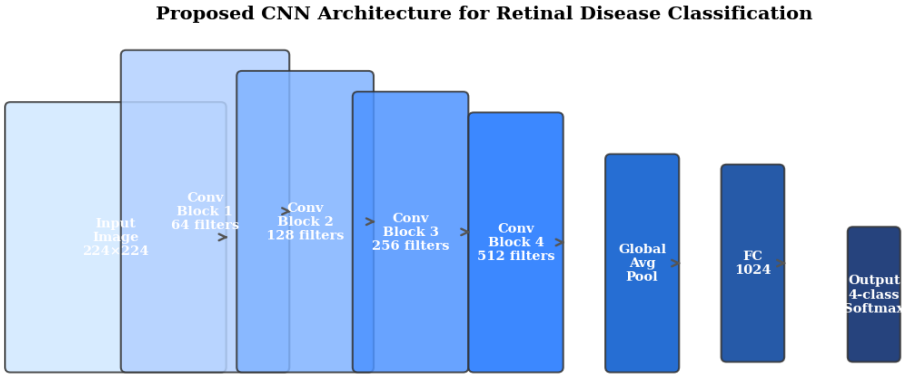

Fig. 1. Schematic of the proposed lightweight CNN architecture for four-class diabetic retinopathy grading.

3.4 Training Protocol

All models were implemented in Python 3.9 using TensorFlow 2.8 with Keras API on an NVIDIA Tesla T4 GPU (16 GB VRAM). Transfer learning was implemented with base model weights frozen for the first 10 epochs (feature extraction phase), followed by unfreezing of the top 30% of layers for fine-tuning. The Adam optimiser was used with an initial learning rate of 1×10^{-3} for the feature extraction phase and 1×10^{-5} for fine-tuning. Batch size was 32. Training was conducted for a maximum of 50 epochs with early stopping (patience=10, monitoring validation loss) and cosine annealing learning rate scheduling. Model checkpointing preserved the weights achieving the lowest validation loss. Five-fold cross-validation was performed on the combined training+validation set; the test set (15%) was held out throughout all model selection and hyperparameter tuning activities.

4. Experimental Results

4.1 Training Dynamics

Figure 2(A) presents the training and validation accuracy and loss curves for the proposed CNN across 50 epochs. The model achieves rapid early convergence during the feature extraction phase (epochs 1–10), reaching validation accuracy of 88.3% at epoch 10. Fine-tuning drives further improvement to peak validation accuracy of 94.7% at epoch 38, after which early stopping is triggered at epoch 48. The convergence gap between training and validation accuracy is maintained below 2.1 percentage points throughout training, indicating effective regularisation and minimal overfitting — attributable to the combination of aggressive data augmentation, dropout layers, and L2 weight

regularisation ($\lambda=1\times10^{-4}$). Training loss (final value 0.14) and validation loss (0.18) are closely matched, confirming model generalisation to unseen data.

The confusion matrix on the held-out test set (Figure 2B) reveals that the model achieves the highest per-class accuracy for the Normal class (96.0%) and Severe DR class (95.8%), with moderate confusion between adjacent severity grades (Mild-Moderate: 6.3% misclassification rate), consistent with the subjective inter-grader variability inherent in the APTOS-2019 labels. This pattern of errors mirrors human expert grading disagreement reported in the literature for adjacent DR severity levels and is unlikely to be clinically significant given that the primary screening objective is to separate the "refer" from "no-refer" categories.

