IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 September 2022, accepted 19 October 2022, date of publication 26 October 2022, date of current version 2 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3217216

RESEARCH ARTICLE

Multi-Stream Deep Neural Network for Diabetic Retinopathy Severity Classification Under a Boosting Framework

HAMZA MUSTAFA[1], SYED FAROOQ ALI[1], MUHAMMAD BILAL[2], AND MUHAMMAD SHEHZAD HANIF[2]

[1]School of Systems and Technology, University of Management and Technology, Lahore 54770, Pakistan
[2]Department of Electrical and Computer Engineering, King Abdulaziz University, Jeddah 21589, Saudi Arabia

Corresponding author: Syed Farooq Ali (farooq.ali@umt.edu.pk)

This work was supported by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, under Grant D-935-135-0062041-1443.

ABSTRACT Diabetic Retinopathy (DR) is an eye disorder in patients with diabetes. Detection of DR presence and its complications using fundus images at an early stage helps prevent its progression to the advanced levels. In the recent years, several well-designed Convolutional Neural Networks (CNN) have been proposed to detect the presence of DR with the help of publicly available datasets. However, these existing CNN-based classifiers focus on utilizing different architectural settings to improve the performance of detection task only i.e. presence or absence of DR. The further classification of the severity and type of the disease, however, remains a non-trivial task. To this end, we propose a multi-stream ensemble deep network to classify diabetic retinopathy severity. The proposed approach takes advantages of the deep networks and principal component analysis (PCA) to learn inter-class and intra-class variations from the raw image features. Ensemble machine learning classifiers are then applied to achieve high classification accuracy and robust performance on the obtained deep features. Specifically, a multi-stream network is made using pre-trained deep learning architectures i.e. ResNet-50 and DenseNet-121 to serve as the main feature extractors. Further application of PCA reduces the dimensionality of features and effectively separates the variation space of inter-class and intra-class images. Finally, an ensemble machine learning classifier using AdaBoost and random forest algorithms is built to further improve classification accuracy. The proposed approach has been compared with multiple conventional CNN-based approaches on Messidor-2 (two categories) and EyePACS (two, five categories) datasets. The experiment results show that our proposed approach achieves superior performance (upto 95.58% accuracy) and can be considered a promising method for automatic diabetic retinopathy detection.

INDEX TERMS Deep learning, ResNet, random forest, diabetic retinopathy, Messidor-2, EyePACS.

I. INTRODUCTION

Diabetic Retinopathy (DR) is an eyes disorder in the patients suffering from diabetes. Damage to the blood veins of the retina causes this disease. Diabetic retinopathy symptoms such as Microaneurysm (MA), Exudate (HE) Hemorrhage (HM), Cotton Wool Spot (CWS) can be seen on color fundus retinal imaging, according to several scientific investigations [1]. Microaneurysm is a swelling in the retinal blood veins that looks as a sharp-edged red spot on the retinal surface. Protein loss from tiny retinal veins causes exudates, which are white or pale yellow patches in the retina. Hemorrhages are deposits that look like red spots with non-uniform borders and are caused by thin and weak blood veins leaking.

Non-proliferative diabetic retinopathy (NPDR) and proliferative diabetic retinopathy (DPDR) are the two types of diabetic retinopathy (PDR). Based on the progression of lesions, the NPDR is then classified as 'mild', 'moderate', or 'severe' [2]. Mild DR is the earliest stage at which Micro Aneurysms form. Blood vessel swelling occurs when the

The associate editor coordinating the review of this manuscript and approving it for publication was Humaira Nisar.

illness advances to a moderate level, resulting in impaired vision. During the severe stage, abnormal blood veins development is observed. The last stage of DR is the PDR, in which extensive retinal fractures and detachment occur, resulting in complete blindness [3]. The DR is dangerous because in some cases, when not identified in early levels it will get the patient permanently blind. The patients suffering from DR have 25% more chances of permanent blindness than the people without DR. As a result, globally in persons aged 20 to 65, the leading cause of blindness is DR [4]. The 103.12 million adult population [5] of the world is affected by the DR in the 2020 as shown in Figure 1.

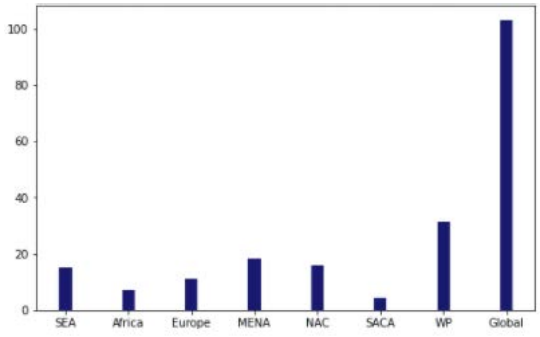

FIGURE 1. Adults population affected from DR in 2020 (in millions) from different regions of the world. MENA=Middle East and North Africa, NAC=North America and Caribbean, SACA=South and Central America, SEA=South East Asia, WP=Western Pacific [5].

The 2013 statistics show that 382 million population was suffering from DR [6]. In 2025, this population will undergo a rapid increase and will reach to 592 million. The Figure 1 shows the number of adults affected by the DR in the year 2020 in different regions of the world. DR-related blindness can be avoided by frequent retinal checkups. The ophthalmologists use manual techniques to detect the DR. They manually look at the color of retinal images of the patient and then identify the level of the DR. This method is very complex and buggy, and it also consumes a lot of time to detect the DR. The timely detection of the DR can save many people from the permanent blindness. Many machine learning (ML) and deep learning (DL) based DR detection techniques have been proposed in recent years.

TABLE 1. Number of DR cases between 2000 and 2010 in the US [7].

Year	Value	Short Value
2000	4,063,247	4,063
2010	7,685,237	7,685

The number of Americans with diabetic retinopathy is anticipated to nearly double between 2010 and 2050, from 7.7 million to 14.6 million shown in the Figure 2.

Following contributions are made in this paper.

- The novel technique for classifying diabetic retinopathy severity is proposed. The suggested technique extracts deep features from ResNet-50's and DenseNet-121's pooling layer, merges them, and then sends them to AdaBoost for classification using random forest (RF).
- The proposed approach (PA) uses ensemble classification that overcomes the problem of overfitting in the datasets with lesser training samples.
- PA outperforms existing approaches in two categories of Messidor-2 dataset in terms of percentage accuracy. These two categories consist of 'No Referable Diabetic Macular Edema Grade (DME)' and 'Referable DME'.
- PA outperforms existing approaches in two and five categories of the EyePACS dataset in terms of percentage accuracy. These five categories include: 'No DR', 'mild', 'moderate', 'severe','PDR'.
- The proposed approach is compared with state-of-the-art deep networks (Xception, Inception-V3, VGG-16, ResNet-50, and DenseNet-121) using EyePACS, Messidor-2, APTOS, and DDR datasets. The proposed approach outperformed these afore-mentioned deep architectures.
- The analysis of the performance classification of the proposed appoarch is made using 2, 3, and 5 categories of all four datasets, namely: EyePACS, Messidor-2, APTOS, and DDR datasets
- We have conducted an Ablation study that shows the effectiveness of the ensemble classifier used in the proposed approach.
- PA extends our previous work [8] in which deep features of ResNet-50 were used along with random forest classifier. This current approach performs the ensemble classification of deep features of ResNet-50, and DenseNet-121 and achieves better accuracy than our previous work.

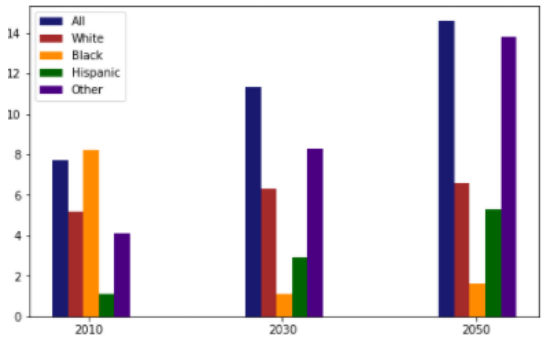

FIGURE 2. DR projection in 2030 and 2050 (in millions) in the US [7].

