Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 22 No. 5 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v22i05.59075

PAPER

Enhanced Diabetic Retinopathy Identification via EfficientNetB3-Based Convolutional Neural Networks and Transfer Learning

Orlando Iparraguirre-Villanueva[1] (✉), **Michael Cabanillas-Carbonell**[2]

[1]Universidad Tecnológica del Perú, Chimbote, Peru

[2]Universidad Privada del Norte, Lima, Peru

c27399@utp.edu.pe

ABSTRACT

Diabetic retinopathy (DR) is one of the most common causes of blindness worldwide, making early detection essential for treating this disease. This research presents the creation and testing of a convolutional neural network (CNN) model based on the EfficientNetB3 architecture to detect signs of DR in typical fundus images. The model was trained with a set of retinal images and tested using metrics such as accuracy, recall, and F1 score. The results show a weighted accuracy of 81%, with high performance in the healthy class (97% accuracy and 98% recall). However, lower accuracy is observed in the advanced stages of DR, mainly attributed to class imbalance in the dataset. It is also necessary to balance the classes and combine the architecture with other models. These findings demonstrate the potential of EfficientNetB3 as a diagnostic support tool and highlight the need to improve data balance and the model for better discrimination of severe cases.

KEYWORDS

computer vision, classification, convolutional neural network (CNN), diabetic retinopathy (DR)

1 INTRODUCTION

Diabetes mellitus, a chronic disease affecting millions of people worldwide, presents several long-term complications that can significantly compromise patients' quality of life [1]. Among these complications, diabetic retinopathy (DR) stands out as a major cause of blindness [2]. It is characterized by progressive damage to the blood vessels in the retina, specifically in the light-sensitive upper part of the eye. According to the World Health Organization (WHO)/International Diabetes Federation (IDF), DR is one of the leading causes of visual impairment in the European region, with more than 950,000 people affected [3], [4]. Over time, hyperglycemia causes weakening of the blood vessel walls, leading to increased leakage of blood and fluid into

Iparraguirre-Villanueva, O., Cabanillas-Carbonell, M. (2026). Enhanced Diabetic Retinopathy Identification via EfficientNetB3-Based Convolutional Neural Networks and Transfer Learning. *International Journal of Online and Biomedical Engineering (iJOE)*, 22(5), pp. 139–155. https://doi.org/10.3991/ijoe.v22i05.59075

Article submitted 2025-10-23. Revision uploaded 2026-02-05. Final acceptance 2026-02-05.

the retina [5]. Risk factors such as duration of diabetes, high blood sugar levels, and hypertension can accelerate damage to blood vessels, and pregnancy can also increase the risk of DR [6]. The disease in its early stages may have no symptoms. However, as the disease progresses, symptoms may manifest with vision loss in certain areas, difficulty seeing at night and blurred vision being the most common symptom [7]. DR comes in two types: non-proliferative and proliferative. The former is the most common form and is characterized by damage to the blood vessels, and the latter is a more advanced stage and occurs when new abnormal blood vessels grow in the retina [8], [9].

Traditionally, the diagnosis of RD has been made by direct visual assessment by an ophthalmologist, who examines fundus images for characteristic signs such as micro-aneurysms, exudates, and hemorrhages [10]. Also, ophthalmologists are known to diagnose RD by looking at fundus images, which may be optical coherence tomography images that allow visualization of details of RD lesions [11]. However, this method has certain limitations, such as diagnostic subjectivity, observer variability, and the need for highly trained specialists [12].

In recent years, the advancement of artificial intelligence (AI), and particularly the development of convolutional neural network (CNNs), has opened new avenues for early detection [13]. CNNs are a type of artificial neural network specifically designed to process visual data [14] and function in a similar way to the way the human brain recognizes patterns in images [15]. It demonstrates great ability to analyze images and extract important features, making it a promising tool for computer-aided diagnosis of various diseases, including DR [16], [17].

This work aims to develop and evaluate a CNN model capable of identifying the characteristic signs of DR in fundus images to provide a clinical diagnostic support tool for healthcare professionals and improve outcomes for patients with DR by reducing the risk of vision loss.

2 LITERATURE REVIEW

Several studies have explored the use of CNNs for DR detection, obtaining significant results. For example, in [18] they worked on an approach using the CNN-based random forest technique to improve classifier accuracy. The results of the proposed approach were promising, achieving an accuracy rate of .9791. Similarly, the authors of the study [19] analyzed two CNN models using transfer learning for DR classification using a highly imbalanced dataset. The effectiveness of the models was demonstrated through metrics, where the MobileNet model performed better than MobileNetV2. MobileNet scored .80 accurately, while MobileNetV2 scored .71. Also, the authors in the paper [20] combined three models, two deep learning models (Inception-v3 and VGG16) and a custom CNN. The weighted average reached a rate of .9506 and a .8788 in the area under the curve (AUC); this result shows that the proposal can help as an assistant to detect and classify DR. Similarly, research [21] proposed a diagnostic system for RD based on CNN. After training, validation, and testing, they chose the ResNet50 model, since it obtained the best accuracy metrics with .929. This result confirms that CNNs are excellent predictors for this type of task. Similarly, in [22] they proposed a DR diagnosis system using CNNs and deep learning (DL). In this proposed system the RestNet50 model is the one that obtained the highest overfitting compared to Inception v3. EfficientNetB4 is the best-performing model in DR detection with a validation accuracy of .7911, followed by InceptionResNetV2, NasNetLarge, and DenseNet169, respectively.

In the same line, in the paper [23], they proposed a technique for DR detection, for which they also used a parallel CNN for feature extraction. They used two datasets: 1 with 34,984 images and 2 with 3,662 images. The results showed that the technique used with the custom CNN achieved accuracies of .9178 for the first dataset and .9727 for the second dataset, respectively. Similarly, in the work [24], a decision-making system was developed to predict DR; this system is composed of a hybrid convolutional network (HCN) and a recurrent neural network (RNN), together with the hummingbird optimizing algorithm (HOA). The classification process was performed by combining the HCN, RNN, and HOA models. The results of the proposed system used different metrics to evaluate the performance of the models, achieving an accuracy of .97. Also, in [25] they used a deep learning (DL) application and a CNN to distinguish the stages of DR, for which they used the following models such as NesNet-10.1, ResNet-50, and VggNet-16. The NesNet-10.1 model achieved the best performance rate with 0.9888 accuracy, with a training loss of 0.3499 and test loss of 0.9882. These results show that the NesNet-10.1 model has the best accuracy for distinguishing DR stages.

