Knowledge-Based Systems 175 (2019) 12–25

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Automated identification and grading system of diabetic retinopathy using deep neural networks

Wei Zhang [a], Jie Zhong [b,*], Shijun Yang [b], Zhentao Gao [a], Junjie Hu [a], Yuanyuan Chen [a], Zhang Yi [a]

[a] Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu 610065, China
[b] The Department of Ophthalmology, Sichuan Academy of Medical Sciences and Sichuan Provincial Peoples Hospital, Chengdu 610072, China

HIGHLIGHTS

- We established a high-quality labelled dataset of DR medical images.
- We developed a novel and well-performing DR recognition and classification system.
- The optimal combination of ensemble model was explored experimentally.

ARTICLE INFO

Article history:
Received 31 October 2018
Received in revised form 14 March 2019
Accepted 15 March 2019
Available online 22 March 2019

Keywords:
Deep learning
Diabetic retinopathy
Ensemble learning
Fundus images
Image classification
Transfer learning

ABSTRACT

Diabetic retinopathy (DR) is a major cause of human vision loss worldwide. Slowing down the progress of the disease requires early screening. However, the clinical diagnosis of DR presents a considerable challenge in low-resource settings where few ophthalmologists are available to care for all patients with diabetes. In this study, an automated DR identification and grading system called DeepDR is proposed. DeepDR directly detects the presence and severity of DR from fundus images via transfer learning and ensemble learning. It comprises a set of state-of-the-art neural networks based on combinations of popular convolutional neural networks and customised standard deep neural networks. The DeepDR system is developed by constructing a high-quality dataset of DR medical images and then labelled by clinical ophthalmologists. We further explore the relationship between the number of ideal component classifiers and the number of class labels, as well as the effects of different combinations of component classifiers on the best integration performance to construct an optimal model. We evaluate the models on the basis of validity and reliability using nine metrics. Results show that the identification model performs best with a sensitivity of 97.5%, a specificity of 97.7% and an area under the curve of 97.7%. Meanwhile, the grading model achieves a sensitivity of 98.1% and a specificity of 98.9%. On the basis of the methods above, DeepDR can detect DR satisfactorily. Experiment results indicate the importance and effectiveness of the ideal number and combinations of component classifiers in relation to model performance. DeepDR provides reproducible and consistent detection results with high sensitivity and specificity instantaneously. Hence, this work provides ophthalmologists with insights into the diagnostic process.

1. Introduction

Diabetic retinopathy (DR) is a chronic complication of diabetes that damages the retina. Notably, the risk of blindness in patients with DR is 25 times that in healthy people; thus, DR is a leading cause of blindness amongst people aged 20–65 years worldwide [1]. The blindness caused by DR can be prevented through regular fundus examinations [2]. A widespread consensus regarding the benefits and cost-effectiveness of screening for DR has been formed amongst western nations [3–5]. Most DR studies use the international clinical disease severity scale to classify DR (Table 1) in accordance with the Early Treatment Diabetic Retinopathy Study (ETDRS). Other details are available in the latest American Association of Ophthalmology Clinical Guidelines: Diabetic Retinopathy (2016 Edition, 2017 Updated) [6]. Nowadays, diabetes screening is common in developed countries;

* Corresponding author.
E-mail addresses: zweiscu@gmail.com (W. Zhang), jiezhong1968maya@163.com (J. Zhong), odalisyang@gmail.com (S. Yang), gaozhentao@stu.scu.edu.cn (Z. Gao), hujunjiescu@gmail.com (J. Hu), chenyuanyuan@scu.edu.cn (Y. Chen), zhangyi@scu.edu.cn (Z. Yi).

https://doi.org/10.1016/j.knosys.2019.03.016
0950-7051/

patients with diabetes are screened from the general population and transferred to DR specialists. Follow-up examinations are performed by these specialists, and medical intervention is implemented when necessary; therefore, the incidence of severe DR in developed countries is low.

However, the situation in China is not as promising. (1) Currently, ophthalmologist-to-patient ratio is nearly 1:1000 in China. (2) The rate of DR screening is less than 10%. (3) The vast majority of patients with DR in China do not know the risks associated with DR; hence, they may not realise they have the disease. Furthermore, a large proportion of diabetic patients ignore this serious complication. Patients with DR in China often undergo late invasive treatment and exhibit serious illness, resulting in poor prognosis and high medical expenses. Therefore, the incidence of severe proliferative DR is much higher in China than in developed countries. Moreover, the eventual blindness from DR is irreversible, thereby placing a heavy burden on Chinese families and the society. The automatic screening and grading of DR is a pressing demand because it can help to solve the abovementioned problems.

Deep neural networks (DNNs), also called deep learning by brain-inspired systems [7,8], can automatically learn numerous abstract high-level features or representations of attribute categories directly from original big data to ascertain a distributed representation of data. A widely used type of DNN is the recurrent neural network, which has shown unprecedented success in academia and industries, including in the areas of speech recognition and machine translation [9–11]. With regard to the characteristics of the spatial coherence of images, convolutional neural networks (CNNs) are preferred because they are highly specialised in views for image recognition, analysis and classification [12]. In recent years, CNNs have also provided insights into various medical studies. Furthermore, they have abilities rivalling those of medical experts [13], especially when applied to skin cancer [14] and breast cancer classification [15] and lung cancer [16] and retinopathy of prematurity detection [17,18].

Nevertheless, challenges remain in the use of CNNs in medical studies. First, sufficient real-world medical images, especially those for some specialised diseases, are difficult to obtain. Furthermore, the availability of labelled medical data is typically limited. Second, DR features are so complex that they are likely to cross-effect with various other lesions, and the minute lesions of DR cannot be detected if images quality is poor. According to medical journals, fundus photographs are labelled by a manual operation process, which is prone to subjectivity. Third, high disease-detection accuracy is difficult to attain effectively by training a single model with a limited scale of medical image data and inevitable image noise. Therefore, two important strategies are used in deep learning: transfer learning [19,20] and ensemble learning [21–23]. The primary concept of transfer learning is knowledge reuse: the migration of big data to small data fields to resolve the problem of data and knowledge scarcity in small data fields. The major conception of classifier ensemble learning is the combination of a series of component classifiers with different learning preferences to resolve the same predictive problem. These ensemble methods enable increased generalisation that outperforms that of each individual component.

In the current work, we developed an automated system called DeepDR for DR screening via deep learning. DeepDR is a complex process composed of three steps: judgment of the existence of retinal lesion characteristics via screening of fundus photographs, evaluation of the severity of DR if lesion features are detected and reporting of the detection of clinical DR. Thus far, DeepDR has been used in some local hospitals to aid primary hospitals in remote areas or clinical communities that lack retinal specialists or appropriate equipment.

Table 1
International clinical DR disease severity scale.

Severity level	Findings observable upon dilated ophthalmoscopy
No DR	No abnormalities.
Mild DR	Microaneurysms only.
Moderate DR	More than just microaneurysms but less than severe NPDR.
Severe DR	Any of the following and no signs of proliferative retinopathy: (1) severe intraretinal hemorrhages and microaneurysms in each of four quadrants; (2) definite venous beading in two or more quadrants; (3) prominent IRMA in one or more quadrants.
Proliferative DR	One or both of the following: (1) neovascularisation; (2) vitreous/preretinal hemorrhage.

IRMA = intraretinal microvascular abnormalities; NPDR = nonproliferative diabetic retinopathy; PDR = proliferative diabetic retinopathy Note :(1) Any patient with two or more of the characteristics of severe NPDR is considered to have very severe NPDR. (2) PDR may be classified as high-risk and non-high-risk.

Table 2
The grading scale of DR in the study.

Grade	Disease severity level
NORMAL	No DR
NPDR	Mild NPDR to moderate NPDR
NPDR2PDR	Severe NPDR to non-high-risk PDR
PDR	High-risk PDR

The contributions of this work are as follows.

(1) We establish a high-quality labelled dataset of DR medical images.

(2) We develop a novel DR identification and grading system. The system performs well in comparison with human evaluation metrics.

(3) We explore the relationship between the number of ideal component classifiers and the number of class labels. Furthermore, the effects of different combination methods on the best integration performance are discussed.

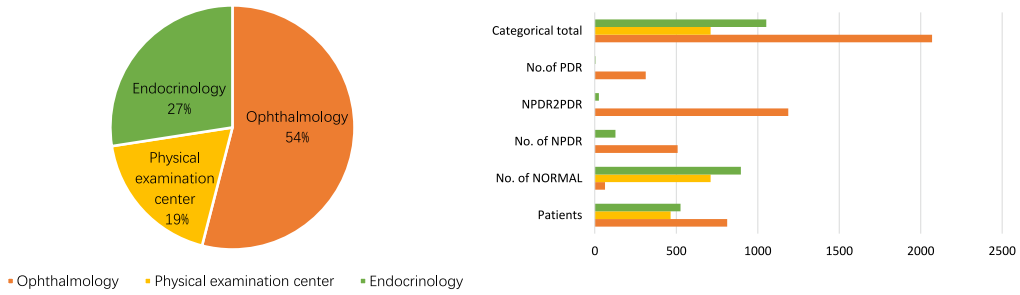
In Section 2, we analyse the related works. In Section 3, we detail the dataset. In Section 4, we describe two novel ensemble models for the two respective tasks. In Section 5, we show the experiments on the two tasks. In Section 6, we provide a discussion of the entire study and future work. Finally, In Section 7, we draw the conclusions.

2. Related works

In the past few decades, the development of automated DR pathology screening has made encouraging progress. From an application perspective, computer-aided detection (CADe) algorithms and computer-aided diagnosis (CADx) algorithms can be viewed as typical representatives in the field. CADe detects lesions at the pixel level with manual segmentations [24]. On the basis of the detected lesions, CADx detects pathologies at the image level [25].

