IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 19, 2021, accepted March 2, 2021, date of publication March 10, 2021, date of current version March 19, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3065273

Automatic Diabetic Retinopathy Diagnosis Using Adaptive Fine-Tuned Convolutional Neural Network

FAHMAN SAEED, MUHAMMAD HUSSAIN, (Senior Member, IEEE), AND HATIM A. ABOALSAMH, (Senior Member, IEEE)
Department of Computer Science, King Saud University, Riyadh 11451, Saudi Arabia

Corresponding author: Muhammad Hussain (mhussain@ksu.edu.sa)

This work was supported by the Deanship of Scientific Research, King Saud University, Riyadh, Saudi Arabia, through research group under Grant RGP-1439-067.

ABSTRACT Diabetic retinopathy (DR) is a complication of diabetes that leads to blindness. The manual screening of color fundus images to detect DR at early stages is expensive and time consuming. Deep learning (DL) techniques have been employed for automatic DR screening on fundus images due to their outstanding performance in many applications. However, training a DL model needs a huge amount of data, which are usually unavailable in the case of DR, and overfitting is unavoidable. Employing a two-stage transfer learning method, we developed herein an intelligent computer-aided system using a pre-trained convolutional neural network (CNN) for automatic DR screening on fundus images. A CNN model learns the domain-specific hierarchy of low- to high-level features. Given this, using the regions of interest (ROIs) of lesions extracted from the annotated fundus images, the first layer of a pre-trained CNN model is re-initialized. The model is then fine-tuned, such that the low-level layers learn the local structures of the lesion and normal regions. As the fully connected layer (FC) layers encode high-level features, which are global in nature and domain specific, we replace them with a new FC layer based on the principal component analysis PCA and use it in an unsupervised manner to extract discriminate features from the fundus images. This step reduces the model complexity, significantly avoiding the overfitting problem. This step also lets the model adopt the fundus image structures, making it suitable for DR feature detection. Finally, we add a gradient boosting-based classification layer. The evaluation of the proposed system using a 10-fold cross-validation on two challenging datasets (i.e., EyePACS and Messidor) indicates that it outperforms state-of-the-art methods. It will be useful for the initial screening of DR patients and will help graders in deciding quickly as regards patient referral to an ophthalmologist for further diagnosis and treatment.

INDEX TERMS Fundus images, diabetic retinopathy, classification, CNN.

I. INTRODUCTION

Diabetes is one of the main health dilemmas worldwide. One complication of diabetes is diabetic retinopathy (DR), which is one of the main causes of blindness [1]. It has different levels of severity [2] and can be controlled if detected at early stages. DR affects the retina, which is responsible for the conversion of light to the electric signal interpreted to create an image. The retina contains a network of blood vessels that provide nutrition to the retina. Diabetes damages the blood vessels; consequently, the retina does not receive blood supply. This affects the health of the retina and ultimately distorts the eyesight of an individual. The earliest stage of DR is referred to as background retinopathy. At this stage, diabetes does not affect the sight, but impairs the blood vessels. The vessels may slightly bulge (microaneurysms - MAs), leak fluid and proteins (exudates - EXs), and leak blood (retinal hemorrhages - HEs), as shown in Figure 1. At a later stage, DR becomes proliferative retinopathy and harms the retina more extensively than background retinopathy. It also increases the chances of vision loss and creates a threat of blindness because most of the retina is starved of proper blood supply. Ophthalmologists or expert graders detect DR manually, which is expensive. Routinely screening

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

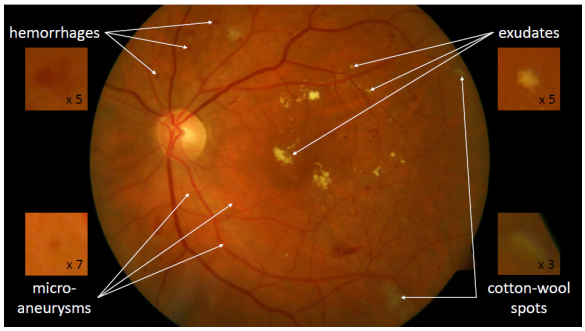

FIGURE 1. Type of lesion in a DR fundus image [1].

a large number of diabetes patients for possible DR prevalence puts a heavy burden on ophthalmologists or expert graders, affecting their efficiency and delaying DR diagnosis and treatment. The diagnosis is also subjective, and the findings of different graders vary [3]–[5]. Given this, there is a great need for an intelligent automatic system for screening patients for DR prevalence, which can help graders to detect DR with confidence at an early stage and refer a patient to an ophthalmologist for further diagnosis and treatment.

Many computer-aided systems have been developed for DR detection and diagnosis in the recent decade. Some of them are computer-aided detection systems (CADe) [1], [6]–[9] that work at the pixel level to detect and segment lesions, while others are computer-aided diagnosis (CADx) systems [1], [10], [11], which work at the image level to detect DR. CADe and CADx systems help reduce the overhead of graders in deciding whether to refer patients to an ophthalmologist. Many methods have been developed for CADx systems using machine [11] and deep learning (DL) [12]. These methods use different techniques for DR classification, which can be broadly grouped into two main categories: hand-engineered features based techniques and DL based methods.

Different hand-engineered techniques have been proposed to extract features from fundus images for their grading. Seoud *et al.* [8] used contrast, Pires *et al.* [10] employed SURF features, Sreejini and Govindan [3] used BoVW based on K-means, local features (IFT, LBP, and LDP), and color features, and Adal *et al.* [13] employed Laplacian of Gaussian (LOG) to extract features from fundus images. The hand-engineered features are not directly learned from the data. As such, this kind of features are not tuned to the lesion structures in fundus images or do not generalize well. In addition, their design involves laborious and exhaustive preprocessing and parameter tuning.

At present, deep learning has been employed for DR diagnosis from fundus images and it has shown promising results. Some authors such as Pratt *et al.* [14], Colas *et al.* [15], Quellec *et al.* [1], Islam *et al.* [16] and Chen *et al.* [17] developed deep models for DR grading. A commonly used DL architecture is convolutional neural network (CNN), it involves millions of learnable parameters and its training needs a huge amount of data, which is usually not available in case of DR, and the overfitting problem is inevitable. To overcome this issue, some authors like Wan *et al.* [18], Gao *et al.* [19] employed fine-tuning and pre-trained CNN models (AlexNet, VggNet, GoogleNet, and ResNet, inception and Inception-V3) for DR grading. The best state-of-the-art pre-trained CNN models like VGG [20] and ResNet [21], DPN [22] are usually trained on ImageNet [23], a big dataset of natural images, and encode the domain-specific hierarchy of low- to high-level features. Natural images have different structures compared to retinal fundus images; thus, fine-tuning these models is essential for their adaptation to the fundus image structures.

Though, the deep learning-based methods perform better than those based on handcrafted features, they do not give as good performance as is expected due to the following reasons: (i) a huge number of DR fundus is not available to train a deep CNN model; hence, overfitting is unescapable; and (ii) a brute force approach to fine-tune a pre-trained model after replacing the classification layer does not properly learn the discriminative structures from the retinal fundus images. The way a model is fine-tuned using a limited number of fundus images has significant impact on its performance. Most of the transfer learning techniques employ fundus images for fine-tuning pre-trained CNN models. As the DR grading depends on the presence of lesions like MAs, EXs, and HMs in fundus images, we propose an effective two-stage method for fine-tuning a pre-trained CNN model for DR grading using lesion ROIs and fundus images. The method takes a retinal fundus image as input, processes it with the fine-tuned model and grades it into normal or DR levels. The main contributions of the proposed work are as follows.

- We introduced an intelligent computer-aided diagnosis system based on DL for the DR grading of retinal fundus images, which does not need any preprocessing technique to preprocess the retinal fundus images.
- We proposed a two-stage fine-tuning method to adapt a pre-trained model to retinal fundus images: in the first stage, it embeds the DR lesion structures in a pre-trained CNN model using lesion ROIs, and in the second stage, it adapts high-level layers to extract the discriminate structures of the retinal fundus images by removing the domain-specific fully connected layer (FC) layers of a pre-trained model and introducing a new PCA layer, which significantly reduces the model complexity and helps overcome the overfitting problem; this also overcomes the limitations of DL model learning due to the small amount of available data for DR detection.
- We validated the proposed method fine-tuned using ROIs on two benchmark public domain challenging datasets: EyePACS [24] and Messidor [25].

This remainder of this paper is structured as follows: Section 1 intoduction; Section 2 describes the proposed method; Section 3 presents experiments result and discussion; and Section 4 concludes the paper.

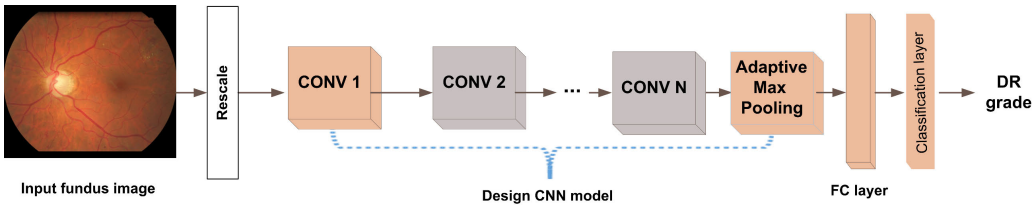

FIGURE 2. Overview of the proposed method.

