EPiC Series in Computing

Volume 101, 2024, Pages 73–82

Proceedings of the 16th International Conference on Bioinformatics and Computational Biology (BICOB-2024)

AI-Driven Diagnostics in Ophthalmology: Tailored Deep Learning Models for Diabetic Retinopathy with XAI Insights

Krishna Mridha[1], Ming Wang[1], and Lijun Zhang[1]

Dept. of Population & Quantitative Health Sciences
Case Western Reserve University, Cleveland, OH 44106.
kxm828@case.edu, mxw827@case.edu, lxz759@case.edu

Abstract

Diabetes Retinopathy, a leading cause of vision impairment, necessitates early and precise detection. To address this, we developed a Convolutional Neural Network (CNN) model and tuned three popular pre-trained models, namely VGG16, Xception, and MobileNetV2, to suit the specific characteristics of our dataset. To better understand the functioning of these deep learning algorithms, Explainable AI (XAI) techniques, such as CAM and Grad CAM++, were employed to highlight the crucial features influencing the model's classifications. This study extends to the realm of imaging analysis, emphasizing the critical importance of carefully selecting and customizing models to ensure precise and dependable diagnosis of complex conditions such as DR. Notably, the VGG16 model exhibited strong performance in identifying cases categorized as 'Moderate' and 'No_DR', achieving accuracies of 0.90 and 0.98, respectively. Similarly, both Xception and MobileNetV2 demonstrated promising results in the DR categories. Remarkably, our custom CNN model, tailored for our dataset, achieved an accuracy of 0.986 in identifying cases without DR ('No_DR'). These results underscore the effectiveness of the trained deep learning models in accurately diagnosing DR.

Keywords: Diabetic Retinopathy, Deep Learning, Convolutional Neural Networks (CNN), Explainable AI (XAI), Grad-CAM, Grad-CAM++, Fundus Imaging.

1 Introduction

Diabetic retinopathy (DR), a complication of diabetes that can lead to blindness, inflicts damage upon the retina by affecting the blood vessels within the tissue, causing fluid leakage and distorting vision. According to statistics from the US, UK, and Singapore (NCHS, 2019; NCBI, 2018; SNEC, 2019), [1,2], DR stands among the eye conditions associated with blindness, alongside cataracts and glaucoma. The global number of patients with retinopathy is projected to rise from 382 million to 592 million by 2025 [3]. A survey conducted in the province of Khyber Pakhtunkhwa (KPK), Pakistan, found that 30% of individuals with diabetes are affected by DR, with 5.6% experiencing blindness. If not controlled during the mild stages, nonproliferative

H. Al-Mubaid, T. Aldwairi and O. Eulenstein (eds.), BICOB-2024 (EPiC Series in Computing, vol. 101), pp. 73–82

diabetic retinopathy (NPDR) can progress to proliferative retinopathy (PDR). Another survey in Sindh, Pakistan, involving 130 patients displaying DR symptoms, revealed that 23.85% of the observed patients had DR, out of which 25.8% were diagnosed with PDR. During the initial phases of DR, patients typically do not experience any symptoms; however, as it advances, they may develop floaters, blurry vision, distortions, and a gradual loss of understanding. Additionally, different stages of DR possess unique characteristics and properties, posing a challenge for doctors to consider them all comprehensively and potentially leading to misdiagnoses [4]. This challenge has prompted the development of an automated solution for detecting DR. Studies indicate that with appropriate treatment and careful eye monitoring, up to 56% of DR cases could be preventable [5]. However, the early stages of DR often lack symptoms, making early detection challenging. Conversely, doctors easily reach a consensus when lesions are visible [6]. Furthermore, the current diagnostic methods are inefficient due to their time-consuming nature and dependence on multiple ophthalmologists addressing patients' issues, contributing to diagnostic discrepancies, and creating an unstable foundation for automated solutions to aid research efforts. Therefore, it becomes imperative to leverage computer vision methods for the automated analysis of fundus images, aiding physicians and radiologists in their diagnostic processes.