Other similar studies, such as [26], proposed a 3D visualization method to identify DR by using optimization parameters to maximize the visibility of important structures. They then integrated the proposal into a system to obtain relevant information and perform the analysis. The results demonstrated that 3D visualization can provide information using multimodal tomography images. Also, in [27], they developed a system combining an LLM module and image-based deep learning for the purpose of making recommendations for DR diabetes management. The results showed that with the LLM module, an average accuracy of 81% was obtained for identifying DR without assistance, and with deep learning assistance, 92.3% accuracy was obtained. The authors in the paper [28] used virtual reality as a support to identify brain cancer, focusing on segmentation of brain tumor classes using magnetic resonance imaging. The approach was validated by 496 MRI images. The test results of the approach demonstrated an accuracy of 98.61%, affirming the proposed strategy. In the paper [29], they developed a deep learning-based system to detect DR in early and late stages. For which they used a dataset of 466247 fundus images, of which 121342 patients had diabetes. The evaluation was performed on 200,136 fundus images. The results ranked the RD in the area under the curve with the following metrics 0.943, 0.955, 0.960, and 0.972, respectively. Supporting the efficiency of the system to classify DR. Similarly, in the work [30], the authors developed a deep learning-based solution to grade and estimate the quality of the DR images, along with ISBI 2020. They used a dataset composed of 2000 regular DR images (500 patients) and 256 ultra-widefield images (128 patients). The findings demonstrated that image quality assessment can be used as a target for feature exploration.

In study [31], a method for automated RD detection using a CNN-based approach was presented. Using a dataset of fundus images, the researchers focused on optimizing the network layers to improve feature extraction, achieving an accuracy of 96%. On the other hand, the authors in [32] proposed a hybrid model that integrates a CNN with the SVM model to classify DR severity levels. This approach leveraged CNN for automated feature representation and SVM for robust classification, achieving an accuracy of 94.5%. Finally, in [33], a system was developed to aid in the early identification of RD using an ensemble of CNN models, including ResNet and Inception architectures. The results showed that the ensemble method outperformed individual models with an accuracy of 95.8%. As can be seen, these studies have reported metrics above 90%, thanks to the use of hybrid approaches that combine image processing with traditional classifiers and the use of balanced datasets.

However, this work uses RGB images without segmentation, which limits the extraction of discriminative features, explaining the difference in results.

3 MATERIALS AND METHODS

This section presents the theoretical foundation of CNN that underpins the research and the development of the proposal.

Convolutional neural networks are made up of several layers of artificial neurons, like the neural cells used by the human brain to transmit various sensory input signals [34], which are mathematical functions used to calculate the sum of various inputs and provide an output in the form of activation [35]. When an input image is fed into a CNN, each of the internal layers generates several activation maps [36]. The output of the first layer is reused as input to the next layer, and so on.

The architecture of a CNN is composed of four main components: 1) Convolution layer that separates and identifies image features for analysis, known as feature extraction [37]. 2) A fully connected layer that acts as the output of the convolution process and predicts based on the previously extracted features [38]. 3) Activation functions, and 4) Pooling, responsible for reducing the dimensionality of the data and capturing translation-invariant features. In CNNs there are two types of pooling: 1) average pooling and 2) maximum pooling. The fully connected layer is responsible for understanding the weights and biases together with the neurons to connect between the separate layers [39], as shown in Figure 1. In CNN training, classification problems are also encountered, and oversampling techniques such as Random oversampling and augmentation are used for this purpose. The first technique, Random oversampling, is used to complement the training data with multiple copies of the minority classes [40]. Augmentation is a technique used to increase the amount of newly created synthetic data [41]. It acts as a regularizer and helps to reduce overfitting when training the EfficientNetB3 model [42].

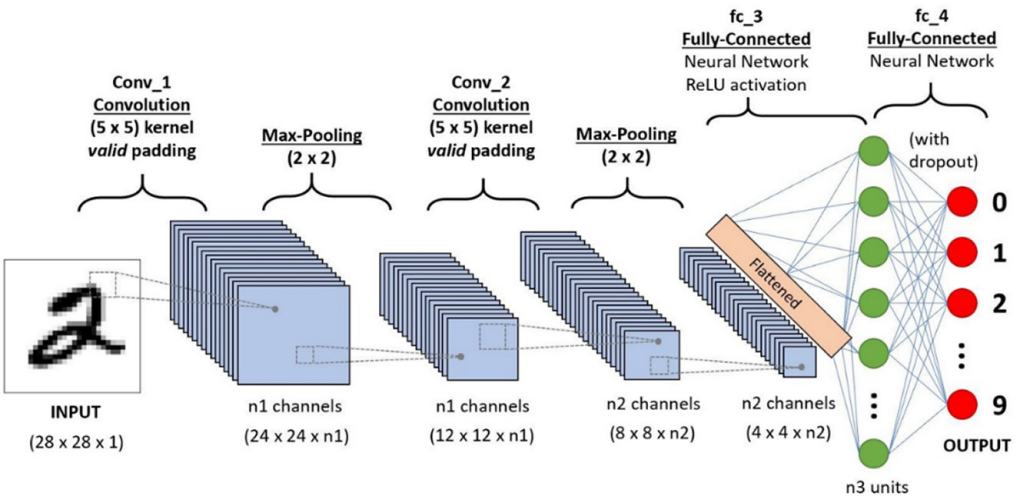

Fig. 1. CNN architecture [43]

This work uses a pre-trained EfficientNetB3 CNN as a starting point. This architecture has demonstrated good performance in various image classification and computer vision tasks [44]. The selection of EfficientNetB3 is based on its ability to extract discriminative features from retinal images, which is critical for accurate DR

detection [45]. In addition, its computational efficiency allows for faster training and the ability to deploy resource-constrained devices. This architecture uses a composite scaling that uniformly adjusts the depth, width, and resolution of the network, which allows finding an optimal architecture for the dataset [46], [47]. Also, the EfficientNetB3 architecture allows the number of output channels in each convolutional layer to be controlled, allowing the capacity of the EfficientNetB3 model to be adjusted. To adapt the model to our case study, a pre-trained version was used, and the last fully connected layers were fine-tuned by fine-tuning.

3.1 Description of the dataset

This work aims to develop and evaluate a CNN model capable of analyzing retinal images and accurately classifying retinal images according to the severity of DR and to provide an automated solution to assist healthcare professionals in diagnosing and classifying DR. To achieve this goal, a dataset [48] consisting of 2750 high-resolution retinal images captured under various conditions will be used. The images are organized according to severity: Healthy (Not DR) with 1000 images, Mild DR with 370 images, Moderate DR with 900 images, Proliferative DR with 290 images, and Severe DR with 190 images, making a total of 2750 images to work with. The architecture will be designed and configured using PD techniques to extract the most important features from the images and to discriminate between the different DR classes. Also, different regulation strategies will be used to improve the generalization of the EfficientNetB3 model and avoid overfitting.

To load the dataset, libraries such as numpy, pandas, seaborn, and matplotlib and other libraries used for training were also imported, such as “tensorflow,” “InputLayer,” “BatchNormalization,” Dropout, Conv2D, MaxPooling2D, Flatten, Dense, Activation, optimizers, Adam, Adamax, metrics, regularizers, callbacks, ModelCheckpoi, image, train_test_split, and confusion_matrix, among other libraries. The dataframe is then used to organize the data in tabular form to analyze the data. It allows complex operations with filtering, aggregation, grouping, and transformation of data. It is also capable of handling heterogeneous data and facilitates data manipulation to eliminate null values and duplicates, among other operations.

