5-1-2020

Hybrid machine learning architecture for automated detection and grading of retinal images for diabetic retinopathy

Barath Narayanan
University of Dayton

Barath Narayanan
University of Dayton

Russell C. Hardie
University of Dayton, rhardie1@udayton.edu

Manawaduge Supun De Silva
University of Dayton

Nathaniel K. Kueterman
University of Dayton

Hybrid machine learning architecture for automated detection and grading of retinal images for diabetic retinopathy

Barath Narayanan Narayanan,[a,b,*] Russell C. Hardie,[a] Manawaduge Supun De Silva,[a] and Nathaniel K. Kueterman[a]

[a]University of Dayton, Department of Electrical and Computer Engineering, Dayton, Ohio, United States

[b]University of Dayton Research Institute, Sensors and Software Systems Division, Dayton, Ohio, United States

Abstract

Purpose: Diabetic retinopathy is the leading cause of blindness, affecting over 93 million people. An automated clinical retinal screening process would be highly beneficial and provide a valuable second opinion for doctors worldwide. A computer-aided system to detect and grade the retinal images would enhance the workflow of endocrinologists.

Approach: For this research, we make use of a publicly available dataset comprised of 3662 images. We present a hybrid machine learning architecture to detect and grade the level of diabetic retinopathy (DR) severity. We also present and compare simple transfer learning-based approaches using established networks such as AlexNet, VGG16, ResNet, Inception-v3, NASNet, DenseNet, and GoogLeNet for DR detection. For the grading stage (mild, moderate, proliferative, or severe), we present an approach of combining various convolutional neural networks with principal component analysis for dimensionality reduction and a support vector machine classifier. We study the performance of these networks under different preprocessing conditions.

Results: We compare these results with various existing state-of-the-art approaches, which include single-stage architectures. We demonstrate that this architecture is more robust to limited training data and class imbalance. We achieve an accuracy of 98.4% for DR detection and an accuracy of 96.3% for distinguishing severity of DR, thereby setting a benchmark for future research efforts using a limited set of training images.

Conclusions: Results obtained using the proposed approach serve as a benchmark for future research efforts. We demonstrate as a proof-of-concept that an automated detection and grading system could be developed with a limited set of images and labels. This type of independent architecture for detection and grading could be used in areas with a scarcity of trained clinicians based on the necessity.

 [DOI: 10.1117/1.JMI.7.3.034501]

Keywords: diabetic retinopathy; computer-aided detection; endocrinology; convolutional neural networks; support vector machine; principal component analysis.

Paper 19320R received Dec. 16, 2019; accepted for publication Jun. 10, 2020; published online Jun. 23, 2020.

1 Introduction

According to the World Health Organization (WHO), ~347 million people suffer from diabetes across the world.[1] Diabetic retinopathy (DR) is an eye disease associated with diabetes.[1] DR is the leading cause of blindness across the globe.[1] Detecting and grading DR at an early stage

*Address all correspondence to Barath N. Narayanan, E-mail: narayananb1@udayton.edu

2329-4302/2020/$28.00 © 2020 SPIE

is critical in preventing permanent vision loss. Currently, in the United States, the prevalence rate of DR is about 28.5%[2] for people with diabetes. DR is the highest cause of new cases of blindness for those between the ages of 20 to 74 years.[3] According to the Wisconsin epidemiologic study of DR, 3.6% of type 1 diabetes patients and 1.6% of type 2 diabetes patients are blind. The blindness caused due to DR during type 1 diabetes and type 2 diabetes are 86% and 33.33%, respectively.[2] About 93 million people are affected by DR around the world.[3] Typically, a trained clinician examines and evaluates photographs of the retina to detect and grade the level of DR severity. However, detecting and grading DR is a time-consuming process.[3] Hence, an automated system would be a valuable tool to enhance the workflow of endocrinologists. This type of computer-aided detection (CAD) and grading system would help in providing a rapid objective second opinion to clinicians. This type of technology could also be utilized in areas with a scarcity of trained endocrinologists.

Computer vision-based detection and grading of DR has been a research area attracting great interest in the past decade.[4–20] Several machine learning approaches are presented in the literature[4–13] for DR study on retinal images. Approaches include traditional feature extraction and classification techniques as well as cutting-edge deep learning algorithms.[21–24] Currently, most researchers are focused on utilizing deep learning for classification of DR. In Refs. 6–13, retinal images are classified using a transfer learning-based approach. In Ref. 6, DR images are classified into referable DR, vision-threatening DR, and proliferative DR using AlexNet.[21] A private dataset is introduced in Ref. 7 comprising 70,000 images for training purposes. In Ref. 7, testing is implemented on the Kaggle test dataset[1] containing 10,000 images collected from 5000 patients. In the Kaggle dataset,[1] retinal images are graded by experts into five different categories based on its corresponding severity. In Ref. 8, a simple deep residual network model is presented to detect DR in retinal images. In addition, they extract features from the proposed convolutional neural network (CNN) model and combine them with metadata features to form an overall feature vector. These features are later classified using a simple tree-based gradient boosting classifier. In Ref. 9, a CNN is proposed to differentiate referable DR and lesions with fundus images. In Ref. 11, several CNN architectures are presented to automate the grading process of DR. In Ref. 12, entropy images are utilized to grade the DR severity level of the given patient. In Ref. 13, inception-v3[24] architecture is utilized to detect DR, and they perform training on images from the Messidor dataset[14] and testing on the Kaggle dataset.[1]

In this paper, we present an innovative hybrid machine learning architecture to detect DR and grade the DR retinal images into different severities solely using the publicly available dataset provided by Asia Pacific Tele-Ophthalmology Society (APTOS) 2019 on Kaggle.[25] These images were captured at Aravind Eye Hospital in India and were then marked by expert clinicians. We divide this architecture into two stages: (i) detection stage: identify whether the given retinal image has DR and (ii) grading stage: determine the severity of DR (mild, moderate, proliferative, or severe). In the detection stage, we study the performance of transfer-learning-based approaches using various established networks. In addition, we study the performance of these architectures under various preprocessing conditions. In the detection stage, we also visualize the class activation mapping results to determine the key image regions.[26] In a class activation mapping technique, discriminative regions are highlighted based on the probability predicted by the network for each class after mapping it back to the final convolutional layer. Visualization using class activation maps helps in enhancing the understanding of the machine learning model and provides more confidence in its predictions for the users. It also allows the expert readers, endocrinologists, and clinicians to see the image regions that are deemed to be highly salient in the detection process. In the grading stage, we study the performance of the transfer learning-based approaches and their architectural variations. We also study the performance of these models with data augmentation. Moreover, we propose a feature extraction approach using CNNs coupled with principal component analysis (PCA) for classification using support vector machine (SVM). This architecture is implemented to overcome the limited availability of images belonging to different DR severities. We believe this type of architecture would help neural networks to extract and learn features that are essential for each stage. This hybrid architecture would also help in retraining a particular stage based on the availability of a new set of images. Our best architecture achieves an overall accuracy of 98.4% for the detection stage and 96.3% for the grading stage thereby setting a new benchmark solely utilizing a small dataset for training

without any data augmentation. These results are obtained using a transfer learning-based approach for the detection stage and our approach of using CNNs with PCA and SVM for the grading stage. This paper provides a proof-of-concept on developing a hybrid approach for detecting and grading DR using limited training images.

The remainder of this paper is organized as follows. Section 2 presents the database utilized for this research along with a brief description of the proposed hybrid approach. Section 3 presents and compares the different transfer learning-based architectures for DR detection in retinal images along with their experimental results. In Sec. 4, we present and compare various architectures for the grading of DR severity along with their experimental results. In Sec. 5, we discuss the results obtained using various approaches in comparison to the existing benchmark. Finally, conclusions are offered in Sec. 6.

