This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2888639, IEEE Access

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Diagnosis of Diabetic Retinopathy Using Deep Neural Networks

ZHENTAO GAO[1], JIE LI[2], JIXIANG GUO[1],(Member, IEEE), YUANYUAN CHEN[1],(Member, IEEE), ZHANG YI[1],(Fellow, IEEE), JIE ZHONG[2] .

[1]Zhentao Gao, Jixiang Guo, Yuanyuan Chen and Zhang Yi are with the Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu 610065, P. R. China

[2]Jie Li and Jie Zhong are with Sichuan Academy of Medical Sciences (Sichuan Provincial People's Hospital)(e-mail: zhongjiellxx@163.com)

Corresponding author:Jie Zhong (e-mail: zhongjiellxx@163.com).

This work was supported by the National Natural Science Foundation of China [Grant numbers 61432012 and U1435213].

ABSTRACT Diabetic retinopathy (DR) is a common eye disease and a significant cause of blindness in diabetic patients. Regular screening with fundus photography and timely intervention is the most effective way to manage the disease. The large population of diabetic patients and their massive screening requirements have generated interest in computer-aided and fully automatic diagnosis of DR. Deep neural networks, on the other hand, have brought many breakthroughs in various tasks in the recent years. To automate the diagnosis of DR and provide the appropriate suggestions to DR patients, we have built a dataset of DR fundus images that have been labeled by the proper treatment method that is required. Using this dataset, we trained deep convolutional neural network models to grade the severities of DR fundus images. We were able to achieve an accuracy of 88.72% for a four degree classification task in the experiments. We deployed our models on a cloud computing platform and provided pilot DR diagnostic services for several hospitals, in the clinical evaluation the system achieved a consistency rate of 91.8% with ophthalmologists, demonstrating the effectiveness of our work.

INDEX TERMS Diabetic Retinopathy, Automatic Diagnosis, Deep Neural Networks

I. INTRODUCTION

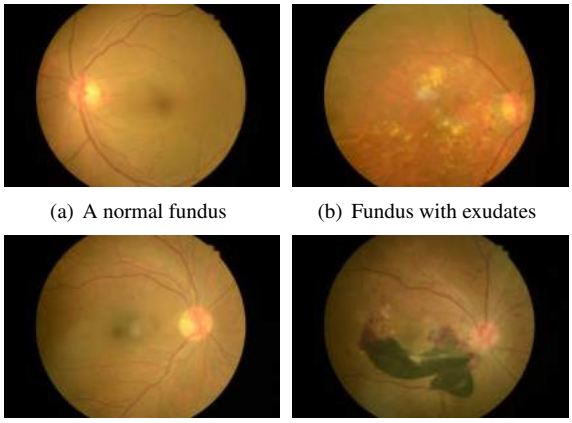
DIABETIC retinopathy (DR) is the most common cause of blindness among diabetic patients [1]. According to the World Health Organization (WHO), there were 422 million diabetic patients in 2014, 35% of whom developed some type of retinopathy owing to the accumulation of damage to small blood vessels in the retina [2]. The prevalence of DR is much higher among special groups of patients. For example, it is estimated that 40% of type II diabetic patients and 86% of type I diabetic patients in the US have DR, and the rate of DR is estimated to be 43% in rural areas of China [3].

The loss of sight can vary during the gradual development of DR. Generally, DR can be separated into two major stages: non-proliferative DR (NPDR) and proliferative DR (PDR), which is characterized by neovascularization or vitreous/preretinal hemorrhage. Up to 10% of diabetic patients who have no DR will develop NPDR annually, and for patients with severe NPDR, the risk of developing PDR in one year is 75%. The shift from normal status(no apparent abnormality in the retina) to PDR commonly takes many years. Thus, NPDR is often divided into three sub-stages: mild, moderate, and severe NPDR. Together, these five stages make up the widely used 'International Clinical Diabetic Retinopathy Disease Severity Scale' [4]. The best treatment options for patients differ between stages. For patients with no DR or mild NPDR, only regular screening is required; for patients with moderate NPDR or worse, the treatment options vary from scatter laser treatment to vitrectomy. Thus, to provide patients with the appropriate treatment, it is important to first grade their DR severity. Clinically, the diagnosis of DR is often made with fundus images, which can be acquired by photographing the fundus directly. The common lesions that indicate DR include hard or soft exudates, microaneurysms and hemorrhages. All of these lesions can be identified from fundus images; FIGURE. 1 shows sample fundus images containing various types of lesions. To make a more accurate diagnosis, fluorescein angiography can be used because it can reveal fine vessel structures in the retina. However, fluorescein dyes can cause an allergic reaction and require functioning kidneys to excrete, and they are usually not available in small hospitals. Currently, fundus images are the most widely used approach for regular screening of DR,

2169-3536 (c) 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2888639, IEEE Access

since the acquisition of such images is convenient and the visibility of most lesions is sufficient.



(a) A normal fundus (b) Fundus with exudates

(c) Fundus with microaneurysms (d) Fundus with hemorrhages

FIGURE 1: Sample fundus images with different type of lesions

Although the equipment for fundus photography can be accessed easily, a qualified ophthalmologist who can analyze the fundus images cannot. The population of diabetic patients is enormous, and the prevalence of diabetes has been rising rapidly—the global prevalence of diabetes among adults has increased from 4.7% in 1980 to 8.5% in 2014 [2]. Yet, experienced ophthalmologists are rare and are distributed unevenly; the total number of ophthalmologists worldwide in 2012 was 210,730 (i.e., 29 ophthalmologists per 1 million persons) [5]. For middle- and low-income countries, the gap between the population of diabetic patients and ophthalmologists can be extremely wide, indicating an urgent need for systems that diagnose DR automatically.

Much work has been done in using computers to make automatic DR diagnoses. Traditional methods often deploy various feature extraction modules to first extract useful information from fundus images. Then, the extracted features are fed into certain classifiers, such as random forests, support vector machine, and the AdaBoost classifier. Such handcrafted feature based methods are laborious and often fail to yield good results.

In the recent decade, deep neural networks (DNNs) have achieved revolutionary results in many areas. They bring about breakthroughs in computer vision, speech recognition, and natural language processing etc. Many applications of deep neural networks have demonstrated performance that can surpass human beings, e.g. face recognition [6], large scale visual recognition [7],the game of Go [8]. The use of DNNs in the diagnosis of DR has also attracted much interest, and much progress has been made. However, despite the many advances that have been made, clinical application of automatic DR diagnosis systems remains unavailable and many works still need to be done.