3.2 Data pre-processing

In this part, we first address oversampling methods for classification problems. For example, random oversampling involves enriching the training data with several layers of certain minority classes. This procedure can be carried out more than once. This is one of the first procedures that has proven to be robust [49], since, instead of duplicating the minority class, some of these can be randomly selected by substitution. The next method employed is augmentation, with the aim of increasing the amount of data by adding slightly altered layers or synthetic data. This works as a regularizer and helps to minimize overfitting by training the EfficientNetB3 model [50].

The data set was segmented into three subgroups, 80% for training and the remaining 20% for validation and testing (10% for validation and 10% for testing). At this stage, the data is reconsidered, and a seed is used for reproducibility, resulting in 2200 for training, 275 for validation, and 275 for testing. Through the function train.Labels. value_counts(): The number of images assigned to each category in the training set is determined: Healthy with 796 images, Moderate DR with 724

images, Mild DR with 290 images, Proliferate DR with 236 images, and Severe DR with 154 images. The parameters were then defined: 20 images were determined for each training cycle. Also, the images were resized to a size of 224×224 pixels, a standard size used in trained models such as EfficientNetB3.

A generator was used for the data-augmented training set, for which the zca_whitening function was used to improve the performance of the EfficientNetB3 model. Also, the rotation_range function was used, which allows randomly rotating images up to 30 degrees, which helps to create new variations of the training images. Multiple methods and functions are then used to extract class labels, such as Healthy, Moderate, Mild, Proliferate_DR, and Severe. As shown in Figure 2.

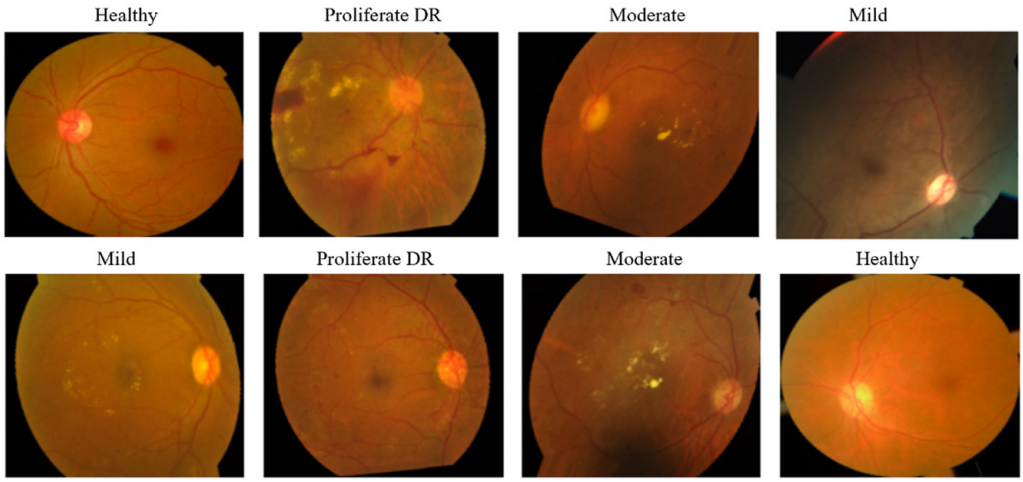

Fig. 2. Sample images from the dataset

Figure 2 shows a set of retinal images with different labels classified into healthy and presented in the upper left and lower right corners. This class represents the retina with no signs of DR. It shows a clean, clear retina with small pink spots indicating the optic disc and no apparent signs of hemorrhages. The Proliferative DR class appears in the second position in both the first and second rows. Advanced signs of proliferative DR can be seen in this class. The images show abnormal blood vessels and signs of hemorrhage indicating significant damage. The moderate class is in the third position in the two rows. A moderate case of DR is evident in this class. The images in this class show some yellow spots and small hemorrhages indicating vascular damage, but less severe than proliferative retinopathy. Finally, the mild class, which is in the fourth position in the first row and in the first position of the second row. In this class mild DR shows minor signs of damage with small dark spots showing early involvement of the retinal blood vessels.

It is important to note that, to minimize overfitting, dropout layers were incorporated into the final dense layers, along with data augmentation techniques such as rotations and normalization. In addition, the model was trained using the Adam optimizer and the categorical cross-entropy loss function, with the aim of analyzing the complete behavior of the learning process throughout the epochs.

3.3 Modeling

In this section, the transfer learning approach is implemented using the EfficientNetB3 model with pre-trained weights. To do so, we first proceed by

importing the necessary libraries such as TensorFlow to work with CNN and the EfficientNetB3 model. To this model we added our own output layers adapted to the specific task. A GlobalAveragePooling2D layer was used to reduce the dimensions, followed by a Dense layer. The EfficientNetB3 model is then compiled, using the Adam optimizer, and a categorical_crossentropy loss function, which is used for multi-class classification tasks. As shown in Figure 3.