2 Materials and Methods

In this section, we present the database utilized for this study. As previously mentioned, we utilize the publicly available APTOS 2019 dataset.[25] This APTOS 2019 blindness detection challenge dataset contains separate training and testing cases. The training dataset is composed of 3662 retinal images marked by expert clinicians and endocrinologists. The testing dataset contains 1928 images and the labels for these images are not publicly available yet. For our validation study, we solely utilize the training dataset provided as part of the APTOS 2019 competition. We believe this is beneficial as it allows other researchers to compare algorithm performances. We also present our results for the APTOS 2019 test dataset in Sec. 5. Distribution of the APTOS 2019 training dataset is provided in Table 1. Pixel dimensions of these retinal RGB images vary from 474×358 to 3388×2588 with a bit depth of 24. Figure 1 shows typical examples marked into different categories by expert clinicians. Note that aspect ratio of these images is unaltered and is provided in the dataset.

Table 1 clearly indicates that there is an uneven distribution of the images belonging to each class. We utilize a hybrid machine learning architecture to address this issue. Figure 2 presents the top-level block diagram of the proposed hybrid machine learning architecture. This architecture contains two blocks, namely, a DR detection algorithm and a grading algorithm. We categorize the input image of a patient suffering from DR or not using the detection algorithm while the grading algorithm is used to determine the level of its severity. We believe this type of architecture would help the machine learning model to extract features and patterns that are highly essential for that particular stage. Note that we also address the class imbalance issue with data augmentation and weighted cross-entropy loss. The results obtained using this approach along with the proposed hybrid approach are discussed in Sec. 5.

For the detection stage, we merge all the retinal images marked as mild, moderate, proliferative, or severe into a single category. We perform hold-out validation in this scenario. We split the dataset into groups of 80% and 20% for training and testing purposes, respectively. We utilize a subset of 10% from our training data for validation purposes in order to fine-tune our hyperparameters. Training, testing, and validation dataset distribution is provided in Table 2. We use the same set of cases for all the architectures presented in the detection stage. We present the

Table 1 APTOS 2019 training dataset distribution.

Class	Number of images
No DR	1805
Mild DR	370
Moderate DR	999
Proliferative DR	295
Severe DR	193

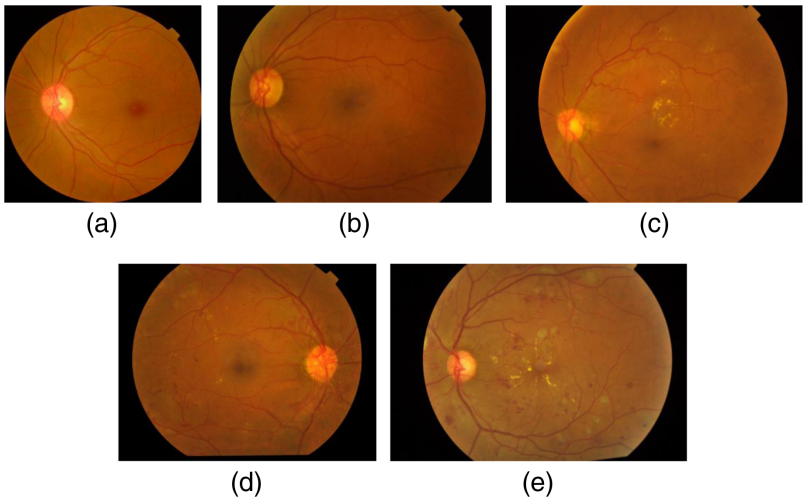


Fig. 1 Retinal images marked by expert clinicians as: (a) no DR, (b) mild DR, (c) moderate DR, (d) proliferative DR, and (e) severe DR.

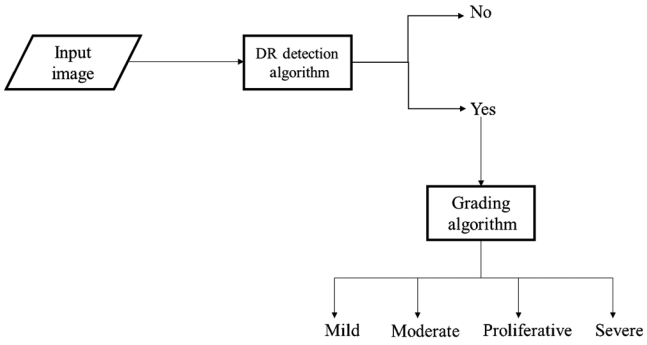


Fig. 2 Proposed top-level hybrid architecture for the automated DR detection and grading.

Table 2 Training, testing, and validation dataset distribution for detection stage.

DR label	# Train images	# Validation images	# Test images
No	1300	144	361
Yes	1337	149	371

details of the architecture along with experimental results and class activation mapping results in Sec. 3.

For the grading stage, due to limited availability of different severities of DR retinal images (mild, moderate, proliferative, and severe) in the APTOS 2019 dataset, we conduct 10-fold validation to analyze our performance. This type of evaluation would provide a reasonable estimate of our performance instead of a typical hold-out validation method. In the 10-fold validation process, we train 10 different networks based on the set of nine-fold images and test the performance on the remaining fold. For each fold, we train and tune our hyperparameters solely based on the images from the training fold. We make sure to exclude the testing fold in any manner to conduct a rigorous study. Note that we utilize the same set of cases in each fold for the different architectures implemented in the grading stage. We present the details of various architectures along with experimental results obtained under different preprocessing conditions in Sec. 4.

3 DR Detection Architecture

In this section, we present various architectures for automated DR detection in retinal images. We use a transfer learning-based approach for DR detection. We utilize the dataset distribution provided in Table 2 for this study. Transfer learning approaches using established deep learning networks have proven to be highly effective for detection in medical imaging applications including DR.[7–13,27,28]

For the detection stage, we adopt the approach presented in Ref. 27 by two of the coauthors of this paper for detection of malaria on cell images. At first, we study the performance of AlexNet,[21] VGG16,[22] ResNet,[23] and Inception-v3[24] networks for this research. We replace the last fully connected layer of the established network with a fully connected layer with 2 units in order to detect DR followed by softmax and classification layers. Figure 3 illustrates the top-level block diagram of this detection approach. To avoid the variations of the color and contrast among images, we preprocess either using (i) a color constancy technique or (ii) histogram equalization after converting to its equivalent grayscale image. Note that the preprocessing technique is applied solely based on the region of interest (black background pixels are ignored for calculation purposes by applying simple morphological operations). Color constancy is implemented as done in Ref. 27. Figure 4 presents the results obtained after the application of such preprocessing techniques.

We study the test performance in terms of overall accuracy and receiver operating characteristic (ROC) curve with and without any preprocessing techniques. We utilize the dataset distribution provided in Table 2 for this study. We resize all images to match with the input size of the network. We choose the Adam optimization technique[29] for all the networks and a validation patience of three with a maximum number of epochs of five. Validation is implemented for every 10 iterations. We choose a minibatch size of 256. Hyperparameters such as learning rate and Adam optimization parameters[29] are determined solely based on validation performance. In addition, we average the posterior probabilities provided by each network and determine the overall probability, and we term this method as "average-of-all." Table 3 summarizes the test performance of the transfer learning-based approaches in terms of overall accuracy along with their 95% confidence interval (CI). CIs are determined by randomly splitting the images into training, validation, and testing, and conducting 10 different hold-out validation experiments. Table 3 summarizes the test performance of the transfer learning-based approaches

