Physical and Engineering Sciences in Medicine (2021) 44:639–653
https://doi.org/10.1007/s13246-021-01012-3

SCIENTIFIC PAPER

Diabetic retinopathy classification based on multipath CNN and machine learning classifiers

S. Gayathri[1] · Varun P. Gopi[1] · P. Palanisamy[1]

Received: 5 October 2020 / Accepted: 6 May 2021 / Published online: 25 May 2021

Abstract

Eye care professionals generally use fundoscopy to confirm the occurrence of Diabetic Retinopathy (DR) in patients. Early DR detection and accurate DR grading are critical for the care and management of this disease. This work proposes an automated DR grading method in which features can be extracted from the fundus images and categorized based on severity using deep learning and Machine Learning (ML) algorithms. A Multipath Convolutional Neural Network (M-CNN) is used for global and local feature extraction from images. Then, a machine learning classifier is used to categorize the input according to the severity. The proposed model is evaluated across different publicly available databases (IDRiD, Kaggle (for DR detection), and MESSIDOR) and different ML classifiers (Support Vector Machine (SVM), Random Forest, and J48). The metrics selected for model evaluation are the False Positive Rate (FPR), Specificity, Precision, Recall, F1-score, K-score, and Accuracy. The experiments show that the best response is produced by the M-CNN network with the J48 classifier. The classifiers are evaluated across the pre-trained network features and existing DR grading methods. The average accuracy obtained for the proposed work is 99.62% for DR grading. The experiments and evaluation results show that the proposed method works well for accurate DR grading and early disease detection.

Keywords DR grading · Retinal fundus images · Retinal lesions · Multipath CNN (MCNN) · Machine Learning classifiers

Introduction

Diabetic retinopathy (DR) is an illness that causes irreversible vision loss in some people with diabetes mellitus. The increasing glucose level in blood enhances its viscosity, which leads to fluid leakage into the surrounding tissues in the retina. This ultimately results in vision loss. As the disease progresses, lesions (MicroAneurysms (MA), Hemorrhage (HM), Exudates, and neovascularization) appear in the retina. These are considered the central components of DR degradation [1]. Non-Proliferative DR (NPDR) and Proliferative DR (PDR) are the main stages of disease severity. Based on lesions, NPDR is further classified as mild, moderate, and severe [2]. The critical phase with lesions and neovascularization is termed PDR. The changes in the retina at different stages are depicted in Fig. 1. The earliest noticeable sign of DR is the presence of small, red dots called MA in the small blood vessels of the retina [3]. Retinal hemorrhage is another complication of DR that occurs due to hypertension and occlusion of retinal veins. Sometimes the small HMs may resemble MAs. The exudates are yellow flicks composed of lipids and proteins residues that filter out from the damaged capillaries. DR, in its severe phase, is hard to cure. Therefore, it is important to detect DR early on to plan and execute efficient management strategies. Thus, several techniques are being developed to detect and determine the severity of DR lesions. The challenging step is to accurately extract the essential features from fundus images for precise classification of DR. Artificial Neural Network (ANN) architectures have provided elegant solutions to image classification problems, including disease detection using biomedical images. Among the ANN techniques, Convolutional Neural Networks (CNN's) are futuristic deep learning architectures that have led to many breakthroughs in automated object detection and classification. These deep

✉ Varun P. Gopi
varun@nitt.edu

S. Gayathri
gsgayathriunnithan@gmail.com

P. Palanisamy
palan@nitt.edu

[1] National Institute of Technology, Tiruchirappalli, Tamil Nadu, India

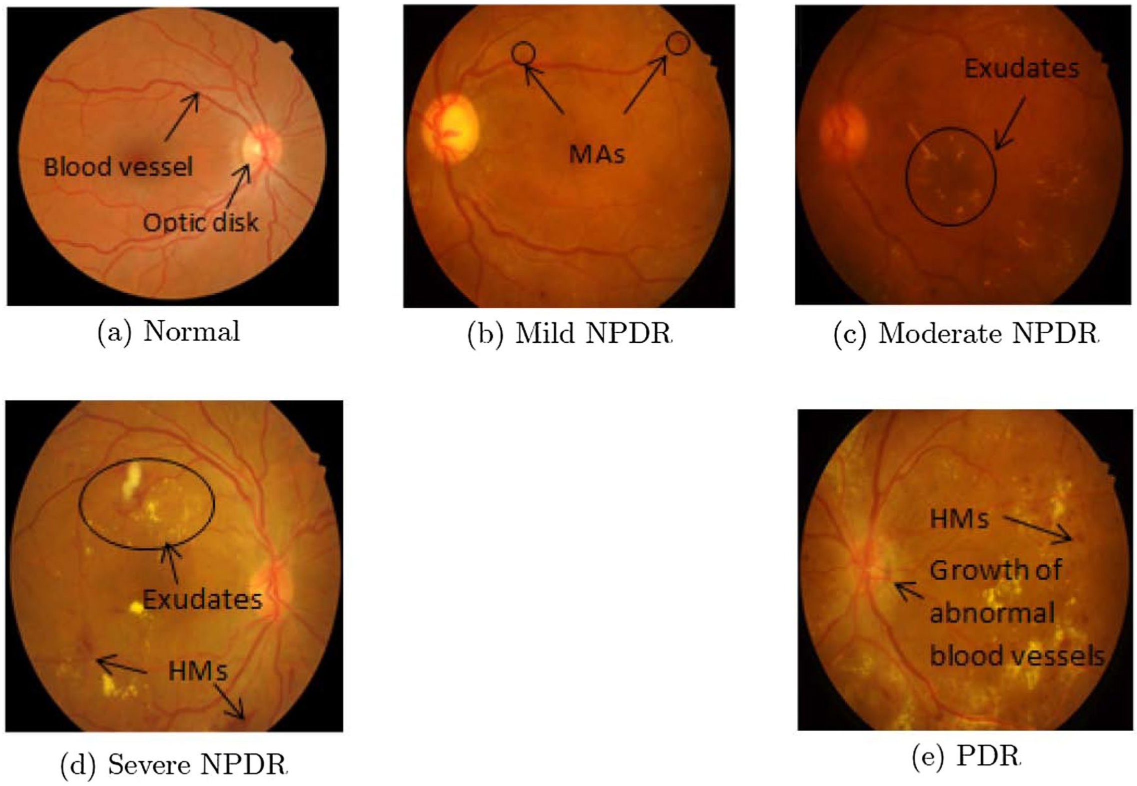

Fig. 1 Different stages of DR in fundus images

learning architectures can extract even minute features that are useful for accurate classification of images. In this work, a Multipath Convolutional Neural Network (M-CNN) is designed to extract DR features from retinal fundus images that can be used in Machine Learning (ML) classifiers for DR classification.

Many research studies are underway to detect and grade DR through neural networking approaches. Some of the groundbreaking DR-classification methods based on different feature extraction techniques are reviewed in this section. A CNN method was proposed by [4] for diagnosing and classifying DR from fundus images based on severity; this method resulted in 75% average accuracy on 5,000 validation images. Another study [5] combined CNN extracted features with support vector machine (SVM) classifiers for lung disease detection (using lung sounds and spectrograms). Yet another study combined CNN with Biometric Pattern Recognition (BPR) [6]. In these two methods, CNN was used as a feature extractor. In [7], the automatic evaluation of the DR severity using ANN was demonstrated. Images of four lesions were extracted and fed into the multilayer feed-forward neural network for grading the disease stages. In [8], a two-stage CNN was used to diagnose abnormal lesions of DR from the fundus image.