This paper is outlined as follows. In Section II, we have discussed the related work. Section IV describes datasets while Section III presents the proposed methodology. Experiments and results are presented in Section V. Section VI concludes the paper and provides future directions.

II. RELATED WORK

The advancement of automated DR pathology screening during the last few decades has been encouraging. In the literature, many deep learning and machine learning-based techniques have been presented. Akram et al. [9] detected the

presence of lesions in the retina using a mixture ensemble classifier built on the Gaussian Mixture Model (GMM) and Support Vector Machine (SVM). By combining the shape enhanced feature set with the intensity features, a similar strategy has been utilized to improve the model's classification accuracy by Akram et al. [10]. Several classification techniques like k-Nearest Neighbors (k-NN), AdaBoost, SVM, and GMM were applied and their performances were evaluated to detect lesions from non-lesions in the provided retinal images [11]. The area of hard exudates, the area of veins and arteries, branching points, texture, and entropy were extracted from retinal images using a hybrid feature extraction technique [12].

The techniques discussed above are not good in performance because they employ traditional classification techniques, which may not be enough for distinguishing between complicated actual data such as lesion and non-lesion pictures. Additionally, the domain knowledge of the input data is required by the approaches utilized in these feature engineering methods. Deep learning, particularly CNNs, offers significant assistance in addressing DR classification issues. Deep learning models can detect minor local characteristics straight from retinal pictures without the need for human assistance or domain expertise. Gulshan et al. [2] used the Incepetion-v3 for the diabetic retinopathy detection. The EyePACS-1 dataset, that includes 9963 images, as well as the Messidor-2 dataset were used to evaluate the model. According to their research the CNN-based models offers great sensitivity and specificity for diagnosing DR. Pratt et al. [13] suggested a CNN that can classify the retinal images in the five stages of the DR and also detect the haemorrhages, micro-aneurysms, and exudates. The CNN was trained on the Kaggle EyePACS data set. They solved the problems of overfitting and skewed datasets as well as suggested a solution. They used the data augmentation to increase the size of data for training. Their CNN used model has 3 fully connected layers and ten convolutional layers. The suggested CNN has 75% accuracy, 30% sensitivity, and 95% specificity.

A DR level classification model with CNN was developed by G. García et al. [14] Their model achieves 93.65% specificity and 83.68% accuracy on EyePACS dataset. Wng et al. [15] used the Kaggle dataset to test the performance of three pre-trained CNNs in classifying all stages of the DR. Inception-Net V3, Alex-Net, and VGG-16 were the three CNN architectures employed. Inception-Net V3 had the highest average accuracy of 63.23%. Esfahan et al. [16] applied a well-known CNN, ResNet-34, for DR classification using the Kaggle dataset and achieved 85% accuracy. To enhance the quality of the images, they used a bunch of image preparation methods. The weighted-addition, gaussian filter, and image-normalisation were used as the image pre-processing techniques. Dutta et al. [17] used the Kaggle dataset to identify and classify DR photos into five DR phases. Using 2000 images, they evaluated the performance of three networks: the CNN, the Deep Neural Network (DNN), and the Back Propagation Neural Network (BNN). Before being input into the networks, a number of filters were implemented. The CNN model was pretrained VGG-16, which has sixteen conv, four maxpooling, and three FC layers, whereas the DNN has three FC layers. Their DNN beats the CNN and the BNN and achieves an accuracy of 86.3%.

Lam et al. [18] for the identification of DR staging, applied convolutional neural networks (CNNs) on colour fundus images. Their network achieved the sensitivity of 95% and the accuracy of 74.5%, 68.8%, and 57.2% for 2, 3, and 4 class classification models. C. Lian et al. [19] used the Alexnet, ResNet-50 and VGG-16 to for the classification of the DR. They focused on network designs, preprocessing, class imbalance, and fine-tuning while using convolutional network to solve the DR classification and achieved an accuracy of 73.19% for Alex-Net, 76.41% for ResNet-50, and 79.04% for VGG-16 on EyePACS dataset. The DR classification CNN model was introduced by Shaban et al. [20], which used the leave-one-out method to test the retinal images and achieved the accuracy, sensitivity, and specificity of 80.2%,78.7%, and 84.6% respectively.

Hongyang et al. [21] used 3 pre-trained CNN architecture to categorize their dataset: Inception-V3, Inception-ResNet-V2, and ResNet-152. The Adam optimizer is also used to adjust CNN's weights during their training. The AdaBoost framework was used to ensemble these models. Their model achieved an accuracy of 88.21%. Wei et al. [22] suggested a technique for detecting the DR using a private dataset that contained 13,767 images divided into four categories. The images were cropped, scaled to fit every network's requirements. They used ResNet-50, Inception-V2, Inception-V3, Xception, and DenseNet to fine-tune pre-trained CNN architectures to identify the DR.

To identify all five DR levels, Harangi et al. [23] combined the existing pre-trained AlexNet with hand-crafted characteristics. The Kaggle dataset was used to train the CNN, while the IDRiD was used to test it. For this study, the accuracy was 90.07%. To identify referable DR images, Yi-Peng et al. [24], also developed a weighted pathways CNN (wp-CNN). In order to remove class imbalance distribution, they augmented the images. Before feeding these images to the CNN, they were sized to 299 x 299 pixels and normalised. The wp-CNN consisted of several conv layers with varying kernel sizes in multiple weighted channels that were fused by averaging. With 94.23% accuracy in their dataset and 90.8% on the STARE dataset, the wp-CNN of 105 layers outperformed pre-trained Resnet, Se-net, and DenseNet models. Anj et al. [25], used different CNN for the DR detection and severity classification on the EyePACS dataset. To achieve better results, they used image processing techniques like local average colour subtraction to help in emphasizing the important characteristics from a fundoscopy, hence improving the Diabetic Retinopathy identification and assessment procedure. They got 71.7 % accuracy, using VGG-16 model. They also applied the VGG-19 and Inception-V3 on the EyePACS dataset and got 79.9% and 70.2% accuracy respectively. ResNet-50, Inception-v3, Xception, DenseNet-121,

and Dense-169 were used to suggest a strategy for DR detection using ensemble classification of deep convolutional neural networks by Qmr et al. [26]. On the publicly accessible Kaggle dataset, the suggested approach outperformed the previous approaches, achieving an accuracy of 80%. Jod et al [27] proposed DR classification technique.By giving a value to each point in the hidden and input spaces, their classifier is capable of explaining the classification outcomes. Their classifier achieves the accuracy of 91% on the Messidor-2 data set for binary classification. Mjr et al [28] proposed a multitasking DL model for DR detection. They created a multitask model that combines a classification and regression model. Both models have their separate loss functions and were trained independently. The multilayer perceptron network takes features as input from the before mentioned models and then categorize the data set images for the diabetic retinopathy. They obtained an accuracy of 82% on the EyePACS dataset and 86% accuracy on the APTOS Dataset.

III. METHODOLOGY

This paper presents a multi-stream deep neural network for classification and grading of diabetic retinopathy using EyePACS, Messidor-2, APTOS, and DDR datasets using 2, 3 and 5 categories. Our proposed multi-stream approach (PA) consists of multiple deep networks including ResNet-50, and DenseNet-121 followed by dimensionality reduction (using PCA) and ensemble classification (boosting). Our approach uses the transfer learning for the deep networks. As shown in Figure 3, PA consists of four steps, namely: pre-processing, feature extraction, dimensionality reduction, and ensemble classification.

The suggested approach uses CNN-based networks for feature extraction. Our approach use two streams of inputs in the form of features extracted from the two deep networks. The features extracted from these networks then fused and fed to the classification model.

A. DEEP FEATURE EXTRACTION

The deep features of densely connected neural network and residual network are extracted from their pooling layers. DenseNets provide a number of compelling advantages, including the elimination of the vanishing gradient problem, improved feature propagation, feature reuse, and a significant reduction in the number of parameters [29]. Without raising the training error percentage, ResNet with a higher number of layers (even thousands) can be trained easily. Using identity mapping, ResNets overcomes the vanishing gradient problem. We implemented Xception, Inception-V3, VGG-16, ResNet-50, and DenseNet-121 on EyePACS, Messidor-2, APTOS, and DDR datasets respectively as shown in Table 2. It has been empirically observed that best accuracies are shown by ResNet-50 and DenseNet-121 respectively. Therefore, we concatenated their features in our PA and got the best results in terms of percentage accuracy.

