# Deep Learning in Ophthalmology: A Study on Diabetic Retinopathy Detection

**Matea Tashkovska, 370319  Stefan Krsteski, 370315**

## Abstract

Diabetic retinopathy (DR) ranks among the most common eye conditions among diabetic patients. Timely and accurate detection of DR is crucial for early intervention and prevention of vision loss. Traditional methods for DR diagnosis, based on manual examination of retinal images by ophthalmologists, are time-consuming and subject to variability between different observers. This project aims to address these limitations by developing a deep learning algorithm capable of automated DR grading from retinal fundus images. Our objective is to apply transfer learning to determine the most effective pretrained model to use with DR images. Furthermore, we reproduce BiRA-Net - an architecture that has previously demonstrated success in addressing this problem. Additionally, we propose a Siamese-Like Network architecture which trains on pairs of eyes, based on the understanding that diabetes usually affects both eyes. The model achieves quadratic weighted kappa score of 0.773 on the public and 0.764 on the private test set on the Kaggle challenge.

## 1. Introduction

Deep learning methods, especially convolutional neural networks (CNNs), have proven to be highly effective in the analysis and interpretation of medical imaging. This approach is particularly useful for detecting diabetic retinopathy. Diabetic retinopathy (DR) is an eye disease that can lead to blindness and vision loss in people who have diabetes (National Eye Institute, 2023). In its early stages, DR often shows no symptoms. However, as it progresses, it can result in a gradual decline in visual sharpness, potentially leading to complete blindness. According to the World Health Organization (WHO), 4.8% of the 37 million cases of blindness worldwide are attributed to DR (World Health Organization, 2005). This percentage is constantly rising, underscoring the critical need for timely and accurate diagnosis of DR. Early detection is essential in preventing the progression of the disease and reducing the risk of severe vision loss, making it a significant public health priority.

DR occurs due to changes in the blood vessels of the retina, which is the light-sensitive tissue located at the back of the inner eye. Microaneurysm (MA) is usually the first symptom of DR that leads to blood leakage in the retina. MAs are very small and appear as red dots with sharp margins. When blood vessels are broken by abnormal swelling, it results in hemorrhage (HM), which is similar to MA, but larger and more noticeable in retinal images. Further leakage from capillaries can lead to the formation of exudates (EX), which are accumulations of protein-rich fluid and cellular matter. These EX typically appear as yellowish spots and are found in or near the outer plexiform layer. These markers are shown in Figure 1.

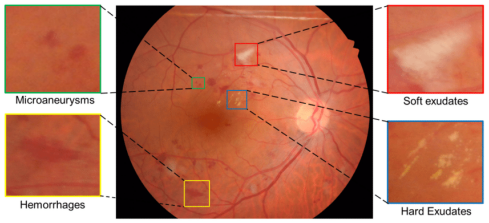


*Figure 1.* Different types of DR lesions.

The traditional approach to diagnosing DR involves a detailed examination of retinal images by ophthalmologists, a process that is not only time-consuming but also subject to variability in interpretation. Developing a deep learning model capable of assisting in the grading of these images could offer significant improvements in the diagnostic process. This could enhance the accuracy and consistency of DR diagnoses and significantly reduce the screening time. This project delves into the specifics of how deep learning models are revolutionizing the detection of DR, exploring the latest advancements and challenges, with the objective to develop a model capable of identifying the level of DR from retinal images.

## 2. Data

The dataset utilized for this project is the EyePACS dataset, sourced from the Diabetic Retinopathy Detection Challenge on Kaggle (Emma Dugas, 2015). It consists of 35.126 high-resolution retinal images, obtained from a group of subjects. A notable feature of dataset is the inclusion of both the left and right eyes for each subject, though its severity can differ between them. The retinal images are gathered from various medical institutions, offering a broad spectrum of patient demographics, disease severities, and image acquisition methods. Each image in the dataset is an RGB fundus photograph, which provides detailed views of the retina and its components. The dimensions of the images differ, but they generally have a resolution of 4000×2000 pixels. Additionally, the images in this dataset vary in quality and contain noise, including artifacts and exposure issues. Experienced ophthalmologists annotated the digital images, supplying reliable labels for the presence and severity of DR. These annotations are divided into five classes - no DR (0), mild (1), moderate (2), severe (3), and proliferative (4) DR. However, the subjective aspect of DR grading, even

among specialists, presents an obstacle in ensuring consistent labeling. Apart from these obstacles, another significant challenge is the major class imbalance, with the "no DR" class representing 70% of the whole dataset.