Features such as area, perimeter, and count of the DR lesions were extracted in [9], and an ANN was implemented for DR classification into mild, moderate, and severe. The authors of [10] used the findings of a validated red lesion detection method to perform automated classification of DR. Assessment was performed using data from a public database by the leave-one-out validation method and to show the viability of automatic DR screening. In [11], the retinal fundus image was first divided into four sub-images. Haar wavelet transformation was applied to extract features, and better feature selection was achieved using Principal Component Analysis (PCA). Then for DR or No DR classification, a backpropagation neural network and one rule classifier were used. DR screening using a four-layer CNN was proposed in [12]. The results were evaluated by performing cross-validation. The essential five-class grading of DR was implemented in [13] by extracting the Hard exudates area, blood vessels area, texture, entropies, and bifurcation points. For classification, a combination of texture and morphological changes was considered. Probabilistic Neural Network (PNN) classifier parameter (σ) was tuned using a genetic algorithm and particle swarm optimization. A bi-channel CNN for DR detection was proposed in [14]. The green channel component was selected using the unsharp masking method from the original input image, and the red channel component was converted into a greyscale image. Further, these two images were given to CNN for detection purposes.

For DR detection, the authors of [15] demonstrated how to tackle blurred retinal images. They used a regularized filter deblurring algorithm to boost the effectiveness of the technique. The blood vessel, MAs, and exudates areas were then computed to give the ANN classifier input. A DR screening using inception-v3 was explained in [16]. The evaluation was carried out using the Kaggle database. In [17], modified Alexnet architecture was used for DR grading from the retinal fundus images. CNN architecture with sufficient pooling layers was suggested to classify the fundus images according to the disease severity. Local features from the retinal fundus images were extracted in [18] using the Local Binary Patterns (LBP) technique. The detection was then made through ML classifiers, particularly Random Forest (RF), Support Vector Machines (SVM), and ANN.

In [19], a SURF-BRISK combined local feature extraction method was implemented, and the most relevant 30 features were selected using the Minimum Redundancy-Maximum Relevance (MR-MR) method. Then the chosen features were fed into different classifiers for DR classification. Another feature extraction technique using a combination of the Haralick and Anisotropic Dual-Tree Complex Wavelet Transform (ADTCWT) was suggested in [20]. This feature extraction method was a time-consuming process as it necessitated the extraction of features using two complex methods. The DR grading method implemented in [21] used a small CNN architecture for feature extraction, followed by ML classifiers for DR classification using different databases. IDx-DR is the first FDA (Food and Drug Administration) approved autonomous AI system for DR screening. In [22], the validation of IDx-DR device is performed for the screening of Referable DR (RDR) and Vision-Threatening Retinopathy (VTDR). The studies in [23], used spanish population to validate the IDx-DR system for DR screening. According to their observation, the system had high specificity of 100% while sensitivity is of 82%. In recent years, multipath and multiscale neural networks have become common for various classification problems. A multipath-multiscale CNN for pulmonary nodule classification from Computed Tomography (CT) images was introduced in [24]. This method reportedly overcome the high variance of nodule characteristics in the CT images during classification. Thus, multipath architectures could be utilized for adequate feature extraction from images. In [25], a multipath ensemble CNN was designed, and the network's evaluation was carried out on different databases. A 3D-multipath neural network for DR grading was reported in [26]. This work combines the features from Optical Coherence Tomography Angiography (OCTA) Scans, Demographic, and Clinical Bio-markers. The machine learning classifiers were used for DR grading, which resulted in an average accuracy of 96.8%.

Methodology

Efficient automated methods for accurate grading of DR from retinal fundus images are required to detect the disease. Conventional CNN has been used in DR detection and grading in recent years. This method uses convolutional kernels (filters), activation function (usually ReLU), pooling layers, and fully connected layers [27]. For the first time in 2012 by Alex Krizhevsky [28], CNN was proposed as a winning entry to the ILSVRC challenge [29]; CNN's have revolutionized the domain of pattern recognition and data inference, especially in the field of computer vision. This work presents a novel M-CNN architecture for extracting features from the retinal fundus images for DR grading, and popular machine learning classifiers are used to grade the disease. The method called transfer learning via feature extraction [30] is adopted. The classifiers are trained with the extracted M-CNN features. Different classifiers (SVM, Random Forest and, J48) are evaluated by calculating the performance metrics from the corresponding confusion matrices for classifiers. After different stages of evaluation, the best classifier for DR grading with M-CNN features chosen. The proposed work flow is demonstrated in Fig. 2.

Architecture specifications

The proposed CNN architecture is shown in Fig. 3. The image size to the CNN input layer is 196×196. In the proposed network, two feed-forward paths are designed. The first is the main path that resembles conventional CNN. The second is intended for multipath feature extraction. It begins after the first CL and concatenates both the feature maps before the fully connected layers. The activation function ReLU is used as it works with better gradient change than sigmoid and tanh functions [31]. The weighted sum of inputs and biases is computed by the Activation Function (AF), which is used to determine whether a neuron can be fired or not. The first convolutional layer uses a 5×5 convolutional operation using eight kernels. Then the path is branched. The previous layer's feature maps are given to the second path that performs a 9×9 convolutional operation with 32 kernels and a max-pooling layer. The main trail leads to a max-pooling layer and, again, to a 5×5 convolution layer. The number of kernels in the network are chosen after the trial and error procedure. After different trials, the prescribed number of kernels in Fig. 3 provides better DR feature extraction. Some of the crucial trials in the kernel count and size selection are demonstrated in Table 1. The minimum

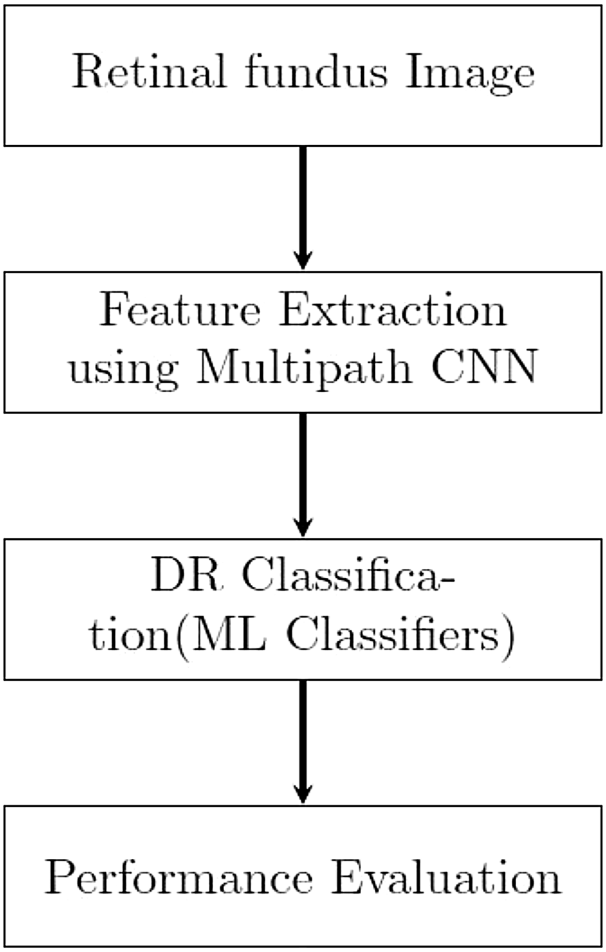

Fig. 2 Proposed method

error rate obtained is 0.97% and those kernel parameters are chosen to build the M-CNN architecture. The max-pooling operation downsamples the input using the maximum value from each cluster of neurons at the previous layer. The downsampling reduces the spatial resolution of the successive layers, helping to preserve the relevant local structures. Then, the main path and the secondary path are eventually concatenated, and the feature maps are given to the weighted transform layer (1×1 without bias) to integrate the features before giving into the Fully Connected Layer (FCL). The first and second FCLs are designed with 128 and 64 hidden neurons, respectively. It was reported in [32] that the softmax classifier degrades the prediction performance of the network. Before taking the features from the second fully connected layer, a dropout (technique to fire out units in a neural network) of 0.5 is used after first fully connected layer to avoid the overfitting during the training of ML classifier. Therefore, a CNN and an ML classifier can improve the entire classification system's efficiency. Hence the 64 features from the second FCL are provided to different classifiers for classification.