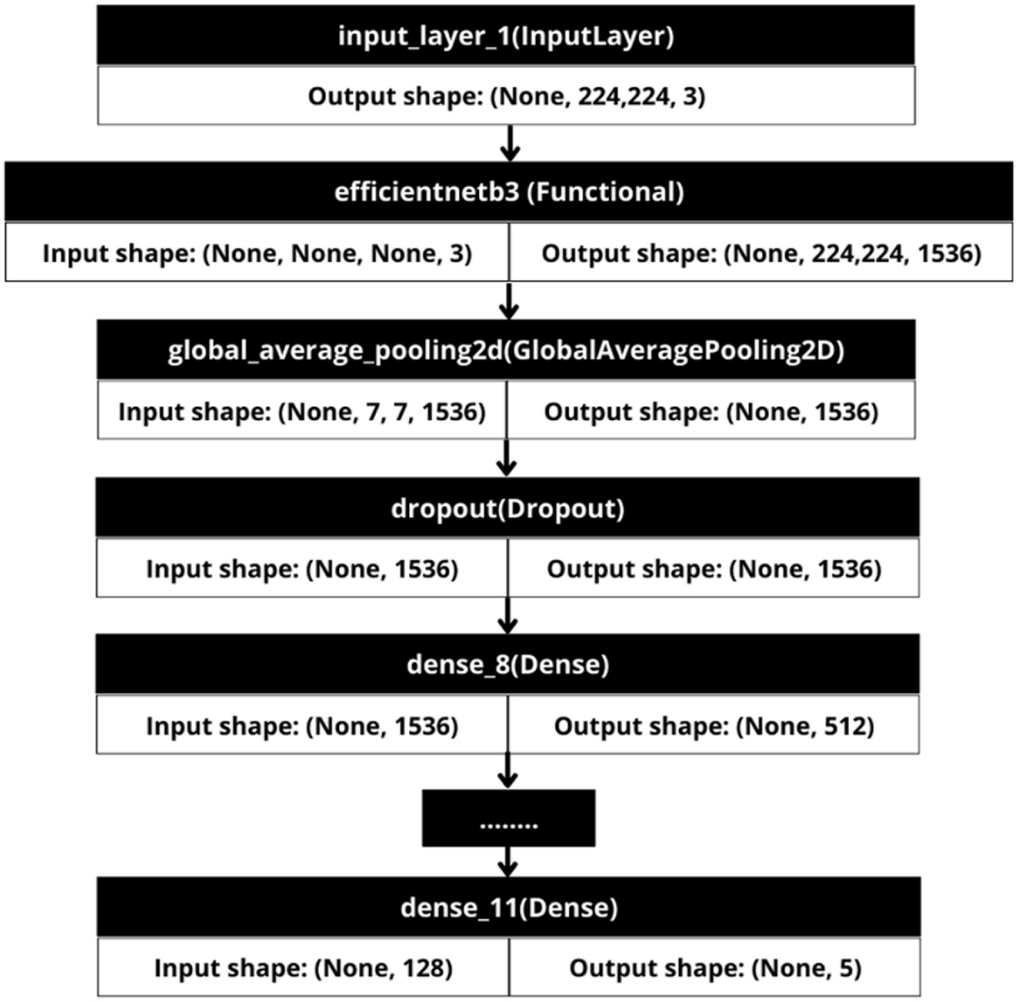

Fig. 3. EfficientNetB3 model architecture

3.4 Training

In this phase, the EfficientNetB3 model is trained using the training data set and a validation data set. The EfficientNetB3 model was trained with 50 epochs, and validations are performed after each epoch. For the initial setup, the following parameters were used: epochs, validation data, validation steps, and the model.Fit() function. In epoch 1: the model starts with a loss of 1.706 in the training set and a precision of 0.49. In the validation set, the loss is 1.532 and the accuracy is 0.465, and so on. At epoch 10, the model achieved an accuracy of 0.875 in the training set and an accuracy of 0.75 in the validation, with a low loss of 0.73; this means that the model is learning well with no signs of overfitting. At epoch 16: While the training

accuracy is improving to 0.772, the accuracy on the validation set is still at 0.738, no longer improving, and the validation loss has plateaued. This may mean that the EfficientNetB3 model is starting to be overfit. At time 50, the model reached a very high accuracy on the training set with 0.948, but the validation loss has increased to 0.044, and the accuracy on the validation set has decreased to 0.712, a clear sign that the EfficientNetB3 model has overfitted the training data. Overall, the model has learned well on the training data and is not generalizing well for the validation data. As can be seen in Figure 4.

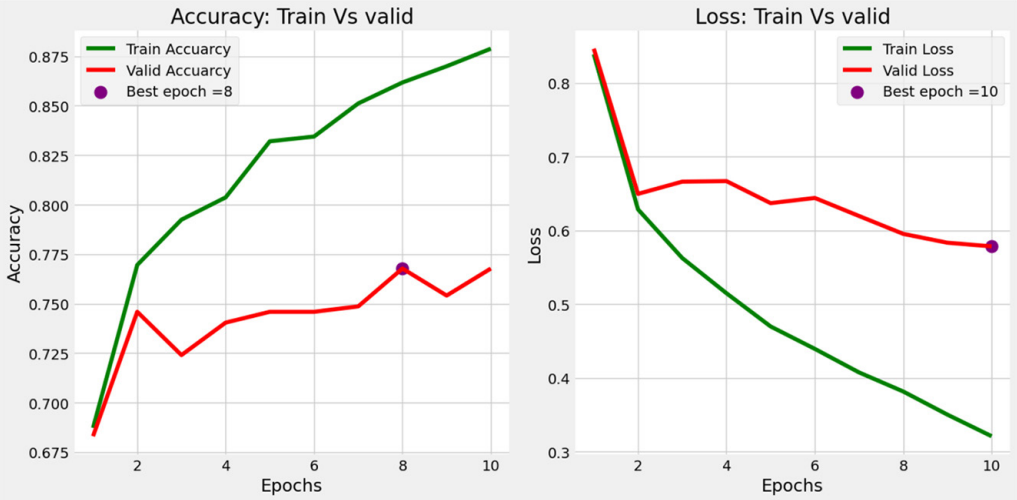

Fig. 4. Training and validation of the EfficientNetB3 model

Starting at Epoch 16, a difference between training and validation is observed, along with a progressive increase in validation loss, indicating clear model overfitting. Although training was extended to 50 epochs, the best balance between generalization and performance is achieved at approximately epoch 10.

Figure 4. It is composed of two graphs: on the left side, the accuracy, and on the right side, the loss. Analyzing the left-hand graph: the green curve represents the accuracy of the model on the training data. As can be seen, this curve follows an upward trend showing that the model improves in prediction as the number of epochs progresses. Achieving a high accuracy of .875. The red curve shows the accuracy of the model with the validation data. Unlike the training curve, the validation accuracy does not improve steadily. At the beginning there is a small decrease in accuracy, which is to be expected, as the model in these early epochs is adjusting. In the graph, it can be seen that from epoch 4, the validation accuracy fluctuates around 0.75, reaching its best point at epoch 8, as indicated by the purple circle.

In the graph on the right, the green curve represents the loss of the model with training data. As usual, the loss decreases as training progresses, indicating that the model is learning to reduce the error in the predictions. The curve stabilizes around a loss of 0.3 around epoch 10. The red curve represents the loss of the model with validation data. Like the accuracy graph, the loss in the validation data also has fluctuating behavior. Although initially, the loss decreases slightly until epoch 2, it then stabilizes and starts to increase slightly in the later epoch, which could be assumed to start to overfit. This happens when the model is fitting too closely to the training data and does not generalize well to the validation data. The graph indicates that the best epoch for validation loss is epoch 10, where it is marked by a purple circle.

Also, the confusion matrix was used to evaluate the performance of the EfficientNetB3 model in its different classes with the test data set. As shown in Figure 5, the vertical axis represents the true classes and the horizontal axis the classes predicted by the EfficientNetB3 model. The numbers on the main diagonal indicate the correct predictions, i.e., how many samples of a class were correctly classified. These values are indicative of good performance. Elements outside the diagonal represent incorrect predictions. The number at the intersection of an off-diagonal row and column is the number of times the model predicted the wrong class. For example, in the "Healthy" class, it correctly predicted 191 healthy retinal samples. Similarly, it classified 4 samples as wrong in that class. One sample is "Severe." Similarly, with the "Mild" class, where the model correctly classified 17 samples. Also, it misclassified 4 samples from "Mild" to "Healthy," 14 to "Moderate," and 1 to "Proliferate DR." In the "Moderate" class, the model correctly predicted 77 samples. It confounded 10 samples with "Moderate" and "Mil," 12 with "Proliferate DR," and 6 with "Severe." This can also be seen in Figure 4. The other class predictions, such as "Proliferate_DR" and "Severe". In general, the EfficientNetB3 model is more accurate in classifying the "Healthy" class with most of the correct predictions. The "Mild" and "Moderate" classes have considerable confusion with each other. This may be because, in these two classes, it is difficult to distinguish DR even for an ML model. Meanwhile, the more advanced classes, such as "Proliferate DR" and "Severe," also present confusion with each other and with the "Moderate" class.