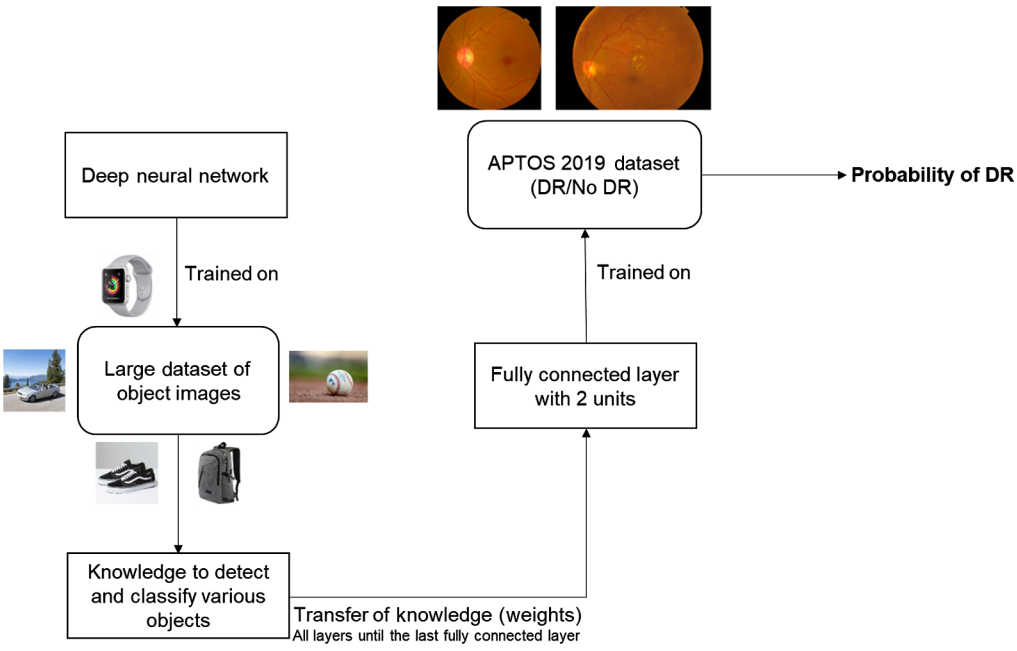


Fig. 3 Top-level block diagram of the transfer learning-based approach for the DR detection stage.

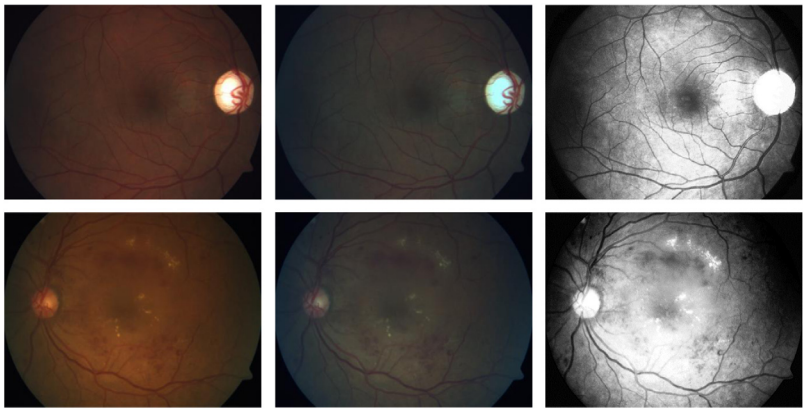


Fig. 4 Preprocessing results: (a) raw image, (b) after color constancy only, and (c) after histogram equalization only.

Table 3 Overall accuracy in percentage with 95% CI for DR detection.

Network	Raw images	Color constancy	Histogram equalization
AlexNet	96.15 ± 1.7	96.8 ± 1.2	96.2 ± 1.8
VGG16	96.23 ± 1.6	97.0 ± 1.3	96.9 ± 1.9
ResNet	96.7 ± 1.7	97.6 ± 1.4	96.8 ± 1.7
Inception-v3	96.6 ± 1.7	98.0 ± 1.3	97.2 ± 1.5
Average-of-all	$\mathbf{96.9 \pm 1.7}$	$\mathbf{98.4 \pm 1.3}$	$\mathbf{97.5 \pm 1.7}$

Note: Bold values represent the best performance among the compared methods.

in terms of overall accuracy. Table 3 indicates that all the presented approaches for the detection of DR perform reasonably well with a minimum performance of 96.2%. Results also indicate that preprocessing using histogram equalization or color constancy provides a modest but consistent boost in the performance. Table 3 results indicate that color constancy outperforms histogram equalization and raw images for all the networks applied. Preprocessing techniques also provide the maximum boost in performance for Inception-v3 architecture when compared to other networks. The average-of-all method provides the best performance among the architectures studied. Figure 5 presents the confusion matrix obtained using the average-of-all method under different preprocessing techniques. Figure 6 presents the ROC curve obtained for DR detection using different networks with a color constancy method of preprocessing. Table 4 summarizes the results in terms of area under the ROC curve (AUC) with 95% CI.

In addition, we also study the performance of modern deep learning networks such as NASNet,[30] DenseNet,[31] and GoogLeNet;[32] however, the performance of these networks is similar to or lower than the networks presented earlier, with a significantly higher training time. Results obtained with these approaches are presented in Table 5. Hence, we adopt AlexNet, VGG16, ResNet, and Inception-v3 throughout this paper.

Figures 7 and 8 present typical CAD system outputs obtained using Inception-v3 network for sample test images. The raw image presented in Fig. 7(a) has been marked by expert clinicians as "no DR." Figure 7(b) presents the result obtained after the application of color constancy on the raw image. Figure 7(c) presents the class activation mapping results and the region that contributed the most for the decision made by the network. These class activation maps are generated based on Ref. 26. Our detection algorithm predicts that DR probability of this retinal image is 0.00.

The raw image presented in Fig. 8(a) has been marked by expert clinicians as mild DR. Figure 8(b) shows the result obtained after the application of color constancy on the raw image.

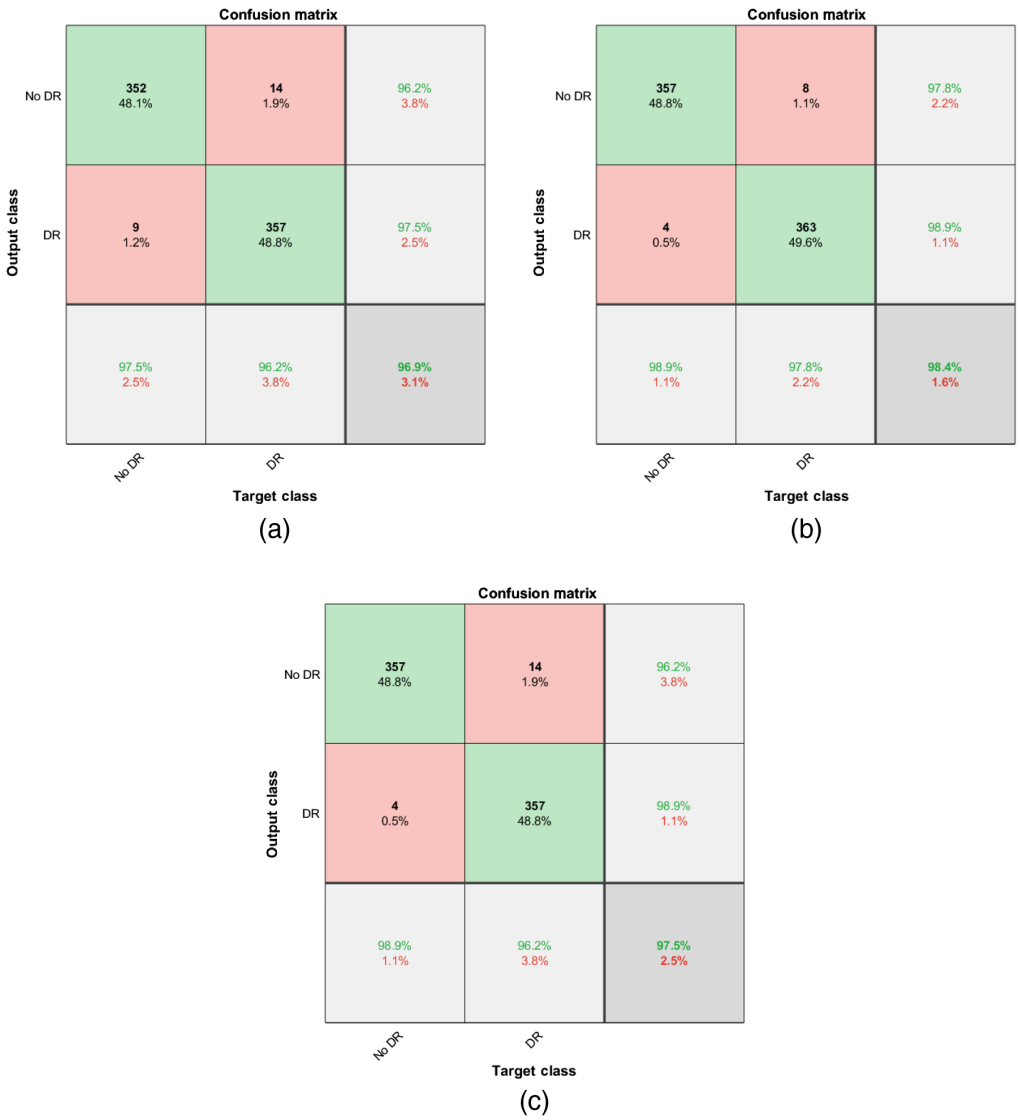


Fig. 5 Confusion matrix obtained using average-of-all method: (a) raw images, (b) after color constancy only, and (c) after histogram equalization only.

