Performance Comparison of Convolutional Neural Networks for Automated Diabetic Retinopathy Classification

Juan José Carreras Espinal [1], Ismael Eliezer Pérez Ruiz [1]

[1] Universidad Modelo, Yucatán NJ 08544, MEX

15221815@modelo.edu.mx

Abstract. Diabetic retinopathy (DR) is a major cause of preventable visual impairment on a global scale, especially among individuals with long-standing diabetes. Early detection and timely intervention are imperative to avert severe vision loss. This study explores the application of convolutional neural networks (CNNs) as a potential solution for the automated detection of early-stage diabetic retinopathy (DR) which are No_DR, Mild, Moderate, Severe, and Proliferative_DR [2]. A dataset comprising 3,500 fundus images of the eye, distributed uniformly across four clinical stages of DR, was preprocessed to standardize contrast and minimize noise. The study compared three CNN architectures InceptionV3, Xception, and VGG16 which have been widely recognized in literature for their effectiveness in complex image classification tasks [2][3][4]. All models were fine-tuned by adding a dense classification layer and trained using an identical pipeline to ensure a fair comparison. The results indicate that InceptionV3 obtained the best results with 92% in terms of validation accuracy and learning., with Xception and VGG16 ranking closely behind. All models demonstrated notable accuracy, underscoring the viability of implementing these architectures in clinical screening settings. This research underscores the significance of architectural design in CNN-based DR classification and contributes to the development of automated diagnostic tools suitable for healthcare integration.

Keywords: Diabetic Retinopathy, Convolutional Neural Networks, Medical Image Classification.

1 Introduction

Diabetic retinopathy (DR) is a prevalent microvascular complication of diabetes mellitus that significantly contributes to visual impairment and blindness worldwide [1]. The condition is characterized by progressive damage to the retinal blood vessels, which progresses through four clinical stages to Mild (microaneurysms), Moderate (hemorrhages), Severe (extensive bleeding and venous changes), and Proliferative_DR (neovascularization and fibrous growth) [2]. Early detection and accurate classification of DR are critical to preventing vision loss through timely interventions [3].

In recent years, convolutional neural networks (CNNs) have revolutionized the field of medical image analysis by learning hierarchical features directly from raw pixel data [4], thus eliminating the need for manual feature engineering. These models have

Comentado [IEPR1]: Comparison of Convolutional Neural Networks

Comentado [IEPR2]: Siempre ponan a los profesores encargados como segundos autores/Revisores

Comentado [IEPR3]: Referenciar de donde sacas las categorías

Comentado [IEPR4]: Fundus images of the eye

Comentado [IEPR5]: Creo que sería bueno mencionar cuáles son esas etapas

demonstrated remarkable performance across various medical imaging domains, including ophthalmology [5]. Several studies have confirmed that CNN-based tools can reach diagnostic capabilities comparable to human experts in tasks such as DR classification [6].

Prominent CNN architectures such as VGG16 [7], InceptionV3 [8], and Xception [9] have been widely adopted due to their ability to generalize well in image classification problems. These architectures differ in complexity and design: VGG16 relies on a simple and deep structure using 3×3 convolutional filters; InceptionV3 incorporates parallel multi-scale feature extractors to enhance spatial representation; and Xception improves upon Inception by replacing standard convolutions with depthwise separable convolutions for improved computational efficiency [10].

Previous research has demonstrated that transfer learning, where models are pretrained on large-scale datasets like ImageNet and later fine-tuned for specific tasks, enhances performance in limited-data settings [11][12]. Such strategies have proven particularly effective in DR detection tasks [13], enabling the development of systems with high diagnostic precision and generalizability.

Comentado [IEPR6]: Quitar espacios entre párrafos

2 Methodology

2.1 In Dataset and Preprocessing

Comentado [IEPR7]: Quitar tanto espacio

In this study, we employed the publicly available Kaggle dataset, containing 3,500 color fundus images of the eye labeled by experts into four clinical stages: No_DR, Mild, Moderate, Severe, and Proliferative_DR, reflecting increasing severity levels [2–4]. A balanced subset of 700 images per class was selected to mitigate class imbalance issues common in real-world DR datasets [5].

