International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

Transfer Learning for Five-Level Diabetic Retinopathy Severity Classification Using InceptionV3 and ResNet50

Tanuja Gannavarapu[1], Sai Prakash Gopalam[2], Bhavana Reddy Paluvai[3]

[1]Department of Computer Science, Kennesaw State University, Marietta, Georgia, USA
Corresponding Author Email: *tanujag796[at]gmail.com*

[2, 3]Department of Computer Science, Kennesaw State University, Marietta, Georgia, USA

Abstract: ***Diabetic retinopathy (DR) is a progressive retinal disease that can cause visual impairment and blindness. This study presents a five-class DR severity classification approach using retinal fundus images from the APTOS 2019 Blindness Detection dataset. The experimental workflow contains 3,662 labeled images representing five severity grades: no DR, mild, moderate, severe, and proliferative DR. Images are resized and processed using retinal-region cropping and Gaussian-based enhancement before being presented to transfer-learning models at 320×320 resolution. ImageNet-pretrained InceptionV3 and ResNet50 architectures are adapted for five-class classification. On the recorded experimental split, InceptionV3 achieved 78.3% test accuracy with a quadratic weighted Cohen's kappa of 0.822, while ResNet50 achieved 81.9% test accuracy with a kappa of 0.860. ResNet50 therefore produced the strongest overall result. The confusion matrix also reveals important class-specific limitations, particularly for the severe-DR class. The study provides a practical baseline for further improvement through stratified data splitting, training-only augmentation, consistent multiclass loss, class-aware learning, and external validation.***

Keywords: Diabetic retinopathy; fundus images; medical image classification; transfer learning; ResNet50

1. Introduction

Diabetic retinopathy is a diabetes-related retinal disease that can progress to visual impairment and blindness. Retinal fundus photography is commonly used for assessment, and automated image analysis has become an important research area for screening and severity classification. Deep learning methods have shown the ability to learn relevant retinal image features directly from fundus photographs and have been investigated for automated diabetic retinopathy assessment [1], [2].

This work was developed as part of a graduate-level project and was subsequently refined through additional experimentation and analysis. The objective is to classify retinal fundus images into five diabetic retinopathy severity levels rather than treating the problem only as binary disease detection. The study evaluates transfer learning using two established convolutional neural network architectures, InceptionV3 and ResNet50.

The main contributions are: (1) a retinal image preprocessing workflow involving resizing, retinal-region cropping, and Gaussian-based enhancement; (2) adaptation of InceptionV3 and ResNet50 to five-class diabetic retinopathy severity classification; (3) comparison using test accuracy and quadratic weighted Cohen's kappa; and (4) analysis of the ResNet50 confusion matrix to identify class-specific strengths and weaknesses.

2. Literature Survey

Deep learning has become an important approach for automated diabetic retinopathy analysis. Gulshan et al. demonstrated the feasibility of deep learning for detecting diabetic retinopathy from retinal fundus photographs [1]. Other studies have investigated convolutional neural networks for screening and staging, demonstrating the potential of deep learning for severity assessment [2].

Transfer learning is useful when a target medical-image dataset is relatively limited compared with large computer-vision datasets. InceptionV3 uses a multi-branch architecture with factorized convolutions to learn image representations at different scales [3]. ResNet introduced residual learning to enable effective training of deeper neural networks [4]. These pretrained architectures can be adapted to retinal image classification by replacing or extending their final classification layers.

In this study, InceptionV3 and ResNet50 are evaluated on the same five-class retinal-image problem. The comparison is intended to determine which of the two transfer-learning architectures provides stronger overall performance while also examining whether accuracy is consistent across the five severity classes.

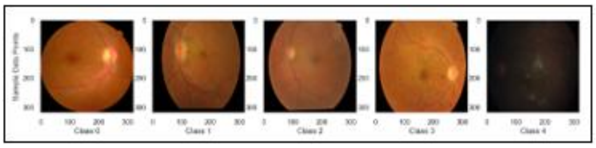

Figure 1: Representative retinal fundus images from the study dataset.

3. Problem Definition

The problem is formulated as supervised five-class classification of retinal fundus images. The target labels are 0 (No DR), 1 (Mild DR), 2 (Moderate DR), 3 (Severe DR), and 4 (Proliferative DR). For an input retinal image, the model produces probabilities for the five classes through a softmax output, and the class with the highest probability is selected as the prediction.

The dataset is imbalanced. The recorded distribution contains 1,805 No-DR images, 370 Mild images, 999 Moderate images, 193 Severe images, and 295 Proliferative-DR images. Therefore, evaluating only overall accuracy can hide poor performance for the less represented severity classes.