In this paper, we propose a new dataset of fundus images for grading DR. In contrast to existing scales that grade fundus images mainly by the pathological changes in the retina, we take into account clinical practice: i.e., we grade a fundus image based on its abnormalities and required treatment method. For example, patients with severe NPDR and mild PDR are recommended to undergo laser scatter treatment; thus, we group them in the same category. With this dataset, several deep convolutional neural network (DCNN) models were trained for the diagnosis of DR. We also propose a new model which is better adapted to small lesions in the fundus images. The experimental results show that our work achieves state-of-the-art performance compared with other works on similar tasks. Our models were also deployed for clinical evaluation in several hospitals, performing nearly as well as ophthalmologists.

II. RELATED WORKS

In the decades-long seeking for automatic DR diagnosis, many works have been done. We summarize related works briefly from three aspects. Firstly, we list the most related datasets that are constructed for DR related tasks. Secondly, we summarize related works using traditional image processing techniques. In the end, we showcase some recent works that use various convolutional neural network (CNN) models.

A. RELATED DATASET

Various datasets of fundus images for DR-related diagnoses have been developed, including

- Standard Diabetic Retinopathy Database Calibration Level 0/1 (DIARETDB0/DIARETDB1) dataset [9], [10],
- Methods to Evaluate Segmentation and Indexing techniques in the field of Retinal Ophthalmology (MESSIDOR) dataset [11],
- Digital Retinal Images for Vessel Extraction (DRIVE) dataset [12],
- STructured Analysis of the Retina (STARE) dataset [13],
- Retinal Vessel Image set for Estimation of Widths (REVIEW) dataset [14],
- Kaggle Diabetic Retinopathy dataset [15],
- E-ophtha dataset [16].

These datasets differ significantly in their annotation, for that they were proposed for different tasks. For example, in the DIARETDB0 dataset, the type and position of each abnormality are labeled in detail, whereas in the Kaggle DR dataset, the label for each image is a simple integer indicating one of five degrees of severity. The sizes of these datasets also differ. The REVIEW dataset contains only 16 images, and the DIARETDB0 dataset contains 130 images, while the Kaggle dataset contains up to 88,702 images. Further, the quality of the annotations varies significantly.

B. TRADITIONAL PRACTICE

Traditional image classification pipelines can be divided roughly into three main stages: image preprocessing, feature

2169-3536 (c) 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2888639, IEEE Access

extraction, and feature classification. For the diagnosis of DR, there are studies for each of these stages. For example, Rubini et al. proposed to apply hessian-based candidate selection before the feature extraction and classification using a support vector machine (SVM) classifier [17]. Mookiah et al. proposed a system that used hybrid features, including exudate/vessel area, texture, and entropy, for DR classification [18]. Bhatkar el al. explored the use of the multilayer perception neural network as the classifier to process extracted features, such as a 64-point discrete cosine transform, and other statistical features, including entropy and Euler's number [19].

Despite the performance of these approaches, the drawbacks of traditional methods are obvious. On one hand, given that simple and direct features are already exploited, crafting new effective features by hand becomes more difficult. On the other hand, the performance of these approaches plateaus, which makes them harder to improve.

C. DEEP CONVOLUTIONAL NEURAL NETWORK APPROACHES

Deep neural networks, especially convolutional ones, have demonstrated their superiority in image classification tasks. Many CNN-based methods have been introduced for making automatic DR diagnoses. For example, Gulshan et al. proposed to use the Inception-v3 as their architecture for detecting DR [20]. Quellec et al. created heatmaps of the image to show the role of each pixel for classification, and used a generalization of backpropagation for training CNNs to create the heatmaps [21]. Yang et al. proposed a two-stage DCNN-based algorithm that detects lesions in fundus images and grades the severity of DR [22]. Chandore et al. trained a DCNN model on a large dataset to detect the symptoms of DR from fundus images [23].

Although much progress has been made with these CNN-based approaches, there are still gaps between the current results and their clinical application. First of all, many works were done on sub tasks of the problem such as vessel segmentation or detecting lesions of specific kinds. Secondly, many experimental results were achieved on very small datasets and thus lack persuasion for real world applications. For example, the accuracies for vessel segmentation on the DRIVE and STARE datasets can be achieved as 97.67% and 98.13% [24], however vessel segmentation alone cannot be used as diagnosis of DR. The accuracies for lesion detection on the DIARETDB0 and DIARETDB1 datasets can be achieved as 96.0% and 94.6% [25], however the presence of certain lesions can only be used as supporting evidence and the diagnosis is up to ophthalmologists who use these facts. It should also be noticed that for these sub tasks the size of datasets are often small, i.e. the number of total images in the STARE, DRIVE, DIARETDB0 and DIARETDB1 datasets are only 20, 40, 130 and 89 respectively. For the DR severity grading task, performance of previous works are relatively low, for example, in [23], only 2 classes were considered (with or without DR), and the accuracies for the two classes were not practical (with DR 0.88, without DR 0.81). In [22], the accuracy for the four grades of NPDR was less than 60%, and the accuracy for another subset of the Kaggle dataset was 75% [26].