II. PROPOSED METHOD

A. PROBLEM FORMULATION AND MOTIVATION

Let $R^{M\times N\times 3}$ be the space of the color retinal fundus images with a resolution of $M \times N$ and Y be the space of labels, such that $Y = \{1, 2, \ldots, c\}$, where c is the number of classes. The problem of predicting the DR grade of the retinal fundus image is related to building a mapping $\phi : R^{M\times N\times 3} \rightarrow Y$ that maps a fundus image $I \in R^{M\times N\times 3}$ to a $y \in Y$ (i.e., $\phi(I) = y$). We employed a deep convolutional neural network (CNN) model to build the mapping ϕ. Figure 2 presents the overall structure of the model that defines this mapping. A CNN model represents the hierarchical structure of images and has shown amazing results in many applications [26]. It has also been used for DR diagnosis [1], [19], [27]. However, the achieved performance was not as expected mainly due to the CNN model involving a huge number of parameters and needing a big amount of data for its training. This much data is not available in the case of DR diagnosis. The overfitting problem cannot be avoided when a CNN model is learned from scratch. Transfer learning was employed to overcome this problem [18], [19], [27], but the performance was unsatisfactory because the hierarchical structure of a CNN model was not properly considered. We introduced a two-stage transfer learning herein to develop a CNN model for modeling the mapping ϕ. State-of-the-art pre-trained models, such as VGGNet, ResNet, and DPN, trained on ImageNet dataset are usually employed for transfer learning. ImageNet consists of natural images, and the internal structure of natural images is entirely different from that of medical images like retinal fundus images. In other words, just fine-tuning a pre-trained model using fundus images will not work due to the significant differences in the domain structures. The hierarchy of features learned by a pre-trained CNN model must be considered. Low-level layers learn low level features, while high-level layers, particularly FC layers, encode domain-specific high-level features. In view of this, using the ROIs extracted from the fundus images, we first fine-tuned the low-level layers and replaced high-level FC layers with a new layer learned in an unsupervised manner. We then modeled the mapping ϕ using a pre-trained CNN model based on this idea (Figure 2). It took fundus images as input and gave a normal or DR grade as the output. It consisted of CONV layers fine-tuned using the ROIs extracted from the fundus images, an FC layer built using PCA, and a classifier layer. The detail of the model design is given in the subsections that follow.

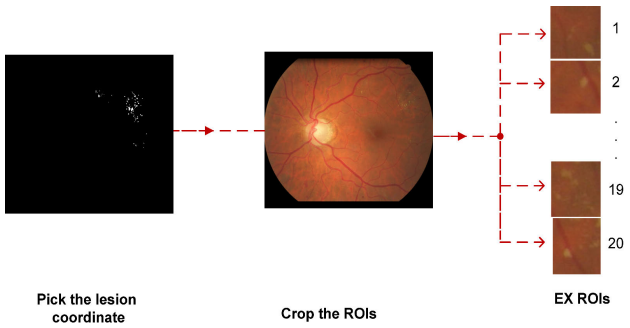

FIGURE 3. Extraction process of the lesion ROIs from the EX fundus image in E_optha (image-C0010492-20).

B. DATASET PREPARATION

We used the lesion ROIs extracted from the fundus images to fine-tune the low-level layers. For this purpose, we used public domain benchmark DR datasets. We employed E-optha [28] for the lesion ROI extraction because it contains a pixel-level lesion annotation. We evaluated the model by employing EyePACS [24] and Messidor [25] because both have image level annotations for DR grading.

E-optha is designed for scientific research on DR diagnosis. It contains two subsets of color fundus images for EX and MA. Each image is manually annotated by three expert ophthalmologists. EXs are contained in 47 images, whereas 148 images contain MAs and small HEs. Using the annotations, we extracted ROIs around each lesion to capture its structure. We then collected 695 MA ROIs and EX ROIs each from the fundus images and used the same number for each type of lesion to keep the data balanced for fine-tuning. We randomly collected the same number of normal ROIs from normal fundus images. The collected ROI dataset had three classes: EX, MA, and normal. Figure 3 illustrates examples of the lesion ROIs.

These ROIs were not enough for fine-tuning. To avoid overfitting, we augmented the training examples such that the internal structures of the ROIs were reflective of the true structures of the fundus images. We employed two data augmentation methods. First, we extracted the ROIs around the lesions with different sizes (i.e., 16×16, 32×32, and 64×64) such that they contain different context information. We then resized them to the same size (i.e., 64×64). Second, we rotated each ROI in four directions [40 °,20 °,180 °,275 °] and flipped it horizontally. Figure 3 shows the examples.

C. DESIGN AND TRAINING OF A CNN MODEL

Training a CNN model from scratch is time consuming and needs a very large dataset, which is unavailable. We introduced a two-stage transfer learning and designed a CNN model using a pre-trained CNN model. We tried three state-of-the-art models, namely VGG19, ResNet152, and DPN107, pre-trained on the ImageNet dataset [29]. The VGGNet achieved an excellent accuracy on ImageNet and in other domains [23], [30]. The ResNet architecture also exhibited excellent results on ImageNet 2015 [21], [31]. The dual path network (DPN) is based on the philosophy of ResNet and explores new features through dual path architectures [22]. Our transfer learning method is different from those employed in [18], [19], and [27] in the sense that we considered the hierarchy of the features learned by a CNN model and fine-tuned and modified it in two stages to adapt it for fundus images. The low-level layers, particularly CONV layers, of a pre-trained CNN model encoded low level and local features that do not heavily depend on domain. We fine-tuned the low-level layers using extracted ROIs in the first stage (Figure 4a). The higher-level layers, particularly the FC layers of a trained model, encoded high-level global features that heavily depend on the domain. These layers also involved a huge number of learnable parameters, which makes the model very complex and lead to an overfitting problem. We removed the FC layers and added a new FC layer using the PCA technique in the second stage (Figure 4b).

1) CONV LAYER TUNING

Using the extracted ROIs, we first re-initialized the filters of the first CONV layer and fine-tuned the model to adapt the CONV layers of a pre-trained model to the structures of the retinal fundus images.

a: RE-INITIALIZATION OF THE WEIGHTS OF FIRST CONV LAYER

Fundus images capture the anatomical features of human eye, and its intrinsic characteristics are entirely different from those of natural images. A pre-trained model is usually learned on ImageNet dataset, which consists of natural images with well-defined micro-structures. As such the CONV layer filters of a pre-trained model are adopted to micro-structures of natural images and are not appropriate for extracting features specific to different lesion structures in fundus images. The filters of the first CONV layer can be directly re-initialized using the lesion ROIs extracted from fundus images. We introduce an algorithm, based on the PCA and the lesion ROIs of MAs and EXs, to re-initialize the filters of the first CONV layer to tune them to the lesion structures. For the re-initialization, we used only the lesion ROIs to emphasize the lesions during DR grading. Algorithm 1 presents the re-initialization details. Reinitializing the filters of the first layer of a pre-trained CNN model has a significant effect on the system performance, as is demonstrated in the Results section.

Algorithm 1 Re-Initialize the Filters of the First CONV Layer Using PCA

Input: Lesion ROIs (MA and EX) and the filters of the first CONV layer of a pre-trained CNN model.
Output: Re-initialized filters of the first CONV layer.

Step 1: Rescale all ROIs to 64×64 and flatten them to $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}$, where n is the number of lesion ROIs, and $\boldsymbol{x_i} \in \boldsymbol{R^d}$, $d = 3 \times 64 \times 64$.

Step 2: Compute the mean m of $\boldsymbol{x_i}$, $i = 1, 2, \ldots, n$, translate them to $\phi_i = \boldsymbol{x_i} - \boldsymbol{m}$, $i = 1, 2, \ldots, n$ and form $A = [\phi_1\ \phi_2\ \ldots\ \phi_n]$, where each ϕ_i is a column vector.

Step 3: Compute the covariance matrix $C = AA^T$.

Step 4: Compute the eigenvalues and eigenvectors $\lambda_j \in \boldsymbol{R}$, *and* $\boldsymbol{u_j} \in \boldsymbol{R^d}$, $j = 1, 2, \ldots, d$ of the covariance matrix C and select K eigenvectors $\boldsymbol{u_k}$, $k = 1, 2, \ldots, K$ corresponding to the largest K eigenvalues, assuming that K filters exist in the first CONV layer. In the case of VGGNet, ResNet, and DPN, K = 64.

Step 5: The dimension of each $\boldsymbol{u_j}$ is d $= 3 \times 64 \times 64$, but the size of each filter in the first CONV layer is $dr = a \times b \times c$, where dr is much smaller than d. For example, in the case of VGGNet, it is $(3 \times 3 \times 3)$, and for ResNet and DPN, it is $(3 \times 7 \times 7)$. To project each $\boldsymbol{u_k} \in \boldsymbol{R^d}$, $k = 1, 2, \ldots, K$ to $\boldsymbol{f_k} \in \boldsymbol{R^{dr}}$, $k = 1, 2, \ldots, K$, repeat steps 2–4 for $\boldsymbol{u_k} \in \boldsymbol{R^d}$, $k = 1, 2, \ldots, K$ and compute the transformation matrix M $= [\boldsymbol{v_1} \ldots \boldsymbol{v_{dr}}]$, where $\boldsymbol{v_j} \in \boldsymbol{R^d}$ denotes the eigenvectors of the respective covariance matrix corresponding to the largest dr eigenvalues.

Step 6: Project each $\boldsymbol{u_k} \in \boldsymbol{R^d}$, $k = 1, 2, \ldots, K$ to $\boldsymbol{f_k} \in \boldsymbol{R^{dr}}$, $k = 1, 2, \ldots, K$, where $\boldsymbol{f_k} = \boldsymbol{M^T}(\boldsymbol{u_k} - \boldsymbol{u})$, and $\boldsymbol{u}$ is the mean of vectors $\boldsymbol{u_k} \in \boldsymbol{R^d}$, $k = 1, 2, \ldots, K$.

Step 7: Reshape each $\boldsymbol{f_k} \in \boldsymbol{R^{dr}}$, $k = 1, 2, \ldots, K$ to the 3D filter of size $dr = a \times b \times c$. As an example, for ResNet and DPN models, $dr = 147$, and it is reshaped to $(3 \times 7 \times 7)$.

b: CNN MODEL FINE-TUNING

After the first CONV layer re-initialization, the pre-trained model was fine-tuned using the extracted ROIs to adapt it to the lesion and normal structures of the retinal fundus images. The ROIs belonged to three classes (i.e., MAs, EXs, and normal); thus, before the model fine-tuning, we replaced the last FC layer (i.e., the classification layer with 1000 neurons based on ImageNet) with a new classification layer having three neurons (Figure 4a).

