Sharper Detection: Enhanced Techniques for Diabetic Retinopathy Grading

Ali RAFIEI [a,1], Osmar R. ZAIANE [a]
[a] *University of Alberta, Edmonton, Canada*
ORCiD ID: Ali Rafiei https://orcid.org/0009-0009-6571-6951,
Osmar R. Zaiane https://orcid.org/0000-0002-0060-5988

Abstract. Diabetic retinopathy (DR) is a leading cause of blindness, and deep learning methods are extensively used for automated DR grading. This paper demonstrates how applying techniques such as oversampling, data augmentation, K-fold cross-validation, learning rate scheduling, and early stopping can significantly improve performance for models such as ResNet18 and ResNet50. Using the APTOS and EyePACS datasets, our results reveal a substantial accuracy improvement compared to prior methods, which rely on oversampling and data augmentation as baseline techniques. We obtained an accuracy of 97.78% on APTOS with ResNet18 and 93.80% on EyePACS with ResNet50, outperforming the baseline. The techniques presented offer a valuable framework that can complement and enhance the state-of-the-art methods for DR grading and beyond.

Keywords. Diabetic Retinopathy, ResNet, Deep Learning, Model Improvement

1. Introduction

Diabetic retinopathy (DR) is the leading cause of blindness in adults worldwide, with millions of individuals at risk [1]. Current screening processes for early-stage DR are insufficient, often allowing the disease to progress undetected until irreversible damage occurs [1]. Retinal fundus images, a cornerstone of DR diagnosis, are often affected by quality issues such as poor lighting, blurry images, and artifacts. These limitations impact both human diagnostic accuracy and the performance of deep learning-based grading models [2], encouraging the need for advanced training techniques to enhance model robustness and reliability. This paper addresses these challenges by investigating strategies to improve deep learning results.

Fundus images visually capture the retinal structure and serve as a critical diagnostic tool for assessing DR severity. As shown in Figure 1, fundus images can display the presence of key features like microaneurysms, hemorrhages, and exudates, which are indicative of severe non-proliferative DR (NPDR). The ability to accurately detect and grade these features is essential for identifying the progression of DR and implementing timely medical interventions.

Deep learning methods have demonstrated promising results in DR grading [4]; however, challenges such as class imbalance, overfitting, and suboptimal training pro-

[1] Corresponding Author: Ali Rafiei, University of Alberta, Edmonton, Canada; E-mail: rafiei@ualberta.ca.

Table 1. International clinical diabetic retinopathy disease severity scale (ICDRDSS) adapted from [3]

Class Label	Proposed Disease Severity Level	Summary of Findings
0	Normal – No apparent retinopathy	No visible abnormalities in the retina.
1	Mild nonproliferative diabetic retinopathy (NPDR)	Presence of small blood vessel changes, such as microaneurysms.
2	Moderate nonproliferative diabetic retinopathy (NPDR)	Worsening blood vessel changes, including more microaneurysms and mild damage to retinal structures.
3	Severe nonproliferative diabetic retinopathy (NPDR)	Advanced retinal damage, characterized by multiple hemorrhages, venous abnormalities, or irregular blood vessel growth, but no signs of proliferative retinopathy.
4	Proliferative diabetic retinopathy (PDR)	Development of new abnormal blood vessels, often accompanied by bleeding into the vitreous or the retina.

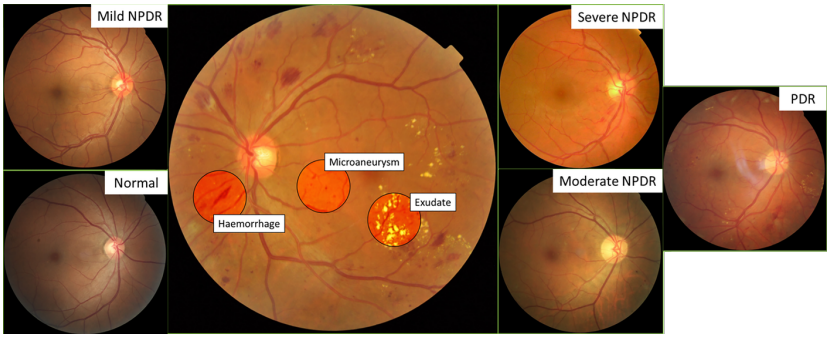


Figure 1. Fundus images illustrating the progression of diabetic retinopathy from normal to proliferative diabetic retinopathy (PDR). The severe non-proliferative diabetic retinopathy (NPDR) image highlights features such as microaneurysms, haemorrhages, and exudates. Adapted from [4].

cesses persist. Previous studies using ResNet50 [5] and ResNet18 [6] have established baselines using oversampling and data augmentation techniques. Building on these baselines, this paper investigates additional strategies to enhance transfer learning effectiveness, including K-fold cross-validation [7], learning rate scheduling [8], and early stopping [9]. These methods aim to address the limitations of previous approaches while providing a generalizable framework for broader applications in medical imaging. By systematically implementing these techniques, we propose a robust approach to improving DR grading outcomes and advancing the field of automated medical diagnostics.