Computer vision methods for DR detection can be broadly categorized into hands-on engineering [7,8] and end-to-end learning [9,10]. Hands-on engineering methods rely on techniques such as HoG [11], SIFT [12], LBP [13], etc., to extract features. However, these approaches encounter challenges in capturing scale, rotation, and illumination variations. On the other hand, end-to-end learning automatically learns hidden features, leading to improved classification performance. Recent studies have explored various augmentation techniques to enhance training datasets, addressing overfitting issues at the expense of increased computational resources. Transfer learning, particularly with architectures like AlexNet and Inception trained on large datasets such as ImageNet, has set state-of-the-art results for tasks with limited training samples [14]. Noteworthy studies include Gao et al. [15], who analyzed 15,599 FFA images from 1,558 eyes of 845 patients using LeNet-5, VGG16, and ResNet18 models, reporting accuracy metrics ranging between 63.67% and 88.88%. In another study [16], the APTOS dataset was examined using the ResNet-50 model, yielding an accuracy of 0.8010, an AUC of 0.86, and an F1 score of 0.6477. A separate study [17] involving 664 patients examined 5992 B-scans from 1201 eyes with CNN models, achieving an accuracy of 88.3%, specificity of 90.0%, and sensitivity of 82.9%. In yet another study [18], the FGADR and IDRiD datasets were investigated using the ViT model, with key performance metrics including an F1-score of 0.825, accuracy of 0.825, AUC of 0.826, and precision and recall of 0.964. Detecting the stage is crucial for intervention in this life-threatening condition.

In this study, we focus on detecting all stages of DR, including the stage using end-to-end ensemble networks. We propose a machine learning model with Explainable AI and Self-Attention Techniques for DR stage detection. Our results demonstrate that our proposed approach surpasses state-of-the-art methods. Our code can be accessed at: Github Code.

2 Materials and Methods

2.1 Dataset

In this study, we utilized a dataset from Kaggle [19], encompassing a diverse array of retinal photographs, as illustrated in Figure 1. The dataset comprises a total of 3,662 samples distributed among five classes. Its value lies in its varied sources, including clinics and the utilization of var-

ious types of cameras. This diversity introduces complexity to training and validating models, making it both challenging and enlightening for investigations and comparisons in the field.

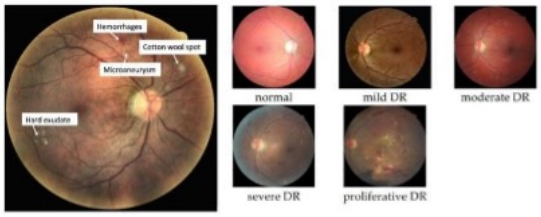


Figure 1: Different DR lesions (left) and DR stages (right). Note: Normal means No_DR.

The dataset categorizes the severity of DR on a scale ranging from 0 to 4, where 0 represents a healthy ('No_DR') state and 4 indicates the presence of DR. Notably, the inclusion of images with varying levels of quality, some containing imperfections or being out of focus, introduces variability. This variability serves as a robust test for the resilience of classifier models and necessitates validation to ensure accurate and reliable results. To comprehend the structure of the dataset and guide model training and evaluation, we divided it into two sets: a training set comprising 2929 samples and a test set consisting of 733 samples, as depicted in Figure 2.

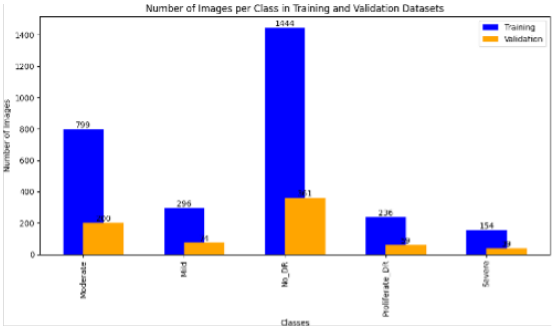


Figure 2: The DR dataset Distribution according to the training and validation samples.

2.2 Data Preprocessing

In the study, we carefully preprocess the data by combining normalization and data augmentation techniques to prepare the dataset for training machine learning models. One crucial step is resizing each image to 224x224 pixels, ensuring that all photos have a size and maintaining consistency across the dataset. To increase the diversity of the dataset and improve the model's ability to generalize, we also incorporate flips as a data augmentation strategy, which helps prevent overfitting. After resizing and augmentation, we apply normalization to scale the values between 0 and 1 by dividing them by 255. Additionally, we refine these values by subtracting RGB values and dividing them by deviation values derived from the ImageNet dataset. This step is crucial for adapting values to optimize network processing. To handle dataset processing, we utilize the TensorFlow function. This function does not extract labels from the directory structure automatically. It also processes images in RGB color mode while setting a batch size of 32, determining how many samples are processed before updating the models' parameters.