Procedure of feature extraction using M-CNN

The extracted features are crucial factors that decide the efficacy of an automated system. A system with the best feature extraction capability can have accurate classification rates. In this work, the M-CNN architecture is designed for extracting the DR features from retinal fundus images for grading the disease stage. So that the issues with the database size can be resolved upto some extent. Generally, the features extracted using a traditional neural network may have losses in the global structures because of the too short or long straight forward path. This can be rectified using shortcut paths. The multipath extracted features help preserve global structures' losses, thereby producing more relevant global and local structures from the M-CNN. After concatenating the feature maps from the two paths, the output competent feature vectors for classification is taken from the second FCL. The main issues facing in multipath feature extraction are (1) A chance to deceive CNN in analyzing the global structures while transferring the features from the current layer to the shortcut path. (2) If the image resolution is poor, the network will become susceptible to global noise interference [24]. These issues can be resolved by including sufficient convolutional layers in the shortcut path. The proposed M-CNN works best with a 9×9 convolutional layer with 32 kernels in the short cut path for DR classification. The implementation of M-CNN is done through Keras [33], a Python-written high-level Application Program Interface (API). An example for the feature maps obtained from the final convolutional layer after concatenation of multi-paths in M-CNN is demonstrated in Fig. 4. for the mild NPDR input image from Messidor database. The 64 feature maps obtained from the final convolutional layer is analyzed to verify the presence of DR features. Even the input size is 196×196 (resized the orginal image size), the M-CNN retains the features that are difficult to visualize with the naked eye.

Feature extraction using pre-trained networks

Pre-trained networks are now available that can be used for classification problems. At the same time, these deep CNNs can be adapted as feature extractors. In such cases, the features are extracted from the intermediate layers. In this work, two pre-trained networks (ResNet-50 [34], VGG-16 [35]) are used for DR feature extraction from retinal fundus images. These networks are used with the pre-trained weights and extract the features from the last pooling layer. Then these features are used in different classifiers for DR classifiers. While analyzing the feature maps in the intermediate layers, it is observed that there is loss of minute features as the network becomes deeper. So, it is necessary to evaluate the effect of such features in DR classification. It is discussed in Sect. 4.3.