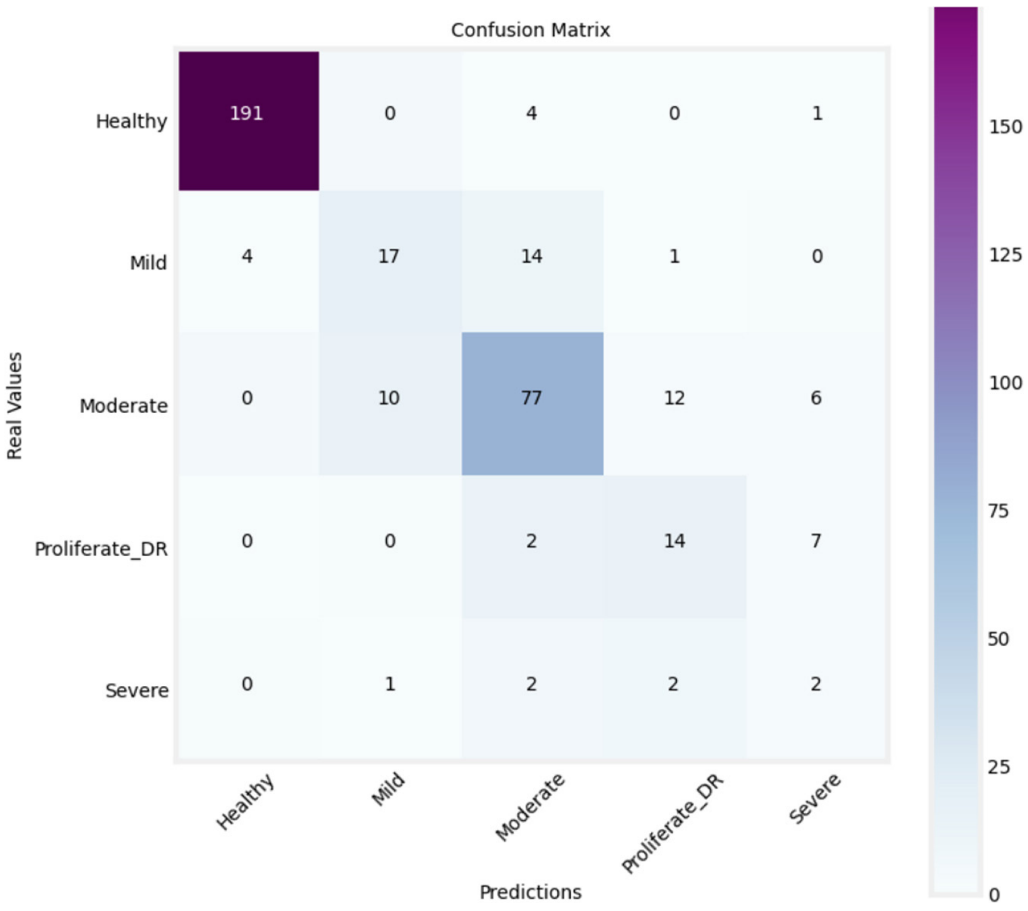

Fig. 5. Evaluation of the model with the confusion matrix

After training and validation, the model is evaluated with the three data sets: training, validation, and test. The accuracy of the model in the test set is somewhat better than in the validation set with 0.763, but still far from the performance in the training set. The loss in the test set of 1.108 is lower than the validation, which is a good sign, but still higher than the loss observed in the training set.

4 RESULTS AND DISCUSSION

This section presents the results obtained during the evaluation of the trained classification model for the detection of RD in the different categories. The model was evaluated with three datasets: training, validation, and test. The metrics evaluated include accuracy, recall, F1-score, support, and loss, both in the training set and in the validation and test sets.

In the Training section, high accuracy is evident in the training set, indicating that the model has learned to identify patterns in the images effectively. However, the results in the validation and test sets show a significant drop in performance, indicating a possible overfitting of the model. Table 1 presents a detailed classification to assess the performance of the model for each of the classes. While the model shows good performance in the classification of healthy cases, its accuracy and detection capability are considerably reduced in the less represented classes, such as the severe, Proliferate_DR classes, which could be inferred to be more biased towards the classes with more data.

To evaluate the performance of the model, metrics such as accuracy, recall, and F1-Score, which are represented in equations (1)–(3) [46], were used. True positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are terms used to calculate the metrics.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TN + FP} \tag{2}$$

$$F1 - Score = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \tag{3}$$

PTs indicate that healthy patients were correctly detected as healthy. NTs indicate that DR-affected patients were correctly identified as DR, while FPs indicate that healthy patients were erroneously detected as DR, and FNs indicate that DR patients were erroneously detected as healthy.

In this context, the accuracy metric is the ability of the classifier not to label a negative sample as positive. Recall represents the ability of the model to find all positive samples. The F1-score is the average between precision and recall. It is a metric that combines both and is useful when there are unbalanced classes. Support represents the number of true samples of each class in the test set. For example, the precision in the "Healthy" class is 0.97, which means that of all the samples that the model predicted were "Healthy." Recall the "healthy" class was 0.98, which shows that all the samples belonging to the "healthy" class were correctly identified. The F1-score of 0.98 means that the balance between accuracy and recall was excellent.

In the DR Mild class, an accuracy of 0.47 was obtained, meaning that, of all the predictions in the class, only 47% were correct. Recall was 0.61, indicating that the

model correctly identified 61% of the real samples in the "Mild" class. The F1-Score was 0.53, suggesting that there was an imbalance between accuracy and recall, indicating that the model has difficulties in correctly identifying the samples in the class, as the pressure is relatively low. However, with the "Moderate" class, it is evident that the precision improved by 0.73. The recall correctly found 0.78 samples. The F1-score with 0.75 indicates a good balance between precision and recall. The Proliferate_DR metrics indicate that the model has difficulties in correctly detecting samples, which could be related to the similarity of features with other advanced classes. The Severe DR Class indicates that the model has serious difficulties correctly classifying the class, possibly due to the lack of sufficient samples with other classes. In general terms, the "macro avg" takes the simple average of the accuracy, recall, and F1-score of each class without considering the number of samples. The "Weighted avg" considers the number of samples of each class, which makes the classes with the highest number of samples predominate, as is the case for the "Healthy" class, which has a significant performance. As can be seen in Table 1.

Table 1. Results of the evaluation of the classification model

	Precision	Recall	F1-score	Support
Healthy	0.97	0.98	0.98	195
Mild	0.47	0.61	0.53	28
Moderate	0.73	0.78	0.75	99
Proliferate DR	0.61	0.48	0.54	29
Severe	0.29	0.12	0.17	16
Accuracy			0.82	367
Macro avg	0.61	0.59	0.60	367
Weighted avg	0.81	0.82	0.81	367

Results obtained using the EfficientNetB3 model showed an overall accuracy of 0.81, with particularly high performance in the "Healthy" class of 0.97 and 0.98 recall, while the more advanced DR classes, such as "Severe" and "Proliferate_DR," showed much lower performances with 0.29 and 0.61 accuracy, respectively. These findings indicate that the model was able to accurately identify healthy retinas; however, it has difficulties in differentiating between the more advanced DR classes. These results align with observations made in other studies that have used CNNs for DR detection but also highlight important areas where improvements could be made.

