Discover Computing

RESEARCH Open Access

Automated seven-stage diabetic retinopathy grading using optimized deep networks through systematic hyperparameter tuning

Yavuz Unal[1*]

*Correspondence:
Yavuz Unal
yunal@sinop.edu.tr
[1]Department of Computer Engineering, Sinop University, Sinop, Turkey

Abstract

Diabetes is a systemic disease that can lead to various pathological changes in the eye. One of these changes, diabetic retinopathy, results from damage to the retinal tissue responsible for light perception and image transmission. High blood glucose levels can damage retinal capillaries, as in other vascular structures, thereby impairing visual function. Early detection of diabetic retinopathy is essential for effective treatment. Accurate classification of disease stages directly affects treatment timing and effectiveness, thereby improving prognosis. Early and accurate staging significantly contributes to preventing disease progression, reducing the risk of complications, and preserving patients' quality of life. In this study, a comprehensive deep learning-based approach was developed to classify the seven stages of diabetic retinopathy using fundus images. The proposed methodology integrates CLAHE (Contrast Limited Adaptive Histogram Equalization) preprocessing to enhance image contrast and reveal subtle pathological features, followed by systematic data augmentation to improve dataset diversity and model generalization. The Optuna optimization framework was employed to systematically identify the best-performing pre-trained deep learning architecture, which was determined to be NASNetLarge. Hyperparameter tuning using Grid Search optimization further refined the model configuration, achieving a classification accuracy of 98.39% with 5-fold cross-validation, demonstrating robust performance ($98.50 \pm 0.21\%$). Furthermore, saliency map visualization confirmed that the model focuses on clinically relevant anatomical structures and pathological features, such as microaneurysms, hemorrhages, and vascular abnormalities, thereby enhancing interpretability and trustworthiness for potential clinical deployment. The results demonstrate that the proposed preprocessing, systematic model selection, and hyperparameter optimization techniques significantly improve classification performance and clinical reliability in diabetic retinopathy diagnosis.

Keywords Diabetic retinopathy, Fundus images, Hyperparameter optimization, Explainable AI

1 Introduction

Diabetic retinopathy (DR) is a diabetes-related condition that causes permanent vision problems. It occurs when diabetes damages the eye's blood vessels, thereby injuring the tissues supplied by those vessels. Long-term hyperglycemia damages the blood vessels in the eye. Diabetes is an ailment that does not show symptoms in its early stages. Following a diabetes diagnosis, eye examinations should be performed regularly. However, only in this way can early-stage diabetic retinopathy be detected [1, 2]. As the disease progresses, blurred vision, floaters, and dark spots in the visual field may develop. It is known that there are millions of diabetes patients in the world. The number of patients with diabetes will increase exponentially in the coming decades. In many of these patients, the development of diabetic retinopathy is inevitable [3]. Diabetic Retinopathy is classified into various stages. These stages are determined based on the severity of the disease and its effects on the retina. Diabetic retinopathy stages are used to characterize disease progression and to guide appropriate treatment strategies [4]. Diabetic Retinopathy (DR) is a severe ocular condition that can advance and result in irreversible visual impairment if not identified promptly. Professional ophthalmologists often conduct conventional diagnostic techniques; despite their incredible accuracy, this approach is both time-intensive and expensive. Moreover, the subjective components of diagnosis may lead to varying interpretations among physicians. This circumstance, particularly in regions with a large patient population, delays the diagnostic process and reduces opportunities for early intervention. Consequently, the swift and accurate identification of diabetic retinopathy is crucial for patients to maintain their vision [5].

Deep learning methodologies have achieved considerable progress in medical image analysis in recent years [6]. Convolutional neural networks (CNNs) and other advanced artificial intelligence models have attained a level of proficiency that enables them to match or exceed human performance in the automated processing, classification, and detection of anomalies in medical images [7, 8]. CNN-based approaches have made particularly significant contributions to the diagnosis of diabetic retinopathy (DR) by analyzing fundus images to detect disease stages based on intraocular vascular abnormalities [9]. Unlike traditional methods, deep learning approaches can process large volumes of fundus images end-to-end in an automated, fast, and objective manner, offering a potential strategy to prevent vision loss through early detection of DR [10].

The accuracy of deep learning models depends on the quality and processing of the input images [11]. Therefore, the Contrast Limited Adaptive Histogram Equalization (CLAHE) method was applied during the preprocessing stage to improve image contrast and obtain distinct features. CLAHE is an effective method for enhancing detail, particularly in low-contrast medical images [12]. Unlike standard histogram equalization or traditional data augmentation methods commonly used in the literature, the proposed study employs the Contrast Limited Adaptive Histogram Equalization (CLAHE) method to achieve greater contrast enhancement in low-contrast diabetic retinopathy images. In this way, the model is expected to learn finer details more effectively. In addition, an automated, systematic approach has been adopted for model selection and hyperparameter optimization, distinct from the manual methods commonly used in the literature. In most studies, model selection is done by manually evaluating a few specific deep learning architectures. In the proposed method, Optuna-based model selection is employed to determine the optimal model configuration across a wide hyperparameter

space, using strategies such as Bayesian optimization and early stopping. Thus, instead of manual trials, the most suitable model for the dataset is automatically found. In existing methods, hyperparameter optimization is typically performed experimentally or via random search, whereas in the proposed method, parameter combinations are evaluated using Grid Search as the search strategy. This enables the acquisition of more stable and reliable hyperparameter settings, particularly with limited datasets.

This study presents several key contributions to automated diabetic retinopathy diagnosis. Unlike most existing studies that focus on binary or five-stage classification, this work implements a clinically relevant seven-stage classification system that distinguishes between all NPDR sub-stages and PDR progression levels, providing more granular diagnostic information for clinical decision-making. The methodology introduces a systematic two-stage optimization framework that combines Optuna-based Bayesian architecture search with grid search for hyperparameter tuning, providing a reproducible approach to optimal model selection. An extensive data augmentation pipeline specifically tailored for fundus images addresses severe class imbalance while preserving pathological feature integrity. The proposed framework ensures robust generalizability through 5-fold cross-validation and provides clinical interpretability through Grad-CAM visualizations, enabling clinicians to understand and trust model decisions. Finally, the optimized NASNetLarge model achieves 98.39% accuracy, outperforming existing approaches while maintaining balanced performance across all severity levels.

In the literature review section, studies on the use of various deep learning approaches for classifying diabetic retinopathy are discussed. The theoretical foundations of the study's dataset and the proposed methods are detailed in the methodology section. The experimental results are presented in the results section, and the recommendations and evaluations for future studies are provided in the discussion and conclusion sections.

2 Literature review

Research on deep learning methodologies for diagnosing diabetic retinopathy has yielded substantial progress in computer vision and medical image analysis. This section will assess current methodologies in the subject by analyzing the techniques documented in the literature, the datasets employed, and the efficacy of deep learning models.

Li et al. [13] developed the Cross-Disease Attention Network (CANet) method for the classification of Diabetic Retinopathy (DR) and Diabetic Macular Edema (DME). They achieved 85% classification accuracy with the model they proposed.

Sajid et al. [14] proposed an automated system called DR-NASNet for classifying diabetic retinopathy (DR) severity into five stages (Normal, Moderate, Mild, Proliferative DR, Severe DR). The method combines Ben Graham and CLAHE-based preprocessing, data augmentation to address class imbalance, dense blocks integrated into a pretrained NASNet architecture, and a linear SVM classifier, achieving 96.05% accuracy on a challenging DR dataset while reducing model size and complexity.

Arora et al. [15] have proposed a deep learning-based model for the diagnosis of diabetic retinopathy (DR). The proposed model aims to classify DR severity levels from retinal images using the EfficientNetB0 architecture. In the study, the "Diagnosis of Diabetic Retinopathy" dataset from Kaggle was used, comprising 35,108 retinal images and covering five DR severity levels. As a result of the experiments conducted, the proposed model achieved 86.53% classification accuracy and a 0.5663 loss rate.

M et al. [16] performed feature extraction on preprocessed (resized and denoised) images from the 5-class IDRiD dataset and used the extracted features as input to the attention mechanism. They used DenseNet-121 and ResNet-50 and achieved the highest accuracy of 90.48% with DenseNet-121.

Shakibania et al. [4] proposed a system that uses pre-trained ResNet-50 and EfficientNet-B0 as dual branches. Using the proposed system, they achieved a classification success rate of 93% on fundus images comprising five stages.

S. N et al. [17] in a dataset consisting of 1065 fundus images from the DR class, analyzed the performance of deep learning models such as DenseNet121, InceptionV3, AlexNet, and ResNet for the detection and classification of diabetic retinopathy. The models were evaluated on retinal fundus images, and DenseNet121 achieved the highest performance with a classification accuracy of 98%.

