Automated retinopathy of prematurity screening using deep neural networks

Jianyong Wang a,1, Rong Ju b,1, Yuanyuan Chen a, Lei Zhang a, Junjie Hu a, Yu Wu a, Wentao Dong c, Jie Zhong d,*, Zhang Yi a,**

a Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu, PR China
b Department of Neonatology, Chengdu Women & Children’s Central Hospital, Chengdu, PR China
c Department of Ophthalmology, Sichuan Academy of Medical Sciences and Sichuan Provincial People’s Hospital, Chengdu, PR China

A R T I C L E   I N F O

Article history:
Received 31 May 2018
Received in revised form 30 July 2018
Accepted 14 August 2018
Available online 27 August 2018

A B S T R A C T

Background: Retinopathy of prematurity (ROP) is the leading cause of childhood blindness worldwide. Automated ROP detection system is urgent and it appears to be a safe, reliable, and cost-effective complement to human experts.

Methods: An automated ROP detection system called DeepROP was developed by using Deep Neural Networks (DNNs). ROP detection was divided into ROP identification and grading tasks. Two specific DNN models, i.e., Id-Net and Gr-Net, were designed for identification and grading tasks, respectively. To develop the DNNs, large-scale datasets of retinal fundus images were constructed by labeling the images of ROP screenings by clinical ophthalmologists.

Findings: On the test dataset, the Id-Net achieved a sensitivity of 96.62%(95%CI, 92.29%–98.89%) and a specificity of 99.32% (95%CI, 96.29%–99.98%) for ROP identification while the Gr-Net attained sensitivity and specificity values of 88.46% (95%CI, 96.29%–99.98%) and 92.31% (95%CI, 81.46%–97.86%), respectively, on the ROP grading task. On another 552 cases, the developed DNNs outperformed some human experts. In a clinical setting, the sensitivity and specificity values of DeepROP for ROP identification were 84.91% (95%CI, 76.65%–91.12%) and 96.90% (95%CI, 95.49%–97.96%), respectively, whereas the corresponding measures for ROP grading were 93.33% (95%CI, 68.05%–99.83%) and 73.63% (95%CI, 68.05%–99.83%), respectively.

Interpretation: We constructed large-scale ROP datasets with adequate clinical labels and proposed novel DNN models. The DNN models can directly learn ROP features from big data. The developed DeepROP is potential to be an efficient and effective system for automated ROP screening.

Fund: National Natural Science Foundation of China under Grant 61432012 and U1435213.
© 2018 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Retinopathy of prematurity (ROP) is an ocular disease observed in infants with low birth weight, especially premature infants. The retinal blood vessels of the infant suffering from ROP develop abnormally [1], infants with low birth weight, especially premature infants. The retinal...
In these ROP screening programs, images of the retinal fundus are captured using fundus photography devices, such as RetCam3, and are used by the ophthalmologist to make a diagnosis. To implement the automated ROP detection system, an algorithm to analyze the retinal fundus images and detect ROP is desired. ROP detection is represented as a classification problem in machine learning. Traditional algorithms [9–12] analyze the retinal fundus images with handcrafted features, such as vessel dilation/tortuosity, venular features, and discern retinal from choroidal images. These methods require the specification of rules, which is unsuitable for ROP detection because of inadequate understanding of ROP symptomatology. Deep neural networks (DNNs) [13–15], a branch of artificial intelligence, have lately yielded impressive results, such as AlphaGo [16]. It avoids the aforementioned “feature engineering” by automatically learning latent features from big data, which is a large set of labeled data consisting of images and corresponding labels. Therefore, DNNs reduce the need for explicit features identified by experts. They have achieved success in a variety of applications of image analysis [3,16–24]. Most relevant research has been limited to ROP with plus–disease [3,21]. For example, DNNs including a pre-processing network with U-net like architecture and a classification network with Inception V1 like architecture achieved a sensitivity of 93% with 94% specificity for diagnosis of plus disease [21]. Generally speaking, ROP with plus-disease is much more distinguishable and many published datasets are limited to such problems. It is clear that plus is an important feature in ROP, however, it is not sufficient to define ROP by itself only. More efforts should be paid to construct datasets by taking more clinical features together, such as stage and zone, into account. Moreover, the number of the images in published datasets is not sufficient to develop robust DNNs models that can deliver satisfactory performance.