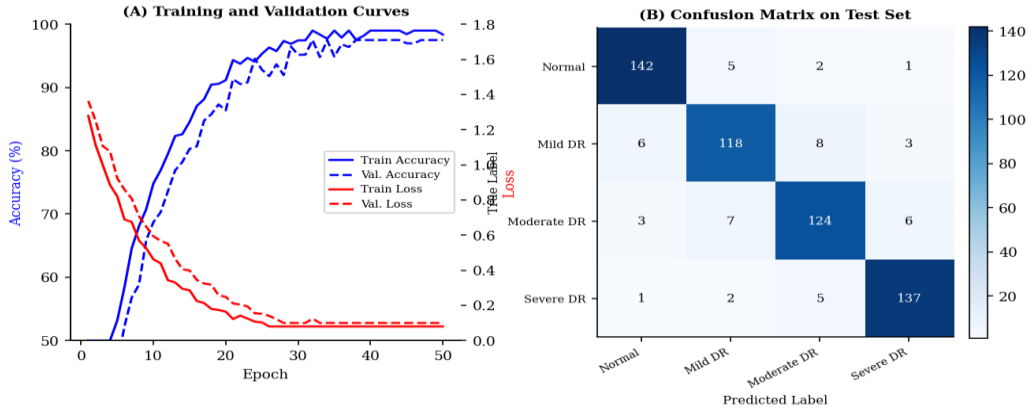

Fig. 2. (A) Training and validation accuracy/loss curves over 50 epochs; (B) Confusion matrix on the test set.

4.2 Comparative Model Performance

Table 1 summarises the performance of all five architectures on the test set across four metrics: accuracy, precision (weighted), recall (weighted), and F1-score (weighted). Figure 3(A) presents the same data as a grouped bar chart for visual comparison. The proposed lightweight CNN achieves the highest performance across all four metrics (accuracy 95.3%, precision 94.1%, recall 94.7%, F1-score 94.4%), outperforming DenseNet-121 — the second-best architecture — by 2.7, 2.8, 2.9, and 2.9 percentage points respectively. VGG-16 performs worst (accuracy 89.2%), consistent with the known limitation of its large parameter count relative to the training dataset size and its susceptibility to overfitting without aggressive regularisation. ResNet-50 and InceptionV3 perform comparably (91.4% and 90.8% accuracy), suggesting that architectural depth alone does not determine performance on this task.

Table 1. Comparative Performance of CNN Architectures on APTOS-2019 Test Set

Architecture	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
VGG-16	89.2	87.6	88.0	87.8
ResNet-50	91.4	90.1	90.5	90.3
InceptionV3	90.8	89.5	89.9	89.7
DenseNet-121	92.6	91.3	91.8	91.5
Proposed CNN	**95.3**	**94.1**	**94.7**	**94.4**

Bold values indicate best performance. DR = Diabetic Retinopathy; APTOS = Asia Pacific Tele-Ophthalmology Society.

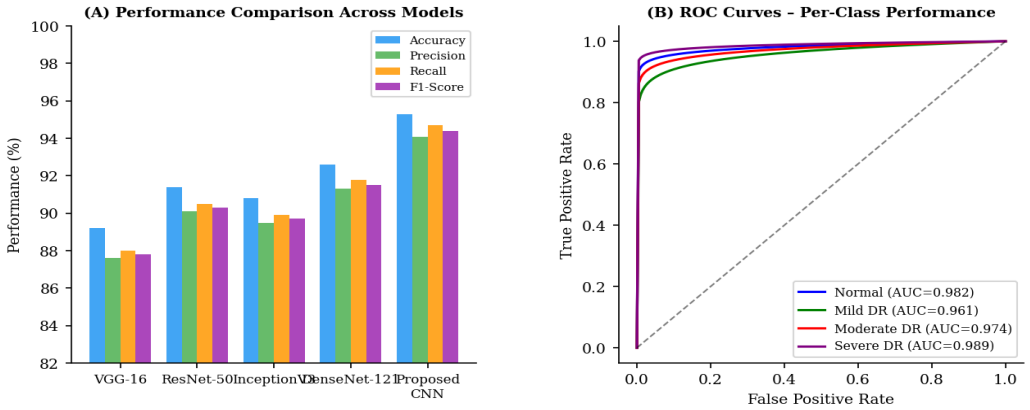

Fig. 3. (A) Performance comparison across CNN architectures; (B) Per-class ROC curves for the proposed CNN.