2.1. Traditional methods for DR diagnosis

From a methodological perspective, automated screening for DR has long focused on pattern recognition or traditional machine learning algorithms. Walter et al. [26] created efficient algorithms for detecting optic disc and exudates; these algorithms yielded a mean sensitivity of 92.8% and a mean predictive value of 92.4% on 30 images. Niemeijer et al. [27] developed a machine learning system capable of detecting exudates, cotton wool spots and drusen; their system can differentiate amongst 300 colour images, and its reporting performance approaches that of retinal



(a) source distribution (b) baseline data distribution

Fig. 1. Characteristics of the dataset for DR.

experts. Faust et al. [28] contributed an important and valuable review of algorithms used for extractions of these features from retinopathy images; they also discussed some reported classification systems. Akram et al. [29] presented a hybrid classifier comprising classifiers that could be used to detect all types of non-proliferative DR (NPDR) lesions and grade different stages of NPDR on the basis of a Gaussian mixture model and m-Mediods; whereas the area under the curve (AUC) values of m-Mediods and Gaussian mixture model are 97.7% and 96.3%, respectively, that of the hybrid classifier achieves reaches 98.1%. Gowda et al. [30] employed the fuzzy c-means clustering method on a dataset consisting of 100 images to identify the exact region of DR and reported an accuracy of 99.01%, sensitivity of 98.38% specificity of 96.36%.

Some obvious shortcomings of these techniques include the following. First, they focus on feature engineering, such that the extraction of features must be specified by experts; fulfilling such requirement is a time-consuming process and increases the burden on clinicians. Second, they show limited scope because the studies present results that were derived from small databases; thus, the generalisation and robustness of systems are limited to a certain extent. Third, the early clinical symptoms of DR are not always obvious, and the sizes of some lesions are insufficient to attract the attention of some graders. Hence, questions arise with regard to the accurate extraction of lesion features and comprehensive diagnosis. This approach is thus transformed into a complex vision issue, although this type of error is relatively understandable. Therefore, one of the ultimate goals of artificial intelligence is to automate this process of feature engineering as much as possible.

2.2. Deep learning for DR diagnosis

Deep learning, especially CNNs, provides powerful support to alleviate the aforementioned problems. Models trained by deep learning can discern subtle local features directly from retinopathy images without human effort or specific domain knowledge.

Litjens et al. [31] contributed an important survey regarding the use of deep learning in image classification, object detection, segmentation, registration and other tasks. They summarised over 300 contributions to the field, most of which appeared in 2016. All top algorithms in the Kaggle machine learning competition [32] in 2015 used CNNs to support an automated DR detection system. Benson et al. [33] repurposed an existing deep network for DR screening via transfer learning and other pre-preprocessing techniques on 979 clinical cases; the repurposed deep network achieves a 95% AUC for identifying severe DR with equal sensitivity and specificity of 90%. Gargeya et al. [34] developed a data-driven deep learning algorithm capable of classifying fundus images as healthy or not; this algorithm identifies relevant cases with high reliability. Gulshan et al. [35] created a deep learning-trained algorithm for detecting referable diabetic retinopathy (RDR) in two separate datasets of 9963 and 1748 images; the algorithm achieved high sensitivity (97.5% and 96.1%) and specificity (93.4% and 93.9%) when applied to the two datasets. A study by researchers at Stanford University (Tamkin et al., [36]) used the InceptionV3 model via transfer learning techniques; the approach achieves 72.96% accuracy in detection of RDR and 92.59% accuracy in detection of stage PDR. Pratt et al. [37] designed a CNN network that enables a classifier to predict the exact DR stage of the sample for a five-class DR detection task. The proposed technique achieves 75% accuracy, 30% sensitivity and 95% specificity.

3. Dataset

3.1. Materials

In our study, macula-centred retinal fundus images were taken from the Sichuan Academy of Medical Sciences and Sichuan Provincial Peoples Hospital between September 2017 and May 2018. The original data comprising 13,767 images of 1872 patients were collected from three sources: ophthalmology, endocrinology and physical examination centres (Fig. 1). In general, almost all patients from the ophthalmology department were diagnosed with DR, and nearly two-thirds of the patients from the endocrinology department had DR; the data from the physical examination centre showed no DR symptoms amongst patients. As almost all patients from ophthalmology department had DR, two types of images, namely, retinal colour fundus photographs and fluorescein angiography fundus photographs (pharmacological pupil dilation) were required, and two or more photographs with a 45° view were captured per eye. This project aims to screen preoperative fundus retinal images and to diagnose the degree of lesions. Therefore, all fluorescence contrast images and postoperative fundus images, that is, 9934 images in 1229 patients, were excluded. For the patients from the endocrinology and physical examination centres, only fundus photos of each eye were taken. Further detailed statistics of the dataset are shown in Table 3.

3.2. Grading standard

Although ETDRS (Table 1) has indispensable reference value for our labelling work, some challenges remained in the grading of DR. First, the interpretation for the reference criteria primarily depends on the ophthalmologist's experience of reading images with reference to the guidelines and is thus qualitative rather than quantitative; thus, the assessment of severity has a degree of subjectivity. Second, most patients with DR in China often neglect this disease, and thus fail to secure timely interventions;

Table 3
Summary of training and validation datasets for DR.

Source	Camera	Assessors	Patients	Images	Images/eye	Normal	Npdr	Npdr2pdr	Pdr	Total[a]
Training										
OPH[b]	Canon	(1,2)[c]	613	1669	2∼6[d]	50	421	953	245	1669
PEC[e]	Canon	(1,2)	379	575	1∼2[f]	575	0	0	0	575
END[g]	Kowa	(1,2)	409	818	2	696	100	19	3	818
$Total_t$[h]			1401	3062		1321 (0.43)	521 (0.17)	972 (0.32)	248 (0.08)	3062
Validation										
OPH	Canon	(1,2)	199	401	2∼6	12	87	234	68	401
PEC	Canon	(1,2)	86	136	1∼2	136	0	0	0	136
END	Kowa	(1,2)	117	234	2	200	27	6	1	234
$Total_v$[i]			402	771		348 (0.45)	114 (0.15)	240 (0.31)	69 (0.09)	771
Total[j]			1803	3833		1619	635	1212	317	3833

[a]Total = Total of images in the department.
[b]OPH = Ophthalmology.
[c](1,2) = 2 ophthalmologists, 1 retinal specialist.
[d]2∼6 = Number of fundus photographs taken per eye per patient.
[e]PEC = Physical Examination Centre.
[f]1∼2 = some patients had only 1 images per eye.
[g]END = Endocrinology.
[h]$Total_t$ = Categorical total of training dataset.
[i]$Total_v$ = Categorical total of validation dataset.
[j]Total = Categorical total of training and validation.

as a result, their cases become aggravated, with symptoms being frequently considered as intermediate to severe NPDR and early PDR. Third, because the transition period above has similar clinical manifestations, the recommended treatment for them, according to the ETDRS, is the same. This situation is also reflected in cases ranging from mild NPDR to moderate NPDR [6].

Therefore, we classified severity into four levels: normal, NPDR, NPDR2PDR and PDR. This annotation strategy was used in our four-class classification task. Table 2 shows the specific classifications.

3.3. Manual grading

All images of the dataset were assessed in stages as the data volume accumulated. The graders included one retinal specialist with more than 27 years of experience in DR research and two seasoned ophthalmologists with more than 5 years of experience in clinical diagnosis and treatment. The entire grading process was divided into three steps. First, the annotators indicated whether a given image was of sufficient quality for grading. Second, the quality of the image was deemed insufficient when it became difficult or impossible to make a confident assessment regarding the characteristics of DR. Then, the image was categorised as normal (absence of DR) and abnormal (presence of DR). Third, the severity of DR in the abnormal images was annotated.

The grading reliability of each image was measured by cross-validation (checking others' grading results per image for every grader). First, almost all patients from the ophthalmology department were found with DR; hence, we used two types of images per eye: retinal colour fundus photographs and fluorescein angiography fundus photographs (pharmacological pupil dilation). The fundus images labelled with inconsistency were corrected using the original diagnosis reports with corresponding photographs of fluorescein angiography because accuracy ultimately came from fluorescent photography. Meanwhile, only retinal colour fundus photos were taken from the patients from the endocrine and physical examination centres. Disagreements were re-examined and resolved via discussion. If no consensus was reached, arbitration was performed by the retinal expert to generate final grading.

3.4. Pre-processing of retinal images

Given the complexity of the retina structure, the characteristics of DR were easily confused with those of other eye diseases. Moreover, we observed a range of imaging noise, such as black space on either side of the eye, low contrast, blurred lens or insufficient light. As a result, some minute lesions could not be accurately identified within the poor photographs. Therefore, some pre-processing was necessary. Fig. 2(b) shows some examples of poor-quality data.

