IEEE Transactions on Medical Imaging Volume: 38 , Issue: 1 , Jan 2019

Automated Analysis for Retinopathy of Prematurity by Deep Neural Networks

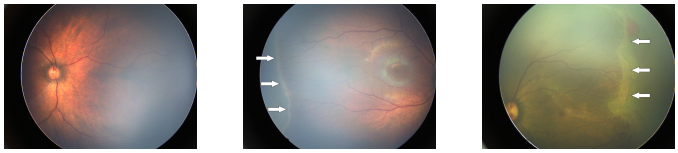
Junjie Hu, Yuanyuan Chen, Jie Zhong, Rong Ju, Zhang Yi, *Fellow, IEEE*

***Abstract*—Retinopathy of Prematurity (ROP) is a retinal vasproliferative disorder disease principally observed in infants born prematurely with low birth weight. ROP is an important cause of childhood blindness. Although automatic or semi-automatic diagnosis of ROP has been conducted, most previous studies have focused on "plus" disease, which is indicated by abnormalities of retinal vasculature. Few studies have reported methods for identifying the "stage" of ROP disease. Deep neural networks have achieved impressive results in many computer vision and medical image analysis problems, raising expectations that it might be a promising tool in automatic diagnosis of ROP. In this paper, convolutional neural networks (CNNs) with novel architecture is proposed to recognize the existence and severity of ROP disease per-examination. The severity of ROP is divided into mild and severe cases according to the disease progression. The proposed architecture consists of two sub-networks connected by a feature aggregate operator. The first sub-network is designed to extract high-level features from images of the fundus. These features from different images in an examination are fused by the aggregate operator, then used as the input for the second sub-network to predict its class. A large dataset imaged by RetCam 3 is used to train and evaluate the model. The high classification accuracy in the experiment demonstrates the effectiveness of proposed architecture for recognizing ROP disease.**

***Index Terms*—Retinopathy of Prematurity, deep neural networks, feature aggregate operator, medical image analysis.**

I. Introduction

As the primary cause of childhood blindness, Retinopathy of Prematurity (ROP) is an eye disease occurs frequently in infants with low birth weight and premature birth [1]. ROP was initially known as Retrolental Fibroplasia (RLF), and originally observed by Terry in the 1940s [2]. Nowadays, it is widely accepted that ROP is closely associated with excessive oxygen use. During gestation, the development of blood vessels begins in the fourth month of gestation, reaching the retinal periphery before birth [3]. For premature infants, relative hyperoxia in the extrauterine environment and the continuous supply of oxygen can slow down the growth rate of retinal vasculature and lead to tissue hypoxia. Retinal neovascularization may then develop at the joint between vascular and avascular areas, producing scar tissue causing retinal detachment through retraction [4].



(a) Normal. (b) Stage 2 of ROP. (c) Stage 3 of ROP.

Fig. 1. Fundus photographs imaged by RetCam 3. From left to right are normal, Stage 2, and Stage 3 of ROP respectively. There is a obvious ridge at the junction between vascularized and avascular retina in both Stage 2 and 3 of ROP. In Stage 3, fibrovascular proliferation can be observed from the ridge into the vitreous.

As the harm caused by this potentially blinding disorder has become clear, an international group formed by ROP experts published a detailed classification guideline in 1984 and 1987 to facilitate the development of clinical treatment and improve understanding of the condition [5], [6]. First, the guideline defined three zones to better describe the location of the illness in ROP with each zone centered on the optic disc. Second, five stages of ROP and a type of ancillary illness called "plus" were proposed. Symptoms of Stage 1 to 5 are listed in Table I. Plus disease occurs in conjunction with ROP, and is characterized by increased dilation and tortuosity in retinal vessels. Fig. 1 shows images of the fundus in normal development, and at Stage 2, and Stage 3 of ROP. According to the reported recommendations [7], any stage of ROP with plus in Zone I, Stage 3 of ROP without plus in Zone I, and Stage 2 or 3 with plus in Zone II require early treatment. There was no appropriate treatment for ROP until the 1980s and 1990s when laser photocoagulation and cryotherapy were shown to be effective methods for preventing blindness in infants [4]. Although these therapies can reduce the incidence of blindness in infants, they also impact patients' visual acuity. Early diagnosis and timely treatment can help to reduce the adverse outcomes and vision loss [8].

