Diabetic Retinopathy Detection Using Retinal Images

Elif Karapınar[1], Elif Keleş[2], Zehra Nur Demir[3]

Computer Engineering Department, Gebze Technical University
Kocaeli, Turkey

[1]ekarapinar@gtu.edu.tr
[2]elif.keles2016@gtu.edu.tr
[3]z.demir2024@gtu.edu.tr

Abstract—**Diabetic Retinopathy (DR) is a progressive retinal disease caused by long-term diabetes and represents one of the leading causes of vision decline in the population. Early detection and accurate grading of DR severity are crucial for preventing irreversible vision loss. Traditional screening relies on manual assessments of fundus images by ophthalmologists, which is time-consuming, subjective, and difficult to scale. Recent advances in computer vision and deep learning have enabled automated DR detection systems capable of analyzing retinal fundus images with high accuracy.**

In this study, a deep learning-based approach is proposed for DR classification using preprocessed fundus images and a convolutional neural network trained via transfer learning.

***Keywords*— Computer Vision, Deep Learning, Diabetic Retinopathy, Digital Image Processing, Transfer Learning**

I. Introduction

Diabetic Retinopathy is a common eye disease caused by diabetes and it occurs due to damage to the small blood vessels of the retina. According to the World Health Organization [2] DR is among the leading causes of vision loss globally, particularly in individuals with long-standing diabetes. The disease often progresses silently in its early stages, making regular screening essential for early intervention.

DR is typically classified into multiple severity levels, ranging from No DR to Proliferative DR, based on the presence of retinal lesions such as microaneurysms, haemorrhages, exudates, and neovascularization. [3] Manual grading of retinal fundus images requires specialized expertise. Therefore, the automated image analysis systems gained significant attention to support large-scale screening programs.

Recent developments in deep learning, especially convolutional neural networks (CNNs), have demonstrated remarkable performance in medical image analysis tasks. CNN-based models can automatically learn discriminative features from retinal images and have shown performance comparable to that of expert ophthalmologists in DR detection tasks.

A. *DR Severity Levels*

The severity of Diabetic Retinopathy is commonly categorized into five classes:

0. **No DR** – No visible retinal abnormalities.
1. **Mild DR** – Presence of microaneurysms.
2. **Moderate DR** – Increased vascular abnormalities and haemorrhages.
3. **Severe DR** – Extensive retinal damage and ischemia
4. **Proliferative DR** – Abnormal new blood vessel growth (highest risk of blindness)

These categories form the basis of most publicly available DR datasets and clinical grading systems.

Table 1: DR Severity Levels and Clinical Characteristics

DR Grade	Clinical Findings
No DR	No visible retinal lesions
Mild DR	Microaneurysms
Moderate DR	Hemorrhages, cotton wool spots
Severe DR	Extensive hemorrhages, venous beading
Proliferative DR	Neovascularization, vitreous hemorrhage