Almost all pre-trained models (e.g., VGGNet, ResNet, and DPN) take a fixed-size image (e.g. 224×224) as input. This constraint leads to a problem due to two reasons: (1) the maximum size of the extracted ROIs is 64×64; and (2) the size of the fundus images to be graded is about 3800×2600 pixels in the EyePACS dataset and 2240×1488 in the Messidor dataset. Most of the state-of-the-art methods resize fundus images to 224×224 [21], 227×227 [32], 299×299 [33], or 320×240 [34]. Rescaling up the ROIs and rescaling down

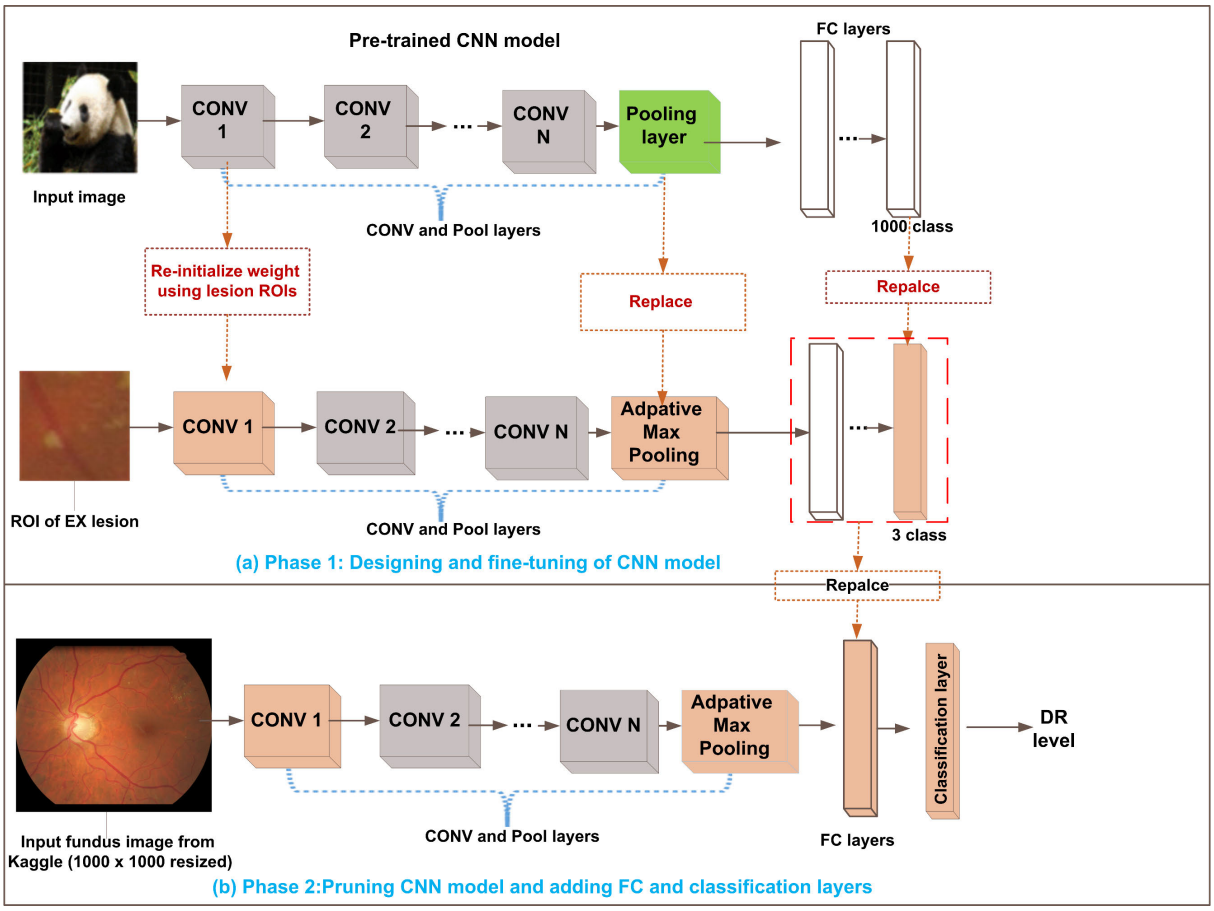

FIGURE 4. Architecture of the proposed system for DR grading.

the fundus images to 224 × 224 are problematic because most of the discriminative structures are destroyed in the fundus images, and extraneous structures are introduced in the ROIs. To overcome this problem, we introduce an adaptive max pooling layer before the FC layers. It adds to the model an ability to predict the label of fundus image or ROI if any size. It takes input of any spatial dimensions $W \times H \times D$ from last CONV layer, reduces it to a fixed output of dimension $1 \times 1 \times D$. It reduces the number of parameters and helps to overcome the problem of overfitting. The number of output features is equal to the number of input planes [35]. After adding this layer, we fine-tuned the pre-trained model using 64 × 64 ROIs. To do so, the ROIs were divided into three subsets: 80% for training, 10% for validation, and 10% for testing. For fine-tuning, we used the cross-entropy loss, Adam optimizer with 0.001 learning rate, 3000 batch size, and 21 epochs. The question was whether the CONV layers learn the lesion structures in the fine-tuned CNN model (*FTCNN*) and whether their output can be interpreted for referral. To answer this question, we created heat maps of the feature maps of the CONV layers using the gradient-weighted class activation mapping (GradCam) visualization method [36] developed based on class activation mapping. method [36] developed based on class activation mapping. Figure 5 shows some heat maps of the CONV layers of ResNet152 with/without the CONV1 layer re-initialization. The maps in Figure 5(a) indicate that an MA lesion structure is hierarchically encoded by the CONV layers and becomes pronounced to graders for diagnosis in the last CONV layer. For the case of a re-initialized CONV1, it is more localized in the last CONV layer. Figure 5(b) depicts the same effect of the last CONV layer for the EX lesion, while Figure 5(c) shows the heat map of the last CONV layer when the input is a fundus image. The lesion structure is depicted in the heat map, which indicates that after fine-tuning the pre-trained model, it learned the lesion structures in the fundus images. The fine-tuned model successfully highlighted the lesion regions in the retinal fundus images, which is more important for clinical examinations and graders for diagnosis.

2) PRUNING AND ADDITION OF FC AND CLASSIFICATION LAYERS

The FC layers of a pre-trained model are learned from the ImageNet dataset and encode the global higher-level features, which are not relevant to the normal and lesion features in the fundus images. Moreover, the FC layers contain a huge number of learnable parameters that is much larger

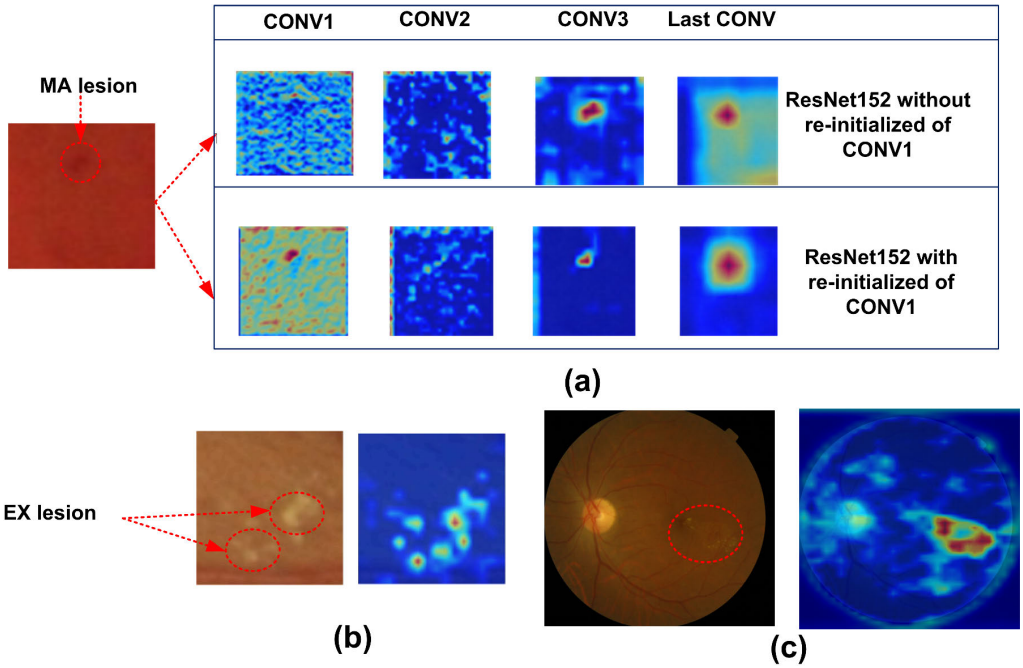

FIGURE 5. Visualizations of activation maps using the GradCam method: (a) visualizations of an MA lesion encoded by the activation maps of the CONV layers of ResNet152 with and without CONV1 re-initialization; (b) visualization of the EX lesion encoded by an activation map of the last CONV layer of ResNet152 with a re-initialized CONV1; and (c) visualization of the lesion structure in a fundus image encoded by an activation map of the last CONV layer of ResNet152 with a re-initialized CONV1.

than the learnable parameters of CONV layers. For example, in VGG-19, the CONV layers contain 20,024,384 weights and biases, and the FC layers have 123,642,856 learnable parameters. As explained in the previous section, fine-tuning a pre-trained model leads to an overfitting problem and an increased time complexity due to the small number of ROIs and the huge number of learnable parameters of the FC layers. We removed all FC layers from the fine-tuned model and added a new FC layer, called the PCA layer, with 153 neurons to reduce the model complexity. The 153 neurons in this layer were selected using the greedy algorithm based on ROIs and the fine-tuned ResNet152 (Table 1). The PCA layer was learned in an unsupervised manner by applying PCA on the activations of the last CONV layer. It was used to extract features from the fundus images (not ROIs, Figure 4b). Algorithm 2 presents how the PCA layer constructed. Removing the FC layers significantly reduced the total number of learnable parameters (e.g., 86% reduction in the pre-trained VGG19). The PCA layer extracted the features relevant to the normal and lesion structures in the fundus images. We classified the extracted features further by adding a classification layer, which predicts whether the input fundus image belongs to a normal person or a DR patient with its severity level. For the classification layer, we considered three tree-based classifiers, namely decision tree (DT) [37], random forests (RF) [38], and gradient boosting (GB) [39], based on their better performances for medical imaging [40]–[42].