Shuffling the dataset is essential during training to maintain randomness among elements while keeping validation and testing phases consistent. We use a fixed seed value throughout to ensure reproducibility and consistent results across runs. To complete the pipeline, we apply the preprocess image function to each item in the dataset. This function takes care of resizing, augmentation, and normalization, ensuring that every image is processed uniformly. As a result, the dataset is well- prepared for model training and improved network performance. This meticulous preprocessing approach plays a role in optimizing the training process and producing model outcomes.

2.3 Feature Extraction and Model Training

In this study, we proposed a DR Detection model with Explainable AI and Self-Attention Techniques, as shown in Figure 3.

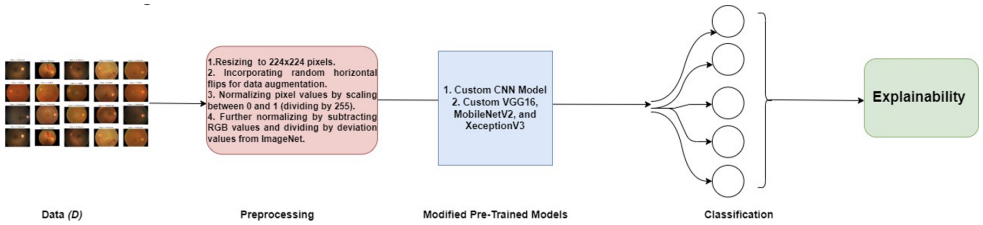


Figure 3: Proposed Model where we have done some preprocessing steps and then gone through the models to extract the features. Then, the model classifies the disease class, and finally, the models visualize the output class using an explainable technique.

Algorithm 1 is designed to classify the DR disease, where we have trained four models. Out of four, three are pre-trained models, and one is a custom CNN model. The models extract the features using convolutional layers, the first layer of every model. This layer extracts features from the input image. Early layers detect simple features (edges, textures), and deeper layers identify more complex features.

$$F_{ij} = Activation\left(\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} I_{(i+m)(j+n)}K_{mn} + b\right) \tag{i}$$

where F_{ij} is the output feature map, I is the input image, K is the kernel, $M \times N$ is the size of the kernel, and b is the bias. The activation function is applied elementwise.

$$\hat{x} = \frac{x - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \tag{ii}$$

Where x is the input, μ_B does the mini-batch mean, σ_B^2 is the mini-batch variance, and ϵ is a small number to prevent division by zero.

$$P_{ij} = \max(region\, corresponding\, to\, (i,j)) \tag{iii}$$

where P_{ij} is the output of the pooling operation, and the max function is applied over a specific input region.

$$y = Activation(Wx + b) \tag{iv}$$

W is the weight, x is the input vector, b is the bias vector, y is the output vector, and the activation function is applied elementwise.

$$Softmax(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \tag{v}$$

Algorithm 1 Custom Image Classification Algorithm

Require: Custom Images (X, Y); where $Y = \{y \mid y \in$ {Mild, Moderate, No_DR, Proliferate DR, Severe}$\}$
Ensure: Trained model for classifying image $x \in X$

1. Perform Preprocessing on each image $x \in X$:
- Resize x to a specified dimension, e.g., 256×256.
- Augment x: Apply random transformations such as rotation, scaling, and flipping.
- Normalize each image based on dataset-specific mean and standard deviation.

2: Import pretrained models $H = \{CustomCNN, VGG16, Xception, MobileNetV2\}$**:**
for all model $h \in H$ **do**
    Replace the last fully connected layer with a layer suitable for (5×1) categories.
end for

3. Train each model:
for all model $h \in H$ **do**
    Set initial learning rate $\alpha = 0.01$ (or another suitable value).
    for epochs = 1 to a specified number, e.g., 30 **do**
        for all mini-batch $(X_i, Y_i) \in (X_{\text{train}}, Y_{\text{train}})$ **do**
            Update parameters of $h(\cdot)$ using a specified optimization equation.
            if validation error does not improve for a set number of epochs, e.g., 3 **then**
                Reduce α, where $\alpha = \alpha \times 0.1$.
            end if
        end for
    end for
end for