In this study, we constructed a large-scale ROP dataset with adequate clinical labels first. The labels are “normal”, “minor ROP”, and “severe ROP”, which are defined on the clinical features of plus, stage and zone together. In the clinical treatment, patients with “minor ROP” are required to have a reexamination after two weeks whereas patients with “severe ROP” are required to be treated by the clinical ophthalmologist immediately. The automated ROP detection system developed on such dataset would be more valuable in the clinical treatment than the systems developed on the dataset with “plus” and “non-plus” labels. Secondly, we developed two specific DNN models on the constructed dataset. Thirdly, we integrated the developed DNN models with cloud platform to implement the automated ROP screening system and tested the system with several methods.

2. Methods

In this study, we developed an automated ROP screening system using DNN models. It could be used for telemedicine ROP screening and facilitate cooperation among hospitals. In this section, the methods to construct the ROP detection datasets, develop and evaluate the DNN models as well as the automated ROP screening systems are presented.

2.1. Image labeling

Automated ROP detection system is urgent and it appears to be a safe, reliable, and cost-effective complement [9,10] to the efforts of ROP specialists, capable of increasing patient access to screening and focusing the resources of the current ophthalmic community on infants with potentially vision-threatening disease.
were constructed: an identification dataset and a grading dataset. In the subsequent non-blind phase, the annotations were evaluated by another more experienced clinical ophthalmologist if the assigned labels varied among the initial experts. The fourth ophthalmologist dropped the cases if diagnoses were not faithfully. The labels, i.e., “normal,” “minor ROP,” and “severe ROP,” used in this study were assigned not only based on the International Committee for the Classification of Retinopathy of Prematurity (ICROP) system [25], but also the requirements of clinical treatment described in Cryotherapy for Retinopathy of Prematurity (CRYO-ROP) and Early Treatment ROP (ETROP) [2,25,26]. The identification labels, “normal” and “ROP,” denote whether the clinical features of ROP found in the retinal fundus images. The grading labels, “minor ROP” and “severe ROP,” indicates the severity of the ROP disease. The clinical features of “minor ROP” are Zone II or III and stage 1 or 2, while the clinical features of “severe ROP” are threshold disease, type I, type II, or AP-ROP, and stage 4, or 5. The detail of the definitions of the clinical features of ROP can be found in the literatures [2,25,26].

2.2. Deep neural networks

ROP detection consists of two subtasks. One is an identification task, which aims to distinguish ROP cases from normal cases. The other is a grading task, where cases of ROP are classified into minor and severe. The two subtasks are binary classification problems in machine learning, which can be formulated as:

\[ F : X = \{ x_i | i = 1, 2, \ldots, n \} \rightarrow y \]

where \( n \) is the number of images of case \( X \) and \( x_i \) is the \( i \)-th image. \( y \in \{0,1\} \) is the corresponding label of \( X \). Our goal is to approximate the function \( F \).

The methods in the literature [3,21] cannot be applied directly into our work since the detection in this study was performed upon each cases rather than each image independently. There are four challenges making it difficult for human eye as well as DNN models to identify and grading the ROP through the retinal fundus images: [1] The number of the images varies in different cases depending on the patient and the photographer; [2] Images of the retinal fundus are usually poor in quality because of motion blur and illumination artifacts. The colors images are with significantly appearance variations; [3] Only a few images in ROP cases contain the distinguishable feature that determines ROP while the others are similar images between the different cases. [4] The feature that determines ROP is in some very small local regions of the image.

To deal with the above problems, an ROP detection algorithm is developed by using deep convolutional neural networks (see Fig. 2). An identification network (Id-Net) and a grading network (Gr-Net) were developed to approximate the mapping \( F \) for the aforementioned two subtasks, respectively. Id-Net and Gr-Net have a similar network architecture (see Fig. 3). It was constructed by introducing a feature-binding block to the well-known Inception–BN17 network, which delivered state-of-the-art performance in the large-scale image classification challenge “ImageNet.”