TABLE 1. Selecting the number of PCA using fine-tuned ResNet152 and ROI dataset.

Number of PCA	ACC
10	98% ± 0.003
50	98.1 % ± 0.005
153	99.20 %± 0.001
183	99.17 % ± 0.002
250	99.14 % ± 0.002
500	99.10 % ± 0.002

The size of the fundus images in EyePACS is near 3800 × 2600 px on average, while that in Messidor is near 2240 × 1488. Although our designed model can take an input fundus image of any size, handling the big size of images is difficult. It requires an excessively large memory size and involves a high computational overhead. Most state-of-the-art methods resize the fundus images into small sizes to overcome this problem: 224 × 224 [21] 227 × 227 [32] 299 × 299 [33], and 320 × 240 [34]. These sizes smooth out the lesion areas, which are usually small (e.g., in the case of MAs and HEs). We resized the fundus images to 1000 × 1000 and extracted the features vectors using the pruned model, *UCNN*, to obtain a compromise between performance and detection accuracy.

III. EXPERIMENTS

A. EVALUATION DATASET

To evaluate the proposed system, we extracted features from the benchmark challenge databases EyePACS [24] and Messidor [43] and classified it using the *ACNN* model. EyePACS contained high-resolution fundus images captured under various conditions. The images were categorized into five DR classes: 0 (no DR), 1 (mild), 2(moderate), 3 (severe), and 4 (proliferative). EyePACS has 88,702 color retinal fundus photographs from 44,351 subjects, and 35,126 images from 17,563 patients were labeled and used for training, while 53,576 images from 26,788 patients were used for testing. The size of the images is a neat 3800 × 2600 px. We used all the labeled samples to evaluate our method using a 10-fold cross-validation (CV). Messidor contained 1200 high-resolution color retinal fundus images acquired at three ophthalmology departments. It had annotations for two grading types: DR grades (four classes) and risk of macular edema grades (three classes). The four DR grades were 0 (normal; μA = 0 AND H = 0), 1 ((0 < μA <= 5) AND (H = 0)), 2 ((5 < μA < 15) OR (0 < H < 5)) AND (NV = 0), and 3 ((μA >= 15) OR (H >= 5) OR (NV = 1)), where μA is the number of MAs; H is the number of HEs; NV = 1 means neovascularization exists; and NV = 0 means no neovascularization exists. We used the DR grades of Messidor for the system evaluation.

B. EVALUATION PROCESS AND METRICS

We evaluated the system using 10-fold CV [44]. The extracted features were divided 10 fold. The model was trained and tested 10 times, taking one fold in turn as the test data and the remaining folds as the training data. We kept the same percentage of subjects of each class in each fold. First, the average performance metrics for each class were calculated over 10 folds, taking the class as positive and all the remaining classes as negative. Finally, the average performance metrics over all classes were computed. Many metrics were used to evaluate the performance of the DR diagnosis systems, such as accuracy (ACC), area under ROC curve (AUC), sensitivity (SE), specificity (SP), precision (PR), recall (RC), F1-score, and kappa [19], [40], [1], [45]. We employed five commonly used metrics (i.e., ACC, SE, SP, AUC, and Kappa) [46], [47] to evaluate the performance of our deep learning-based system.

To determine whether a statistically significant difference exists between any two methods, we used the nonparametric Mann–Whitney–Wilcoxon test (WMW) [48] with a 5% significance level with the null hypothesis method A = method B an alternative hypothesis that method A is better than method B. We used WMW because the data distribution is unknown.

C. MODEL SELECTION

Our CNN-based system has many hyper-parameters: (i) backbone pre-trained CNN model; (ii) CONV1 with or without re-initialization; and (iii) method for the classification layer. In the sequel, we provide an empirical analysis of these hyper-parameters and select the best choices. All models were implemented and fine-tuned using PyTorch [3] on a system equipped with processor Intel® Core i9-7900X CPU at 3.3 GHz, 64 GB RAM, and NVIDIA GeForce GTX 1080 Ti.

Algorithm 2 Adaptation of the Fine-Tuned CNN Model and Extract Features From Fundus Images

Input: ROI lesions from the E-optha dataset $I_1, I_2, \ldots, I_n$ and fundus images $f_1, f_2, \ldots, f_n$ (from EyePACS or Messidor) and the fine-tuned CNN model (*FTCNN*).
Output: Adapted the CNN model (*ACNN*).

Step 1: Load the *FTCNN* model and remove the FC layers to obtain a pruned CNN model (PCNN).
Step 2: Resize the ROI images $I_1, I_2, \ldots, I_n$ to $I'_1, I'_2, \ldots, I'_n$ to 64 × 64 px each.
Step 3: Compute the activations $a_1, a_2, \ldots, a_n$ of $I'_1, I'_2, \ldots, I'_n$, such that $a_i = PCNN\left(I'_i\right)$, $i = 1, 2, \ldots, n$.
Step 4: Flatten the activations $a_1, a_2, \ldots, a_n$ to vectors $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}; \boldsymbol{x_i} \in \boldsymbol{R^d}$.
Step 5: Add a new FC layer with L neurons.
Step 6: Using vectors $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}$ and PCA, learn the weights $\boldsymbol{W}$ and biases $\boldsymbol{b}$ of the neuron, such that $\boldsymbol{W} = \boldsymbol{M^T}$ and $\boldsymbol{b} = -\boldsymbol{M^T}\mu$ where, $M = [\boldsymbol{u_1} \ldots \boldsymbol{u_L}], \boldsymbol{u_i} \in \boldsymbol{R^d}$, $i = 1, 2, \ldots, L$ are the eigenvectors of the covariance matrix computed from the vectors $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}$ corresponding to the largest L eigenvalues (i.e., the best L is 153 is shown in Table 1), and μ is the mean vector.
Step 7: The updated CNN (UCNN) model is obtained after the FC layer addition.
Step 8: Add the classification layer to the UCNN to obtain the ACNN model.
Step 9: Resize the fundus images $f_1, f_2, \ldots, f_n$ to $f'_1, f'_2, \ldots, f'_n$ to 1000 × 1000 px each.
Step 10: Compute the activations $a'_1, a'_2, \ldots, a'_n$ of $f'_1, f'_2, \ldots, f'_n$, such that $a'_i = UCNN(f'_i)$, $i = 1, 2, \ldots, n$, and predict it using the classification layer

1) EFFECT OF THE METHOD FOR THE CLASSIFICATION LAYER

We considered three tree-based classifiers for the classification layer: DT, RF, and GB. We used the pre-trained ResNet152 model as the backbone CNN model with a re-initialized CONV1 layer to test the effect of these methods. Figure 6 depicts the results of the three methods on the EyePACS dataset. The GB method exhibited the best performance in terms of all performance metrics. GB exceled because it builds trees in such a way that each new tree helps correct the errors made by the previously added trees. GB has much flexibility and can be optimized using different loss functions. We used GB with a maximum depth

TABLE 2. Comparison between classification methods on the EyePACS dataset using the WMW test – ResNet152 with re-initialized CONV1.

Classifier	ACC ± STD	P	SE ± STD	P	SP ± STD	P	KAPPA ± STD	P
GB	99.73 ± 0.0003	-	96.04 ± 0.005	-	99.81 ± 0.0003	-	98.45 ± 0.002	-
RF	99.05 ± 0.001	0.001	83.33 ± 0.014	0.001	99.37 ± 0.0009	0.001	94.46 ± 0.006	0.001
DT	99.52 ± 0.0009	0.001	93.71 ± 0.01	0.005	99.7 ± 0.006	0.001	97.18 ± 0.005	0.003

TABLE 3. Effect of the re-initialized CONV1 on the method with the ResNet152 model and GB (EyePACS dataset).

Model	ACC ± STD	p	SE ± STD	p	SP ± STD	p	KAPPA ± STD	p
Re-initialized	99.73 ± 0.0003	-	96.04 ± 0.005	-	99.81 ± 0.0003	-	98.45 ± 0.002	-
Non-re-initialized	90.18 ± 0.005	0.001	75.43 ± 0.014	0.001	93.87 ± 0.004	0.001	69.26 ± 0.017	0.001

TABLE 4. Effect of the re-initialized CONV1 on the method with the ResNet152 model and GB (messidor dataset).

Model	ACC ± STD	p	SE ± STD	p	SP ± STD	p	KAPPA ± STD	p
Re-initialized	98.88 ± 0.007	-	97.32 ± 0.016	-	99.26 ± 0.004	-	96.67 ± 0.016	-
Non-re-initialized	90.87 ± 0.017	0.001	78.09 ± 0.03	0.001	94.28 ± 0.049	0.001	73.37 ± 0.05	0.001

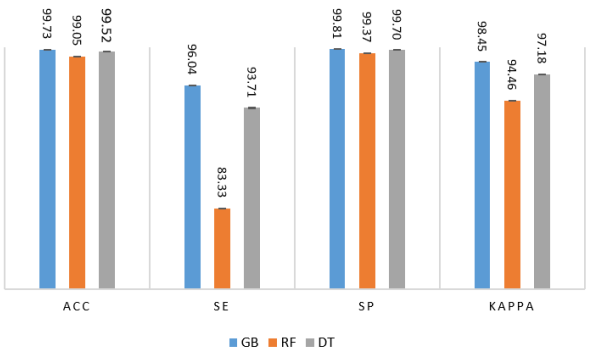

FIGURE 6. Comparison between the classification methods along with STD on the EyePACS dataset when ResNet152 with a re-initialized CONV1 is used as the backbone model.

of 3 and least squares regression as a loss function. We also performed a statistical significance test to check whether a significant difference exists among GB, RF, and DT. The distributions of the outcomes of various performance metrics were unknown; therefore, we used the nonparametric Mann–Whitney–Wilcoxon test (WMW) [48] with an alternative hypothesis that GB was better than RF (or DT). Table 2 shows the results. The p-values are less than 0.05 in both cases; hence, the null hypothesis is rejected, and we accept the alternative hypothesis i.e. the GB is significantly better than RF and DT at a 5% significance level.