First, an algorithm was devised to remove the invalid areas of black space by cropping a fixed number of pixels from each of the four sides of each image whilst avoiding a large amount of computational overhead caused by the black space. Second, the images exhibited resolutions in the range of 1631 × 1879 to 1823 × 1650; we standardised the resolution by downsizing all images to a uniform size in accordance with the input requirements of specific models. Furthermore, for measuring the light intensity at each pixel in a single image, we converted all images to greyscale. For the images with excessively bright or dark background and foreground, we used histogram equalisation (HE) to improve the visual effects of the images and to discover hidden messages. To improve local contrast and enhance edge definition in each image region, we used adaptive HE, which was initially proposed by Stark [38,39], as a part of the pre-processing steps. Mathematically, the adaptive HE can be written as follows:

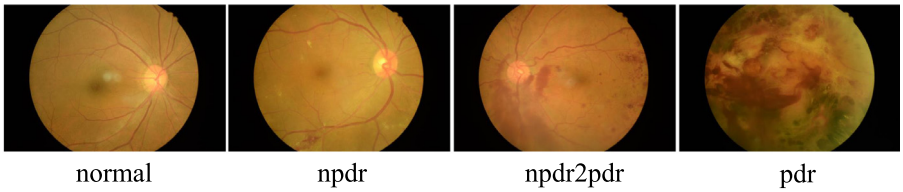
$$\begin{aligned} f_c(u,v) &= q(u-v,\alpha) - \beta q(u-v,1) + \beta uq(d,\alpha) \\ q(d,\alpha) &= \frac{1}{2}\sin(d)\,|2d|^{\alpha}, \end{aligned} \tag{1}$$

where $0 \le \alpha$ and $\beta \le 1$, and we set $\alpha = \beta$. f_c is called an accumulation function, and HE is given if $\alpha = 0$. These equations are explained in detail in [39].

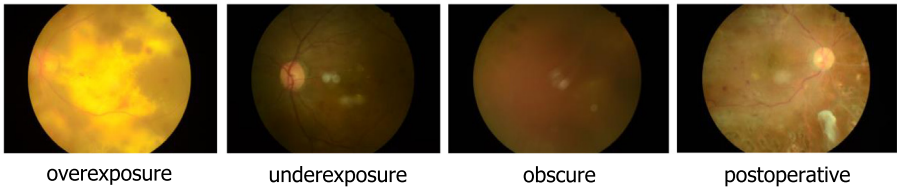
For dark images, we provided a contrast stretching algorithm to enhance the contrast effect of each area of the image. This algorithm was performed by using the following equation:

$$I(x,y) = \frac{I(x,y) - I_{min}}{I_{max} - I_{min}} \times 255, \tag{2}$$

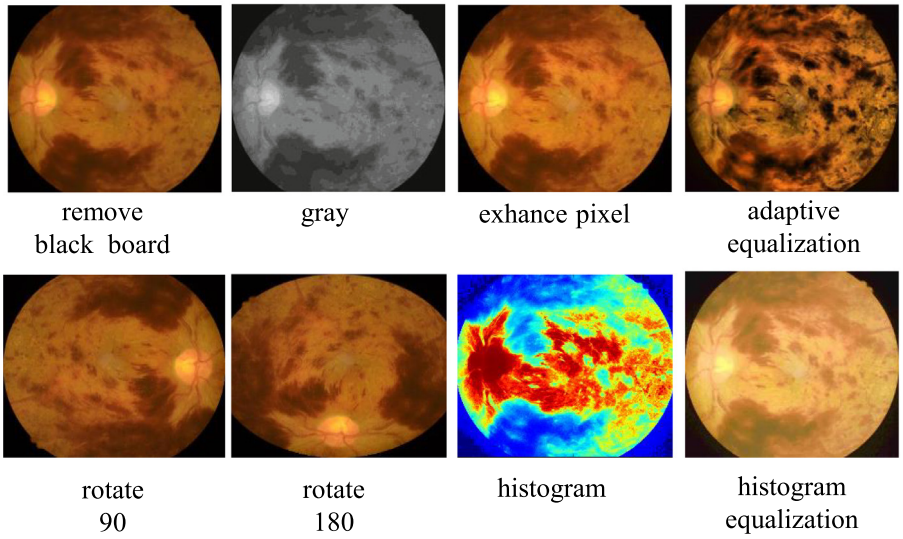
where $I(x,y)$ is the grey value of a pixel in the original image and I_{min} and I_{max} are the real minimum and maximum greyscale values of the original image, respectively.



(a) Normal and DR level



(b) Example images of poor quality



(c) Example images of data preprocessing

Fig. 2. Example images of poor quality and preprocessing in the dataset.

3.5. Performance comparison

Fig. 2(c) shows some examples of the pre-processing in our dataset. The comparative experiment was executed via Xception. The model performed well in the binary task. The model was trained via resizing only and no other pre-processing. Then, it was compared with the models trained with all the pre-processing methods. The accuracy of the model without pre-processing reached 94.79% until 300 epochs. This model's accuracy did not exceed such level even after fine turning. By contrast, the model with pre-processing converged well after 220 epochs and achieved an accuracy of 95.68%. This accuracy rate improved to 97.15% after fine turning (Fig. 5). Fig. 3 and Table 7 provide further details.

3.6. Data augmentation

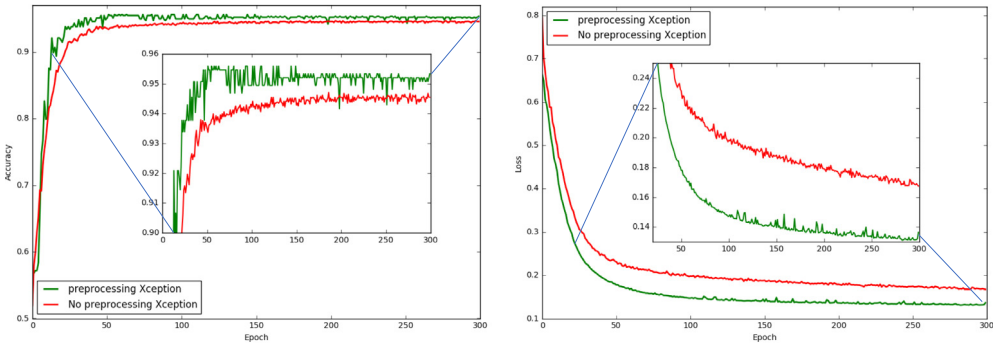
The amount of PDR only accounted for 9% of the total data (Table 3), and the inter-grader variability was serious in the pathological features. The model had difficulties in learning the characteristics of PDR, and PDR was over identified as NPDR2PDR. Therefore, we used data augmentation technology, such as random rotation by 0°–180°, random horizontal or vertical flips, to enhance the size of the training set in real time. Generally, augmented images retained the major features from their original images. Therefore, the technologies ensured that the training set was expanded whilst the images were not copied completely.

4. Model and methodology

4.1. Aim and objective

The following two aims motivated this study and were realised with corresponding ensemble models that were designed in this work.

(1) To build an early DR automatic screening system. This aim is a binary classification task to identify the presence of DR. Currently, the manual DR screening method is labour intensive and suffers from inconsistencies across sites [40]; moreover, the number of people who can master DR treatment skills is still



(a) Comparison of Accuracy (b) Comparison of Loss

Fig. 3. The comparison before and after preprocessing of the identification system.

Table 4
Hyper-parameter configuration.

Configuration	Value
Optimisation function	RMSprop
Epoch	300 (training), 100 (fine turning)
Batch size	32
Learning rate	2.00E−04
Drop out	5.00E−01
Class_weight	auto
ReduceLROnPlateau	monitor = 'val_acc', factor = 0.5, patiencer = 10, epsilonr = 0.0001
EarlyStopping	monitorr = 'val_loss', patiencer = 15
ModelCheckpoint	monitorr = 'val_acc', moder = 'auto', periodr = 1

Table 5
Class distribution and classifications report of the identification system.

	Training	Validation	Testing	
Normal	1189	132	348	
Abnormal	1567	174	423	
Total	2756	306	771	
	Precision	Recall	F1_score	Support
Normal	0.97	0.98	0.97	348
Abnormal	0.98	0.97	0.98	423
ave/total	0.98	0.98	0.98	771

small amongst most grassroots health workers. Therefore, the first task was primarily applied to communities at the grassroots level because patients with DR may be identified timely and referred to ophthalmology for further diagnosis and treatment whilst potentially alleviating the burden for ophthalmologists.

(2) To build an automatic grading system. This aim is a multi-classification task to predict the level of DR severity. The grading system was primarily used in hospitals to provide ophthalmologists with auxiliary diagnostic references whilst avoiding human subjectivity.

4.2. Architecture and strategy of ensemble model

In this study, 'learner', 'basic learner', 'component classifier' and 'component' refer to an independent neural network used in an ensemble.

Several components comprised a corresponding ensemble model. Each component was a two-part neural network. The first part was a feature extractor comprising a pre-training model that was initialised via transfer learning which involved the removal of the top layer. The second part was a classifier that made predictions on the basis of the aforementioned features; it was realised by a customised stand deep neural network with training from scratch.

4.3. Ensemble strategy

Many studies have largely realised ensemble methods with many components at the expense of time and memory. Therefore, the number of component classifiers to be included in the ensemble is an important issue in integration model experiments [41]. The issue further triggers deep thinking, namely, the choice of combination method and the training mechanism of components, which largely characterise the ensemble method [42–44]. The former indicates which learners can be combined, whereas the latter describes how to combine them.

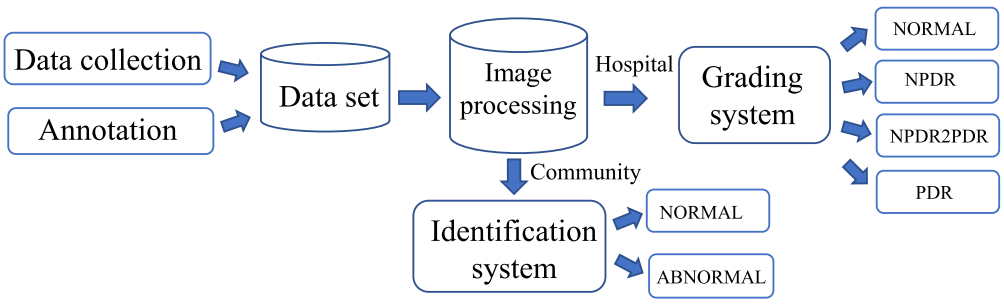
For 'which learners', extensive literature has shown that the performance of the overall classifier can be enhanced if a not-so-weak learner is used as a base learner [45]. Meanwhile, the predictive power of a learner is closely related to its ability to extract high-quality features; thus, a high-performance pre-model must be selected as the feature extractor in the bottom part of the component.

