Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 20 No. 1 (2024) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v20i01.45247

PAPER

Classification of Diabetic Retinopathy by Deep Learning

Roaa Al-ahmadi[1],
Hatoon Al-ghamdi[1](✉),
Lobna Hsairi[1,2]

[1]University of Jeddah, Jeddah, Saudi Arabia

[2]University of Tunis El Manar, Tunis, Tunisia

2200634@uj.edu.sa

ABSTRACT

Diabetic retinopathy (DR), which is a leading cause of adult blindness, primarily affects individuals with diabetes. The manual diagnosis of DR, with the assistance of an ophthalmologist, has proven to be a time-consuming and challenging process. Late detection of DR is a significant factor contributing to the progression of the disease. To address this issue, the present study utilizes deep learning (DL) and transfer learning algorithms to analyze different stages of DR and precisely detect the condition. Using a large dataset comprising approximately 60,000 images, this study employs ResNet-101, DenseNet121, InceptionResNetV2, and EfficientNetB0 DL models to automatically assess the progression of DR. Images of patients' eyes are inputted into the models, and the DL architectures are adapted to extract relevant features from the eye images. The study's findings demonstrate that DenseNet121 outperforms ResNet-101, InceptionResNetV2, and EfficientNetB0 in accurately classifying the five stages of DR. The accuracy of the models was 97%, 96%, 95%, and 94%, respectively. These results underscore the effectiveness of DL in achieving an accurate and comprehensive classification of retinitis pigmentosa. By enabling accurate and timely diagnosis of DR, the application of DL techniques significantly contributes to the field of ophthalmology, facilitating improved treatment decisions for patients.

KEYWORDS

diabetic retinopathy (DR), deep learning (DL), classification, ResNet101, DenseNet121, EfficientNetB0, InceptionResNetV2

1 INTRODUCTION

Deep learning (DL) shows significant promise in medical image analysis because of its ability to recognize complex patterns within data. Research has demonstrated that DL models can accurately detect diabetic retinopathy (DR), an eye complication of diabetes that can lead to blindness, from retinal images [11]. DR is characterized by retinal changes such as hemorrhages, abnormal blood vessel growth, aneurysms, and cotton wool spots. Early detection and treatment of DR are critical for preserving vision. Thus, developing a DL model that can automatically screen retinal images for DR severity could facilitate timely interventions [5]. Detecting the

Al-ahmadi, R., Al-ghamdi, H., Hsairi, L. (2024). Classification of Diabetic Retinopathy by Deep Learning. *International Journal of Online and Biomedical Engineering (iJOE)*, 20(1), pp. 74–88. https://doi.org/10.3991/ijoe.v20i01.45247

Article submitted 2023-09-14. Revision uploaded 2023-10-22. Final acceptance 2023-10-28.

severity of DR could assist doctors in monitoring the progression of the disease and determining suitable treatment plans to minimize the risk of vision loss. This study aims to develop an accurate DL model to classify the severity of DR in retinal images. DR is classified into five stages, including mild non-proliferative retinopathy (MNR), moderate non-proliferative retinopathy, severe non-proliferative retinopathy (SNR), and proliferative diabetic retinopathy (PDR). MNR is the mildest stage and is identified by the presence of microaneurysms that generally do not cause significant visual impairment. Moderate non-proliferative retinopathy, on the other hand, is characterized by blocked blood vessels that often result in swelling, distortion, and impaired vision. SNR is characterized by persistent blockage of blood vessels, leading to a lack of oxygen supply to the retina and the potential stimulation of abnormal blood vessel formation. PDR is the most advanced stage and is characterized by the growing of abnormal blood vessels on the surface and inside the eye, which leads to severe vision loss. Additionally, diabetic macular edema (DME) is a complication that can occur at any stage of DR, but it is more prevalent in later stages. DME is characterized by fluid leakage in the central retina, which results in blurred or distorted vision. In our study, four state-of-the-art DL models have achieved the highest accuracy in the binary classification of DR. We utilized these four models to evaluate their capabilities for multi-class classification of DR based on its five severity levels. The four models—ResNet-101, DenseNet121, InceptionResNetV2, and EfficientNetB0—have previously shown promising results for binary DR detection tasks. However, classifying DR into its five stages is more challenging due to the subtle variations between stages. The ResNet-101 model achieved an accuracy of 0.82 thanks to its residual connections, which facilitate flow gradients and enable deeper networks for improved accuracy. The DenseNet121 model achieved an accuracy of 0.87 by connecting layers to each other, which strengthens feature propagation and alleviates the vanishing gradient problem. The InceptionResNetV2 model achieved an accuracy of 0.85 by combining the Inception and ResNet architectures to leverage their respective strengths. The EfficientNetB0 model achieved an accuracy of 0.73 while maintaining efficiency through neural architecture search, which optimizes network parameters to strike a balance between accuracy and efficiency. The potential benefits of an accurate DL model for DR classification are significantly high. It enables the automatic screening of retinal images for early detection of DR and timely intervention. It also alleviates the burden on physicians by detecting various levels of DR severity. Furthermore, it could serve as a foundation for applying DL to other tasks in medical image analysis, ultimately improving patient outcomes.

This paper's structure is as follows: Section 2 provides an overview of studies that are relevant to our research. Section 3 describes the dataset, preprocessing techniques, and models used, as well as the training and hyperparameter tuning approaches. Section 4 presents findings, including performance metrics and a comparison of the models. Section 5 tests the models' predictions on retinal scan images. Section 6 summarizes the main findings and proposes potential avenues for future research.

2 LITERATURE REVIEW

Several studies have been conducted to explore the use of DL for DR classification. This research may assist in the selection and utilization of machine learning techniques for detecting and diagnosing diabetic eye disease. Dutta et al. [4] conducted a study to enhance image preprocessing and improve classification accuracy.