4. Methodology/ Approach

A. Dataset: The study uses the APTOS 2019 Blindness Detection dataset [5]. The supplied experimental notebook contains 3,662 labeled retinal fundus images. The recorded transfer-learning experiment uses an 80:20 random train/test split with random_state=51, producing 2,929 training images and 733 test images. The training partition is further divided into 2,344 training images and 585 validation images using a 20% validation split.

B. Preprocessing: Images are initially processed at 512×512 pixels. The preprocessing workflow removes low-intensity background regions, applies a circular mask to retain the retinal field, and performs Gaussian-based enhancement. The resulting images are supplied to the transfer-learning networks at 320×320×3 resolution, with pixel values rescaled to the [0,1] range.

C. Augmentation: The recorded transfer-learning pipeline applies horizontal flipping through the image generator to increase variation in the training data. However, the same generator configuration was also applied to the validation subset in the recorded notebook. This is a methodological inconsistency, since augmentation is intended for the training partition only; validation data should use deterministic preprocessing so that reported validation performance is not affected by random transformations. A corrected experiment should apply stochastic augmentation only to training images and use deterministic preprocessing for validation and test images.

D. InceptionV3: The ImageNet-pretrained InceptionV3 network is adapted to five-class classification. The recorded architecture uses 320×320×3 input, global average pooling, dropout, a 2,048-unit ReLU dense layer, a second dropout layer, and a five-unit softmax output. Training is performed in stages, beginning with most base layers frozen and followed by fine-tuning.

E. ResNet50: The ImageNet-pretrained ResNet50 network is adapted using global average pooling, dropout, a 2,048-unit ReLU dense layer, a second dropout layer, and a five-unit softmax output. The recorded workflow begins with most layers frozen and then performs fine-tuning. Adam is used as the optimizer, with learning-rate changes between the initial training and fine-tuning stages. Early stopping and learning-rate reduction are used to control overfitting and improve optimization. The fine-tuning stage in the recorded notebook uses binary cross-entropy as the loss function, which is a methodological inconsistency for this five-class softmax formulation; categorical cross-entropy is the appropriate multiclass objective and should be used consistently across all training stages in a corrected experiment.

F. Evaluation: Test accuracy is used to measure overall classification performance. Quadratic weighted Cohen's kappa is also reported because the five DR categories are ordered severity levels. The confusion matrix is examined to understand how predictions are distributed across individual severity classes.

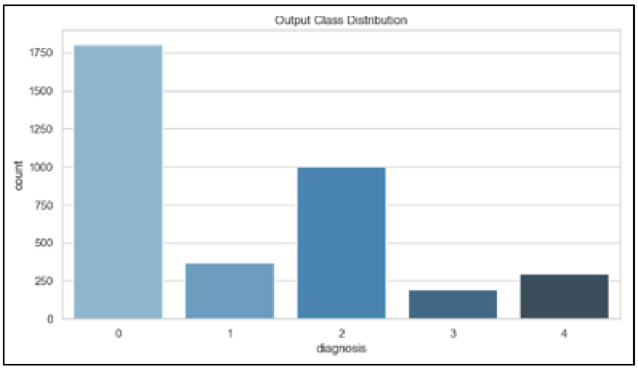

Figure 2: Distribution of the five diabetic retinopathy severity classes

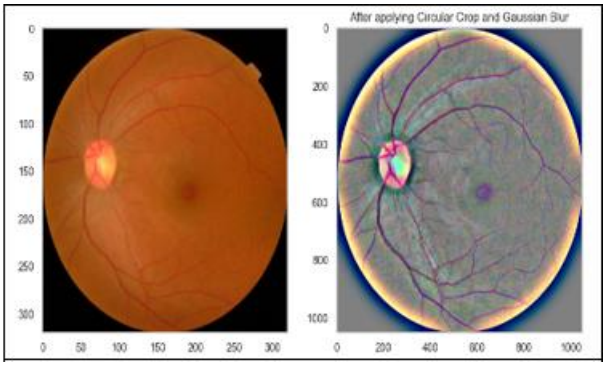

Figure 3: Example of circular retinal cropping and enhancement

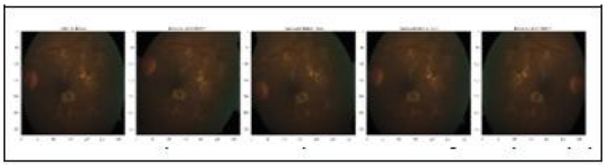

Figure 4: Example augmentation sequence from the original workflow

5. Results and Discussion

The verified transfer-learning results are shown in Table I. InceptionV3 achieved 86.5% training accuracy and 78.3% test accuracy, with a quadratic weighted Cohen's kappa of 0.822. ResNet50 achieved 86.5% training accuracy and 81.9% test accuracy, with a kappa of 0.860. ResNet50 therefore produced the strongest overall result between the two transfer-learning models evaluated.