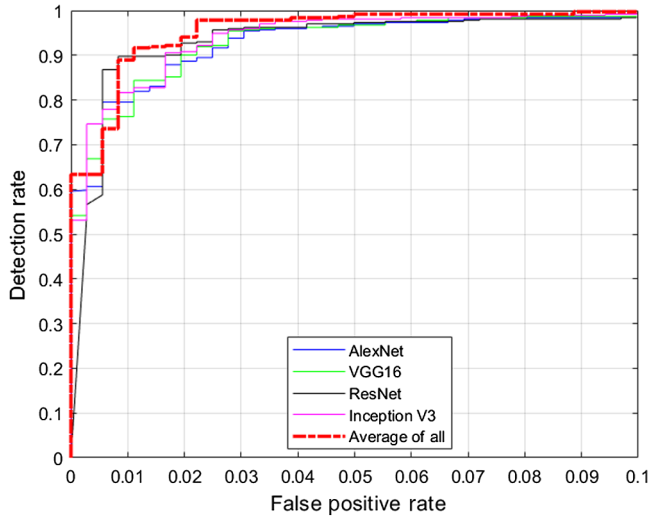


Fig. 6 ROC curves obtained for DR detection using color constancy method of preprocessing.

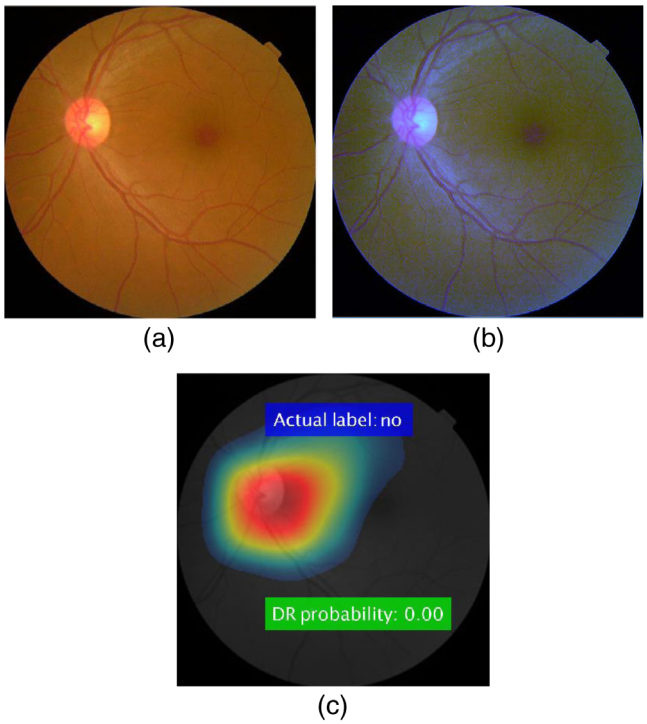
Table 4 Overall AUC with 95% CI for DR detection.

Network	Raw images	Color constancy	Histogram equalization
AlexNet	0.981 ± 0.05	0.988 ± 0.03	0.982 ± 0.04
VGG16	0.982 ± 0.04	0.989 ± 0.03	0.988 ± 0.03
ResNet	0.984 ± 0.02	0.990 ± 0.02	0.984 ± 0.02
Inception-v3	0.988 ± 0.03	0.993 ± 0.03	0.989 ± 0.03
Average-of-all	$\mathbf{0.989 \pm 0.02}$	$\mathbf{0.995 \pm 0.01}$	$\mathbf{0.991 \pm 0.02}$

Note: Bold values represent the best performance among the compared methods.

Table 5 Overall accuracy in percentage with 95% CI for DR detection using modern networks.

Network	Raw images	Color constancy	Histogram equalization
NASNet	95.9 ± 2.0	96.7 ± 1.8	96.2 ± 1.9
DenseNet	96.0 ± 1.4	96.3 ± 1.3	96.2 ± 1.9
GoogLeNet	96.2 ± 1.6	96.7 ± 1.0	96.1 ± 1.5



(a) (b) (c)

Fig. 7 Typical CAD system output for a no DR case: (a) raw image, (b) preprocessed using color constancy, and (c) class activation mapping result.

Figure 8(c) presents the region that contributed the most for our Inception-v3 architecture to detect DR and our algorithm provides a probability of 1.00 for the same. We believe this kind of visualization of our predictions would assist expert clinicians in making fast and accurate decisions.

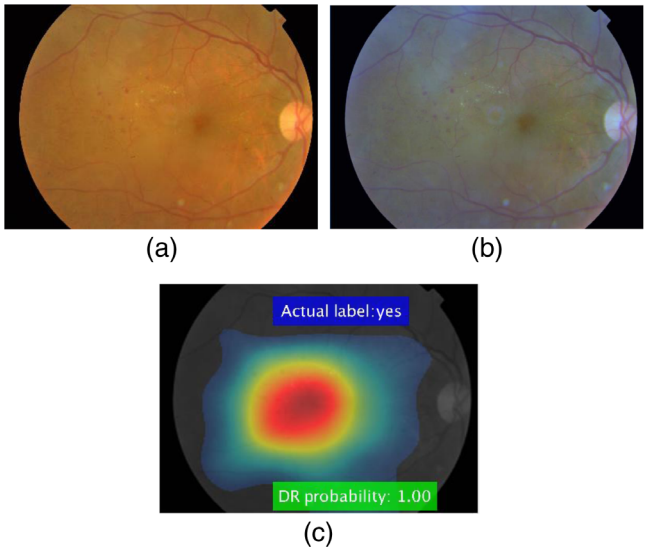


Fig. 8 Typical CAD system output for a DR case: (a) raw image, (b) preprocessed using color constancy, and (c) class activation mapping result.

4 DR Grading Architecture

In this section, we present various architectures for distinguishing the severity of DR (mild, moderate, proliferative, or severe) along with our proposed architecture.

4.1 *Transfer Learning Approach*

At first, we study the performance of transfer learning-based approaches using the same set of established networks. We adopt the same transfer learning approach as shown in Fig. 3. However, we replace the last fully connected layer of the network with a fully connected layer with 4 units in order to distinguish the DR severities. As mentioned in Sec. 2, we estimate the performance using a 10-fold validation technique to study the robustness of the algorithm. For each fold, training is implemented with a validation patience of three. We repeat the same process as is done in the detection stage and study the performance under different preprocessing conditions. Table 6 presents the overall accuracy obtained for these networks with 95% CI. CI values are obtained based on each fold performance. Results clearly indicate that the performance of these algorithms is relatively poor and this could be attributed to the low number of training images from each severity class. In the subsequent section, we study the performance of these algorithms after data augmentation. Results also indicate that distinguishing among different grades of DR is a difficult process in comparison to DR detection.

Table 6 Overall accuracy in percentage for DR grading using transfer learning.