Diagnosis of ROP requires inspecting the fundus of premature infants from different views using imaging systems like RetCam 3, which is a digital retinal camera with high image quality. The imaging data are then interpreted by experienced ophthalmologists to determine the presence of the symptoms of ROP or plus disease described above. However, the diagnosis can be challenging for several reasons. First, developing countries such as China and India have an insufficient number of qualified ophthalmologists to match the number of premature infants [9], [10]. In these countries, there is a pressing

Junjie Hu, Yuanyuan Chen, and Zhang Yi are with the Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu 610065, P. R. China (e-mail:zhangyi@scu.edu.cn).

Jie Zhong is with the Department of Ophthalmology, Sichuan Academy of Medical Sciences and Sichuan Provincial Peoples Hospital, Chengdu 610072, P. R. China.

Rong Ju is with Department of Neonatology, Chengdu Women & Children's Central Hospital, Chengdu 610031, P. R. China.

J. Hu and Y. Chen are co-first authors. J. Zhong, R. Ju, and Z. Yi are co-corresponding authors.

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

TABLE I
SYMPTOMS OF STAGE 1 TO 5 OF ROP

Stage	Symptoms
1	a thin demarcation line separates vascularized and avascular areas
2	line in Stage 1 evolves to a ridge
3	extraretinal fibrovascular proliferation in the ridge
4	partial retinal detachment
5	total retinal detachment

need for ophthalmologists. Second, the quality of imaging is affected by many factors (e.g., focus, illumination, and eyes movement). Third, the classification guidelines provide only qualitative signs rather than quantitative descriptions. Thus, clinical assessment mainly depends on the ophthalmologist's subjective interpretation of the symptoms [11]. As a result, there exists disagreement in diagnoses between different experts assessing the same examination. This uncertainty has been reported when diagnosing the presence of plus disease and the stage of ROP [12], [13] . To assist ophthalmologists in the diagnosis of ROP, a number of computer-aided diagnosis systems have been proposed. Most of the proposed systems have focused on detecting plus disease, which is an illness that co-occurs with ROP that can be quantified [14]. However, few of them focused on the automated stages classification of ROP.

In the current study, a novel methodology based on deep neural networks is proposed to automatically recognize ROP in fundus images. Assessment of ROP is accomplished in two steps. The first step is designed to recognize whether ROP disease is present. If ROP is recognized, the second step is to assess the severity of the disease. Both steps are required to analyze examination data, which involve variable number of images in different views. This recognition method faces several challenges: 1) Unlike the characteristics of plus disease indicated by the dilation and tortuosity of posterior veins, the stage of ROP is characterized by the demarcation line between vascularized and avascular areas. The locations and shapes of the lines in the fundus vary significantly in each examinations. 2) A massive amount of annotated data is required to learn the features of ROP from the data. However, the datasets in current ROP-related researches cannot meet this requirement. 3) Each examination of the fundus contains a number of images with different views. To obtain the ROP recognition result, the model must jointly analyze multiple images in an examination.

To address the first challenge, convolutional neural networks (CNNs) with powerful abstract ability are used. CNNs have been widely applied in computer vision related problems, and have been shown to learn high-level features directly from data. Meanwhile, transfer learning is used in the paper to facilitate the training phase. Transfer learning is an effective method to train the very deep CNNs when the target dataset is small. It has been widely used in medical imaging applications, and performs better than models trained from scratch [15]. To

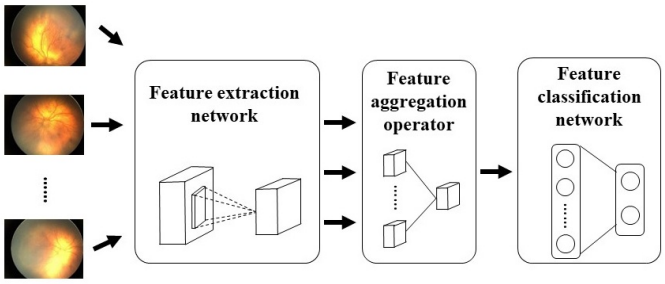


Fig. 2. The proposed architecture with two sub-networks for recognizing of ROP. The input is multiple fundus images of an examination. The first sub-network is used to extract features in multiple images in an examination and the second one is used to classify the ROP disease. The extracted features from the multiple images in first sub-network is aggregated before fed to the second sub-network.

