

## Diabetic Retinopathy Classification Using a Hybrid Deep Learning and Machine Learning Model

Motahareh Barzegari<sup>1</sup>, Fatemeh Zare Mehrjardi<sup>2</sup>, Mohsen Sardari Zarchi<sup>3\*</sup>

<sup>1</sup>M.Sc. Student of Artificial Intelligence and Robotics, Faculty of Engineering, Department of Computer Engineering, Meybod University, Meybod, Iran.

<sup>2</sup>Assistant Professor, Faculty of Engineering, Department of Computer Engineering, Meybod University, Meybod, Iran.

<sup>3</sup>Associate Professor, Faculty of Engineering, Department of Computer Engineering, Meybod University, Meybod, Iran

### Abstract

**Objective:** Among diabetic patients, diabetic retinopathy (DR) remains one of the most common causes of preventable blindness and vision loss, making its early detection crucial for preventing irreversible complications. Manual evaluation of fundus photographs is a lengthy process. Additionally, it requires specialized training that is not always available in all clinical settings. Consequently, artificial intelligence-based automated retinal image analysis systems have emerged as complementary tools to enhance diagnostic accuracy and efficiency. This study proposes an ensemble learning-based framework to improve the accuracy and robustness of automated DR detection. In the first stage, pretrained convolutional neural network (CNN) models extract high-level features from fundus images, capturing complex patterns and DR-related lesions. These features are then fed into several classical machine-learning classifiers, including Support Vector Machine (SVM), Random Forest, and XGBoost. To further boost discriminative power and reduce classification errors, a stacking ensemble strategy integrates the predictions of the individual classifiers within a meta-learning framework, enabling the model to learn the optimal combination for DR detection and grading. This hybrid approach effectively combines the strengths of deep learning and classical machine learning, yielding improved performance in DR detection and classification. Experimental results show that the stacking ensemble achieves higher accuracy and F1-score compared to individual models, underscoring its potential as an auxiliary tool for early diabetic retinopathy detection.


**Keywords:** Diabetic retinopathy, Ensemble learning, Deep learning, Machine learning

### QR Code:



**Citation:** Barzegari M, Zare Mehrjardi F, Sardari Zarchi M. Diabetic Retinopathy Classification Using a Hybrid Deep Learning and Machine Learning Model. IJDO 2026; 18 (2) :98-112

**URL:** <http://ijdo.ssu.ac.ir/article-1-1037-en.html>

 10.18502/ijdo.v18i2.21647

### Article info:

**Received:** 5 January 2026

**Accepted:** 20 May 2026

**Published in June 2026**



This is an open access article under the (CC BY 4.0)

### Corresponding Author:

**Mohsen Sardari Zarchi**, Associate Professor, Department of Computer Engineering, Meybod University, Meybod, Iran.

**Tel:** (98) 913 454 4437

**Email:** [sardari@meybod.ac.ir](mailto:sardari@meybod.ac.ir)

**Orcid ID:** 0000-0003-0831-3426

## Introduction

**D**iabetes mellitus is one of the most prevalent chronic metabolic diseases worldwide, affecting millions of individuals and recognized as a leading cause of adult blindness. Among the various complications of diabetes, diabetic retinopathy (DR) stands out as particularly dangerous because it damages retinal tissue. Without timely diagnosis and treatment, DR can result in permanent vision loss or severe visual impairment (1).

Diagnosis of diabetic retinopathy primarily relies on fundus images; however, manual examination by specialists is time-consuming, costly, and susceptible to human error (2). Therefore, artificial intelligence, particularly machine learning and deep learning, is now widely used to automate DR diagnosis. These systems can rapidly and accurately analyze fundus images, extract complex disease-related features, and facilitate early detection (3). Their deployment not only reduces specialists' workload but also improves diagnostic accuracy and accessibility, especially in underserved and remote regions.

Despite notable advances, several challenges persist. These include the need for models that achieve high accuracy, strong generalization capability, and sufficient computational efficiency, especially when dealing with low-quality images or imbalanced datasets with unequal sample distribution across DR stages. Traditional diagnostic approaches depend on manual feature extraction and expert interpretation. In contrast, deep convolutional neural networks (CNNs), especially pretrained models, enable more accurate and fully automated DR detection without manual intervention, thereby streamlining the diagnostic process (4).

In this study, we investigate a deep-learning-based approach to automatically detect and grade diabetic retinopathy through fundus image analysis. By leveraging CNNs, classical machine-learning models, and data-augmentation techniques, the proposed method

aims to enhance diagnostic accuracy and robustness against variations in image quality and noise. The goal is to develop an assistive tool for ophthalmologists that contributes effectively to early DR detection and vision-loss prevention.

This article consists of five main sections: Section 2 presents a review of related works and recent advancements in the field of diabetic retinopathy detection using deep learning and machine learning techniques; Section 3 describes the proposed method, including details about data collection, data preprocessing, model training, and feature extraction; Section 4 reports the experimental results, which evaluate the performance of the models and provide a comparison between different methods; and finally, Section 5 provides a discussion of the findings' importance and offers directions for future work.

## Review of related works

In recent years, deep learning, particularly Convolutional Neural Networks (CNNs), has revolutionized the diagnosis of ocular diseases due to their remarkable ability to extract complex features from medical images. A landmark study by Gulshan and colleagues provided early evidence that deep CNNs can reach performance levels comparable to clinical experts. Their large-scale study, utilizing over 120,000 fundus images from the EyePACS dataset, was among the first to show that a deep CNN could rival ophthalmologists in diagnostic accuracy, underscoring the inherent strength of CNNs in processing intricate medical imagery (2). Following this, Ting et al. conducted a comprehensive review of AI and deep learning applications in ophthalmology, emphasizing the critical role of multicenter and diverse datasets in enhancing model generalization. Their findings confirmed that deep learning algorithms could accurately detect DR and other diabetes-related eye conditions, and importantly, that data from

heterogeneous populations and varying geographical locations significantly boosted model performance and generalizability (3). Further illustrating the trend towards automated diagnosis and practical clinical deployment, Karsaz et al. employed a pretrained GoogleNet architecture trained on the Kaggle Diabetic Retinopathy dataset. Their approach achieved high DR detection accuracy on real-world clinical images, substantially reducing the need for manual feature extraction (4).