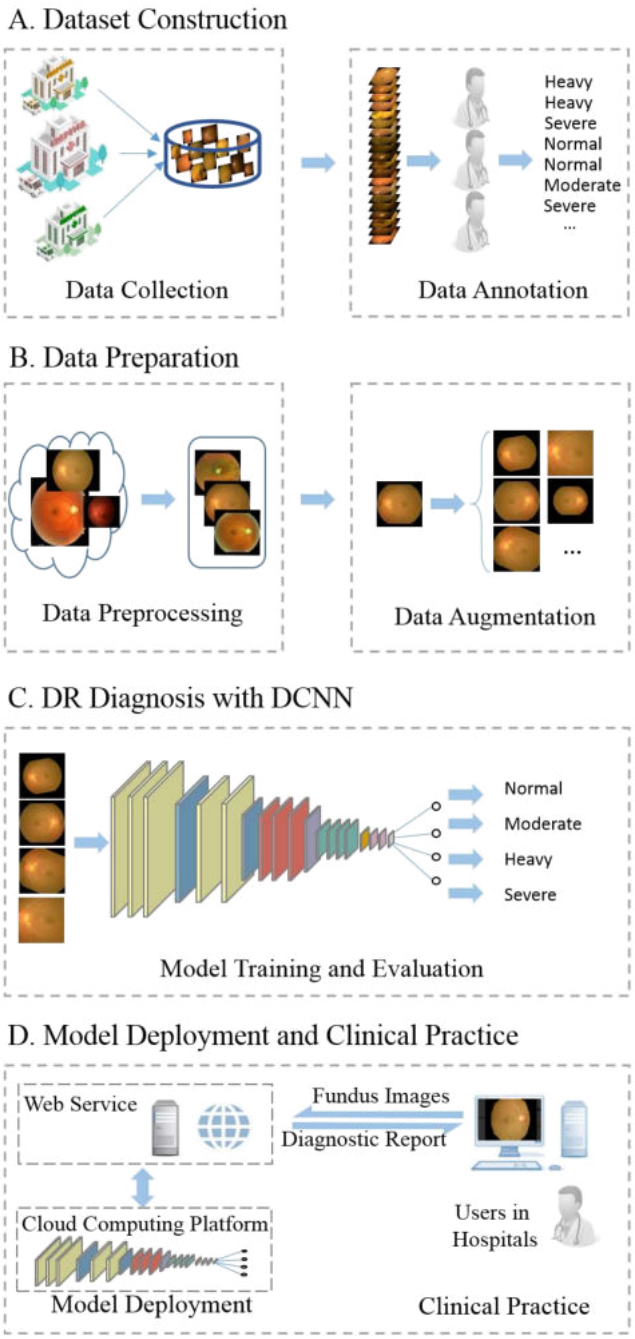


FIGURE 2: The framework of this work

III. OVERVIEW

Our aim was to build a model that can grade the severity of DR in a given fundus image. To this end, we performed the following steps, as shown in FIGURE. 2:

1) Data collection

2169-3536 (c) 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

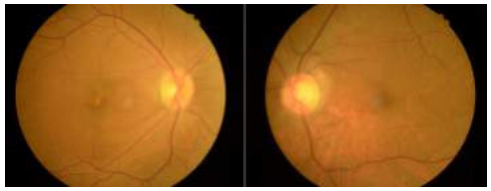
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2888639, IEEE Access

2) Data annotation
3) Data preprocessing
4) Data augmentation
5) Model setup and evaluation
6) Model deployment
7) Clinical evaluation

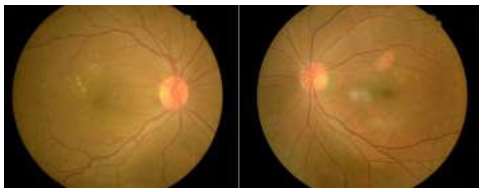
To describe these steps in detail, the rest of this paper is organized into five sections. In the dataset construction section, we describe how our data were collected and annotated and provide a detailed annotation scheme and our rationale for such a scheme. In the data preprocessing and augmentation section, the preprocessing pipeline for unifying images from different sources is described. The augmentation techniques in our work are also introduced in this section. In the model section, basic ideas and computational principles of DCNNs are described. Then, the models we use are described. In the experiments section, detailed descriptions for various experiments are included, and the experiment results are discussed and analyzed using visualization techniques. To evaluate our work in real clinical environments, we deploy our models on a cloud computing platform and provide a pilot diagnostic service to several hospitals in nearby cities. These works are introduced in the model deployment and clinical evaluation section. Finally, we present the conclusion section.

IV. DATASET CONSTRUCTION

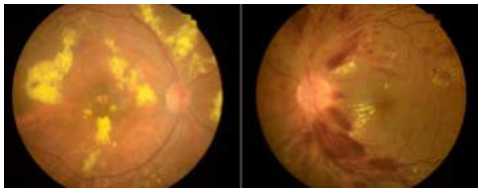
In addition to the existing datasets mentioned above, we have constructed a novel dataset of fundus images. Each image in the dataset is labeled with one of four degrees of DR severity. Our dataset is moderate in size, totaling 4476 images from three clinical departments (the Ophthalmology Department, Health Management Center, and Endocrinology & Metabolism Department) in Sichuan Provincial People's Hospital, which ranks second in Ophthalmology in Sichuan Province. After the data collection, three senior ophthalmologists were invited to label each of these images. Each image is labeled by them independently, for images labeled with inconsistency, discussions were held to achieve a final result. For patients from the Ophthalmology Department, the original diagnosis reports with fluorescein angiography were also used to facilitate the discussions. The fundus images in our dataset were labeled with one of four degrees: Normal, Moderate, Heavy, and Severe. The relationship between these degrees and those degrees in [4] is listed in TABLE 1, as is the corresponding treatment suggestions for each degree. The annotation criteria that we used differed slightly from that of other datasets. There are two major reasons for this choice. The first is that, for ophthalmologists, the progression from severe NPDR to mild PDR is sometimes obscure or even unrecognizable from fundus images. If necessary, a fluorescein angiography is proper for a more precise diagnosis. The second is that the suggested treatments are same for mild and moderate NPDR patients or severe NPDR and mild PDR patients. Thus, it is convenient to group them together for clinical practice. In FIGURE. 3, we show sample images with the corresponding labels from our annotation.



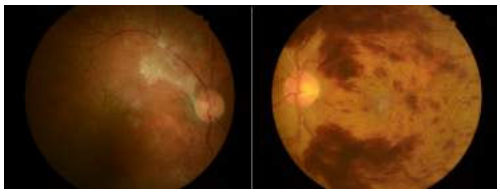
(a) Normal



(b) Moderate



(c) Heavy



(d) Severe

FIGURE 3: Sample fundus images and corresponding labels

V. DATA PREPROCESSING AND AUGMENTATION

A. PREPROCESSING

One problem that we face is the variety of fundus image capture devices; because the techniques that are required for the manufacture of such devices are mature and open, various companies have developed dozens of such products, each of which generates fundus images to a particular standard. To ensure that machines learn the true features of DR rather than device-specific information, we need to process the acquired fundus images from different sources and change them into a uniform format through the following steps:

1) Size normalization

The first step is to resize different images into a uniform scale so that all fundus areas in different images have the same diameter. The black borders on each side of the fundus image are removed at the outset by summing the images horizontally and vertically and discarding regions that correspond to values under a selected threshold. Then, the images are resized to fixed dimensions.