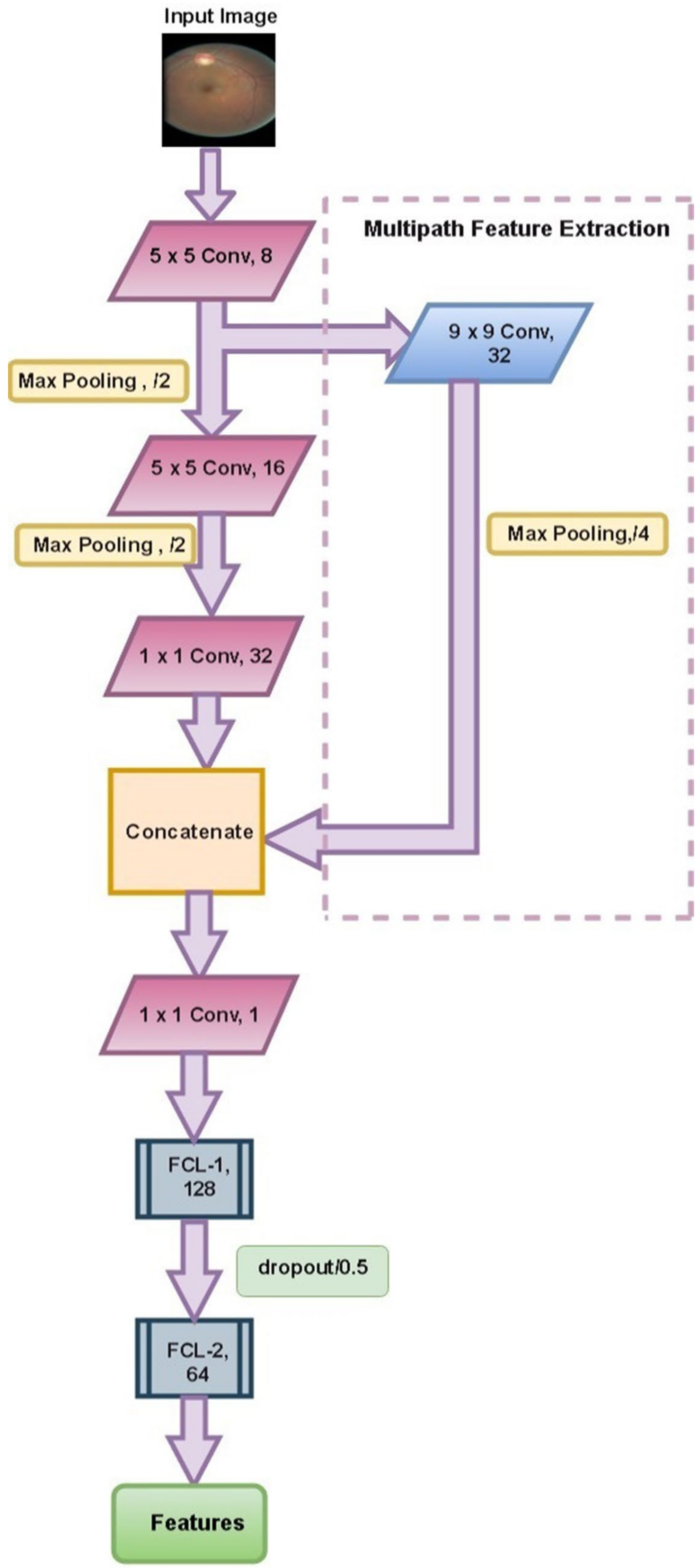

Fig. 3 Proposed M-CNN architecture

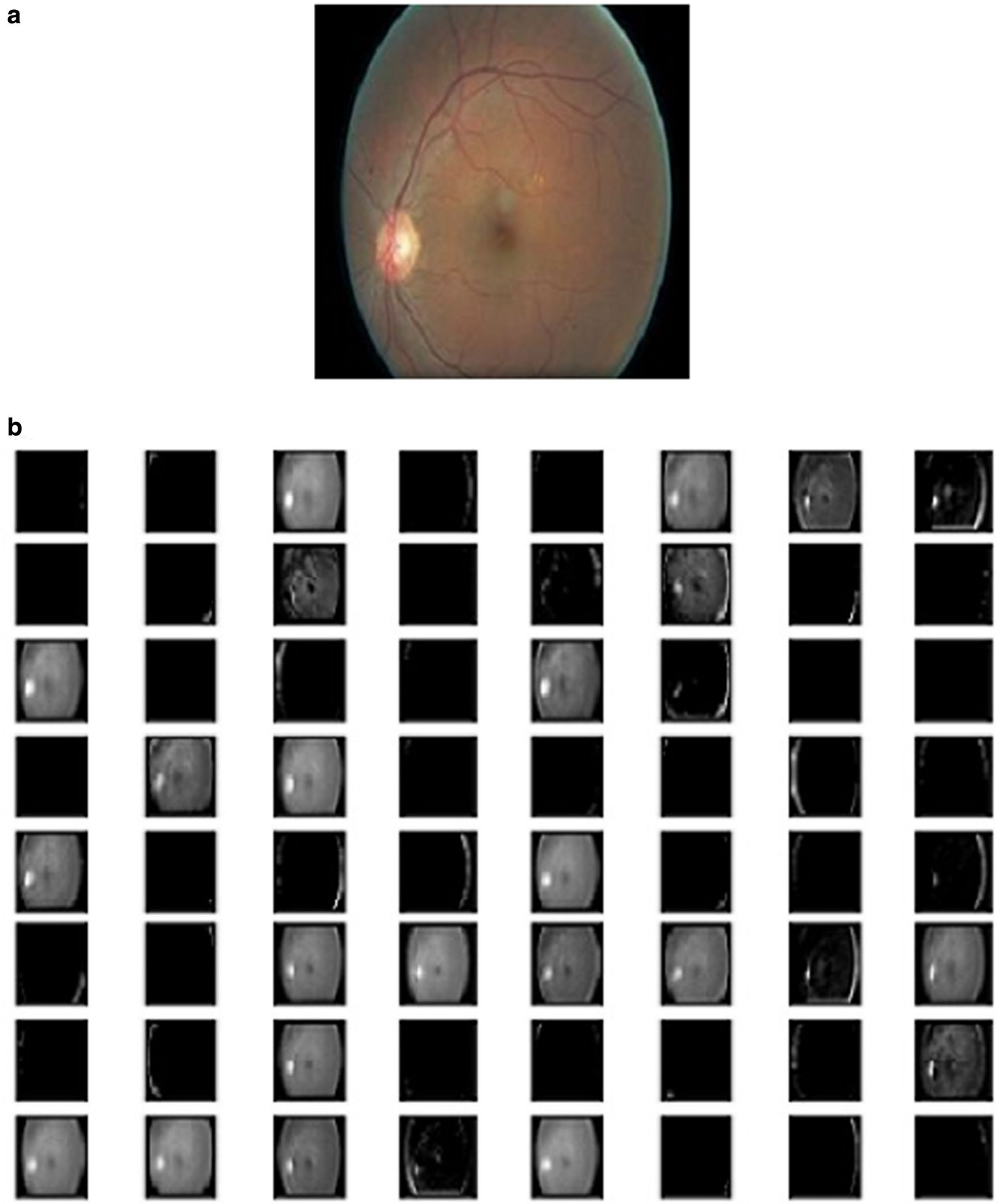

Fig. 4 Visualization of M-CNN Feature map from a mild NPDR image **a** Original Image (Mild NPDR category from Messidor database), **b** Feature Maps

Table 1 Evaluation of different Kernel parameters in the M-CNN layers for DR classification

Convolutional Layer	Kernel Size	# of Kernels	Error Rate (%)
1	9×9	16	3.47
2	9×9	16	
3	1×1	32	
Secondary path	11×11	32	
1	9×9	8	3.12
2	9×9	16	
3	1×1	32	
Secondary path	11×11	32	
1	7×7	8	2.11
2	7×7	16	
3	1×1	32	
Secondary path	9×9	32	
1	5×5	8	1.01
2	5×5	16	
3	1×1	32	
Secondary path	11×11	32	
1	5×5	8	0.97
2	5×5	16	
3	1×1	32	
Secondary path	9×9	32	
1	3×3	8	0.99
2	3×3	16	
3	1×1	32	
Secondary path	9×9	32	

Classifiers

After the feature extraction using the M-CNN method, the images are classified into different categories using machine learning classifiers such as SVM, Random Forest, and J48. The classifiers are trained and validated using the extracted M-CNN features.

Support vector machine (SVM)

The Support Vector Machine (SVM) [36] is a supervised learning strategy relevant to binary classification tasks. The SVM classifier is helpful in situations in which the input data are non-linearly separable in space. It is also suitable for many multiclass classification problems. For this, it uses a "one vs. all" scheme. In this method, the multiclass problem is split into a binary classification problem for each class. The steps for creating the classifier pseudo-code are adopted from [19].

Table 2 Database description

Database	Category	Each category size
IDRiD	Normal	134
	Mild NPDR	20
	Moderate NPDR	136
	Severe NPDR	74
	PDR	49
Kaggle	Normal	25810
	Mild NPDR	2443
	Moderate NPDR	5292
	Severe NPDR	873
	PDR	708
MESSIDOR	Normal	546
	Mild DR	153
	Moderate DR	247
	Severe DR	254

Random forest

Random Forest [37] is an ensemble model classifier with a collection of decision trees structured to have different random vectors [38] for each of them. The steps for obtaining pseudo-code are as described in [19].

J48

J48 is the java version of C4.5 [39] Decision Tree (DT) intended for data mining. In DT, the information gain is the fundamental parameter in the design. The equations related to the J48 classifier are adopted from [20]. During the decision tree construction, the most substantial information gain is picked as the test feature for the current node. To make the classifier more efficient, we use the maximum depth parameter value as 3, finalized through experiments with random values. The decision tree is used for reduced error pruning [40] to reduce the complexity with fewer power nodes. The classification of input feature vectors is carried out according to the conditions during validation/testing.

Performance analysis

A system's performance appraisal is critical as it establishes the efficiency of a new system. In the following steps, the consistency of the proposed model is assessed.

K-fold cross validation

The performance of the classifier is evaluated using the technique called K-fold cross-validation [41]. In order to clear up the issue with imbalanced database in Table 2, stratified random sampling is involved in the cross validation method. In this method of cross validation, the data splitting assures same class distribution in each subset. The cross-validation procedure is followed as in [19].

Evaluation metrics

The evaluation results are stored in the confusion matrix format [42]. For example, consider a "$C \times C$" matrix with P_{ij} as elements (where, i, j $= 1, 2, 3 \ldots$, no. of classes). In this matrix, let J represent True Positives (TP) count, K-False Negatives (FN) count, M and N denote the False Positives (FP) and True Negatives (TN) count respectively. TP and TN present correctly classified data, while FP and FN are the incorrectly classified information. In multiclass classification, the TPs and FPs can be acquired for each actual class i by taking p predicted classes through equation (6). Then the model performance can be analyzed by calculating different evaluation metrics from the confusion matrices.

$$\begin{aligned}
&\# \quad TPs, \quad J_i = P_{ii} \\
&\# \quad FNs, \quad K_i = \Sigma_{j=1}^{p} P_{ij} - J_i \\
&\# \quad FPs, \quad M_i = \Sigma_{j=1}^{p} P_{ji} - J_i \\
&\# \quad TNs, \quad N_i = \Sigma_{j=1}^{p} \Sigma_{k=1}^{n} P_{ik} - J_i - M_i - K_i
\end{aligned} \tag{1}$$

Accuracy determines the overall strength of the system. It is established using Eq. 2:

$$Accuracy_i = \frac{J_i}{J_i + K_i + M_i + N_i} \tag{2}$$

False Positive Rate (FPR) describes the incorrect positive predictions rate during the classification. For an ideal classifier, the FPR is 0.0. It is evaluated from the confusion matrix using the following equation:

$$FPR_i = \frac{M_i}{M_i + N_i} \tag{3}$$

Precision represents the efficiency with which the system makes perfect positive predictions. It is measured as,

$$Precision_i = \frac{J_i}{J_i + M_i} \tag{4}$$

Recall, also called as sensitivity explains how a model prevents FNs effectively.