Network	Raw images	Color constancy	Histogram equalization
AlexNet	60.2 ± 13.8	61.3 ± 10.7	61.0 ± 12.3
VGG16	61.1 ± 14.1	62.8 ± 9.3	62.2 ± 11.4
ResNet	64.5 ± 12.7	65.2 ± 8.3	65.9 ± 9.7
Inception-v3	65.6 ± 11.5	65.4 ± 7.3	64.3 ± 9.8
Average-of-all	$\mathbf{65.3 \pm 9.7}$	$\mathbf{66.7 \pm 6.8}$	$\mathbf{65.8 \pm 8.0}$

Note: Bold values represent the best performance among the compared methods.

Table 7 Overall accuracy in percentage for DR grading using transfer learning after data augmentation.

Network	Raw images	Color constancy	Histogram equalization
AlexNet	61.5 ± 13.6	60.8 ± 10.1	61.0 ± 12.1
VGG16	62.1 ± 13.2	62.9 ± 9.2	62.2 ± 10.1
ResNet	63.9 ± 12.1	65.3 ± 8.2	65.3 ± 9
Inception-v3	65.6 ± 9.6	66.1 ± 7.1	66.8 ± 8.9
Average-of-all	$\mathbf{65.7 \pm 8.5}$	$\mathbf{66.9 \pm 6.9}$	$\mathbf{66.9 \pm 7.5}$

Note: Bold values represent the best performance among the compared methods.

4.2 *Data Augmentation*

Data augmentation has been a popular technique for deep learning algorithms with limited training samples.[33] Data augmentation techniques typically involve rotation, translation, scaling, color contrast techniques, etc. However, all these techniques may not be applicable for retinal images. We follow the footsteps of Ref. 13 to augment the dataset as these techniques have proven to be effective for DR detection. We augment the dataset such that each class type has ~3000 images. Note that we solely augment the training images for that particular fold. This would also help in tackling class imbalance. Table 7 presents the overall accuracy obtained for each technique after the application of data augmentation. Table 7 results indicate that data augmentation does not considerably improve the performance. Maximum accuracy obtained for distinguishing the severities of DR is only 66.9%.

4.3 *Modified Transfer Learning Architecture*

In this section, we present results after making a slight variation to the existing transfer learning architecture. We introduce an additional fully connected layer with 100 units before the last fully connected layer with 4 units for all the networks mentioned in Sec. 4.1. We also add dropout layers to avoid overfitting. Figure 9 presents the top-level block diagram of this approach. We train these updated networks for each fold and study the performance of these networks for each type of preprocessing. Table 8 presents the results obtained using these networks. Note that we perform the same set of data augmentation techniques as mentioned in Sec. 4.2. Results indicate that there is an increase in the performance when compared to Tables 6 and 7. We obtain a maximum accuracy of 74.2% using this technique. However, there is still scope for further improvement.

4.4 *CNN as Feature Extractors*

Combination of traditional machine learning approaches such as SVM for classification and CNN as a feature extractor has been an effective approach for limited training samples. We extract features from the newly added fully connected layer with 100 units from each network and later classify the features using SVM with a linear kernel and study the performance for each methodology. Figure 10 presents the top-level block diagram of this approach. Note that SVM has the ability to classify images with limited samples,[34] hence we do not perform any data augmentation for this approach. Table 9 summarizes the results obtained using this approach. Table 9 clearly indicates that this type of feature extraction using CNN is highly effective. We achieve an overall accuracy of 85.7% using this approach with a color constancy preprocessing technique.

4.5 *Proposed Architecture for Grading*

In this section, we propose an approach for automated DR grading. Results from Sec. 4.4 indicate that CNN performs well as a feature extractor. In this section, we combine all these salient

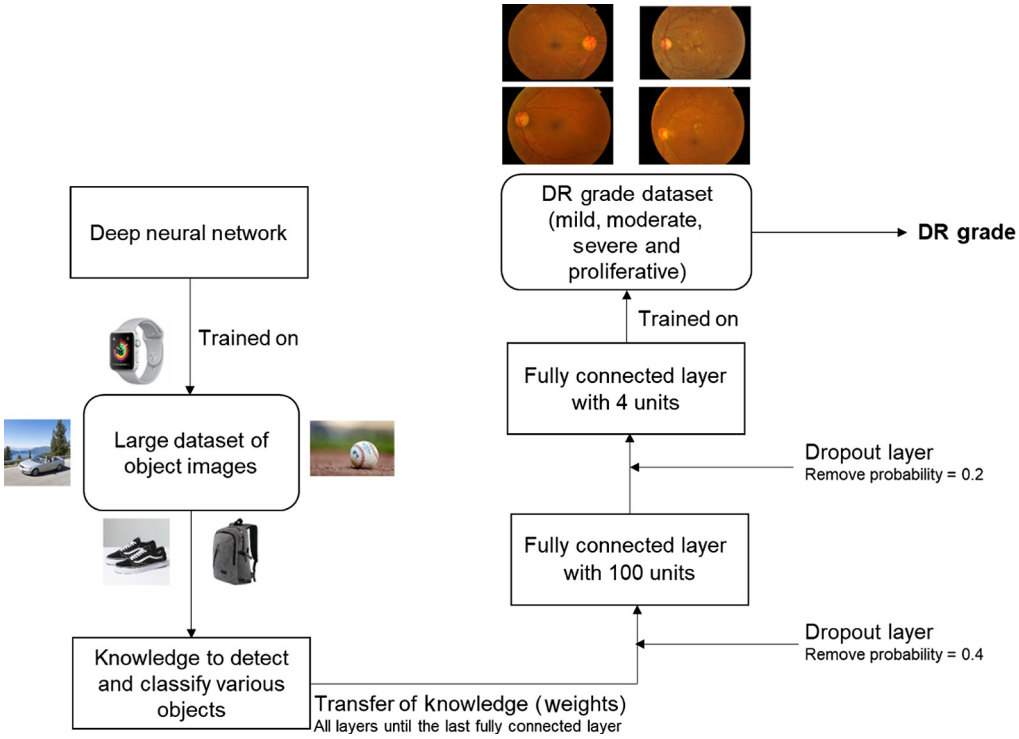


Fig. 9 Top-level block diagram of the modified transfer learning-based approach for DR grading.

Table 8 Overall accuracy in percentage for DR grading with additional fully connected layer and data augmentation.

Network	Raw images	Color constancy	Histogram equalization
AlexNet	70.3 ± 9.3	72.1 ± 5.3	71.8 ± 6.1
VGG16	71.1 ± 8.1	71.1 ± 5.1	71.7 ± 5.3
ResNet	72.8 ± 8.0	72.1 ± 3.9	72.1 ± 5.0
Inception-v3	72.9 ± 7.6	73.2 ± 2.9	73.2 ± 4.3
Average-of-all	$\mathbf{73.0 \pm 5.1}$	$\mathbf{74.2 \pm 3.9}$	$\mathbf{73.7 \pm 4.5}$

Note: Bold values represent the best performance among the compared methods.

features from each CNN architecture to form a vector of 400 features for each retinal image. Some of the features extracted from various CNNs could be correlated, hence, we perform the PCA method of dimensionality reduction[35] on this pool of features to represent the data with a minimal set of features for classification. We choose an explained variance of 90% for PCA. Different numbers of features are extracted for each fold. The number of features typically ranged from 15 to 20. After extracting PCA features, we classify using SVM with a linear kernel as implemented in Sec. 4.4. Figure 11 presents the block diagram of the proposed approach.

Table 10 summarizes the overall accuracy for each preprocessing method. As mentioned in Sec. 4.4, no data augmentation is performed in this scenario. This approach provided a significant boost in the classification performance and we achieve an overall accuracy of 96.3% using the color constancy method of preprocessing. Figure 12 presents the confusion matrices obtained using these approaches.