2169-3536 (c) 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2888639, IEEE Access

TABLE 1: The annotation scheme adopted

Severity Level	Observable findings on fundus images	Corresponding category in [4]	Suggested treatment
Normal	No abnormalities	Normal	Recheck in 12 months
Moderate	Microaneurysm, exudation, but less than Heavy	Mild or moderate NPDR	Recheck in 6 months
Heavy	Any of the following: 1. More than 20 intraretinal hemorrhages in each of 4 quadrants 2. Definite venous beading in 2+ quadrants 3. Prominent IRMA in 1+ quadrant 4. PDR but less than Severe	Severe NPDR or mild PDR	Scatter laser treatment
Severe	One or more of the following: Pre-macular hemorrhage, vitreous hemorrhage, severe retinal proliferative	Severe PDR	Laser or vitrectomy treatment

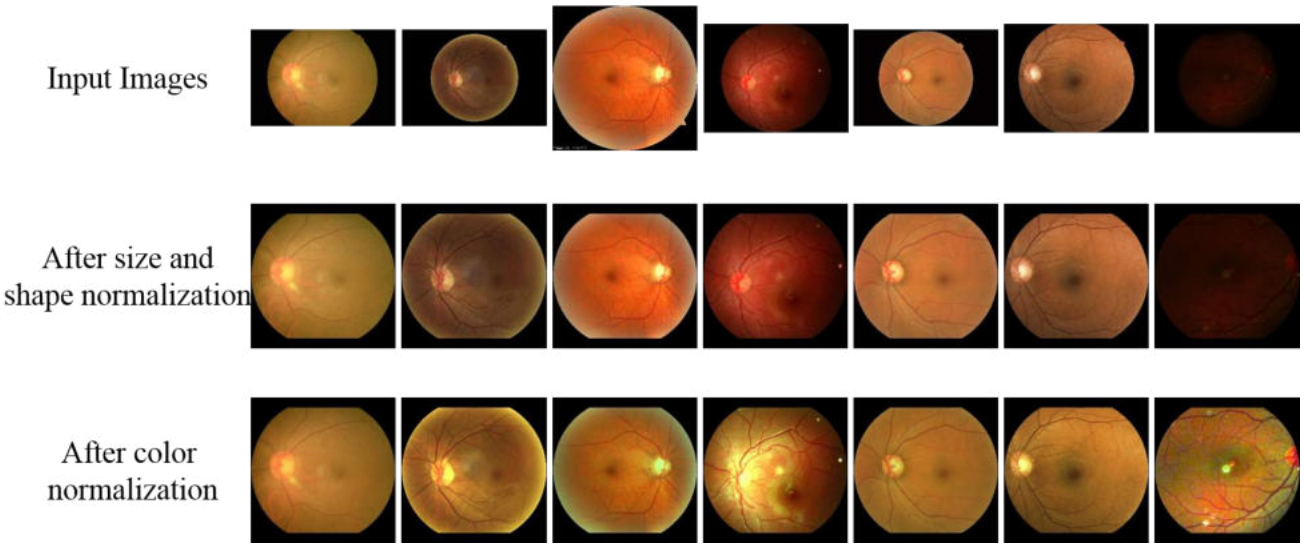


FIGURE 4: The preprocessing pipeline

2) Shape normalization

Some fundus images are complete circles, whereas others may lack the top and bottom margins. In addition, many devices capture a small notch on the edge of the circle. To unify the shapes of these images, we use a mask that contains the largest common area of all images from different sources to obscure unwanted parts of the image.

3) Color normalization

After the shape of each image is normalized, its color must be tuned because different devices may produce images with different color temperatures, and the illumination conditions can vary. Our method of color tuning is simple: we shift each of the RGB channels of a fundus image to a pre-calculated mean and truncate the values above 255, as follows:

$$\begin{cases} R_i = min\{\frac{R_i}{mean(R)} \cdot r^*, 255\} \\ G_i = min\{\frac{G_i}{mean(G)} \cdot g^*, 255\} \\ B_i = min\{\frac{B_i}{mean(B)} \cdot b^*, 255\}, \end{cases} \quad (1)$$

where R, G, B represent the RGB channels of the fundus image being processed, R_i, G_i, B_i represent each pixel value in the corresponding channel, and r^*, g^*, b^* are mean values of the RGB channels calculated from over 1000 fundus images that have been captured well with illumination.

This color normalization method is direct and effective, for example, the rightmost fundus image in FIGURE. 4 was enhanced significantly after color normalization. Another advantage of this method is that it ensures that the total input values of each channel are approximately the same for different fundus images, thus providing the learning model with a more stable input range, benefiting such models as neural networks.

The overall pipeline for preprocessing is shown in FIGURE. 4, illustrated with several sample images throughout the pipeline.

B. AUGMENTATION

Having sufficient training data is the key for training a neural network successfully; unfortunately, this requirement is seldom satisfied in most neural network applications. For medical imaging applications, the lack of data is more significant because of the cost of the annotations, and because of the imbalance in the occurrence between diseases. To mitigate shortages in data and fully utilize the data that are available, certain data augmentation techniques must be carried out; in our experiment, we used the Augmentator software package [27]. Specifically, we augmented our data through the following means:

- flip the image horizontally

2169-3536 (c) 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2888639, IEEE Access

- flip the image vertically
- randomly rotate the image in the range of [-25,25] degrees
- randomly zoom in or out in the range of [0.85,1.15]
- randomly distort the image

All of these methods were combined for augmenting each image, and a probability of 0.5 was used to determine whether or not to perform each of them.

VI. MODEL

Neural networks are computational models that are formed by connecting simple computational units, called neurons, in certain patterns. They can, in principle, mimic behaviors of any given function if there are enough neurons. CNNs are a special type of neural networks that were proposed by LeCun in 1990 [28]. Because of their superior performance on image-oriented tasks, they are now the mainstream model for image-related tasks. Generally, a CNN contains three basic components: the convolutional layers, the in-place activation operation, and the pooling layers. For classification tasks, there may be several fully connected layers and a classification layer at the end.

Formally, given an input $X \in \Re^{c\times m\times n}$, which is commonly an image or a feature map of c channels and $m \times n$ size, a convolutional layer computes a function f of X, which gives a 2-D matrix X' such that

$$X' = f(X) = b + \sum_{j=0}^{c} W_j * X_j, \tag{2}$$

where $W \in \Re^{c\times K\times K}$ is the convolution kernel of shape $c \times K \times K$, b is the bias term, and $*$ denotes a 2-D cross-correlation operator. A max pooling layer performs down-sampling on a given input image or feature map by dividing the input into small patches in a sliding window manner and computing the maximum of each patch. The max operation can be replaced by the average or min operation, giving different types of pooling layers. For the in-place activation operation, a non-linear function is used; a common choice for CNNs is the Rectified Linear Unit (ReLU) function:

$$ReLU(x) = \begin{cases} 0 & \text{if } x < 0, \\ x & \text{otherwise.} \end{cases} \tag{3}$$

For the convolutional layer and pooling layer, the sliding window on the input can slide in an overlapping or non-overlapping manner, but in common practice, convolutional layers are overlapping and pooling layers are non-overlapping. By interlacing convolutional layers and pooling layers, a CNN of arbitrary depth can be built. But, it is generally difficult to find a structure that works optimally for a given task.