The proposed Id-Net and Gr-Net consist of three blocks. The first is a feature extractor block, which is identical to the subnetwork from the input layer to the second Inception-B block in the Inception–BN17 network. Given an case \( X \), the feature extractor block takes each image \( x_i \) one by one and outputs the corresponding feature map \( s_i \). The second block is the feature-binding block, which aims to bind the set of feature maps \( S = \{ s_i | i = 1, 2, \ldots, n \} \) into one feature map \( S(X) \). The feature fusing method used in our networks is the max operation, defined by \( S(X) = \max \{ s_j | j = 1, 2, \ldots, n \} \). The \( j \)-th element of the \( i \)-th feature map, and \( S(X) \) is the \( j \)-th element of feature map \( S(X) \). The feature-binding block can reduce negative effects of the varying number of images in each case \( X \), and can enhance the feature maps of distinguishable images while suppressing similar images that are found in many cases belonging to different categories. The third block is the classification block, which consists of several convolution layers and a softmax layer. It learns deep abstract features from feature map \( S(X) \), and maps the final feature into the corresponding label of the input case.

Id-Net and Gr-Net need to be trained on the large-scale dataset to obtain the network parameters. The training of a neural network refers to changing the parameters in it to improve network performance by deploying a learning algorithm, such as the backpropagation (BP) algorithm. The network performance is described by a cost function that measures the distance between the network’s output and the real label. After training, the neural network is expected to produce an output that is close to the corresponding real label. The training processes of Id-Net and Gr-Net were identical except for the training dataset used. Id-Net was trained on the identification dataset whereas the Gr-Net was trained on the grading dataset. The training process consisted of two phases. The first was offline pre-training. An Inception–BN network was trained on Image-Net. The learned parameters were used as
the initial parameters of Id-Net and Gr-Net, which were then fine-tuned on the corresponding ROP dataset. Here, the retinal fundus images were resized from 1200*1600*3 to 240*320*3 and the values of the pixels were normalized into range \([-1, 1]\). Cross-Entropy was used as cost function. The optimization method was AdaDelta [27] with the default parameters \(\text{lr} = 1.0, \rho = 0.9, \text{eps} = 1e^{-06}\). The proposed DNN models were implemented using Tensorflow (https://www.tensorflow.org) and the source code is available in Github as https://github.com/qmiwang/deeprop.

2.3. Experts comparisons

To further evaluate the performance of the developed DNN models, the retinal fundus images captured during routine clinical ROP Screenings in Jan. 2018 in Chengdu Women & Children’s Central Hospital were used to compare the DNN models’ predictions with the diagnoses of three human experts.

Having developed a labeling system, we instructed three experts with adequate clinical experience in ROP detection to make a referral decision on each case using only the retinal fundus image independently. Given a case, the true label was given by the voting of the labels made by three experts. It means that the label of a given case is obtained by choosing the label given by at least two experts. The confusion table and error rate of the DNN model and the three experts were calculated by referring the true labels. Furthermore, we computed the KAPPA values between the DNN model and the three experts respectively.

2.4. Testing of the DeepROP system under clinical settings

The ROP detection algorithm was integrated into a cloud computing platform for automated ROP detection in images of the retinal fundus. The platform is called DeepROP. It can be used for telemedicine ROP screening and cooperation among hospitals. In general, the workflow consists of the following stages (see Fig. 4):

Step1: Uploading images of the retinal fundus. In a hospital that has ROP screening devices, prematurely born infants undergo ROP screening programs. The photographer, doctor, or anyone who can obtain images of the retinal fundus can access the automated ROP detection platform and upload the images through a website.

Step2: Automated ROP detection. The automated ROP detection platform saves the retinal fundus images and automatically divides them into two cases: images of the left eye and those of the right eye. Given a case, Id-Net takes the images as input and generates the ROP identification results. The detection process is complete if the given case is determined to be normal. Otherwise, Gr-Net is executed to determine the severity of ROP. Finally, an initial diagnostic report is generated automatically. This process does not require any human assistance. It eliminates the drawbacks stemming from a lack of adequately trained clinicians for ROP identification.