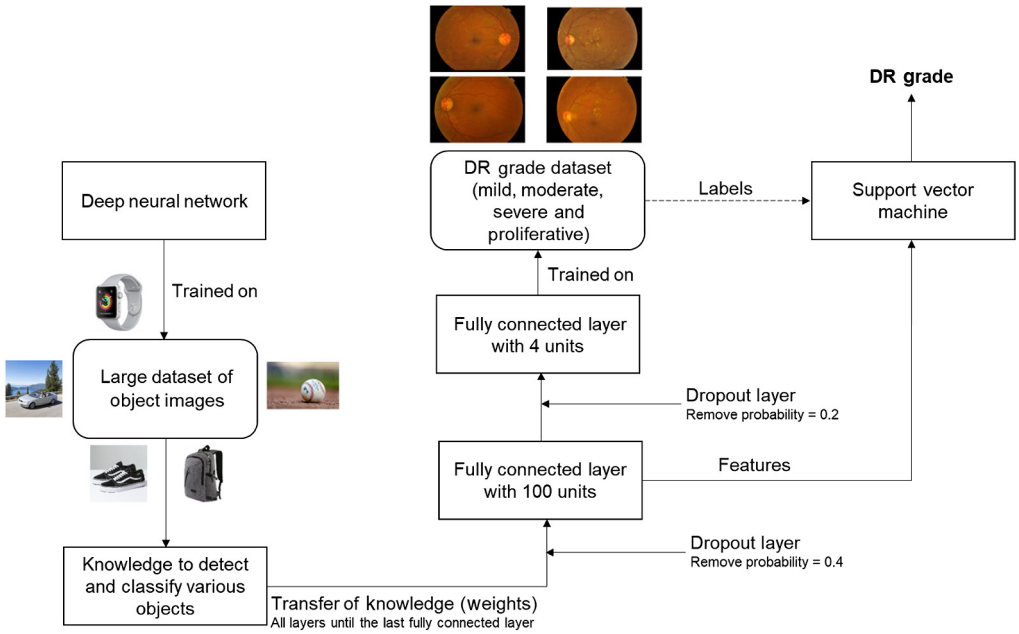


Fig. 10 Top-level block diagram with CNN features and SVM classifier for DR grading.

Table 9 Overall accuracy in percentage for DR grading using CNN features and SVM classifier.

Network	Raw images	Color constancy	Histogram equalization
AlexNet	75.7 ± 6.8	81.6 ± 5.4	80.5 ± 6.2
VGG16	76.8 ± 5.3	83.5 ± 6.1	82.3 ± 5.7
ResNet	78.9 ± 7.1	84.9 ± 8.4	83.7 ± 7.7
Inception-v3	79.8 ± 6.4	$\mathbf{85.7 \pm 5.4}$	83.7 ± 6.0

Note: Bold values represent the best performance among the compared methods.

5 Discussion

Several papers have addressed the study of classification of retinal images, but not much research work (with the exception of Ref. 36) has been implemented on the recent APTOS 2019 dataset. This dataset has its own set of challenges with a limited supply of images belonging to each class. We presented a hybrid architecture to detect and grade DR. Having a separate architecture for detection and grading helps the network in learning patterns/features specific to that application.

Table 11 presents the summary of results for different datasets for the detection of DR. Although it is not fair comparing different algorithms for different datasets, it is good to note the accuracy and AUC values in order to establish a benchmark for the APTOS 2019 dataset. Based on the results, our proposed algorithm for detection of DR in APTOS 2019 performs reasonably well and sets up a good benchmark for future research efforts for the same dataset.

In this research, we have also addressed the issue of class imbalance and limited training samples. We presented a different architecture for grading of DR due to the limited availability of training samples, which revealed that rich features are essential in order to classify the severity of DR. We estimated our performance using the 10-fold validation technique. We have also analyzed the performance of various cutting-edge deep learning algorithms and observed their limitations. We studied the performance of these networks after the application of data augmentation and observed a slight increase in the performance. Extracting features from CNN with classification using SVM provided an accuracy boost of nearly 11%. Our proposed method of extracting and combining features from various CNNs, applying the PCA method of dimensionality reduction, and later, classifying using SVM provided the best performance among all

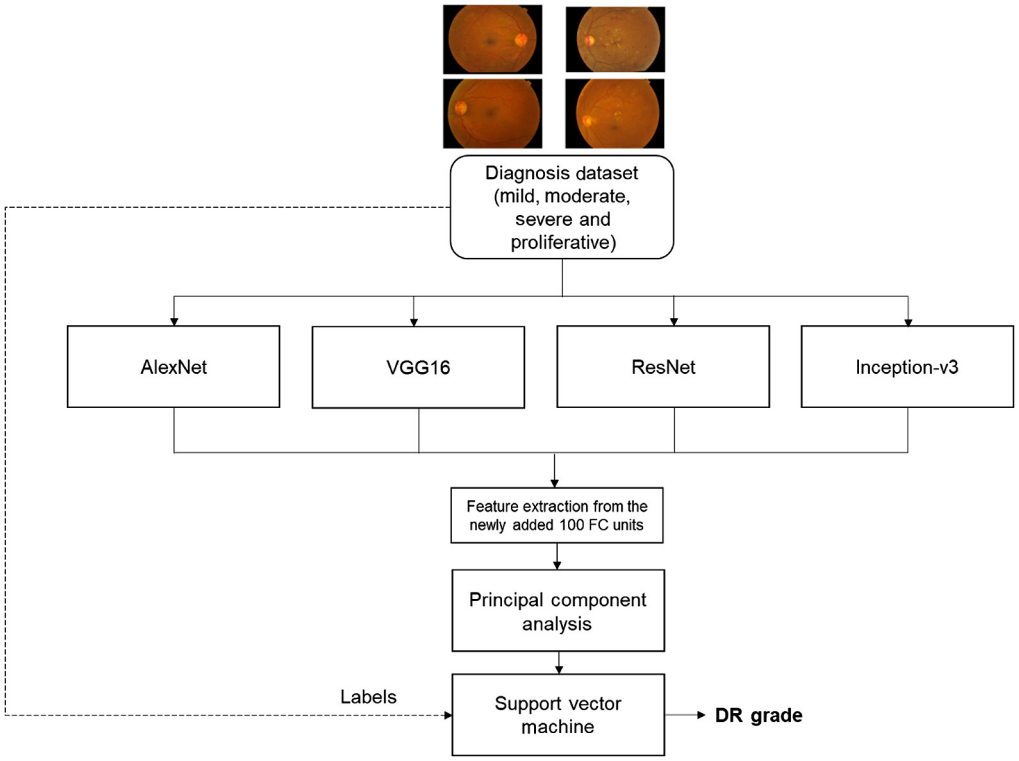


Fig. 11 Proposed approach for DR grading.

Table 10 Overall accuracy in percentage for DR grading using proposed architecture.

Preprocessing technique	Accuracy
Raw image	89.7 ± 2.8
Color constancy	$\mathbf{96.3 \pm 1.3}$
Histogram equalization	95.0 ± 1.8

Note: Bold values represent the best performance among the compared methods.

the methods presented. We achieved an accuracy of 96.3% for DR grading. This type of approach can be applied to other applications with limited training samples.

We submitted our results for the APTOS 2019 test dataset on Kaggle using different architectures under different preprocessing conditions. Tables 12 and 13 summarize the quadratic weighted kappa score[37] obtained for our different submissions on the public and private leaderboard. CIs are obtained after 10 different submissions based on each network on Kaggle. Single-stage architecture represents having one network to detect DR and distinguish its severities. This is implemented using Inception-v3 architecture with a weighted cross-entropy loss and data augmentation as described in Sec. 4.2. Average-of-all and XGBoost[38] refer to application of "average-of-all" method implemented in Sec. 4.1 using different architectures for detection and later applying XGBoost after extracting 100 features from each of the CNN architectures mentioned in Sec. 4.4. The stacking of layers approach refers to use a random forest-based classifier after obtaining the predictions from all four architectures (AlexNet, VGG16, Inception-v3, and ResNet) for DR detection and the proposed grading model. The proposed hybrid architecture refers to using the average-of-all method for detection stage, SVM after extracting features from all CNNs, and PCA for the grading stage. Tables 12 and 13 indicate that the best performance is achieved using the proposed hybrid approach with a color constancy preprocessing technique. Our performance falls under the top 1% among 2932 other submissions across public and private leaderboards.