Various CNN structures have been proposed, starting from Lenet in 1998 [29]. Important breakthroughs in network structures include the Alex network [30], the VGG network [31], the ResNet [32], and the Inception network [33]. Many variants have been derived from each of these typical structures. The Inception network and its derivatives are among the best image classification models and are widely used in different applications. We use the Inception-V3 network as our base model. The detailed structure of this network is given in TABLE 2. The parameters for each layer are denoted (kernel size, stride, number of channels) for convolutional layers and (type, kernel size, stride) for pooling layers in the Layer Configuration column. The shape of the data that pass through the network is denoted (channels, width, height) in the Data Shape column. All convolution layers in the network are followed by a batch norm and an activation calculation and are thus omitted from the table. The Inception A/B/C/D/E modules are blocks that combine convolution layers with different kernel sizes. When using this base model directly, the output units in the last softmax output layer are reduced to four as we only have four degrees of DR severity.

We notice that some microaneurysms and exudates can be really small in their size. So, when a fundus image is resized to a size of 299 by 299 pixels (the size commonly used by mainstream CNN models for image classification), small lesions in the image may be decreased to one or two pixels, making them harder to be detected. To mitigate this problem, we further propose a model for processing larger size fundus images. Specifically, in the first step of preprocessing, we resize the fundus images to 600 by 600 pixels, then we cut each image into four 300 by 300 pieces and feed these pieces into four different Inception-V3 models. In the end, we concatenate the features of the global average pooling layers from the four Inception-V3 models into one vector and feed this vector to a softmax output layer. We will denote this model as Inception@4 in the rest of this paper, and the structure of this model is illustrated in FIGURE. 5. Since the area of the image is enlarged four times, it is easier to detect small lesions in this larger image. Comparing to applying an Inception-V3 model directly on a 600 by 600 image, there are three advantages of this strategy. Firstly, more free parameters can give the model better capacity. Secondly, the four models can focus on different locations of a fundus image and learn location related lesions such as diabetic macular edema. Thirdly, since the original Inception-V3 model is designed for images of 299 by 299 pixels, when the input image is enlarged too much, the feature maps inside the model will be enlarged too, thus the depth of the model may no longer be optimal. In the experimental results part, we can see that this strategy improves the grading accuracy and the recall rate compared to the original Inception-V3 model.

As mentioned before, the data we have at hand is very limited comparing to the tremendous amount of model parameters. To mitigate this problem, we have already used several data augmentation techniques. Another popular technique aiming at data shortage is transfer learning, i.e. transfer knowledge from domains with sufficient data to domains with insufficient data. There are many ways for doing transfer learning. For example, one can map the data from different domains to one common space and use the mapped samples as training data, or use the samples from different domains by

2169-3536 (c) 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2888639, IEEE Access

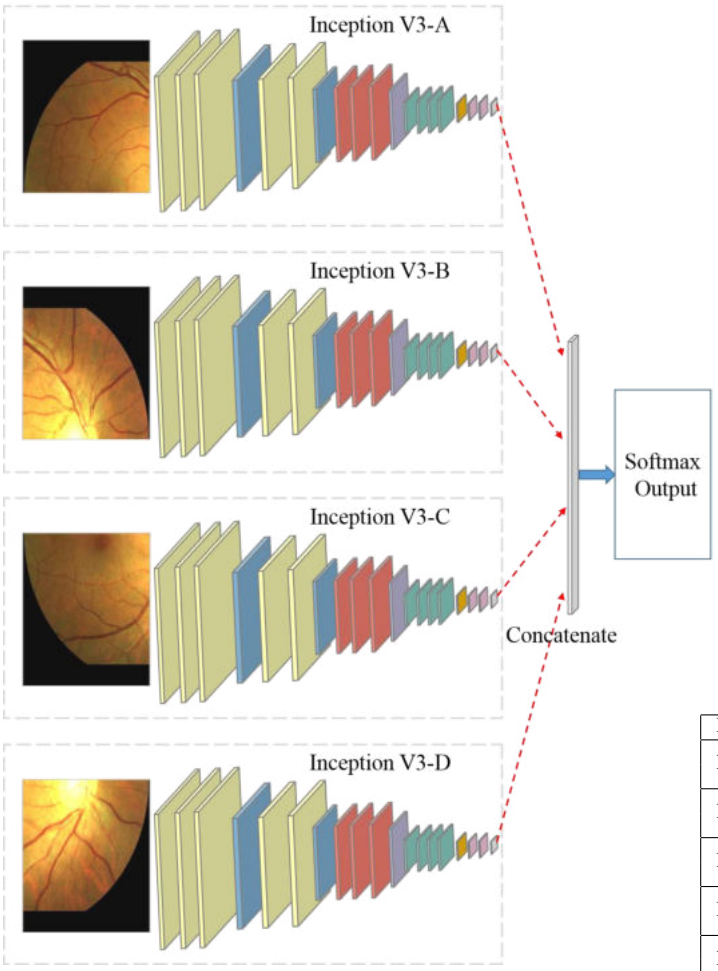


FIGURE 5: The structure of Inception@4

TABLE 2: Network structure of Inception-V3 network

Layer Index	Layer Type	Layer Configuration	Data Shape
0	Data	-	(3,299,299)
1	Convolution	(3x3,2x2,32)	(32,149,149)
2	Convolution	(3x3,1x1,32)	(32,147,147)
3	Convolution	(3x3,1x1,64)	(64,147,147)
4	Pooling	(max,3x3,2x2)	(64,73,73)
5	Convolution	(1x1,1x1,80)	(80,73,73)
6	Convolution	(3x3/1x1,192)	(192,71,71)
7	Pooling	(max,3x3,2x2)	(192,35,35)
8	Inception A	-	(256,35,35)
9	Inception A	-	(288,35,35)
10	Inception A	-	(288,35,35)
11	Inception B	-	(768,17,17)
12	Inception C	-	(768,17,17)
13	Inception C	-	(768,17,17)
14	Inception C	-	(768,17,17)
15	Inception C	-	(768,17,17)
16	Inception D	-	(1280,8,8)
17	Inception E	-	(2048,8,8)
18	Inception E	-	(2048,8,8)
19	Pooling	(avg,8x8,1x1)	(2048,1,1)
20	Softmax Output	-	4

TABLE 3: Training/Test set statistics

Dataset	Set	Normal	Moderate	Heavy	Severe	Total
Dataset 1	Training	1513	760	975	321	3569
	Test	385	190	247	85	907
Dataset 2	Training	1518	758	970	322	3568
	Test	380	192	252	84	908
Dataset 3	Training	1515	759	964	320	3558
	Test	383	191	258	86	918
Dataset 4	Training	1515	759	965	327	3566
	Test	383	191	257	79	910
Dataset 5	Training	1518	760	976	318	3572
	Test	380	190	246	88	904

assigning different weights. However, the most popular and ready to use technique is to use parameters of a model trained on problems that have sufficient data for model initialization. The purpose of this initialization is to provide the model a good starting point and avoid bad local minima. In our work, we take this approach, the parameters of models that have been pre-trained on the ImageNet [34] dataset are used for initialization of the networks.