Despite these successes, single deep learning models often grapple with limitations such as overfitting, sensitivity to imbalanced datasets, and reduced accuracy in classifying severe DR stages. To surmount these challenges, ensemble learning methods have emerged as a robust solution, combining multiple models to achieve more stable and reliable performance (5,6). Alauthman et al., for instance, proposed an ensemble-based DR detection system integrating several well-known classification algorithms. By evaluating different feature combinations, they demonstrated that their ensemble approach significantly outperformed individual classifiers in accuracy, offering a fast, reliable, and fully automated tool for DR detection (5). Similarly, Revathy et al. extracted key retinal features and employed a hybrid classification strategy combining Support Vector Machines (SVM), k-Nearest Neighbors (k NN), Random Forest, Logistic Regression, and Multi-Layer Perceptron (MLP). Their hybrid method achieved satisfactory accuracy, sensitivity, and F1 score, proving that combining multiple machine learning algorithms can enable reliable automated diagnosis while reducing reliance on manual expert assessment. Notably, their approach outperformed single CNN models, particularly on imbalanced data and severe DR stages (7).

In a broader evaluation, Abushawish et al. systematically compared 26 pretrained architectures (including ResNet, DenseNet, VGG, Inception, MobileNet, Xception, EfficientNet, and Vision Transformer) for DR detection and grading. Their study highlighted

that transfer learning combined with interpretability techniques like Grad CAM can substantially improve classification accuracy, model transparency, and the identification of discriminative image features (8). Complementing this, Bidwai et al. demonstrated that integrating multimodal retinal images with an optimized LightGBM algorithm yielded superior performance compared to conventional single-image classification, providing an effective and reliable solution for automated DR diagnosis in clinical practice (9).

While traditional deep learning methods like CNNs have shown remarkable success in DR grading, they often rely on extensive labeled data and face challenges with subtle lesions and imbalanced datasets (2,7). To address the data scarcity issue, recent advancements explore self-supervised learning (SSL). The DR-MAE framework, for example, overcomes limitations of random masking in standard SSL approaches by proposing an anatomy-guided masked autoencoder. This method intelligently generates masks based on optic disc localization and vascular topology, preserving critical regions instead of relying on random destruction. By incorporating a Hybrid Dimensional Convolution (HDC) block for multi-scale and local-global feature interaction, DR-MAE achieves state-of-the-art performance, showcasing the potential of structured self-supervision for robust DR classification (10).

Building upon the strengths of different architectures, Vision Transformers (ViTs) have emerged as a powerful alternative to CNNs, adept at modeling global contextual information and capturing long-range dependencies often missed by CNNs. Specifically, the Task-Optimized Vision Transformer (TOViT) was developed to enable high-performance DR grading in resource-constrained settings. Through optimization strategies such as layer-wise learning rate scheduling, attention head tuning, and embedding dimension refinement, TOViT enhances feature extraction efficiency.

Furthermore, structured pruning and 8-bit quantization enable real-time inference on low-cost hardware like Raspberry Pi-4, demonstrating the scalability of advanced transformer-based diagnostics for portable, point-of-care screening devices without compromising clinical accuracy (11).

Further advancing hybrid architectures, a CNN-Transformer fusion model has been proposed to tackle limitations like inconsistent performance and poor interpretability inherent in standalone deep learning models. This approach synergistically leverages CNNs for local feature extraction and transformers for capturing long-range dependencies, leading to enhanced accuracy and generalizability. The model achieved high performance on the APTOS 2019 dataset and on the IDRiD dataset, confirming its robustness across varied scenarios. This work presents a powerful AI-driven diagnostic tool by effectively combining the strengths of both architectures, significantly improving clinical DR management through early and reliable detection (12).

Among ensemble techniques, stacking represents a highly advanced strategy that combines the outputs of multiple base learners via a meta-classifier to amplify overall predictive performance (6). Motivated by these diverse studies, the present research introduces a hybrid framework designed for improved diabetic retinopathy detection. Specifically, three pretrained deep learning models (ResNet50, DenseNet121, and ConvNeXt Base) are employed as feature extractors for retinal images. The extracted features are subsequently fed into various classical machine learning classifiers, including SVM, Random Forest, XGBoost, and LightGBM, whose individual performances are rigorously compared. Finally, a stacking strategy is implemented to integrate both the CNN-based features and the classical model predictions, utilizing XGBoost as the meta-classifier. Experimental results underscore that this comprehensive hybrid ensemble approach achieves superior accuracy and robustness in

classifying different DR stages compared to individual models and other ensemble methods.

## Proposed method

The main objective of this study is to present an efficient approach for the detection of diabetic retinopathy from fundus images. Given the critical importance of early diagnosis and the limitations associated with traditional image assessment methods performed by specialists, the application of artificial intelligence techniques particularly deep learning can play a significant role in automating the diagnostic process while improving accuracy and computational efficiency. In this research, a hybrid model based on deep learning and machine learning is designed to extract complex features from fundus images and to classify diabetic retinopathy accordingly. An overview of the architecture and main stages of the proposed method is shown in Figure 1, while detailed descriptions of each stage are provided in the subsequent sections.

### 1. Data collection

In this study, the “Diabetic Retinopathy 224×224 Gaussian Filtered” dataset available on the Kaggle platform is utilized. This dataset is a preprocessed version of the APTOS 2019 Blindness Detection dataset. The images have been processed using a Gaussian filter and resized to 224×224 pixels for noise reduction and enhancement of features relevant to diagnosis. The use of this dataset enables effective training and evaluation of deep learning models for the detection and grading of diabetic retinopathy. The dataset consists of 3,662 retinal fundus images collected from diabetic patients, which are categorized into five distinct classes corresponding to different stages of diabetic retinopathy. Table 1 presents the class labels along with the number of images available for each class in the employed dataset. The classes are defined as follows.