solve the second challenge, a large scale dataset annotated by some experienced clinical ophthalmologists is used. To our best knowledge, the training dataset used in the current study is larger than that of the previous ROP-related studies by an order of magnitude. The large dataset contribute to learn the disease-related characteristics in CNNs and reduce the overfitting. To address the third challenge, a novel architecture of CNNs is proposed. Two sub-networks are included in the model: the first is designed to extract features from multiple images in an examination and the second is for classification. To jointly analyze the extracted features, an aggregate operator is used for binding features from the first sub-network. The architecture of the proposed model is shown in Fig. 2. Several architectures of CNNs pre-trained on ImageNet are explored, including the VGG (Visual Geometry Group) [16], Inception [17], [18] and Residual Networks (ResNets) [19], [20].

The current study makes several new contributions:

(i) A large dataset labeled by some experienced clinical ophthalmologist is used for automatically diagnosing ROP. The training dataset is larger than that of the previous ROP-related studies by an order of magnitude.
(ii) A novel ROP recognition architecture is proposed. The architecture contains two sub-networks which aim to extract and classify high-level features in a data-driven manner.
(iii) Feature aggregation operator is used to bind the features from different images in an examination. CNNs using the operator have superior accuracy compared with other state-of-the-art methods.
(iv) Two tasks are performed in the current study, including recognition of the existence and severity of ROP. The classification and visualization results revealed the proposed architecture successfully learned the characteristics of ROP, providing a potentially useful tool to aid clinicians in diagnosing ROP disease.

II. RELATED WORKS

In this section, an overview of previous studies using traditional methods for ROP diagnosis is presented, followed by a brief introduction to deep neural networks and their application in ROP.

A. Traditional Methods for Diagnosis of ROP

The vast majority of automated or semi-automated methods for ROP diagnosis are focused on the recognition of plus disease, which is important for identifying infants with severe ROP disease. Because the existence and severity of plus disease are defined by the abnormality of vessels, most of these methods have attempted to measure statistics of vessels in fundus, such as diameter and tortuosity. Typically, three main steps are involved: (a) vessels segmentation; (b) measurement of the vessel's diameter (thickness); (c) measurement of the vessel's tortuosity [11]. The segmentation step requires accurate identification of the vascular tree from the retinal image, and the following two steps are based on the segmented vessels. Combined with these three steps, many computer-aided systems (CAD) have been proposed to assist ophthalmologists in improving diagnostic accuracy of ROP.

For example, a system called "ROPTool" has been proposed [21] to assist ophthalmologist in diagnosing plus disease. Using this system, operators first determine the area containing the vessels to be analyzed, which enables the system to track the vessels automatically using the "ridge/valley traversal" method [22]. To ameliorate overestimation, tortuosity was calculated as the total length of the vessel divided by the generated smooth curve instead of a straight line. Dilation was calculated by the average of the width over its length, divided by the area of the optic nerve. Based on the calculated values, operators can diagnose the existence of plus disease in a more quantitative way. "i-ROP" [23] was a system designed to grade plus disease into three types: normal, pre-plus, and plus. Principal spanning forests [24] algorithm was used to extract the vessels. It go a step further beyond "ROPTool", using 11 indices to quantify the tortuosity and dilation, including Cumulative Tortuosity Index (CTI), Integrated Curvature (IC), Integrated Squared Curvature (ISC), etc.

Although traditional methods have been found to aid diagnosis of ROP, there remain several challenges. Most to be solved, the precision of the measurement heavily relies on the vessel segmentation, meaning that errors in segmentation may be amplified in subsequent measurements. For measuring the diameter and tortuosity of the segmented vessels, the location of measurement and the effects of magnification differences may also impact the results [11].

B. Deep Neural Networks for Diagnosis of ROP

Deep neural networks [16], [25] have received much interest in the field of machine learning. Two types of neural networks, including feed-forward neural networks (FNNs) [26], [27] and recurrent neural networks (RNNs) [28]–[31], have been heavily studied during the last decades. A recent interesting result on RNNs can be found in [32]. Since 2012, when AlexNet [25] has won the ILSVRC-2012 competition [33], many important breakthroughs in computer vision have been achieved using deep neural networks. Several critical factors contribute to these achievements, including novel network architecture [16], [17], [19], [31], powerful computation ability by utilizing graphics processing units (GPUs), large-scale annotated dataset, etc. Compared with traditional classification methods using hand-craft features, it extracts different levels of features from low to high as the networks going deeper in a data-driven manner. Numerous studies have explored deep neural networks in a range of medical image analysis applications, including mitosis detection [34], lymph node detection [35], lung pattern classification [36], and breast cancer classification [37], etc.