2. Methodology

We fine-tune ResNet18 and ResNet50 [10], pre-trained on ImageNet, using transfer learning for DR grading. ResNet18 is trained on the APTOS 2019 dataset [11], and ResNet50 is trained on the EyePACS dataset [12]. To establish baselines, we reference prior works [5,6] that use oversampling and data augmentation. Our enhanced results build on these by incorporating k-fold cross-validation, learning rate scheduling, and early stopping. We limit our focus to ResNet18 with APTOS and ResNet50 with EyePACS due to the availability of baseline results for these combinations in prior work [4]. This approach ensures direct and fair comparisons, highlighting the performance improvements achieved by our techniques.

DR disease severity is classified into five distinct levels, ranging from no apparent retinopathy to proliferative diabetic retinopathy (PDR). Table 1 outlines these levels along with their observable features on dilated ophthalmoscopy. Accurately distin-

guishing between these stages is crucial for effective diagnosis and treatment planning, making this classification integral to training and evaluating our models.

Both datasets used in this study feature five DR severity levels but present unique challenges due to their variability and complexity. The APTOS 2019 dataset comprises 3,662 retinal fundus images, while the EyePACS dataset contains 35,126 retinal fundus images, both manually labeled [11,12]. To address class imbalances, in our enhanced version, we oversample each dataset such that every class contains the same number of samples as the class with the most images. This ensures balanced representation across all DR severity levels. Furthermore, we replicate an 80:20 train/validation and test split, with 64% of the data used for training, 16% for validation, and 20% for testing, as seen in [5]. These steps are critical to ensure robust and generalizable model performance.

Our proposed techniques build upon established methods [5,6], beginning with oversampling to mitigate class imbalance. To achieve uniform class distribution, we resample the training set by duplicating random samples from underrepresented classes until all classes match the size of the largest. For data augmentation, we apply a consistent sequence of transformations to training images: (1) resize the image, (2) apply random horizontal flips, (3) convert the image to a tensor, and (4) normalize using ImageNet mean and standard deviation. We choose these augmentations to introduce natural variability that reflects real-world fundus images while avoiding artificial distortions.

To further improve model robustness, we incorporate k-fold cross-validation, learning rate scheduling, and early stopping. K-fold cross-validation ensures the model is evaluated across multiple train-validation splits, reducing the likelihood of overfitting to a specific subset. Learning rate scheduling dynamically adjusts the learning rate during training, enabling faster convergence initially and fine-tuning in later stages for improved stability. Early stopping monitors validation loss and terminates training once performance plateaus, preventing overfitting and minimizing unnecessary computation.

We assessed model performance using a classification report that includes precision, recall, and F1-score for each class, along with macro and weighted averages. We also measured inter-rater agreement using Cohen's quadratic weighted kappa, which provides a robust evaluation of model reliability. Additionally, we calculated overall accuracy and per-class accuracy metrics to evaluate performance across all DR severity levels.

3. Results

Our enhanced models consistently outperform the baselines on both datasets, achieving up to a 0.98% to 4.8% relative improvement. Table 2 highlights the improved scores of the models across all key metrics, demonstrating the effectiveness of incorporating k-fold cross-validation, learning rate scheduling, and early stopping. These strategies improve generalization and prevent overfitting during training. Class-level accuracy, presented in Table 3, shows substantial gains, particularly in the more challenging DR severity classes such as Class 4 (Severe NPDR) and Class 0 (PDR), which exhibit significant class imbalance in both datasets, with Class 4 having some of the fewest labeled images and Class 0 the most, making strong performance in these classes particularly noteworthy.

The confusion matrices in Figure 2 illustrate these improvements in detail. On the left, the ResNet18 model achieves similar progress on the APTOS dataset, with near-perfect classification in the no-retinopathy (Class 0) and mild NPDR (Class 1) categories.

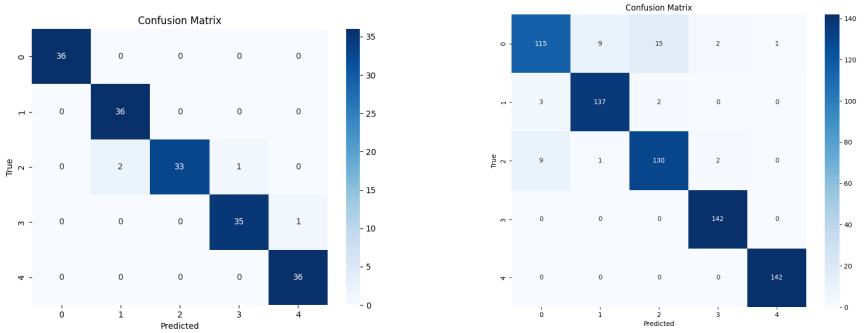
On the right, the enhanced ResNet50 model trained on the EyePACS dataset significantly reduces misclassifications across all severity levels, especially in Classes 1 and 2. These results validate the robustness of our proposed enhancements and highlight their effectiveness in overcoming critical challenges such as class imbalance and overfitting. However, due to the use of random undersampling to ensure fair comparisons between classes and the variability introduced by random seed, the precision observed at 100% in some classes should be interpreted with caution rather than as a definitive measure.