#### 1.1. No DR: No diabetic retinopathy

The patient exhibits no visible signs of diabetic retinopathy.

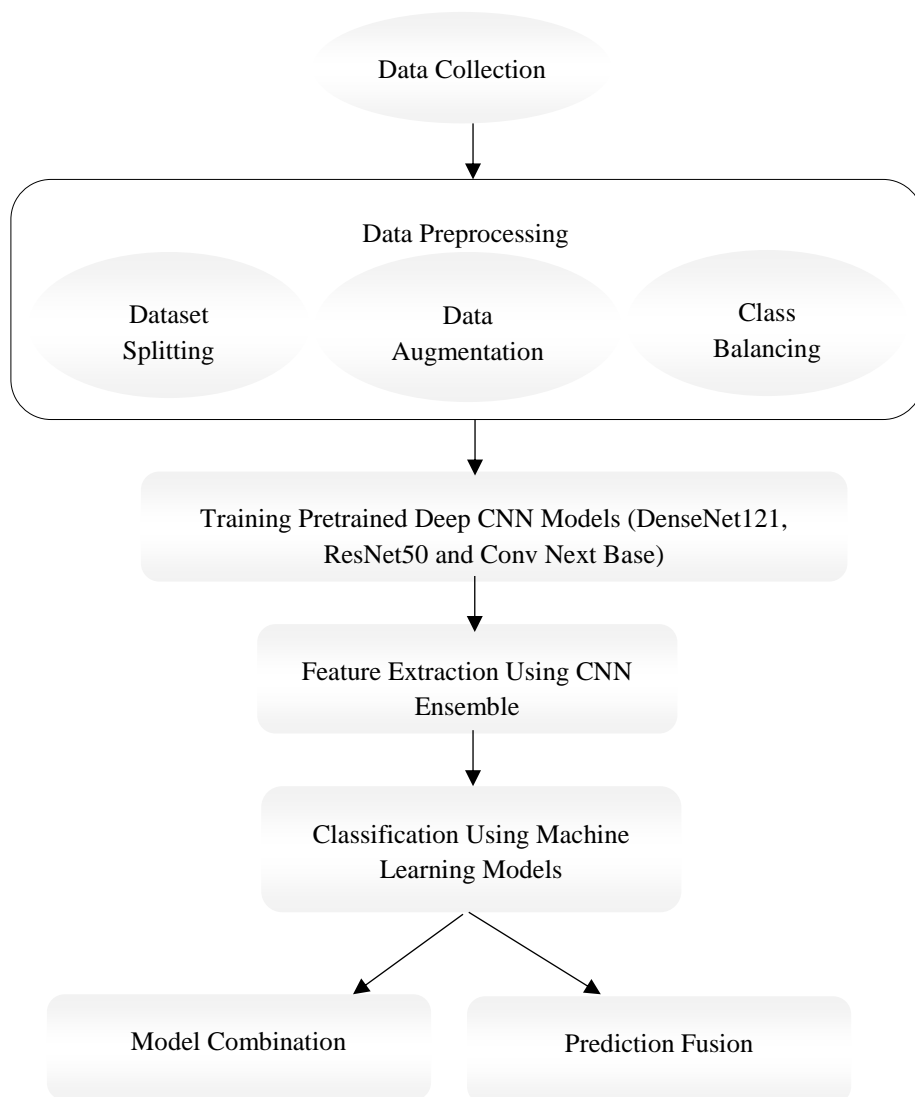


Figure 1. An overview of the proposed method

Table 1. Number of images per class in the dataset

Class	Number of Images
No DR	1805
Mild DR	370
Moderate DR	999
Severe DR	193
Proliferative DR	295
Total	3662

**1.2. Mild DR:** Mild diabetic retinopathy

Presence of small hemorrhages or a limited number of microaneurysms in the retina.

**1.3. Moderate DR:** Moderate diabetic

Retinopathy more extensive lesions, hemorrhages, and vascular abnormalities are observed; however, the disease has not yet reached an advanced stage.

**1.4. Severe DR:** Severe diabetic

Retinopathy a large number of lesions and hemorrhages are present in the retina, with a high risk of visual impairment.

**1.5. Proliferative DR:** Proliferative diabetic

retinopathy abnormal growth of new blood vessels in the retina, associated with a high risk of vision loss.

## 2. Data preprocessing

The first stage of the proposed method involves data preprocessing, which consists of several essential steps. In this section, these preprocessing steps are described in detail.

### 2.1. Dataset Splitting

In the first step, the dataset was divided into three independent subsets: training, validation, and testing, with a split ratio of 70/15/15. This partitioning was performed to preserve the class distribution in each subset and to ensure that the samples were balanced across the Train, Validation, and Test sets. Subsequently, each image was resized to 224×224×3 dimensions, and its pixel values were normalized between 0 and 1.

### 2.2. Data augmentation

To increase the diversity of training samples and enhance the generalizability of deep networks, data augmentation techniques were applied exclusively to the training set. In this stage, operations such as pixel intensity normalization, image rotation, and flipping were performed. This process enriched the feature space and reduced the risk of overfitting, whereas for the validation and test sets, only image resizing and normalization were applied to ensure that the data were evaluated in their raw form, without artificial modifications.

### 2.3. Class balancing

A key feature of the current implementation is the use of an on-the-fly class balancing mechanism during the model training process. In this approach, instead of manually repeating or artificially augmenting data, a sample from each class is randomly selected during each dataset iteration, ensuring that the class distribution remains balanced throughout training. The number of instances belonging to each class is implicitly adjusted to match the size of the majority class; thus, in each epoch, the model observes an equal number of images from all classes. Simultaneously, data augmentation techniques including rotation,

flipping, brightness, and contrast adjustments are applied to these images to increase the visual diversity of the samples. This approach allows the model to encounter different and diverse versions of the data at each training step, thereby enhancing its sensitivity to minority classes. Consequently, the model's bias toward frequently occurring classes is reduced, and the network's performance in detecting rare stages of diabetic retinopathy is significantly improved.