2) EFFECT OF CONV1 RE-INITIALIZATION

We used ResNet152 as the backbone model and GB for the classification layer based on the results of the previous subsection to test the effect of the CONV1 re-initialization of a pre-trained backbone CNN model on our CNN-based system. Figure 7 exhibits the results from the EyePACS and Messidor datasets. The method with the re-initialized CONV1 was significantly better than that without in terms of all performance metrics on both datasets.

Tables 3 and 4 present the statistical significance results computed using the WMW test on both datasets. The p-value < 0.05 for all metrics corresponding to both datasets; hence, we rejected the null hypothesis and accepted the alternative that a significant difference exists between the methods with and without re-initialized CONV1 at a 5% significance level. The method with the re-initialized CONV1 performed better because the CONV1 re-initialization made the backbone CNN model directly learn the patterns specific to the DR lesions.

3) EFFECTS OF BACKBONE CNN MODELS

We considered three backbone pre-trained CNN models (i.e., VGG19, ResNet152, and DPN107) based on their superior performance in many applications. We considered the system with re-initialized CONV1 and GB based on the findings of the previous subsections to observe the effects of the pre-trained backbone CNN models. Figure 8 shows the results from the EyePACS and Messidor datasets with the three models. Among the three models, the ResNet152 model gave a better performance in terms of all performance metrics. DPN107 ranked second. Tables 5 and 6 show the WMW test results to verify whether the difference among ResNet152, VGG19, and DPN107 was statically significant. The p-values < 0.05 on both datasets for all metrics, indicating that the ResNet152 model was significantly better than VGG19 and DPN107 at a 5% significance level. The superior performance of ResNet152 was probably caused by the two following reasons: 1) it is deeper than the other two CNN models; and 2) it is based on residual learning, which helps overcome the overfitting problem by properly fine-tuning the low-level layers. Although DPN107 is also based on residual learning,

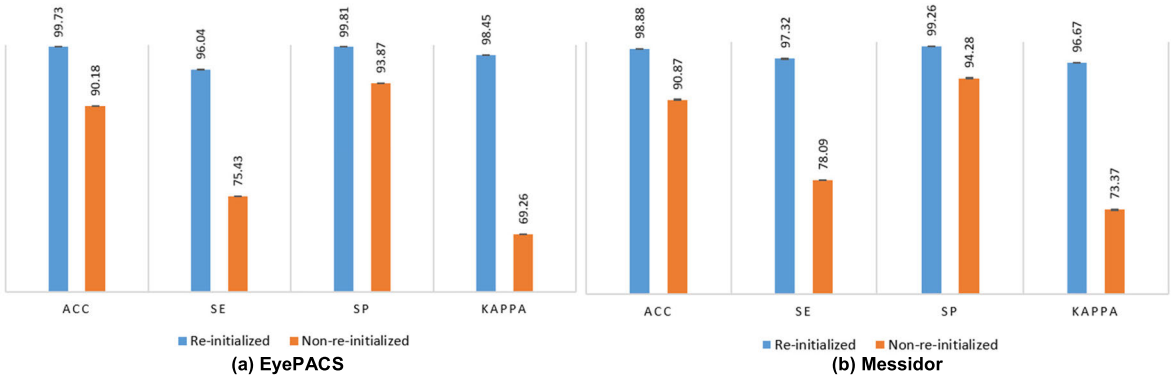

FIGURE 7. The effect of the re-initialization of CONV1 on the performance of the proposed method with ResNet152 and GB along with STD.

TABLE 5. Selection of the CNN model (EyePACS dataset).

Model	ACC ± STD	p	SE ± STD	p	SP ± STD	p	KAPPA ± STD	P
RESNET152	99.73 ± 0.0003	-	96.04 ± 0.005	-	99.81 ± 0.0003	-	98.45 ± 0.002	-
VGG19	95.04 ± 0.006	0.001	87.55 ± 0.015	0.001	96.91 ± 0.004	0.001	84.45 ± 0.018	0.001
DPN107	96.81 ± 0.004	0.001	92.07 ± 0.009	0.001	98.01 ± 0.002	0.001	90.03 ± 0.012	0.001

TABLE 6. Selection of the CNN model (messidor dataset).

Model	ACC ± STD	p	SE ± STD	P	SP ± STD	P	KAPPA ± STD	P
RESNET152	98.88 ± 0.007	-	97.32 ± 0.016	-	99.26 ± 0.004	-	96.67 ± 0.016	-
VGG19	98.21 ± 0.01	0.001	95.69 ± 0.03	0.001	98.87 ± 0.007	0.001	94.78 ± 0.03	0.001
DPN107	98.29 ± 0.005	0.001	96.23 ± 0.02	0.001	98.87 ± 0.004	0.001	95 ± 0.02	0.001

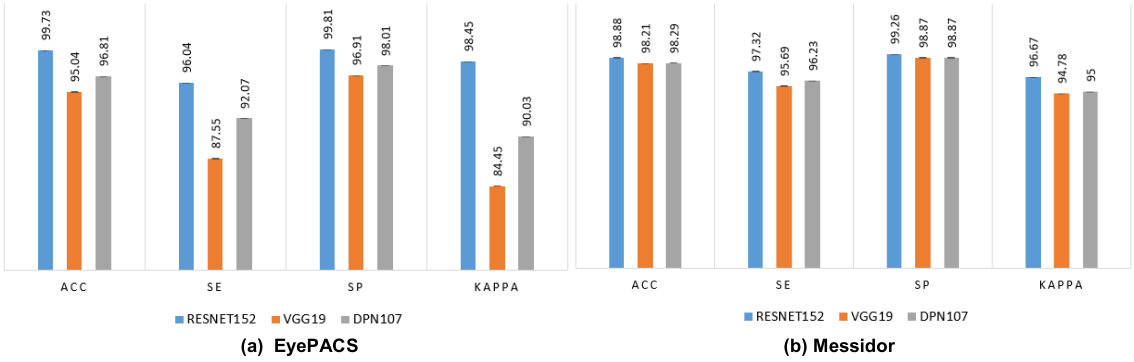

FIGURE 8. Comparison between the effects of the three CNN models on the method with re-initialized CONV1 and GB along with STD.

it is not as deep as ResNet152, which is why DPN107 gives a better performance compared to VGG19.

4) CORRECTNESS OF THE MODEL

The discussion in the previous subsections revealed that the best configuration of the proposed system involved ResNer152 with re-initialized CONV1 and a GB classification layer. We will refer to this as ResNetGB moving forward. It is very important to ensure that the model is not suffering from overfitting and underfitting. A deep model suffers from overfitting (or underfitting) if the bias is low (high), but the variance is high (or low) [49], [50]. Overfitting means a model memorizes the data, gives a good performance on the training data, but a poor performance on the test data. The performance is also not consistent when the model is trained and tested over different datasets. It can be tested using a 10-fold cross-validation (CV). Table 7 lists the 10-fold CV results along with the bias and the variance of the proposed model on EyePACS and Messidor. The bias and the variance of each fold were very low; the average bias and variance over 10 fold were also very low; and the results of 10 fold were consistent, indicating that the proposed model was robust

TABLE 7. Accuracy results showing bias and variance of ResNetGB.

	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10	AVG
EyePACS Dataset											
Training	99.99%	99.98%	99.99%	99.98%	99.99%	99.99%	99.98%	99.99%	99.99%	99.9%	99.9%±0.0003
Testing	99.78%	99.74%	99.66%	99.71%	99.76%	99.73%	99.74%	99.69%	99.76%	99.7%	99.7%±0.0003
Bias	0.01	0.02	0.01	0.02	0.01	0.01	0.02	0.01	0.01	0.01	0.01
Variance	0.21	0.24	0.33	0.27	0.23	0.26	0.24	0.30	0.23	0.27	0.26
Messidor Dataset											
Training	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Testing	99.17%	98.33%	98.33%	98.75%	98.75%	99.58%	98.75%	98.33%	99.16%	99.6%	98.88%
Bias	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Variance	0.83	1.67	1.67	1.25	1.25	0.42	1.25	1.67	0.84	0.42	1.13

TABLE 8. Results with the best configuration for the multi class in the EyePACS and messidor datasets.

	ACC	SE	SP	Kappa	AUC
EyePACS	99.73%	96.04%	99.81%	98.45%	0.98
Messidor	98.88%	97.32%	99.26%	96.67%	0.98

and did not suffer from overfitting or underfitting on both datasets.

5) BEST CONFIGURATION MODEL: RESULTS AND DISCUSSION

The following subsections provide the results of the proposed system with the best configuration for three scenarios, namely multiclass case, normal vs DR, and referable vs non-referable, which are commonly addressed in the existing work on DR diagnosis. We also present a comparison with the state-of-the-art works.

a: MULTICLASS CASE

The EyePACS dataset has five classes: normal, mild DR, moderate DR, severe DR, and proliferative DR. Messidor has four classes: normal and three DR grades based on the number of MAs and HEs. Table 8 lists the results of the system with the best configuration on each dataset.

Overall, the results on both datasets were high and almost comparable in terms of all metrics. Both datasets were challenging datasets, and the proposed model achieved above 95% ACC, SE, SP, kappa, and AUC, implying model robustness. Figure 9 illustrates the confusion matrices for both datasets computed using the 10-fold cross-validation. These give insights into the system performance for different classes on different challenging datasets. The class level results in Table 9 (a) and the confusion matrix for EyePACS in Figure 9(a) indicate that the maximum rate of misclassified images (i.e., 10.28%) is related to the class proliferative DR (DR grade 4). A total of 66 out of 708 patients with proliferative DR were misclassified as mild DR (20), moderate DR (11), and severe DR (35). The next maximum rate of the misclassified images (7.5%) belonged to severe DR

TABLE 9. Performance for each class based on EyePACS and messidor datasets.