Step3: ROP expert evaluation. The diagnostic report contains a confidence score for the automated ROP detection platform. If the confidence score is below a threshold, the results should be evaluated by human experts. They can check the report online, and one report can be evaluated by several experts. Our cloud-based platform thus supports cooperation among hospitals for ROP diagnosis.

The website was assessed in six hospitals for routine ROP screening programs. The clinical ophthalmologist uploaded images captured during routine clinical ROP screening programs without any data selection or preprocessing. The DeepROP system would generate the diagnose automatically. By referring the diagnosis by the clinical ophthalmologist as the standard label, the specificity and sensitivity of the model were presented.

3. Results

In this study, we constructed a large-scale ROP datasets of retinal fundus images and developed two DNN models for the detection of ROP. One is Id-Net which aims to distinguish the normal case from ROP case and the other is Gr-Net which classify the ROP cases into either Minor ROP case or Severe ROP case. The two DNN models were trained and evaluated on the constructed dataset and then integrated into a cloud based platform for automated ROP detection, which was assessed in the routine ROP screening in multiple hospitals in China.

3.1. Large-scale ROP datasets

We collected lots of images of ROP screenings from the Chengdu Women & Children’s Central Hospital. Following the proposed image labeling method, there were 349 cases drop in the identification dataset and 222 cases drop in the grading task. The proposed datasets consist of three parts: developing data to build the DNN models, data for expert comparison, and data collected from web. The number of cases and images are listed in the Table 1. The cases in developing data and data for

![Fig. 3. The architecture of the DNNs used in the algorithm. The feature-binding block was implemented by an element-wise max operation. The architectures of the Conv Factory, Inception A, and Inception B modules are identical to the corresponding modules in the Inception–BN17 network.](image)

![Fig. 4. The workflow of developed DeepROP system in the clinical test. The cloud-based platform consists of the proposed DNN models and the DeepROP website as the user interface.](image)
expert comparison were collected from screenings of 869 infants. The summaries of the birth weights and gestation ages of the infants are presented in Fig. 5. As expected, most of them had low birth weights or were premature. There were 93 boys and 78 girls while the genders of other infants were unknown since they were not recorded during the ROP screening.

To the best of our knowledge, this is the largest dataset of images of the retinal fundus for ROP detection. A comparison of our dataset with published datasets is listed in Table 2. Our dataset was more heterogeneous as it contained images of more patients and more cases of ROP screening. It would be a good tutorial for ROP education because it ensures coverage of the various instances of different ROP severities. Furthermore, the scale of the dataset is key to the success of automated ROP grading methods, such as DNNs. The labels used in our dataset were designed according to the clinical treatment of ROP patients in Sichuan, China. It is thus more suitable for patients in China.

3.2. Performance of the model

To train and evaluate DNN models, the developing dataset was split randomly. 298 cases (149 normal and 149 ROP) and 104 cases (52 minor ROP and 52 severe ROP) were selected randomly from the developing dataset for evaluating the performance of Id-Net and Gr-Net, respectively. Other cases were used to train Id-Net and Gr-Net. To deal with the imbalance problems, during the training, we oversampled the classes with less cases to ensure the number of the cases are same in each task.

By referring to annotations provided by four ophthalmologists, the proportions of false positives and missed cases of the optimized DNNs were recorded. The results were presented in Fig. 6, Table 3, and Table 4. In the identification task, the developed Id-Net distinguished normal cases from ROP cases with a sensitivity of 96.64% (95%CI, 92.34%–98.90%) and a specificity of 99.33% (95%CI, 96.32%–99.98%). The area under the ROC curve was 99.49%. Using Gr-Net, the algorithm provided ROP grading attained sensitivity and specificity values of 88.46% (95%CI, 76.56%–95.65%) and 92.31% (95%CI, 81.46%–97.86%), respectively. The area under the ROC curve was 95.08%.

3.3. Comparison of the model with human experts

The dataset for expert comparison were the retinal fundus images captured during routine clinical ROP in Jan. 2018 in Chengdu Women & Children's Central Hospital to compare the developed DNN’s referral decisions with the decisions made by human experts. There are 2361 images of 552 cases from 264 patients. The distributions of the normal cases, minor ROP cases, and severe ROP cases were presented in Table 1. The inter-rater KAPPA values among the DNN models and three experts were presented in the Table 5. Furthermore, the confusion tables and error rates of the DNN models and the three experts are presented in Tables 6–9, and Fig. 7. The developed DNN models outperformed the expert 1.