## 3. Methods and Models

### 3.1. Data Processing

**Resizing and cropping** Our initial data preprocessing involved resizing retinal images to a uniform height of 512 pixels while maintaining their aspect ratios. This was followed by a central crop to extract a 512x512 pixel area. However, this method occasionally led to loss of crucial retinal information in images with larger retinal circles. To address this, we introduce a second approach. We first converted images to grayscale and then generated a binary matrix by using a threshold in order to identify non-black pixels. This helped in differentiating the retinal area from the black space. We then identified the cropping boundaries by locating rows and columns in the binary matrix that had above a certain percentage of non-black pixels. This allowed us to crop out only the retinal regions without loss of retinal information.

**Image resolutions** We experimented with different image resolutions in order to determine the most effective resolution for capturing detailed retinal features. Initially, we explored a lower resolution of 120x120 pixels, which offered the advantage of faster processing times. However, the primary limitation of this lower resolution was the potential loss of critical retinal details which are crucial for accurate classification. In order to find a balance between processing speed and detail retention, we incorporated a 300x300 pixel resolution into our experiments. This offers improved detailed view of retinal features compared to the 120x120 pixel format. Finally, our final experiments focused on using images with a 400x400 resolution, allowing a more detailed examination of the retina and potentially leading to better identification.

**Image normalization** We applied image normalization to ensure that the pixel values of the retinal images were on a consistent scale. This normalization was performed on the whole dataset and was based on the training dataset's mean and standard deviation.

**Data augmentation** With the aim of enhancing our dataset, we implemented data augmentation techniques to increase the robustness of our model. These augmentations included random rotations of up to 10 degrees as well as random horizontal and vertical flips. This way we introduce variability in the dataset while preserving the integrity of the retinal images. Random cropping was avoided as it carried a risk of excluding critical regions of the retina where signs of disease may be present. Similarly, gaussian blur was not employed, as it could potentially alter the appearance of disease indicators, leading to a change in the disease label or making it challenging to differentiate between various classes.

### 3.2. Models

**Transfer Learning** In this research, we have employed various neural network architectures to address the challenge of DR grading. We experimented with transfer learning techniques by using multiple well-known architectures in order to assess the performance of more complex models such as ResNet18 (He et al., 2016), InceptionV3 (Szegedy et al., 2014), EfficientNetB3 (Tan & Le, 2019), and Google's Vision Transformer (Dosovitskiy et al., 2020). These models, along with their pre-trained weights, were taken from the `torchvision.models` subpackage, which contains models for addressing different tasks, including image classification. Transfer learning proves to be beneficial for this task because it allows us to leverage the knowledge gained from large models pre-trained on the ImageNet dataset (Deng et al., 2009), and transfer this knowledge to the DR classification. Initially, we began by unfreezing and training only the final layer of each model. This approach was then extended to fine-tuning the last layers of every model, allowing a more in-depth evaluation of their performance. In addition, for the best performing model that we decided to fine-tune further, we trained the entire network. With this strategy, we intend to identify the best-performing model, which we will later fine-tune and use as a backbone for our more complex architectures.

**BiRA-Net** In our pursuit to develop a network inspired by successful models in DR grading, we focused on leveraging and adapting the BiRA-Net architecture (Zhao et al., 2019), which was specifically designed for this purpose. BiRA-Net combines an attention model for feature extraction and bilinear model for fine-grained classification. In our adaptation, we modified the standard BiRA-Net by replacing its ResNet component with our best-performing model for improved feature extraction.

**Siamese-Like Network** A Siamese-Like Network for classification was designed in order to process and combine information from pairs of images, combining left and right eye information. This method is based on the idea that combining characteristics from both eyes can improve the predictions for DR grading since DR-related alterations frequently occur in both eyes, although with differing degrees. This network has two principal components - the feature extraction branch and the classification head, as shown in Figure 2. The feature extraction is conducted using two branches of the best-performing model. These branches are identical in structure but operate independently, processing the left and right eye images separately. The classification head of our network is a custom-designed neural network built on top the feature extraction branches. The Siamese Network is trained to concatenate the features from both eyes using learnable weights. These weights are crucial as they allow the model to dynamically adjust the influence of each eye's features based on their relevance to DR grading.