3 Result and Discussion

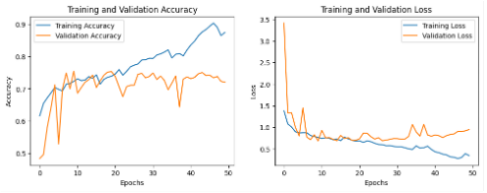
In evaluating DR prediction and detection models, four key performance metrics are commonly used [20]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{vi}$$

$$Recall = \frac{TP}{TP + FN} \tag{vii}$$

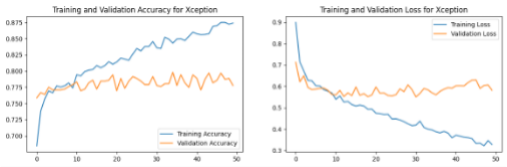
$$Precision = \frac{TP}{FP + TP} \tag{viii}$$

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{ix}$$



(a) Training and Loss curve for Custom Model

(b) Training and Loss curve for VGG16

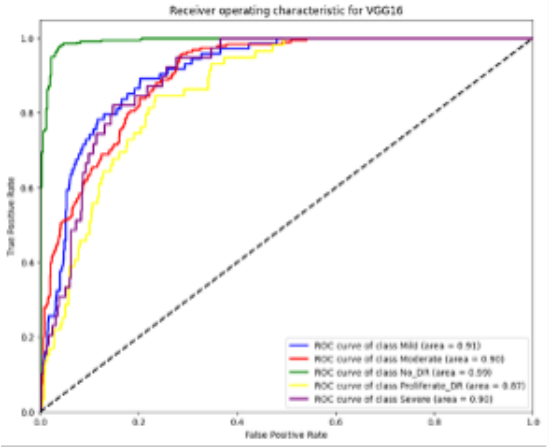


(c) Training and Loss curve for XceptionV3

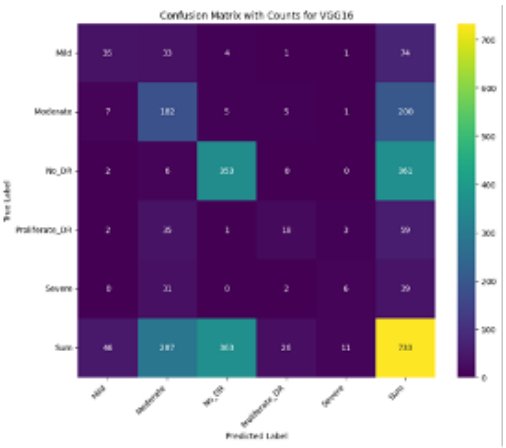
(d) Training and Loss curve for MobileNetV2

Figure 4: Training and Loss curve for all Models Custom Model (a), VGG16 (b), Xception (c), and MobileNetV2 (d)

We have trained our models with 50 epochs. The learning curve of training and validation data is shown in Figure 4. As we can see, the VGG16 model performs better than other models. That's why we are only working with VGG16 now. The results we are showing below are only from VGG16.



(a) AUC-ROC Curve



(b) Confusion Matrix

Figure 5: AUC-ROC Curve (a), Training and Confusion Metrix (b)

Figure 5 showing the AUC-ROC curve for the VGG16 as this model performance better according to the Figure 4 and Table 1 results. The VGG16 models effectiveness, in identifying DR is showcased. On the side there are ROC curves exhibiting AUC values for various DR severity categories indicating the models strong performance especially for the 'No_DR' category (AUC = 0.99). The right side features a confusion matrix that demonstrates how well the model

predicts classes. It is worth noting that while the model accurately predicts cases of 'No_DR' it does exhibit some misclassifications in DR stages like labeling 'Mild' cases, as 'Moderate' on multiple occasions.