Metric	Normal	Mild DR	Moderate DR	Severe DR	Proliferative DR
SE	99.9%	97.2%	99.35%	93.01 %	90.67%
SP	99.7%	99.8%	99.76%	99.87 %	99.95

(a) EyePACS

Metric	Normal	DR-level 1	DR-level 2	DR-level 3
SE	98.90%	96.73%	98.38%	95.28%
SP	99.52%	99.41%	98.86%	99.25%

(b) Messidor

(DR grade 3) probably because many images belonging to these classes were dark and had a low contrast (Figure 10(a)). Such MA and EX regions in these images were not clear, and the model failed to classify them correctly. The normal images had the smallest misclassified rate and high SP (Table 9)(a). The small number of images, which were misclassified as normal, came from classes 1 and 2 (i.e., 0.2% and 0.4% of the images belonging to mild (DR grade 1) and moderate (DR grade 2) DR, respectively) likely because the images belonging to the adjacent classes might be labeled incorrectly due to the continuity of the DR progression. This is hard to separate even for experts [51]. Given the clinical protocols, some state-of-the-art methods classified normal and DR grade 1 as non-referable [15], [16]; hence, it was not a big concern. Table 9 (a) presents the SE and the SP of each class in the EyePACS dataset based on the confusion matrix in Figure 9(a). Proliferative DR and severe DR had lower sensitivities than the other DR grades. The SP of all classes was very high, indicating that the FNR of the proposed system with the best configuration was very small. Based on the confusion matrix in Table 9 (b) and Figure 9(b) for the Messidor dataset, the maximum misclassification rates were of DR grade 3 (4.96%) and grade 1 (3.38%). This was probably caused by the same reason found for the case of EyePACS (i.e., some images belonging to

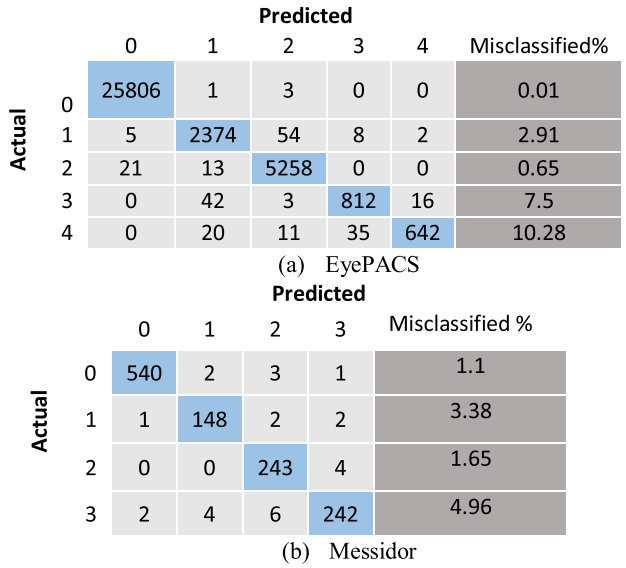

FIGURE 9. Confusion matrix of the best model on the (a) EyePACS (five classes) and (b) Messidor (four classes) datasets.

TABLE 10. Non DR vs DR stages result based on the EyePACS and messidor datasets.

Dataset	SE	SP
EyePACS	99.72%	99.99%
Messidor	99.54%	98.90%.

these classes had poor contrast for MAs and HEs because of the dominant red color in Figure 10(b)). The model failed to classify them correctly. The normal class had the lowest misclassification rate of 1.1% as in the case of EyePACS. Meanwhile, the number of images misclassified as normal was 3 (0.25%). Table 9 (b) lists the sensitivities and the specificities of the four classes in the Messidor dataset showing that classes DR-level 1 and 3 had lower sensitivities. Normal and all DR levels had a high SP. The performance of the proposed system on both datasets revealed that it was robust and provided very low FPR and FNR for the DR grading scenario.

b: NORMAL VS DR

In this scenario, we considered the non-DR vs DR stages (all DR grades). In the case of the EyePACS dataset, normal was taken as non-DR, and all DR severities (i.e., mild DR, moderate DR, severe DR, and proliferative DR) were considered as DR [52]. For the Messidor dataset, the DR grades (1, 2, and 3) were considered as DR [8]. Figure 11(a) depicts the 10-fold cross-validation confusion matrix for EyePACS, while Figure 11(b) shows that for the Messidor dataset. Table 10 presents the corresponding sensitivities and specificities.

From the confusion matrix related to EyePACS, only four normal images were misclassified as DR (0.01%) (i.e., FPR was very low, while SP was very high at 99.99%) (Table 10).

TABLE 11. Result for non-referable vs referable cases on the EyePACS and messidor datasets.

Dataset	SE	SP
EyePACS	98.60%	99.76%.
Messidor	98.80%	98.86%.

Only 26 DR images were misclassified as normal (i.e., FNR was very low at 0.27%, and SE was very high at 99.72%). The confusion matrix corresponding to the Messidor dataset indicated that only 1.1% of the normal images were misclassified as DR, and 0.4% of the DR images were misclassified as normal. In summary, the model had very low FPR and FNR and very high SE (99.54%) and SP (98.90%) (Table 10). The performance of the proposed model on both datasets revealed its robustness and provided very high SE and SP for the normal vs. DR scenario.

c: REFERABLE VS NON-REFEREABLE

We considered the non-referable vs referable case in this scenario. For the EyePACS dataset, non-referable was taken as normal and mild DR, while all other DR severities (i.e., moderate DR, severe DR, and proliferative DR) were considered referable [15], [16]. In the case of Messidor, DR grades 2 and 3 were considered referable, while DR grades 0 and 1 were non-referable [10]. Figure 12(a) illustrates the 10-fold cross-validation confusion matrix for EyePACS, which indicates that 0.24% of the non-referable images were misclassified as referable, and 1.4% of the referable cases were misclassified as non-referable. This result indicates that the proposed model has high SE (98.60%) and SP 99.76% (Table 11). The confusion matrix of the Messidor dataset in Figure 12(b) depicts that 1.2% of the non-referable images were misclassified as referable, and 1.2% of the referable cases were misclassified as non-referable. In other words, the proposed model obtained very high SE (98.80%) and SP (98.86%) (Table 11). The results and the discussion above implied that the proposed method was robust because it had very high SE and SP on both challenging datasets for the referable vs non-referable scenario.

d: DISCUSSION

This section presents a comparison of ResNetGB with the state-of-the-art methods, which focus on the DR grading of retinal fundus images and have been evaluated on the EyePACS and Messidor datasets. The comparison is made with hand-engineered and deep learning-based methods for the three scenarios, a global view is given in Table 12. ResNetGB was evaluated by employing all labeled images from the EyePACS dataset using 10-fold CV for three scenarios: DR grading, normal vs DR, and non-referral (normal and DR grade 1) vs referral (DR grade 2 to max DR grade). For the DR grading, the state-of-the-art method by Wan *et al.* [18] achieved the maximum ACC of 95.86%, SE of 86.47%, SP of 97.43, and AUC of 0.9786 on the EyePACS dataset. This method used VGGNet-s with transfer learning.

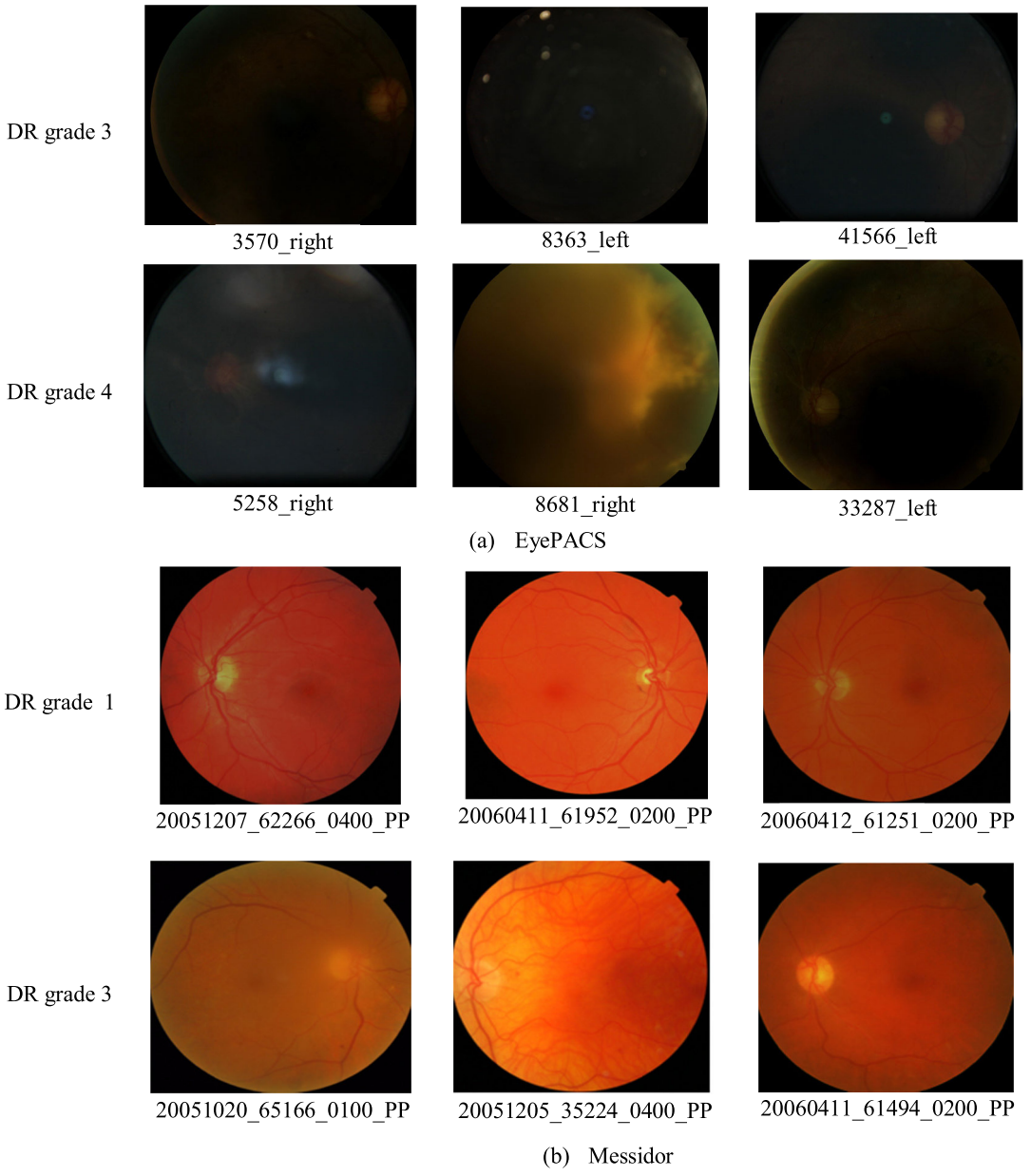

FIGURE 10. Example of misclassified images from the (a) EyePACS and (b) Messidor datasets.