$$Recall_i = \frac{J_i}{J_i + K_i} \tag{5}$$

F1-score is required to evaluate the model's accuracy when there is imbalanced data input. It determines the harmonic mean of precision and recall.

$$(F1 - score)_i = \frac{2J_i}{2J_i + M_i + K_i} \tag{6}$$

Specificity valuates the effectiveness of preventing false positives(FPs) during the classification. The FPR and specificity total equals 1.

$$Specificity_i = \frac{N_i}{M_i + N_i} \tag{7}$$

Kappa-score (K-score) is the classifier's consistency metric that measures the inter-observer reliability. This is a ratio of observed accuracy (R_O) to predicted accuracy (R_E) and is estimated as:

$$K - score = \frac{(R_O - R_E)}{(1 - R_E)} \tag{8}$$

Except for the FPR, high values of the other measurements reflect an excellent classifier performance. There was some difficulty in using the detailed class efficiency measures for the analysis of the model. Therefore, the weighted average values of the evaluation metrics are calculated for easy evaluation of the system. If M_1 indicates the class 1 (C_1) evaluation metric and M_2 indicates class 2 (C_2) evaluation metric, then the weighted average of metric W_{em} can be written as:

$$W_{em} = \frac{(M_1 * |C_1|) + (M_2 * |C_2|)}{|C_1| + |C_2|} \tag{9}$$

Experimental results

This section evaluates how the M-CNN features influence the performance of SVM, Random Forest, and J48 classifiers for DR grading and to select the classifier that functions best for the DR grading while using M-CNN features. For that, different evaluation metrics are calculated from the confusion matrices obtained after the model's cross-validation.

Database description

The databases used in this work are IDRiD [43], Kaggle (for DR detection) [44], MESSIDOR [45]. The IDRiD contains 413 images, Kaggle database containing 35126 images. The MESSIDOR database includes 1200 images. The database description is given in Table 2. The DR features are extracted from the images in these databases separately and those features are used to train the ML classifiers. The ML classifiers doesn't require a very large database like what a deep neural network requires.

Proposed method of DR multiclass classification

The proposed system for DR grading consists of an M-CNN for feature extraction and ML classifier for DR multiclass classification/severity grading. The designed network is treated as an arbitrary feature extractor. In order to extract the most efficient features, it is required to pre-train the designed M-CNN with fundus images of DR. The M-CNN is pre-trained with a total of 53679 (not used for performance evaluation of the system) DR category fundus images taken from Kaggle, Messidor and IDRiD databases. The best learning rate has been found experimentally to be 0.003. A momentum factor of 0.9 is used to make the training less noisy and converge faster to the objective. The problem of overfitting is compromised by using a dropout factor of 0.5. The network is pre-trained for 100 epochs with multiple iterations to get the optimized weights that make the network a good DR feature extractor. After that the the images are given into the M-CNN pre-trained network for forward propagation. Then from the second fully connected layer the output features are collected. The classifier can be selected only after evaluating different classifiers using these M-CNN features from the images of different databases. The advantage of multiple path CNN is the extraction of local as well as global features. Already deep neural networks are computationally expensive to train. So, this issue is solved by using the M-CNN extracted features in the ML classifiers. The stratified 10-fold cross validation is then applied to evaluate the performance of the mentioned classifiers. More deeper and wider network leads to loss of minute features in the DR image. In this work the minute features are important for mild NPDR classification. So, this specified network architecture paves a way for better DR feature extraction.

Confusion matrices of each classifiers

The primary fact that required in the assessment of a model is the confusion matrix. The confusion matrices that shows the classifier's efficiency are obtained by performing 10- fold cross validation using extracted M-CNN features in each classifier. The performance metrics are further calculated from the corresponding confusion matrices using the basic equations mentioned in Sect. 2.5.2. The confusion matrices obtained for three classifiers using M-CNN extracted features from IDRiD, Kaggle and MESSIDOR databases are provided in Tables 3, 4 and 5 respectively.

Table 3 Confusion matrix for evaluation using IDRiD Database

	Normal	Mild NPDR	Moderate NPDR	Severe NPDR	PDR
(a) SVM					
Normal	127	0	5	2	0
Mild NPDR	0	20	0	0	0
Moderate NPDR	4	0	132	0	0
Severe NPDR	7	0	0	64	3
PDR	0	0	0	19	30
(b) Random Forest					
Normal	126	0	5	3	0
Mild NPDR	1	2	17	0	0
Moderate NPDR	7	1	128	0	0
Severe NPDR	15	0	0	57	2
PDR	13	0	0	23	13
(c) J48					
Normal	134	0	0	0	0
Mild NPDR	0	19	1	0	0
Moderate NPDR	1	0	135	0	0
Severe NPDR	0	0	0	73	1
PDR	0	0	0	0	49

Performance metrics calculation using proposed feature extraction

The evaluation metrics described in Sect. 2.5.2 are calculated from the confusion matrices obtained for each classifier. The different measures used for better classification efficacy are FPR, Specificity, Precision, Recall, F1-Score, and Accuracy. The specificity and recall are needed to understand the test's strength of the classifier. Specificity determines the proportion of the actual negatives, and recall determines the proportion of actual positives that are predicted correctly. The F1-score is the combined metric of precision and recall. In giving weightage to both precision and recall values, this F1-score can be considered an evaluation metric rather than an accuracy metric. The detailed metrics are tabulated in Table 6 for IDRiD database, Table 7 for Kaggle database, and Table 8 for MESSIDOR database. The weighted average values for each metric are calculated from the detailed efficiency measures and shown in Table 9. The other performance metrics for each classifier, such as validation accuracy and K-score, are depicted in Table 10.

Table 4 Confusion matrix for evaluation using Kaggle database

	Normal	Mild NPDR	Moderate NPDR	Severe NPDR	PDR
(a) SVM					
Normal	25706	0	84	13	7
Mild NPDR	0	2388	55	0	0
Moderate NPDR	61	80	5151	0	0
Severe NPDR	504	0	0	303	66
PDR	37	0	0	432	239
(b) Random forest					
Normal	25810	0	0	0	0
Mild NPDR	1	2392	50	0	0
Moderate NPDR	2	0	5290	0	0
Severe NPDR	520	0	0	350	3
PDR	469	0	0	9	230
(c) J48					
Normal	25809	1	0	0	0
Mild NPDR	0	2443	0	0	0
Moderate NPDR	1	0	5291	0	0
Severe NPDR	0	0	0	872	1
PDR	0	0	0	0	708

The inter-rater reliability implied by K-score represents the extent to which the values gathered in the experiment are accurate representations of the calculated data. The K-score ranges from 0.81 to 1.00 represent almost perfect agreement [46]. Training a CNN with small database doesn't produce a good classification model. At the same time, CNNs are capable of extracting minute features from images. When comparing the results obtained using different database, it is clear that the database size has an important role in modeling a perfect classifier. The IDRiD database contains only 413 images. This is not enough to train the M-CNN. In this work the proposed M-CNN is used for DR feature extraction and the extracted features from 413 images are used to train the ML classifiers. The same is done for the other two databases.