## 3. Model training with CNNs

In this study, three pre-trained deep networks DenseNet121, ResNet50, and ConvNeXt-Base were employed, with each model trained independently on the training set. The selected loss function, FocalLabelSmoothLoss, is a combination of Focal Loss and Label Smoothing, designed to increase sensitivity to rare samples while preventing overfitting and enhancing the generalizability of the networks. The choice of this loss function was carefully considered based on the characteristics of our dataset. To justify this selection, we provide a comparison among Cross-Entropy Loss, Focal Loss, and FocalLabelSmoothLoss. Cross-Entropy Loss, while simple and widely used in classification tasks, tends to favor majority classes in imbalanced datasets, leading to poor performance in identifying less prominent classes. Focal Loss was designed to reduce the impact of easy samples and focus more on challenging ones, performing well in imbalanced scenarios; however, it may still struggle with hard predictions. FocalLabelSmoothLoss integrates both Focal Loss and Label Smoothing, preventing hard predictions and utilizing a probabilistic distribution to enhance model generalization, particularly in multi-class problems where class boundaries may not be well-defined. Given the specific conditions of our problem, which include imbalanced data and multi-class nature, FocalLabelSmoothLoss proves to be a more effective loss function compared to Cross-Entropy and even Focal Loss. This function not only aids in focusing on difficult samples but

also enhances model generalization and reduces error rates in less prominent classes. Furthermore, cross-validation results indicate that using FocalLabelSmoothLoss significantly improved model performance. Hence, this loss function was selected as an optimal strategy to enhance model efficacy in this research. The outputs of each model were subsequently used for the next stage of feature extraction.

#### 4. Feature extraction from CNN models

This stage is aimed at extracting high-level features from the images processed by the trained CNNs. These features capture complex and rich patterns in the images and are subsequently used by classical machine learning models. Extracting features from trained CNNs allows classical models to operate on high-level information without the need for direct training on raw images. Moreover, the dimensionality reduction compared to the original images helps retain essential information for classification. The ability to combine multiple networks and obtain more comprehensive features further enhances the accuracy and generalizability of the system. This step serves as a bridge between deep learning methods and classical machine learning algorithms, forming the core of the hybrid ensemble approach. In this study, features were extracted from the last convolutional layer of each network, prior to the final classification layer. This layer performs a Global Average Pooling (GAP) operation. In transfer learning architectures, this part of the network typically contains high-level, abstract representations of the retinal structure, capturing key spatial information that reflects complex semantic patterns associated with diabetic retinopathy lesions (such as hemorrhages and microaneurysms) and critical retinal regions. Specifically, after removing the final fully connected layer of each model, the feature outputs were extracted from the GAP layer and stored as numerical vectors (feature vectors). These vectors were then used as inputs for classical machine learning models (e.g., SVM, Random Forest, XGBoost, LightGBM).

This choice was motivated by the GAP layer's strong ability to provide abstract features while reducing dimensionality without losing semantic information. The selection of this extraction point was also supported by prior studies in medical computer vision, which have shown that layers near the final classifier exhibit the highest discriminative power for distinguishing disease severity, whereas intermediate layers primarily capture textural patterns and edges. Additional experiments were conducted to ensure the optimality of the feature extraction point, demonstrating that extracting features from this layer outperformed intermediate layers, yielding approximately a 2–3% improvement in F1-score and AUC. Consequently, this layer was selected as the final feature extraction point.

#### 5. Data classification using machine learning models

The objective of this stage was to evaluate the capability of non-deep models in classifying images based on features extracted from CNNs. The advantages of using classical models include faster training compared to deep networks, reduced requirement for large amounts of data for direct image training, and easy integration with ensemble and voting strategies. Applying classical models to CNN-extracted features creates a hybrid approach that leverages both the feature extraction power of deep networks and the speed and simplicity of classical models. The outputs of this stage, in addition to being evaluated independently, were used for Voting Ensemble and decision fusion with CNNs. In this study, a set of classical machine learning models was employed to investigate the effect of high-level features extracted from CNNs. These models, due to their structural and learning diversity, allow a more precise assessment of the role of extracted features in multi-class classification. To ensure optimal performance and reproducibility, the hyperparameter search ranges for all classical models used in this study were selected based on recommendations from

previous studies and preliminary experiments (Table 2).

These ranges provide a balanced trade-off among model complexity, training time, and the risk of overfitting, and have also been reported in similar research in the domain of diabetic retinopathy detection. The selected ranges were designed to offer sufficient diversity for discovering optimal configurations while avoiding unnecessary computational cost. The main criterion for selecting the best configuration was maximization of the weighted F1-score on the validation set.

The description of each model is provided as follows:

### **Random Forest (RF)**

Random Forest (RF) is a powerful ensemble learning method primarily used for classification and regression tasks (10). It builds multiple decision trees during training and merges their outputs to achieve a more accurate and stable prediction. The Bagging (Bootstrap Aggregating) technique is utilized, where each tree is trained on a random subset of the dataset, which helps in reducing variance and improving model robustness. In image

processing, RF effectively handles nonlinear and complex features extracted from Convolutional Neural Networks (CNNs) and is particularly beneficial when the dataset is small or imbalanced, as it can maintain performance in the presence of noise (13,14).

### **HistGradientBoosting (HistGB)**

HistGradientBoosting (HistGB) is a variant of the gradient boosting method that optimizes both training time and memory efficiency. It achieves this by binning continuous features into discrete bins, which reduces the complexity of the data and speeds up the training process. HistGB is particularly well-suited for large datasets with a high number of dimensions, as it can efficiently model complex patterns learned from features extracted by CNNs. The model benefits from tree-based learning to provide accurate predictions while maintaining low computational costs.