A recent study used an ImageNet pre-trained GoogLeNet to classify the existence of plus disease in ROP [14], constituting the first attempt to use deep neural networks to diagnose plus disease. Two types of classification tasks were explored in the study, including the per-image and per-examination classifications. In the per-image classification, the researchers fine-tuned the convolutional kernel in the 9th inception block along the last fully-connected layer. Based on the per-image classification, the researchers also proposed a per-examination classifier by assuming the Beta distribution prior over the probability that an examination is diagnosed with plus disease. Both the per-image and per-examination classifiers have superior performance than those of previous methods, demonstrating that CNNs may provide a promising tool for diagnosing ROP. Meanwhile, the visualization results revealed that CNNs have successfully learned the abnormalities in vessels correlated to plus disease.

The main characteristic in the ROP recognition is that there are variable number of fundus images in an examination. To solve this problem, Worrall et al. [14] assumed the Beta distribution prior in per-examination classification. However, the assumption of Beta distribution is unfavorable in the generalization of the learned classifier because the parameters α and β in Beta distribution can't be tuned by the gradient-based optimization method.

According to previous study [7], besides plus disease, the stage of ROP is an important factor indicating whether early treatment is needed. In the current study, a novel architecture based on the neural network model is proposed to identify ROP disease in per-examination manner. Unlike Worrall et al.'s method [14], the proposed model learns to recognize ROP directly from the data, instead of using a predefined assumption. The recognition is accomplished in two steps. The first step is to classify the existence of the ROP disease, and the second is to identify its severity. The proposed model has been trained and evaluated on a dataset of 2668 examinations, which is larger than the data in previous studies by an order of magnitude. To our best knowledge, the current study is the first attempt to use a large dataset recognizing the existence and severity of ROP using deep neural networks.

III. Data and Methodology

In this section, the used dataset is first described in detail. The characteristics of the dataset provide a better understanding of the task. Then the proposed model used in the two-steps classification tasks are illustrated, including the architecture of networks and the feature aggregation operator.

A. Data

Before the data is used to train and evaluate the proposed model, three steps are needed: data imaging, data annotation,

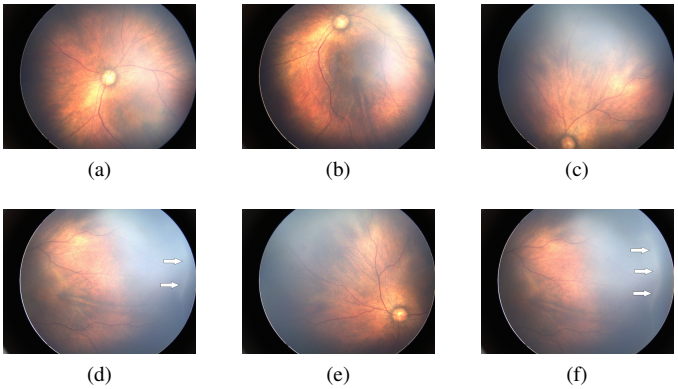


Fig. 3. Multiple fundus images from different shooting angles in an examination of the left eye. An obvious ridge can be seen in Fig.3d and Fig.3f at the marked arrows, while Fig.3a, Fig.3b, Fig.3c, and Fig.3e appeared normal from visual inspection.

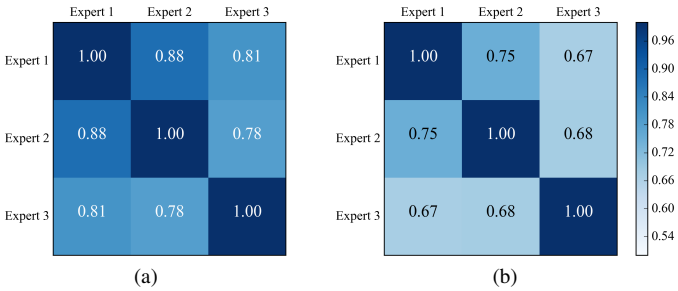


Fig. 4. The quadratic-weighted kappa score between the three ophthalmologists. (a) and (b) represent the first and second annotations phases, respectively.