They achieved this by utilizing statistical features and fuzzy C-means (FCM) clustering to address the challenges associated with image resizing. They compared three neural network models for their performance and trained both images and statistical data using FCM for class labeling. Moreover, another study by Saxena et al. [13] aims to expedite the initial screening process for DR by utilizing cutting-edge DL models that are based on a convolutional neural network (CNN). The authors utilized a dataset of 56,839 fundus images from EyePACS to train and validate their models. In addition, the paper discusses the challenges of automated disease detection in medical images using CNNs and compares the results with those of other prominent studies. A more recent study by Sungheetha Sharma. [15] proposes a CNN framework for detecting hard exudates in fundus images, aiming to enable early and accurate detection of diabetic conditions and overcome the limitations of traditional methods. Their methodological framework achieves higher accuracy than other detection algorithms and classifies the severity of the diabetic condition using the results of confusion matrix detection. Similarly, Khan et al. [7] aimed to develop and evaluate a CNN and image processing-based approach for the automatic detection and classification of DR in fundus images. The authors implemented several state-of-the-art deep neural network (DNN) architectures, including InceptionV3, VGGNet, and ResNet, using transfer learning. Moreover, Oh et al. [10] investigated the potential of using CNNs for the early detection of DR based on ultra-widefield (UWF) fundus images. The authors evaluated the performance of automated detection of DR systems using two types of segmentation images: ETDRS 7SF and F1-F2 images. Their findings highlight the potential of using DL algorithms and UWF fundus images for early DR detection. On the other hand, Umamageswari et al. [18] considered this a valuable contribution because it presents a novel approach to classifying diabetic patients by analyzing retinal images using DL techniques. Using CNN, the authors analyze retinal images to identify features associated with DR. The CNN model is trained and tested on a large dataset of retinal images, which includes both healthy and diabetic subjects. The proposed DL-based approach demonstrates high accuracy and efficiency in detecting DR, surpassing traditional methods and other DL models. The study highlights the potential of DL techniques for improving the diagnosis and treatment of complications related to diabetes, particularly DR. Furthermore, Menaouer et al. [9] propose a hybrid DL strategy using CNN and two VGG network models (VGG16 and VGG19) for the detection of DR. The study utilizes data augmentation to equalize the distribution of photos across the classes that represent the different levels of disease severity. The authors demonstrate the effectiveness of the proposed hybrid DL strategy for DR detection. Another study by Kobat et al. [8] introduced a novel technique for the automated classification of DR pictures in fundal photographs using horizontal and vertical patch division. Using the DenseNet201 architecture combined with transfer learning, they created a model for patch division that is not fixed in size. The model achieved high accuracy in classifying DR images, indicating its potential for use in the automated detection of DR. Moreover, the study by Asia et al. [2] focuses on the use of CNN classification models for detecting DR in retinal fundus images. They apply methods to preprocess the images and improve their quality, as well as utilize data augmentation techniques to expand the dataset and prevent overfitting. Through experimentation with various CNN architectures, they achieved high accuracy in detecting DR and further enhanced performance through transfer learning. Yasashvini et al. [19] utilized CNN, a hybrid CNN with ResNet, and a hybrid CNN with DenseNet on a large dataset containing approximately 3,662 training images to automatically identify the stage of progression of DR. Image augmentation techniques were used to improve robustness, while brightness and contrast were enhanced using image processing. In the study by S et al. [12], a new deep learning model is proposed for grading the severity of DR. The model combines a CNN architecture (VGG-16) with an attention module to accurately capture deteriorated regions in retinal fundus images. The study demonstrates the effectiveness of the model by evaluating various preprocessing pipelines. It achieves a high level of accuracy in grading the severity of DR. Finally, the paper by Al-Hazaimeh et al. [1] presents a method that combines image processing and artificial intelligence for the accurate detection of DR in retinal fundus images. The proposed approach surpasses existing techniques, achieving a detection accuracy higher than 98.80 percent. It addresses the limitations of current methods and provides improved performance and accuracy in diagnosing DR. Our study builds upon previous research that has achieved high accuracy in classifying diseases by simplifying disease complexity through subclassification. However, we encountered challenges related to the size of the dataset and issues with convergence. To tackle these challenges, we fine-tuned multiple CNN models on the retinal image dataset using techniques such as learning rate scheduling, learning rate reduction, and callback functions to save the best-performing model. We also increased the number of convolutional filters and utilized data augmentation techniques to enhance the performance of the model. Our study provides insights into the challenges and solutions of using DL for disease classification. Table 1 provides a summary of the references.

Table 1. Overview of previous studies concerning the use of DL to classify DR images

Rev. No.	Dataset	Methods	Best Method	Year
[4]	Kaggle	BNN CNN DNN	DNN Accuracy 89.6%	2018
[13]	EyePACS, Messidor-2, and Messidor-1.	CNN EyePACS CNN Messidor-1 CNN Messidor-2	CNN Messidor-1 Sensitivity 88.8% Specificity 89.9% AUC 95.8%	2020
[10]	Catholic Kwandong University International St. Mary's Hospital, South Korea.	CNN ETDRS 7SF CNN F1–F2	CNN ETDRS 7SF Accuracy 83.38% Sensitivity 83.38% Specificity 83.41% AUC 91.50%	2021
[18]	Mendeley diabetes dataset	SVM Nave Bias RBF Kernel SVM KNN CNN	CNN Accuracy 94% Specificity 100% Sensitivity 92%	2022
[9]	APTOS 2019 DR	CNN	Accuracy 90.60% Recall 95% F1 score 94% Precision 94.66%	2022
[8]	APTOS 2019 DR	DenseNet201	Accuracy Case 1 94.06% Case 2 87.43%	2022
[2]	Xiangya No. 2 Hospital Ophthalmology (XHO), Changsha, China	ResNet-101 ResNet-50 VggNet-16	ResNet-101 Accuracy 98% Loss 34%	2022

(Continued)

Table 1. Overview of previous studies concerning the use of DL to classify DR images *(Continued)*

Rev. No.	Dataset	Methods	Best Method	Year
[19]	Kaggle	CNN CNN with ResNet CNN with DenseNet	CNN with DenseNet Accuracy 96.22%	2022
[1]	Kaggle	SVMGA	Accuracy 98.8% Specificity 96.4% Sensitivity 99.2%	2022
[12]	Kaggle EyePacs dataset	VGG-16	Accuracy 89.20%	2022

3 METHODOLOGY

This section provides an overview of the system architecture implemented in this study, which aims to assess the severity of DR based on the size and distribution of exudates. A visual representation of the primary components of the system can be found in Figure 1. It outlines the five stages of DR and the classification system. In more advanced cases, the severity of DR is further influenced by the presence and extent of exudates and hemorrhages. The system employs image filtering and other preprocessing methods to segment components, extract features, and classify images. A CNN is utilized for this task, analyzing fundus retinal images to determine the stage of DR. The system incorporates four distinct CNN models: ResNet-101, EfficientNetB0, InceptionResNetV2, and DenseNet121, providing a comprehensive approach to disease diagnosis.