Several studies related to this research topic have shown high levels of accuracy in DR classification using CNN architecture, obtaining significant results in contexts like this research. For example, the work [16] achieved an accuracy of 97.91% when combining CNN with a Random Forest approach, a method that could provide a direction for improving our model. While our EfficientNetB3 model showed competitive performance in certain classes, this higher accuracy indicates that combining other techniques, such as tree-based algorithms, can help improve performance. Another important work is that of [17], where they compared two CNN models and transfer learning on a highly imbalanced dataset, obtaining better results with the MobileNet model with 80% accuracy compared to MobileNetV2 with 71% accuracy. Our EfficientNetB3 model achieved 81% accuracy, indicating that lightweight architectures such as MobileNet or MobileNetv2 may be suitable for classification tasks

with imbalanced datasets. These studies highlight the importance of balance in the data, a key factor that could have affected our model's ability to correctly identify underrepresented classes such as "Severe" and "Proliferate_DR."

Research [18] demonstrated high effectiveness by combining multiple models, including Inception-V3, VGG16, and a custom CNN, achieving an accuracy of 95.06% and an AUC of 87.88%. Compared to the findings in our EfficientNetB3 model that achieved a weighted F1-score of 81%, it can be presumed that the overall performance could be improved by using a combination of deep learning architectures, as proposed in this work. Also, the use of the ResNet50 model in the study [19] achieved an accuracy of 92.9%, which reinforces the effectiveness of deep architecture for DR detection. Although the EfficientNetB3 model showed competitive performance, ResNet has been considerably a more robust model in this task, as also mentioned by [20], where EfficientNetB4 obtained an accuracy of 79.11%. This indicates that more advanced versions of EfficientNet, such as B4 or B7, can provide higher predictive power, especially in more complex classes such as the advanced stages of DR.

Regarding the handling of large data volumes, in the study [21], where they used a custom CNN with a dataset of more than 34k, they achieved accuracies of 91.78% and 97.27% on two different datasets. This shows that increasing the data volume and turning the models to work efficiently can lead to better results with respect to accuracy. Our work could implement data augmentation techniques and thereby use larger and more balanced datasets to reduce performance variabilities between classes. Finally, the work [22] presented a hybrid decision-making system based on CNN, RNN, and the hummingbird optimizer algorithm, where it achieved an accuracy of 97%. These hybrid approaches, which also combine different CNN techniques, could be useful to address problems of overfitting observed in small classes. These findings lead us to conclude that RNN or optimization algorithms can improve the ability of models to learn patterns. While it is true that our model, based on EfficientNetB3, showed good overall results, with an accuracy of 81%, previous studies have shown that models such as ResNet50, combinations of CNN architecture or additional techniques, and optimization methods can provide better results, especially in detecting more advanced DR classes.

The quality of the images is a relevant point to address, considering that they play an important role in the identification of features. This work achieved significant results despite not using optimization techniques to improve the quality of the images. The authors in the works [24] and [28] suggest the use of 3D visualization methods and optimization parameters to identify the DR. The use of these methods could have improved the quality of the images, and with it, the results. The findings of this work have a very close relationship to the results of the work [25], where they obtained 81% accuracy in the identification of the DR, supported with deep learning. This is due to the similarities of the dataset and the techniques used. However, in the work [27], the results obtained were higher than our findings; this could be due to the size of the dataset, considering that they used more than 46 thousand fundus images, which determines that deep learning responds efficiently for this type of task.

Although the average performance obtained in this study is 81%, which is lower than that reported in other studies (91–97%), this is mainly due to the small size of the dataset, the imbalance between classes, and the absence of advanced segmentation techniques. Therefore, the results should be interpreted as a feasibility assessment rather than a direct comparison of performance with studies in the literature review. It is important to note that class imbalance significantly affects model performance, particularly in the Severe and Proliferate DR classes. Although random

oversampling and data augmentation techniques were applied, they were not sufficient to compensate for the low representation of these categories.

It is recommended to implement more advanced techniques, such as SMOTE and class-weighted loss functions, which have proven to be more effective in similar imbalance scenarios and should be considered as future lines of work to improve sensitivity in clinical classes.

5 CONCLUSION

This work aimed to develop and evaluate a CNN model based on the EfficientNetB3 architecture capable of identifying the characteristic signs of DR in fundus images to provide a clinical diagnostic support tool for healthcare professionals. The results obtained were encouraging, showing a competitive overall performance, but also evidenced areas for improvement.

The EfficientNetB3 model achieved an accuracy of 81%, indicating a good ability to identify DR in retinal images, especially in the "Healthy" class. However, it was also observed that the model has difficulties in accurately classifying more advanced stages of DR, such as "Proliferative DR" and "Severe," indicating the need for additional adjustments in data processing. While the model achieved 97% accuracy and 98% recall in the "Healthy" class, the more advanced classes such as "Severe" performed poorly, with 29% accuracy and 12% recall. These metrics could be due to an imbalance in the dataset, where advanced DR classes are less represented. This finding is in line with some previous studies that have highlighted the importance of balancing classes to achieve improved accuracy across all classes.

One of the main challenges encountered was the unbalanced distribution of classes in the dataset, which negatively affected the model's ability to correctly identify underrepresented classes, such as "Severe." Studies such as [19], which used large balanced datasets, achieved accuracies of over 91%, highlighting the need to apply data augmentation techniques to improve class performance.

To improve the results obtained, it is recommended to explore the combination of CNN models with specialized architectures, such as the use of ResNet50 or hybrid learning techniques, which could help to reduce the overfitting problems observed in the more advanced classes. Also, it is suggested to implement regularization techniques such as dropout and hyperparameter optimization to improve the generalizability of the model.

With respect to the limitations observed in the more advanced classes, the model could be very helpful as a complementary tool for early detection of DR, especially to identify healthy retinas with high accuracy. However, for the model to be implemented in clinical practice, it is necessary to improve performance in the advanced stages of the disease, where accuracy and recall are essential for diagnosis.

6 REFERENCES

[1] O. Iparraguirre-Villanueva, K. Espinola-Linares, R. O. Flores Castañeda, and M. Cabanillas-Carbonell, "Application of machine learning models for early detection and accurate classification of type 2 diabetes," *Diagnostics 2023*, vol. 13, no. 14, p. 2383, 2023. https://doi.org/10.3390/diagnostics13142383

[2] M. A. dos Reis *et al.*, "Advancing healthcare with artificial intelligence: Diagnostic accuracy of machine learning algorithm in diagnosis of diabetic retinopathy in the Brazilian population," *Diabetology and Metabolic Syndrome*, vol. 16, no. 1, p. 209, 2024. https://doi.org/10.1186/s13098-024-01447-0

[3] IDF, "Diabetes around the world." Accessed: Sep. 23, 2024. [Online]. Available: https://diabetesatlas.org/

[4] WHO, "World Health Organization (WHO)." Accessed: Sep. 23, 2024. [Online]. Available: https://www.who.int/europe/activities/promoting-diabetic-retinopathy-screening

[5] Z. Liu *et al.*, "Cost-effectiveness of incorporating self-imaging optical coherence tomography into fundus photography-based diabetic retinopathy screening," *npj Digital Medicine*, vol. 7, no. 1, 2025. https://doi.org/10.1038/s41746-024-01222-5

[6] A. M. Mutawa, K. Al-Sabti, S. Raizada, and S. Sruthi, "A deep learning model for detecting diabetic retinopathy stages with discrete wavelet transform," *Applied Sciences (Switzerland)*, vol. 14, no. 11, p. 4428, 2024. https://doi.org/10.3390/app14114428