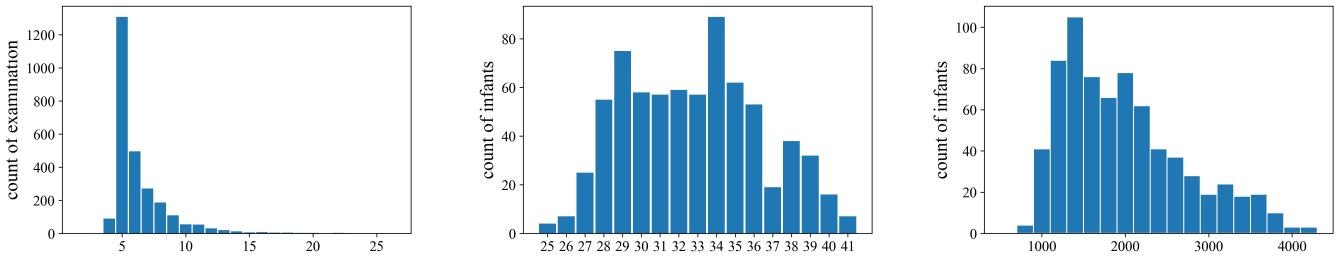
and data partition. The first step acquires fundus images in examinations, and the second annotates the examinations by experienced ophthalmologists. The annotated examinations are then partitioned into training, validation, and testing datasets in the third step. These three steps are described in detail in the following sections.

Data Imaging: Images of the fundus are obtained using RetCam 3 from the Chengdu Women and Children's Central Hospital (WCCH) from 2014 to 2017. The RetCam 3 can only capture one fixed view of the fundus at a time. To observe the fundus thoroughly, the operator typically take multiple images of the fundus of the infant's eye in an examination. An examination of the left eye in Stage 2 of ROP is presented in Fig. 3. In this examination, the ridge in the fundus can be observed in (d) and (f). In the current study, 3017 examinations are collected and 349 examinations are excluded in the annotation phase, which yield the final 2668 examinations from 720 infants. Each examination contains a variable number of images per eye. The resolution of the image is 1600 $\times$ 1200. The numbers of images per examination, gestation age, and birth weight are plot as histogram in Fig. 5. The number of images per examination varies from 2 to 26, and the most frequently occurring number of images is 5. The gestation ages varies from 25 to 41 weeks, with a mean value of 32 weeks. 45% infants' gestation age is under 32 weeks. The maximum, minimum, and mean birth weights are 4250, 700, and 1994 grams, respectively. 32% of the infants' birth weight is less than 1500 grams.

Data Annotation: The reference standard of the annotation is comply with the symptoms described in Table I. The process of the annotation is splitted into two phases: first, the ophthalmologists annotated examinations into normal and ROP types, followed by annotation of the severity of ROP. Both the two phases are annotated by three experienced ophthalmologists from the department of ophthalmology in Sichuan Academy of Medical Sciences and Sichuan Provincial Peoples Hospital. One of the annotator is the chief physician that have more than ten years of clinical experience of ROP. The other two annotators are both doctors that have over five years of clinical experience. In the first phase, the examinations with the consistent labels among the three ophthalmologists are picked out. The intersection process can minimize the subjective bias and reduce the risks caused by carelessness. Base on the examinations annotated as ROP in the first phase, the second phase requires the ophthalmologists to identify these examination's stage. Similar with two previous studies [12], [38], a high diagnosis variability was observed among experts due to the subjective assessment. To ameliorate the potential affects of the bias, only the examinations with consistent labels among the three ophthalmologists are used. The quadratic-weighted kappa score between the three ophthalmologists in the two annotation phases are shown in the Fig. 4. It can be seen that the first two ophthalmologists have higher agreement than that of the third ophthalmologist in both annotation phases.