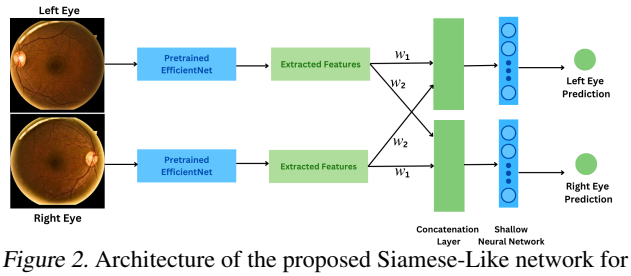


*Figure 2.* Architecture of the proposed Siamese-Like network for DR detection using both eyes.

### 3.3. Experimental setup

The dataset was divided into training and validation set, with 90% for training and 10% for validation. After the split, distribution of different DR classes was similar in both training and validation sets. For testing our model's performance, we used a separate test dataset provided by the competition on Kaggle, split into public and private part.

In addressing the multi-class classification challenge of DR grading, we used Cross-entropy loss. In our Siamese-Like Network design, which processes left and right eye images, we used the average of the loss values from both eyes, ensuring a balanced contribution from each image. We also addressed the issue of unbalanced data by introducing penalty weights in the loss function. These weights are set in reverse proportion to how often each DR grade appears in our training data, helping our model to learn more effectively from rarer cases. The network is trained using the Adam optimizer with an initial learning rate of 3e-5 and a weight decay factor of 1e-5. Each network is trained for 20 epochs with batch size of 32.

### 3.4. Metrics

Our primary metric for evaluating the performance of our models is the Quadratic Weighted Kappa (QWK) which is calculated when quadratic weights are applied to Cohen's kappa (Cohen, 1960). This metric was selected due to its use in the Kaggle competition, which aligns with our testing framework. The QWK is particularly suited for our problem of DR grading as it penalizes predictions based on how far they are from the actual class labels. In other words, the greater the difference between the predicted and the actual class, the more severe the penalization. This property of QWK is crucial in medical data contexts where the proximity of predictions to the true label is also important. This score ranges from -1, indicating total disagreement, to 1, representing complete agreement. Additionally, we utilize Receiver Operating Characteristic (ROC) analysis for each class to provide a more detailed assessment of model performance. This helps in understanding the trade-offs between sensitivity and specificity for each class.

## 4. Results and Analysis

| Model | Only Last Layer | Last Layers |
|---|---|---|
| ResNet18 | 0.381 | 0.507 |
| EfficientNetB3 | 0.350 | 0.592 |
| InceptionV3 | 0.314 | 0.520 |
| Vision Transformer | 0.468 | 0.564 |

*Table 1.* Comparison of QWK on the validation set using transfer learning models by training only the final layer vs. multiple layers

Table 1 shows the performance of the transfer learning models on the validation dataset. All results presented in the table are based on images with a resolution of 300x300 pixels. Initially, training only the last classification layer of each model resulted in low QWK scores. However, by strategically unfreezing and training additional layers, the performance improved significantly. The problem with using large pretrained models on biomedical data, is that they are initially trained on ImageNet, a dataset primarily composed of non-biomedical images. This finding suggests that fine-tuning more layers can be more effective. Among all the models, EfficientNet showed the best results. The Vision Transformer also performed well, but due to its larger size and longer training time, we chose to further refine and work with EfficientNet for this project.

| Step | Experiment | Validation | Public | Private |
|---|---|---|---|---|
| 1 | Central Crop, 120x120 | 0.521 | 0.545 | 0.539 |
| 2 | Improved-Crop, 120x120 | 0.547 | 0.561 | 0.562 |
| 3 | Data Augmentation, 120x120 | 0.589 | 0.587 | 0.580 |
| 4 | High-Res, 300x300 | 0.689 | 0.721 | 0.726 |
| 5 | Higher-Res, 400x400 | 0.696 | 0.740 | 0.733 |

*Table 2.* Comparison of QWK of different image preprocessing and image resolutions on the validation, private and public test set

Table 2 shows the performance of different image processing techniques used while training all layers of the EfficientNet network. The numerous improvements shown in the table were performed in iterative steps, where the model from step $k-1$ served as the pre-trained model for step $k$. Our initial step involved resizing the retinal images to 120x120 pixels in order to minimize computational time, and applying a central crop. However, the results were unsatisfactory, probably due to significant loss of crucial information during the cropping process. Furthermore, while keeping the same resolution and improving the cropping of the retina part of the images, the model's QWK increased. Adding data augmentation to the 120x120 images further improved the scores. This is because of the increased diversity and robustness of the training dataset which allows the model to learn more generalized features. Additionally, increasing the image resolution to 300x300 significantly improved the model's performance, with a QWK score of 0.726 on the private test set. The model performed even better with 400x400 resolution, achieving a QWK score of 0.733. These results indicate that higher resolution images help the model detect DR lesions more accurately.