The recorded ResNet50 confusion matrix provides additional information that is not captured by overall accuracy. The approximate class recalls are 99% for No DR, 51% for Mild, 92% for Moderate, 0% for Severe, and 40%

International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

for Proliferative DR. The result indicates that the model performs well for some classes but has substantial difficulty identifying severe and some minority-class cases. In particular, zero recall for severe DR in the recorded test matrix is an important limitation.

The class imbalance is a likely contributor to this behavior, since the Severe class contains only 193 images compared with 1,805 No-DR images and the recorded experiment does not explicitly use class weighting or a class-aware objective. The validation-augmentation and loss-function inconsistencies noted in the Methods section (IV-C and IV-E) further limit how directly these results can be interpreted and should be corrected in a follow-up experiment.

Table I: Verified transfer-learning results from the supplied notebook

Model	Train Accuracy	Test Accuracy	QWK
InceptionV3	86.5%	78.3%	0.822
ResNet50	86.5%	81.9%	0.860

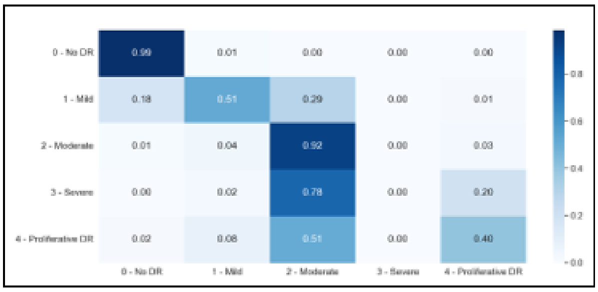

Figure 5: Normalized ResNet50 confusion matrix from the recorded test run

6. Application Prototype

A simple GUI prototype was developed to load the trained ResNet50 model and predict one of the five DR severity classes for an uploaded retinal image. This demonstrates the practical application of the trained model. However, the current GUI preprocessing does not reproduce the complete training preprocessing pipeline. Before presenting the GUI as a deployment-ready application, it should use the same retinal cropping and enhancement steps used during model training. The interface should also display human-readable severity labels and clearly indicate that the output is a research prediction rather than a medical diagnosis.

7. Limitations

The study has several limitations. First, the reported experiment uses a single random 80:20 split rather than repeated stratified evaluation or an independent external test set. Second, the dataset is substantially imbalanced. Third, the recorded validation generator applies horizontal flipping. Fourth, binary cross-entropy is used during the later fine-tuning stage for a five-class softmax task. Fifth, the recorded ResNet50 confusion matrix shows zero recall for the severe-DR class. These limitations mean that the current result should be considered a research baseline rather than a clinically validated diagnostic system.

8. Future Scope

Future work should use a stratified train/validation/test split and apply augmentation only to the training partition. A consistent multiclass loss should be used throughout training. Class weighting, balanced sampling, focal loss, or ordinal-learning approaches should be evaluated to improve performance on minority and severe disease classes. Future evaluations should report per-class precision, recall, F1-score, specificity, balanced accuracy, quadratic weighted kappa, and multiclass ROC-AUC where appropriate.

External validation on an independent retinal dataset is also important to assess generalization across cameras, clinics, and patient populations. The GUI should be updated to reproduce the training preprocessing pipeline exactly and should include a clear research-use disclaimer.

9. Conclusion

This study presents a five-class diabetic retinopathy severity classification approach using retinal fundus images and transfer learning. On the recorded experimental split, InceptionV3 achieved 78.3% test accuracy with a quadratic weighted kappa of 0.822, while ResNet50 achieved 81.9% test accuracy with a kappa of 0.860. ResNet50 therefore provided the strongest overall performance between the two verified transfer-learning experiments.

At the same time, the confusion matrix shows that overall accuracy does not represent uniform performance across all five classes, with zero recall for severe DR in the recorded test matrix. The work therefore provides a useful baseline and identifies clear directions for improvement rather than making a claim of clinical readiness.

References

[1] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, "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, doi: 10.1001/jama.2016.17216.

[2] M. Shaban, Z. Ogur, A. Mahmoud, A. Switala, A. Shalaby, H. Abu Khalifeh, M. Ghazal, L. Fraiwan, G. Giridharan, H. Sandhu, and A. S. El-Baz, "A Convolutional Neural Network for the Screening and Staging of Diabetic Retinopathy," PLOS ONE, vol. 15, no. 6, e0233514, 2020, doi: 10.1371/journal.pone.0233514.

[3] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[5] APTOS 2019 Blindness Detection, Kaggle, 2019. [Online]. Available: https://www.kaggle.com/competitions/aptos2019-blindness-detection