Table 2: Dataset Class Distribution

Diagnosis	Class Distribution
0	1434
1	300
2	808
3	154
4	234

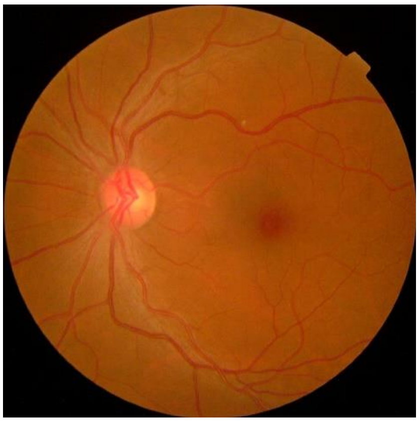

Figure 1: Sample Image for Class 0 in dataset

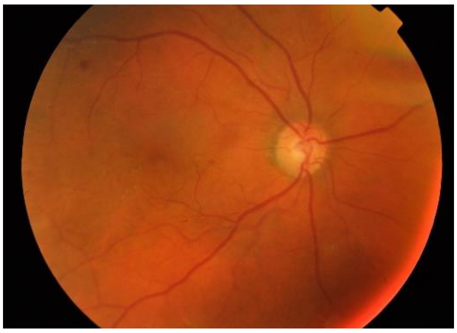

Figure 2: Sample Image for Class 1 in dataset

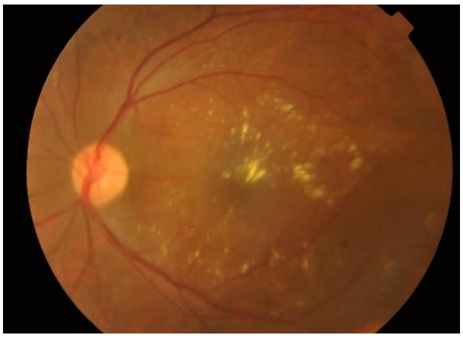

Figure 3: Sample Image for Class 2 in dataset

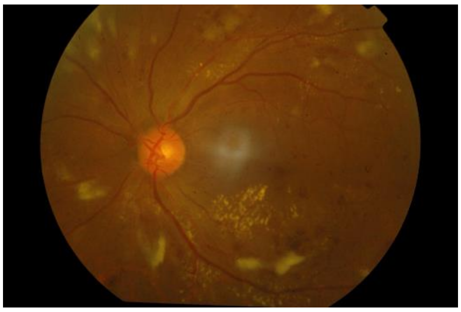

Figure 4: Sample Image for Class 3 in dataset

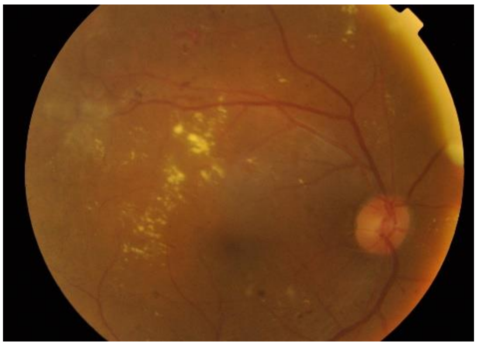

Figure 5: Sample Image for Class 4 in dataset

Figures 1-5 illustrate retinal fundus images corresponding to the five diabetic retinopathy (DR) severity grades, taken from the test dataset [1].

In the No DR category (Figure 1), the retinal vasculature appears normal, with no visible microaneurysms, hemorrhages, or exudates.

Mild DR (Figue 2) is characterized by the presence of small microaneurysms, which are the earliest clinically detectable signs of retinal vascular damage.

As the disease progresses to Moderate DR (Figure 3), additional abnormalities such as intraretinal haemorrhages and cotton wool spots become apparent, indicating increasing vascular leakage and ischemia.

In Severe DR (Figure 4), extensive hemorrhages, venous beading, and intraretinal microvascular abnormalities are observed, reflecting significant retinal ischemia.

The most advanced stage, Proliferative DR (Figure 5), is marked by abnormal neovascularization and fibrous tissue growth, which in fact, increases the risk of vitreous hemorrhage and retinal detachment. These visual patterns show the progressive nature of DR also the need for automated systems to classify interclass differences.

II. Methodology

A. *Dataset and Class Distribution*

This research uses the APTOS 2019 Blindness Detection dataset, which consists of retinal fundus images labeled according to five diabetic retinopathy severity levels: No DR (0), Mild (1), Moderate (2), Severe (3), and Proliferative DR (4). After dataset validation and preprocessing, a total of 2,930 images were used for training and 366 images for validation.

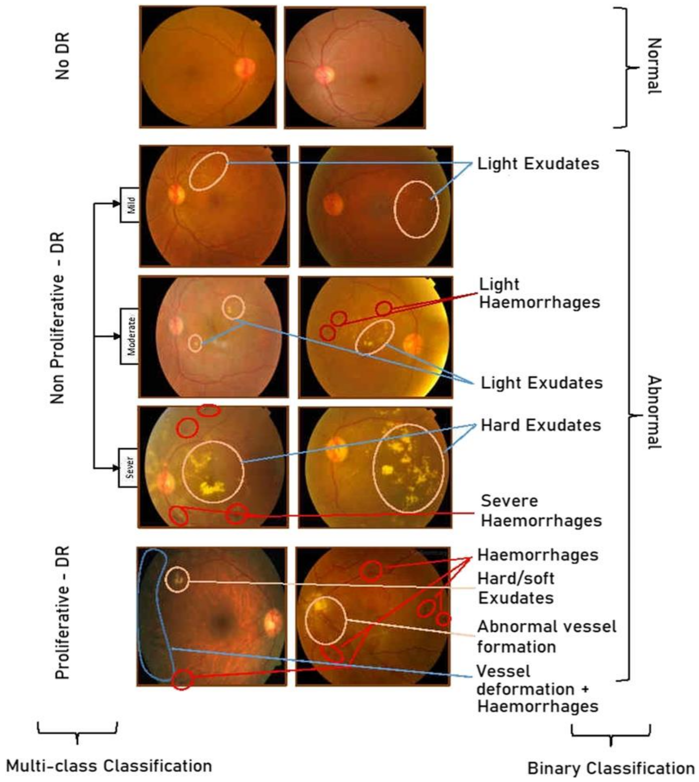

Figure 6: Diabetic retinopathy stages with their symptoms[8]

In this dataset, there is a significant amount of class imbalance, but this is expected for real-world medical image datasets. To solve this problem, class weights were computed using a balanced weighting strategy, ensuring that under-represented severity classes contribute more strongly to the training loss.