Once we fine-tuned the model, we advanced to utilizing more complex architectures, such as BiRA-Net. However, this approach did not enhance the outcomes beyond the best results contained in Table 2. It achieved a QWK of 0.700 on the validation set, and scores of 0.739 and 0.740 on the public and private test sets, respectively. We can observe that the results remain unchanged, which might be due to the altercation of the BiRA-Net architecture, suggesting compatibility issues between the two architectures.

| Step | Model | Validation | Public | Private |
|---|---|---|---|---|
| 6 | Siamese Network | 0.746 | 0.773 | 0.764 |
| 7 | Siamese Network, weights | 0.688 | 0.725 | 0.725 |

*Table 3.* Comparison of QWK on Siamese-Like networks with and without weights in the loss function on the validation, private and public test set

The results of employing a Siamese-Like Network are shown in Table 3. By utilizing a dual-input architecture that combines features from both the left and right eye, the Siamese Network achieved the highest QWK score of 0.764 on the private test set. This finding aligns with what was outlined in the Methods section, indicating that the characteristics of one eye can provide valuable information about the level of DR for the other eye.

Furthermore, integrating penalty weights in the loss function to focus on less represented classes led to a more balanced performance across different classes. This approach, while slightly reducing the overall QWK scores, is crucial in medical applications where reducing false negatives is paramount to avoid severe consequences in misdiagnosis. This can also be observed from Figure 3, where we can see that the model using weights in the loss function displays a distribution of predictions that is more concentrated along the diagonal.

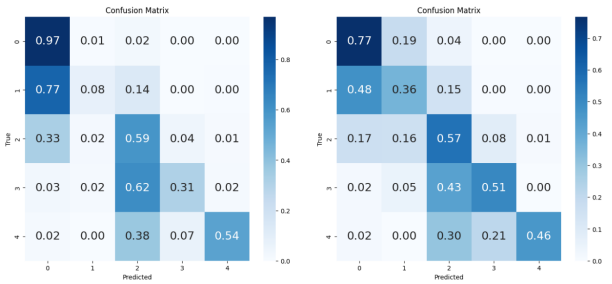


*Figure 3.* Confusion matrices of Siamese-Like networks without (left) and with weights (right) in the loss function for each class - no DR (0), mild (1), moderate (2), severe (3), and proliferative (4)

The ROC curve analysis reveals varying model performance across different DR stages. Class 0 (no DR) shows a high accuracy with an AUC of 0.86, indicating effective identification of non-DR cases. In contrast, the model exhibits challenges in detecting mild DR (class 1) with a lower AUC of 0.61, suggesting that early stages of DR are more difficult to be correctly detected. However, for moderate to proliferative DR (classes 2, 3, and 4), the model demonstrates a significantly improved capability, as evidenced by AUCs of 0.88, 0.96, and 0.95, respectively. This enhanced performance in the higher classes can be attributed to the more pronounced symptoms in later DR stages.

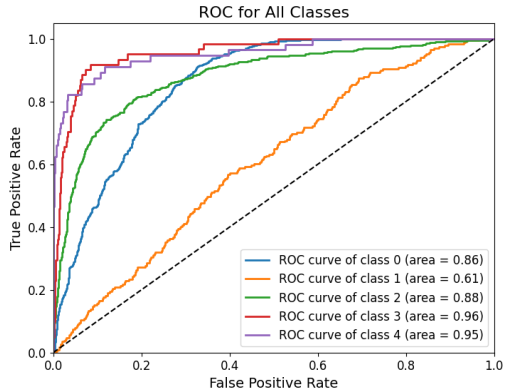


*Figure 4.* Receiver Operating Characteristic for each class - no DR (0), mild (1), moderate (2), severe (3), and proliferative (4)

Figure 5 shows heatmaps of our final model for multiple examples of proliferative DR retinas, using Grad-CAM (Selvaraju et al., 2017) in order to make the model's decisions more interpretable. These heatmaps are visual representations of the areas in the retinal images that the model identifies as significant for diagnosing DR. It's evident that the model accurately identifies the EX, HM, and MA as indicators of DR in the images of the retina. This demonstrates the efficacy of our deep learning approach in detecting signs of DR as well as its high performance in grading DR.