For 'how to', because DNNS has high variance and low bias, the variance of models can be dramatically reduced via averaging if the models are independent [46,47]. In the current work, we averaged the softmax scores of all models to solve the second issue above because these probabilities from different models might have varying output magnitudes. Comparative experiments between the averaging and max methods applied to the ensemble models were added for objectivity. The results demonstrated that the averaging method was more effective in our study (Tables 8 and 9).

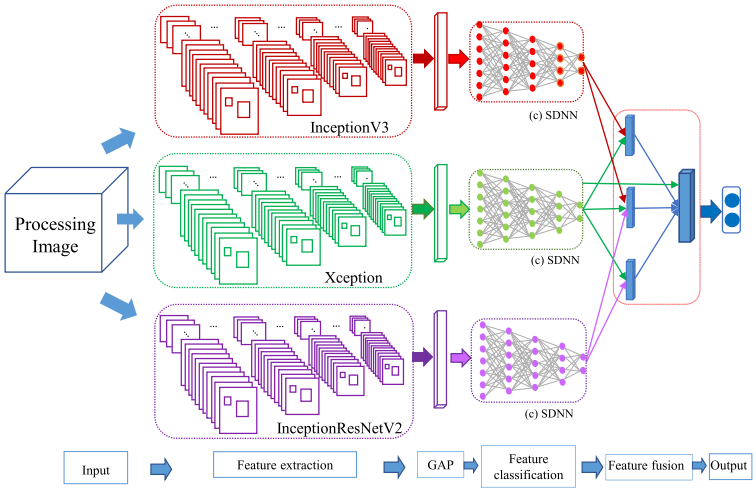
4.4. Transfer learning at the first part

Several standout CNNs that remove top layers made up the first stage of different ensembles. We used these CNNs to produce a compact feature vector representation in a DR image. In consideration of different tasks, we performed some analyses of respective CNN characteristics as follows to facilitate our selection of a strong feature extractor (learner) for each component of the ensemble.

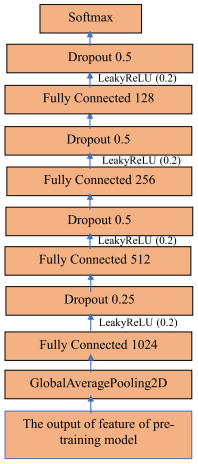
To enhance the speed of the calculation, we used ResNet50 [48] and InceptionV3 [49] as alternatives. InceptionResNetV2 [50] was also attempted because it can make the network deep and fast by mitigating the problem of gradient disappearance using jumper connections. Similarly, Xception [51] was considered as one improvement to InceptionV3 due to its ability to improve the effects of the model without increasing the complexity of the network. DenseNets [52] were attractive because they can fully use the features and further reduce gradient disappearance; furthermore, they performed well in our experiments.



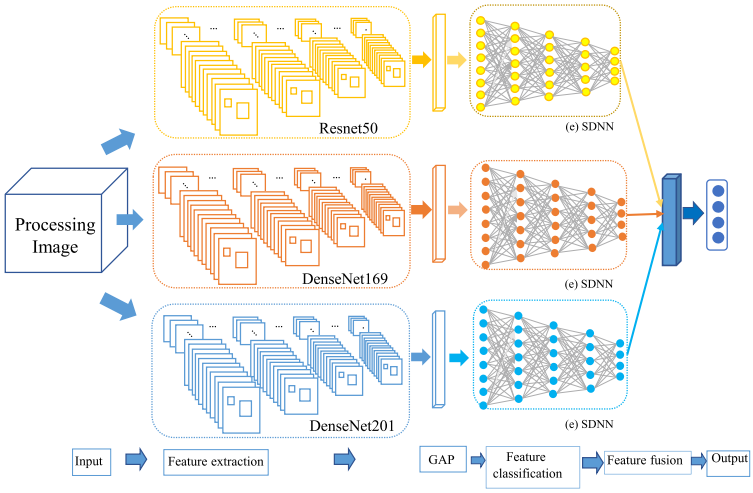
(a) The overall framework



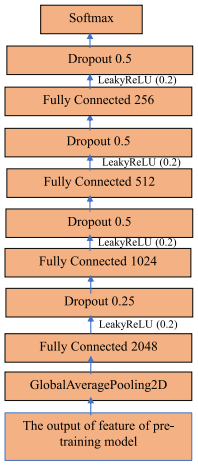
(b) the Identification system



(c) SDNN



(d) the Grading system



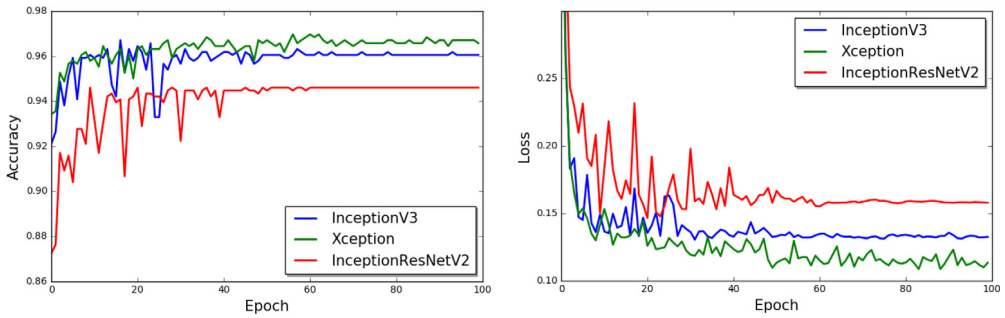
(e) SDNN

Fig. 4. The whole system.

4.5. Design of customised standard deep neural network (SDNN) at the second part

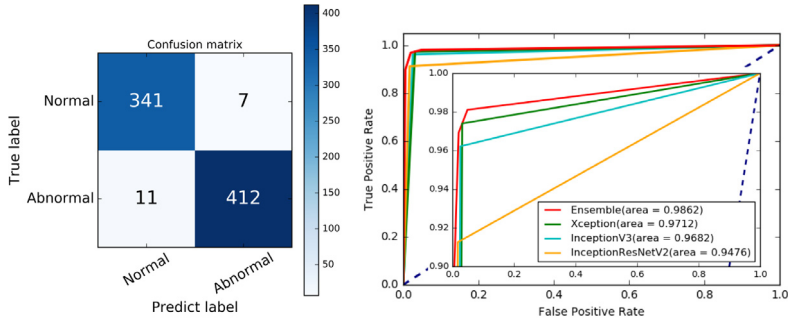
On the basis of the specific data distributions and difficulties of the two tasks, two respective types of SDNNs were defined as component classifiers at the second stage (Fig. 4). Input to the SDNNs was given by the output of the feature extractors. SDNNs have the same network depth but different parameters. This faint distinction led to important changes in the prediction performance: through experiments and observations, we found that adding or removing a layer could reduce the learning capacity of the network regardless of the task. Given the similarity of designs of SDNN frameworks, we showed only the development of the SDNN model in the four-class classification task.

Notably, the feature maps from the forward layer are spatial, but they can be normalised via global average pooling (GAP). Thus, we designed a GAP layer as the first layer of the SDNN. The second layer was a fully connected layer with 2048 hidden neurons. Proper nonlinearity is an important factor in the incremental performance of a classification model, especially for a



(a) training accuracy curves (b) training loss curves

Fig. 5. Training curves of the components of the identification system.



(a) Confusion matrix (b) ROC curves of components

Fig. 6. Confusion matrix and ROC curves of the identification system.

limited dataset. Thus, we closely monitored the effects of non-linearity between the rectified linear unit function (ReLU) and its variant, leaky rectified linear unit function (leaky ReLU), on SDNN performance. In the same condition, we conducted multiple comparison experiments and found that leaky ReLU was significantly faster than ReLU with respect to convergence and shortening of training time. According to prior literature [43–45], the difference may be caused by the potential disadvantage of ReLU during optimisation: the gradient is 0 whenever the neuron is dead (saturated and not active). This occurrence may cause the unit to remain inactive because the gradient-based optimisation algorithm does not adjust the weight of a dead neuron. Therefore, the speed of training ReLU networks is slow when gradients remain at zero. By contrast, leaky ReLU slightly adjusts the weight of dead neurons into small and non-zero gradients. On the basis of the above analysis, we applied leaky ReLU layers to the output of all inner fully connected layers, except the output layer, to achieve nonlinearity in the SDNN. The two functions of ReLU and leaky ReLU are defined in Eqs. (3) and (4), respectively.

$$h^{(i)} = max(w^{(i)(T)}x, 0) = \begin{cases} w^{(i)T}x, & \text{if } w^{(i)T}x > 0 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

$$h^{(i)} = max(w^{(i)(T)}x, 0) = \begin{cases} w^{(i)T}x, & \text{if } w^{(i)T}x > 0 \\ 0.01w^{(i)T}x, & \text{otherwise} \end{cases} \tag{4}$$

The formulas above show the similarities and differences between the two functions. The latter function may achieve a more robust gradient than the former by sacrificing sparsity.