B. *Image Preprocessing*

All retinal images were applied to a standardized preprocessing steps designed to enhance diagnostically relevant features while reducing the background noise. First, black borders surrounding the retinal region were removed by threshold-based cropping. After that, images were spatially aligned by centering the retinal region using intensity-based centroid estimation. To enhance vascular structures (the veins on the eyes) and lesion visibility, Contrast Limited Adaptive Histogram Equalization (CLAHE) [7] was applied to the green channel of each image, as this channel provides the highest contrast for retinal blood vessels. Finally, all images were resized to 224×224 pixels to match the input requirements of the deep learning model.

Preprocessed images were saved to disk and reused during training to ensure computational efficiency and reproducibility.

C. *Data Augmentation and Input Pipeline*

To improve the model, on-the-fly data augmentation was applied during training, including random horizontal flipping and small rotational transformations. All images were normalized using the EfficientNet preprocessing function, to ensure consistency with ImageNet pretraining statistics. Separate generators were used for training and validation to prevent data being leaked.

Training was conducted on NVIDIA T4 GPU, providing significant acceleration.

D. *Model Architecture*

The proposed model is based on EfficientNetB0, a convolutional neural network architecture known for its favorable accuracy-to-parameter ratio. The ImageNet-pretrained EfficientNetB0 backbone was used as a feature extractor, followed by a global average pooling layer, a dropout layer with a rate of 0.3 to reduce overfitting, and a fully connected softmax layer producing five output probabilities corresponding to DR severity levels.

E. *Training Strategy*

A two-stage transfer learning strategy was employed:

i. *Warm-up Phase:*
During the initial three epochs, the EfficientNet backbone was frozen, and only the classification head was trained using the Adam optimizer with a learning rate of 1×10^{-3}.

ii. *Fine-tuning Phase:*
In the second stage, the backbone was active, and the entire network was fine-tuned using a reduced learning rate of 1×10^{-4}. Adaptive learning rate scheduling was applied using the ReduceLROnPlateau strategy, and early stopping was employed to prevent overfitting.

Class weights were incorporated into the loss function to reduce the effects of class imbalance.

F. *Evaluation Metrics*

Model performance was evaluated using multiple complementary metrics, including categorical accuracy, top-2 categorical accuracy, and area under the ROC curve (AUC). Additionally, Quadratic Weighted Kappa (QWK)[6] was used as the primary evaluation metric, as it measures agreement between predicted and true DR grades while penalizing larger misclassification errors more heavily. QWK is widely accepted in clinical grading tasks involving ordinal labels.

1. Quadratic Weighted Kappa
   a. **Ordinal Nature of DR Classification**: Diabetic retinopathy severity labels are ordinal, meaning that the classes follow a natural progression from No DR to Proliferative DR. Misclassifying a Mild case as Moderate is clinically less severe than misclassifying it as Proliferative. Standard

accuracy metrics treat all misclassifications equally and therefore fail to capture the clinical impact of prediction errors.

b. **Definition and Properties of QWK**: Quadratic Weighted Kappa (QWK) measures the level of agreement between two raters while accounting for agreement occurring by chance. Unlike unweighted kappa, QWK applies a quadratic penalty to disagreements based on the distance between predicted and true labels, making it particularly suitable for multi-class ordinal classification problems.

c. **Clinical Relevance**: QWK has been widely adopted in medical image analysis tasks where grading severity is involved, including diabetic retinopathy screening challenges. It reflects not only whether predictions are correct but also how incorrect they are, aligning closely with clinical decision-making processes.

d. **Comparison to Accuracy**: While accuracy provides a general measure of correctness, it does not distinguish between minor and severe grading errors. QWK, on the other hand, penalizes large discrepancies more heavily, making it a more informative metric for assessing diagnostic reliability. For this reason, QWK is commonly used as the primary evaluation metric in DR detection benchmarks and competitions.

e. **Literature Support**: Previous studies and public challenges have consistently employed QWK as the standard evaluation metric for diabetic retinopathy classification, further validating its appropriateness for this task.