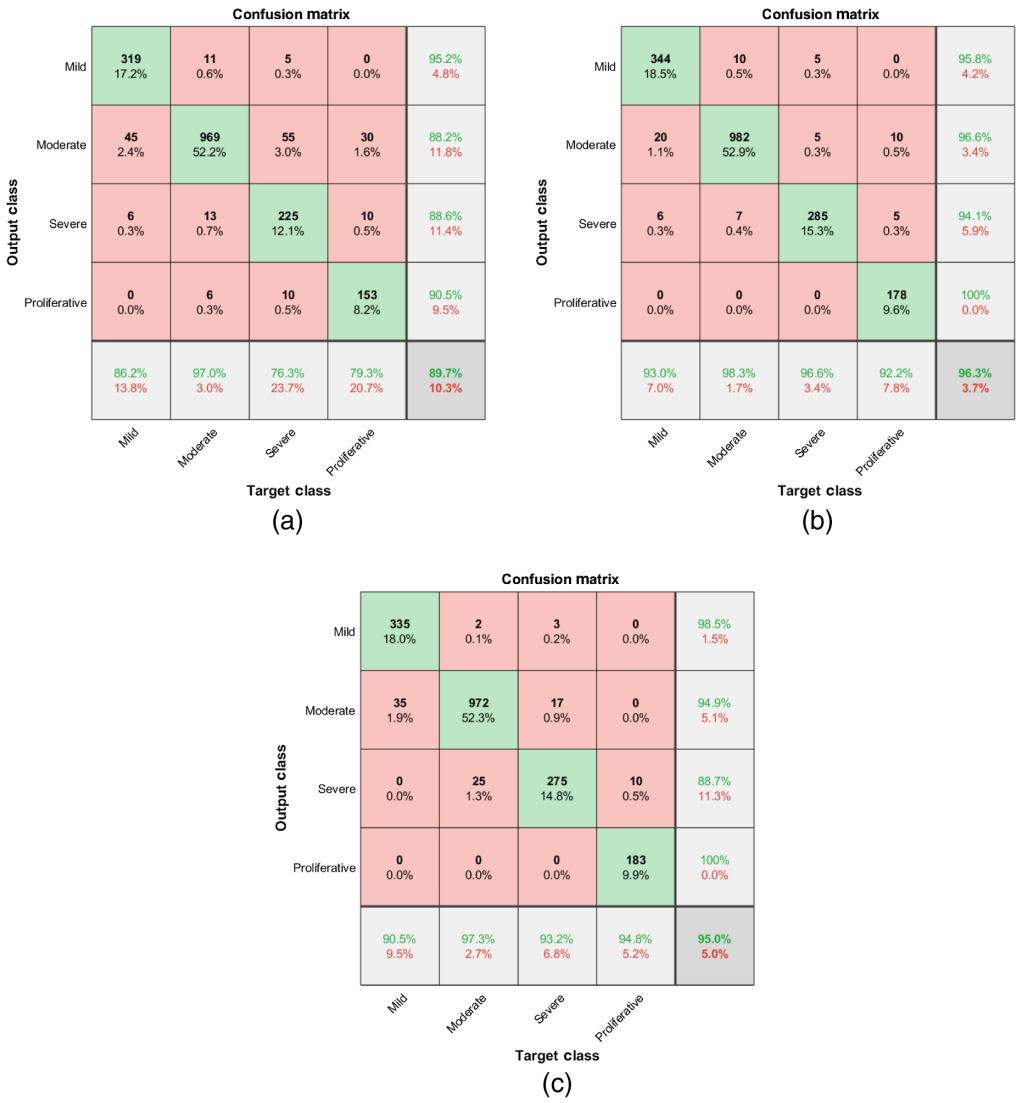


Fig. 12 Confusion matrix obtained using proposed approach for DR grading: (a) raw image, (b) after color constancy only, (c) after histogram equalization only.

Table 11 Summary of results for DR detection.

Reference	Dataset	Accuracy (%)	AUC
Gargeya et al.[8]	EYEPACS, MESSIDOR	96.3	0.97
Sengupta et al.[13]	EYEPACS	86.7	0.88
Sengupta et al.[13]	MESSIDOR	90.4	0.91
Proposed detection method	**APTOS**	**98.4 ± 0.5**	**0.994 ± 0.02**

Note: Bold values represent the best performance among the compared methods.

6 Conclusions

In this research, we have proposed and presented a hybrid machine learning architecture for automated detection and grading of DR on retinal images. Tables 12 and 13 clearly indicate that having separate architectures for detection and grading provides a significant boost in the performance when compared to a single-stage architecture. In fact, the proposed hybrid architecture outperforms the single-stage architecture without any data augmentation. This could be

Table 12 Quadratic weighted kappa score on APTOS 2019 public leaderboard.

Architecture	Raw images	Color constancy	Histogram equalization
Single stage	0.72 ± 0.01	0.74 ± 0.005	0.73 ± 0.005
Average-of-all and XGBoost	0.76 ± 0.01	0.79 ± 0.005	0.77 ± 0.004
Stacking of Layers	0.79 ± 0.006	0.80 ± 0.002	0.81 ± 0.005
Proposed hybrid	0.82 ± 0.004	$\mathbf{0.84 \pm 0.002}$	0.83 ± 0.005

Note: Bold values represent the best performance among the compared methods.

Table 13 Quadratic weighted kappa score on APTOS 2019 private leaderboard.

Architecture	Raw images	Color constancy	Histogram equalization
Single stage	0.83 ± 0.01	0.8103 ± 0.005	0.805 ± 0.005
Average-of-all and XGBoost	0.84 ± 0.01	0.87 ± 0.009	0.86 ± 0.008
Stacking of layers	0.89 ± 0.003	0.91 ± 0.009	0.86 ± 0.004
Proposed hybrid	0.91 ± 0.004	$\mathbf{0.9298 \pm 0.003}$	0.921 ± 0.005

Note: Bold values represent the best performance among the compared methods.

attributed to the fact that having separate architectures helps in learning features and patterns specific to that stage thereby enhancing the performance of machine learning algorithms. We have presented these results for a publicly available dataset thereby setting a benchmark for future research efforts. At first, we studied the performance of transfer learning-based approaches utilizing AlexNet, VGG16, ResNet, Inception-v3, NASNet, DenseNet, and GoogLeNet. Results indicate that transfer-learning-based approaches are effective for detection of DR in retinal images. Preprocessing methods such as histogram equalization and color constancy also provided a modest but consistent boost in the performance. In APTOS 2019 dataset, there is a slight variation in terms of illumination across images. Employing the color constancy method of preprocessing helps to make the input more invariant to the color of the illumination source.[39] This assists the deep learning algorithms in focusing on the true underlying physical differences that are relevant for distinguishing classes and helps to boost the performance.[39] Average-of-all method using color constancy achieved the best accuracy of 98.4% for the detection of DR with an AUC value of 0.994. We also visualized class activation mapping results to enhance the understanding of the detection model for both data science researchers and clinicians.

For DR grading of retinal images, we studied the performance of various CNN architectures with and without data augmentation. We also studied their performance with minor variations. In addition, we studied the performance of CNNs as feature extractors and SVM for classification. Later, we proposed a method to fuse PCA with CNN and SVM. This type of architecture can be utilized for scenarios with limited training samples. Tables 9 and 10 indicate that the SVM method of classification is highly effective for grading of DR severity. Our proposed algorithm of using CNN as feature extractors along with the PCA method of dimensionality reduction and SVM for classification achieved the best accuracy of 96.3% for grading of DR severity. To combat limited training data, we utilize various CNNs as feature extractors to form a pool of salient features. Some of these features determined by various CNNs could be redundant and applying PCA helps in transforming them to an uncorrelated feature space and using a minimal set to represent those features. These transformed features assist SVM in classifying the different grades of DR. Having separate networks for detection and grading could help in retraining a particular architecture with new labeled data as necessary. This paper provides a proof-of-concept study of an automated DR detection and grading system developed using a limited set

of labeled images. Detection and/or grading networks can be implemented in regions with a dearth of trained clinicians based on the necessity.