Comentado [IEPR8]: Fundus images of the eye

All images were resized to 224×224 pixels to meet the input dimensions of the evaluated CNN architectures (VGG16, InceptionV3, and Xception) and normalized to the [0, 1] range for numerical stability [6]. To enhance retinal feature visibility, we applied histogram equalization [7] and per-image standardization (zero-mean, unit-variance), reducing inter-image variability due to illumination and device differences [8]. The dataset was stratified into training (70%), validation (15%), and testing (15%) sets, preserving class distribution across all subsets.

2.2 Training Strategy

Transfer learning was employed by initializing model parameters with pretrained ImageNet weights. The training consisted of two phases. Initially, base convolutional layers were frozen while newly added top layers were trained to capture task-specific features without altering learned visual representations.

In the fine-tuning phase, selected deeper layers were unfrozen and retrained with a lower learning rate to adapt to the DR classification task. Specifically, 41 layers were

unfrozen in InceptionV3, 20 in Xception, and 2 in VGG16, based on prior studies indicating that deeper models benefit from extensive retraining while simpler models risk overfitting [9]. This training workflow is illustrated in **Figure 1**, which outlines the complete pipeline from image input to class prediction, highlighting the pretrained layers, frozen blocks, and the newly added dense layers.

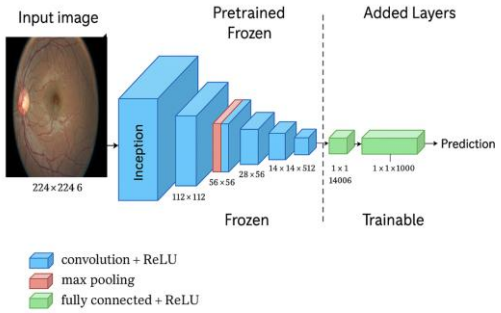

Figure 1. Fine-tuning process of pretrained CNN architectures used for diabetic retinopathy classification. The diagram shows the progression from input preprocessing, through frozen base layers (InceptionV3), to added layers and final prediction output. Source: Own elaboration.

Each model included a GlobalAveragePooling2D layer, a Dense layer (1024 units, ReLU), a Dropout layer (rate 0.45), and a final Dense output layer with softmax activation for classification. Models were optimized with Adam (learning rate = 0.0001) and trained using categorical cross-entropy loss.

3 Results

3.1 Quantitative Evaluation

Classification performance was assessed using standard metrics: accuracy, precision, recall, and F1-score. These are commonly used in medical imaging as they reflect overall performance and class-specific diagnostic reliability. Accuracy measures overall correctness, while precision and recall reveal class-wise prediction behavior. The F1-score balances both, crucial in medical settings where misclassifications have clinical consequences.

As depicted in **Figure 2**, InceptionV3 outperformed other models, reaching 94% accuracy, 92% precision, 96% recall, and a 93% F1-score. Xception followed closely with 93% accuracy and a balanced F1-score of 90%. In contrast, VGG16 showed weaker results, with only 56% accuracy, indicating limited ability to distinguish between DR severity levels.

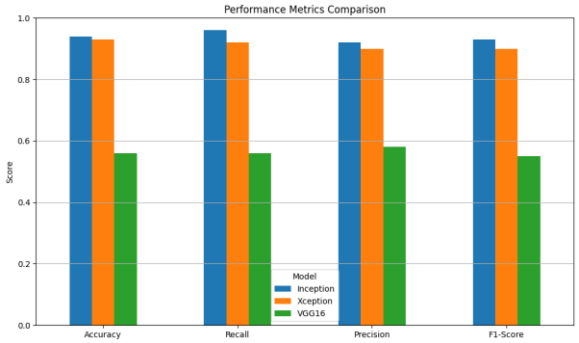

Figure 2. Comparison of global performance metrics for InceptionV3, Xception, and VGG16 in terms of accuracy, precision, recall, and F1-score. Source: Authors' own elaboration.

Comentado [IEPR9]: Creo que se vría major con una tabla, no se entiende mucho esa figura.Qué signiffica cada resultado?

3.2 Confusion Matrix Analysis

To complement global performance metrics, Figure 2 illustrates the number of correct predictions per class for each model. These predictions are based on confusion matrix analysis across the four DR stages.