A dropout layer was added after each dense layer. The addition can effectively omit many neurons of hidden layers during training and ensure the validity of the data; it can also mitigate or prevent data overfitting if the network shows excessive reliance on certain nodes in one layer. We updated each node with probability p = 0.5 whilst updating each layer. Then, we left it unchanged with probability 1−p. Following these multiple layers, the final layer was a standard softmax classifier with cross entropy as the loss function. The softmax function took an N-dimensional vector of arbitrary real values and produced another N-dimensional vector with real values in the range (0,1), thereby adding up to 1.0; each value of the output vector represented the probability that the sample belonged to each class. Cross entropy served as a loss function that revealed the distance or degree of closeness between the true labels and the predicted labels of the network. It is defined as follows:

$$L_j = -\log\left(\frac{e^{s_j}}{\sum_{k=1}^{N} e^{s_k}}\right), \tag{5}$$

where N indicates the number of classes, s_j is the score for the sample label j, and s_k is the score for a particular label k. Softmax ensured that the prediction probabilities exhibited a proper probability distribution.

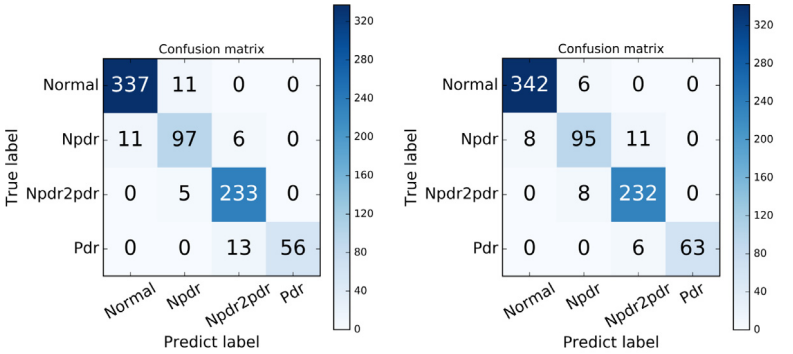
5. Experiments

5.1. Configuration

The algorithms were implemented using Keras (http://keras.io/). All experiments were performed on a high-end workstation with an Intel Xeon E5-2620 CPU and NVIDIA Tesla K40 GPU with 64 GB of RAM. The dataset was split into 70% for training, 10% for validation and 20% for testing. The configuration of the hyper-parameter and the class distributions of the two tasks are shown in Tables 4–6.

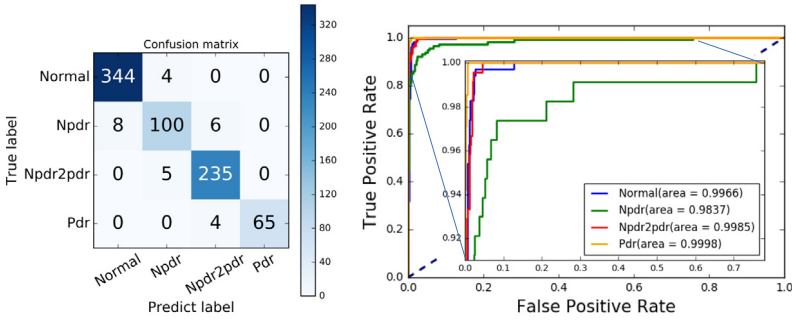
5.2. Strategy

The experiment process consisted of six steps: input data, data pre-processing, single-model feature extraction, single-model



(a) the two-stages strategy (b) the four-class strategy

Fig. 7. Comparison of confusion matrices between two strategies of the grading system.



(a) Confusion matrix (b) ROC curves

Fig. 8. Confusion matrix and ROC curves of the grading system.

Table 6
Class distribution and classifications report of the grading system.

	Training	Validation	Testing	
Normal	1189	132	348	
npdr	469	52	114	
npdr2pdr	875	97	240	
pdr	224	25	69	
Total	2756	306	771	
	Precision	Recall	F1_score	Support
Normal	0.98	0.99	0.98	348
npdr	0.92	0.88	0.9	114
npdr2pdr	0.96	0.98	0.97	240
pdr	1	0.94	0.97	69
ave/total	0.96	0.96	0.96	771

feature classification, multi-model feature fusion and output results (Fig. 4). During one component development, the first phase focused on the separate pre-training of the SDNNs; the key points were the number of layers needed in each SDNN with corresponding parameters to achieve optimal performance. The second phase was to combine each pre-processing model with the corresponding SDNN model with retraining. Lastly, the component was fine-tuned by determining the suitable layer to be frozen. For convenience, each component classifier was named as the components' internal pre-training model.

Models were designed in parallel and independently whilst taking them as alternative component classifiers. The choice of the most suitable component classifier and the size of the ensemble have a great impact on the accuracy of the prediction results of the model. Therefore, we attempted rich experiments to explore the relationship amongst the ideal size of the ensemble model (the number of basic component classifiers), the optimal width of the ensemble model (the combination method of component classifiers) and the number of class labels for the two systems on the basis of our dataset.

5.3. Metrics

We utilised reliability and validity to evaluate the models and their ensembles. Reliability measures the degree of stability obtained by repeated tests under the same conditions. It can be assessed using the Kappa value. Kappa $>$ 0.8 indicates excellent consistency. Validity reflects the degree of conformity between measured and actual values, as indicated by the seven metrics below.

Accuracy indicates the proportion of samples classified correctly. Precision is the proportion of positives correctly predicted. For early screening systems, sensitivity and specificity are important reference indicators of the referral decisions of screening options that directly indicate the effectiveness of the system. Sensitivity measures the proportion of positives correctly predicted, whereas specificity is the proportion of true negatives correctly identified as such. Given a typical balance between the two measures, a receiver operating characteristic curve with AUC can represent this balance graphically. The F1_score is the harmonic average of precision and recall. An F1_score close to 1 indicates good performance. F_β is the weighted harmonic average of F1_score. As the Youdens index approaches 1, the screening system exhibits enhanced authenticity; the reverse relationship is equally valid. These measures are presented in the following:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{6}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \tag{7}$$

Table 7
Metrics of the identification system ensembles.

Model	Accuracy	Precision	Sensitivity	Specificity	Auc	Kappa	F1_score	F_β_score	Youden's index
Xception[a]	0.9521	0.9491	0.9321	0.9782	0.9551	0.9032	0.9523	0.9503	0.9101
Baseline:Resnet 50	0.9342	0.9333	0.9332	0.9351	0.9342	0.9231	0.9322	0.9223	0.9223
0:Xception[b]	0.9715	0.9711	0.9740	0.9683	0.9712	0.9424	0.9715	0.9712	0.9424
1:InceptionV3	0.9676	0.9667	0.9622	0.9741	0.9682	0.9676	0.9669	0.9346	0.9363
2:IncepresV2[c]	0.9442	0.9435	0.9125	0.9828	0.9476	0.9444	0.9433	0.8883	0.8953
Ens(01)[d]	**0.9767**	0.9759	0.9716	**0.9828**	0.9772	0.9529	0.9767	0.9761	**0.9544**
Ens(02)[e]	0.9689	0.9677	0.9574	**0.9828**	0.9701	0.9373	0.9689	0.9680	0.9402
Ens(12)[f]	0.9598	0.9585	0.9456	0.9770	0.9613	0.9191	0.9599	0.9588	0.9226
3:Ens(012)[g]	0.9702	0.9691	0.9622	0.9799	0.9710	0.9340	0.9702	0.9694	0.9421
Ens(012,01)[h]	0.9689	0.9678	0.9598	0.9799	0.9698	0.9373	0.9689	0.9681	0.9397
Ens(012,02)[i]	0.9662	0.9652	0.9574	0.9770	0.9672	0.9321	0.9663	0.9655	0.9345
Ens(0,01,02)[j]	0.9753	0.9750	**0.9764**	0.9742	0.9752	0.9503	0.9753	0.9750	0.9505
Ens(0,2,01,02)[k]	0.9702	0.9691	0.9621	0.9799	0.9710	0.9399	0.9710	0.9694	0.9420
Ens(0,1,01,02)[l]	0.9740	0.9732	0.9693	0.9799	0.9746	0.9477	0.9741	0.9735	0.9492
Ens(0,01,02,3)[m]	**0.9767**	0.9760	0.9740	0.9799	0.9769	0.9529	0.9767	0.9762	0.9539
Ens(0,2,01,02,3)[n]	0.9701	0.9691	0.9622	0.9799	0.9710	0.9399	0.9702	0.9694	0.9421
Ens(0,1,01,02,3)[o]	0.9741	0.9732	0.9693	0.9799	0.9745	0.9477	0.9741	0.9735	0.9492
Ens (0,01,02,012)[p]	**0.9767**	**0.9760**	**0.9764**	0.9800	**0.9862**	**0.9530**	**0.9769**	**0.9762**	0.9540

[a]Xception = Comparison of performances before preprocessing of Xception.
[b]Xception = Comparison of performances after preprocessing of Xception.
[c]IncepresV2 = InceptionResNetV2.
[d]Ens(01) = Average of Xception and InceptionV3.
[e]Ens(02) = Average of Xception and InceptionResNetV2.
[f]Ens(12) = Average of InceptionV3 and InceptionResNetV2.
[g]Ens(012) = Average of Xception,InceptionV3 and InceptionResV2.
[h]Ens(012,01) = Average of Ens(012) and Ens(01).
[i]Ens(012,02) = Average of Ens(012) and Ens(02).
[j]Ens(0,02,02) = Average of Xception, Ens(02) and Ens(02).
[k]Ens(0,2,02,12) = Average of Xception, InceptionResNetV2, Ens(02) and Ens(12).
[l]Ens(0,1,01,02) = Average of Xception, InceptionV3, Ens(01) and Ens(02).
[m]Ens(2,02,12,3) = Average of Xception, Ens(01), Ens(02) and Ens(012).
[n]Ens(0,2,01,02,3) = Average of Xception, InceptionResNetV2, Ens(01), Ens(02), and Ens(012).
[o]Ens(0,1,01,02,3) = Average of Xception, InceptionV3, Ens(01), Ens(02) and Ens(012).
[p]Ens(0,01,02,012) = Average of Xception, Ens(01), Ens(02) and Ens(012).