VII. EXPERIMENTS

A. EXPERIMENTAL SETUP

We separate our data randomly into a training set and a test set at a ratio of 4:1; fundus images that are captured on each eye (one or two images per eye) are treated as a single sample. Random separation is carried out five times independently to prevent poor randomization. This results in five pairs of training and test sets. Because we separate the data by eye rather than image, the number of samples in each category varies slightly in different training/test set pairs. The statistics of each training/test set pair are listed in TABLE 3.

We evaluate the Inception-V3 and Inception@4 models on each of these dataset pairs; several mainstream models, such as the Resnet model and the VGG model, are evaluated as a comparison. The evaluation metrics we use are as follows. First, we calculate the 4-classification accuracy of the models that are evaluated. Because the rate of missed diagnoses is important for clinical applications, we further calculate the precision and recall rate of the models, for the calculation of these two metrics, the three degrees besides normal are considered as a with DR degree. The precision of a model indicates the reliability of the model diagnosing a patient as ill, and the recall rate reflects the sensitivity of a model to the disease being diagnosed. These metrics are calculated as

$$P = \frac{TP}{TP + FP} \tag{4}$$

and

$$R = \frac{TP}{FN + TP}, \tag{5}$$

where TP, FP, and FN denote True Positive, False Positive, and False Negative, respectively.

All of these evaluations are carried out independently. The platform we use to conduct the experiments is a workstation with 2 Xeon E5-2620 CPUs, 2 Tesla K40 GPUs, and 64G of RAM. The network implementation is done using the MXNET package [35].

2169-3536 (c) 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2888639, IEEE Access

B. DETAILS OF LEARNING

For training, we use a batch size of 32, for the Inception@4 model the batch size is reduced to 4 since it requires more GPU RAM, the optimizer we use is an Adam optimizer with a learning rate of 1e-05, weight decay rate is set as 0.2 to prevent overfitting, other parameters of the optimizer are set using the default values in MXNET, i.e. beta1=0.9, beta2=0.999, epsilon=1e-08.

All parameters in these models are initialized using models pre-trained on the ImageNet dataset. For parameters not in the pre-trained models i.e. the parameters of the output layers, we use a Xavier initializer with magnitude of 1 for initialization.

During training we found that the I/O operation for reading the images from the disk takes a lot of time. In order to reduce this time, we set up a separate process to cache all the preprocessed images in the RAM, the training process can then retrieve each minibatch of data from the caching process. This strategy boosts the training significantly, and is particularly useful when a number of models need to be evaluated simultaneously for many times.

C. RESULTS

The evaluation results (averaged over five datasets) for each model are listed in TABLE 4. The Inception-V3 model and the adapted Inception@4 model surpass all other models. It is notable that by classifying more Normal samples(together with samples with Mild DR) as Mild DR, the recall rate of VGG-19 is 0.6% higher than that of Inception-V3, however this is at the cost of 2.51% of precision and 2.85% overall accuracy. By a similar reason, the Inception@4 model has a lower precision rate than the Inception-V3 model, however, the increase in accuracy and recall is more significant and preferable.

TABLE 4: The results of the models evaluated

	Accuracy	Precision	Recall
Resnet-18	87.61%	95.76%	92.52%
Resnet-101	87.26%	95.63%	92.48%
VGG-19	85.50%	94.00%	93.01%
Inception-V3	88.35%	**96.51%**	92.41%
Inception@4	**88.72%**	95.77%	**94.84%**

D. VISUALIZATION

The success of neural network models is largely based on their revolutionary performance and simplicity in the model setup. The only requirement for training a neural network is a dataset of input-output pairs; no feature engineering is required, and the network will automatically learn the mapping between inputs and outputs. This black box character renders neural networks easy to use; as a trade-off, the explainability of neural networks is poor. Much effort has been made to demystify how neural networks function. For the CNN models that were used in the image classification problem, many techniques have been proposed to visualize what the model really learns; such approaches include the occlusion test, feature visualization, deconvolutional networks [36], and classification activation map (CAM) [37]. Here, we used the CAM method for the visualization analysis. A CAM is obtained by weighting the feature maps before the global average pooling layer at the end of the network. By projecting the weights of the output layer onto the feature maps, the contribution of each region that is activated in these feature maps can be combined to indicate the regions in the input that count toward the prediction.

In FIGURE. 6, we highlight representative results that were obtained using CAM on the trained Inception@4 model. For the CAM and Combined image columns, four patches from the four sub-models are combined. We can see clearly that for these images, the model learns to focus on the main lesions during the classification process.

To further analyze the performance of the trained models, we plotted the confusion matrices of the Inception@4 model on five test sets in FIGURE. 7. From the confusion matrices, we can see that most of the misclassifications lay between adjacent categories, indicating that such samples are hard to separate for the model. We conjecture that this is because the progression of DR is continuous; thus, the category to which a patient belongs in certain stages is ambiguous. One important issue we must consider is the impact of false negative diagnoses. From the confusion matrices we can see that all false negative diagnoses are the ones that should be graded as Moderate. From TABLE 1 we can see that the suggested treatment for a Moderate DR patient is recheck in 6 months, and the suggested treatment for a diabetic patient without DR is recheck in 12 months. Thus, for the false negative diagnoses, delayed treatment of 6 months can be caused. From the confusion matrices we can estimate that the rate of Moderate DR be graded as Normal is 14.25%, we can also see that the probability of Moderate DR change into high risk PDR within one year is at most 8.1% [4]. Thus, the impact of such false negative diagnoses is limited.

VIII. MODEL DEPLOYMENT AND CLINICAL EVALUATION

To meet the requirements of clinical applications, we have built a system that diagnoses DR via the internet. The models[1] are first deployed on a cloud computing platform to compute the diagnosis; then, a web server is deployed to wrap the models and provide a user interface to users from different hospitals via the internet. The pilot runs are carried out in four hospitals in different locations. During the pilot runs, users from these hospitals upload images to the system and get diagnosis results from the system. In the backstage, ophthalmologists from the annotation team login to the system and retrieve all uploaded cases. Then, all uploaded cases are annotated again by the ophthalmologists. The annotation results are then treated as the ground truth to judge the performance of the system. Thus far, the diag-

[1]Currently, we ensemble five Inception-V3 models on the cloud rather than five Inception@4 models for the sake of saving GPU RAM.