TABLE 2. Comparison of % accuracy of state-of-the-art deep architectures with PA on EyePACS, Messidor-2, APTOS, and DDR.

DataSets	Xception	Inception-V3	VGG-16	ResNet-50	DenseNet-121	PA
EP	72.00	75.38	75.03	81.95	82.48	85.46
M-2	78.77	74.09	79.32	78.21	81.10	86.78
APTOS	65.90	65.98	61.10	62.51	64.35	72.53
DDR	62.80	59.60	62.30	63.00	61.40	68.24

1) RESIDUAL NETWORK

Kaiming He et al. [30] developed the residual neural network (ResNet). The performance of the deep network is dependent on the depth of the network. For the same dataset, various depth levels might produce different outcomes. The model performs better as the number of layers increases. Network depth cannot be simply increased by adding more layers, because of the vanishing gradient problem. Because the gradient is transmitted back to prior levels, repeated multiplication may result in a very small gradient. To solve this issue, the residual neural network was developed. ResNet offers a variety of architectures, including 18, 34, 50, 101, and even 152 layer architectures.

The classification results will be determined by passing the results of each filter in the ResNet architecture through average pooling and entering the fully connected layer network with softmax activation function [31]. The Figure 4 presents the visual representation of the ResNet-50 deep features using EyePACS dataset.

In comparison to shallow networks, images processed across deep networks have a higher chance of obtaining more precise and abstract information. Nonetheless, training a deep CNN while preserving gradient flow across deep layers to make it converge in a reasonable amount of time is quite difficult. It is natural to expect that when the number of layers within a network grows, the network's accuracy will begin to decline and degradation might arise. Additionally, some typical network training problems during back-propagation, such as gradient vanishing, convergence time, are common in deeper networks. To address these difficulties, residual learning emerges as an effective technique for training deep CNNs with a faster convergence and increased accuracy of the network. In this technique, some of the alternative training layers are skipped or bypassed by learning identity function. Another advantage is the information of the previous layer can be added up in the subsequent layers as well.

The input $x_{[n-1]}$ is added to the output $y_{[n-1](x[n-1])}$ in the next layer as shown in Equation 1.

$$x_{[n]} = y_{[n]}\left(x_{[n-1]}\right) + x_{[n-1]}, \tag{1}$$

$x_{[n]} - x_{[n-1]}$ becomes the final prediction. Learning residual images rather than the actual input images are much easier for the network.

The features map extracted from ResNet-50 for a single eye image from EyePACS dataset is presented in Figure 5.

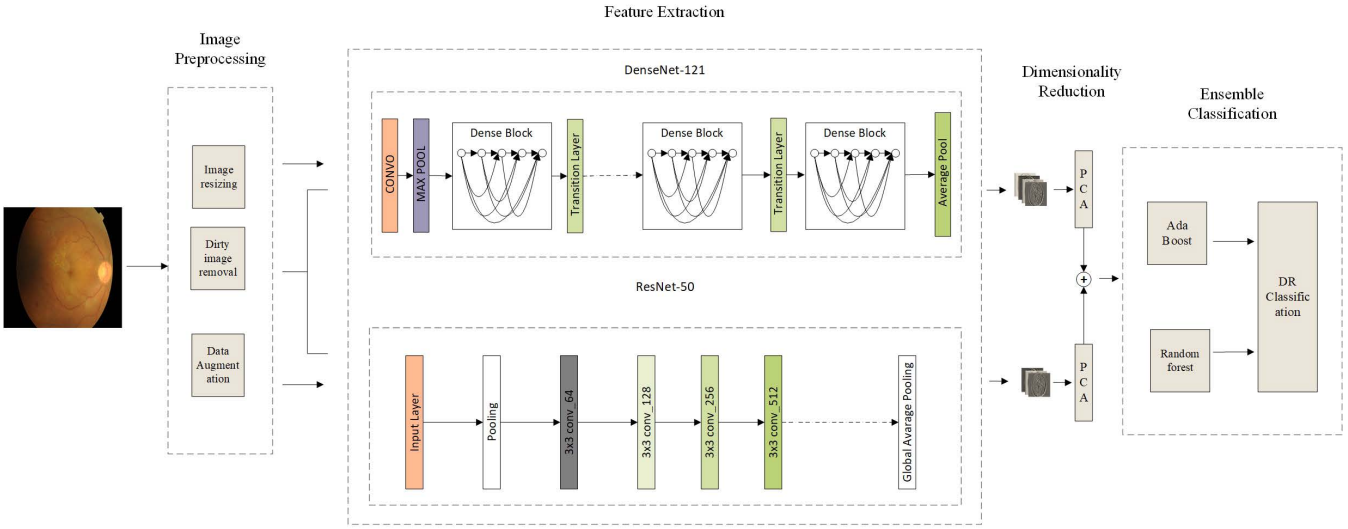

FIGURE 3. Architecture of PA with features extracted from DenseNet-121 and ResNet-50 followed by dimensionality reduction by PCA and then its ensembled classification(Boosting).

2) DENSELY CONNECTED NEURAL NETWORK

The DenseNet network focuses on deepening the DL networks and it increases the training efficiency by decreasing the layers connections. All layers are interconnected with each other, such as second, third, and fourth layers are connected to the first layer. The same goes for other layers as well. This layer's interconnection allows the maximal flow of information between network layers. Every layer, take input from its preceding layer and sends its feature maps to all subsequent layers. DensNet is different from the ResNet in a way that it uses the concatenation operation to combine the features rather than using summation. Aside from the fundamental convolutional and pooling layers, DenseNet has two key components, namely: dense-blocks and transition layers. The Figure 6 presents the visual representation of the DenseNet-121 architecture.

As discussed earlier, in the standard CNN, the input image is processed through numerous convolutional (Conv) layers to extract high-level information. As CNNs go deeper, the information disappears as it approaches its destination because of the longer route between the input and output layers. In case of ResNet model, identity mapping was used to enable gradient propagation, in which element-by-element addition is used. It can be thought of as procedures with a state transmitted from one ResNet unit to the other. On the contrary, in DenseNet, all preceding layers send the additional inputs to the next layers, which then passes its own features-maps to all succeeding layers. Because each layer receives feature maps across all previous layers, the framework could become more compact and lighter, ending up in fewer channels. In contrast to ResNets, we never sum features before passing them into a layer; instead, we concatenate them before passing them into a layer. As a result, the n-th layer has inputs, that are feature-maps from all the previous Conv blocks. All L following layers receive their own feature-maps. In an L-layer network, this offers $L(L+1)/2$ connections rather than just L, as in conventional frameworks. The DenseNet is organized in *db, each of which has a different set of filters but the same dimensions. Therefore, it has better performance accuracy and memory efficiency.

Each layer with extra number of channels is shown by the growth rate k. The l-th layer's generalization is aided by the growth rate (k). It determines how much information is incorporated to each layer. $k_{(l)} = (k_0 + k(l-1))$ is utilized to compute the growth rate of DB. The input tensor proceeds through a sequence of Conv operations with a predefined number of filters (k) in every dense block, with the output of each one being concatenated to the original tensor. As a result, at every internal step of the dense block, the feature maps of the given tensor grow arithmetically by k feature vectors per stage. DenseNet has many advantages over traditional networks.

The feature map extracted from the afore-mentioned deep architectures, ResNet-50 and DenseNet-121, are larger in size. Therefore, we applied principal component analysis (PCA) to reduce its dimensionality. PCA minimizes the dimensionality of a dataset using lot of linked variables while keeping as much variance as feasible. We used first 10 eigen vectors corresponding to largest eigen values. The feature vector of one stream (ResNet-50) was reduced from 14500×10 dimensions to 300×10 dimensions while the features of other stream (DenseNet-121) were reduced from 14500×10 dimensions to 300×10 dimensions. This dimensionality reduction was achieved by applying principal component analysis (PCA).