Table 5 Confusion matrix for evaluation using MESSIDOR database

	Normal	Mild DR	Moderate DR	Severe DR
(a) SVM				
Normal	539	0	7	0
Mild DR	0	128	25	0
Moderate DR	23	2	222	0
Severe DR	13	0	0	241
(b) Random Forest				
Normal	546	0	0	0
Mild DR	6	112	35	0
Moderate DR	6	1	240	0
Severe DR	49	0	1	204
(c) J48				
Normal	545	1	0	0
Mild DR	0	153	0	0
Moderate DR	1	0	246	0
Severe DR	0	0	0	254

Performance metrics calculation using pre-trained network feature extraction

There are many existing CNNs that produce good results in different classification problems. The features extracted using the ResNet-50 and VGG-16 networks are fed into each classifier. Then the system is evaluated by applying K-fold cross valuation. The K-score and the accuracy obtained for each classifier are given in Table 11. The IDRiD and MESSIDOR databases are used for the experiments.

The DR classification performance of ResNet-50 and VGG-16 via the transfer learning method is also analyzed. For fine tuning the network, the fully connected layers in the ResNet-50 and VGG-16 are removed. Then a global average pooling layer is added to the output of the backbone model. The overfitting is avoided at this layer as there is no parameter to optimize in the global average pooling. Then three fully connected layers are used with batch normalization between them. The first fully connected layer consists of 512 nodes, second one with 64 nodes. A drop out of 0.5 is used after the second fully connected layer. Then the last fully connected layer using 'softmax' activation with number of nodes equal to the number of classes is added for the final DR classification. Table 12 shows the results of transfer learning based DR classification for different databases.

Table 6 Detailed efficiency measures calculated from confusion matrix of IDRiD Database

Classifier	Precision	Recall	Specificity	F1-Score	FPR	Class
SVM	0.920	0.948	0.961	0.934	0.039	Normal
	–	0.00	1.00	–	0.00	Mild NPDR
	0.841	0.971	0.910	0.901	0.090	Moderate NPDR
	0.753	0.865	0.938	0.805	0.062	Severe NPDR
	0.992	0.732	0.612	0.909	0.008	PDR
Random Forest	0.778	0.940	0.871	0.851	0.129	Normal
	0.667	0.100	0.997	0.174	0.003	Mild NPDR
	0.853	0.941	0.921	0.895	0.079	Moderate NPDR
	0.687	0.770	0.923	0.726	0.077	Severe NPDR
	0.005	0.995	0.867	0.265	0.406	PDR
J48	**0.993**	**0.993**	**0.996**	**0.993**	**0.004**	**Normal**
	1.000	**0.950**	**1.000**	**0.974**	**0.000**	**Mild NPDR**
	0.993	**0.993**	**0.996**	**0.993**	**0.004**	**Moderate NPDR**
	0.986	**0.986**	**0.997**	**0.986**	**0.003**	**Severe NPDR**
	0.980	**1.00**	**0.997**	**0.990**	**0.003**	**PDR**

Table 7 Detailed efficiency measures calculated from confusion matrix of Kaggle Database

Classifier	Precision	Recall	Specificity	F1-Score	FPR	Class
SVM	0.977	0.996	0.935	0.986	0.065	Normal
	0.968	0.977	0.973	0.998	0.002	Mild NPDR
	0.974	0.973	0.995	0.974	0.005	Moderate NPDR
	0.405	0.347	0.987	0.374	0.013	Severe NPDR
	0.766	0.338	0.998	0.469	0.002	PDR
Random Forest	0.963	1.00	0.894	0.981	0.106	Normal
	1.00	0.979	1.00	0.989	0.00	Mild NPDR
	0.991	1.00	0.998	0.995	0.002	Moderate NPDR
	0.975	0.401	1.00	0.568	0.00	Severe NPDR
	0.987	0.325	1.00	0.489	0.00	PDR
J48	**1.00**	**1.00**	**1.00**	**1.00**	**0.00**	**Normal**
	1.000	**1.00**	**1.00**	**1.00**	**0.00**	**Mild NPDR**
	1.00	**1.00**	**1.00**	**1.00**	**0.00**	**Moderate NPDR**
	0.999	**0.999**	**1.00**	**0.999**	**0.000**	**Severe NPDR**
	0.999	**1.00**	**1.00**	**0.999**	**0.00**	**PDR**

Table 8 Detailed efficiency measures calculated from confusion matrix of MESSIDOR Database

Classifier	Precision	Recall	Specificity	F1-Score	FPR	Class
SVM	0.937	0.987	0.945	0.962	0.055	Normal
	0.985	0.837	0.998	0.905	0.002	Mild DR
	0.874	0.899	0.966	0.886	0.034	Moderate DR
	1.00	0.949	1.00	0.974	0.00	Severe DR
Random Forest	0.90	1.00	0.907	0.947	0.093	Normal
	0.991	0.732	0.999	0.842	0.001	Mild DR
	0.870	0.972	0.962	0.918	0.038	Moderate DR
	1.00	0.803	0.891	1.00	0.00	Severe DR
J48	**0.998**	**0.998**	**0.998**	**0.998**	**0.002**	**Normal**
	1.00	**0.993**	**1.00**	**0.997**	**0.00**	**Mild DR**
	0.996	**0.996**	**0.999**	**0.996**	**0.001**	**Moderate DR**
	0.996	**1.00**	**0.999**	**0.998**	**0.001**	**Severe DR**

Table 9 Weighted average values from the detailed efficiency measures

Database	Classifier	Precision	Recall	Specificity	F1 Score	FPR
IDRiD	SVM	–	0.855	0.945	–	0.055
	Random Forest	0.792	0.789	0.917	0.758	0.083
	J48	**0.990**	**0.990**	**0.997**	**0.990**	**0.003**
Kaggle	SVM	0.957	0.962	0.951	0.958	0.049
	Random Forest	0.971	0.970	0.921	0.964	0.079
	J48	**1.00**	**1.00**	**1.00**	**1.00**	**0.00**
MESSIDOR	SVM	0.944	0.942	0.968	0.941	0.032
	Random Forest	0.926	0.918	0.95	0.916	0.050
	J48	**0.998**	**0.998**	**0.999**	**0.997**	**0.001**

Table 10 Validation accuracy and Kappa Score for each classifier with M-CNN feature extraction using different databases

Database	Classifier	Validation Accuracy	Kappa Score
IDRiD	SVM	85.47	0.798
	Random Forest	78.93	0.704
	J48	99.03	0.991
Kaggle	SVM	96.18	0.909
	Random Forest	96.99	0.927
	J48	99.9	0.999
MESSIDOR	SVM	94.16	0.915
	Random Forest	91.83	0.879
	J48	99.75	0.998

Discussions

Confusion matrix analysis

From the confusion matrices, it is seen that the J48 classifier is capable of more effective DR grading than the others. Suppose the mild NPDR and the Normal categories are analyzed, in that case, it is clear that the proposed M-CNN feature extraction with the J48 classifier is effective in detecting DR. Another important factor is that the classification using J48 classifier gives better results in early detection of DR. While analyzing the confusion matrices the mild NPDR and normal images are almost classified correctly. So, the model has the capability of early detection of DR.