[7] M. Saxena, P. Narra, M. Saxena, and R. Saxena, "Deep learning ensemble framework for multiclass diabetic retinopathy classification," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 22, no. 3, pp. 665–672, 2024. https://doi.org/10.12928/telkomnika.v22i3.25794

[8] P. Kakati, S. G. Quek, G. Selvachandran, T. Senapati, and G. Chen, "Analysis and application of rectified complex t-spherical fuzzy Dombi-Choquet integral operators for diabetic retinopathy detection through fundus images," *Expert Syst. Appl.*, vol. 243, p. 122724, 2024. https://doi.org/10.1016/j.eswa.2023.122724

[9] N. Gharaibeh, O. M. Al-Hazaimeh, A. Abu-Ein, and K. M. O. Nahar, "A hybrid SVM NAÏVE-BAYES classifier for bright lesions recognition in eye fundus images," *International Journal on Electrical Engineering and Informatics*, vol. 13, no. 3, pp. 530–545, 2021. https://doi.org/10.15676/ijeei.2021.13.3.2

[10] X. Xu, D. Liu, G. Huang, M. Wang, M. Lei, and Y. Jia, "Computer aided diagnosis of diabetic retinopathy based on multi-view joint learning," *Comput. Biol. Med.*, vol. 174, p. 108428, 2024. https://doi.org/10.1016/j.compbiomed.2024.108428

[11] J. Tang *et al.*, "TSNet: Task-specific network for joint diabetic retinopathy grading and lesion segmentation of ultra-wide optical coherence tomography angiography images," *Visual Computer*, vol. 40, no. 9, pp. 5935–5946, 2024. https://doi.org/10.1007/s00371-023-03145-w

[12] Z. Hai *et al.*, "A novel approach for intelligent diagnosis and grading of diabetic retinopathy," *Comput. Biol. Med.*, vol. 172, p. 108246, 2024. https://doi.org/10.1016/j.compbiomed.2024.108246

[13] V. Guevara-Ponce, O. Roque-Paredes, C. Zerga-Morales, A. Flores-Huerta, M. Aymerich-Lau, and O. Iparraguirre-Villanueva, "Detection of breast cancer using convolutional neural networks with learning transfer mechanisms," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, pp. 574–580, 2023. https://doi.org/10.14569/IJACSA.2023.0140661

[14] M. V. Cicinelli *et al.*, "Assessing diabetic retinopathy staging with AI: A comparative analysis between pseudocolor and LED imaging," *Translational Vision Science & Technology*, vol. 13, no. 3, 2001. https://doi.org/10.1167/tvst.13.3.11

[15] Z. Wang, S. Chen, T. Liu, and B. Yao, "Multi-branching temporal convolutional network with tensor data completion for diabetic retinopathy prediction," *IEEE J. Biomed. Health Inform.*, vol. 28, no. 3, pp. 1704–1715, 2024. https://doi.org/10.1109/JBHI.2024.3351949

[16] X. N. Wang, Z. Guan, B. Qian, T. Chen, and Q. Wu, "A deep learning system for the detection of optic disc neovascularization in diabetic retinopathy using optical coherence tomography angiography images," *Visual Computer*, vol. 41, no. 2, pp. 1293–1302, 2024. https://doi.org/10.1007/s00371-024-03418-y

[17] N. Gharaibeh, O. M. Al-Hazaimeh, B. Al-Naami, and K. M. O. Nahar, "An effective image processing method for detection of diabetic retinopathy diseases from retinal fundus images," *International Journal of Signal and Imaging Systems Engineering*, vol. 11, no. 4, pp. 206–216, 2018. https://doi.org/10.1504/IJSISE.2018.093825

[18] V. Thanikachalam, K. Kabilan, and S. K. Erramchetty, "Optimized deep CNN for detection and classification of diabetic retinopathy and diabetic macular edema," *BMC Med. Imaging*, vol. 24, no. 1, p. 227, 2024. https://doi.org/10.1186/s12880-024-01406-1

[19] P. K. Das and S. Pumrin, "Diabetic retinopathy classification: Performance evaluation of pre-trained lightweight CNN using imbalance dataset," *Engineering Journal*, vol. 28, no. 7, pp. 13–25, 2024. https://doi.org/10.4186/ej.2024.28.7.13

[20] K. Nazir, J. Kim, and Y. C. Byun, "Enhancing early-stage diabetic retinopathy detection using a weighted ensemble of deep neural networks," *IEEE Access*, vol. 12, pp. 113565–113579, 2024. https://doi.org/10.1109/ACCESS.2024.3432867

[21] M. Dhouibi, A. K. Ben Salem, A. Saidi, and S. Ben Saoud, "Acceleration of convolutional neural network based diabetic retinopathy diagnosis system on field programmable gate array," *International Journal of Informatics and Communication Technology*, vol. 12, no. 3, pp. 214–224, 2023. https://doi.org/10.11591/ijict.v12i3.pp214-224

[22] D. Das, S. K. Biswas, and S. Bandyopadhyay, "Detection of Diabetic Retinopathy using Convolutional Neural Networks for Feature Extraction and Classification (DRFEC)," *Multimed. Tools Appl.*, vol. 82, no. 19, pp. 29943–30001, 2023. https://doi.org/10.1007/s11042-022-14165-4

[23] M. Nahiduzzaman *et al.*, "Diabetic retinopathy identification using parallel convolutional neural network based feature extractor and ELM classifier," *Expert Syst. Appl.*, vol. 217, p. 119557, 2023. https://doi.org/10.1016/j.eswa.2023.119557

[24] E. Dhiravidachelvi, S. Senthil Pandi, R. Prabavathi, and C. Bala Subramanian, "Artificial humming bird optimization–based hybrid CNN-RNN for accurate exudate classification from fundus images," *J. Digit. Imaging.*, vol. 36, no. 1, pp. 59–72, 2023. https://doi.org/10.1007/s10278-022-00707-7

[25] A. O. Asia *et al.*, "Detection of diabetic retinopathy in retinal fundus images using CNN classification models," *Electronics (Switzerland)*, vol. 11, no. 17, p. 2740, 2022. https://doi.org/10.3390/electronics11172740

[26] M. Li, Y. Jung, M. Fulham, and J. Kim, "Importance-aware 3D volume visualization for medical content-based image retrieval-a preliminary study," *Virtual Reality and Intelligent Hardware*, vol. 6, no. 1, pp. 71–81, 2024. https://doi.org/10.1016/j.vrih.2023.08.005

[27] J. Li *et al.*, "Integrated image-based deep learning and language models for primary diabetes care," *Nat. Med.*, vol. 30, pp. 2886–2896, 2024. https://doi.org/10.1038/s41591-024-03139-8

[28] M. A. Guerroudji, K. Amara, M. Lichouri, N. Zenati, and M. Masmoudi, "A 3D visualization-based augmented reality application for brain tumor segmentation," *Comput. Animat. Virtual Worlds*, vol. 35, no. 1, 2024. https://doi.org/10.1002/cav.2223