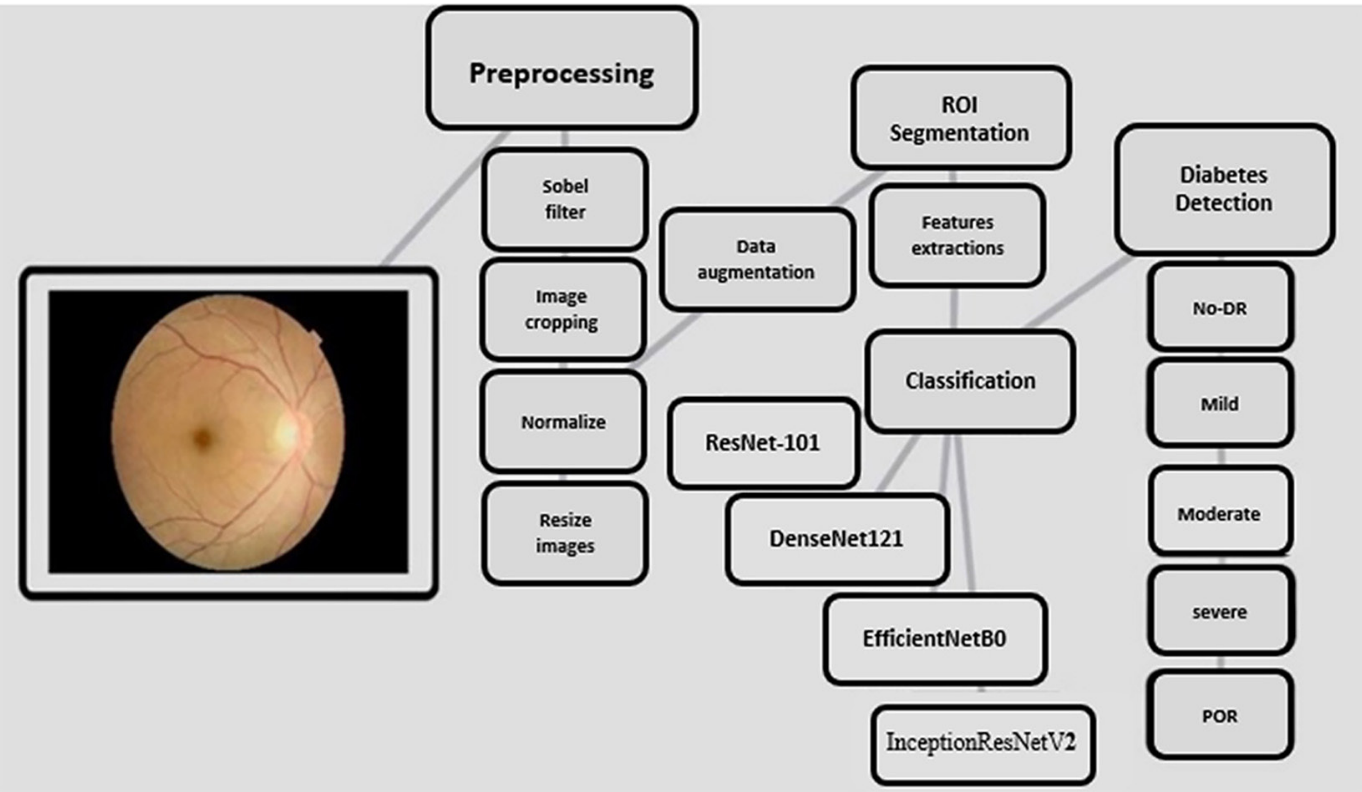

Fig. 1. General architecture

3.1 Datasets

A total of 64,000 fundus images were collected from three distinct datasets obtained from Kaggle. The first dataset comprised 35,000 fundus images; the second dataset contained 25,000 fundus images; and the third dataset included 4,000 fundus images. These images were divided into five classes: proliferative DR, mild DR, MDR, SDR, and no DR. Figure 2 shows sample images from both datasets. For training the model, 60,000 images were utilized, with 80% of the images used for training and 20% for validation. The remaining images from the third dataset were reserved for testing the model.

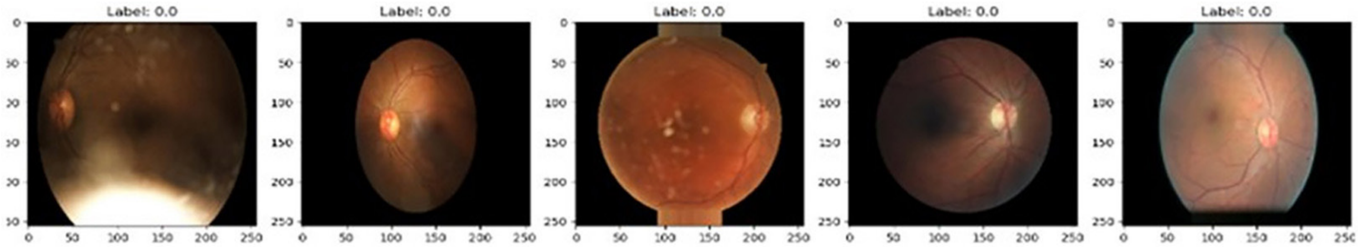

Fig. 2. Samples of dataset retinal images

3.2 Data preparation and pre-processing

The dataset contains images taken with various camera models and lighting conditions, which can affect their appearance. The images also vary in size, orientation, and the presence of off-centered circular disks. To prepare the images for CNN models, we preprocessed them by removing black borders and applying various techniques. To maintain a consistent aspect ratio and ensure that the entire fundus is visible in the image, the images were cropped to a square shape using the minimum dimension as a reference. Random cropping was avoided, and a technique was employed to dynamically extract the color of the leftmost border pixel. Using this color information, the whitespace around the fundus image was trimmed while preserving the original aspect ratio [1]. Next, we applied illumination correction by using a Gaussian blur and subtracting the local average color to minimize variations in illumination. We also applied the Sobel filter to enhance the edges and boundaries in the image. Additionally, the output was normalized to ensure consistent intensities. We then rescaled the pixel intensities from 0–255 to 0–1 using min-max normalization. Finally, we applied data augmentation by randomly zooming and shifting the images up to 20% to generate modified versions and improve the generalization of the DL models. Figure 3 displays a sample image before and after preprocessing.