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \tag{8}$$

$$\text{F1\_score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{9}$$

$$\text{F}_\beta = \frac{\left(1+\beta^2\right) \times \text{Precision} \times \text{recall}}{\beta^2 \times \text{Precision} + \text{recall}} \tag{10}$$

$$\begin{aligned}\text{Youden's index} &= \text{Sensitivity} + \text{Specificity} - 1 \\ &= \frac{\text{TP} \times \text{TN} - \text{FN} \times \text{FP}}{(\text{TP+FN}) \times (\text{TN+FP})}.\end{aligned} \tag{11}$$

Where, the samples can be divided into true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) in accordance with the combination of the real class and the classifier prediction category.

5.4. Identification system

The situation of classification imbalance was not serious. To reduce the computational cost, we selected three models (InceptionV3, InceptionResNetV2 and Xception) as alternative feature extractors, with ResNet50 as the baseline. These models largely represent the best learning abilities available. Whilst making multiple combinations of these models, we expected the identification system to achieve the best performance.

After training these models independently (Fig. 5), we marked the models from strong to weak: Xception, InceptionV3 and InceptionResNetV2 as 0, 1 and 2, respectively. First, the models were combined directly to an ensemble named Ens(012) by averaging their softmax scores. However, the performance did not match that of Xception. We speculated that the ideal number of component classifiers for binary classification problems may not simply be a result of the direct integration of the three basic models above. Thus, ensemble models consisting of two basic models were generated. The experiments showed that Ens(01) performed the best amongst all ensembles. We widen the ensemble models by combining these basic models with Ens(01) and Ens(02). Notably, Ens (0,01,02) exceeded well the previous best model Ens(01), and the result indicated that the optimal width of the ensemble may sometimes be useful. Therefore, the ensembles were further widened by combining the basic models with other ensemble models, such as Ens (0,2,02,01). The new results were almost better than those of Ens(012), but they were not as good as those of Ens(01). We replaced these basic models with the ones that performed well in the above experiments. Ens (0,01,02,012) showed a standout performance; hence, it was taken as the ensemble model of the task, for which it yielded the highest performance (Fig. 4(b)). Fig. 6 and Table 5 show the performance of the final ensemble. Table 7 presents the primary ensembles attempted and their performances under as many combinations as possible.

5.5. Grading system

5.5.1. Two alternative strategies

Regarding the grading system, we considered two alternative implementation strategies. The first strategy was a two-stage system: the first stage was a binary classification that distinguished between abnormal images and normal images via Xception, and the second stage was a ternary classification process to predict the level of DR severity on the basis of the abnormal data above

Table 8
Metrics of the grading system ensembles: the first group, the input is (299 × 299).

Model	Accuracy	Recall	Precision	Sensitivity	Specificity	Kappa	F1_score	F_β_score	Youden's index
0: InceptionV3	0.9455	0.9260	0.9376	0.9787	0.9655	0.8811	0.9449	0.9349	0.9442
1: IncepresV2	0.9416	0.9072	0.9072	0.9764	0.9741	0.9123	0.9410	0.9306	0.9505
2: Xception	0.9209	0.8809	0.9001	0.9622	0.9828	0.8811	0.9201	0.9959	0.9449
Ens(01)[a]	0.9442	0.9101	0.9408	0.9788	0.9770	0.9436	0.9436	0.9336	0.9460
Ens$(01)_m$	0.9442	0.9101	0.9408	0.9788	0.9770	0.9162	0.9436	0.9336	0.9557
Ens(02)[b]	0.9507	0.9332	0.9431	0.9811	0.9684	0.9262	0.9503	0.9409	0.9495
Ens$(02)_m$	0.9468	0.9267	0.9396	0.9787	0.9684	0.9203	0.9461	0.9367	0.9471
Ens(12)[c]	0.9429	0.9094	0.9387	0.9787	0.9741	0.9143	0.9424	0.9318	0.9529
Ens$(12)_m$	0.9416	0.9072	0.9377	0.9764	0.9741	0.9123	0.9411	0.9306	0.9505
3: Ens(012)[d]	0.9520	0.9701	0.9677	0.9811	0.9770	0.9280	0.9517	0.9423	0.9581
Ens$(012)_m$	0.9442	0.9101	0.9408	0.9787	0.9770	0.9162	0.9436	0.9336	0.9557
Ens(3,0)[e]	0.9455	0.9262	0.9377	0.9787	0.9626	0.9184	0.9449	0.9350	0.9414
Ens(3,1)[f]	0.9416	0.9072	0.9376	0.9764	0.9741	0.9123	0.9410	0.9306	0.9505
Ens(3,2)[g]	0.9494	0.9217	0.9406	0.9764	0.9770	0.9241	0.9489	0.9365	0.9534
Ens$(3, 2)_m$	0.9416	0.9072	0.9377	0.9764	0.9741	0.9123	0.9410	0.9306	0.9505
Ens(3,01)[h]	**0.9546**	**0.9282**	**0.9529**	**0.9811**	**0.9799**	**0.9318**	**0.9542**	**0.9473**	**0.9610**
Ens$(3, 01)_m$	0.9533	0.9260	0.9518	0.9810	0.9799	0.9299	0.9528	0.9459	0.9610
Ens(3,02)[i]	0.9494	0.9281	0.9432	0.9787	0.9713	0.9242	0.9488	0.9398	0.9500
Ens$(3, 02)_m$	0.9494	0.9296	0.9423	0.9787	0.9684	0.9242	0.9490	0.9396	0.9471
Ens(01,02)[j]	0.9494	0.9310	0.9421	0.9787	0.9684	0.9242	0.9489	0.9396	0.9471
Ens$(01, 02)_m$	0.9520	0.9324	0.9463	0.9787	0.9741	0.9281	0.9513	0.9432	0.9529
Ens(3, *Ens*$(01, 02)_m$)[k]	0.9533	0.9306	0.9489	0.9787	0.9741	0.9300	0.9528	0.9449	0.9529
Ens(3, *Ens*$(01, 02)_m)_m$	0.9520	0.9310	0.9465	0.9787	0.9741	0.9280	0.9515	0.9431	0.9529

[a]Ens(01) = Average of InceptionV3 and IncepresV2; *Ens*$(01)_m$ = max of them.
[b]Ens(02) = Average of InceptionV3 and Xception; *Ens*$(02)_m$ = max of them.
[c]Ens(12) = Average of InceptionV3, IncepresV2 and Xception; *Ens*$(12)_m$= max of them.
[d]Ens(012) = Average of IncepresV2 and Xception; *Ens*$(012)_m$ = max of them.
[e]Ens(3,0) = Average (max) of Ens(012) and InceptionV3.
[f]Ens(3,1) = Average (max) or max of Ens(012) and IncepresV2.
[g]Ens(3,2) = Average of Ens(012) and Xception; *Ens*$(3, 2)_m$ = max of them.
[h]Ens(3,01) = Average of Ens(012) and Ens(01); *Ens*$(3, 01)_m$ = max of them.
[i]Ens(3,02) = Average of Ens(012) and Ens(02); *Ens*$(3, 02)_m$ = max of them.
[j]Ens(01,02) = Average of Ens(01) and Ens(02); *Ens*$(01, 02)_m$ = max of them.
[k]Ens(3, *Ens*$(01, 02)_m$) = Average of Ens(012), *Ens*$(01, 02)_m$; *Ens*$(3, Ens(01, 02)_m)_m$ = max of them.

via Resnet50. The alternative strategy was a quaternary classification model for predicting the level of DR severity on the basis of all testing images via Resnet50. The accuracy of the former was 94.1%, and the accuracy of the latter was 95.2%. The intuitive comparison (Fig. 7) showed that the strategy of the four-class classification worked the best.

5.5.2. Four-class classification

The seven models discussed above were used in the experiments because of the thin granularity of multiple classification and the small amount of PDR data. The four-class experiments were divided into two groups by the models on the basis of their different input sizes (299×299, 224×224). From strong to weak, three models, i.e. InceptionV3, InceptionResNetV2 and Xception, were marked as 0, 1 and 2, respectively, in group one; the others, i.e. ResNet50, DenseNet169, DenseNet201 and DenseNet121, were marked as 4, 5, 6, 7, respectively. First, we evaluated various ensembles in each group. Second, the outstanding models from the two groups were combined as much as possible to determine the final ensemble framework of the grading system.

Galton's theory states that combining many simple predictions is a force for accurate predictions. Hence, we combined all basic models directly. This step yielded Ens (0124567) with a remarkable accuracy rate of 96.36% relative to the other base learners above. Notably, the sensitivity and specificity of Ens (0124567) were 98.10% and 98.56%, respectively; however, it costs more than the pre-trained models do. Inspired by the binary experiments above, we reduced the size of the ensembles to four. Moreover, the optimal number of component classifiers should be similar to the number of class labels in some studies [42–44]. The results showed that Ens (4567) had a high accuracy of 96.23% under the condition involving the same number of base learners. Sequentially, we reduced the size of the ensembles to three. The top three single models (0,4,5) were integrated into Ens (045); however, its accuracy was only 94.94%. Subsequently, we replaced the third-ranked model 2: InceptionV3 with the fourth-ranked model 5: DenseNet201; this step yielded Ens (456) with an accuracy of 96.50% and other relatively high indicators. A variety of integrations were further attempted, and similar performances were achieved; however, the results did not exceed the current level. Therefore, our judgment was correct. In other words, an optimal number of component classifiers exists and might be near the number of class tags with subtle adjustments depending on the specific task. Fig. 4(d) shows the Ens (456) framework, and Fig. 8 shows its performance. Further information about the evaluation metrics is shown in Tables 8–10.