G. *Challenges And Limitations*

In this study, we faced a few challenges throughout the process of development. Here are some to worth mention:

1. **Class Imbalance:** Since we used a world-dataset, the dataset was highly imbalanced. Class weights were used to reduce this problem, but it cannot be fully removed.
2. **Image Variability**: Fundus images were taken under different conditions, such as different cameras and lighting. In some images, the retina is not centered. Preprocessing steps like cropping, centering, and contrast enhancement help reduce this variability.
3. **Small Visual Differences Between Classes**: The visual differences between neighboring DR stages, especially between Mild and Moderate DR, are very small. This causes most classification errors to occur between adjacent classes.
4. **Single Dataset Usage**: The model was trained and tested only on the APTOS 2019 dataset. Its performance on other datasets was not evaluated, so generalization to different data sources is uncertain.
5. **Limited Image Resolution**: All images were resized to 224×224 pixels. While this helps training efficiency, small lesions may be harder to detect at this resolution.

III. Experimental Results

A. *Training Performance*

During the warm-up phase, the model showed rapid convergence, with validation accuracy improving from 62.0% to 72.7%, and validation QWK increasing from 0.7519 to 0.8257. This shows that the pretrained feature representations were effective for the DR classification task.

Following fine-tuning, further performance gains were observed. The best validation performance was achieved at epoch 5 of the fine-tuning phase, yielding a final QWK score of 0.8444 and a validation accuracy of 75.7%. Early stopping was triggered after performance stabilized, preventing overfitting.

B. *Confusion Matrix Analysis*

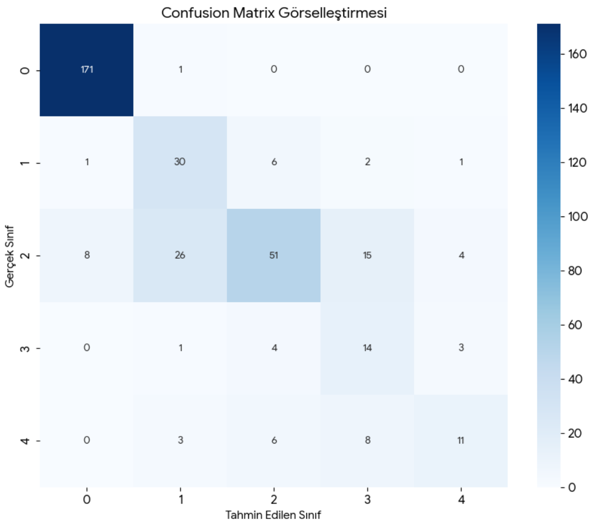

Figure 7: Confusion Matrix

The confusion matrix reveals strong classification performance for the No DR (class 0) category, with 171 out of 172 images correctly classified. Most misclassifications occurred between adjacent severity levels, particularly between Mild and Moderate DR, which is consistent with being the diseases growing states have similar outcomes.

Severe and Proliferative DR classes exhibited higher confusion rates, largely due to their limited representation in the dataset. However, the model demonstrated the ability to correctly identify a substantial portion of advanced cases, which is critical for clinical screening applications.

C. *Prediction Distribution*

The prediction distribution across classes indicates that the model does not collapse into majority-class predictions and can produce outputs across all five severity levels. This behaviour reflects the effectiveness of the applied class-weighting strategy and the ordinal-aware evaluation using QWK.

D. *Quantitative Results Summary*

Table 3: Best Results (Epoch 5, restored with early stopping)

Metric	Value
Validation Accuracy	75.68%
Validation AUC	0.9207
Top 2 Accuracy	91.26%
Final QWK	**0.8444**

E. *Discussion of Results*

The achieved QWK score of 0.8444 demonstrates strong agreement between predicted and true DR severity grades, approaching performance levels reported in prior deep learning-based DR studies. The majority of classification errors occur between neighboring classes, indicating that the model captures the ordinal nature of DR severity progression. These results suggest that the proposed system is well-suited for automated DR screening and severity estimation.

IV. Discussion

The experimental results demonstrate that the proposed deep learning-based system is capable of classifying diabetic retinopathy severity levels from retinal fundus images accurately. The achieved Quadratic Weighted Kappa (QWK) score of 0.8444 shows a high-level agreement between model predictions and expert-provided ground truth labels. These results highlight the effectiveness of the employed preprocessing and training strategies.