[29] L. Dai *et al.*, "A deep learning system for detecting diabetic retinopathy across the disease spectrum," *Nat. Commun.*, vol. 12, no. 1, p. 3242, 2021. https://doi.org/10.1038/s41467-021-23458-5

[30] R. Liu *et al.*, "DeepDRiD: Diabetic retinopathy—grading and image quality estimation challenge," *Patterns*, vol. 3, no. 6, p. 100512, 2022. https://doi.org/10.1016/j.patter.2022.100512

[31] M. M. Al-Nawashi, O. M. Al-Hazaimeh, and M. K. Khazaaleh, "A new approach for breast cancer detection-based machine learning technique," *Applied Computer Science*, vol. 20, no. 1, pp. 1–16, 2024. https://doi.org/10.35784/acs-2024-01

[32] O. M. Al-hazaimeh, A. A. Abu-Ein, N. M. Tahat, M. A. 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 (iJOE)*, vol. 18, no. 13, pp. 131–151, 2022. https://doi.org/10.3991/ijoe.v18i13.33985

[33] N. Gharaibeh, A. A. Abu-Ein, O. M. Al-hazaimeh, K. M. O. Nahar, W. A. Abu-Ain, and M. M. Al-Nawashi, "Swin transformer-based segmentation and multi-scale feature pyramid fusion module for Alzheimer's disease with machine learning," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 19, no. 4, pp. 22–50, 2023. https://doi.org/10.3991/ijoe.v19i04.37677

[34] V. Thanikachalam, K. Kabilan, and S. K. Erramchetty, "Optimized deep CNN for detection and classification of diabetic retinopathy and diabetic macular edema," *BMC Med Imaging*, vol. 24, no. 1, p. 227, 2024. https://doi.org/10.1186/s12880-024-01406-1

[35] S. Liu, W. Wang, L. Deng, and H. Xu, "CNN-trans model: A parallel dual-branch network for fundus image classification," *Biomed Signal Process Control*, vol. 96, p. 106621, 2024. https://doi.org/10.1016/j.bspc.2024.106621

[36] M. Saxena, P. Narra, M. Saxena, and R. Saxena, "Deep learning ensemble framework for multiclass diabetic retinopathy classification," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 22, no. 3, pp. 665–672, 2024. https://doi.org/10.12928/telkomnika.v22i3.25794

[37] J. Caicho *et al.*, "Diabetic retinopathy: Detection and classification using AlexNet, GoogleNet and ResNet50 convolutional neural networks," in *Communications in Computer and Information Science*, vol. 1532, CCIS, 2022, pp. 259–271. https://doi.org/10.1007/978-3-030-99170-8_19

[38] F. M. J. Mehedi Shamrat *et al.*, "An advanced deep neural network for fundus image analysis and enhancing diabetic retinopathy detection," *Healthcare Analytics*, vol. 5, p. 100303, 2024. https://doi.org/10.1016/j.health.2024.100303

[39] Z. Hai *et al.*, "A novel approach for intelligent diagnosis and grading of diabetic retinopathy," *Comput. Biol. Med.*, vol. 172, p. 108246, 2024. https://doi.org/10.1016/j.compbiomed.2024.108246

[40] C. X. Ling and C. Li, "Data mining for direct marketing: Problems and," Accessed: Oct. 11, 2024. [Online]. Available: www.aaai.org

[41] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big. Data*, vol. 6, no. 1, pp. 1–48, 2019. https://doi.org/10.1186/s40537-019-0197-0

[42] R. Van Den Goorbergh, M. Van Smeden, D. Timmerman, and Ben Van Calster, "The harm of class imbalance corrections for risk prediction models: Illustration and simulation using logistic regression," *Journal of the American Medical Informatics Association*, vol. 29, no. 9, pp. 1525–1534, 2022. https://doi.org/10.1093/jamia/ocac093

[43] Phani, "Introduction to convolutional neural network — analytics vidhya." Accessed: Oct. 10, 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/10/what-is-the-convolutional-neural-network-architecture/

[44] Y. Yang, Z. Cai, S. Qiu, and P. Xu, "Vision transformer with masked autoencoders for referable diabetic retinopathy classification based on large-size retina image," *PLoS ONE*, vol. 19, no. 3, p. e0299265, 2024. https://doi.org/10.1371/journal.pone.0299265

[45] P. Macsik, J. Pavlovicova, S. Kajan, J. Goga, and V. Kurilova, "Image preprocessing-based ensemble deep learning classification of diabetic retinopathy," *IET Image Process*, vol. 18, no. 3, pp. 807–828, 2024. https://doi.org/10.1049/ipr2.12987

[46] R. Gayathri, S. Karthikeyan, T. Kausheekraj, and S. Keerthana, "Diabetic retinopathy detection using CNN with Res-LSTMDN," in *Proceedings of the 2024 10th International Conference on Communication and Signal Processing, ICCSP 2024*, 2024, pp. 254–259. https://doi.org/10.1109/ICCSP60870.2024.10543841

[47] Aryan, R. Chaudhuri, and S. Deb, "Precise lesion analysis to detect diabetic retinopathy using Generative Adversarial Network (GAN) and Mask-RCNN," *Procedia Comput. Sci.*, vol. 235, pp. 520–529, 2024. https://doi.org/10.1016/j.procs.2024.04.051

[48] S. R. Rath, "Diabetic retinopathy 224×224 (2019 Data)." Accessed: Oct. 13, 2024. [Online]. Available: https://www.kaggle.com/datasets/sovitrath/diabetic-retinopathy-224x224-2019-data

[49] R. Gunawan, H. A. Ghani, N. Khamis, and H. F. Adah Amran, "Imbalanced data handling in multiclass distributed denial of service attack detection using deep learning," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 22, no. 6, pp. 1396–1404, 2024. https://doi.org/10.12928/telkomnika.v22i6.26005

[50] F. Muftic, M. Kadunic, A. Musinbegovic, A. A. Almisreb, and H. Jaafar, "Deep learning for magnetic resonance imaging brain tumor detection: Evaluating ResNet, EfficientNet, and VGG-19," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 6, pp. 6360–6372, 2024. https://doi.org/10.11591/ijece.v14i6.pp6360-6372

7 AUTHORS

Orlando Iparraguirre-Villanueva is a systems engineer with a master's degree in information technology management and a PhD in systems engineering from the Universidad Nacional Federico Villarreal, Peru; certified in ITIL®, with specialization in Business Continuity Management (SBCM) and SCRUM certification; Lecturer and panelist in national and international events, with extensive experience in undergraduate and graduate teaching in various universities in the country (E-mail: c27399@utp.edu.pe).

Michael Cabanillas-Carbonell is a systems engineer with a PhD in systems and telecommunications engineering from the Polytechnic University of Madrid He is a Research Professor at the Facultad de Ingeniería, Universidad Privada del Norte. President of the IEEE Peru EIRCON International Engineering Research Conference. He is an International speaker specializing in software development, artificial intelligence, machine learning, business intelligence, and augmented reality. He has also authored more than 100 scientific articles indexed in IEEE Xplore, Scopus, and WoS (E-mail: alejandro.cabanillas@upn.edu.pe).