Efficiency evaluation

While analyzing the evaluation metrics results in Sect. 3.4, it is clear that the SVM classifier does not perform well in DR grading for all the databases. The Random Forest classifier performs better than the SVM classifier. SVM uses the "one vs. all" method in multiclass problems, which induces difficulties in analyzing the output. Random Forests can handle categorical features well, and therefore, in multiclass problems, it outperforms SVM to some extent. J48, which gives the highest specificity, precision, recall, and F1-score for our classification problem, works best with M-CNN features for DR grading. The FPR obtained for the J48 classifier is nearly 0 in the case of all databases. According to the initial evaluation, the M-CNN features with the J48 classifier are suitable

Table 11 Evaluation of the classifiers using pre-trained network extracted features for DR Grading

Pre-trained network	Database	Classifier	Kappa statistic	Accuracy(%)
ResNet-50 [34]	IDRiD	SVM	0.595	70.46
		Random Forest	0.434	60.29
		J48	0.901	92.46
	MESSIDOR	SVM	0.76	83.75
		Random Forest	0.251	56.25
		J48	0.892	91.22
VGG-16 [35]	IDRiD	SVM	0.594	70.46
		Random Forest	0.434	60.29
		J48	0.701	79.18
	MESSIDOR	SVM	0.763	81.75
		Random Forest	0.251	56.25
		J48	0.734	83.23

Table 12 Evaluation of pre-trained networks via transfer learning for DR Grading

Pre-trained network	Database	Kappa statistic	Accuracy(%)
ResNet-50 [34]	IDRiD	0.42	47
	MESSIDOR	0.58	65
	Kaggle	0.71	76
VGG-16 [35]	IDRiD	0.44	51.5
	MESSIDOR	0.61	68
	Kaggle	0.73	79.5

for DR grading. In the evaluation of performance metric for each classifier in Table 10, the J48 classifier can be seen to have a validation accuracy of above 99% using all databases. The K-score is also higher for the J48 classifier. When the other classifiers show K- score of less than 0.95, the J48 classifiers offer K- score above 0.98, which shows the proposed model's markable efficacy. The results are highlighted in the performance evaluation tables to notice the efficiency of J48 classifier than the other classifiers.

DR grading using pre-trained networks

Evaluation of DR grading using pre-trained networks is performed in this section. The pre-trained networks are used as feature extractors of DR and also used as a DR multi-class classifier utilizing the transfer learning method. On analyzing the Table 11, ResNet-50 features using IDRiD, the J48 classifier performs well with a K-score of 0.901 and an accuracy of 92.46%. But for the MESSIDOR database, the J48 performs with a K-score of 0.892 and an accuracy of 91.22%. On analyzing the VGG-16 features using IDRiD and MESSIDOR databases, the SVM classifier performs better than the other classifiers. However, the K-score and classifier accuracy for SVM is low. The Table 12 shows the performance of pre-trained CNNs in the DR grading. The Kaggle, MESSIDOR and IDRiD databases are used for evaluation. The results shows that the difference in the number of layers and the database size affects the DR classification performance of existing pre-trained CNNs. Compare to ResNet-50, VGG-16 shows more accuracy using Kaggle database. In the case of medical images, as the network becomes deeper there might be chance of losing the useful features. So the classification performance degrades in deeper networks [47]. Then also the performance of proposed method is better than the VGG-16.

Comparison of the proposed (M-CNN + J48) system with existing DR classification methods

Many methods are implemented for DR classification using different techniques. The relevant existing methods are compared with the efficiency of the proposed method, and the comparisons are tabulated in Table 13. The technique proposed in [20] shows almost the same ability as in the proposed work. But the time complexity for extracting features is less in the proposed method because we use two different feature extraction methods to obtain the features. The ADTCWT and Haralick are functional feature extractors,

Table 13 Comparison of proposed work with existing methods

Database	Methods	Precision	Recall	Specificity	F1-Score	K-score	Accuracy(%)
IDRiD [43]	ResNet-50 [34] + J48	–	–	–	–	0.901	92.46
	VGG-16 [35] +J48	–	–	–	–	0.701	79.18
	S. Gayathri et al. [19]	0.925	0.920	0.969	0.908	–	92.01
	S. Gayathri et al. [21]	0.990	0.990	0.997	0.990	0.986	99.03
	Proposed work(M-CNN+J48)	**0.990**	**0.990**	**0.997**	**0.990**	**0.987**	**99.03**
MESSIDOR [45]	Shanthi et al. [17]	92.07	0.923	0.97	–	–	96.25
	ResNet-50 [34] + J48	–	–	–	–	0.892	91.22
	VGG-16 [35] + J48	–	–	–	–	0.734	81.23
	Gayathri et al. [19]	0.982	0.983	0.993	0.982	–	98.28
	Gayathri et al. [21]	0.998	0.998	0.999	0.998	0.996	99.75
	Gayathri et al. [20]	0.998	0.998	0.999	0.998	–	99.75
	Proposed work(M-CNN+J48)	**0.998**	**0.998**	**0.999**	**0.998**	**0.996**	**99.75**
Kaggle [44]	Y. Yang et al [48]	–	–	–	–	0.75	56
	Zeng et al. [49]	–	0.822	70.7	–	0.82	–
	Li et al. [50]	–	–	–	–	–	86.17
	Gayathri et al. [21]	1.00	1.00	1.00	1.00	–	99.9
	Gayathri et al. [20]	0.99	0.99	1.00	0.99	–	99.9
	Proposed work(M-CNN+J48)	**1.00**	**1.00**	**1.00**	**1.00**	0.99	99.9

but it requires more time for extracting the features than our M-CNN feature extraction method. When considering both the time consumption and classification efficacy, it can be concluded that the proposed model is the fastest method that can be used for DR grading. The proposed system works better than using the M-CNN alone for the DR classification. The features are collected from the second fully connected layers and replaced the third fully connected layer and softmax function with ML classifiers; this helps reduce the model's time complexity. The efficiency measures obtained for the proposed work are almost the same as those reported in [21]. While considering the early detection of DR, the proposed method works best in classifying normal and mild NPDR images, which is a milestone in DR classification. From the comparative analysis and the efficiency measures of the proposes system, it can be seen that the proposed system works best for early and fast DR classification.

Conclusion

DR is vision-threatening morbidity that has become widely prevalent in recent times. This work proposes an automated early DR diagnosis and fast grading technique. M-CNN extraction and ML classifier are used to extract relevant features from fundus images and classify the lesions according to their severity levels. The model is analyzed using IDRiD, Kaggle, and MESSIDOR databases. The ML classifiers used in the experiments are SVM, Random Forest, and J48. The features extracted using pre-trained networks are also used for evaluation. The proposed method exhibits an average validation accuracy of 99.62% and a K-score of 0.995. After many experiments, it is seen that the M-CNN features show the best performance with the J48 classifier. The M-CNN and J48 classifier combination can thus be used for early and fast automatic prediction, and grading of DR. This multipath network can be modified to predict other retinal diseases, enhancing the retinal health care monitoring system.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval For this type of study, formal consent is not required.