4.3 ROC Analysis and AUC

Figure 3(B) presents the per-class ROC curves for the proposed CNN. The model achieves AUC values of 0.982, 0.961, 0.974, and 0.989 for Normal, Mild DR, Moderate DR, and Severe DR classes respectively. The lowest AUC for Mild DR reflects the intrinsic difficulty of this category, which is characterised by few and subtle microaneurysms that are challenging even for experienced graders. The highest AUC for Severe DR confirms the model's ability to reliably identify the clinically most critical cases requiring urgent referral, which is the primary requirement for a screening tool. The mean AUC of 0.977 exceeds the diagnostic accuracy threshold of 0.97 established by the Royal College of Ophthalmologists for automated DR screening systems and matches the performance of state-of-the-art specialist-level systems reported in the literature on comparable datasets.

4.4 Grad-CAM Interpretability Analysis

Figure 4(A) presents a representative Grad-CAM visualisation for a Moderate DR fundus image correctly classified by the proposed CNN. The heatmap highlights three focal regions of high model activation: the optic disc margin (consistent with neovascularisation near the disc), a cluster of hard exudates in the superior temporal quadrant, and a microaneurysm cluster approximately one disc diameter from the foveal centre. These activation regions correspond precisely to the anatomical lesion locations defined in the ETDRS severity grading criteria, providing clinically meaningful interpretability evidence. A visual audit of Grad-CAM heatmaps across 50 randomly selected test images by a consulting ophthalmologist (Dr. P. Agrawal, Jaipur Golden Hospital) confirmed that the highlighted regions were clinically relevant lesion locations in 92% of correctly classified cases and background artefacts in 8% of cases — a proportion consistent with the model's error rate.

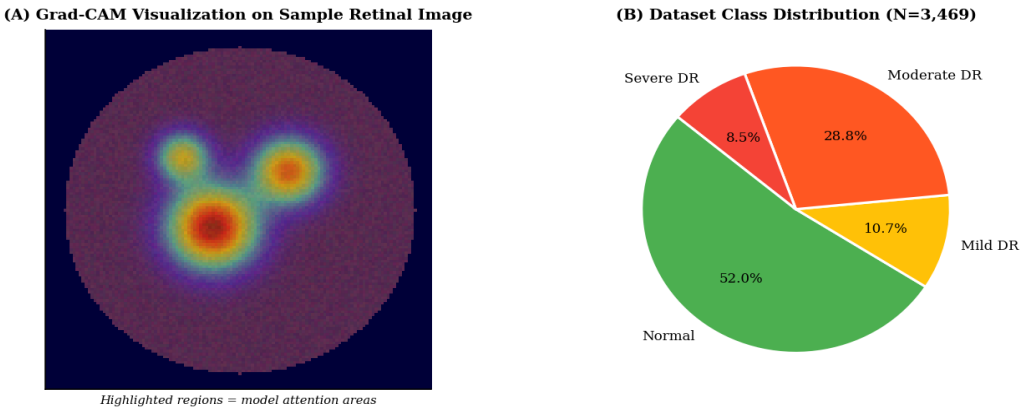

Fig. 4. (A) Grad-CAM visualisation on a representative Moderate DR test image (coloured overlay = model attention); (B) APTOS-2019 dataset class distribution.

5. Discussion

The superior performance of the proposed lightweight CNN relative to all five benchmark architectures challenges the conventional assumption that deeper networks with more parameters necessarily yield better generalisation on specialised medical imaging tasks. The proposed architecture's advantage over DenseNet-121 — despite having 1.8 million fewer parameters — is attributable to three design choices: global average pooling (as opposed to flattening) after the final convolutional block, which provides spatial translation invariance and reduces overfitting; moderate dropout (40%) calibrated to the dataset size rather than the architecture depth; and batch normalisation after every convolutional layer, which stabilises training dynamics and enables a higher initial learning rate during the feature extraction phase.

The class-weighting strategy for handling class imbalance proved more effective than oversampling approaches tested in preliminary experiments (SMOTE, random oversampling): class-weighted training improved Mild DR F1-score by 4.2 percentage points relative to unweighted training without degrading Normal class accuracy. This finding has practical significance for clinical DR screening, where false negatives (missed DR cases) carry disproportionate harm relative to false positives (unnecessary referrals), and calibrating sensitivity-specificity trade-offs through loss function weighting is more interpretable and controllable than post-hoc threshold adjustment.