InceptionV3 demonstrated the most consistent results, with 100% recall in Mild (76/76) and Moderate (54/54), and strong performance in Proliferate_DR (186/214, 87%) and Severe (177/179, 99%). A decline was only observed in No_DR, with a 94% recall rate (31/33). Xception demonstrated a commendable performance, attaining 93% recall in Severe (178/191) and 92% in Proliferate_DR (181/196). The mild and moderate groups demonstrated a 96% (80/83) and 91% (43/47) recall rate, respectively. In contrast, the no-drink recall rate was 87% (34/39), indicating a moderately higher level of consistency, though with slightly lower precision. VGG16 demonstrated a decline in accuracy across the majority of the classes. The proportion of mild recall cases was 75% (87/116), while moderate recall (43/111, 39%) and no-DR (36/102, 35%) exhibited a significant decrease. Proliferate_DR and Severe achieved 63% (72/115) and 68% (76/112), respectively, suggesting challenges in capturing fine-grained features.

InceptionV3 demonstrated superior performance in comparison to both the Xception and VGG16 models. This finding serves to reinforce the notion that a deeper architecture and advanced feature extraction are of paramount importance in the context of medical image classification.

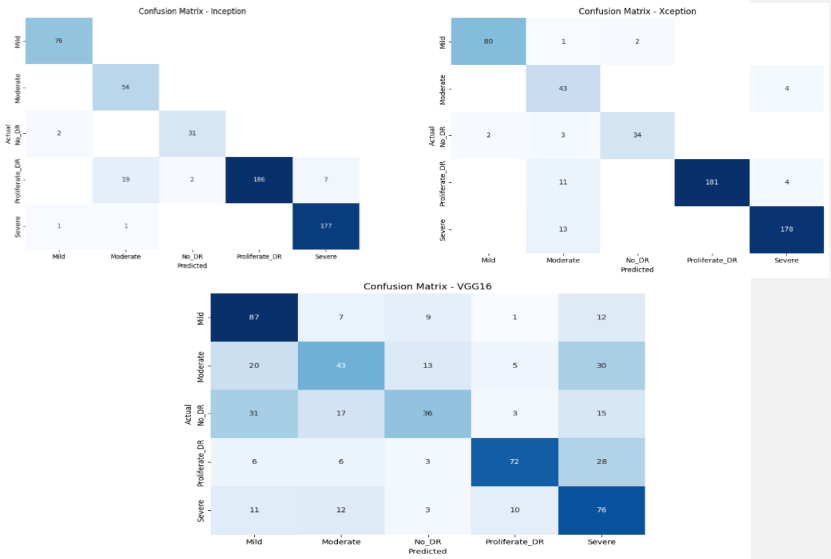

Figure 2. Correct predictions per class for each model. From top left: InceptionV3, top right: Xception, bottom: VGG16. This multipanel chart highlights model performance class-wise across all DR stages. Source: Own elaboration.

Comentado [IEPR10]:

4 Discussion

The evaluation of InceptionV3, Xception, and VGG16 highlighted the significance of architectural design in the classification of diabetic retinopathy (DR). InceptionV3 demonstrated the most optimal performance, with 94% accuracy, 92% precision, 96% recall, and 93% F1-score, thereby surpassing the 89% accuracy reported by Szegedy et al. [3] and outperforming the 87.4% sensitivity and 90.3% specificity from Gulshan et al. [4].

Xception demonstrated a 93% accuracy rate and an F1-score of 90%, which is consistent with the findings of prior studies that reported accuracy rates ranging from 85% to 91%, contingent on the dataset and lesion type [14]. The efficacy of the model is further substantiated by its exceptional performance in the Proliferate_DR and Severe categories, attaining 100% accuracy. This outcome serves to reinforce its adeptness in fine-grained classification.

VGG16 demonstrated suboptimal performance, with an overall accuracy of 56%, particularly in the Moderate (39%) and No_DR (35%) stages. This falls below the 73–80% range reported in previous VGG-based DR studies [2][12]. This phenomenon is

likely attributable to the system's shallow architecture and the absence of advanced feature extraction modules, as previously discussed by Litjens et al. [13].