Table 1: Accuracy, Precision, Recall, and F1-Score for VGG16

Class	Metric	VGG16	Xception	MobileNetV2	Custom Model
Mild	Accuracy	0.4595	0.5676	0.5405	0.3514
	Precision	0.6800	0.5385	0.5714	0.5532
	Recall	0.4595	0.5676	0.5405	0.3514
	F1 Score	0.5484	0.5526	0.5556	0.4298
	Specificity	0.6800	0.5385	0.5714	0.5532
Moderate	Accuracy	0.9000	0.7000	0.8200	0.6800
	Precision	0.6383	0.6635	0.6805	0.6126
	Recall	0.9000	0.7000	0.8200	0.6800
	F1 Score	0.7469	0.6813	0.7438	0.6446
	Specificity	0.6383	0.6635	0.6805	0.6126
No_DR	Accuracy	0.9806	0.9668	0.9834	0.9862
	Precision	0.9725	0.9614	0.9647	0.8436
	Recall	0.9806	0.9668	0.9834	0.9862
	F1 Score	0.9766	0.9641	0.9739	0.9093
	Specificity	0.9725	0.9614	0.9647	0.8436
Proliferate_DR	Accuracy	0.2881	0.4237	0.3390	0.1864
	Precision	0.6296	0.4098	0.7692	0.4231
	Recall	0.2881	0.4237	0.3390	0.1864
	F1 Score	0.3953	0.4167	0.4706	0.2588
	Specificity	0.6296	0.4098	0.7692	0.4231
Severe	Accuracy	0.1026	0.3333	0.3077	0.1538
	Precision	0.4000	0.6500	0.4286	0.3750
	Recall	0.1026	0.3333	0.3077	0.1538
	F1 Score	0.1633	0.4407	0.3582	0.2182
	Specificity	0.4000	0.6500	0.4286	0.3750

Table 1 provides the performance of four classification models: VGG16, Xception, MobileNetV2, and a Custom Model. They were used to analyze datasets of photographs. These models were assessed across categories such as 'Mild', 'Moderate', 'No_DR', 'Proliferate_DR', and 'Severe'. The evaluation focused on accuracy, precision, recall, F1 score, and specificity. In the 'No_DR' category specifically, all models demonstrated accuracy and precision. This is important for identifying images without retinopathy and shows their ability to distinguish between normal and pathological conditions effectively. However, regarding the Severe categories, the models generally scored lower. This indicates difficulty detecting cases or handling the complexities of advanced stages of diabetic retinopathy. Each model has its strengths and weaknesses. VGG16 performed well regarding recall for the category, which suggests its effectiveness in identifying cases. On the other hand, Xception excelled in precision, which means it made accurate predictions overall. MobileNetV2 precision was particularly notable in the category, which is crucial for reducing false positives in medical diagnostics. The Custom Model had metrics overall. It still provided valuable insights, especially in the 'No_DR' and 'Moderate' categories, which suggests it might offer a different analytical perspective.

The analysis underscores the significance of selecting a model for medical image analysis, highlighting the necessity of choosing a model that matches requirements and dataset characteristics, especially when diagnosing diabetic retinopathy.

In Figure 6, we have shown five samples taken from testing data. The model VGG16 performs four samples correctly out of five. The original class was 'No-DR' but was predicted to be 'Mild'.

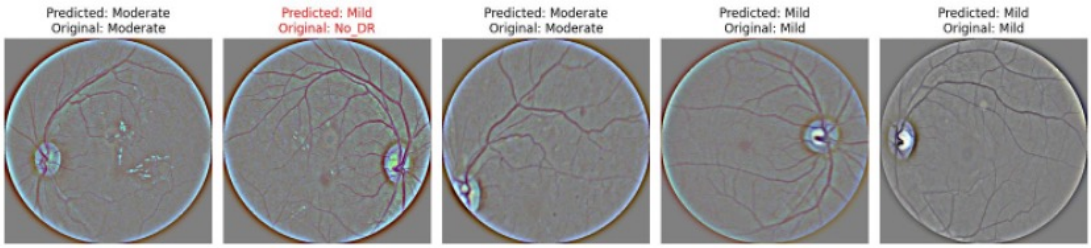


Figure 6: Validation Samples from RetiaVisionNet. Out of five samples, four are correct.