Table 2. Performance Metrics for ResNet18 (APTOS) and ResNet50 (EyePACS)

Model / Method	Dataset	Precision	Recall	F1-Score	Kappa	Overall Accuracy
ResNet18 Baseline [6]	APTOS	93.75%	94.57%	94.16%	-	96.8%
ResNet18 Enhanced (ours)	APTOS	97.85%	97.78%	97.76%	99.45%	**97.78%**
ResNet50 Baseline [5]	EyePACS	68.0%	73.0%	70.4%	62.0%	89.0%
ResNet50 Enhanced (ours)	EyePACS	93.75%	93.80%	93.71%	0.947%	**93.80%**

Table 3. Class-Level Accuracy for our Enhanced Models

Model	Class 0	Class 1	Class 2	Class 3	Class 4
ResNet18 Enhanced (APTOS)	100.0%	100.0%	91.67%	97.22%	100.0%
ResNet50 Enhanced (EyePACS)	80.99%	96.48%	91.55%	100.0%	100.0%



(a) ResNet18 on the APTOS dataset

(b) ResNet50 on the EyePACS dataset.

Figure 2. Balanced confusion matrices showing enhanced model performance. The left matrix corresponds to ResNet18 on the APTOS dataset, and the right matrix corresponds to ResNet50 on the EyePACS dataset.

4. Discussion and Future Work

Our results demonstrate the effectiveness of the proposed techniques—k-fold cross-validation, learning rate scheduling, and early stopping—in improving model performance and providing practical value for medical imaging applications. These techniques address key challenges such as overfitting and class imbalance, leading to more reliable and adaptable models. Their seamless integration with ResNet architectures highlights their versatility and potential for application across diverse use cases.

Future work can focus on applying these techniques to advanced architectures like Vision Transformers and DenseNets. This would offer valuable insights into their ability to address complex challenges in medical imaging. Testing these methods on larger and more diverse datasets could further validate their scalability and robustness. Additionally, incorporating explainability tools, such as saliency maps, could improve model interpretability and build trust among clinicians.

5. Conclusion

This study shows that combining oversampling, data augmentation, k-fold cross-validation, learning rate scheduling, and early stopping can significantly enhance DR grading performance in deep neural networks. By systematically applying these strategies, we achieve meaningful improvements that complement state-of-the-art methodologies and pave the way for better diagnostic automation.

These techniques lay a strong foundation for advancing deep learning in medical imaging. Future research can be encouraged to apply these training techniques to emerging architectures and expand their use to other medical imaging tasks. This would tackle common challenges in medical imaging datasets, providing a scalable and efficient solution to improve diagnostic accuracy and reliability.

References

[1] Kropp M, Golubnitschaja O, Mazurakova A, Koklesova L, Sargheini N, Vo TTKS, de Clerck E, Polivka Jr J, Potuznik P, Polivka J, et al. Diabetic retinopathy as the leading cause of blindness and early predictor of cascading complications—risks and mitigation. Epma Journal. 2023;14(1):21-42.

[2] Laurik-Feuerstein KL, Sapahia R, Cabrera DeBuc D, Somfai GM. The assessment of fundus image quality labeling reliability among graders with different backgrounds. PloS one. 2022;17(7):e0271156.

[3] Wilkinson CP, Ferris III FL, Klein RE, Lee PP, Agardh CD, Davis M, Dills D, Kampik A, Pararajasegaram R, Verdaguer JT, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology. 2003;110(9):1677-82.

[4] Tajudin NMA, Kipli K, Mahmood MH, Lim LT, Awang Mat DA, Sapawi R, Sahari SK, Lias K, Jali SK, Hoque ME. Deep learning in the grading of diabetic retinopathy: A review. IET Computer Vision. 2022;16(8):667-82.

[5] Reguant R, Brunak S, Saha S. Understanding inherent image features in CNN-based assessment of diabetic retinopathy. Scientific Reports. 2021;11(1):9704.

[6] Khalifa NEM, Loey M, Taha MHN, Mohamed HNET. Deep transfer learning models for medical diabetic retinopathy detection. Acta Informatica Medica. 2019;27(5):327.

[7] Geisser S. The predictive sample reuse method with applications. Journal of the American statistical Association. 1975;70(350):320-8.

[8] LeCun Y, Bottou L, Orr GB, Müller KR. Efficient backprop. In: Neural networks: Tricks of the trade. Springer; 2002. p. 9-50.

[9] Caruana R, Lawrence S, Giles C. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. Advances in neural information processing systems. 2000;13.

[10] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition; 2016. p. 770-8.

[11] Karthik, Maggie, Dane S. APTOS 2019 Blindness Detection; 2019. Kaggle. `https://kaggle.com/competitions/aptos2019-blindness-detection`.

[12] Dugas E, Jared, Jorge, Cukierski W. Diabetic Retinopathy Detection; 2015. Kaggle. `https://kaggle.com/competitions/diabetic-retinopathy-detection`.