Furthermore, the averagetime per case used by the DNNs is around 2s while it is around 30–60 s for human experts. The detection prediction by the DNN models is very efficient.

3.4. DeepROP website and clinical test

The proposed ROP detection algorithm was integrated with a cloud computing platform for automated ROP detection and report generation.
by the proposed networks, sample search and download, report evaluation and interaction between patients and ophthalmologists.

The website was assessed in six hospitals for routine ROP screening programs, including Chengdu Women & Children’s Central Hospital, Xin Hua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Ya An People’s Hospital, People’s Hospital of De Yang City, Affiliated Hospital of North Sichuan Medical College, Affiliated Hospital of Southwest Medical University. The clinical ophthalmologist uploaded images captured during routine clinical ROP screening programs without any data selection or preprocessing. The developed DNN models achieved good performance. From March 3 to December 28, 2017, 472 ROP screenings were conducted on 404 infants in Chengdu Women & Children’s Central Hospital. The clinical ophthalmologist uploaded images from these 472 screenings to the website. DeepROP automatically divided the images into 944 cases according to eye. For each case, it made two predictions for identification and grading, respectively, and generated an initial diagnostic report that was evaluated by a clinical ophthalmologist. The ophthalmologist edited the report if his/her diagnosis was inconsistent with the initial report by DeepROP. For comparison, the diagnosis by the clinical ophthalmologist was referred to as the standard label. The ophthalmologist diagnosis and DeepROP predictions were then recorded. The performance of DeepROP is shown in Table 10.

### 4. Discussions

Automated ROP grading algorithms are desirable for ROP screening and treatment. Dataset plays a crucial role in the training of automated ROP grading algorithms. In general, a dataset used for DNNs must contain a large number of images of ROP screenings that cover various ROP features and different severities, and have been captured from an adequately large number of patients to reduce the individual effect and avoid the over-fitting of the learning algorithm to some specific

---

**Table 3**

| Predicted label | Normal | ROP |
|-----------------|--------|-----|
| True label      | Normal | 148 | 1 |
|                 | ROP    | 5   | 144 |

**Table 4**

| Predicted label | Minor ROP | Severe ROP |
|-----------------|-----------|------------|
| True label      | Minor ROP | 48         | 4 |
|                 | Severe ROP | 6          | 46 |

**Table 5**

|                   | DNN Models | Expert1 | Expert2 | Expert3 | Label |
|-------------------|------------|---------|---------|---------|-------|
| DNN Models        | 1          | 0.34    | 0.70    | 0.70    | 0.68  |
| Expert 1          | 0.34       | 1       | 0.32    | 0.42    | 0.53  |
| Expert 2          | 0.70       | 0.32    | 1       | 0.69    | 0.80  |
| Expert 3          | 0.70       | 0.42    | 0.69    | 1       | 0.89  |
| Label             | 0.68       | 0.53    | 0.80    | 0.89    | 1     |

**Table 6**

| Predicted Label | Normal | Minor | Severe |
|-----------------|--------|-------|--------|
| True Label      | Normal | 487   | 11     | 3      |
|                 | Minor  | 6     | 16     | 9      |
|                 | Severe | 2     | 1      | 17     |

**Table 7**

| Predicted label | Normal | Minor | Severe |
|-----------------|--------|-------|--------|
| True Label      | Normal | 499   | 0      | 2      |
|                 | Minor  | 15    | 16     | 0      |
|                 | Severe | 5     | 15     | 0      |

**Table 8**

| Predicted label | Normal | Minor | Severe |
|-----------------|--------|-------|--------|
| True Label      | Normal | 492   | 9      | 0      |
|                 | Minor  | 1     | 20     | 10     |
|                 | Severe | 0     | 0      | 20     |
feature. Otherwise, the performance of the learning algorithm worsens when applied to new images of ROP screening. The datasets constructed in this study are advantageous in terms of the above properties over other datasets. Moreover, both the identification dataset and grading dataset were provided, which enabled the process to grade cases of ROP for accurate treatment, rather than simply identify ROP.