Singh and Dobhal [18] proposed a new system for classifying diabetic retinopathy (DR) stages. Fundus images obtained from India were analyzed and classified into four categories: Microaneurysms (MA), Soft Exudates (SE), Hard Exudates (EX), and Hemorrhages (HE). For feature extraction, the VGG16 architecture was used, and for classification, Logistic Regression was employed. The proposed system achieved an accuracy of 90.4%, improving upon existing models.

Romero-Oraa et al. [19] proposed a new deep learning-based framework for the automatic classification of diabetic retinopathy (DR) stages. The proposed method uses an attention mechanism that separately focuses on red lesions (microaneurysms and hemorrhages) and bright lesions (hard exudates) in retinal images. This mechanism generates attention maps that are independently explainable using only image-level labels. The method extracts features using the Xception architecture and employs focal loss to address data imbalance. In evaluations on the Kaggle DR dataset, the model achieved an accuracy of 83.7% and a quadratic weighted kappa of 0.78.

Suo et al. [20] proposed the CS-ResNet-101 model and the Fully Convolutional Spatial Attention Module (FCSAM) for the classification of diabetic retinopathy (DR). Additionally, an image enhancement algorithm was developed to improve DR image quality, and the cross-entropy loss function was modified to reduce overfitting. The EyePACS dataset was used for training and testing, and the APTOS 2019 dataset for external validation. The results show that the proposed model achieved 98.1% accuracy, 99.6% specificity, 98.1% sensitivity, and an F1 score of 98.1%.

Khan et al. [1], in the model they proposed for diabetic retinopathy classification, applied advanced data augmentation techniques to a 7-class dataset and classified this dataset using deep learning architectures such as VGG16-TL, VGG19, AlexNet, and SqueezeNet. As a result of the study, they achieved an accuracy rate of 96.95% with the VGG16-TL architecture, demonstrating higher performance compared to other architectures.

Vij and Arora [21] developed Deep Inductive Transfer Learning Based Diagnostic Systems (DITL-DS) for Multiclass DR Severity Classification. They used a deep inductive transfer learning-based (DITL) model (InceptionV3, ResNet34, EfficientNet B0, VGG16, and Xception) and achieved 98.40% classification accuracy.

Hassan and Ismail [22] developed a convolutional neural network (CNN)-based model for classifying diabetic retinopathy (DR), and Bayesian CNN methods have been applied to this model. Bayesian methods, such as Variational Inference (VI) and Monte Carlo

Dropout (MC-dropout), were employed. The proposed models were evaluated on the APTOS 2019 and Messidor-2 datasets, achieving accuracy rates of 94.70% and 77.00% (CNN), 94.00% and 86.00% (BCNN-VI), and 93.30% and 81.00% (BCNN-MC-dropout), respectively.

Priya and Kumar [23] propose DRD-MHSAGGCN-BCOA, a novel method for detecting Diabetic Retinopathy (DR) using Multi-Head Self-Attention-based Gated Graph Convolutional Networks and a Binary Chimp Optimization Algorithm. Applied to the DIARETDB1 dataset, the method preprocesses images, segments abnormalities, and extracts features using advanced techniques. MHSAGGCN classifies DR stages (normal, mild, moderate, severe). Evaluated on metrics such as accuracy and F1-score, the method outperforms existing approaches, achieving 23.34%-34.33% higher accuracy and 33.21%-66.67% higher F1-score.

Shoaib et al. [24] proposed an innovative method that uses deep learning and transfer learning to diagnose Diabetic Retinopathy (DR). They performed data augmentation using the DiaGAN model and improved the DiaGAN-CNN model by fine-tuning InceptionResNetV2 and InceptionV3. In the evaluations conducted on the ODIR dataset, the model achieved 98% accuracy and 98% precision, surpassing existing methods.

Gargi et al. [25] proposed a new method based on deep learning for the classification of diabetic retinopathy severity. Fundus images from the APTOS-2019 and DDR datasets were preprocessed using a contrast-enhanced Non-local Means Gaussian filter to reduce noise and enhance contrast. For feature extraction, they employed the Excess Level Attention Axial Spectral Transformer, and for feature selection, the Dual-Phase Sine Equilibrium Optimization Algorithm. They classified DR severity using the Compact Pyramidal Dense Mixed Attention Network (CPDenA) model. The model achieved 97.95% accuracy on the APTOS-2019 dataset and 98.59% accuracy on the DDR dataset.

3 Materials and methods

In this study, the model developed to classify the seven stages of diabetic retinopathy (DR) from fundus images is presented in detail. The proposed model includes image processing and deep learning techniques. Additionally, this section explains the dataset used, pre-processing and optimization steps, the model architecture, and performance metrics.

3.1 Dataset

The study used the Diabetic Retinopathy Public Dataset [19], which was created to detect the seven stages of diabetic retinopathy from fundus images. The images in the dataset are 2124 × 2056 pixels and are in JPG format, with RGB color space. Figure 1 presents some examples of the original fundus images included in the dataset.

In the dataset, seven stages representing different severity levels of DR have been defined. These data, along with the number of images they contain, are provided in Table 1. Each stage comprises a different number of images, reflecting the progression of the disease and retinal changes.

The dataset contains seven different stages of diabetic retinopathy (DR). In grade 1 DR, there are no retinal abnormalities. Grade 2 DR is referred to as Mild NPDR and is characterized by only microaneurysms (small balloon-like swellings in blood vessels). Grade 3 DR is defined as Moderate NPDR and includes hemorrhages and mild vascular

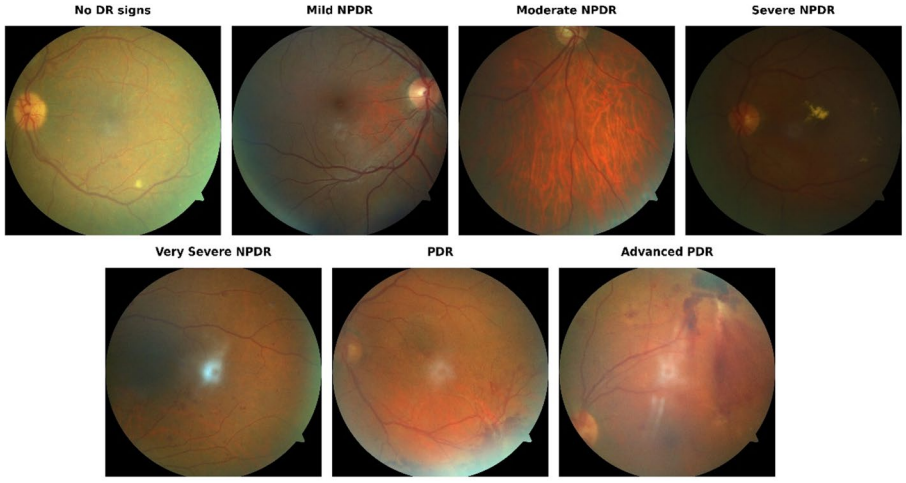

Fig. 1 Fundus image DR sample images are included in the dataset

Table 1 Augmented 7-stage dataset distribution of classes

Classes	DR stages	DR 7 Stages (Version 3)
1-Stage	No DR signs	711
2-Stages	Mild (or early) NPDR	6
3-Stages	Moderate NPDR	110
4-Stages	Severe NPDR	210
5-Stages	Very severe NPDR	139
6-Stages	PDR	116
7-Stages	Advanced NPDR	145
Total		1437

Table 2 Data augmentation parameters

Augmentation	Parameter name	Value ranges
Horizontal shift	X-axis translation	[-30,30]
Vertical shift	Y-axis translation	[-30,30]
Horizontal scaling	X-axis scaling	[0.7,1.3]
Vertical scaling	Y-axis scaling	[0.7,1.3]
Image rotation	image rotation	[-30,30]
Horizontal reflection	Reflect X-axis	True/False
Vertical reflection	Reflect Y-axis	True/False

anomalies. Grade 4 DR is known as Severe NPDR and is characterized by serious retinal hemorrhages and venous (vein) abnormalities. Grade 5 DR is defined as Very Severe NPDR and causes serious retinal damage. Grade 6 DR is called Proliferative DR (PDR) and includes advanced symptoms such as neovascularization (the formation of new blood vessels) and bleeding. Grade 7 DR is the most advanced and dangerous stage. At this stage, extensive neovascularization and severe damage to the retina are observed.

3.2 Data augmentation

Data augmentation is a technique that artificially diversifies training data to enhance the model's generalization [26]. The original dataset used in the study (version V3) contains 1437 images. These images were augmented using the data augmentation techniques and related parameters specified in Table 2.