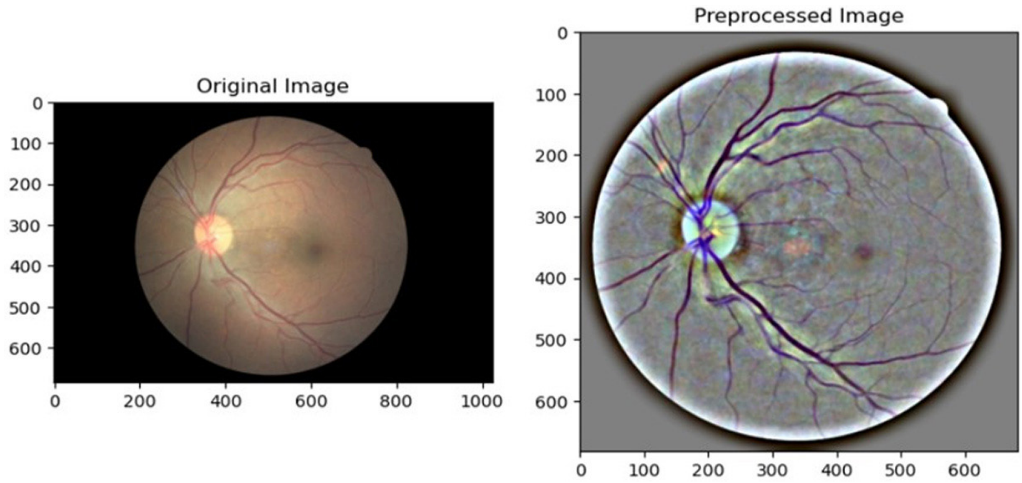

Fig. 3. Original image and the preprocessed image

3.3 Models' details

The models used in this study, EfficientNetB0, ResNet-101, DenseNet121, and InceptionResNetV2, have shown success in image classification tasks. They were selected to evaluate their effectiveness in diagnosing diabetic retinopathy.

EfficientNetB0 model. EfficientNetB0 is a family of CNN architectures optimized for accuracy and efficiency. We utilized the EfficientNetB0 variant, which was fine-tuned on the target dataset, for DR classification. The pre-trained model with ImageNet weights was utilized, and the final fully connected layer was replaced using a new layer containing five output units with SoftMax activation. Figure 4a shows the structure of EfficientNetB0, including the number of parameters for each level and all layers.

ResNet-101 model. ResNet-101 is a deep residual neural network proposed by He et al. (2016) [6] to address the vanishing gradient problem in deep neural networks. It contains 101 layers and utilizes skip connections to reuse activations from previous layers. The pre-trained model with ImageNet weights was utilized, and the final fully connected layer was substituted with a new layer that consists of five output units with SoftMax activation. The model structure is displayed in Figure 4b.

DenseNet121 model. DenseNet121 is a popular CNN architecture that utilizes dense connections to promote feature reuse and enhance gradient flow. The network is structured in dense blocks, where each layer is connected to every other layer within the block. Bottleneck layers and transition layers are used to reduce computational complexity. DenseNet121 utilizes global average pooling instead of fully connected layers prior to the final classification layer. The model structure, the number of parameters for each level, and all layers are displayed in Figure 4c.

InceptionResNetV2 model. The InceptionResNetV2 model is a DL model used for image classification. It combines the Inception architecture with residual connections. The Inception architecture utilizes multiple parallel convolutional layers with varying filter sizes and a concatenation operation to enhance the depth and width of the network, all while ensuring computational efficiency. Residual connections mitigate the vanishing gradient problem and improve the flow of gradients during back propagation. The InceptionResNetV2 model includes several Inception modules with residual connections, resulting in a robust and efficient model. The model's components include the stem, Inception-ResNet-A blocks, a Reduction-A block, Inception-ResNet-B blocks, and Inception-ResNet-C blocks, as well as global average pooling and SoftMax. The integration of Inception and residual connections enables the network to learn complex feature representations and maintain gradient flow, thereby allowing for the training of deeper models with improved performance. Figure 4d shows the network structure of the InceptionResNetV2 model.