The results showed that DNNs can be trained using large-scale datasets, without the need to specify features by experts, to automatically detect ROP in images of the retinal fundus with high sensitivity and accuracy. Compared with human experts with adequate clinical experience on ROP diagnosis, the developed DNNs obtained comparable performance and made the detection prediction very efficiently. Further, it provides the same diagnosis on a given image every time, which is difficult for a human ophthalmologist. Multiple DNN models can be trained and integrated to simulate the consultation of several experts as well. This will be tested in our future work. Furthermore, the proposed method in this study is more suitable to be deployed in the automated ROP screening system than the method in the literature [21]. Because our method detects the ROP per case no matter how many images there are.

The clinical setting test highlighted the impressive performance of the DeepROP platform. Results suggest that whilst ROP identification worked very well prospectively, the ROP grading does not do so well. Maybe there are two reasons. One is that the features between “minor ROP” cases and “severe ROP” cases are less distinguishable than the features between “Normal” cases and “ROP” cases. The other is that the number of cases used to develop the Gr-Net is smaller than the number of cases used to develop the Id-Net so that the generalization performance of the Gr-Net is poorer than the Id-Net. The clinical setting test is an important step in telemedicine ROP screening programs and multihospital collaboration for several reasons. First, it enables ROP pre-screening and evaluation of patients in non-specialized hospitals. Second, it breaks the bottleneck of data usage due to the isolation of data in individual hospitals and accelerates the process of collecting a large-scale dataset for fine-grained ROP grading. Third, the cloud-based platform can collect new data, which will increase the coverage of ROP features, and improve the ROP detection algorithm. For example, in the Web-based test in this study, 4908 images of 944 cases of 404 infants were collected in 10 months. The size of those data items was larger than published datasets.

The limitations of our works lie on three aspects. Firstly, the number of “severe ROP” cases is not enough to ensure the generalization performance in the clinical test. Secondly, the developed system can only evaluate the severity of ROP. In future work, we plan to collect more data and grade ROP in more fine-grained classes, such as the “plus-disease,” “stage,” and “zone.” Thirdly, an open question is whether the predictions of the cloud platform will influence the diagnoses of ophthalmologists in comparison with clinical testing without the platform. More attention should be paid to test the platform more widely in future.

Acknowledgment

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

Declarations of interest

The authors declare no competing financial interests.

Author contributions

Zhang Yi, Jie Zhong, and Rong Ju designed the research; Zhang Yi, Jianyong Wang, Yuanyuan Chen, Lei Zhang, Junjie Hu, and Yu Wu implemented the detection system and analyzed the data; Rong Ju and Jie Zhong were responsible for the retinal fundus images collection and labeling, respectively; Jianyong Wang and Rong Ju co-wrote the manuscript; all authors discussed the results and commented on the manuscript.