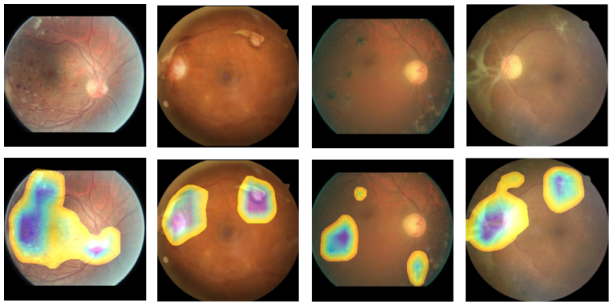

*Figure 5.* Heatmaps of proliferative DR examples illustrated with Grad-CAM.

## 5. Conclusion

Our project on DR grading using deep learning models highlighted several key findings. Using a Siamese-Like Network to analyze images from both eyes proved beneficial. This method capitalized on the bilateral nature of DR, as indicated by the highest QWK score achieved by the Siamese Network. Additionally, we observed that incorporating penalty weights to the loss function reduces false negatives and produces a model that is better at differentiating between the classes. Furthermore, from the evaluation of various pre-trained networks, we concluded that the utilization of EfficientNet yielded the highest level of performance. However, it's worth noting that in scenarios where computational resources are more abundant, exploring the use of a larger and more complex architectures may yield even better performance. For instance, the Vision Transformer showed promise, although its extensive training time presented practical constraints for this project and can be explored further in the future. Moreover the resolution of the images was sufficient for the current analysis, however, it could be up-scaled to capture finer details of the retinal images. It is reasonable to assume that higher resolution images would provide more detailed information, likely leading to improved model performance. A measure that can be experimented in the following work could also include improved loss function, such as focal loss, which applies a modulating term to the cross entropy loss in order to focus learning on hard misclassified examples. Finally, our model can be improved by enriching it with data collected from different available sources, resulting in a larger, more balanced dataset further enhancing its performance and generalization.

## References


Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104. URL https://doi.org/10.1177/001316446002000104.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. URL http://arxiv.org/abs/2010.11929.

Emma Dugas, Jared, J. W. C. Diabetic retinopathy detection, 2015. URL https://kaggle.com/competitions/diabetic-retinopathy-detection.

He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '16, pp. 770–778. IEEE, June 2016. doi: 10.1109/CVPR.2016.90. URL http://ieeexplore.ieee.org/document/7780459.

National Eye Institute. Diabetic retinopathy, 2023. URL https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/diabetic-retinopathy. Accessed: November 22, 2023.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions, 2014. URL http://arxiv.org/abs/1409.4842. cite arxiv:1409.4842.

Tan, M. and Le, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks, 2019. URL http://arxiv.org/abs/1905.11946. cite arxiv:1905.11946Comment: Published in ICML 2019.

World Health Organization. Prevention of blindness from diabetes mellitus: report of a who consultation in geneva, switzerland. https://apps.who.int/iris/bitstream/handle/10665/43576/924154712X_eng.pdf?sequence=1&isAllowed=y, 2005.

Zhao, Z., Zhang, K., Hao, X., Tian, J., Heng Chua, M. C., Chen, L., and Xu, X. Bira-net: Bilinear attention net for diabetic retinopathy grading. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 1385–1389, 2019. doi: 10.1109/ICIP.2019.8803074.


## 6. Appendix

### 6.1. Quadratic Weighted Kappa

This appendix provides a more comprehensive explanation of the QWK, a key metric previously introduced in the metrics section of this paper, to offer a deeper understanding of its significance in our analysis.

To contextualize the use of QWK, it is essential to understand Cohen's kappa (Cohen, 1960). Cohen's kappa is a statistical measure that computes inter-annotator agreement for classification tasks. It is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{1}$$

where $p_o$ is the empirical probability of agreement on the label assigned to any sample, and $p_e$ is the expected agreement when both annotators assign labels randomly. When quadratic weights are applied to Cohen's kappa, it becomes QWK, emphasizing the penalization of predictions that are further from the actual label.

In the context of DR detection, this metric quantifies the agreement between the model's prediction and the diagnoses provided by ophthalmologists, who serve as the trusted source for labeling these images.