### **Logistic Regression (LR)**

Logistic Regression (LR) is a widely used statistical model for binary classification that can be extended to multi-class classification through techniques such as one-vs-rest.

**Table 2. Hyperparameters of the machine learning models**

Model	Parameters	Reason for Selection
<b>Random Forest (RF)</b>	n_estimators=200 random_state=42	Using 200 trees improves accuracy and stability. RF is robust to noise and imbalanced data. Not specifying a maximum depth allows the model to learn complex patterns.
<b>HistGradientBoosting (HistGB)</b>	max_iter=200 random_state=42	HistGB is a faster version of GBDT. Setting max_iter= 200 provides a good balance between speed and accuracy and is suitable for large datasets.
<b>Logistic Regression</b>	max_iter=500 random_state=42	Increasing max_iter ensures model convergence, especially for datasets with high-dimensional features. It is also simple and fast.
<b>K-Nearest Neighbors (KNN)</b>	n_neighbors=5	Choosing 5 neighbors typically balances bias and variance, preventing overfitting to local decisions.
<b>Support Vector Classifier (SVC)</b>	kernel='linear' probability=True random_state=42	A linear kernel is appropriate for data represented by CNN-extracted features.
<b>LightGBM</b>	n_estimators=200 objective='multiclass' num_class=5 random_state=42	LightGBM is fast and suitable for high-dimensional data. Using 200 trees provides a suitable balance between predictive accuracy and efficiency.
<b>XGBoost</b>	n_estimators=200 use_label_encoder=False eval_metric='mlogloss' random_state=42	The mlogloss metric is appropriate for multi-class classification. XGBoost performs strongly on features extracted from CNNs.

It estimates the probability of a sample belonging to a class using the logistic function to map predicted values to probabilities (13, 14). Given its simplicity, LR is highly interpretable, making it a valuable benchmark for assessing the performance of more complex models like CNNs. It is particularly effective when the relationship between features is linear, but may struggle when applied to more complex, nonlinear datasets.

### ***K-Nearest Neighbors (KNN)***

K-Nearest Neighbors (KNN) is a non-parametric, instance-based learning algorithm that classifies samples based on the majority class among their k-nearest neighbors in the feature space (13). It is particularly effective in capturing the local structure of the data, making it suitable for datasets with complex patterns. One of its significant advantages is its simplicity and ease of implementation. However, KNN can be memory-intensive, especially as the dataset grows, and its prediction speed can decrease significantly with larger datasets. In this study, the use of compact features from CNNs helps to mitigate these limitations.

### ***Support Vector Classifier (SVC)***

Support Vector Classifier (SVC) is a powerful supervised learning algorithm that focuses on finding the optimal hyperplane that separates different classes with the maximum margin. It utilizes support vectors, which are the data points closest to the hyperplane, to define the decision boundary. SVC can employ various kernel functions, such as linear, polynomial, and radial basis function (RBF), to capture complex, nonlinear relationships between features (14). This flexibility allows SVC to perform well in high-dimensional spaces, particularly when the data is appropriately normalized.

### ***LightGBM (LGBM)***

LightGBM (Light Gradient Boosting Machine) is a highly efficient gradient boosting framework designed to handle large datasets

with high dimensionality. It employs a unique histogram-based learning technique that bins continuous feature values, enabling faster training and reduced memory usage. LGBM optimizes the growth of trees using a leaf-wise strategy rather than the traditional depth-wise approach. This method enhances the model's accuracy, particularly for complex datasets, making it a strong candidate for multi-class classification.

### ***XGBoost (XGB)***

XGBoost (Extreme Gradient Boosting) is one of the most widely used and effective boosting algorithms, known for its speed and performance. It incorporates various techniques such as regularization, which prevents overfitting, and parallel processing, which accelerates computation. XGBoost builds trees in a sequential manner, optimizing the loss function at each step to improve accuracy. It is particularly effective for classification and regression tasks involving CNN-extracted features due to its capability to handle complex patterns and interactions in the data while maintaining reasonable training times (14,15).

## **6. Integration of deep and classical models (Hybrid Ensemble)**

### ***6.1. Ensemble voting for prediction fusion***

The fusion of predictions from trained models is performed to improve the accuracy, robustness, and generalizability of the system. In this study, the ensemble integrates both CNN-based models and classical machine learning models, allowing the strengths of each model type to contribute to the final decision. The adopted fusion strategy is known as Voting Ensemble, which can be implemented in three different forms:

**Soft Voting:** The predicted class probabilities from each model are averaged, and the class with the highest combined probability is selected.

**Hard Voting:** The final class label is determined by the majority vote of the individual models for each sample.

**Weighted Voting:** Model predictions are combined using predefined weights (e.g., assigning different weights to CNNs and classical machine learning models).

## 6.2. Hierarchical model aggregation (Stacking Ensemble)

The objective of the stacking ensemble approach is to combine the predictions of classical base models trained on CNN-extracted features using a meta-model (meta-learner). The meta-learner learns in which classes or samples each base model performs better and makes the final decision accordingly. In this study, XGBoost was selected as the meta-model, receiving as input the class probability outputs of each base model along with the original features. Compared to voting ensembles, this approach offers an advantage by explicitly modeling the dependencies, strengths, and weaknesses of the individual models, thereby improving overall classification performance. The stacking ensemble produces the final class label for each sample, and its performance is evaluated using metrics such as accuracy, F1-score, and the confusion matrix.

## Results

To evaluate the performance of the proposed system, a series of experiments were conducted on diabetic retinopathy screening data. Initially, the data were prepared through several preprocessing steps, including class balancing, data augmentation (rotation, flipping, brightness adjustment, and scaling), and partitioning into training, validation, and test sets. For this stage, the NumPy and Pandas libraries were used for data management, OpenCV and scikit-image for image processing, and the Albumentations library for data augmentation. This preprocessing pipeline played a crucial role in preventing overfitting and enhancing the generalization capability of the models. Subsequently, three deep architectures ResNet50, DenseNet121, and ConvNeXt-Base were employed as base models using PyTorch and the torchvision

library. Each model was trained independently on the training set, and its performance was evaluated on the validation set. The initial results indicated that all three architectures were capable of extracting meaningful patterns from the images; however, their standalone classification performance did not reach the desired level for certain classes. To further improve performance, deep feature representations were extracted from these networks and combined into a unified feature set (CNN Feature Ensemble).