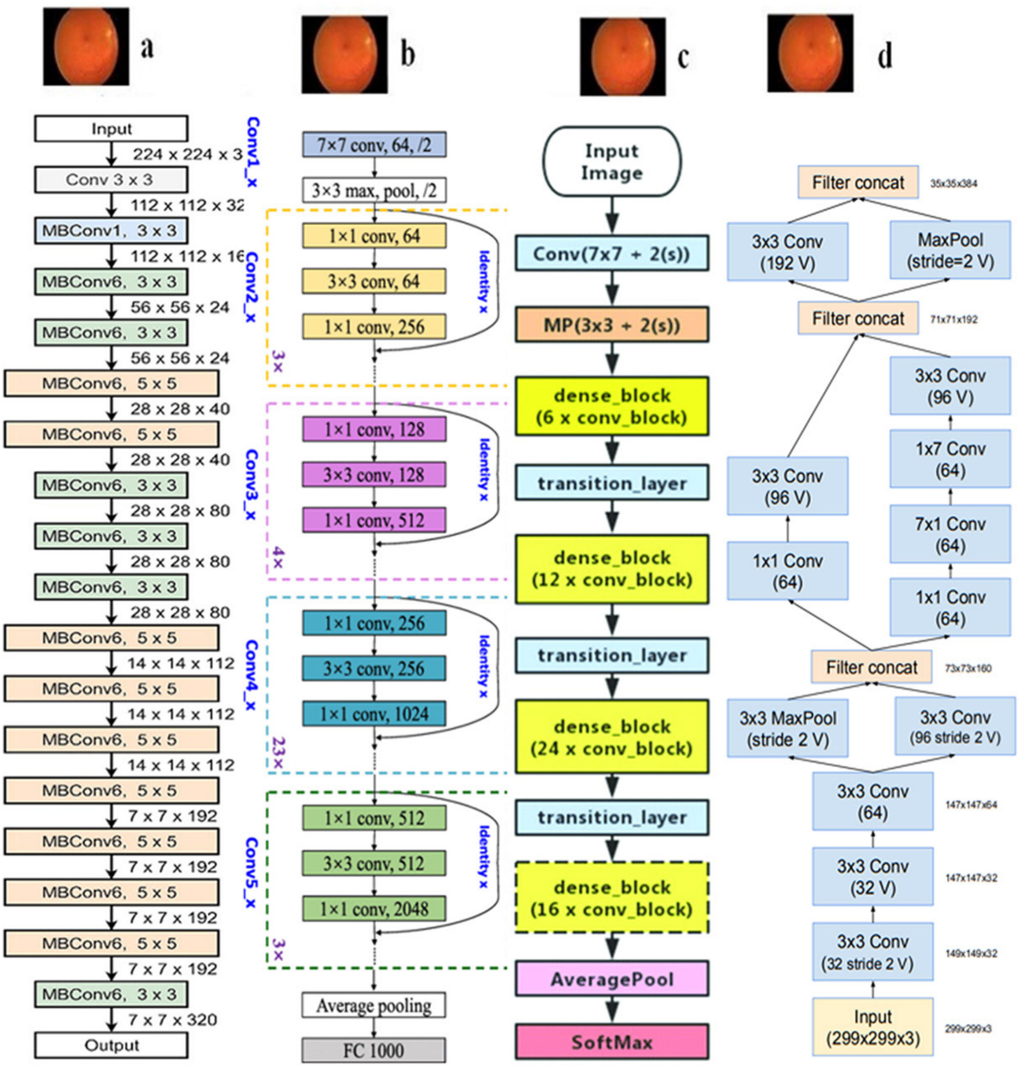

Fig. 4. Architecture of (a) EfficientNetB0, (b) ResNet-101, (c) DenseNet121, and (d) InceptionResNetV2

3.4 Training and hyperparameters tuning

Experiments were conducted to compare the performance of the ResNet-101, DenseNet121, InceptionResNetV2, and EfficientNetB0 models. The same model structures, hyperparameters, and training procedures were used to ensure a fair comparison. These experiments, which maintain a high standard of transparency and integrity as recommended by [3], are discussed in detail in the following subsections.

Initial hyperparameters experiment. Transfer learning was employed in this study by using pre-trained models as the base model. A new classification layer was added on top of the base model, which was randomly initialized and then frozen. Freezing the layers of the base model resulted in a reduction of parameters, which led to faster training and prevented overfitting. The models were trained for 40 epochs with a batch size of 16, using the Adam optimizer with a learning rate of 0.0001. The results in Table 2 show that the DenseNet121 model achieved the highest accuracy of 0.58 in classifying the five DR severity stages. However, the InceptionResNetV2 and ResNet-101 models achieved relatively low accuracies of 0.48 and 0.39, respectively. The EfficientNetB0 model performed the worst, with an accuracy of 0.24, which was lower than the other two models.

Table 2. Initial hyperparameters experiment setup and evaluation

Architecture	Learning Rate	Batch Size	Epochs	Accuracy
EfficientNetB0	0.0001	16	40	0.24
ResNet-101	0.0001	16	40	0.39
DenseNet121	0.0001	16	40	**0.58**
InceptionResNetV2	0.0001	16	40	0.48

First hyperparameters experiment. Several techniques were employed to optimize the performance of the deep learning models. A global average pooling layer and a dropout layer were added to the models in order to mitigate overfitting. A learning rate schedule was defined to dynamically adjust the learning rate during training. For the first 10 epochs, the learning rate is set to 0.0001. From epochs 10 to 15, the learning rate increased to 0.0005. From epochs 15 to 30, the learning rate is reset to 0.0001. After epoch 30, the learning rate is reduced by a factor of (epoch/(1+epoch)), which helps the models converge to better local minima. Additionally, a model checkpoint callback was created to save the model weights corresponding to the lowest validation loss, capturing the best-performing model. The models were trained for 70 epochs with a batch size of 16, aiming to strike a balance between training time and model performance. The Adam optimizer was utilized because of its effectiveness in training deep neural networks. The DenseNet121 model achieved the highest accuracy of 0.86 for classifying the five DR severity stages, as shown in Table 3. The accuracy of the InceptionResNetV2 model is 0.82. The accuracy of the ResNet101 model improved to 0.79 compared to the first experiment, but it still remained lower. The EfficientNetB0 model achieved an accuracy close to 0.80.

Table 3. First hyperparameters experiment setup and evaluation

Architecture	Dropout Layer	Learning Rate	Batch Size	Accuracy
EfficientNetB0	0.5	0.0001,0.0005	16	0.80
ResNet-101	0.5	0.0001,0.0005	16	0.79
DenseNet121	0.5	0.0001,0.0005	16	**0.86**
InceptionResNetV2	0.5	0.0001,0.0005	16	0.82

Second hyperparameters experiment. The experiment showed improved performance compared to previous experiments due to several significant modifications. A significant improvement was made by adding a new dense layer with 2048 neurons. This allowed the models to learn more complex features from the data. Only the top layers of the pre-trained models were fine-tuned, leveraging the knowledge from the pre-trained weights while allowing the models to learn task-specific features. A fixed learning rate of 0.0001 was used to stabilize the training. Regularization techniques, including the EarlyStopping and ReduceLROnPlateau callbacks, were implemented to mitigate overfitting. EarlyStopping stops training when the validation loss stops improving, while ReduceLROnPlateau decreases the learning rate when the validation loss plateaus. The models were trained for 40 epochs with a batch size of 8, and the parameters were chosen to optimize performance. These methodological changes helped prevent overfitting and improve the models' ability to generalize. Based on the results in Table 4, the DenseNet121 model demonstrated the highest accuracy at 87%, while InceptionResNetV2, EfficientNetB0, and ResNet-101 achieved accuracies of 86%, 84%, and 82%, respectively.