In Table 2, the horizontal and vertical shift operations translate the images by ± 30 pixels along the X and Y axes, respectively. Horizontal and vertical scaling operations generate inputs of different sizes by randomly varying the image's scale factors along the X and Y axes between 0.7 and 1.3. The image rotation process allows the visuals to be viewed from different angles by rotating them between − 30 and 30 degrees. Additionally, horizontal and vertical reflection operations mirror images about the X- and Y-axes, respectively. These reflection operations determine whether they are applied by returning "True" or "False". These data augmentation techniques help prevent overfitting by allowing the model to see a greater variety of examples than those in the training data.

The augmentation process increased the number of images from 1437 to 15,879. Table 3 presents the number of duplications from which the DR stages were derived.

The number of images for all classes except the Mild (or early-stage) NPDR class has increased ninefold. In the original dataset, there are only six images in the Mild (or early-stage) NPDR class. Therefore, the data augmentation rate for this class differs from the other classes.

3.3 Image pre-processing

Image preprocessing encompasses various techniques and operations in image processing and computer vision that enhance raw image data for analysis or model training. This phase aims to improve image quality, diminish noise, optimize size, or accentuate specific features. Image preprocessing is an essential preliminary step in improving the performance of machine learning models [27].

3.4 Contrast-limited adaptive histogram equalization (CLAHE)

In image processing, a contrast enhancement technique called CLAHE (Contrast Limited Adaptive Histogram Equalization) can be used. The Adaptive Histogram Equalization (AHE) method was developed specifically to enhance detail in low-contrast images. MRI is used to see details better in low-contrast medical images, like X-rays [28, 29]. In Fig. 2, fundus images pre-processed with contrast-limited adaptive histogram equalization (CLAHE) are presented.

In this study, preprocessing was performed using the Contrast-Limited Adaptive Histogram Equalization (CLAHE) method to improve the classification performance of fundus images. As demonstrated in Fig. 2, this method increases local image contrast by dividing the image into 8 × 8-pixel tiles and applying histogram equalization independently to each tile, thereby making details in low-contrast regions more pronounced. Unlike global histogram equalization, CLAHE prevents over-amplification by limiting the contrast enhancement within each tile, preserving the natural appearance of retinal

Table 3 Augmented images and classes

Classes	Augmented images count
1-Stage	6399
2-Stages	3000
3-Stages	990
4-Stages	1890
5-Stages	1251
6-Stages	1044
7-Stages	1305
Total	15.879

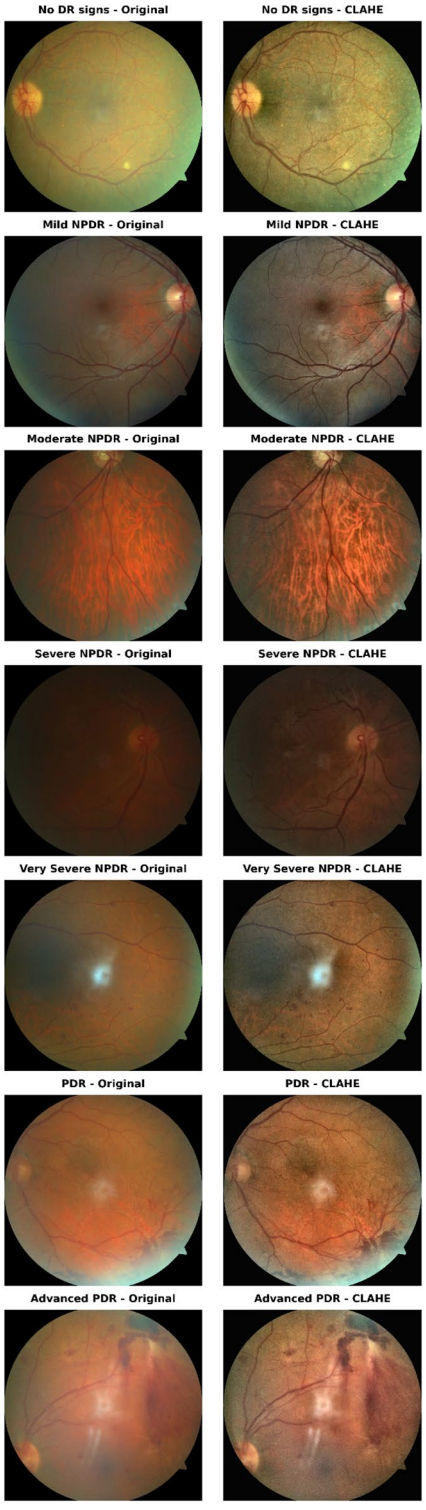

Fig. 2 Comparison of original (left) and CLAHE-preprocessed (right) fundus images across seven DR severity stages: **a** No DR signs, **b** Mild NPDR, **c** Moderate NPDR, **d** Severe NPDR, **e** Very Severe NPDR, **f** PDR, **g** Advanced PDR

structures while significantly improving the visibility of pathological features such as microaneurysms, hemorrhages, and exudates. This preprocessing step is essential for enabling the classification model to learn important visual features more effectively and to increase diagnostic accuracy, particularly for early-stage DR detection where lesions are subtle and difficult to identify.

3.5 Optuna software framework

Optuna is an open-source Python library developed by Akiba and colleagues for hyperparameter optimization. Using the Define-by-Run principle, it performs hyperparameter search in a flexible, efficient, and automatic manner. It supports dynamic creation and modification of the hyperparameter search space during the experiment, early termination of unnecessary trials, and operation on large-scale systems. Additionally, it can determine which pre-trained networks yield the best results on a dataset by evaluating them on that dataset [30].

In this study, using Optuna, we determined which of the nine pre-trained networks achieves the best classification performance on the given dataset. The evaluated pre-trained models are ResNet50, DenseNet121, EfficientNetB0, Xception, InceptionV3, VGG16, MobileNetV2, NASNetLarge, and InceptionResNetV2. The analyses conducted using Optuna-based optimization identified that the NASNetLarge architecture achieved the highest classification performance.

3.6 Two-stage optimization strategy

This study employs a systematic two-stage optimization strategy to achieve optimal model performance. In the first stage, Optuna is used to select the high-level architecture and to explore feasibility across nine pre-trained CNN architectures (ResNet50, DenseNet121, EfficientNetB0, Xception, InceptionV3, VGG16, MobileNetV2, NASNetLarge, and InceptionResNetV2). Optuna's Bayesian optimization approach efficiently explores the architecture space by learning from previous trials and adaptively selecting promising candidates, making it highly suitable for comparing fundamentally different model architectures. Once the optimal architecture (NASNetLarge) is identified via Optuna, the second stage applies grid search to perform deterministic, fine-grained hyperparameter tuning of the selected model. While Optuna excels at broad exploration across diverse architectures, Grid Search provides exhaustive evaluation of specific hyperparameter combinations within a defined search space, ensuring that no potentially optimal configuration is overlooked for the final model. This sequential approach combines the efficiency of Bayesian optimization for architecture search with the comprehensiveness of Grid Search for final hyperparameter refinement, thereby maximizing both computational efficiency and model performance.

3.7 Hyperparameters tuning and optimization

The Grid Search algorithm has been used to improve the performance of the proposed model. Grid Search is a commonly used systematic approach for hyperparameter optimization. This algorithm aims to determine the hyperparameter combination that yields the best performance by systematically scanning the hyperparameter space of a specific model. Hyperparameters are parameters that are predefined during model training and directly affect the model's learning behavior. Grid Search tests all possible combinations

of these parameters and selects the most suitable one in terms of a specific performance metric [31]. The algorithm operates on a hyperparameter grid defined by the user. This grid contains a list of possible values for each hyperparameter. Grid Search evaluates the model's performance for each combination by sequentially trying all possible combinations of these values using the cross-validation method. Cross-validation divides the dataset into multiple subsets to more reliably measure the model's generalization ability and performs training and testing operations on each subset. In this way, the consistency and performance of the model on different datasets can be measured more clearly [32]. The advantage of Grid Search is that it has a high probability of finding the best hyperparameter combination by conducting a comprehensive search. However, this method can be computationally expensive, especially when the hyperparameter space is large or the model training is time-consuming. In this case, alternative methods such as Random Search or more advanced optimization techniques (for example, Bayesian Optimization) can be used. However, Grid Search offers a simple and effective solution, especially in cases where the hyperparameter space is relatively small [33].

The hyperparameters used in the optimization process of the study are presented in Table 4.