References

[1] Aetaer-Cansizoglu E, Kalpathy-Cramer J, You S, Keck K, Endogmus D, Chiang MF. Analysis of underlying causes of inter-expert disagreement in retinopathy of prematurity diagnosis. Application of machine learning principles. Methods Inf Med 2015;54(1): 93–102.
[2] Early Treatment for Retinopathy of Prematurity Cooperative Group. Revised indications for the treatment of retinopathy of prematurity: results of the early treatment for retinopathy of prematurity randomized trial. Arch Ophthalmol 2004;121(12): 1684–96.
[3] Worrall DE, Wilson CM, Brostow GJ. Automated retinopathy of prematurity case detection with convolutional neural networks. In: Carneiro G, et al, editors. Deep Learning and Data Labeling for Medical Applications. LABELS 2016, DLIMA 2016. Lecture Notes in Computer Science. Springer, 2016, p. 68–76.
[4] American Academy of Ophthalmology. Ophthalmologists warn of shortage in specialists who treat premature babies with blinding eye condition, 2008.
[5] Kemper AR, Wallace DK. Neonatologists/practitioners’ practices and experiences in arranging retinopathy of prematurity screening services. Pediatrics 2007;120(3): S27–31.
[6] Pararajasegaram R. Vision 2020 Global Initiative for the Elimination of Avoidable Blindness. Action Plan 2006–2011. Commun Eye Health 1998;11(26): 29.
[7] Steinikuller PG, Du L, Gilbert C, Foster A, Collins ML, Coats DK. Childhood blindness. JAAPSO 1999;3(1): 26–32.
[8] Gilbert C, Fielder A, Cordillo L, et al. Characteristics of infants with severe retinopathy of prematurity in countries with low, moderate, and high levels of development: Implications for screening programs. Pediatrics 2005;115(5): 518–25.
[9] Rajkumar ER, Rani P, Rajanani KT, Kandemir M, Singh D. Multiple instance learning: Robust validation on retinopathy of prematurity. IJCTA 2016;9(36): 451–9.
[10] Aetaercansizoglu E, Boloncano V, Campbell JP, et al. Computer-based image analysis for plus disease diagnosis in retinopathy of prematurity: Performance of the “i-rop” system and image features associated with expert diagnosis. Transl Vis Sci Technol 2015;4(6): 5.
[11] Wilson CM, Cocker KD, Moseley MJ, et al. Computerized analysis of retinal vessel width and tortuosity in premature infants. Invest Ophthalmol Vis Sci 2008;49(8): 3577–85.

Table 9
The confusion table of the expert 3 on the dataset used for the comparison of the model with experts.

| True label | Normal | Minor | Severe |
|------------|--------|-------|--------|
| Normal     | 489    | 12    | 0      |
| Minor      | 0      | 31    | 0      |
| Severe     | 0      | 0     | 20     |

Fig. 7. The error rates of the experts and the developed DNN model.

Table 10
Summaries of the performance of DNNs used in DeepROP website in clinical test.

| Id-Net         | Gr-Net         |
|----------------|----------------|
| Sensitivity    | 84.91%(95%CI, 76.65%–91.12%) | 93.33%(95%CI, 68.05%–99.83%) |
| Specificity    | 96.50%(95%CI, 95.49%–97.96%) | 73.63%(95%CI, 68.05%–99.81%) |
| Accuracy       | 95.55%(95%CI, 94.03%–96.77%) | 76.42%(95%CI, 67.18%–84.12%) |
[12] Binenbaum G. Algorithms for the prediction of retinopathy of prematurity based on postnatal weight gain. Clin Perinatol 2013;40(6):261–70.
[13] Lecun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44.
[14] Werbos P. Generalization of backpropagation with application to a recurrent gas market model. Neural Netw 1988;1(4):339–56.
[15] Zhang L, Yi Z, Amari S, et al. Theoretical study of oscillator neurons in recurrent neural networks. IEEE Trans Neural Netw Learn Syst 2018. https://doi.org/10.1109/TNNLS.2018.2793911 (Online Early Access).
[16] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of go without human knowledge. Nature 2017;550(7676):354–9.
[17] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. ICML 2015:448–56.
[18] Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: a neural image caption generator. CVPR 2015:3156–64.
[19] Er R, Rani P. Classification of retinopathy of prematurity using back propagation neural network. Int J Biomed Eng Technol 2016;22(04):338.
[20] Guishan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316(22):2402–10.
[21] Brown JM, Campbell J, Beers A, et al. Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. JAMA Ophthalmol 2018;136(7):803–10.
[22] Kermansy D, Goldbaum M, Cai W, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-based Deep Learning. Cell 2018;172:1122–31.
[23] Lee S, Baeghman M, Lee Y. Deep learning is effective for the classification of OCT images of normal versus Age-related Macular Degeneration. Ophthalmology Retina 2017;1(4):322–7.
[24] Poplin R, Varadarajan A, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nature Biomed Eng 2018;2:158–64.
[25] Moskowitz A, Hansen RM, Fulton AB. Retinal, visual, and refractive development in retinopathy of prematurity. Eye Brain 2016;8(1):103–11.
[26] Quinn G. The international classification of retinopathy of prematurity revisited. Arch Ophthalmol 2005;123(7):591–9.
[27] Zeiler MD. ADADELTA: An Adaptive Learning Rate Method. arXiv preprint arXiv: 1212.5701; 2012.