Disclosures

The authors declare that there are no conflicts of interest related to this article.

Acknowledgments

The authors would like to thank Kaggle and Aravind Eye Hospital, India, for contributing toward a publicly available dataset utilized for this research. They would also like to thank anonymous reviewers for further strengthening the paper.

References

1. "Kaggle diabetic retinopathy dataset," https://www.kaggle.com/c/diabetic-retinopathy-detection/overview (accessed 27 November 2019).
2. A. S. Krolewski et al., "Risk of proliferative diabetic retinopathy in juvenile-onset type I diabetes: a 40-yr follow-up study," *Diabetes Care* **9**(5), 443–452 (1986).
3. D. S. Fong et al., "Retinopathy in diabetes," *Diabetes Care* **27**, s84–s87 (2004).
4. H. Leopold, J. Zelek, and V. Lakshminarayanan, "Deep learning methods for retinal image analysis in signal processing and machine learning for biomedical big data," in *Machine Learning for Biomedical Big Data*, E. Sejdic and T. H. Falk, Eds., pp. 329–365, CRC Press, Taylor & Francis Group, Boca Raton, Florida (2018).
5. S. Sengupta et al., "Ophthalmic diagnosis and deep learning—a survey," arXiv:1812.07101, https://export.arxiv.org/pdf/1812.07101 (2018).
6. M. D. Abràmoff et al., "Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning," *Invest. Ophthalmol. Vis. Sci.* **57**(13), 5200–5206 (2016).
7. E. Colas et al., "Deep learning approach for diabetic retinopathy screening," *Acta Ophthalmol.* **94**(S256) (2016).
8. R. Gargeya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," *Ophthalmology* **124**(7), 962–969.
9. G. Quellec et al., "Deep image mining for diabetic retinopathy screening," *Med. Image Anal.* **39**, 178–193 (2017).
10. H. Takahashi et al., "Applying artificial intelligence to disease staging: deep learning for improved staging of diabetic retinopathy," *PLoS One* **12**(6), e0179790 (2017).
11. G. García et al., "Detection of diabetic retinopathy based on a convolutional neural network using retinal fundus images," *Lect. Notes Comput. Sci.* **10614**, 635–642 (2017).
12. G. M. Lin et al., "Transforming retinal photographs to entropy images in deep learning to improve automated detection for diabetic retinopathy," *J. Ophthalmol.* **2018**, 1–6 (2018).
13. S. Sengupta et al., "Cross-domain diabetic retinopathy detection using deep learning," *Proc. SPIE* **11139**, 111390V (2019).
14. E. Decencière et al., "Feedback on a publicly distributed image database: the Messidor database," *Image Anal. Stereol.* **33**(3), 231–234 (2014).
15. Z. Zhang et al., "ACHIKO-K: database of fundus images from glaucoma patients," in *IEEE 8th Conf. Ind. Electron. and Appl. (ICIEA)*, IEEE, pp. 228–231 (2013).
16. E. J. Carmona et al., "Identification of the optic nerve head with genetic algorithms," *Artif. Intell. Med.* **43**(3), 243–259 (2008).
17. J. Sivaswamy et al., "Drishti-GS: retinal image dataset for optic nerve head (ONH) segmentation," in *IEEE 11th Int. Symp. Biomed. Imaging (ISBI)*, IEEE, pp. 53–56 (2014).
18. A. Budai et al., "Robust vessel segmentation in fundus images," *Int. J. Biomed. Imaging* **2013**, 154860 (2013).

19. Y. Zheng et al., "How much eye care services do Asian populations need? Projection from the Singapore Epidemiology of Eye Disease (SEED) study," *Invest. Ophthalmol. Vis. Sci.* **54**(3), 2171–2177 (2013).
20. Z. Zhang et al., "ORIGA-light: an online retinal fundus image database for glaucoma analysis and research," in *Annu. Int. Conf. IEEE Eng. Med. and Biol.*, Buenos Aires, IEEE (2010).
21. A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Adv. Neural Inf. Process. Syst.*, pp. 1097–1105 (2012).
22. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556, https://arxiv.org/abs/1409.1556 (2014).
23. K. He et al., "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 770–778 (2016).
24. C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 1–9 (2015).
25. "APTOS 2019 blindness detection," https://www.kaggle.com/c/aptos2019-blindness-detection/overview (accessed 27 November 2019).
26. B. Zhou et al., "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, Las Vegas, Nevada, pp. 2921–2929 (2016).
27. B. N. Narayanan, R. Ali, and R.C. Hardie, "Performance analysis of machine learning and deep learning architectures for malaria detection on cell images," *Proc. SPIE* **111390**, 111390W (2019).
28. S. Namuduri et al., "Automated quantification of DNA damage via deep transfer learning based analysis of comet assay images," *Proc. SPIE* **111390**, 111390Y (2019).
29. P. D. Kingma and J. Ba, "Adam: a method for stochastic optimization," arXiv:1412.6980, https://arxiv.org/abs/1412.6980 (2014).
30. B. Zoph et al., "Learning transferable architectures for scalable image recognition," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 8697–8710 (2018).
31. G. Huang et al., "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, Honolulu, Hawaii, pp. 4700–4708 (2017).
32. T. Fang, "A novel computer-aided lung cancer detection method based on transfer learning from GoogLeNet and median intensity projections," in *Proc. IEEE Int. Conf. Comput. Commun. Eng. Technol. (CCET)*, pp. 286–290 (2018).
33. L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," arXiv: 1712.04621, https://arxiv.org/abs/1712.04621 (2017).
34. J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.* **9**(3), 293–300 (1999).
35. S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemom. Intell. Lab. Syst.* **2**(1–3), 37–52 (1987).
36. N. Sikder et al., "Early blindness detection based on retinal images using ensemble learning," (2019) https://www.researchgate.net/profile/Niloy_Sikder/publication/336829911_Early_Blindness_Detection_Based_on_Retinal_Images_Using_Ensemble_Learning/links/5db4589b92851c577ec9fd75/Early-Blindness-Detection-Based-on-Retinal-Images-Using-Ensemble-Learning.pdf
37. A. Arora, "Quadratic weighted kappa rating,," https://www.kaggle.com/aroraaman/quadratic-kappa-metric-explained-in-5-simple-steps (accessed 28 November 2019).
38. T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. and Data Mining*, San Francisco, pp. 785–794 (2016).
39. G. Kaur and N. Bawa, "A review on color constancy based image enhancement," *Int. J. Comput. Trends Technol.* **26**(1), 6–11 (2015).

Barath Narayanan Narayanan received his bachelor's degree (with distinction) in electrical and electronics engineering from SRM University, Chennai, India, in 2012 and his MS and PhD degrees in electrical engineering from University of Dayton (UD) in 2013 and 2017, respectively. He holds a joint appointment as a research scientist at the University of Dayton Research Institute's (UDRI) Software Systems Group and as an adjunct faculty in the Electrical and

Computer Engineering Department at UD. His research interests include machine learning, computer vision, medical imaging, cyber security, and predictive analytics. He is a member of SPIE.

Russell C. Hardie is a full professor in the Department of Electrical and Computer Engineering at the University of Dayton, with a joint appointment in the Department of Electro-Optics. He received the University of Dayton's top university-wide teaching award in 2006 and the Rudolf Kingslake Medal and Prize from SPIE in 1998. He also received the School of Engineering Award of Excellence in teaching in 1999.

Manawaduge Supun De Silva received her BSc (honors) degree in engineering physics from the University of Colombo, Sri Lanka, in 2014, and her MS degree in electrical engineering from the University of Dayton in 2016. She is a graduate teaching assistant at the University of Dayton. She is currently pursuing her PhD in electrical engineering. Her research interests include medical image processing, machine learning, and computer vision.

Nathaniel K. Kueterman received his BS degree in electrical engineering from the University of Dayton in 2018. He is currently pursuing his MS degree in electrical engineering, with a focus on machine learning and image processing. He is a graduate research assistant at the University of Dayton.