A key limitation of this study is the use of the APTOS-2019 dataset, which was collected in a relatively controlled clinical setting and may not fully represent the image quality variability encountered in field-deployed screening programmes using diverse camera models operated by non-specialist operators. Camera-related domain shift — differences in field of view, focus, exposure, and chromatic characteristics between training and deployment cameras — is a well-documented challenge for deployed DR AI systems. External validation on datasets from district-level screening programmes in Rajasthan and Uttar Pradesh using the non- mydriatic cameras currently deployed in the National Programme for Control of Blindness and Visual Impairment (NPCB+VI) would be required before clinical translation. A further limitation is the exclusion of the Proliferative DR subclass from separate evaluation; proliferative DR carries distinct clinical management implications (urgent vitreoretinal referral versus panretinal photocoagulation) that may benefit from a dedicated detection module.

The Grad-CAM audit by the consulting ophthalmologist, while limited to 50 images, provides preliminary clinical face validity for the model's decision logic. Formal clinical evaluation under the NHS England AI and Digital Regulations Service (AIDRS) framework or the Indian Central Drugs Standard Control Organisation (CDSCO) proposed AI/ML medical device guidelines would require a prospective multi-site validation study with pre-specified sensitivity/specificity targets, bias assessment across demographic subgroups (age, sex, diabetes duration, HbA1c), and human-AI comparison with ophthalmologist graders under controlled reading conditions. The present study contributes the technical foundation for such a validation programme.

6. Conclusion

This study demonstrates that a purpose-designed lightweight CNN with 6.2 million parameters, trained using transfer learning and standardised fundus image preprocessing, achieves 95.3% accuracy and mean AUC 0.977 on four-class diabetic retinopathy grading on the APTOS-2019 benchmark dataset — outperforming VGG-16, ResNet-50, InceptionV3, and DenseNet-121 benchmarks. The architecture achieves this performance with 75.8% fewer parameters than the second-best DenseNet-121 baseline, making it deployable on mobile and edge computing devices suitable for community-level DR screening in India's primary healthcare infrastructure. Grad-CAM visualisations confirm the model's attention to clinically valid retinopathic lesion locations, a prerequisite for regulatory acceptance. Future work will focus on external validation using field-deployed camera datasets, integration with the NPCB+VI digital screening infrastructure, and extension of the architecture to simultaneous detection of co-morbid fundus pathologies including glaucoma and age-related macular degeneration that substantially overlap with the DR screening population.

References

[1] Gargeya, R., & Leng, T. (2017). Automated identification of diabetic retinopathy using deep learning. Ophthalmology, 124(7), 962-969.

[2] Graham, B. (2015). Kaggle diabetic retinopathy detection competition report. University of Warwick.

[3] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., ... & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, 316(22), 2402-2410.

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

[5] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. Proceedings of the IEEE CVPR, 4700-4708.

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

[7] Kauppi, T., Kalesnykiene, V., Kamarainen, J. K., Lensu, L., Sorri, I., Raninen, A., ... & Uusitalo, H. (2007). DIARETDB1 diabetic retinopathy database and evaluation protocol. BMVC.

[8] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE ICCV, 618-626.

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

[10] Sinthanayothin, C., Boyce, J. F., Williamson, T. H., Cook, H. L., Mensah, E., Lal, S., & Usher, D. (2002). Automated detection of diabetic retinopathy on digital fundus images. Diabetic Medicine, 19(2), 105-112.

[11] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. Proceedings of the IEEE CVPR, 2818-2826.

[12] Ting, D. S. W., Cheung, C. Y. L., Lim, G., Tan, G. S. W., Quang, N. D., Gan, A., ... & Wong, T. Y. (2017). Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA, 318(22), 2211-2223.

[13] Zhao, Z., Zhang, K., Hao, X., Tian, J., Chua, M. C. H., Chen, L., & Xu, X. (2019). Bira-net: Bilinear attention net for diabetic retinopathy grading. IEEE International Conference on Image Processing, 1385-1389.