Informed consent This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Abramoff MD, Garvin MK, Sonka M (2010) Retinal imaging and image analysis. IEEE Rev Biomed Eng 3:169–208
2. Zachariah S, Wykes W, Yorston D (2015) Grading diabetic retinopathy (dr) using the Scottish grading protocol. Commun Eye Health 28:72–73
3. Cheung N, Jin Wang J, Klein R, Couper DJ, Richey Sharrett A, Wong TY (2007) Diabetic retinopathy and the risk of coronary heart disease. Diabetes Care 30(7):1742–1746
4. Pratt H, Coenen F, Broadbent DM, Harding SP, Zheng Y (2016) Convolutional neural networks for diabetic retinopathy. Proc Comput Sci 90:200–205 (20th Conference on Medical Image Understanding and Analysis (MIUA 2016))
5. Demir F, Sengur A, Bajaj V (2020) Convolutional neural networks based efficient approach for classification of lung diseases. Health Inf Sci Syst 8(1):4
6. Zhou L, Li Q, Huo G, Zhou Y (2017) Image classification using biomimetic pattern recognition with convolutional neural networks features. Comput Intell Neurosci 2017
7. James J, Sharifahmadian E, Shih L (2018) Automatic severity level classification of diabetic retinopathy. Int J Comput Appl 180:30–35
8. Yang Y, Li T, Li W, Wu H, Fan W, Zhang W (2017) Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks. In: MICCAI
9. Paing MP, Choomchuay S, Rapeeporn Y (2016) Detection of lesions and classification of diabetic retinopathy using fundus images. In: 2016 9th biomedical engineering international conference (BMEiCON), pp 1–5
10. Seoud L, Chelbi J, Cheriet F (2015) Automatic grading of diabetic retinopathy on a public database
11. Prasad DK, Vibha L, Venugopal KR (2015) Early detection of diabetic retinopathy from digital retinal fundus images. In: 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS), pp 240–245
12. Andonová M, Pavlovičová J, Kajan S, Oravec M, Kurilová V (2017) Diabetic retinopathy screening based on cnn. In: 2017 International Symposium ELMAR, pp 51–54
13. Mookiah MRK, Rajendra Acharya U, Joy Martis R, Chua CK, Lim CM, Ng EYK, Laude A (2013) Evolutionary algorithm based classifier parameter tuning for automatic diabetic retinopathy grading: A hybrid feature extraction approach. Knowl-Based Syst 39:9–22
14. Pao S-I, Lin H-Zin, Chien K-H, Tai M-C, Chen J-T, Lin G-M (2020) Detection of diabetic retinopathy using bichannel convolutional neural network. J Ophthalmol
15. Shahin EM, Taha TE, Al-Nuaimy W, El Rabaie S, Zahran OF, El-Samie FEA (2012) Automated detection of diabetic retinopathy in blurred digital fundus images. In: 2012 8th International Computer Engineering Conference (ICENCO), pp 20–25
16. Kanungo YS, Srinivasan B, Choudhary S (2017) Detecting diabetic retinopathy using deep learning. In: 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT), pp 801–804
17. Shanthi T, Sabeenian RS (2019) Modified alexnet architecture for classification of diabetic retinopathy images. Comput Electr Eng 76:56–64
18. de la Calleja J, Tecuapetla L, Auxilio Medina M, Bárcenas E, Urbina Nájera AB (2014) LBP and machine learning for diabetic retinopathy detection 8669:110–117

19. Gayathri S, Gopi Varun P, Palanisamy P (2020) Automated classification of diabetic retinopathy through reliable feature selection. Phys Eng Sci Med pp 1–19
20. Gayathri S, Krishna AK, Gopi VP, Palanisamy P (2020) Automated binary and multiclass classification of diabetic retinopathy using Haralick and multiresolution features. IEEE Access 8:57497–57504
21. Gayathri S, Gopi VP, Palanisamy P (2020) A lightweight CNN for diabetic retinopathy classification from fundus images. Biomed Signal Process Control 62:102115
22. Van Der Heijden AA, Abramoff MD, Verbraak F, van Hecke MV, Liem A, Nijpels G (2018) Validation of automated screening for referable diabetic retinopathy with the idx-dr device in the hoorn diabetes care system. Acta Ophthalmol 96(1):63–68
23. Shah A, Clarida W, Amelon R, Hernaez-Ortega MC, Navea A, Morales-Olivas J, Dolz-Marco R, Verbraak F, Jorda PP, van der Heijden Amber A, et al (2020) Validation of automated screening for referable diabetic retinopathy with an autonomous diagnostic artificial intelligence system in a Spanish population. J Diab Sci Technol 1932296820906212
24. Wang Y, Zhang H, Chae KJ, Choi Y, Jin GY, Ko S-B (2020) Novel convolutional neural network architecture for improved pulmonary nodule classification on computed tomography. Multidimensional Systems and Signal Processing 1–21
25. Wang X, Bao A, Cheng Y, Yu Q (2018) Multipath ensemble convolutional neural network. IEEE Trans Emerg Topics Comput Intell
26. Eladawi N, Elmogy M, Ghazal M, Fraiwan L, Aboelfetouh A, Riad A, Sandhu H, El-Baz A (2019) Diabetic retinopathy grading using 3d multi-path convolutional neural network based on fusing features from octa scans, demographic, and clinical biomarkers. In: 2019 IEEE International conference on imaging systems and techniques (IST), IEEE. pp 1–6
27. O'Shea K, Nash R (2015) An introduction to convolutional neural networks. ArXiv e-prints
28. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
29. Russakovsky O, Deng J, Hao S, Krause J, Satheesh S, Ma S et al (2015) Imagenet large scale visual recognition challenge. Int J Comput Vision 115(3):211–252
30. Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E (2018) Deep learning for computer vision: A brief review. Computational intelligence and neuroscience 2018
31. Nwankpa C, Ijomah W, Gachagan A, Marshall S (2018) Activation functions: Comparison of trends in practice and research for deep learning. arXiv:1811.03378
32. Tang Y (2013) Deep learning using linear support vector machines. arXiv preprint arXiv:1306.0239
33. Chollet François (2015) keras. https://github.com/fchollet/keras
34. Targ S, Almeida D, Lyman K (2016) Resnet in resnet: Generalizing residual architectures. arXiv:1603.08029
35. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. CoRR, arXiv:1409.1556
36. Pradeep KJ, Balamurali S, Kadry R, Lakshmana K (2019) Diagnosis of diabetic retinopathy using multi level set segmentation algorithm with feature extraction using svm with selective features. Multimedia Tools and Applications 1573–7721
37. Roychowdhury A, Banerjee S (2018) Random forests in the classification of diabetic retinopathy retinal images. In: Bhattacharyya S, Gandhi T, Sharma K, Dutta P (eds) Advanced Computational and Communication Paradigms, vol 475. Springer, Singapore, pp 168–176
38. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
39. Sharma S, Agrawal J, Sharma S (2013) Classification through machine learning technique: C4. 5 algorithm based on various entropies. Int J Comput Appl 82:28–32
40. Elomaa T, Kaariainen M (2001) An analysis of reduced error pruning. J Artif Intell Res 15:163–187
41. Yadav S, Shukla S (2016) Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In: 2016 IEEE 6th International Conference on Advanced Computing (IACC), pp 78–83
42. Visa S, Ramsay B, Ralescu A, Knaap E (2011) Confusion matrix-based feature selection. CEUR Workshop Proc 710:120–127
43. Porwal P, Pachade S, Kamble R, Kokare M, Deshmukh G, Sahasrabuddhe V, Meriaudeau F (2018) Indian diabetic retinopathy image dataset (IDRiD): A database for diabetic retinopathy screening research. Data 3:1–8
44. Kaggle and EyePacs (2015) Kaggle diabetic retinopathy detection
45. Decencière E, Zhang X, Cazuguel G, Lay B, Cochener B, Trone C, Gain P, Ordonez R, Massin P, Erginay A, Charton B, Klein J-C (2014) Feedback on a publicly distributed database: the messidor database. Image Anal Stereol 33(3):231–234
46. McHugh M (2012) Interrater reliability: The kappa statistic. Biochemia medica: časopis Hrvatskoga društva medicinskih biokemičara / HDMB 22:276–82
47. Study of convolutional neural networks for early detection of diabetic retinopathy (2020)
48. Yang Y, Li T, Li W, Wu H, Fan W, Zhang W (2017) Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp 533–540
49. Zeng X, Chen H, Luo Y, Ye W (2019) Automated diabetic retinopathy detection based on binocular Siamese-like convolutional neural network. IEEE Access 7:30744–30753
50. Li Y-H, Yeh N-N, Chen S-J, Chung Y-C (2019) Computer-assisted diagnosis for diabetic retinopathy based on fundus images using deep convolutional neural network. Mobile Information Systems 2019