Table 4. Second hyperparameters experiment setup and evaluation

Architecture	Regularization Techniques	Learning Rate	Batch Size	Accuracy
EfficientNetB0	Early Stopping, ReduceLROnPlateau	1e-4	8	0.84
ResNet-101	Early Stopping, ReduceLROnPlateau	1e-4	8	0.82
DenseNet121	Early Stopping, ReduceLROnPlateau	1e-4	8	**0.87**
InceptionResNetV2	Early Stopping, ReduceLROnPlateau	1e-4	8	0.86

4 RESULTS

The behavior of the four models (EfficientNetB0, ResNet-101, DenseNet121, and InceptionResNetV2) during training is similar in several aspects. Initially, the models exhibit low accuracy on both the training and validation sets, indicating a poor fit to the data. However, as training progresses, the models gradually learn from the data, resulting in improved accuracy during training. A notable trend observed is that the validation accuracy reaches a peak and then starts to decline, indicating that the models are overfitting to the training data. For EfficientNetB0, overfitting becomes apparent around epoch 19. To prevent further overfitting, early stopping is triggered at epoch 24. ResNet-101 demonstrates rapid learning in the initial 7 epochs and achieves the highest validation performance. However, overfitting starts to occur after epoch 7, and to prevent this early stopping, it is implemented at epoch 12 to revert back to weights that performed the best at epoch 7. Similarly, DenseNet121 demonstrates rapid learning and achieves its highest validation performance at epoch 15. Overfitting sets in after Epoch 15, and early stopping at Epoch 20 is used to restore the best weights from Epoch 15. InceptionResNetV2 performs well on the training set, with accuracy steadily increasing over the 10 epochs. However, the performance of the validation set is comparatively lower, suggesting overfitting to the training set and a limited ability to generalize to new data. These observations highlight the significance of monitoring both training and validation accuracy during training in order to detect importance and prevent excessive training. Early stopping proves to be an effective technique for preventing overfitting and restoring the best-performing weights, as shown in Figure 5.

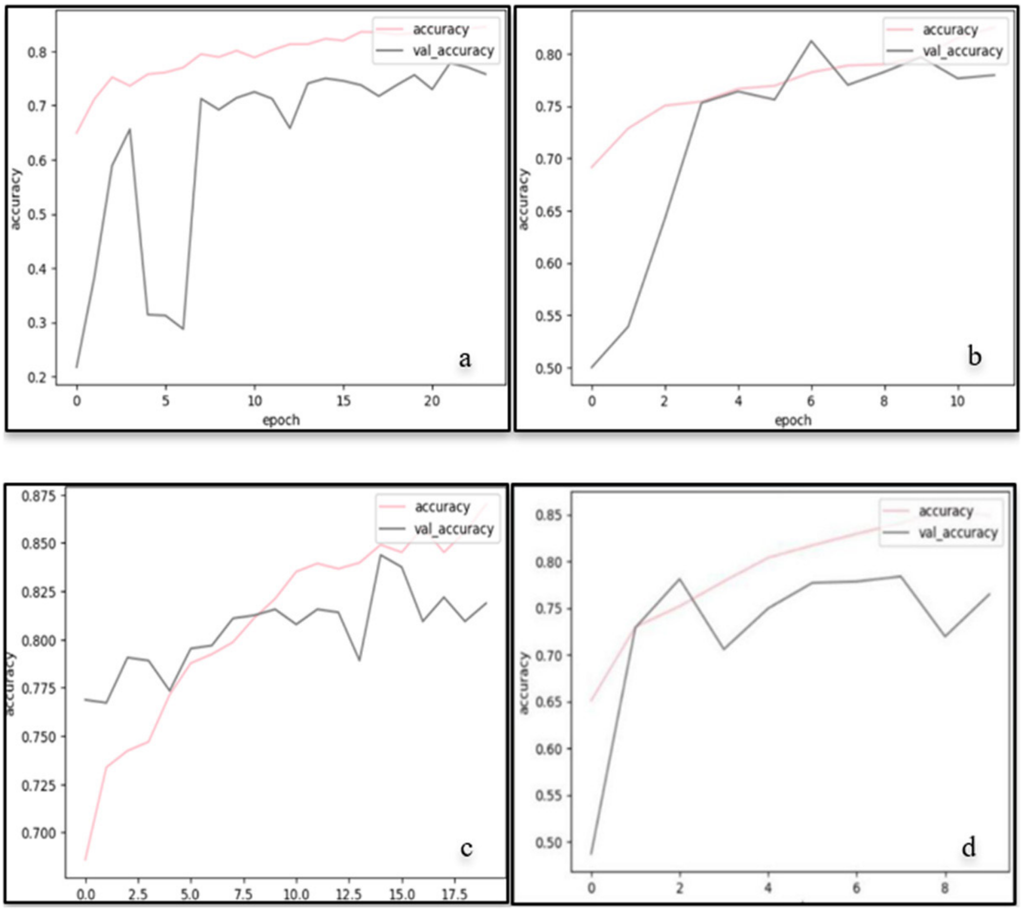

Fig. 5. Learning curves of (a) EfficientNetB0; (b) ResNet-101; (c) DenseNetB0; and (d) InceptionResNetV2

Based on the results in Table 5, DenseNet121 emerged as the best-performing model overall, achieving the highest training accuracy of 0.87, precision of 0.89, and AUC of 0.98. These metrics indicate that DenseNet121 was the most effective in correctly classifying images while minimizing incorrect predictions. InceptionResNetV2 achieved the second-highest results, with a training accuracy of 0.86, precision of 0.87, and AUC of 0.95. Although its performance was still good, it was lower than that of DenseNet121, suggesting that DenseNet121 had better generalization to the data. ResNet-101 has a lower training accuracy of 0.82 and precision of 0.85 compared to InceptionResNetV2, but it has a higher AUC of 0.97. EfficientNetB0, on the other hand, exhibited the lowest performance, with a training accuracy of 0.73, precision of 0.75, and an AUC of 0.93. This implies that, despite being a more efficient model, EfficientNetB0 was not able to achieve as high a performance on this dataset and task compared to the other models.