Dropout Rate is a technique used to prevent overfitting and varies in steps of 0.1 between a minimum of 0.3 and a maximum of 0.7. Batch Size determines the number of data samples to be used during training and has been selected as 16, 32, and 64. Optimizer refers to the algorithm used to update the model's weights, and options such as Adam, SGD, and RMSprop have been offered. Learning Rate is an important hyperparameter that determines the model's learning rate, and it was tested with three values: 0.01 (1e-2), 0.001 (1e-3), and 0.0001 (1e-4). Dense Units determine the number of neurons in the fully connected layer and are set to values ranging from 64 to 512 in steps of 64. Finally, Freeze Layers refer to the training or freezing of specific layers during the transfer learning process. It is set to a minimum of 0 and a maximum of the total number of layers in the model, and it can be changed in increments of 1. This setting is important for determining the extent to which transfer learning will be applied.

3.8 Proposed model

In this section, the proposed model for classifying diabetic retinopathy is presented. The block diagram of the model is presented in Fig. 3. The model uses a DR dataset consisting of seven stages as input. The CLAHE (Contrast Limited Adaptive Histogram Equalization) method is applied to the input images to enhance contrast, thereby making the details in the retina images more pronounced. The Optuna-based model selection mechanism identifies the model with the highest classification performance by comparing different deep learning architectures. Among the pre-trained models evaluated in

Table 4 Determined Values for Hyperparameter Optimization

Hyperparameter	Values
Dropout Rate	min value = 0.3, max value = 0.7 step = 0.1
Batch Size	16,32,64
Optimizer	Adam, SGD, RmsProp
Learning Rate	1e-2, 1e-3, 1e-4
Dense Units	min value = 64, max value = 512, step = 64
Freeze Layers	min value = 0, max value= total number of layers in the model, step = 1

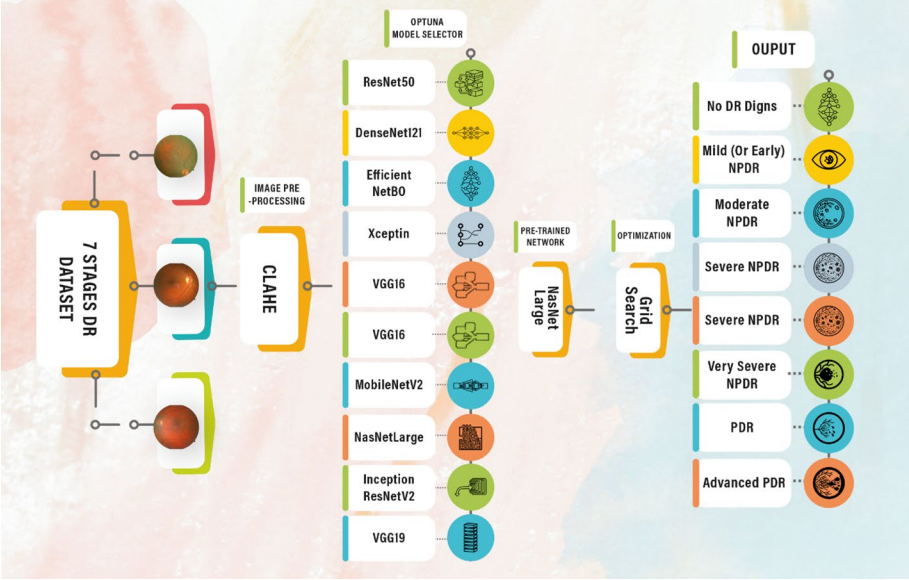

Fig. 3 Proposed Model

Table 5 Per-Class Performance Metrics Based on Average Confusion Matrix

DR stage	Precision	Recall	F1-score
No DR signs	0.9992	1.0000	0.9996
Mild (or early) NPDR	1.0000	1.0000	1.0000
Moderate NPDR	0.9588	0.9894	0.9739
Severe NPDR	0.9893	0.9867	0.9880
Very Severe NPDR	0.9841	0.9960	0.9900
PDR	0.9799	0.9512	0.9654
Advanced PDR	0.9922	0.9884	0.9903

this context are ResNet-50, DenseNet-121, EfficientNet-B0, Xception, VGG-16, MobileNetV2, NASNet-Large, Inception-ResNet-V2, and VGG-19.

At the end of the model selection process, NasNetLarge was identified as the best-performing network and was used for the classification task. Subsequently, hyperparameter optimization was performed using grid search to improve the model's overall accuracy. The model classifies diabetic retinopathy into seven different stages: "No DR Signs," "Mild NPDR," "Moderate NPDR," "Severe NPDR," "Very Severe NPDR," "PDR," and "Advanced PDR.

The model was trained using standard categorical cross-entropy loss without class weighting. Although the dataset exhibits severe class imbalance, particularly for the Mild NPDR class (only six original images), we relied solely on extensive data augmentation to mitigate this issue. The results in Table 5 demonstrate that augmentation alone effectively addressed the imbalance, as evidenced by perfect classification performance (Precision = 1.00, Recall = 1.00, F1-Score = 1.00) for the Mild NPDR class. However, future work could explore class-weighted loss functions or focal loss to provide additional safeguards against potential overfitting to augmented samples.

3.9 Performance evaluation metrics

In the study, various metrics were used to evaluate the classification performance of the model. In the process of determining the stages of diabetic retinopathy, performance metrics such as accuracy, precision, recall, and F1-score have been utilized. To calculate these metrics, it is necessary to know the values of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) obtained from the confusion matrix. These terms quantitatively express the classification success of the model and allow for a comprehensive evaluation of its performance [33].

Accuracy refers to the degree to which a model or measurement provides correct results; that is, it is a measure of the alignment between actual values and predicted values. The accuracy formula is provided below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision is a metric that measures how many of the results predicted as positive by a model are correct. In other words, it is the ratio of "true positives" to the sum of "true positives" and "false positives." The precision formula is as follows.

$$Prescision = \frac{TP}{TP + FP} \tag{2}$$

Recall is a metric that measures how accurately a model identifies the actual positive examples. In other words, it is the ratio of "true positives" to the sum of "true positives" and "false negatives." Recall is given in Formula 3.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1-score is the harmonic mean of the precision and recall metrics and combines these two metrics in a balanced manner. The F-score formula is as follows.

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

4 Results

4.1 Experimental setup

In this study, the applications were carried out using the Google Colab platform. To accelerate computational processes and ensure high performance during the training of deep learning models, the NVIDIA T4 graphics processing unit (GPU) has been used. The specified GPU has 16 GB of memory and provides sufficient capacity for efficient model training and inference when working with large datasets. Thanks to the cloud-based resources provided by Google Colab, it has been possible to minimize hardware limitations during the execution of experiments and to test different hyperparameter configurations.

During model training, various geometric and pixel-based transformations were applied to the dataset using the ImageDataGenerator class. First, the pixel values were normalized by scaling them to the range of 0 to 1. Subsequently, transformations such as rotation (range = 40), width shift (range = 0.2), height shift (range = 0.2), shear (range = 0.2), zoom (range = 0.2), and horizontal flip (True) were applied to increase

data diversity. Gaps were filled using the 'nearest' method, ensuring that image distortions after transformation were corrected. These methods have improved the model's performance by enabling it to encounter more diverse and realistic data samples during training.

The computational requirements and training efficiency of the optimized NASNet-Large model are as follows: For the final optimized configuration (713 frozen layers, dropout 0.3, batch size 64, learning rate 0.0001), each fold in the 5-fold cross-validation required approximately 1.8–2.2 h of training on the Nvidia T4 GPU, with early stopping typically terminating training between 29 and 45 epochs rather than the maximum 50 epochs. The complete 5-fold cross-validation process took approximately 10 h. The model achieved stable convergence, with each training epoch requiring approximately 170 s (2.8 min) after the initial model loading phase. During inference, the model processed validation sets (approximately 3,176 images per fold) efficiently within minutes, demonstrating practical feasibility for clinical deployment. The NASNetLarge architecture, while computationally intensive during training due to its 88.9 million parameters, achieves a reasonable processing speed suitable for batch processing in screening programs. GPU memory usage peaked at approximately 12–14 GB during training with batch size 64, remaining well within the T4's 16 GB capacity. These computational characteristics indicate that the model can be trained and deployed on commonly available cloud-based GPU infrastructure without requiring specialized high-end hardware, making it accessible for clinical institutions with moderate computational resources.

The model was trained for 50 epochs; however, due to early stopping, training was sometimes completed after fewer epochs. The results of the analyses of the NasNetLarge model across different hyperparameter settings are shown in Table 6.

SGD provides the lowest accuracy and F1-score values compared to other optimizers. Especially for a batch size of 32, the accuracy (69.60%) is significantly lower. Adam and RMSprop provide higher accuracy compared to SGD. The highest accuracy was achieved with RMSprop (batch size 32, 87.76%). When the batch size is set to 32 for Adam and RMSprop, the accuracy increases somewhat. Especially for Adam, a significant improvement is observed from 85.66% to 87.20%. The best result was obtained with RMSprop (batch size 32) (Accuracy: 87.76%, Precision: 0.88, Recall: 0.88, F1-score: 0.87). Batch size 32 generally produced better results, but it may vary depending on the type of optimizer.