5.6. Analysis of experiments

The UK National Institute for Clinical Excellence guidelines state that a DR screening test should have at least a sensitivity of 80% and a specificity of 95%. In our work, the identification model performed well with a sensitivity of 97.5%, a specificity of 97.7% and an accuracy of 97.7%. By contrast, the grading model achieved a sensitivity of 98.1%, a specificity of 98.9% and an accuracy of 96.5%. Therefore, the models achieved satisfactory performance on our dataset.

From the experiments, we found that the stronger the base learner was, the higher the performance was generally; moreover, the effects of the ensembles with multiple-ensemble classifiers were stronger than those of the dull ensemble models in some cases. We noticed that some results degenerated after the ensemble phase in the four-class experiments. We conjectured this degeneration was primarily caused by the model selection strategy in the ensemble phase: various base learners of different

Table 9
Metrics of the grading system ensembles: the second group, the input is (224 × 224).

Model	Accuracy	Recall	Precision	Sensitivity	Specificity	Kappa	F1_score	F_β_score	Youden's index
4: Resnet50	0.9494	0.9240	0.9451	0.9811	0.9828	0.9241	0.9491	0.9404	0.9638
5: DN_169	0.9469	0.9195	0.9466	0.9551	0.9914	0.9199	0.9464	0.9406	0.9465
6: DN_201	0.9429	0.9140	0.9333	0.9811	0.9741	0.9144	0.9421	0.9289	0.9552
7: DN_121	0.9364	0.9162	0.9222	0.9740	0.9741	0.9052	0.9372	0.9203	0.9481
Ens(45)[a]	0.9507	0.9261	0.9463	0.9811	0.9828	0.9261	0.9504	0.9418	0.9638
Ens(46)[b]	0.9546	0.9295	0.9534	0.9811	0.9828	0.9318	0.9540	0.9480	0.9638
Ens(47)[c]	0.9507	0.9273	0.9456	**0.9811**	0.9826	0.9261	0.9506	0.9415	0.9638
Ens(56)[d]	0.9494	0.9190	0.9450	0.9787	0.9856	0.9239	0.9485	0.9390	0.9644
Ens(57)[e]	0.9494	0.9241	0.9450	0.9645	0.9885	0.9240	0.9491	0.9403	0.9530
Ens(67)[f]	0.9481	0.9186	0.9428	0.9787	0.9798	0.9220	0.9472	0.9373	0.9586
8:Ens(456) [g]	**0.9650**	**0.9467**	**0.9635**	**0.9811**	0.9885	**0.9475**	**0.9647**	**0.9599**	**0.9696**
$Ens(456)_m$	0.9546	0.9295	0.9534	0.9811	0.9828	0.9318	0.9540	0.9480	0.9638
Ens(457)[h]	0.9571	0.9319	0.9561	0.9787	0.9885	0.9357	0.9567	0.9508	0.9508
Ens(467)[i]	0.9611	0.9402	0.9572	0.9834	0.9828	0.9417	0.9610	0.9535	0.9662
Ens(567)[j]	0.9611	0.9347	0.9589	0.9764	0.9914	0.9415	0.9605	0.9535	0.9508
9:Ens(4567)[k]	0.9623	0.9412	0.9616	0.9787	0.9856	0.9435	0.9620	0.9572	0.9644
$Ens(4567)_m$	0.9546	0.9295	0.9534	0.9811	0.9828	0.9318	0.9540	0.9480	0.9638
Ens(8,9,467)[l]	0.9611	0.9405	0.9595	0.9787	0.9828	0.9416	0.9609	0.9554	0.9615
Ens(8,9,567)[m]	0.9637	0.9434	0.9639	0.9763	0.9885	0.9454	0.9632	0.9594	0.9649
Ens(7,56,9,567)[n]	0.9650	0.9387	0.9635	0.9764	**0.9913**	0.9473	0.9645	0.9580	0.9678

[a]Ens(45) = Average of Resnet50 and DN_169.
[b]Ens(46) = Average of Resnet50 and DN_201.
[c]Ens(47) = Average of Resnet50 and DN_121.
[d]Ens(56) = Average of DN_169 and DN_201.
[e]Ens(57) = Average of DN_169 and DN_121.
[f]Ens(67) = Average of DN_201 and DN_121.
[g]Ens(456) = Average of Resnet50, DN_169 and DN_201; $Ens(456)_m$ = max of them.
[h]Ens(457) = Average of Resnet50, DN_169 and DN_121.
[i]Ens(467) = Average of Resnet50, DN_201 and DN_121.
[j]Ens(567) = Average of DN_169, DN_201 and DN_121.
[k]Ens(4567) = Average of Resnet50, DN_169, DN_201 and DN_121; $Ens(4567)_m$ = max of them.
[l]Ens(456,8,467) = Average of Ens(456), Ens(4567), Ens(467).
[m]Ens(456,8,567) = Average of Ens(456), Ens(4567), Ens(567).
[n]Ens(7,56,8,567) = Average of DN_121, Ens(56), Ens(4567) and Ens(567).

depths can implicitly learn different levels of semantic image representation. On the basis of model complementarity, the posterior probabilities of these weak learners can be fused to predict the modalities of unseen images. Compared with the max method in Table 8, the averaging method was effective in our work because it reduced the variances of the components substantively [47,53].

6. Discussion

During the design of the two ensemble models, we made several considerations in the following aspects.

(1) Combined strategy of components: The frameworks of the two classification tasks searched for the ideal combination of component classifiers on the basis of the number of class tags in the dataset as a guide. We assumed that the basic components used in our experiments were all independent. In the experiments, we found that arbitrarily increasing or decreasing the number of component classifiers would reduce the performance of the model. Moreover, different combinations of methods of the component classifiers were important in achieving the best integration performance. However, the diversity of the real-world dataset, the complexity of the task and the degree of independence of existing component classifiers under the constraint of computational resource requirements did not guarantee this assumption in most cases; thus, determining the integration framework of a given set classifier on real data remains a challenging problem.

(2) Model optimisation: In view of the limited dataset and the deep models used in the system, we should note the problem of gradient disappearance. On the basis of the unsupervised layer-wise training [54], we proposed a supervised block-wise training strategy. Each SDNN of a component classifier was briefly used as an independent model, and it was trained with high-level features as input via the corresponding pre-trained feature extractor model. After the SDNN was separately trained, it could be connected to the corresponding feature extractor to form a component classifier; the feature extractor module could be initialised with the pre-trained weight, and the optimal weight obtained after the SDNN independent training could be used as its pre-trained weight to initialise itself. Subsequently, the entire component classifier could be trained using fine-tuning after the training. This training arrangement increased the speed of the convergence of the whole classification component. In sum, the entire component was divided into several blocks for training; a good setting of weight parameters was found for each block, and the component was then globally optimised in accordance with the local optimal weight of each block whilst minimising the training cost.

(3) Several obvious advantages: First, the reproducibility and consistency of diagnostic results could provide clinicians with insights into the diagnostic process. Additionally, the two different systems could be used to match different application requirements. When screening populations with substantial diseases, achieving high sensitivity and high specificity is critical to minimise false-positive and false-negative results. Finally, a quick reporting of auxiliary diagnostic results could improve clinicians' efficiency.

(4) Some limitations: The annotation work was based on the clinical experience of the ophthalmologist graders. Therefore, the algorithm may perform differently when used in images with subtle findings that a majority of clinicians could not identify. Another fundamental limitation arises from the black box, which is the nature of deep networks. The network automatically learns the features from the images and associated grade; however, the

Table 10
Metrics of the grading system ensembles: integration.

Model	Accuracy	Recall	Precision	Sensitivity	Specificity	Kappa	F1_score	F_β_score	Youden's index
Ens(04)[a]	0.9559	0.9292	0.9716	0.9828	0.9885	0.9336	0.9552	0.9506	0.9544
Ens(045)[b]	0.9494	0.9242	0.9448	0.9645	0.9885	0.9240	0.9491	0.9403	0.9530
Ens(0456)[c]	0.9494	0.9242	0.9448	0.9645	0.9885	0.9240	0.9491	0.9403	0.9530
10:Ens(0124567)[d]	0.9637	0.9411	0.9647	0.9810	0.9856	0.9454	0.9632	0.9594	0.9667
Ens(01567)[e]	0.9624	0.9404	0.9634	0.9812	0.9856	0.9435	0.9618	0.9582	0.9667
Ens(01,56)[f]	0.9624	0.9406	0.9632	0.9811	0.9856	0.9435	0.9615	0.9577	0.9667
Ens(8,10) [g]	0.9624	0.9406	0.9632	0.9811	0.9856	0.9435	0.9615	0.9577	0.9667
Ens(Ens(3,01),8,10) [h]	0.9611	0.9353	0.9627	0.9811	0.9856	0.9414	0.9605	0.9564	0.9667
Ens(3,9,567) [i]	0.9611	0.9353	0.9627	0.9811	0.9856	0.9414	0.9605	0.9564	0.9667
Ens(Ens(3,01),9,567) [j]	0.9637	0.9412	0.9644	0.9834	0.9856	0.9454	0.9632	0.9564	0.9690
Ens(9,567)	0.9611	0.9428	0.9584	0.9787	0.9856	0.9416	0.9608	0.9550	0.9644

[a]Ens(04) = Average of InceptionV3 and Resnet50.
[b]Ens(045) = Average of InceptionV3, Resnet50, DN_169.
[c]Ens(0456) = Average of InceptionV3, Resnet50, DN_169,and DN_201.
[d]Ens(0124567) = Average of all models.
[e]Ens(01567) = Average of InceptionV3, InceptionV3Resnet50, DN_169, and DN_201.
[f]Ens(01,56) = Average of Ens(01) and Ens(56).
[g]Ens(8,10) = Average of **Ens(456)** and Ens(01567).
[h]Ens(Ens(3,01),8,10) = Average of Ens(3,01), Ens(456) and Ens(01567).
[i]Ens(3,8,567) = Average of Ens(012), Ens(4567) and Ens(567).
[j]Ens(Ens(3,01),9,567) = Average of Ens(012), Ens(4567) and Ens(567).

specific features by which the networks are formed are unknown. Understanding the aspects used by deep neural networks to make predictions is an active area of research.