According to the annotation results, a high level of data imbalance is observed, where most ROP data are in Stage 2 and 3. There are relatively few ROP data in Stages 1, 4, and 5. There are two potential explanations for this imbalance: 1) The demarcation line separates the avascular and vascularized areas in both Stage 1 and 2, although the line is wider in Stage 2. These relative differences are determined subjectively, and the ophthalmologists tended to annotate them as Stage 2. 2). In Stage 4 and 5, retinal detachment can be observed. This phenomenon is rarely found in the current dataset because effective intervention will be carried out before the disease become severe. To solve the problem of imbalance, the ROP data are further divided into mild and severe cases according to the phase of stage. Stage 1 and 2 are classified as mild, while Stage 3, 4, and 5 are classified as severe. This type of grading is consistent with previous study [5], [39], in which Stage 3 is an important phase between the growth of the demarcation line and the detachment of the retina. The last row in Table II shows the number of annotated examinations. There are 2668 identified examinations from 720 infants, including 1484 and 1184 examinations in normal and ROP respectively. It should be noted that only examinations with consistent labels among the three annotators are included in the table, so the sum of mild and severe ROP cases is less than 1184.

Data Partition: The dataset used for training, evaluation, and testing the model are split in random and shown in the first three rows in Table II. In the classification of normal and



(a) Number of images in examinations. (b) Examined infant's gestation age. (c) Examined infant's birth weight.

Fig. 5. Histograms of the number of images per examination, gestation age, and birth weight.

TABLE II
DATASET USED FOR TRAINING, EVALUATION, AND TESTING THE MODEL

	Normal	**ROP**	**Mild**	**Severe**
Train set	1184	884	225	241
Validation set	150	150	50	50
Test set	150	150	50	50
Total	1484	1184	325	341

ROP, 150 examinations of normal and ROP are used as the validation and test dataset, respectively. The examinations on the left are used as training data. In the classification of mild and severe ROP, 50 examinations of mild and severe cases are used as the validation and test datasets, and the remaining ROP examinations are used as training data.

B. Methodology

The inputs of our model are the annotated dataset $D = \{\boldsymbol{x}_{i,j}, \boldsymbol{y}_j; i = 1, ..., \tilde{N}, j = 1...M\}$ containing M instances of examination, with corresponding labels. $\boldsymbol{x}_{i,j} \in \mathbb{R}^{w \times h \times c}$ where w, h, and c denote the width, height, and channels of the fundus images, and i, j denote the i-th image in j-th examination. The $\tilde{N}$ varied from j since each examination could contain different numbers of fundus images. The goal is to learn a robust model $f(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$ parameterized with $\boldsymbol{\theta}$, which mapping the input image space $\mathcal{X}$ to the target space $\mathcal{Y}$.

Architecture of Networks: The model explored in this work is based on the CNNs. Three types of layers are typically included in CNNs: convolutional, pooling, and fully-connected layers. In the convolutional layers, the parameters to learn are the kernels that connect with the input locally. The convolutional operation for single channel can be formulated as:

$$z_{v,u}^{l+1} = \sum_{p=0}^{P_l-1} \sum_{q=0}^{Q_l-1} \sum_{k=0}^{K_l-1} w_{p,q,k}^{l} \cdot a_{v+p,u+q,k}^{l}, \tag{1}$$

where the P, Q, and K denote the dimensions of the kernel, and the lower-case letters denote the cursor in the kernel. v and u denote the spatial location of the output $\boldsymbol{z}^{l+1}$. The shared kernel $\boldsymbol{w}^l$ convolves the input $\boldsymbol{a}^l$ along its dimensions of width and height to obtain the output. Then, a non-linear activation function F is applied to $\boldsymbol{z}^{l+1}$,

$$\boldsymbol{a}^{l+1} = F(\boldsymbol{z}^{l+1}). \tag{2}$$

Pooling is another important layer in CNNs, aiming to reduce the dimensionality of the inputs, thus decreasing the computational complexity. Two types of pooling are commonly used, including the max and mean pooling. In the pooling operation, the filter with fixed size slides over the spatial dimensions of input feature maps in a certain stride. During each slide process, the max or mean is calculated when max or mean pooling is used, respectively. The fully-connected layer usually appears in the bottom of the CNNs. Differ from the convolutional layer, which is locally connected, each neuron in the fully-connected layer has connections with all the neurons in upper layer.