2169-3536 (c) 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2888639, IEEE Access

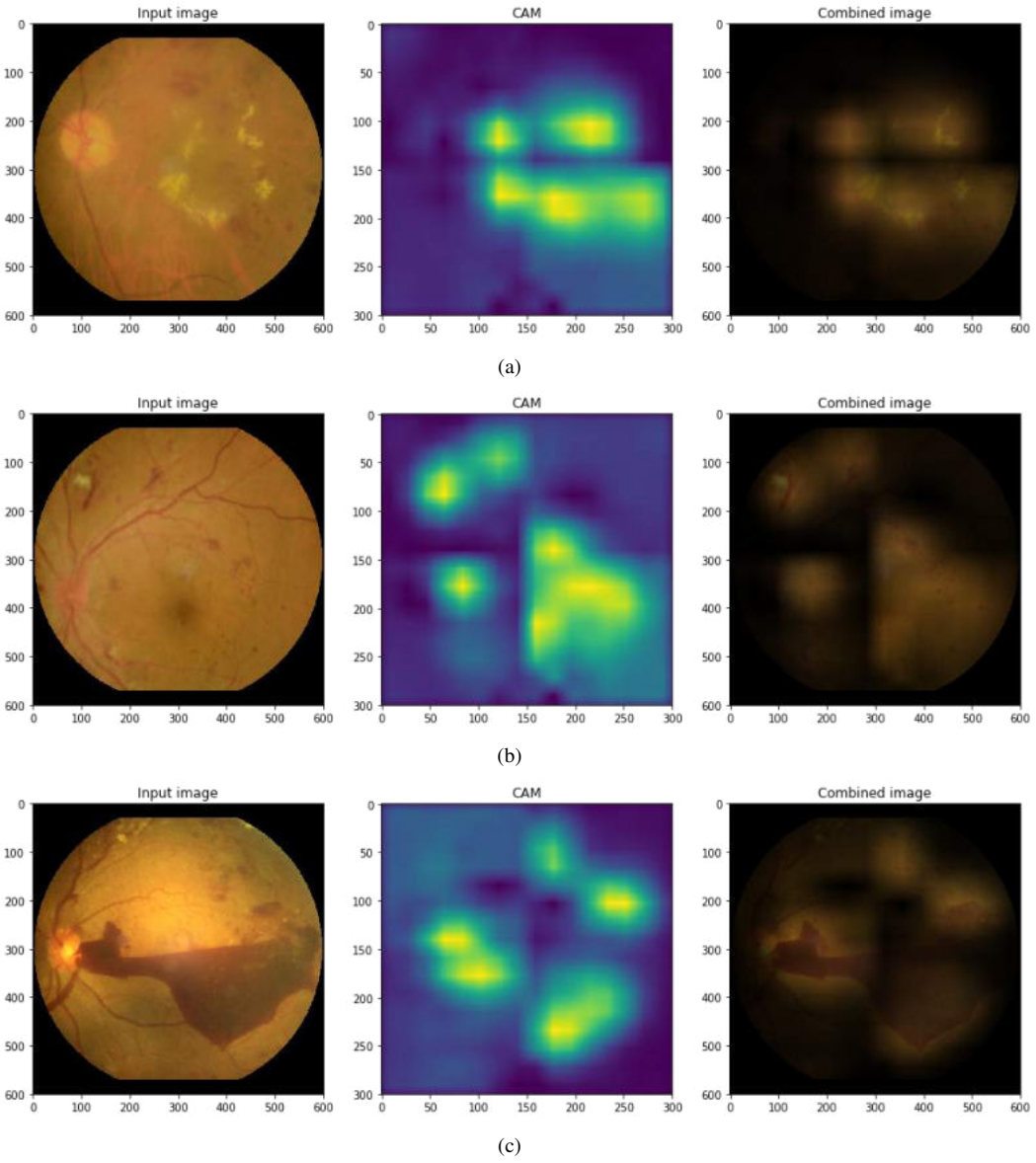


FIGURE 6: Typical results for CAM on the trained Inception@4 model

nose consistency between the system and ophthalmologists reaches 91.8% (259 out of 282 valid diagnoses, all testing cases new to the models), demonstrating the reliable performance of the system in a clinical environment. Similar to results in the experimental section, all false negative cases are cases that should be graded as Moderate which require no urgent treatments.

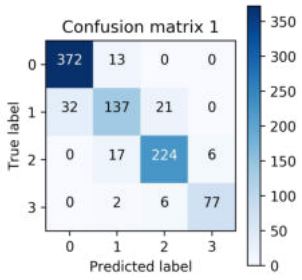
IX. CONCLUSION

The huge population of diabetic patients and the prevalence of DR among them have fostered a great demand in automatic DR diagnosing systems. So far, a lot of achievements have been made and satisfactory results have been achieved in many sub problems like vessel segmentation, lesion detection. However, these results are obtained on datasets relatively small and are steps away from real world applications.

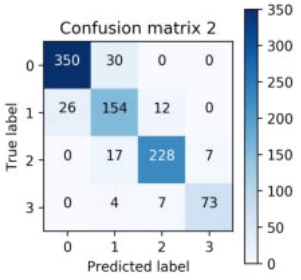
For clinical application, systems that can give DR severity directly are more favorable and practical. However, current results for multi-class severity grading are still not good enough for clinical application. In this work, we investigated the automatic grading of DR using deep neural networks. We proposed a novel dataset that is moderate in size and annotated with a new labeling scheme that is more useful for clinical practice. We proposed a preprocessing pipeline to change fundus images into a uniform format. We used the Inception-V3 network and a proposed modification of it as our diagnostic models and evaluated the performance of them with several mainstream CNN models. The experimental results demonstrate the efficiency of the models in diagnosing DR. Visualization and analysis of the trained models provide

2169-3536 (c) 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

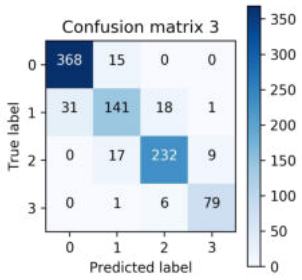
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2888639, IEEE Access



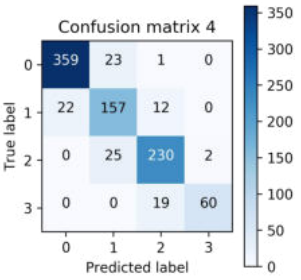
(a) Confusion matrix on Dataset 1



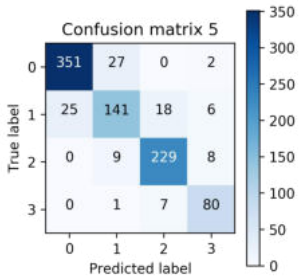
(b) Confusion matrix on Dataset 2



(c) Confusion matrix on Dataset 3



(d) Confusion matrix on Dataset 4



(e) Confusion matrix on Dataset 5

FIGURE 7: The confusion matrix of Inception@4 network on each test set

insights into how the models make diagnoses using given fundus images and justify the diagnostic ability of the models from a different viewpoint. For clinical applications, the trained models are deployed on a cloud computing platform and provide pilot diagnostic services to several hospitals via the internet. The performance of the system in the clinical evaluation demonstrates the efficiency of this work. In the future, data from more equipments will be included, and a broader pilot study will be launched. The accumulated data will be further used to improve the accuracy of the models.