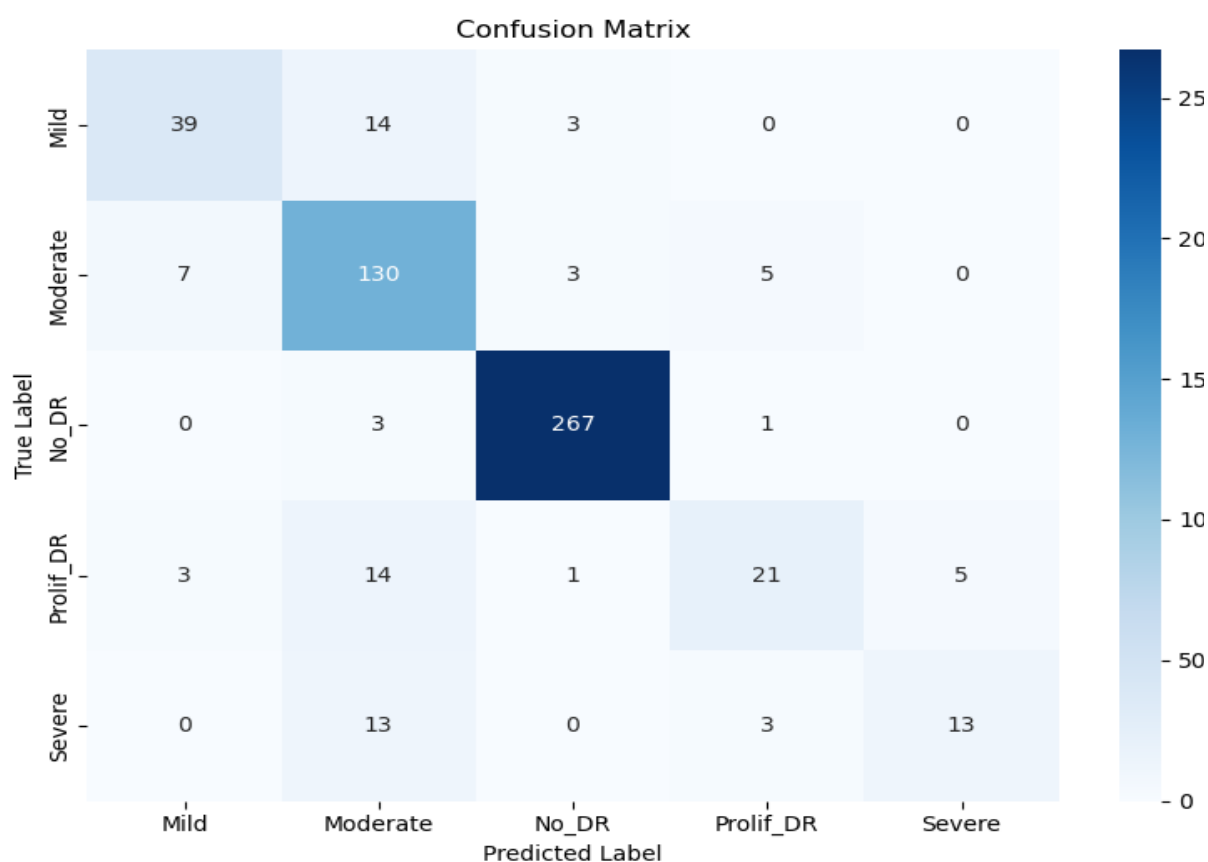
These aggregated features were then used as inputs to a range of classical machine learning models, including Random Forest, KNN, SVM, Logistic Regression, and ensemble voting methods, as well as XGBoost and LightGBM. The performance results of the classical and ensemble models are summarized in Tables 3 and 4.

The results presented in the tables 3 and 4 and Figure 2 indicate that the individual classical models achieve an accuracy of approximately 83-84% and a weighted F1-score of about 82-83%. Subsequently, the adoption of ensemble strategies leads to a noticeable performance improvement. In the ensemble voting approach, predictions from the classical models were combined using three strategies: Soft Voting, Hard Voting, and Weighted Voting.

As shown in Table 4, Hard Voting slightly outperforms the other voting strategies, achieving an accuracy of 83.64% and a weighted F1-score of approximately 83.02%; however, its performance remains inferior to that of the stacking approach. Next, the stacking ensemble method was applied, in which the outputs of the classical models were treated as base learners and an XGBoost model was employed as the final meta-learner. This hierarchical ensemble achieved an accuracy of 85.45% and a weighted F1-score of approximately 84.91%, delivering the best overall performance among all evaluated methods. Figure 2 presents the corresponding confusion matrix. Figure 2 illustrates that the stacking ensemble approach is capable of achieving a high level of accuracy in diabetic

**Table 3. Numerical values of the model comparison across classes based on the f1-score metric**

Model	Level	F1-score Metric				
		Mild	Moderate	No_DR	Proliferate_DR	Severe
RandomForest		0.66	0.8	0.97	0.57	0.43
HistGB		0.62	0.8	0.98	0.58	0.47
LogisticRegression		0.65	0.8	0.97	0.53	0.43
KNN		0.67	0.79	0.97	0.54	0.49
SVC		0.65	0.79	0.98	0.55	0.49
LightGBM		0.65	0.79	0.98	0.58	0.44
XGBoost		0.66	0.79	0.98	0.58	0.45
Stacking		0.72	0.81	0.98	0.58	0.49
Voting (soft)		0.63	0.79	0.97	0.56	0.45
Voting (hard)		0.64	0.79	0.97	0.57	0.45
Voting (weighted)		0.66	0.79	0.97	0.56	0.45



**Figure 2. Confusion matrix of the stacking ensemble method**

retinopathy classification by simultaneously leveraging multiple complementary models.