The Figure 7 and represent the linear projection features after the application of PCA on the EyePACS and Messidor-2 respectively.

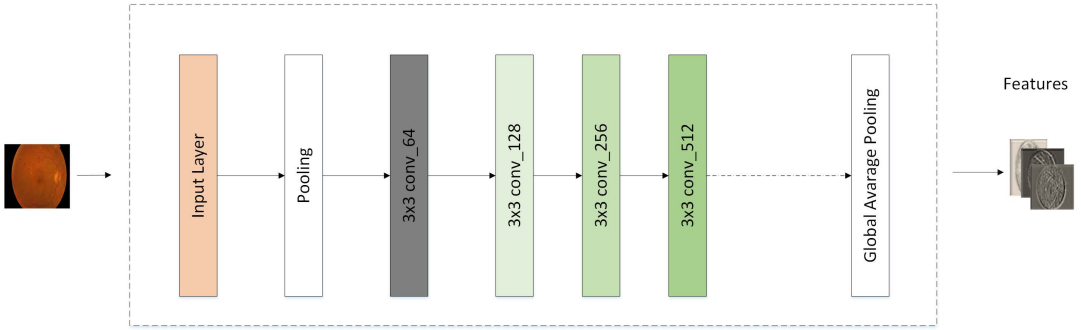

FIGURE 4. Key layers of ResNet-50 architecture.

B. ENSEMBLE CLASSIFICATION

Ensemble classification involves learning a group of classifiers, referred to as an ensemble of classifiers, and then combining their results for the classification of unknown cases to use some type of voting. The ensemble classification technique makes use of multiple classifiers and has become famous due to its excellent performance. Bagging, blending and boosting are the two methods available. Multiple classification methods are used in bagging, and the most accurate one is voted at the end. The last classifier is the one who receives the most votes.When there is bias or underfitting in the data, bagging isn't useful.Another problem of bagging is that we cannot comprehend which characteristics are being picked during sampling, which means that some features may never be utilised, resulting in the loss of critical data. Boosting is a technique that employs a sequence of different classifiers. Each classifier's weights are modified based on the preceding classifier. The data is first separated into several parts, then one of them is confirmed with the help of others, and so on.When dealing with bias or underfitting in a data set, the boosting approach is suitable.We have ensembled the random forest and AdaBoost classifier in our PA.

1) AdaBoost ALGORITHM

Freund and Schapire proposed boosting in 1990 [32]. It is a common way of coordinated learning and an effective instrument for increasing the learning system's prediction capacity. AdaBoost is a self-adaptive boosting technique that uses a set of multiple classifiers to improve the performance of weak classifiers [33]. It adjusts to the basic algorithm's mistake rate during training by dynamically changing its weight for every input.

Boosting approach is based on a conceptual examination of the Probably Approximately Corect (PAC) learning method. The ideas of strong and weak learning were suggested by Kearns and Valiant. If a polynomial learning algorithm exists to classify the set of concepts in the PAC learning approach, and the recognition accuracy is high, then this set of concepts is considered strong learning (SL). If the probability of correct identification for this learning algorithm is only slightly higher than random guessing, this group is considered as weak learning (WL). The question of similarity between WL and SL algorithm, suggested by Kearns and Valiant, is if the WL algorithm could be upgraded to SL algorithm or not?. If the two are comparable, we can boost an algorithm that is marginally superior than random guessing to a strong learning algorithm only if we uncover it when learning the concepts. Before and after the training, boosting will generate a sequence of classifiers.

Each classifier's training set is a subset of the overall training set, and whether each sample appears in that subset or not is determined by the performance of the preceding classifiers. The samples that are judged to be incorrect by the existing classifiers will have a higher probability of appearing in the new training subset, causing resulting classifiers to focus more on the issue of differentiating samples, which appears to be quite difficult for the existing classifiers. The AdaBoost technique was commonly used to combine numerous weak classifiers into a single strong classifier.

In this study, we presented a multi-stream network that uses the features of afore-mentioned CNNs and classifies the stages of DR using random forest under a boosting framework. The three steps of the AdaBoost algorithm are as follows [21]. Given a set of samples (a_i, b_j) with j=1,…n, a_i is the feature vector and b_j is the label of a_i. The samples' distribution is set up as follows:

$$d_1(j) = \frac{1}{n} \tag{2}$$

Then it select hypothesis model h_t with the weighted error for t=1,…,T:

$$\epsilon_t = P_{j\ D_t}[h_t(a_j) \neq) b_j] \tag{3}$$

Each h_t's weight is calculated using Equation 4.

$$\epsilon_t = \frac{1}{2} \ln(\frac{1-\epsilon_t}{\epsilon_t}) \tag{4}$$

The sample distribution has been modified as shown in Equation 5.

$$D_{t+1}(j) = \frac{D_{t+1}(j) \exp -\alpha_t b_j h_t(a_j))}{Z_t} \tag{5}$$

Z_t stands for the normalisation factor.

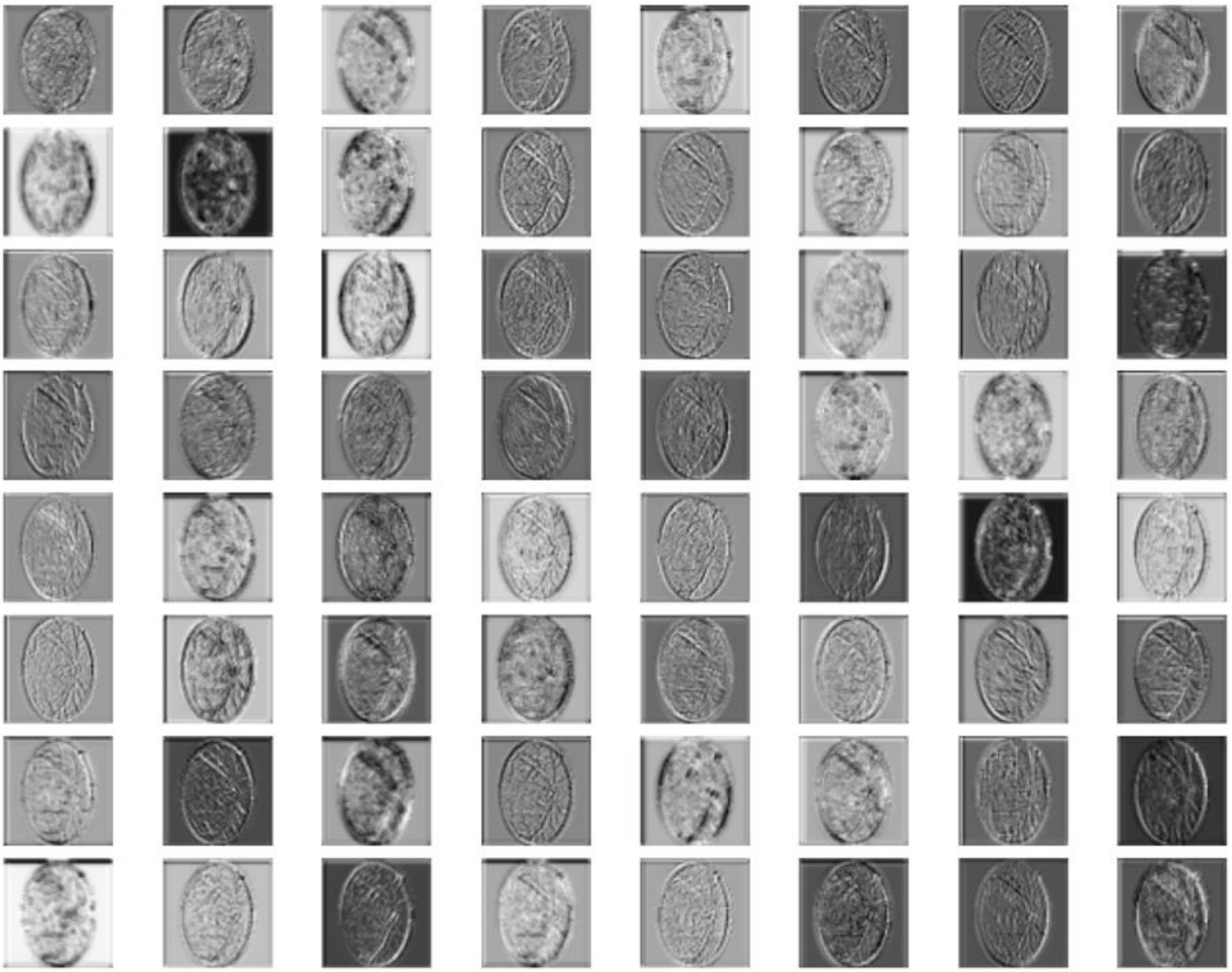

FIGURE 5. Visual representation of ResNet's deep features for a single retinal image from EyePACS dataset having diabetic retinopathy with 'moderate' severity level.