Table 5. Training results for the models

Training	Accuracy	Precision	AUC
EfficientNetB0	0.73	0.75	0.93
ResNet-101	0.82	0.85	0.97
DenseNet121	**0.87**	**0.89**	**0.98**
InceptionResNetV2	0.86	0.87	0.95

Based on Table 6, the DenseNet121 model achieves the best overall performance, with the highest validation accuracy of 0.84, precision of 0.87, and AUC of 0.97. This suggests that it is the top-performing model for this classification task among the four models. The ResNet101 model achieves reasonably high validation precision and AUC values for the five classes, although its accuracy of 0.82 is slightly lower than that of DenseNet121. The study conducted by [2] achieved an accuracy of 0.98 for a ResNet-101 binary classification, which indicated that it performs well for multiclass tasks. The InceptionResNetV2 model has a precision value of 0.79, which is the lowest among the models. However, it achieves a relatively high validation accuracy of 0.83 and an AUC of 0.95. The EfficientNetB0 model performs the worst, with the lowest validation accuracy of 0.80 and precision of 0.81. However, it still achieves a reasonably high AUC of 0.95.

Table 6. Validation results for the models

Validation	Accuracy	Precision	AUC
EfficientNetB0	0.73	0.76	0.93
ResNet-101	0.81	0.84	0.96
DenseNet121	**0.84**	**0.87**	**0.97**
InceptionResNetV2	0.79	0.82	0.96

You can observe from Table 7 that DenseNet121 achieves the highest test accuracy, precision, and AUC among the four models. It has a test accuracy of 0.83, a precision of 0.85, and an AUC of 0.97. ResNet-101 achieved the second highest results, with a test accuracy of 0.81, a precision of 0.83, and an AUC of 0.96. Inception ResNetV2 has a test accuracy of 0.76, a precision of 0.78, and an AUC of 0.94. EfficientNetB0 has the lowest test accuracy, precision, and AUC among the four models, with a test accuracy of 0.75, precision of 0.77, and AUC of 0.95.

Table 7. Test results for the models

Test	Accuracy	Precision	AUC
EfficientNetB0	0.75	0.77	0.95
ResNet-101	0.81	0.83	0.96
DenseNet121	**0.83**	**0.85**	**0.97**
InceptionResNetV2	0.76	0.78	0.94

5 IMAGE PREDICTION

To assess the performance of our DL models in classifying the severity of DR, we evaluated their predictions on retinal scan images that were not included in the model training (see Figure 6). While the models were generally accurate, they misclassified some of the images in the test set, indicating that there is room for improvement in their performance. The misclassifications suggest that the models still struggle with certain aspects of classifying the severity of DR. Though the models showed good overall accuracy, the misclassified test scans indicate that performance could still be optimized to reduce errors.

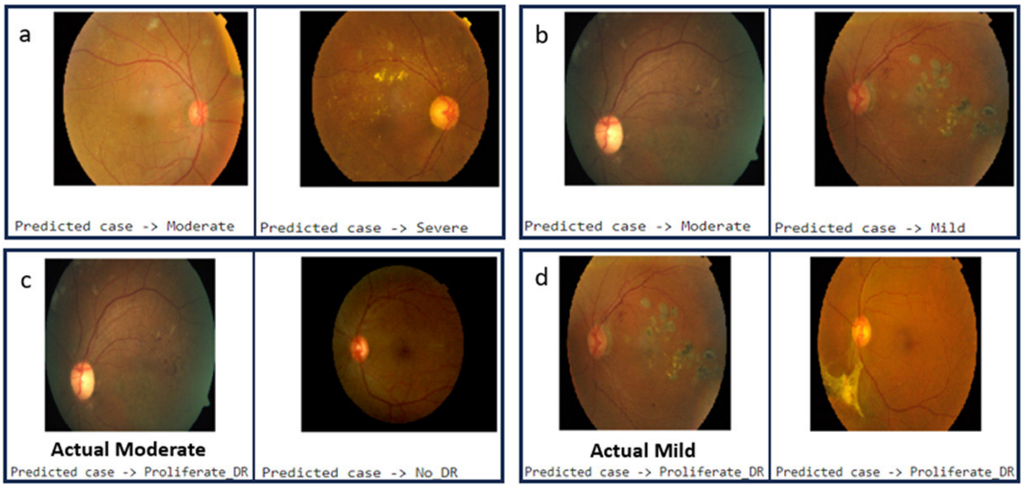

Fig. 6. Models' prediction (a) DenseNetB0, (b) ResNet-101, (c) EfficientNetB0, and (d) InceptionResNetV2

6 CONCLUSION

The objective of the study was to develop a classification method for categorizing the five stages of retinitis pigmentosa, a condition that can lead to vision loss if not detected early, particularly in individuals with diabetes. The researchers utilized DenseNet-121, ResNet-101, Inception-ResNetV2, and EfficientNetB0 models to automate and enhance the efficiency of the classification process. Among the models evaluated, DenseNet-121 achieved the highest accuracy of 0.87, followed by Inception-ResNetV2 with 0.86, ResNet-101 with 0.82, and EfficientNetB0 with 0.73. It is worth noting that all models performed well in terms of accuracy. These results demonstrate the potential of DL techniques for disease classification tasks. However, the study also revealed the challenges associated with achieving higher accuracy, such as convergence issues and the requirement for more extensive training. To address these challenges, the researchers suggest exploring preprocessing techniques and ensemble learning, which have shown promise for improving classification accuracy. The study highlights the importance of utilizing advanced machine learning methods to improve disease diagnosis and management. To further enhance accuracy, future research could focus on ensemble methods that combine the most effective models. Additionally, addressing convergence challenges and conducting more extensive training can improve the performance of DL models. The implications of these findings extend beyond retinitis pigmentosa and can be applied to other disease classification tasks. This emphasizes the importance of utilizing advanced ML techniques to advance disease diagnosis and management.