REFERENCES

[1] F. DS, A. L, G. TW, K. GL, B. G, C. JD, K. R, and A. D. Association, "Diabetic retinopathy," Diabetes care, vol. 26, no. 1, pp. 226–229, 2003.

[2] W. H. Organization et al., Global report on diabetes. World Health Organization, 2016.

[3] T. Y. W. Ning Cheung, Paul Mitchell, "Diabetic retinopathy," Lancet, vol. 376, no. 9735, pp. 124–136, 2010.

[4] D. Ophthalmoscopy and E. Levels, "International clinical diabetic retinopathy disease severity scale detailed table," 2002.

[5] S. Resnikoff, W. Felch, T.-M. Gauthier, and B. Spivey, "The number of ophthalmologists in practice and training worldwide: a growing gap despite more than 200 000 practitioners," British Journal of Ophthalmology, vol. 96, no. 6, pp. 783–787, 2012. [Online]. Available: http://bjo.bmj.com/content/96/6/783

[6] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: Face recognition with very deep neural networks," arXiv preprint arXiv:1502.00873, 2015.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1026–1034.

[8] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot et al., "Mastering the game of go with deep neural networks and tree search," nature, vol. 529, no. 7587, p. 484, 2016.

[9] T. Kauppi, V. Kalesnykiene, J.-K. Kamarainen, L. Lensu, I. Sorri, H. Uusitalo, H. Kälviäinen, and J. Pietilä, "Diaretdb0: Evaluation database and methodology for diabetic retinopathy algorithms," Machine Vision and Pattern Recognition Research Group, Lappeenranta University of Technology, Finland, vol. 73, 2006.

[10] T. Kauppi, V. Kalesnykiene, J.-K. Kamarainen, L. Lensu, and A. R. Sorri, Iiris, "Diaretdb1 diabetic retinopathy database and evaluation protocol," in Medical Image Understanding and Analysis, vol. 2007. Citeseer, 2007, p. 61.

[11] E. DecenciÍĺre, 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 Analysis and Stereology, vol. 33, no. 3, pp. 231–234, 2014. [Online]. Available: https://www.ias-iss.org/ojs/IAS/article/view/1155

[12] J. Staal, M. D. Abramoff, M. Niemeijer, M. A. Viergever, and B. van Ginneken, "Ridge-based vessel segmentation in color images of the retina," IEEE Transactions on Medical Imaging, vol. 23, no. 4, pp. 501–509, April 2004.

[13] A. D. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," IEEE Transactions on Medical Imaging, vol. 19, no. 3, pp. 203–210, March 2000.

[14] B. Al-Diri, A. Hunter, D. Steel, M. Habib, T. Hudaib, and S. Berry, "Review - a reference data set for retinal vessel profiles," in 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Aug 2008, pp. 2262–2265.

[15] "Diabetic retinopathy detection," https://www.kaggle.com/c/diabetic-retinopathy-detection/data.

[16] E. DecenciÍĺre, 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. La?, and A. Chabouis, "Teleophta: Machine learning and image processing methods for teleophthalmology," IRBM, vol. 4161, no. 2, pp. 91–203, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1959-0318(13)00023-7

[17] S. S. Rubini and D. A. Kunthavai, "Diabetic retinopathy detection based on eigenvalues of the hessian matrix," Procedia Computer Science, vol. 47, pp. 311–318, 2015.

[18] M. Mookiah, U. R. Acharya, R. J. Martis, C. K. Chua, C. Lim, E. Ng, and A. Laude, "Evolutionary algorithm based classifier parameter tuning for automatic diabetic retinopathy grading: A hybrid feature extraction approach," Knowledge-Based Systems, vol. 39, pp. 9 – 22, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950705112002651

[19] A. P. Bhatkar and G. U. Kharat, "Detection of diabetic retinopathy in retinal images using mlp classifier," in 2015 IEEE International Symposium on Nanoelectronic and Information Systems, Dec 2015, pp. 331–335.

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

[21] G. Quellec, K. Charrière, Y. Boudi, B. Cochener, and M. Lamard, "Deep image mining for diabetic retinopathy screening," Medical image analysis, vol. 39, pp. 178–193, 2017.

[22] Y. Yang, T. Li, W. Li, H. Wu, W. Fan, and W. Zhang, "Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks," in International Conference on Medical Image Computing and Computer-Assisted Intervention, vol. 10435. Springer, Cham, 2017, pp. 533–540.

2169-3536 (c) 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2888639, IEEE Access

[23] V. Chandore and S. Asati., "Automatic detection of diabetic retinopathy using deep convolutional neural network," International Journal of Advance Research, Ideas and Innovations in Technology, vol. 3, pp. 633–641, 2017.
[24] S. Wang, Y. Yin, G. Cao, B. Wei, Y. Zheng, and G. Yang, "Hierarchical retinal blood vessel segmentation based on feature and ensemble learning," Neurocomputing, vol. 149, Part B, pp. 708–717, 2015.
[25] P. Adarsh and D. Jeyakumari, "Multiclass svm-based automated diagnosis of diabetic retinopathy," in 2013 International Conference on Communication and Signal Processing, April 2013, pp. 206–210.
[26] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional neural networks for diabetic retinopathy," Procedia Computer Science, vol. 90, pp. 200–205, 2016.
[27] M. D. Bloice, C. Stocker, and A. Holzinger, "Augmentor: An image augmentation library for machine learning," arXiv preprint arXiv:1708.04680, 2017.
[28] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," in Advances in Neural Information Processing Systems 2, D. S. Touretzky, Ed. Morgan-Kaufmann, 1990, pp. 396–404. [Online]. Available: http://papers.nips.cc/paper/293-handwritten-digit-recognition-with-a-back-propagation-network.pdf
[29] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, Nov 1998.
[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf
[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016, pp. 770–778.
[33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015, pp. 1–9.
[34] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, June 2009, pp. 248–255.
[35] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang, "Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems," arXiv preprint arXiv:1512.01274, 2015.
[36] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in Computer Vision – ECCV 2014, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 818–833.
[37] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016, pp. 2921–2929.

2169-3536 (c) 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.