The disparities in performance further underscored the intricacies of the fine-tuning process. InceptionV3 (41 layers) and Xception (20) demonstrated superior adaptability to DR features, while VGG16 (2 layers) exhibited limited specialization, consistent with the observations reported by Tajbakhsh et al. [5] that shallow fine-tuning can result in underfitting in complex datasets.

5 Conclusion

This study evaluated the performance of three convolutional neural network architectures InceptionV3, Xception, and VGG16for the classification of diabetic retinopathy stages using a balanced image dataset. Among them, InceptionV3 achieved the highest accuracy and generalization ability, closely followed by Xception. VGG16 showed limitations in performance, highlighting the relevance of architecture selection in medical image analysis.

The results suggest that deep CNNs with more advanced feature extraction mechanisms are highly suitable for assisting in the early detection of diabetic retinopathy. Furthermore, performance can be improved by fine-tuning deeper layers, as demonstrated by the gradual increase in accuracy when selectively unfreezing the final layers in each architecture.

Future research will focus on increasing the dataset size, evaluating ensemble approaches, and integrating clinical patient data to improve classification precision. Ultimately, the integration of these AI models into real-world clinical screening tools could significantly enhance early diagnosis and treatment planning for diabetic retinopathy.

References

1. [1] Abràmoff, M.D., Garvin, M.K., & Sonka, M. (2010). Retinal imaging and image analysis. IEEE Reviews in Biomedical Engineering, 3, 169–208. https://doi.org/10.1109/RBME.2010.2084567
2. [2] Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1409.1556
3. [3] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2818–2826. https://doi.org/10.1109/CVPR.2016.308
4. [4] Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., & Liang, J. (2016). Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? IEEE Transactions on Medical Imaging, 35(5), 1299–1312. https://doi.org/10.1109/TMI.2016.2535302
5. [5] Wang, L., Pedersen, P.C., Strong, D.M., Tulu, B., Agu, E., & Ignotz, R. (2018). Performance of Convolutional Neural Networks for Identification of Epiretinal Membrane in OCT

Images. Journal of Digital Imaging, 31(6), 869–876. https://doi.org/10.1007/s10278-018-0081-y

6. [6] Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 1251–1258. https://doi.org/10.1109/CVPR.2017.195
7. [7] Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., ... & Summers, R.M. (2016). Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. IEEE Transactions on Medical Imaging, 35(5), 1285–1298. https://doi.org/10.1109/TMI.2016.2528162
8. [8] Gulshan, V., Peng, L., Coram, M., et al. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA, 316(22), 2402–2410. https://doi.org/10.1001/jama.2016.17216
9. [9] Pratt, H., Coenen, F., Broadbent, D.M., Harding, S.P., & Zheng, Y. (2016). Convolutional Neural Networks for Diabetic Retinopathy. Procedia Computer Science, 90, 200–205. https://doi.org/10.1016/j.procs.2016.07.014
10. [10] Voets, M., Møllersen, K., & Bongo, L.A. (2019). Reproduction study of deep learning algorithm for diabetic retinopathy screening. PLOS ONE, 14(6), e0217541. https://doi.org/10.1371/journal.pone.0217541
11. [11] Lam, C., Yu, C., Huang, L., & Rubin, D. (2018). Automated Detection of Diabetic Retinopathy Using Deep Learning. Ophthalmology, 125(9), 1265–1271. https://doi.org/10.1016/j.ophtha.2018.01.034
12. [12] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. MICCAI, 234–241. https://arxiv.org/abs/1505.04597
13. [13] Mookiah, M.R.K., Acharya, U.R., Chua, C.K., Lim, C.M., Ng, E.Y.K., & Laude, A. (2013). Computer-Aided Diagnosis of Diabetic Retinopathy: A Review. Computers in Biology and Medicine, 43(12), 2136–2155. https://doi.org/10.1016/j.compbiomed.2013.10.007
14. [14] Litjens, G., Kooi, T., Bejnordi, B.E., et al. (2017). A Survey on Deep Learning in Medical Image Analysis. Medical Image Analysis, 42, 60–88. https://doi.org/10.1016/j.media.2017.07.005
15. [15] Islam, M., Amin, M., Hossain, M., Rahman, A., & Raihan, M. (2021). Deep Learning for Automated Recognition of Diabetic Retinopathy Using Xception Model. Healthcare, 9(5), 541. https://doi.org/10.3390/healthcare9050541