The strong performance observed in the frequent classes particularly No\_DR and Moderate demonstrates the effectiveness of the meta-learner in integrating information extracted from heterogeneous architectures. Moreover, the fusion of local features extracted

by CNNs contributes to improved model stability and reduced overfitting. In contrast, the model exhibits limitations in classifying advanced disease stages (Severe and Proliferative). A considerable proportion of samples from these classes are misclassified into adjacent categories.

**Table 4. Comparison of models across classes based on different evaluation metrics**

Model	Metric	Mild	Moderate	No_DR	Proliferate_DR	Severe
<b>Random Forest</b>	F1 Score	0.66	0.8	0.97	0.57	0.43
	Precision	0.68	0.73	0.97	0.77	0.5
	Recall	0.64	0.88	0.97	0.45	0.38
	Accuracy	0.8382				
	Weighted F1	0.8320				
<b>Hist GB</b>	F1 Score	0.62	0.8	0.98	0.58	0.47
	Precision	0.66	0.73	0.97	0.72	0.55
	Recall	0.59	0.87	0.98	0.48	0.41
	Accuracy	0.8382				
	Weighted F1	0.8320				
<b>Logestic Regression</b>	F1 Score	0.65	0.8	0.97	0.53	0.43
	Precision	0.69	0.72	0.97	0.75	0.56
	Recall	0.61	0.89	0.98	0.41	0.34
	Accuracy	0.8382				
	Weighted F1	0.8283				
<b>KNN</b>	F1 Score	0.67	0.79	0.97	0.54	0.49
	Precision	0.69	0.74	0.97	0.73	0.54
	Recall	0.66	0.85	0.98	0.43	0.45
	Accuracy	0.8400				
	Weighted F1	0.8340				
<b>SVC</b>	F1 Score	0.65	0.79	0.98	0.55	0.49
	Precision	0.65	0.73	0.97	0.82	0.6
	Recall	0.66	0.87	0.98	0.41	0.41
	Accuracy	0.8400				
	Weighted F1	0.8326				
<b>Light GBM</b>	F1 Score	0.65	0.79	0.98	0.58	0.44
	Precision	0.69	0.73	0.97	0.72	0.48
	Recall	0.62	0.85	0.98	0.48	0.41
	Accuracy	0.8364				
	Weighted F1	0.8315				
<b>XGBoost</b>	F1 Score	0.66	0.79	0.98	0.58	0.45
	Precision	0.68	0.73	0.97	0.72	0.55
	Recall	0.64	0.85	0.98	0.48	0.38
	Accuracy	0.8400				
	Weighted F1	0.8336				
<b>Stacking</b>	F1 Score	0.72	0.81	0.98	0.58	0.49
	Precision	0.75	0.76	0.97	0.72	0.54
	Recall	0.7	0.87	0.99	0.48	0.45
	Accuracy	0.8545				
	Weighted F1	0.8491				
<b>Voting (Soft)</b>	F1 Score	0.63	0.79	0.97	0.56	0.45
	Precision	0.67	0.73	0.97	0.71	0.5
	Recall	0.59	0.85	0.98	0.45	0.41
	Accuracy	0.8327				
	Weighted F1	0.8265				
<b>Voting (Hard)</b>	F1 Score	0.64	0.79	0.97	0.56	0.45
	Precision	0.68	0.73	0.97	0.74	0.5
	Recall	0.61	0.86	0.98	0.45	0.41
	Accuracy	0.8364				
	Weighted F1	0.8302				
<b>Voting (Weighted)</b>	F1 Score	0.63	0.79	0.97	0.56	0.45
	Precision	0.67	0.73	0.97	0.71	0.5
	Recall	0.59	0.85	0.98	0.45	0.41
	Accuracy	0.8327				
	Weighted F1	0.8265				
<b>CNN[4]</b>	F1 Score	0.62	0.78	0.94	0.61	0.47
	Precision	0.66	0.73	0.93	0.72	0.52
	Recall	0.58	0.84	0.95	0.49	0.42
	Accuracy	0.825				
	Weighted F1	0.813				
<b>MLP[2]</b>	F1 Score	0.47	0.73	0.97	0.45	0.46
	Precision	0.56	0.71	0.95	0.42	0.52
	Recall	0.41	0.74	0.98	0.48	0.41
	Accuracy	0.79				
	Weighted F1	0.78				

This behavior can primarily be attributed to three factors, which are expected to be addressed in future research:

1. Severe class imbalance, leading to insufficient learning of discriminative patterns associated with advanced stages.
2. Visual overlap and structural similarity among different DR grades, which complicates the delineation of diagnostic boundaries.
3. Variable retinal image quality, including blur, improper illumination, and low contrast, which results in the attenuation or loss of critical features in the Severe and Proliferative stages.

Despite these limitations, the experimental results indicate that the proposed system based on the combination of three CNN architectures for feature extraction and their integration with classical machine learning models within a hybrid ensemble framework achieves a significant improvement in both accuracy and F1-score, providing an effective approach for the automated detection of diabetic retinopathy.