As a result of hyperparameter optimization performed using the Grid Search method, the highest classification performance was achieved with the parameters given in Table 7.

The results in Table 7 show that the NasNetLarge model demonstrated a very high performance under certain hyperparameters. The model was trained using the Adam optimizer and a learning rate of 0.0001. The dropout rate was set at 30%, and 713 layers

Table 6 Results of the analysis conducted with various hyperparameters

Model	Optimizer	Learning rate	Batch Size	Accuracy	Precision	Recall	F1-score
NasNetLarge	SGD	0.001	16	72.47%	0.71	0.72	0.70
NasNetLarge	SGD	0.001	32	69.60%	0.68	0.70	0.66
NasNetLarge	Adam	0.001	16	85.66%	0.87	0.86	0.86
NasNetLarge	Adam	0.001	32	87.20%	0.87	0.87	0.87
NasNetLarge	RMSprop	0.001	16	86.50%	0.86	0.86	0.86
NasNetLarge	RMSprop	0.001	32	87.76%	0.88	0.88	0.87

Table 7 The performance results of the model optimized with Grid Search

Model	Optimizer	Learning Rate	Dropout	Froze Layer	Batch Size	Accuracy	Precision	Recall	F1-score
NasNet-Large	Adam	0.0001	0.3	713	64	98.39%	0.99	0.99	0.99

Table 8 5-Fold Cross-Validation Results of the Proposed Model

Fold	Accuracy	AUC	F1-score	Loss
1	0.9846	0.9981	0.9846	0.0582
2	0.9817	0.9983	0.9817	0.0642
3	0.9884	0.9979	0.9884	0.0472
4	0.9852	0.9996	0.9852	0.0378
5	0.9849	0.9988	0.9852	0.0487
Mean ± Std	0.9850 ± 0.0021	0.9986 ± 0.006	0.9850 ± 0.0021	0.0512 ± 0.0092

were frozen. This indicates that the transfer learning approach was used and that the model benefited from pre-trained weights. Compared to the previously tested values of 16 and 32, the accuracy of the model trained with a batch size of 64 has significantly increased (98.39%). This indicates that the larger batch size contributes to the model undergoing a more stable learning process. The model, with an accuracy rate of 98.39%, performs significantly better than previous configurations. The Precision, Recall, and F1-score values are 0.99, indicating that the model offers high overall performance and distinguishes the classes well. Compared to the previous results, it is observed that the combination of a low learning rate, a greater number of frozen layers, and an increased batch size has significantly improved the model's performance.

To further validate the robustness and generalization capability of the proposed model, 5-fold cross-validation was performed on the entire dataset. This approach provides a more reliable performance estimate by training and evaluating the model on different data partitions, thereby reducing the risk of overfitting and ensuring that the results are not dependent on a specific train-test split. The results of the 5-fold cross-validation are presented in Table 8.

The 5-fold cross-validation results demonstrate that the proposed model achieves consistently high performance across all folds. The mean accuracy of 98.50% ± 0.21% and mean AUC of 0.9986 ± 0.0006 indicate excellent classification performance. The F1-score of 98.50% ± 0.21% confirms the balanced precision-recall trade-off across all diabetic retinopathy stages. Notably, the low standard deviation values ($\sigma < 0.003$) across all metrics indicate that the model exhibits stable performance regardless of the data partition, confirming its strong generalization capability. These results validate that the proposed NASNetLarge-based model is robust and reliable for clinical diabetic retinopathy classification.

To provide deeper insight into the training dynamics and convergence behavior of the model across different data partitions, Fig. 4 presents the training and validation performance curves for all five folds throughout the training process.

Figure 4 illustrates the training dynamics of the proposed NASNetLarge-based model across all five folds during the cross-validation process. Panel (a) shows the training accuracy progression, where all folds demonstrate rapid convergence within the first 10 epochs, reaching accuracy levels above 90%. The curves exhibit consistent learning

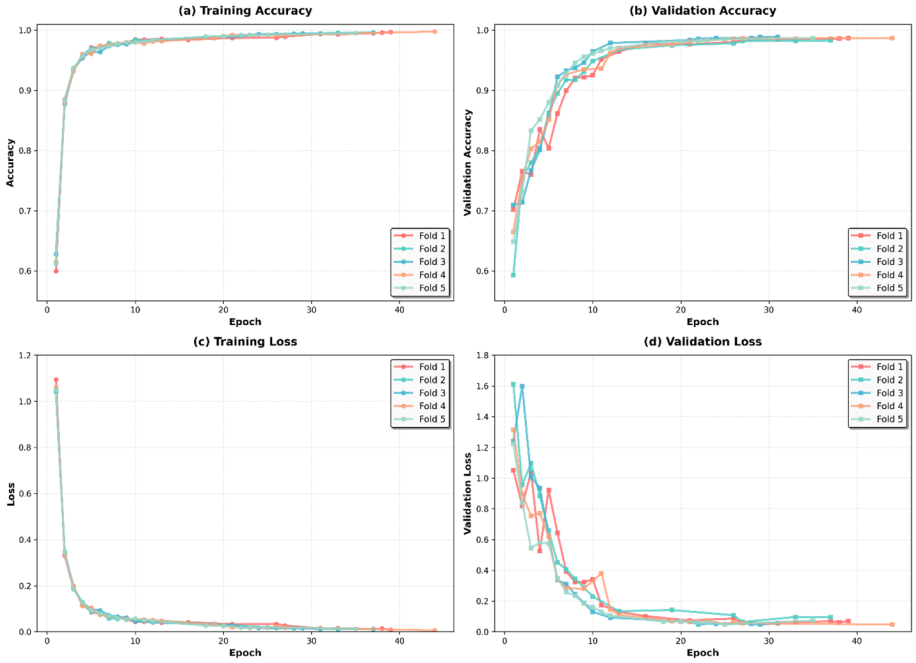

Fig. 4 Training and Validation Performance Curves for 5-Fold Cross-Validation. **a** Training accuracy, **b** Validation accuracy, **c** Training loss, **d** Validation loss

patterns across different data partitions, with final training accuracies ranging from 99.47% to 99.76%, indicating effective feature learning regardless of the training subset. Panel (b) presents the validation accuracy curves, which reveal stable performance across all folds with minimal variance. All folds achieve final validation accuracies above 98.2%, with Fold 3 reaching the highest validation accuracy of 98.90%. The smooth and parallel progression of validation curves across folds demonstrates that the model's performance is reproducible and not dependent on specific train-validation splits. The training loss curves in panel (c) show smooth and consistent optimization across all folds, with final loss values ranging from 0.0066 to 0.0133. The rapid initial descent in the first 5–7 epochs, followed by gradual refinement, indicates efficient gradient-based learning without oscillations or instability. Similarly, panel (d) displays the validation loss progression, where all folds converge to values below 0.1 by epoch 20, with final validation losses ranging from 0.0476 to 0.0953. The consistent patterns observed across all four panels confirm several important characteristics of the proposed model: [1] rapid convergence without requiring excessive training epochs [2], stable performance across different data partitions with low variance ($\sigma < 0.21\%$) [3], minimal overfitting as evidenced by the small gap between training and validation metrics (0.57%-1.34%), and [4] smooth optimization trajectories without training instabilities. These observations validate that the model exhibits strong generalization capability and is robust to variations in training data, making it suitable for potential future clinical applications in diabetic retinopathy classification, pending external validation.

To address potential class-imbalance concerns, particularly for the Mild NPDR class, which had limited original samples (6 images before augmentation), Table 5 presents detailed per-class performance metrics derived from the average confusion matrix across all folds. Table 5. Per-Class Performance Metrics Based on Average Confusion Matrix.

The results demonstrate that despite the severe class imbalance in the Mild NPDR category, the model achieved perfect classification performance (Precision = 1.00, Recall = 1.00, F1-Score = 1.00) for this class, indicating that the combined effect of data augmentation and the model's learning capacity effectively mitigated the imbalance issue. All seven DR stages demonstrate F1-scores above 0.96, with the lowest performance observed in the PDR class (F1 = 0.9654), likely due to overlapping features with adjacent severity levels. These per-class metrics confirm that the model maintains high discriminative ability across all diabetic retinopathy stages, including those with limited training samples.

Figure 5 presents the confusion matrices obtained from each fold of the 5-fold cross-validation, providing a detailed view of the model's classification performance across different data partitions.

The confusion matrices across all five folds demonstrate consistent classification performance. The diagonal elements show high values indicating correct predictions, while off-diagonal misclassifications remain minimal. The No DR signs and Mild NPDR classes achieve near-perfect classification, while Moderate NPDR and PDR classes show slightly higher but clinically acceptable error rates due to overlapping features between adjacent stages. The average confusion matrix confirms that the model reliably distinguishes between all seven diabetic retinopathy stages with minimal inter-class confusion.