The final integrated model is derived in Equation 6 when the weighted error is convergent and almost unchanged.

$$H_a(j) = \sum_{t=1}^{T} \alpha_t h_t(x) \tag{6}$$

2) RANDOM FOREST

Several trees are produced together in the Random Forest classification approach. The new example is added down towards each of the trees for classifying test data. Every tree creates a class for test data, which is known as class voting [34]. The classifier chooses the most popular class as the test data's final classification. Because it works on big datasets in a time-saving method, random forest classification is a widely used ensemble model classifier. For machine learning, a random forest classifier with a percentage split is utilized. The training dataset, which comprises two-thirds of the entire data, is used to aid tree growth. Cases are chosen at random and replaced, which means that a case evaluated for a tree might be reallocated to another tree. The square root of the total number of feature variables is typically chosen at random from all of the specified feature values. During the forest's growth, its value remains constant. To divide a node, the best split on these specified feature variables is utilized. The remaining one-third of the data is referred to as the test dataset. Out-of-bag (OOB) data is the name given to this type of data.

Each tree creates a class for each test data that is used to calculate the vote for that test data class. The test data is assigned to the class with the most votes. This classifier is connected with the term 'random' in two ways i.e., sample data and feature variables both were chosen at random. Random forest classifier has an advantage of not requiring a separate test dataset. Internally, at the time of building, the OOB data is utilized to calculate the inaccuracy. When each forest tree has reached maturity, OOB instances are hung from the tree, and the number of votes for the proper class is computed.

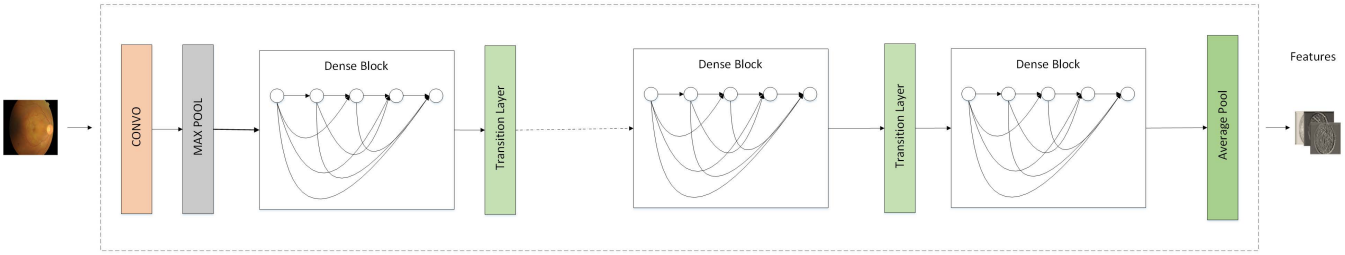

FIGURE 6. Key layers of DenseNet-121 architecture.

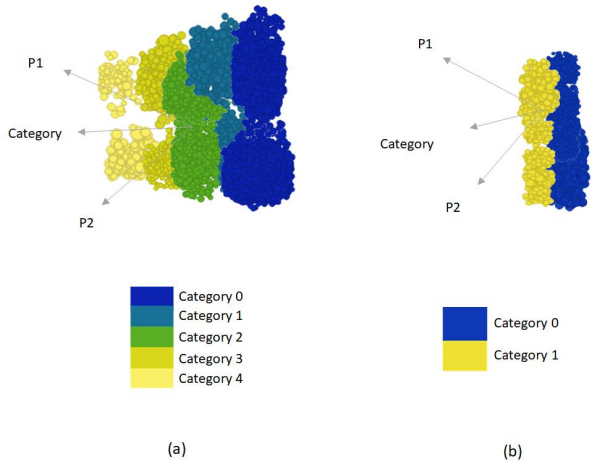

FIGURE 7. Linear Projections of features using first two principal components (P1 and P2) for a) EyePACS and b) Messidor-2.

- Set the total number of classes and feature variables to A and B, respectively.
- Let b represent the number of chosen feature variables at a node ($b = B$ in most cases).
- Select a subset of the dataset comprising A distinct classes at random with a replacement for each decision tree.
- To determine the optimum split and decision at each node of a decision tree, choose b feature variables at random for each node.

IV. DATASET

EyePACS[1] is the largest retinal image dataset that is publicly available on the Kaggle website. California Health Care Foundation provided this dataset for the DR competition on Kaggle. It contains around 80,000 retinal images. This data is collected from California and other parts of the world, primary care clinics, and contains the left and right images of the retina. But this data is a bit noisy, out of focus, and has some exposure issues. The data is labeled on the ICDRDSS scale. Key retinal images of this dataset with different DR levels are shown in Figure 8.

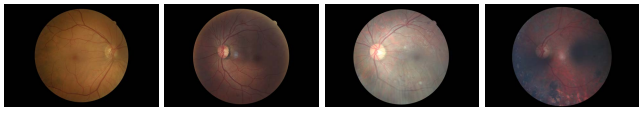

FIGURE 8. Key retinal images from EyePACS dataset showing different severity levels of DR. a=Mild, b=Moderate, c=Severe, d=Proliferative.

Messidor-2 dataset[2] [35], [36] includes 1058 pictures from the Messidor-1 dataset, as well as 690 additional photographs taken in the Brest University Hospital's Ophthalmology department between year 2009 and 2010. The third data set we used for experimentation is the APTOS. To detect the DR automatically, Aravind Eye Hospital from India's remote areas, gathered the data to develop sophisticated tools to detect DR automatically and enhance the hospital's capability to identify new patients. We increased the size of Messidor-2 from 1748 to 3000 images and APTOS from 3363 to 6000 images by applying data augmentation.

The second-largest dataset that is publically available is DDR, which contains 12522 pictures. DDR is relatively new dataset, collected from Chinese hospitals from 2016 to 2018. The data is scaled on the ICDRDSS by various specialists.

We used two, three and five categories of the aforementioned datasets. The categories are basically the labelling of the data-set images with respect to the severity of the DR. In two category classification, the retinal images are labelled as the No-DR and DR images while in case of three, they are labelled as No-DR, Mild-DR, and Severe-DR. In five category classification, the retinal images are labelled as No-DR, Mild-Dr, Moderate-DR, Severe-DR, and Proliferative-DR.

V. EXPERIMENTS AND RESULTS

In this section, we have analyzed our PA by doing various experiments using publicly available datasets. For datasets with fewer images including Messidor-2 (1700 images), and APTOS (3000 images), we have performed data augmentation to increase the size of training data. The PA was compared with various state-of-the-art deep architectures (Xception, Inception-V3, VGG-16, ResNet-50 and DenseNet-121) and approaches (Wng [15], Lia [19], Anj [25], Qmr [26], Mjr [28], Lam [18], Gab [14],

[1]https://www.kaggle.com/c/diabetic-retinopathy-detection/data

[2]http:/www.adcis.net/en/third-party/messidor2/

TABLE 3. Performance measures of PA for EyePACS dataset for each class.