## Conclusion

Diabetic retinopathy is one of the leading causes of vision loss among patients with diabetes; therefore, early and accurate detection of this disease represents a critical challenge in improving ophthalmic care and preventing blindness. Timely diagnosis can effectively prevent the progression of retinal lesions and enable appropriate therapeutic interventions, whereas delayed detection may result in irreversible damage. With the increasing prevalence of diabetes worldwide, the development of automated and intelligent diabetic retinopathy detection systems particularly those based on deep learning and model ensemble strategies has become increasingly important and can play a significant role in enhancing rapid and accurate patient screening. In this study, an automated diabetic retinopathy detection system was proposed based on the integration of three CNN architectures ResNet50, DenseNet121, and ConvNeXt-Base for feature extraction, followed by their fusion with classical machine

learning models. The experimental results demonstrated that combining features from multiple deep networks and employing ensemble techniques, particularly the stacking ensemble, leads to a substantial improvement in performance compared to individual models. Specifically, standalone classical models achieved an accuracy of approximately 83–84% and a weighted F1-score of 82–83%, whereas the stacking ensemble reached an accuracy and weighted F1-score of 85.45%. This performance gain highlights the importance of leveraging complementary feature representations and hierarchical model aggregation to enhance system generalizability. Furthermore, the use of well-established libraries such as PyTorch, scikit-learn, XGBoost, and LightGBM, along with robust data augmentation and image preprocessing techniques, facilitated model development and yielded reliable results. Nevertheless, certain limitations remain, including dependence on the quality and size of the available datasets and the lack of extensive evaluation in real-world clinical environments. Although the obtained results on publicly available benchmark datasets indicate promising performance, external validation using real patient data is essential to assess the model's generalizability and clinical applicability. As part of future work, collaboration with ophthalmology clinics is planned to conduct a pilot clinical study using newly acquired, previously unseen retinal images. This evaluation will aim to assess the robustness, reliability, and practical effectiveness of the proposed model under real clinical conditions. Such efforts are expected to play a crucial role in preparing the system for deployment in real-world screening and diagnostic workflows. Future research directions may include the adoption of more advanced CNN architectures and Vision Transformers for richer feature extraction, optimization of ensemble strategies through blending or boosted ensembles, integration of retinal images with patient clinical data, expansion of dataset size and diversity, and the deployment of the system in clinical

environments to evaluate real-world performance and physician interaction. Overall, this study demonstrates that a hybrid ensemble framework combining CNNs with classical machine learning models constitutes an effective and practical approach for automated diabetic retinopathy detection, with strong potential for application in clinical screening systems to improve diagnostic accuracy and efficiency.

### **Acknowledgments**

During the preparation of this work, the authors used AI-based tools solely for language refinement, including grammar and spelling checks. All scientific content, methodology, analyses, and conclusions were developed independently by the authors. The AI tools served only as writing assistants. The authors acknowledge and thank these tools for their support in proofreading the final manuscript.

### **Funding**

This paper has received no external funding.

### **Conflict of Interest**

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancy, have been completely witnessed by the authors.

### **Authors' contributions**

M.B conceptualized the study, developed the methodology, and conducted software development, validation, formal analysis, writing, review and editing. F.ZM contributed to the investigation, data curation, writing, review and editing. M.SZ supervised the project, provided manuscript review and feedback, and handled project administration.

## References

1. Lee R, Wong TY, Sabanayagam C. Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss. *Eye and vision*. 2015;2(1):17.
2. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Journal of the American Medical Association*. 2016;316(22):2402-10.
3. Wang Z, Keane PA, Chiang M, Cheung CY, Wong TY, Ting DS. Artificial intelligence and deep learning in ophthalmology. *Artificial intelligence in medicine*. 2022:1519-52.
4. Karsaz A, Mohammadian Roshan S. Deep Convolutional Neural Networks for Diabetic Retinopathy Screening. *Advanced Signal Processing*. 2020;4(2):225-37.
5. Odeh I, Alkasasbeh M, Alauthman M. Diabetic retinopathy detection using ensemble machine learning. *In2021 international conference on information technology (ICIT) 2021*:173-178.
6. Bakasa W, Viriri S. Stacked ensemble deep learning for pancreas cancer classification using extreme gradient boosting. *Frontiers in Artificial Intelligence*. 2023;6:1232640.
7. Revathy R, Nithya BS, Reshma JJ, Ragendhu SS, Sumithra MD. Diabetic retinopathy detection using machine learning. *International Journal of Engineering, Research and Technology*. 2020;9(6):122-6.
8. I. Y. Abushawish, E. Abdel-Raheem, S. Modak, S. A. Mahmoud, and A. J. Hussain, "Deep learning in automatic diabetic retinopathy detection and grading systems: A comprehensive survey and comparison of methods," *IEEE Access*. 2024;12:12345-12367.doi: 10.1109/ACCESS.2024.3415617.
9. Bidwai P, Gite S, Pahuja N, Pahuja K, Kotecha K, Jain N, et al. Multimodal image fusion for the detection of diabetic retinopathy using optimized explainable AI-based Light GBM classifier. *Information Fusion*. 2024;111:102526.
10. Ren Y, Shao D, Yi S. DR-MAE: Self-supervised learning for diabetic retinopathy grading based on masked autoencoder. *Journal of King Saud University Computer and Information Sciences*. 2025;37(8):217.
11. Bhoopalan R, Sekar P, Nagaprasad N, Mamo TR, Krishnaraj R. Task optimized vision transformer for diabetic retinopathy detection and classification in resource constrained early diagnosis settings. *Scientific Reports*. 2025;15(1):39047.
12. Rezaee K, Farnami F. Innovative approach for diabetic retinopathy severity classification: An ai-powered tool using CNN-transformer fusion. *Journal of Biomedical Physics & Engineering*. 2025;15(2):137.
13. Ahmadabadi JZ, Mehrjardi FZ, Ghanbary M, Mirzaei M. Identification of Effective Factors and Prediction of Ischemic Heart Disease Using Machine Learning Methods and Data from the Yazd Health Study (YaHS). *Journal of Shahid Sadoughi University of Medical Sciences*. 2024, 32(7):8067-8079.(in Persian)
14. Akbari Podineh M, Zare Mehrjardi F, Sardari Zarchi M. Multimodal analysis of ECG signals for cardiac arrhythmia detection using machine learning and deep learning methods. *Applied and basic Machine intelligence research*. 2025:17-34.
15. Zare Mehrjardi F, Yazdian-Dehkordi M, Latif A. Evaluating classical machine learning and deep-learning methods in sentiment analysis of Persian telegram message. *Soft Computing Journal*. 2022;11(1):88-105