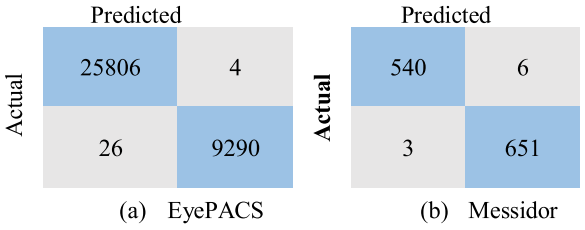

FIGURE 11. Confusion matrices of the system with the best configuration on the (a) EyePACS and (b) Messidor datasets for the non DR vs DR stages.

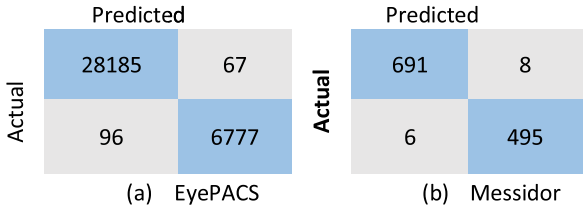

FIGURE 12. Confusion matrix of ResNet152 with a re-initialized CONV1 model on the (a) EyePACS dataset and the (b) Messidor dataset for non-referable vs referable cases.

Meanwhile, Quellec et al. [1] built a CNN model from scratch using 80% of the EyePACS dataset as the training set and 20% as the test set and achieved an ACC of 95.4%. ResNetGB outperformed all existing methods by a big margin, exhibiting 3.87% ACC improvement, 9.57% SE improvement, and 2.38% SP improvement. The difference was statistically

TABLE 12. Comparison of ResNetGB and the state-of-the-art methods.

Paper	Method	Dataset	Performance				
			ACC (%)	SE (%)	SP (%)	AUC (%)	Kappa (%)
DR grading (multiclass)							
[14]	CNN model	EyePACS (train: 96%, test: 4%)	75.00	30.00	95.00		
[18]	Transfer learning (VGGNet-s)	EyePACS	95.68	86.47	97.43	0.9786	
[16]	CNN model	EyePACS (train:96%, test:4%)				0.844	85.1
[1]	CNN model	EyePACS (train:80%, test:20%)	95.4				
[17]	SI2DRNet-v1	Messidor	91.2			0.965	
Proposed	ResNetGB	EyePACS (10-fold CV) Messidor (10-fold CV)	99.73 98.88	96.04 97.32	99.81 99.26	0.98 0.98	98.45 96.67
Normal vs DR (all DR levels) (two classes)							
[8]	Handcrafted features	Messidor		93.9	50	0.899	
[3]	Bag of visual words and SVM	Messidor		98	97		
[16]	CNN model	EyePACS (train:96%, test:4%)		94.5	90.2		
Proposed model	ResNetGB	EyePACS (10-fold CV) Messidor (10-fold CV)		99.72 99.54	99.99 98.90	0.9986 0.9913	
Non-referral (normal and DR grade 1) vs referral (DR grade 2 to the highest grade) (two classes)							
[8]	Handcrafted features	Messidor		96.2	50	0.916	
[10]	Handcraft features	Messidor				0.863	
[15]	CNN model End to end training	EyePACS (train:89%, test:11%)		96.2	66.6	0.946	
[16]	CNN model	EyePACS (train:96%, test:4%)		98	94		
Proposed model	ResNetGB	EyePACS (10-fold CV) Messidor (10-fold CV)		98.60 98.80	99.76 98.86	0.9918 0.9876	

significant at $p < 5.8E{-}10$ using 95% CL for SE and $p < 0.02$ using 97% CL for SP.

The best Kappa achieved so far on the EyePACS dataset was 85.1% [16], which is significantly lower than that obtained by ResNetGB (98.45%). The difference was statistically significant at $p < 7.00149E{-}18$ for Kappa using 95% CL. On the Messidor dataset, ResNetGB also outperformed the existing method with an improvement of 8.68% in ACC and 0.017 in AUC. The difference was statistically significant at $p < 4.7E{-}07$ using 95% CL for ACC and $p < 0.027$ using 97% CL.

For the normal vs DR case, the ResNetGB model performed better than the method by Sreejini and Govindan [3], which was based on handcrafted features, on the Messidor dataset. The performance increase of ResNetGB over this method was 1.19% in SE and 1.47% in SP. The difference was statistically significant at $p < 0.03$ using 99% CI for SE and $p < 0.03$ using 96% CL for SP.

For the non-referable vs. referable scenario, the best performing deep learning-based method by Colas *et al.* [15] provided a maximum SE of 98% and an SP of 94% on the EyePACS dataset. These values were less than those achieved by ResNetGB (SE: 98.60%; SP:99.76%). The difference was significant at $p < 0.02$, as obtained using 97% CL for SE and $p < 0.04$ using 96% CL for SP. The ResNetGB model also exhibited an enhanced performance on the Messidor dataset by approximately 2.6% in SE and 48.86% in SP compared to the best existing method by Seoud *et al.* [5]. The difference was significant at $p < 0.03$ using 95% CI for SE. Moreover, ResNetGB improved by 0.124 in terms of the AUC on the Messidor dataset compared to the method by Pires method, which was based on handcrafted features (Fisher Vector) [10]. The difference was significant at $p < 1.3E{-}14$ using 95% CL.

The above discussion indicates that ResNetGB outperformed the existing deep learning-based methods [1], [14]–[18] on both EyePACS and Messidor datasets for three diagnosis scenarios. The superiority of ResNetGB is attributed to its consideration of the hierarchical structure of the CNN model and its adaptation of it to the lesion structures by reinitializing the CONV1 weights using lesion ROIs and by its fine-tuning with ROIs. Another important point is that we did not resize the fundus images to a small size like most state-of-the-art deep learning methods because it destroys the lesion structures.

IV. CONCLUSION

We developed herein an automatic system based on deep learning for grading retinal fundus images and referring a DR patient to an ophthalmologist at an early stage. The system was built on a pre-trained model using a two-stage transfer learning method because of the limited available dataset and a huge number of parameters in a deep CNN model. First, we tried three state-of-the-art CNN models pre-trained on ImageNet. Natural images in ImageNet have structures different from those of fundus images; thus, we adapted the hierarchical structure of a pre-trained CNN model to the fundus images by reinitializing the filters of its CONV1 layer using the lesion ROIs extracted from the annotated E-optha dataset and then fine-tuned it using the ROIs. Second, the FC layers encoded the high-level features relevant to the natural images and have a very large number of learnable parameters. To tune them to high-level features, reduce the model complexity, and avoid overfitting, we replaced the FC layers with a PCA layer learned using ROIs and used it to extract discriminate features from the fundus images. We then added a classification layer to predict the DR grades of fundus images. Consequently, ResNet152 with re-initialized CONV1 and GB layer (ResNetGB) achieved the best results. A comparison with the state-of-the-art methods showed that ResNetGB outperformed the others in three DR diagnosis scenarios, namely DR grading, normal vs DR, and non-referable vs referable on the two challenging datasets of EyePACS and Messidor. The difference was statistically significant. The ResNetGB model did not employ any preprocessing or enhancement step. Its shortcoming is that it failed to accurately predict the DR grade of a fundus image when it has saturation and a very low contrast. For the fundus images with a good quality, it accurately predicted the DR level. How the model can reliably predict the DR level from poor quality, very low contrast, and saturated fundus images will be the subject of the future work. The system will help graders to screen DR patients reliably without any delay at the early stages.

REFERENCES

[1] G. Quellec, K. Charrière, Y. Boudi, B. Cochener, and M. Lamard, "Deep image mining for diabetic retinopathy screening," *Med. Image Anal.*, vol. 39, pp. 178–193, Jul. 2017.

[2] N. Cheung, P. Mitchell, and T. Y. Wong, "Diabetic retinopathy," *Lancet*, vol. 376, no. 9735, pp. 124–136, 2010.

[3] K. S. Sreejini and V. K. Govindan, "Retrieval of pathological retina images using bag of visual words and pLSA model," *Eng. Sci. Technol., Int. J.*, vol. 22, no. 3, pp. 777–785, Jun. 2019.

[4] D. J. Hemanth, O. Deperlioglu, and U. Kose, "An enhanced diabetic retinopathy detection and classification approach using deep convolutional neural network," *Neural Comput. Appl.*, vol. 32, no. 3, pp. 707–721, 2020.

[5] S. Stolte and R. Fang, "A survey on medical image analysis in diabetic retinopathy," *Med. Image Anal.*, vol. 64, Aug. 2020, Art. no. 101742.

[6] S. Morales, K. Engan, V. Naranjo, and A. Colomer, "Retinal disease screening through local binary patterns," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 1, pp. 184–192, Jan. 2017.

[7] R. Kamble, M. Kokare, G. Deshmukh, F. A. Hussin, and F. Mériaudeau, "Localization of optic disc and fovea in retinal images using intensity based line scanning analysis," *Comput. Biol. Med.*, vol. 87, pp. 382–396, Aug. 2017.

[8] L. Seoud, T. Hurtut, J. Chelbi, F. Cheriet, and J. M. P. Langlois, "Red lesion detection using dynamic shape features for diabetic retinopathy screening," *IEEE Trans. Med. Imag.*, vol. 35, no. 4, pp. 1116–1126, Apr. 2016.