Figure 7 explains three samples using our model through Grad-CAM and Grad-CAM++. This visualization enables humans to trust and comprehend their decision-making processes. Explainable Artificial Intelligence (XAI) aims to develop AI systems that can clarify their decision-making and predictions to ensure accountability and human understanding. In XAI, model explanation techniques such as feature importance and visual explanations are used to achieve this objective.

$$L^{\text{Mild}}_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^{\text{Mild}} A_k^{\text{Mild}}\right) \tag{x}$$

where

$$\text{Grad-CAM}^{\text{Mild}} = \alpha_1 A^1 + \alpha_2 A^2 + \alpha_3 A^3 \quad \text{i.e.,} \quad \sum_k \alpha_k^{\text{Mild}} A_k^k \tag{xi}$$

$$y^{\text{Mild}} = \sum_k \left\{ \sum_a \sum_b \alpha_{ab}^{K,\text{Mild}} . \text{relu}\left(\frac{\delta y^{\text{Mild}}}{\delta A_{ab}^{k,\text{Mild}}}\right) \left[\sum_i \sum_j A_{ij}^k\right] \right\} \tag{xii}$$

4 Conclusion

The study marks a significant advancement in using AI for DR diagnosis. It demonstrates that customized machine learning models, coupled with XAI techniques like Grad CAM and Grad CAM++, can significantly enhance accuracy and reliability in DR detection. Models like VGG16 and Xception showed commendable performance, with the custom CNN model excelling in identifying 'No_DR' cases. This underscores the value of tailoring models to specific needs and the crucial role of XAI in making AI decisions understandable in medical diagnostics. The research contributes vitally to imaging analysis in ophthalmology, emphasizing the importance of model selection, adaptation, and interpretability in AI-driven diagnostics. It paves the way for future advancements in AI applications in healthcare, offering a route to more efficient, accurate, and understandable diagnostic tools.

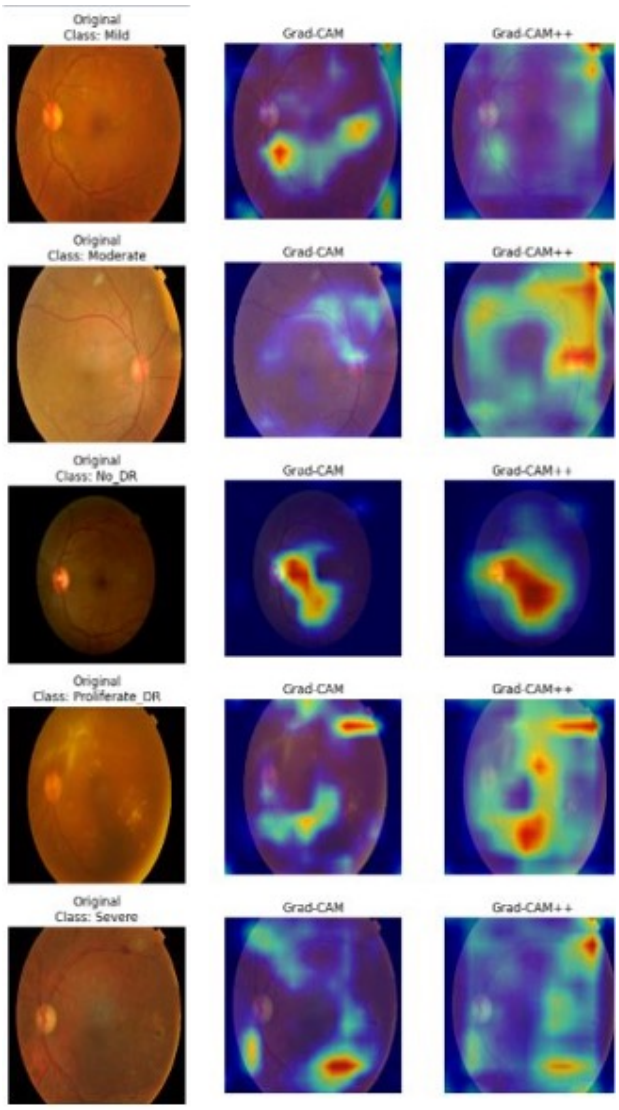


Figure 7: Explain the ability output of the three samples using Grad-CAM and Grad-CAM++.

References

[1] Diabetes Control and Complications Trial Research Group, Nathan, D.M., Genuth, S., Lachin, J., Cleary, P., Crofford, O., Davis, M., Rand, L., Siebert, C. (1993). *The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus.* N Engl J Med, 329(14), 977–986. Massachusetts Medical Society.

[2] UK Prospective Diabetes Study (UKPDS) Group. (1998). *Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33).* Lancet, 352(9131), 837–853. Elsevier.