In the future, a large training dataset containing tens of thousands of abnormal cases must be collected from other hospitals via various types of cameras to improve the models' generalisation. Second, the visualisation of the aspects learned by CNNs is important as it can improve the interpretability of diagnostic results by identifying the source regions of features associated with a specified classification result, as well as the magnitude of the feature intensity. Additionally, doctors can make an accurate diagnosis on the basis of visualisation results. Third, in the case of a medical dataset of limited scale, further discussing the design and research of ensemble frameworks from a theoretical perspective is necessary.

7. Conclusion

In conclusion, a high-quality labelled medical imaging DR dataset was built, and an identification and grading system of DR called DeepDR was proposed. The relationship between the number of ideal component classifiers and the number of class labels was verified and explored. Using nine medical metrics, we evaluated the models in terms of validity and reliability. The results demonstrated that DeepDR worked satisfactorily.

Acknowledgements

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

References

[1] R. Klein, B.E. Klein, S.E. Moss, M.D. Davis, D.L. DeMets, The wisconsin epidemiologic study of diabetic retinopathy: Ii. Prevalence and risk of diabetic retinopathy when age at diagnosis is less than 30 years, Arch. Ophthalmol. 102 (4) (1984) 520–526.
[2] Y.H. Ma, Diabetic retinopathy screening rate of less than 10% in China, Chin. J. Med. Sci. 3 (2016).
[3] J.C. Javitt, L.P. Aiello, Y. Chiang, F.F. Rd, J.K. Canner, S. Greenfield, Preventive eye care in people with diabetes is cost-saving to the federal government. implications for health-care reform, Diabetes Care 17 (8) (1994) 909–917.
[4] D.E. Singer, Screening for diabetic retinopathy, J. Intern. Med. 240 (1) (1996) 45.
[5] J.K. Kristinsson, E. Stefánsson, F. Jónasson, I. Gíslason, S. Bjirnsson, Systematic screening for diabetic eye disease in insulin dependent diabetes, Acta Ophthalmol. 72 (1) (2010) 72–78.
[6] Diabetic Retinopathy, American Academy of Ophthalmology https://www.aao.org/preferred-practice-pattern/diabetic-retinopathy-ppp-updated-2017/, 2016.
[7] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.
[8] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436.
[9] V. Sze, Y.-H. Chen, T.-J. Yang, J.S. Emer, Efficient processing of deep neural networks: A tutorial and survey, Proc. IEEE 105 (12) (2017) 2295–2329.
[10] Z. Yi, Foundations of implementing the competitive layer model by lotka-volterra recurrent neural networks, IEEE Trans. Neural Netw. 21 (3) (2010) 494–507, http://dx.doi.org/10.1109/TNN.2009.2039758.
[11] L. Zhang, Z. Yi, S.-i. Amari, Theoretical study of oscillator neurons in recurrent neural networks, IEEE Trans. Neural Netw. Learn. Syst. (2018).
[12] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 2012 (2012) 1097–1105.
[13] R. Fakoor, F. Ladhak, A. Nazi, M. Huber, Using deep learning to enhance cancer diagnosis and classification, in: Proceedings of the International Conference on Machine Learning, vol. 28, 2013.
[14] A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (7639) (2017) 115.
[15] D. Wang, A. Khosla, R. Gargeya, H. Irshad, A.H. Beck, Deep learning for identifying metastatic breast cancer, 2016, arXiv preprint arXiv:1606.05718.
[16] A.M. Rossetto, W. Zhou, Deep learning for categorization of lung cancer CT images, in: 2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologie, CHASE, 2017, pp. 272–273.
[17] J. Wang, R. Ju, Y. Chen, L. Zhang, J. Hu, Y. Wu, W. Dong, J. Zhong, Z. Yi, Automated retinopathy of prematurity screening using deep neural networks, Ebiomedicine (2018).
[18] J. Hu, Y. Chen, J. Zhong, R. Ju, Z. Yi, Automated analysis for retinopathy of prematurity by deep neural networks, IEEE Trans. Med. Imaging (2018).
[19] S.J. Pan, Q. Yang, et al., A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359.
[20] R.S. Michalski, A theory and methodology of inductive learning, in: Machine Learning, Springer, 1983, pp. 83–134.
[21] L. Rokach, Ensemble-based classifiers, Artif. Intell. Rev. 33 (1–2) (2010) 1–39.
[22] R. Polikar, Ensemble based systems in decision making, IEEE Circuits Syst. Mag. 6 (3) (2006) 21–45.
[23] D. Opitz, R. Maclin, Popular ensemble methods: An empirical study, J. Artificial Intelligence Res. 11 (1999) 169–198.
[24] R.J. Winder, P.J. Morrow, I.N. Mcritchie, J.R. Bailie, P.M. Hart, Algorithms for digital image processing in diabetic retinopathy, Comput. Med. Imaging Graph. 33 (8) (2009) 608–622.

[25] M.D. Abràmoff, M.K. Garvin, M. Sonka, Retinal imaging and image analysis, IEEE Rev. Biomed. Eng. 3 (2010) 169–208.
[26] T. Walter, J.-C. Klein, P. Massin, A. Erginay, A contribution of image processing to the diagnosis of diabetic retinopathy-detection of exudates in color fundus images of the human retina, IEEE Trans. Med. Imaging 21 (10) (2002) 1236–1243.
[27] M. Niemeijer, B. van Ginneken, S.R. Russell, M.S. Suttorp-Schulten, M.D. Abramoff, Automated detection and differentiation of drusen, exudates, and cotton-wool spots in digital color fundus photographs for diabetic retinopathy diagnosis, Invest. Ophthalmol. Vis. Sci. 48 (5) (2007) 2260–2267.
[28] O. Faust, R. Acharya, E.Y.-K. Ng, K.-H. Ng, J.S. Suri, Algorithms for the automated detection of diabetic retinopathy using digital fundus images: A review, J. Med. Syst. 36 (1) (2012) 145–157.
[29] M.U. Akram, S. Khalid, A. Tariq, S.A. Khan, F. Azam, Detection and classification of retinal lesions for grading of diabetic retinopathy, Comput. Biol. Med. 45 (2014) 161–171.
[30] L. Gowda, K. Viswanatha, Automatic diabetic retinopathy detection using FCM, 2018.
[31] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. van der Laak, B. Van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88.
[32] I. Kaggle, Diabetic Retinopathy Detection, American Academy of Ophthalmology, 2015, URL: https://www.kaggle.com/c/diabetic-retinopathy-detection.
[33] J. Benson, H. Carrillo, J. Wigdahl, S. Nemeth, J. Maynard, G. Zamora, S. Barriga, T. Estrada, P. Soliz, Transfer learning for diabetic retinopathy, in: Image Processing, 2018, p. 70.
[34] R. Gargeya, T. Leng, Automated identification of diabetic retinopathy using deep learning, Ophthalmology 124 (7) (2017) 962–969.
[35] V. Gulshan, L. Peng, M. Coram, M.C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, et al., Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, JAMA 316 (22) (2016) 2402–2410.
[36] A. Tamkin, I. Usiri, C. Fufa, Deep CNNs for diabetic retinopathy detection.
[37] H. Pratt, F. Coenen, D.M. Broadbent, S.P. Harding, Y. Zheng, Convolutional neural networks for diabetic retinopathy, Procedia Comput. Sci. 90 (2016) 200–205.
[38] J.A. Stark, Adaptive image contrast enhancement using generalizations of histogram equalization, IEEE Trans. Image Process. 9 (5) (2002) 889–896.
[39] J.A. Stark, W.J. Fitzgerald, Model-based adaptive histogram equalization, Signal Process. 39 (1–2) (1994) 193–200.
[40] J.K.H. Goh, C.Y. Cheung, S.S. Sim, P.C. Tan, G.S.W. Tan, T.Y. Wong, Retinal imaging techniques for diabetic retinopathy screening, J. Diabetes Sci. Technol. 10 (2) (2016) 282–294.
[41] G. Tsoumakas, I. Partalas, I. Vlahavas, A taxonomy and short review of ensemble selection, in: Workshop on Supervised and Unsupervised Ensemble Methods and Their Applications, 2008, pp. 1–6.
[42] D. Hernández-Lobato, G. Martínez-Muóz, A. Suárez, How large should ensembles of classifiers be? Pattern Recognit. 46 (5) (2013) 1323–1336.
[43] H.R. Bonab, F. Can, A theoretical framework on the ideal number of classifiers for online ensembles in data streams, 2016, pp. 2053–2056.
[44] H.R. Bonab, F. Can, Less is more: A comprehensive framework for the number of components of ensemble classifiers, CoRR (2017).
[45] T.G. Dietterich, Ensemble methods in machine learning, in: Multiple Classifier Systems, Springer, Berlin, Heidelberg, 2000, pp. 1–15.
[46] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[47] C. Ju, A. Bibaut, M. van der Laan, The relative performance of ensemble methods with deep convolutional neural networks for image classification, J. Appl. Stat. (2018) 1–19.
[48] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (9) (2015) 1904–1916, http://dx.doi.org/10.1109/TPAMI.2015.2389824.
[49] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: The IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016.
[50] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, 2016.
[51] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.
[52] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[53] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[54] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, Greedy layer-wise training of deep networks, Adv. Neural Inf. Process. Syst. (2007) 153–160.