Class	Precision	Recall	F1-score
0	0.90	0.92	0.91
1	0.90	0.80	0.84
2	0.89	0.69	0.78
3	0.49	0.84	0.62
4	0.62	0.81	0.70

Jod et al. [27]) using datasets, namely; EyePACS, Messidor-2, APTOS, and DDR. The results showed that it outperformed these. In order to conduct component-wise analysis, we conducted the ablation study as well. Moreover, we implemented and compared our PA, in terms of various performance measures, on various categories of datasets (2, 3, and 5). These performance measures include :accuracy measure refers to the percentage of properly predicted samples and is widely used in classification tasks.The ratio of True Positives to all Positives is precision,whereas recall is the percentage of genuine positives predicted as true positives.The f1-score is the harmonic mean of precision and recall. The equations used to calculate accuracy, precision, recall and F1-score [37] are mentioned below.

$$A = \frac{tp + tn}{tp + tn + fp + fn} \tag{7}$$

$$P = \frac{tp}{tp + fp} \tag{8}$$

$$recall = \frac{tp}{fp + fn} \tag{9}$$

$$F1 - score = \frac{2 * precision * recall}{precision + recall} \tag{10}$$

The precision accuracy and the F1-score for the EyePACS dataset are shown in the Table 3.

A. ENVIRONMENT

For feature extraction from the deep architectures, we used Google Colab while the classification was done using Weka. The implementation of these deep architectures was done using python. Furthermore, we used the Keras and various other python libraries including numpy, pandas, and tensorflow.

B. EXPERIMENT: DATA AUGMENTATION

In order to increase the size of training data, we applied data augmentation on those datasets that have fewer samples, namely; Messidor-2 and APTOS. After applying the data augmentation, the size of these aforementioned datasets increased from 1748 and 3668 images to 2600 and 7000 respectively. We used rotation and flipping operations for performing data augmentation using Keras library. In the Table 4, that the PA shows better accuracy on data set with augmentation for 2, 3, and 5 categories respectively. The reason of lower accuracy on datasets without augmentation was due to overfitting problem, which was later resolved by applying augmentation.

TABLE 4. Categories wise % accuracy of PA on Messidor-2 and APTOS datasets with or without augmentation. M-2= Messidor-2, 2-cat=2-category, 3-cat=3-category, 5-cat=5-category, A=Augmentation, WA=Without Augmentation.

Category	M-2 ataset		APTOS dataset	
	WA	A	WA	A
2-cat	93.14	95.58	87.10	89.00
3-cat	85.47	89.47	73.54	77.90
5-cat	83.12	86.78	69.37	72.53

TABLE 5. Comparison of % accuracy of different ensemble classification techniques with PA using EyePACS, Messidor-2, APTOS and DDR datasets. EP = EyePACS.

Data-Sets	Bagging	Blending	Boosting-(PA)
EP	83.94	84.23	85.46
M-2	84.73	85.90	86.78
APTOS	71.81	70.60	72.53
DDR	67.35	66.34	68.24

C. EXPERIMENT: ABLATION STUDY

For ablation study, we replaced ensemble classification of boosting in PA with the other two ensembling techniques including bagging, and blending as shown in Table 5. It can be seen that PA outperformed aforementioned techniques on all the four datasets, namely; EyePACS, Messidor-2, APTOS, and DDR. The reason include that AdaBoost handles the overfitting problem by using various weak learners, and it fully considers the weight of each classifier. Bagging gives lowest percentage accuracy on EyePACS, and Messidor-2 datasets, because bagging ignores the values with the highest and lowest results that may have wide difference and provides an average result. Blending gave lowest percentage accuracy on APTOS, and DDR datasets because blending do not handle well the class imbalance and over-fitting issues available in these datasets.

We compared the percentage accuracy of our PA by replacing its classifier with others including J-48, random forest, SVM, AdaBoost, naive bayes, and decision tree as shown in Table 6. It can be seen that PA outperformed these classifiers on all four datasets, namely; EyePACS, Messidor-2, APTOS, and DDR. AdaBoost uses multiple weak classifiers and fully considers the weight of each classifier, and hence, less prone to over-fitting. The worse performance was shown by naive bayes on Messidor-2 and DDR datasets. Naive bayes is the bad estimator due to the zero frequency problem. J-48 and decision tree showed the lowest accuracy for EyePACS and APTOS datasets because of the class imbalance issue in these datasets.

D. CATEGORY-WISE PERFORMANCE

This subsection provides the accuracy of proposed approach on all four datasets (Messidor-2, EyePACS, APTOS, DDR) for two, three and five categories of diabetic retinopathy. As shown in Table 7, the PA shows best performance in terms of percentage accuracy on Messidor-2 dataset. This

TABLE 6. Comparison of % accuracy of different classification models with PA on EyePACS, Messidor-2, APTOS and DDR datasets. DT=decision table, RF=random forest, AdB=AdaBoost.

Data-Sets	J48	RF	SVM	AdB	NB	D-Tree	PA
EP	76.00	81.70	80.38	83.53	79.22	79.57	85.46
M-2	79.82	80.90	81.32	81.48	78.84	82.21	86.78
APTOS	70.63	69.00	70.25	67.10	67.60	65.40	72.53
DDR	65.90	65.98	64.10	66.70	63.86	64.31	68.24

TABLE 7. Category-wise % accuracy of PA on EyePACS, Messidor-2, APTOS and DDR datasets, 2-cat=2-category, 3-cat=3-category, 5-cat=5-category.

Category	M-2	EP	APTOS	DDR
2-cat	95.58	89.20	89.00	76.81
3-cat	89.47	86.81	77.90	72.49
5-cat	86.78	85.46	72.53	68.24

dataset has lesser noisy images and their quality is better than other datasets. PA has shown least performance on DDR because of class imbalance and noisy images in this dataset. It can be observed that as the categories of the dataset increase, the performance of PA decreases. The accuracy of PA on Messidor-2 reduces from 95.58% to 86.78% as categories increases from 2 to 5. The reason includes curse of dimensionality.

E. EXPERIMENT: COMPARISON WITH DEEP ARCHITECTURES

In this subsection, we have compared our PA with state-of-the-art architectures, namely; Xception, Inception-V3, VGG-16, ResNet-50, and DenseNet-121. As shown in Table 2, PA, using features of ResNet-50 and DenseNet-121, can deliver more accurate classification results than the other models deployed independently on EyePACS, Messidor-2, APTOS, and DDR. The reasons includes ensemble classification of deep features that reduces overfitting. Inception and VGG-16 showed worst performance on Messidor-2, and APTOS. The reason might include overfitting as both these datasets have lesser training data.

F. EXPERIMENT WITH EXISTING APPROACHES

In this subsection, we have compared PA with state-of-the-art approaches using EyePACS, and Messidor-2 datasets.

1) EXPERIMENT: EyePACS

PA is compared with Mjr, Wng, Anj, Qmr, and Lam on EyePACS dataset using 5 categories as shown in Figure 9. It can be observed that PA outperformed the other approaches and achieved an accuracy of 85.46%. EyePACS has class imbalance problem and boosting used in PA solves this problem. Wng gives the lowest accuracy of 63.23% on this dataset, because it uses Inception-V3,which has the problem of overfitting on lesser training data.

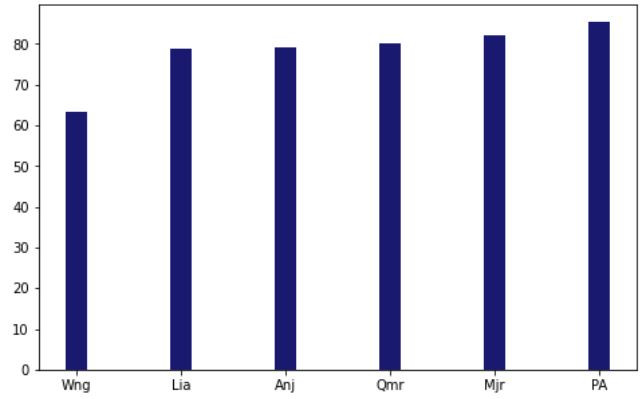

FIGURE 9. Comparison of PA with existing approaches in terms of accuracy on EyePACS dataset using 5 categories.

TABLE 8. Comparison of PA accuracy for 2 categories with existing approaches for EyePACS dataset.