[3] Jinfeng, G., Qummar, S., Junming, Z., Ruxian, Y., Khan, F.G. (2020). *Ensemble Framework of Deep CNNs for Diabetic Retinopathy Detection*. Comput Intell Neurosci, 2020, 8864698. DOI: 10.1155/2020/8864698. PMID: 33381160, PMCID: PMC7755466.

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

[5] Padhy, S.K., Takkar, B., Chawla, R., Kumar, A. (2019). *Artificial intelligence in diabetic retinopathy: A natural step to the future*. Indian J Ophthalmol, 67(7), 1004–1009. DOI: 10.4103/ijo.IJO_1989_18. PMID: 31238395, PMCID: PMC6611318.

[6] Maxim, L.D., Niebo, R., Utell, M.J. (2014). *Screening tests: a review with examples*. Inhal Toxicol, 26(13), 811–828. DOI: 10.3109/08958378.2014.955932. PMID: 25264934, PMCID: PMC4389712.

[7] Welikala, R.A., Fraz, M.M., Dehmeshki, J., Hoppe, A., Tah, V., Mann, S., Williamson, T.H., Barman, S.A. (2015). *Genetic algorithm-based feature selection combined with dual classification for the automated detection of proliferative diabetic retinopathy*. Computerized Medical Imaging and Graphics, 43, 64–77.

[8] Amin, J., Sharif, M., Yasmin, M., Ali, H., Fernandes, S.L. (2017). *A method for detecting and classifying diabetic retinopathy using structural predictors of bright lesions*. J. Comput. Sci., 19, 153–164.

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

[10] Quellec, G., Charrière, K., Boudi, Y., Cochener, B., Lamard, M. (2017). *Deep image mining for diabetic retinopathy screening*. Med. Image Anal., 39, 178–193.

[11] Dalal, N., Triggs, B. (2005). *Histograms of oriented gradients for human detection*. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), 1, 886–893.

[12] Lowe, D.G. (1999). *Object recognition from local scale-invariant features*. Proc. ICCV, 2, 1150–1157.

[13] Ahonen, T., Hadid, A., Pietikainen, M. (2006). *Face description with local binary patterns: Application to face recognition*. IEEE Trans. Pattern Anal. Mach. Intell., 28(12), 2037–2041.

[14] Abbas, Q., Albathan, M., Altameem, A., Almakki, R.S., Hussain, A. (2023). *Deep-Ocular: Improved Transfer Learning Architecture Using Self-Attention and Dense Layers for Recognition of Ocular Diseases*. Diagnostics (Basel), 13(20), 3165. PMID: 37891986, PMCID: PMC10605427.

[15] Gao, Z., Pan, X., Shao, J., Jiang, X., Su, Z., Jin, K., Ye, J. (2023). *Automatic interpretation and clinical evaluation for fundus fluorescein angiography images of diabetic retinopathy patients by deep learning*. British Journal of Ophthalmology, 107(12), 1852–1858.

[16] Ohri, K., Kumar, M. (2023). *Supervised fine-tuned approach for automated detection of diabetic retinopathy*. Multimedia Tools and Applications, 1–22.

[17] Singh, R., Singuri, S., Batoki, J., Lin, K., Luo, S., Hatipoglu, D., Anand-Apte, B., Yuan, A. (2023). *Deep Learning Algorithm Detects Presence of Disorganization of Retinal Inner Layers (DRIL)–An Early Imaging Biomarker in Diabetic Retinopathy*. Translational Vision Science & Technology, 12(7), 6–6.

[18] Nazih, W., Aseeri, A.O., Atallah, O.Y., El-Sappagh, S. (2023). *Vision Transformer Model for Predicting the Severity of Diabetic Retinopathy in Fundus Photography-Based Retina Images*. IEEE Access, 11, 117546–117561. DOI: 10.1109/ACCESS.2023.3326528.

[19] Dataset, APTOS 2019 Blindness Detection. Available online: https://www.kaggle.com/c/aptos2019-blindness-detection. Accessed: 2023-05-28.

[20] Mridha, K., Kumbhani, S., Jha, S., Joshi, D., Ghosh, A., Shaw, R.N. (2021). *Deep Learning Algorithms are used to Automatically Detection Invasive Ducal Carcinoma in Whole Slide Images*. 2021 IEEE 6th International Conference on Computing, Communication and Automation (ICCCA), 123–129. DOI: 10.1109/ICCCA52192.2021.9666302.