Figure 6 presents the per-class one-vs-rest ROC curves obtained from the 5-fold cross-validation, together with the mean ROC over all folds for the seven diabetic retinopathy severity levels. In all folds, the ROC curves lie very close to the upper-left corner, and the corresponding AUC values remain consistently high (≈ 0.998–1.000) for every class, indicating excellent discriminative ability. The mean AUC scores across folds are 1.000 for No DR signs, Mild (or early) NPDR, Severe NPDR, Very Severe NPDR, and Advanced PDR, 0.999 for Moderate NPDR, and 0.999 for PDR, demonstrating that the proposed model can reliably distinguish all stages of diabetic retinopathy with minimal performance variation between folds.

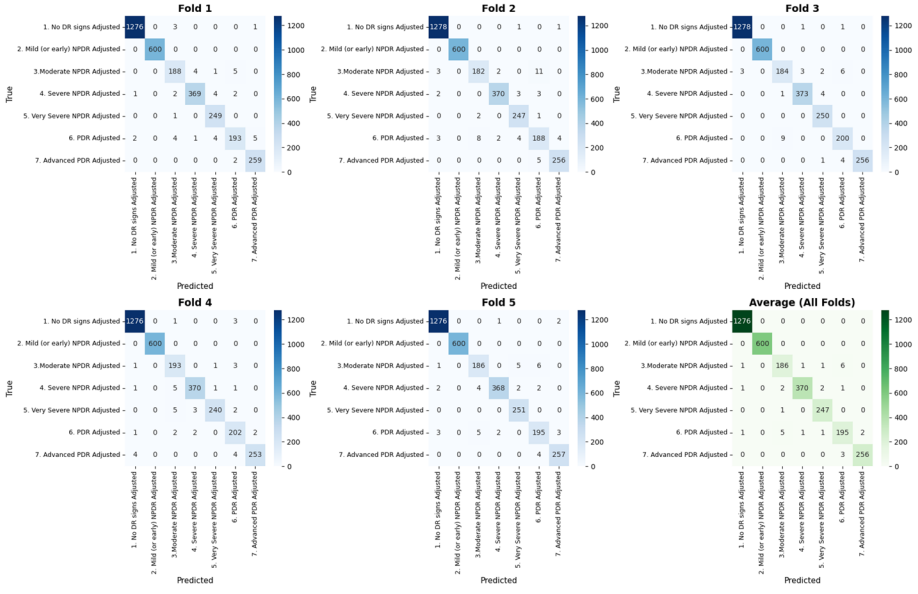

Fig. 5 Confusion Matrices for 5-Fold Cross-Validation

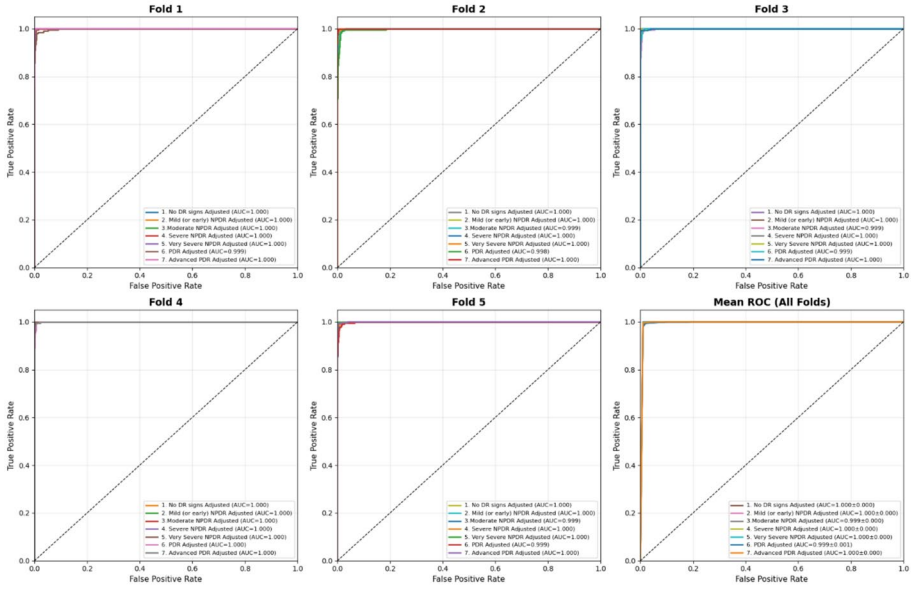

Fig. 6 Per-Class ROC Curves for 5-Fold Cross-Validation

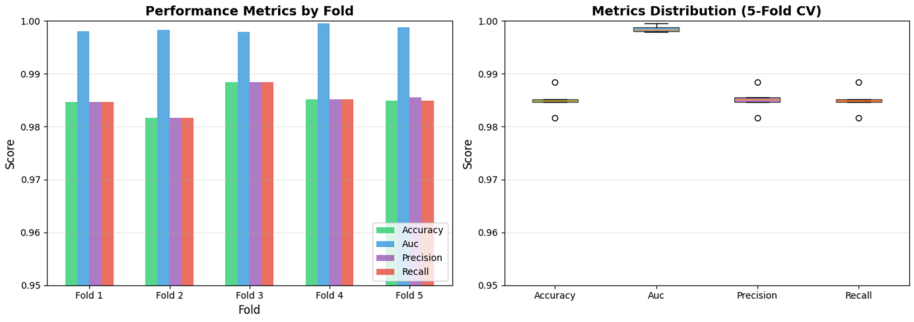

Fig. 7 Overall Performance Metrics for 5-Fold Cross-Validation

While Fig. 6 focuses on the class-wise discrimination ability through ROC analysis, Fig. 7 summarizes the global performance of the proposed model across all folds.

Figure 7 presents the overall classification performance of the proposed method over 5-fold cross-validation using accuracy, AUC, precision, and recall metrics. The bar plot shows that all folds achieve very high scores, with AUC values close to 1.0 and accuracy, precision, and recall consistently above 0.98, indicating robust predictive capability in each split. The box plots on the right further illustrate the tight distribution of these metrics, with small interquartile ranges and only minor deviations across folds, demonstrating that the model's performance is both strong and stable without significant variability between different training–validation partitions.

Comparative results with other studies in the literature conducted on the same dataset are presented in Table 9.

Khan et al. [1] used VGG16-TL, VGG19, AlexNet, and SqueezeNet models in the study, which achieved an accuracy of 96.95%. The proposed model, Optimized Nasnet-Large, achieved an accuracy of 98.39% on the same dataset, outperforming prior studies. These results indicate that the proposed model provides a more effective approach to classifying diabetic retinopathy. The application of optimization techniques has

Table 9 Comparison of our study with the previous studies

Proposed by	Model Used	Dataset Used	Results
Khan et al.	Vgg16-TL VGG19 AlexNet SqueezeNet	7 Stages	96,95%
Our Model	Optimized NasnetLarge	7 Stages	98.39%

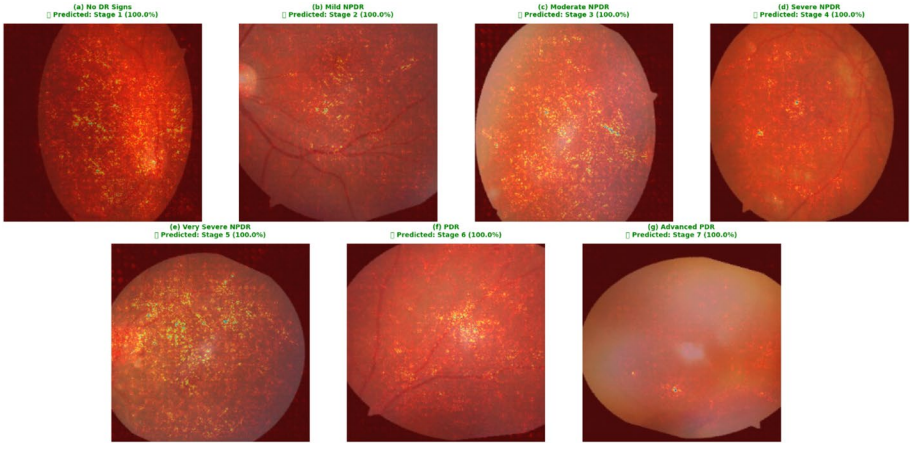

Fig. 8 Saliency map visualization of model attention across seven DR stages

increased the model's accuracy and achieved a more successful result than existing approaches.