Dataset	Lam	Anj	Gab	PA
EP	74.50	80.4	83.68	89.20

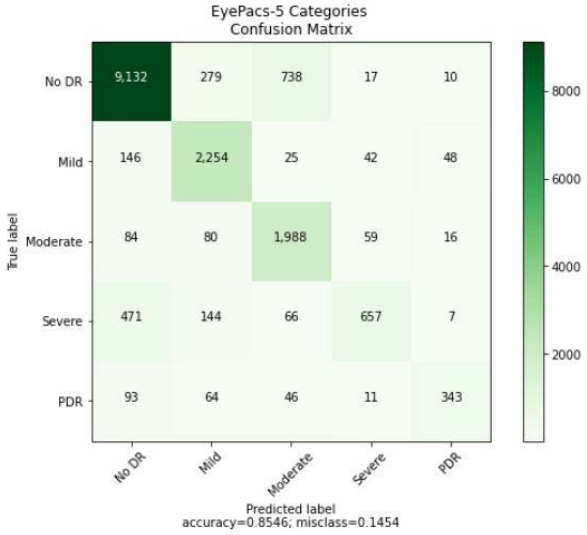

FIGURE 10. Proposed method's confusion matrix for EyePACS Dataset using 5 categories.

PA is compared with Anj, Lam, and Gab on the same dataset using 2 categories as shown in Table 8. It can be observed that PA outperformed these approaches and achieved an accuracy of 89.20%. Lam showed least performance on this dataset as depth of AlexNet (used in it) is very less.

The confusion matrix is the most important factor to consider when evaluating a model. Using the fundamental equations described above, the performance measures are further computed from the associated confusion matrix. The proposed model's confusion matrix for the task of DR classification on the EyePacs dataset is shown in Figure 10. We can see that the number of retinal images classified correctly is lower, which makes it suitable for deploying in healthcare facilities and hospitals. Moreover, the false negative for proliferative DR categories is very low i.e., 0.012% which

TABLE 9. Comparison of PA accuracy for 2 categories with existing approaches for Messidor-2 dataset.

Method	Year	Accuracy
Jod	2020	91.00
PA	2021	95.58

reduces the chances of mistreatment of a patient suffering from this disease.

2) EXPERIMENT: MESSIDOR-2

PA have achieved the accuracy of 95.8% for the Messidor-2 dataset for 2-category DR classification as shown in the Table 9. PA yields better percentage accuracy than Jod as ensemble classification used in PA reduces overfitting problem of Messidor-2 that arises due to its lesser training size.

VI. CONCLUSION AND FUTURE WORK

In this study, we introduced a novel ensemble techniques for automated grading and classification of diabetic retinopathy, which is built upon deep learning models, namely: ResNet-50, and DenseNet-121. To overcome the problem of high dimensional feature vector of afore-mentioned deep models, we applied PCA for dimensionality reduction. In order to justify the usage of ResNet-50 and DenseNet-121, we computed the percentage accuracies of each state-of-the-art deep networks (Xception, Inception-V3, VGG-16, ResNet-50,and DenseNet-121) using EyePACS, Messidor-2, APTOS, and DDR. We found out that the best accuracies are shown by ResNet-50 and DenseNet-121. That is why, the proposed method, using the features of these two afore-mentioned deep network, outperformed the other networks on all four datasets. The ablation study was also performed on proposed method to see the effect of replacing the boosting with other classifiers (J48, Random Forest, SVM, AdaBoost, Naive Bayes, and Decision Tree) and ensembling methods (blending and bagging). In order to ensure rigorous experimentation, we compared our proposed method with 9 state-of-the-art approaches. Results showed that our proposed approach outperformed the other approaches and achieved an accuracy of 95.58%, 89.20%, 89%, and 76.81% on Messidor-2, EyePACS, APTOS, and DDR dataset respectively.

It can be observed that percentage accuracy increases with the decrease in number of categories and vice versa. In case of Messidor-2, when we move from two to five categories, the percentage accuracy also reduces from 95.58% to 86.78% respectively. Similarly, in case of EyePACS, it changes from 89.20% to 86.78%. The reason includes the curse of dimensionality. In order to avoid reduction in percentage accuracy with the increase in categories, the dataset need to be increased exponentially. Hence, the addition of large data repositories in future would be helpful for generating more promising results.

We hope to extend our deep learning algorithm in future to work in an uncontrolled environment and will replace our current dimensionality reduction PCA technique with auto encoders to increase the accuracy of our proposed approach. More testing on real-world circumstances is necessary for clinical applications, and the system should be made more robust to run on low cost devices for quick response. As compared to manual diagnosis, the automated approaches are quicker and enables the doctors to consult more patients in lesser time. In near future, compact deep learning solutions for multiple devices with better accuracy will be in great demand.

ACKNOWLEDGMENT

The authors would like to thank DSR for their technical and financial support.

REFERENCES

[1] U. R. Acharya, M. R. K. Mookiah, J. E. W. Koh, J. H. Tan, S. V. Bhandary, A. K. Rao, H. Fujita, Y. Hagiwara, C. K. Chua, and A. Laude, "Automated screening system for retinal health using bi-dimensional empirical mode decomposition and integrated index," *Comput. Biol. Med.*, vol. 75, pp. 54–62, Aug. 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482516301044

[2] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, J. Cuadros, and R. Kim, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *J. Amer. Med. Assoc.*, vol. 316, no. 22, pp. 2402–2410, Dec. 2016.

[3] R. Williams, M. Airey, H. Baxter, J. Forrester, T. Kennedy-Martin, and A. Girach, "Epidemiology of diabetic retinopathy and macular oedema: A systematic review," *Eye*, vol. 18, pp. 963–983, Jul. 2004.

[4] R. Klein, B. E. K. Klein, S. E. Moss, M. D. Davis, and D. L. DeMets, "The Wisconsin epidemiologic study of diabetic retinopathy: III. Prevalence and risk of diabetic retinopathy when age at diagnosis is 30 or more years," *Arch. Ophthalmol.*, vol. 102, no. 4, pp. 527–532, Apr. 1984, doi: 10.1001/archopht.1984.01040030405011.

[5] Z. L. Teo, Y.-C. Tham, M. Yu, M. L. Chee, T. H. Rim, N. Cheung, M. M. Bikbov, Y. X. Wang, Y. Tang, Y. Lu, I. Y. Wong, D. S. W. Ting, G. S. W. Tan, J. B. Jonas, C. Sabanayagam, T. Y. Wong, and C.-Y. Cheng, "Global prevalence of diabetic retinopathy and projection of burden through 2045: Systematic review and meta-analysis," *Ophthalmology*, vol. 128, no. 11, pp. 1580–1591, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0161642021003213

[6] N. Kaur, S. Chatterjee, M. Acharyya, J. Kaur, N. Kapoor, and S. Gupta, "A supervised approach for automated detection of hemorrhages in retinal fundus images," in *Proc. 5th Int. Conf. Wireless Netw. Embedded Syst. (WECON)*, Oct. 2016, pp. 1–5.

[7] NEI. *Dr Datar*. Accessed: Oct. 26, 2022. [Online]. Available: https://www.nei.nih.gov/learn-about-eye-health/outreach-campaigns-and-resources/eye-health-data-and-statistics/diabetic-retinopathy-data-and-statistics/diabetic-retinopathy-tables

[8] M. K. Yaqoob, S. F. Ali, M. Bilal, M. S. Hanif, and U. M. Al-Saggaf, "ResNet based deep features and random forest classifier for diabetic retinopathy detection," *Sensors*, vol. 21, no. 11, p. 3883, Jun. 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/11/3883

[9] M. U. Akram, S. Khalid, and S. A. Khan, "Identification and classification of microaneurysms for early detection of diabetic retinopathy," *Pattern Recognit.*, vol. 46, no. 1, pp. 107–116, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S003132031200297X

[10] M. U. Akram, S. Khalid, A. Tariq, S. A. Khan, and F. Azam, "Detection and classification of retinal lesions for grading of diabetic retinopathy," *Comput. Biol. Med.*, vol. 45, pp. 161–171, Feb. 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S003132031200297X

[11] S. Roychowdhury, D. D. Koozekanani, and K. K. Parhi, "DREAM: Diabetic retinopathy analysis using machine learning," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 5, pp. 1717–1728, Sep. 2014.