[9] N. Gori, H. Kadakia, V. Kashid, and P. Hatode, "Detection and analysis of microanuerysm in diabetic retinopathy using fundus image processing," *Int. J. Adv. Res., Ideas Innov. Technol.*, vol. 3, no. 2, pp. 907–911, 2017.

[10] R. Pires, S. Avila, H. F. Jelinek, J. Wainer, E. Valle, and A. Rocha, "Beyond lesion-based diabetic retinopathy: A direct approach for referral," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 193–200, Jan. 2017.

[11] G. Quellec, M. Lamard, A. Erginay, A. Chabouis, P. Massin, B. Cochener, and G. Cazuguel, "Automatic detection of referral patients due to retinal pathologies through data mining," *Med. Image Anal.*, vol. 29, pp. 47–64, Apr. 2016.
[12] P. S. Grewal, F. Oloumi, U. Rubin, and M. T. S. Tennant, "Deep learning in ophthalmology: A review," *Can. J. Ophthalmol.*, vol. 53, no. 4, pp. 309–313, Aug. 2018.
[13] K. M. Adal, P. G. Van Etten, J. P. Martinez, K. W. Rouwen, K. A. Vermeer, and L. J. van Vliet, "An automated system for the detection and classification of retinal changes due to red lesions in longitudinal fundus images," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 6, pp. 1382–1390, Jun. 2018.
[14] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional neural networks for diabetic retinopathy," *Procedia Comput. Sci.*, vol. 90, pp. 200–205, Jan. 2016.
[15] E. Colas, A. Besse, A. Orgogozo, B. Schmauch, N. Meric, and E. Besse, "Deep learning approach for diabetic retinopathy screening," *Acta Ophthalmol.*, vol. 94, Oct. 2016. [Online]. Available: https://onlinelibrary.wiley.com/toc/17553768/2016/94/S256
[16] S. M. S. Islam, M. M. Hasan, and S. Abdullah, "Deep learning based early detection and grading of diabetic retinopathy using retinal fundus images," 2018, *arXiv:1812.10595*. [Online]. Available: http://arxiv.org/abs/1812.10595
[17] Y.-W. Chen, T.-Y. Wu, W.-H. Wong, and C.-Y. Lee, "Diabetic retinopathy detection based on deep convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 1030–1034.
[18] S. Wan, Y. Liang, and Y. Zhang, "Deep convolutional neural networks for diabetic retinopathy detection by image classification," *Comput. Electr. Eng.*, vol. 72, pp. 274–282, Nov. 2018.
[19] Z. Gao, J. Li, J. Guo, Y. Chen, Z. Yi, and J. Zhong, "Diagnosis of diabetic retinopathy using deep neural networks," *IEEE Access*, vol. 7, pp. 3360–3370, 2019.
[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[21] 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.
[22] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, "Dual path networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–11.
[23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
[24] Kaggle. (Feb. 23, 2019). *Diabetic Retinopathy Detection (Kaggle)*. [Online]. Available: https://www.kaggle.com/c/diabetic-retinopathy-detection/data
[25] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, B. Charton, and J.-C. Klein, "Feedback on a publicly distributed image database: The Messidor database," *Image Anal. Stereol.*, vol. 33, no. 3, pp. 231–234, 2014.
[26] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018.
[27] C. Lam, C. Yu, L. Huang, and D. Rubin, "Retinal lesion detection with deep learning using image patches," *Invest. Ophthalmol. Vis. Sci.*, vol. 59, no. 1, pp. 590–596, 2018.
[28] E. Decencière, G. Cazuguel, X. Zhang, G. Thibault, J.-C. Klein, F. Meyer, B. Marcotegui, G. Quellec, M. Lamard, R. Danno, D. Elie, P. Massin, Z. Viktor, A. Erginay, B. Laÿ, and A. Chabouis, "TeleOphta: Machine learning and image processing methods for teleophthalmology," *IRBM*, vol. 34, no. 2, pp. 196–203, 2013.
[29] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural Comput.*, vol. 29, no. 9, pp. 2352–2449, 2017
[30] Y. Yu, H. Lin, J. Meng, X. Wei, H. Guo, and Z. Zhao, "Deep transfer learning for modality classification of medical images," *Information*, vol. 8, no. 3, p. 91, 2017.
[31] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.
[32] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and < 0.5 MB model size," 2016, *arXiv:1602.07360*. [Online]. Available: http://arxiv.org/abs/1602.07360
[33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.
[34] J. Hu, Y. Chen, J. Zhong, R. Ju, and Z. Yi, "Automated analysis for retinopathy of prematurity by deep neural networks," *IEEE Trans. Med. Imag.*, vol. 38, no. 1, pp. 269–279, Jan. 2019.
[35] Alexis Cook. (2017). *Global Average Pooling Layers for Object Localization*. Accessed: Aug. 19, 2019. [Online]. Available: https://alexisbcook.github.io/2017/globalaverage-poolinglayers-for-object-localization/
[36] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 618–626.
[37] L. Breiman, *Classification and Regression Trees*. Belmont, CA, USA: Wadsworth, 1984, p. 432.
[38] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[39] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.
[40] A. R. Chowdhury, T. Chatterjee, and S. Banerjee, "A random forest classifier-based approach in the detection of abnormalities in the retina," *Med. Biol. Eng. Comput.*, vol. 57, no. 1, pp. 193–203, 2019.
[41] D. Lavanya and K. U. Rani, "Performance evaluation of decision tree classifiers on medical datasets," *Int. J. Comput. Appl.*, vol. 26, no. 4, pp. 1–4, 2011.
[42] A. Mayr, H. Binder, O. Gefeller, and M. Schmid, "The evolution of boosting algorithms," *Methods Inf. Med.*, vol. 53, no. 6, pp. 419–427, 2014.
[43] J. C. Klein *et al.*, "Methods to evaluate segmentation and indexing techniques in the field of retinal ophthalmology (MESSIDOR)," Ecole des Mines de Paris, Paris, France, 2016.
[44] G. Forman and M. Scholz, "Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement," *ACM SIGKDD Explor. Newslett.*, vol. 12, no. 1, pp. 49–57, 2010.
[45] 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.
[46] S. Haghighi, M. Jasemi, S. Hessabi, and A. Zolanvari, "PyCM: Multiclass confusion matrix library in Python," *J. Open Source Softw.*, vol. 3, no. 25, p. 729, May 2018.
[47] D. M. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," 2020, *arXiv:2010.16061*. [Online]. Available: https://arxiv.org/abs/2010.16061
[48] M. W. Fagerland and L. Sandvik, "The Wilcoxon–Mann–Whitney test under scrutiny," *Statist. Med.*, vol. 28, no. 10, pp. 1487–1497, 2009.
[49] C. Schaffer, "Overfitting avoidance as bias," *Mach. Learn.*, vol, 10, no. 2, pp. 153–178, 1993.
[50] NG Andrew, Technical Strategy for AI Engineers Draft 2018. *Machine Learning Yearning*. Accessed: Aug. 2018. [Online]. Available: http://www.mlyearning.org
[51] C. I. Sánchez, M. Niemeijer, A. V. Dumitrescu, M. S. A. Suttorp-Schulten, M. D. Abramoff, and B. van Ginneken, "Evaluation of a computer-aided diagnosis system for diabetic retinopathy screening on public data," *Invest. Ophthalmol. Vis. Sci.*, vol. 52, no. 7, pp. 4866–4871, 2011.
[52] R. Gargeya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," *Ophthalmology*, vol. 124, no. 7, pp. 962–969, 2017.

FAHMAN SAEED received the master's degree from King Saud University, where he is currently pursuing the Ph.D. degree with the Computer Science Department, College of Computer and Information Science. His research interests include image processing and pattern recognition, deep learning, and medical image recognition.

MUHAMMAD HUSSAIN (Senior Member, IEEE) received the M.Sc. and M.Phil. degrees from the University of the Punjab, Lahore, Pakistan, in 1990 and 1993, respectively, and the Ph.D. degree in computer science from Kyushu University, Fukuoka, Japan, in 2003. From April 2003 to September 2005, he worked as a Postdoctoral Researcher with Kyushu University and received funding from the Japan Science and Technology Agency (JST). He is currently a Professor with the Department of Computer Science, King Saud University, Saudi Arabia. His current research interests include deep learning, image forensics, digital watermarking, medical imaging (mammograms, diabetic retinopathy, EEG brain signals), and biometrics (face recognition, fingerprint recognition). In these research areas, he has published more than 100 research articles in ISI indexed journals, and the proceedings of refereed international conferences. He has received several research grants from the Japan Science and Technology Agency (JST), the National Science Technology and Innovation Plan (NSTIP) of Saudi Arabia, and the Research Center of College of Computer and Information Sciences, KSU, Saudi Arabia. He is a member of editorial boards, an advisor, and a Reviewer of many famous ISI journals, international conferences and funding agencies. He was an Editor of the *Journal of Computer and Information Sciences, King Saud University* (Elsevier). He has served on the program committees of various international conferences.

HATIM A. ABOALSAMH (Senior Member, IEEE) received the Ph.D. degree in computer engineering and science from the University of Miami, Miami, FL, USA, in 1987. He is currently a Professor and a Chairman of the Department of Computer Science, King Saud University, Saudi Arabia. His research interests include pattern recognition, biometrics, probability modeling, and machine learning. He is a Fellow Member of the British Computer Society (BCS). He is a Senior Member of the Association of Computing Machinery, USA (ACM), and the International Association of Computer Science and Information Technology (IACSIT), Singapore. He is a Professional Member of the IEEE. He is also the Editor-in-Chief of the Saudi Computer Society Journal. He held leading posts of Vice Rector for Development and Quality-King Saud University (KSA), Riyadh, Saudi Arabia, from 2006 to 2009, and the Dean of the College of Computer and Information Sciences, the Editor-in-Chief of the KSU-Journal of Computer Sciences, and the Vice President of the Saudi Computer Society.