Analysis of the confusion matrix reveals that the majority of misclassifications occur between adjacent severity classes. This behavior aligns with clinical observations, as the visual distinction between neighboring stages of DR is often subtle and subject to inter-observer variability, even among experienced ophthalmologists. Importantly, the model indicates strong performance in identifying advanced disease stages, which is critical for referral based screening systems.

The incorporation of class weighting played a significant role in minimizing the effects of class imbalance, preventing the model from over-predicting the majority class. Additionally, the two-stage training strategy allowed the network to benefit from pretrained ImageNet representations while gradually adapting to the features of retinal images. The observed plateau in performance during later fine-tuning epochs suggests that the model reached a stable convergence point without overfitting.

Overall, the results indicate that the proposed approach successfully balances accuracy, robustness, and clinical relevance, making it a viable candidate for automated DR screening applications.

V. Conclusion

In this study, an automated diabetic retinopathy classification system based on deep convolutional neural networks was presented.

A comprehensive pre-processing pipeline was designed to enhance retinal image quality by removing background noise, aligning anatomical structures, and improving contrast in diagnostically significant regions. Using a transfer learning approach with EfficientNetB0, the model was trained and evaluated on the APTOS 2019 dataset across five DR severity levels.

The proposed method achieved a final validation QWK score of 0.8444 and demonstrated strong classification performance across all classes despite significant dataset imbalance. The use of QWK as the primary evaluation metric enabled an ordinal-aware assessment of model predictions, providing a clinically meaningful measure of agreement.

The experimental findings show that deep learning-based approaches, when combined with proper preprocessing stages and evaluation strategies, can effectively support diabetic retinopathy screening and classification. Such systems have the potential to reduce diagnostic workload and improve access to early detection, particularly in large-scale screening programs.

VI. Future Work

Several promising research directions emerge from this study: Future research directions may include the integration of attention mechanisms or lesion-level localization techniques to improve interpretability and clinical trust. Expanding the training dataset with additional fundus images from diverse populations could further enhance generalization.

Moreover, applying classification orders losses or regression-based severity modeling may improve performance in borderline cases.

Finally, prospective clinical validation and deployment-oriented optimization would be necessary steps toward real-world application.

ACKNOWLEDGMENT

We would like to thank or instructor of the Digital Image Processing course at Gebze Technical University, Habil Kalkan, for their guidance throughout the semester. We also acknowledge the APTOS 2019 Blindness Detection competition for providing the retinal fundus dataset used in this research. Also open-source Image Processing community for providing accessible pretrained models which make experimentation and comparison feasible. This project is conducted collaboratively by the authors as part of their graduate coursework.

APPENDIX A

The implementation and the source code used in this study, including preprocessing, model training, and evaluation, is publicly available on GitHub:

- https://github.com/ElifKeles/CSE564-Digital-Image-Processing-Project

REFERENCES

[1] Kaggle, “APTOS 2019 Blindness Detection Dataset,” 2019.

[2] World Health Organization, “Promoting diabetic retinopathy screening,”.

[3] American Academy of Ophthalmology, “Diabetic Retinopathy: Causes, Symptoms, Treatment”, 2025.

[4] V. Gulshan et al., “Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs,” JAMA, vol. 316, no. 22, pp. 2402–2410, 2016.

[5] M. Tan and Q. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” in Proc. ICML, 2019.

[6] J. Cohen, “Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit,” Psychological Bulletin, vol. 70, no. 4, pp. 213–220, 1968.

[7] M. Hayati, K. Muchtar, R. Roslidar, N. Maulina, I. Syamsuddin, G. N. Elwirehardja, and B. Pardamean, “Impact of CLAHE-based image enhancement for diabetic retinopathy classification through deep learning,” Expert Systems with Applications, 2021.

[8] S. Ather at al., “A novel vessel extraction technique for a three-way classification of diabetic retinopathy using cascaded classifier,” Multimedia Tools and Applications, vol. 83, no. 28, pp. 1–21, Feb. 2024, doi: 10.1007/s11042-024-18407-5.

[9] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” in *Int. Conf. Learn. Representations (ICLR)*, 2015.

[10] [5] J. Cuadros and G. Bresnick, “EyePACS: An Adaptable Telemedicine System for Diabetic Retinopathy Screening,” *Journal of Diabetes Science and Technology*, vol. 3, no. 3, pp. 509-516, 2009.