7 REFERENCES

[1] O. M. Al-Hazaimeh, A. Abu-Ein, N. Tahat, M. Al-Smadi, and M. M. Al-Nawashi, "Combining artificial intelligence and image processing for diagnosing diabetic retinopathy in retinal fundus images," *International Journal of Online and Biomedical Engineering*, vol. 18, no. 13, pp. 131–151, 2022. https://doi.org/10.3991/ijoe.v18i13.33985

[2] A.-O. Asia, C.-Z. Zhu, S. A. Althubiti, D. Al-Alimi, Y.-L. Xiao, P.-B. Ouyang, and M. A. A. Al-Qaness, "Detection of diabetic retinopathy in retinal fundus images using CNN classification models," *Electronics*, vol. 11, no. 17, p. 2740, 2022. https://doi.org/10.3390/electronics11172740

[3] C. Closa, "Planning, implementing and reporting: Increasing transparency, replicability and credibility in qualitative political science research," *European Political Science*, vol. 20, pp. 270–280, 2021. https://doi.org/10.1057/s41304-020-00299-2

[4] S. Dutta, B. Manideep, S. M. Basha, R. D. Caytiles, and N. Ch. S. N. Iyengar, "Classification of diabetic retinopathy images by using deep learning models," *International Journal of Grid and Distributed Computing*, vol. 11, no. 1, pp. 99–106, 2018. https://doi.org/10.14257/ijgdc.2018.11.1.09

[5] R. Gargeya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," *Ophthalmology*, vol. 127, no. 7, pp. 919–927, 2020.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90

[7] A. Khan, N. Kulkarni, A. Kumar, and A. Kamat, "D-CNN and image processing based approach for diabetic retinopathy classification," in *Advances in Intelligent Systems and Computing*, Springer, Singapore, 2022, vol. 1354, pp. 283–291. https://doi.org/10.1007/978-981-16-2008-9_27

[8] S. G. Kobat, N. Baygin, E. Yusufoglu, M. Baygin, P. D. Barua, S. Dogan, O. Yaman, U. Celiker, H. Yildirim, R.-S. Tan, T. Tuncer, N. Islam, and U. R. Acharya, "Automated diabetic retinopathy detection using horizontal and vertical patch division-based pre-trained DenseNET with digital fundus images," *Diagnostics*, vol. 12, no. 8, p. 1975, 2022. https://doi.org/10.3390/diagnostics12081975

[9] B. Menaouer, Z. Dermane, N. El Houda Kebir, and N. Matta, "Diabetic retinopathy classification using hybrid deep learning approach," *SN Computer Science*, vol. 3, no. 357, 2022. https://doi.org/10.1007/s42979-022-01240-8

[10] K. Oh, H. M. Kang, D. Leem, H. Lee, K. Y. Seo, and S. Yoon, "Early detection of diabetic retinopathy based on deep learning and ultra-wide-field fundus images," *Scientific Reports*, vol. 11, no. 1897, 2021. https://doi.org/10.1038/s41598-021-81539-3

[11] R. Poplin, A. V. Varadarajan, K. Blumer *et al.*, "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning," *Nature Biomedical Engineering*, vol. 2, pp. 158–164, 2018. https://doi.org/10.1038/s41551-018-0195-0

[12] S. Geetha, M. Parashar, J. S. Abhishek, R. V. Turaga, I. A. Lawal, and S. Kadry, "Diabetic retinopathy grading with deep visual attention network," *International Journal of Online and Biomedical Engineering*, vol. 18, no. 9, pp. 160–177, 2022. https://doi.org/10.3991/ijoe.v18i09.30075

[13] G. Saxena, D. K. Verma, A. Paraye, A. Rajan, and A. Rawat, "Improved and robust deep learning agent for preliminary detection of diabetic retinopathy using public datasets," *Intelligence-Based Medicine*, vols. 3–4, p. 100022, 2020. https://doi.org/10.1016/j.ibmed.2020.100022

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[15] A. Sungeetha and R. R. Sharma, "Design an early detection and classification for diabetic retinopathy by deep feature extraction based convolution neural network," *Journal of Trends in Computer Science and Smart Technology*, vol. 3, no. 2, pp. 81–94, 2021. https://doi.org/10.36548/jtcsst.2021.2.002

[16] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017. https://doi.org/10.1609/aaai.v31i1.11231

[17] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning,* ICML 2019, Long Beach, 2019, pp. 6105–6114.

[18] A. Umamageswari, S. Deepa, and L. S. Beevi, "Novel approach for classification of diabetics from retinal image using deep learning technique," *International Journal of Health Sciences*, vol. 6, no. 1, pp. 2729–2736, 2022. https://doi.org/10.53730/ijhs.v6nS1.5196

[19] R. Yasashvini, M. V. Raja Sarobin, R. Panjanathan, G. Jasmine, and L. J. Anbarasi, "Diabetic retinopathy classification using CNN and hybrid deep convolutional neural networks," *Symmetry*, vol. 14, no. 9, p. 1932, 2022. https://doi.org/10.3390/sym14091932

8 AUTHORS

Roaa Al-ahmadi received her bachelor's degree in mathematics from Umm Al-Qura University in 2010 which launched over a decade of experience as a mathematics teacher, during which time she also earned a master's degree in data science from the University of Jeddah in 2023 to expand her expertise in data analysis. Currently, she works as a mathematics teacher where she shares her passion while continuing her scholarly work (E-mail: alahmadiroaa@yahoo.com).

Hatoon Al-ghamdi earned a Bachelor of Science degree in Statistics from King Abdulaziz University in 2018 and a Master of Computer Science and Engineering degree in Data Science from the University of Jeddah in 2023 (E-mail: 2200634@uj.edu.sa).

Dr. Lobna Hsairi is an Assistant Professor in the Information System and Technology Department in the College of Computer Science and Engineering at the University of Jeddah (KSA). She focuses her research on data science, intelligent agents, machine learning, Deep Learning, analysis, and intelligent control of large-scale complex systems (E-mail: Lalhabib@uj.edu.sa).