To ensure the clinical validity and interpretability of the proposed model, saliency map visualization was performed to identify the regions that most influence the classification decisions across different DR stages (Fig. 8). The visualization reveals that the Optimized NasNetLarge model focuses on clinically relevant anatomical structures and pathological features. In early stages (No DR to Severe NPDR), the model concentrates on central retinal regions, optic disc, and vascular areas where microaneurysms, hemorrhages, and exudates typically manifest. Notably, in advanced stages (PDR and Advanced PDR), the attention maps show more diffuse patterns, reflecting widespread retinal changes and potential vitreous complications characteristic of these severe stages. This explainability analysis demonstrates that the model's superior performance stems from learning meaningful pathological features rather than spurious correlations, thereby enhancing its trustworthiness for potential clinical deployment.

5 Discussion

The proposed Optimized NASNetLarge model achieved 98.39% accuracy on the 7-stage diabetic retinopathy classification task, surpassing the previous state-of-the-art result of 96.95% reported by Khan et al. [1] on the same dataset. The 5-fold cross-validation results (mean accuracy: $98.50 \pm 0.21\%$) demonstrate strong generalization capability with minimal variance across different data partitions. Several factors contributed to this performance improvement. First, the CLAHE-based preprocessing enhanced contrast in low-visibility regions, enabling the model to detect subtle pathological features such as microaneurysms and early exudates. Second, the systematic Optuna-based model selection identified NASNetLarge as the optimal architecture for this task, avoiding

Table 10 Related works in diabetic retinopathy diagnosis with deep learning

Study	Dataset	Model/Method	Accuracy
Li et al. [13]	Messidor DR (2 Class) IDRiD DR (5 Class)	Cross-Disease Attention Network (CANet)	85%
Sajid et al. [14]	Multiple DR-dataset 5-Stages	DR-NASNet (Improved NASNet + Dense Blocks + SVM)	96.05%
Arora et al. [15]	Kaggle DR (35,108 images, five classes)	EfficientNetB0	86.53%
M et al. [16]	IDRiD DR (5 Class)	DR and DME detection	90.48%
Shakibania et al. [4]	Fundus images (5 stages)	Dual-branch (ResNet50 + EfficientNetB0)	93%
S. N et al. [17]	1,065 fundus images	DenseNet121, InceptionV3, AlexNet, ResNet	DenseNet121: 98%
Singh & Dobhal [18]	Indian Fundus Images	VGG16 + Logistic Regression	90.4%
Romero-Oraa et al. [26]	Kaggle DR	Attention Mechanism + Xception	83.7%
Suo et al. [20]	EyePACS, APTOS 2019	CS-ResNet-101 + FCSAM	98.1%
Khan et al. [1]	7-Stages	Vgg16-TL VGG19 AlexNet SqueezeNet	96.95%
Vij & Arora [21]	IDRiD (5 class) APTOS-2018	DITL-DS (InceptionV3, ResNet34, EfficientNetB0, VGG16, Xception)	98.40%
Hassan & Ismail [22]	APTOS 2019, Messidor-2	CNN + Bayesian CNN (VI, MC-Dropout)	CNN: 94.7%, 77.0% BCNN-VI: 94.0%, 86.0% BCNN-MC: 93.3%, 81.0%
Priya & Kumar [23]	DIARETDB1	DRD-MHSAGGCN-BCOA	23.34%-34.33% higher than existing methods
Shoaib et al. [24]	ODIR	DiaGAN-CNN (InceptionResNetv2, InceptionV3)	98%
Gargi et al. [25]	APTOS-2019, DDR	CPDenA (Excess Level Attention Axial Spectral Transformer)	APTOS-2019: 97.95% DDR: 98.59%

the trial-and-error approach commonly used in literature. Third, Grid Search hyperparameter optimization with 713 frozen layers, 0.3 dropout, and batch size 64 prevented overfitting while maintaining high discriminative capability. The saliency map analysis (Fig. 8) validated that the model's decisions are based on clinically relevant features. In early stages, the model concentrated attention on central retinal regions where microaneurysms and hemorrhages typically appear. In advanced PDR stages, attention patterns became more diffuse, correctly reflecting widespread pathological changes. This interpretability is crucial for clinical acceptance, as it demonstrates that the model is not relying on dataset artifacts or spurious correlations. Compared to other recent studies (Table 10), our approach achieves competitive performance while providing better explainability. While some studies [20, 23] reported similar or slightly higher accuracies on different datasets, our model's strength lies in its combination of high accuracy, cross-validated stability, and interpretable decision-making on the challenging 7-stage classification task. However, certain limitations should be acknowledged. The dataset exhibits class imbalance, particularly in the Mild NPDR stage (only 6 original images). Although data augmentation mitigated this issue, future work should incorporate focal loss or class-weighted training to further improve performance on minority classes.

Additionally, external validation on independent datasets from different populations and imaging devices would strengthen claims of generalizability.

5.1 Limitations

Despite achieving perfect classification performance for the Mild NPDR class (Precision = 1.00, Recall = 1.00, F1-Score = 1.00), the severe class imbalance in this category represents a significant limitation of this study. The Mild NPDR class contained only 6 original images before augmentation, which was then expanded through data augmentation techniques. While our results demonstrate that augmentation successfully enabled the model to learn discriminative features for this underrepresented class, there remains a potential risk that the model may have learned augmentation-induced artifacts rather than true pathological features specific to Mild NPDR. To validate the robustness of these findings, future work should include: [1] collection of additional original Mild NPDR samples to create a more naturally balanced dataset [2], external validation on independent datasets with naturally occurring class distributions to confirm that the model generalizes to real-world Mild NPDR cases beyond augmented variations, and [3] incorporation of class-weighted loss functions or focal loss during training to provide explicit mechanisms for handling class imbalance beyond augmentation alone. These steps would further strengthen confidence in the clinical applicability of the model, particularly for early-stage diabetic retinopathy detection where Mild NPDR identification is critical for timely intervention.

A significant limitation of this study is the absence of external validation on independent datasets from different populations, imaging devices, and acquisition protocols. While the 5-fold cross-validation demonstrates robust internal consistency ($98.50 \pm 0.21\%$), this evaluation is limited to a single dataset collected under specific conditions. External validation is essential before clinical deployment to ensure that the model generalizes across diverse clinical settings, including variations in camera equipment, image quality, demographic characteristics, and disease prevalence. Without multi-center validation studies, it cannot be conclusively determined whether the model's high performance will translate to real-world screening programs where imaging conditions and patient populations may differ substantially from the training data. Future work must prioritize external validation on independent datasets to establish the model's true generalizability and clinical readiness. Additionally, prospective evaluation in real clinical workflows, including assessment of integration challenges, user acceptance, and impact on diagnostic accuracy and workflow efficiency, will be crucial steps toward practical implementation.

6 Conclusions

In this study, a comprehensive deep learning-based approach was developed for the classification of seven diabetic retinopathy stages using fundus images. The proposed methodology integrating CLAHE-based preprocessing, systematic model selection via Optuna, and Grid Search hyperparameter optimization achieved a classification accuracy of 98.39%, with 5-fold cross-validation demonstrating robust and stable performance ($98.50 \pm 0.21\%$). The application of CLAHE-based contrast enhancement and data augmentation techniques effectively improved image quality and diversity, enhancing the model's generalization capability. The Optuna framework systematically identified

NASNetLarge as the optimal architecture, and subsequent Grid Search optimization fine-tuned hyperparameters to maximize classification performance. The confusion matrix and classification metrics indicate minimal misclassification rates, confirming the model's robust capability in distinguishing between different DR stages. Furthermore, saliency map analysis demonstrated that the model's superior performance stems from learning clinically meaningful pathological features rather than spurious correlations. The visualization confirmed that the model focuses on diagnostically relevant regions including the optic disc, vascular abnormalities, microaneurysms, hemorrhages, and lesion sites across all DR stages. In early stages (No DR to Severe NPDR), attention concentrated on central retinal regions where pathological changes typically manifest, while in advanced stages (PDR and Advanced PDR), the model correctly identified diffuse patterns characteristic of widespread retinal damage. This explainability component enhances the model's clinical trustworthiness and provides confidence for potential deployment in real-world diagnostic workflows, addressing a critical requirement for clinical acceptance of AI-based diagnostic systems. High classification performance combined with interpretability is critically important for early diagnosis and accurate staging, directly affecting treatment decisions and improving patient prognosis. Considering both the quantitative performance achieved and the qualitative validation through saliency analysis, the proposed model shows promise for diabetic retinopathy classification and could potentially be integrated into screening programs and clinical decision support systems following comprehensive external validation across diverse clinical settings. In future studies, several directions could further enhance model performance and clinical applicability. First, alternative loss functions such as Focal Loss could be investigated to better handle class imbalance and improve performance on minority classes, particularly Mild NPDR. Second, expanding the dataset with additional examples for underrepresented classes or applying more comprehensive data augmentation strategies could enhance the model's sensitivity and overall robustness. Third, external validation on independent datasets from diverse populations and imaging devices would strengthen generalizability claims and provide evidence for real-world deployment. Finally, integration with clinical workflows and prospective evaluation in screening settings would be essential steps toward practical implementation.