[12] M. R. K. Mookiah, U. R. Acharya, R. J. Martis, C. K. Chua, C. M. Lim, E. Y. K. Ng, and A. Laude, "Evolutionary algorithm based classifier parameter tuning for automatic diabetic retinopathy grading: A hybrid feature extraction approach," *Knowl.-Based Syst.*, vol. 39, pp. 9–22, Feb. 2013.

[13] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional neural networks for diabetic retinopathy," *Proc. Comput. Sci.*, vol. 90, pp. 200–205, Jan. 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050916311929

[14] G. García, J. Gallardo, A. Mauricio, J. López, and C. Del Carpio, "Detection of diabetic retinopathy based on a convolutional neural network using retinal fundus images," in *Artificial Neural Networks and Machine Learning–(ICANN)*, A. Lintas, S. Rovetta, P. F. Verschure, and A. E. Villa, Eds. Cham, Switzerland: Springer, 2017, pp. 635–642.
[15] X. Wang, Y. Lu, Y. Wang, and W.-B. Chen, "Diabetic retinopathy stage classification using convolutional neural networks," in *Proc. IEEE Int. Conf. Inf. Reuse Integr. (IRI)*, Jul. 2018, pp. 465–471.
[16] M. T. Esfahani, M. Ghaderi, and R. Kafiyeh, "Classification of diabetic and normal fundus images using new deep learning method," *Leonardo Electron. J. Pract. Technol.*, vol. 17, no. 32, pp. 233–248, 2018.
[17] S. Dutta, B. C. Manideep, S. M. Basha, R. D. Caytiles, and N. C. S. N. Iyengar, "Classification of diabetic retinopathy images by using deep learning models," *Int. J. Grid Distrib. Comput.*, vol. 11, no. 1, pp. 89–106, Jan. 2018.
[18] C. Lam, D. Yi, M. Guo, and T. Lindsey, "Automated detection of diabetic retinopathy using deep learning," *AMIA Summits Transl. Sci.*, vol. 2018, no. 1, p. 147, 2018.
[19] C. Lian, Y. Liang, R. Kang, and Y. Xiang, "Deep convolutional neural networks for diabetic retinopathy classification," in *Proc. 2nd Int. Conf. Adv. Image Process. (ICAIP)*, New York, NY, USA, 2018, pp. 68–72, doi: 10.1145/3239576.3239589.
[20] M. Shaban, Z. Ogur, A. Shalaby, A. Mahmoud, M. Ghazal, H. Sandhu, H. Kaplan, and A. El-Baz, "Automated staging of diabetic retinopathy using a 2D convolutional neural network," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol. (ISSPIT)*, Dec. 2018, pp. 354–358.
[21] H. Jiang, K. Yang, M. Gao, D. Zhang, H. Ma, and W. Qian, "An interpretable ensemble deep learning model for diabetic retinopathy disease classification," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 2045–2048.
[22] W. Zhang, J. Zhong, S. Yang, Z. Gao, J. Hu, Y. Chen, and Z. Yi, "Automated identification and grading system of diabetic retinopathy using deep neural networks," *Knowl.-Based Syst.*, vol. 175, pp. 12–25, Jul. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705119301303
[23] B. Harangi, J. Toth, A. Baran, and A. Hajdu, "Automatic screening of fundus images using a combination of convolutional neural network and hand-crafted features," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 2699–2702.
[24] Y.-P. Liu, Z. Li, C. Xu, J. Li, and R. Liang, "Referable diabetic retinopathy identification from eye fundus images with weighted path for convolutional neural network," *Artif. Intell. Med.*, vol. 99, Aug. 2019, Art. no. 101694. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365718307747
[25] A. Jain, A. Jalui, J. Jasani, Y. Lahoti, and R. Karani, "Deep learning for detection and severity classification of diabetic retinopathy," in *Proc. 1st Int. Conf. Innov. Inf. Commun. Technol. (ICIICT)*, Apr. 2019, pp. 1–6.
[26] S. Qummar, F. G. Khan, S. Shah, A. Khan, S. Shamshirband, Z. U. Rehman, I. Ahmed Khan, and W. Jadoon, "A deep learning ensemble approach for diabetic retinopathy detection," *IEEE Access*, vol. 7, pp. 150530–150539, 2019.
[27] J. de la Torre, A. Valls, and D. Puig, "A deep learning interpretable classifier for diabetic retinopathy disease grading," *Neurocomputing*, vol. 396, pp. 465–476, Jul. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231219304539
[28] S. Majumder and N. Kehtarnavaz, "Multitasking deep learning model for detection of five stages of diabetic retinopathy," *IEEE Access*, vol. 9, pp. 123220–123230, 2021.
[29] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2016, *arXiv:1608.06993*.
[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[31] P. Napoletano, F. Piccoli, and R. Schettini, "Anomaly detection in nanofibrous materials by CNN-based self-similarity," *Sensors*, vol. 18, no. 2, p. 209, Jan. 2018.
[32] R. E. Schapire, "The strength of weak learnability," *Mach. Learn.*, vol. 5, no. 2, pp. 197–227, Jun. 1990.
[33] P. Wu and H. Zhao, "Some analysis and research of the adaboost algorithm," in *Proc. Int. Conf. Intell. Comput. Inf. Sci.* Berlin, Germany: Springer, 2011, pp. 1–5.
[34] A. Roychowdhury and S. Banerjee, "Random forests in the classification of diabetic retinopathy retinal images," in *Advanced Computational and Communication Paradigms*. Singapore: Springer, 2018, pp. 168–176.
[35] E. Decencière, X. Zhang, G. Cazuguel, and B. Lay, "Feedback on a publicly distributed image database: The Messidor database," *Image Anal. Stereol.*, vol. 33, no. 3, pp. 231–234, 2014.
[36] M. D. Abrámoff, J. C. Folk, D. P. Han, J. D. Walker, D. F. Williams, S. R. Russell, P. Massin, B. Cochener, P. Gain, L. Tang, M. Lamard, D. C. Moga, G. Quellec, and M. Niemeijer, "Automated analysis of retinal images for detection of referable diabetic retinopathy," *JAMA Ophthalmol.*, vol. 131, no. 3, pp. 351–357, Mar. 2013.
[37] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manage.*, vol. 45, no. 4, pp. 427–437, Jul. 2009.

HAMZA MUSTAFA received the B.S. degree in information technology from the University of the Punjab and the M.S. degree in computer science from the University of Management and Technology (UMT), Lahore, Pakistan. He is a Graduate Research Scholar with the School of Systems and Technology, UMT. He is currently working as a Software Engineer at a private firm. His research interests include machine learning, computer vision, and deep learning.

SYED FAROOQ ALI received the M.S. degree (Hons.) in CS from LUMS, Lahore, Pakistan and the Ph.D. degree in CS from UMT, Pakistan. He did his Ph.D. course work, the Ph.D. Comprehensive exam and the M.S. degree in CS from Ohio State University, Columbus, USA. He received the LUMS Fellowship for M.S. degree. He is currently working as an Assistant Professor with UMT. His research interests include computer vision, digital image processing, and medical imaging. He is a reviewer for various IEEE conferences and journals.

MUHAMMAD BILAL is an Educator, a Researcher, and a Maker. He was a Postdoctoral Researcher at KAIST, South Korea. He is an Associate Professor with the Department of Electrical and Computer Engineering, KAU. His research interests include digital image/signal processing, machine learning/AI, digital/analog circuit design, embedded systems, and robotics.

MUHAMMAD SHEHZAD HANIF received the B.Sc. degree in electrical engineering from the University of Engineering and Technology, Lahore, Pakistan, in 2001, and the M.S. degree in engineering sciences and the Ph.D. degree in computer engineering from the Université Pierre et Marie Curie, France, in 2006 and 2009, respectively. He is currently an Assistant Professor with the Department of Electrical and Computer, King Abdulaziz University, Jeddah, Saudi Arabia. His research interests include machine learning, image analysis, information fusion, and object detection and tracking.