Author contributions
Yavuz Unal: Conceptualization, methodology, software, validation, formal analysis, investigation, writing - original draft, writing - review and editing, visualization.

Funding
No funding was received.

Data availability
This research utilized a publicly available dataset concerning diabetic retinopathy classification and severity, released in 2021. The dataset is available on Zenodo, and the direct link for access is https://zenodo.org/records/4891308 .

Declarations

Ethics approval and consent to participate
Not applicable. This study utilized a publicly available dataset; therefore, no ethical approval was required. This study used an existing publicly available dataset; no participants were directly involved.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 9 December 2025 / Accepted: 1 April 2026
Published online: 08 April 2026

References

1. Khan AQ, Sun G, Khalid M, Farrash M, Bilal A. Multi-Deep Learning Approach With Transfer Learning for 7-Stages Diabetic Retinopathy Classification. Int J Imaging Syst Tech. 2024;34(6):e23213.
2. Khan KB, Khaliq AA, Jalil A, Iftikhar MA, Ullah N, Aziz MW, et al. A review of retinal blood vessels extraction techniques: challenges, taxonomy, and future trends. Pattern Anal Applic. 2019;22(3):767–802.
3. Li X, Wen L, Du F, Yang L, Wu J. Recognition of Diabetic Retinopathy Grades Based on Data Augmentation and Attention Mechanisms. Int J Imaging Syst Tech. 2024;34(6):e23201.
4. Shakibania H, Raoufi S, Pourafkham B, Khotanlou H, Mansoorizadeh M. Dual branch deep learning network for detection and stage grading of diabetic retinopathy. Biomed Signal Process Control. 2024;93:106168.
5. Kropp M, Golubnitschaja O, Mazurakova A, Koklesova L, Sargheini N, Vo TTKS, et al. Diabetic retinopathy as the leading cause of blindness and early predictor of cascading complications—risks and mitigation. EPMA J. 2023;13(1):21–42.
6. Tan TE, Wong TY. Diabetic retinopathy: Looking forward to 2030. Front Endocrinol. 2023;9:13:1077669.
7. Dai L, Sheng B, Chen T, Wu Q, Liu R, Cai C, et al. A deep learning system for predicting time to progression of diabetic retinopathy. Nat Med. 2024;30(2):584–94.
8. Tsuneki M. Deep learning models in medical image analysis. J Oral Biosci. 2022;64(3):312–20.
9. Lim JI, Rachitskaya AV, Hallak JA, Gholami S, Alam MN. Artificial intelligence for retinal diseases. Asia-Pacific J Ophthalmol. 2024;13(4):100096.
10. Suedumrong C, Phongmoo S, Akarajaka T, Leksakul K. Diabetic Retinopathy Detection Using Convolutional Neural Networks with Background Removal, and Data Augmentation. Appl Sci. 2024;14(19):8823.
11. Luca AR, Ursuleanu TF, Gheorghe L, Grigorovici R, Iancu S, Hlusneac M, et al. Impact of quality, type and volume of data used by deep learning models in the analysis of medical images. Inf Med Unlocked. 2022;29:100911.
12. Bhan B, Patel S. Efficient Medical Image Enhancement using CLAHE Enhancement and Wavelet Fusion. IJCA. 2017;167(5):1–5.
13. Li X, Hu X, Yu L, Zhu L, Fu CW, Heng PA. CANet: Cross-Disease Attention Network for Joint Diabetic Retinopathy and Diabetic Macular Edema Grading. IEEE Trans Med Imaging. 2020;39(5):1483–93.
14. Sajid MZ, Hamid MF, Youssef A, Yasmin J, Perumal G, Qureshi I, et al. DR-NASNet: Automated System to Detect and Classify Diabetic Retinopathy Severity Using Improved Pretrained NASNet Model. Diagnostics. 2023;13(16):2645.
15. Arora L, Singh SK, Kumar S, Gupta H, Alhalabi W, Arya V, et al. Ensemble deep learning and EfficientNet for accurate diagnosis of diabetic retinopathy. Sci Rep. 2024;14(1):30554.
16. A PS MMPP, Vijayan ESK. D. Diabetic Retinopathy Detection in Fundus Images Using Deep-Learned Features. In: 2024 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER) [Internet]. Mangalore, India: IEEE; 2024 [cited 2025 Jan 29]. pp. 365–70. Available from: https://ieeexplore.ieee.org/document/10750630/
17. S N S MA. Deep Learning-Based Diabetic Retinopathy Severity Classification and Progression Time Estimation. IFAC-PapersOnLine. 2024;58(3):78–83.
18. Singh D, Dobhal DC. A Deep Learning-based Transfer Learning Approach Fine-Tuned for Detecting Diabetic Retinopathy. Procedia Comput Sci. 2024;233:444–53.
19. Romero-Oraá R, Herrero-Tudela M, López MI, Hornero R, García M. Attention-based deep learning framework for automatic fundus image processing to aid in diabetic retinopathy grading. Comput Methods Programs Biomed. 2024;249:108160.
20. Suo Y, He Z, Liu Y. Deep learning CS-ResNet-101 model for diabetic retinopathy classification. Biomed Signal Process Control. 2024;97:106661.
21. Vij R, Arora S. Modified deep inductive transfer learning diagnostic systems for diabetic retinopathy severity levels classification. Biomed Signal Process Control. 2025;99:106885.
22. Hassan MM, Ismail HR. Bayesian deep learning applied to diabetic retinopathy with uncertainty quantification. Heliyon. 2025;11(2):e41802.
23. Bindu Priya M, Manoj Kumar D. MHSAGGCN-BCOA: A novel deep learning based approach for diabetic retinopathy detection. Biomed Signal Process Control. 2025;105:107569.
24. Shoaib MR, Emara HM, Mubarak AS, Omer OA, Abd El-Samie FE, Esmaiel H. Revolutionizing diabetic retinopathy diagnosis through advanced deep learning techniques: Harnessing the power of GAN model with transfer learning and the DiaGAN-CNN model. Biomed Signal Process Control. 2025;99:106790.
25. Gargi M, Krishna Eluri R, Prakash Samantray O, Hajarathaiah K. Compact Pyramidal dense mixed attention network for Diabetic retinopathy severity prediction under deep learning. Biomed Signal Process Control. 2025;100:106960.
26. Cui C, Yao J, Xia H. Data Augmentation: A Multi-Perspective Survey on Data, Methods, and Applications. CMC. 2025;85(3):4275–306.
27. Eshwaran Ramachandran M, Babu Balachandran G, Velladurai P, Rajakumar A. Deep learning for solar PV fault classification using RGB imaging and comparison of preprocessing techniques. Sol Energy. 2025;301:113959.
28. Pizer SM, Johnston RE, Ericksen JP, Yankaskas BC, Muller KE. Contrast-limited adaptive histogram equalization: speed and effectiveness. In: [1990] Proceedings of the First Conference on Visualization in Biomedical Computing [Internet]. Atlanta, GA, USA: IEEE Comput. Soc. Press; 1990 [cited 2025 Feb 1]. pp. 337–45. Available from: http://ieeexplore.ieee.org/document/109340/
29. Khan KB, Siddique MS, Ahmad M, Mazzara M. A Hybrid Unsupervised Approach for Retinal Vessel Segmentation. Parodi MB, editor. BioMed Research International. 2020;2020(1):8365783.
30. Akiba T, Sano S, Yanase T, Ohta T, Koyama M, Optuna. : A Next-generation Hyperparameter Optimization Framework [Internet]. arXiv; 2019 [cited 2025 Jan 28]. https://arxiv.org/abs/1907.10902
31. Kılıçarslan S, Diker A, Közkurt C, Dönmez E, Demir FB, Elen A. Identification of multiclass tympanic membranes by using deep feature transfer learning and hyperparameter optimization. Measurement. 2024;229:114488.
32. Rehman F, Rehman M, Anjum M, Hussain A. Optimized YOLOV8: An efficient underwater litter detection using deep learning. Ain Shams Eng J. 2025;16(1):103227.

33. Gupta S, Panwar A, Goel S, Mittal A, Nijhawan R, Singh AK. Classification of Lesions in Retinal Fundus Images for Diabetic Retinopathy Using Transfer Learning. In: 2019 International Conference on Information Technology (ICIT) [Internet]. Bhubaneswar, India: IEEE; 2019 [cited 2025 Feb 13]. pp. 342–7. https://ieeexplore.ieee.org/document/9031884/

Publisher's Note