Fine-tuning with pre-trained networks has been proved as an effective method for training CNNs. The pre-training of CNNs denotes the use of another dataset (e.g ImageNet) to train a model parameterized with $\tilde{\boldsymbol{\theta}}$ by minimizing the loss function. Then, the parameters in the pre-trained model are used to initialize the model $f(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$ in the current task. In the current study, several ImageNet pretrained networks are explored, including VGG [16], Inception [17], and Residual Networks [19]. The VGG network is stacked with multiple convolutional layers with very small kernel (3×3) and max pooling. The VGG network shows that the very small kernels are efficient for constructing CNNs. Inception is a kind of module that consists of max pooling and convolutional layer with the kernel size of 1×1, 3×3, and 5×5. To save computational resources, 1×1 convolutions are adhered to max pooling, 3×3, and 5×5 convolutions. The intuition behind the Inception module is let the model itself to learn the optimal structure among the different kinds of operations. To overcome the difficulties in training very deep CNNs, residual network has proposed another type of module that composed of the residual and cross-layer shortcut connection. The module can be formulated as $\boldsymbol{a}^{l+1} = F^l(\boldsymbol{a}^l) + \boldsymbol{a}^l$, where $F^l(\boldsymbol{a}^l)$ and $\boldsymbol{a}^l$ denote the residual and shortcut connections, respectively.

Feature Aggregate Operator: Unlike traditional image classification task that the input to CNNs is a single image, the input in the ROP recognition is an examination containing $\tilde{N}$

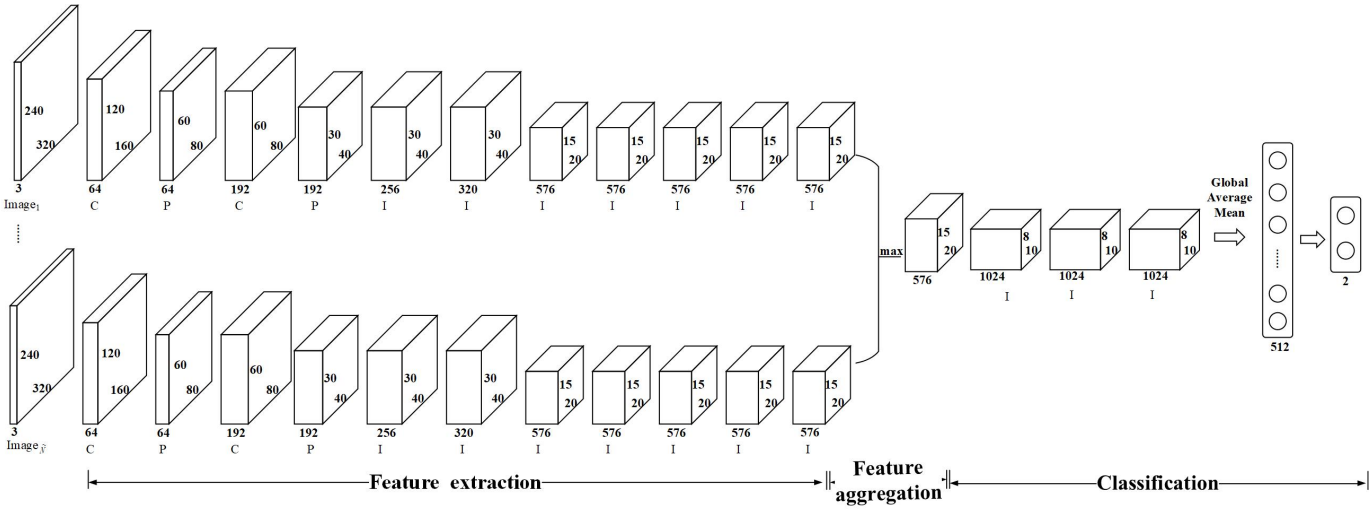


Fig. 6. Architecture of Inception-V2 with feature aggregate operator of max in module 2. The capital letter C, P, and I denote the Convolutional, Pooling, and Inception operations respectively. The values represent the number of channel, width, and height of the feature maps.

variable images. This requires the model to make decisions based on multiple images. Consider the examination of ROP shown in Fig.3, misdiagnose will happen when the model extracts features only from images in Fig.3a, Fig.3b, Fig.3c, or Fig.3e. To fully utilize the information in each image of an examinations, the model should learn to aggregate features from all of those images. Then the model is supposed to complete the recognition with the use of the aggregated features. However, *how* and *where* to aggregate features are two problems should be solved. *How* represent the strategies used to aggregate the features. This is critically important for the model's accurate recognition because the disease-related characteristics may only appeared in some of the images in an examination. *Where* denote the aggregate location in the model. This is another major factor since it is unclear which is the optimal abstract level of the disease-related characteristics.

For the problem of *how*, inspired by the aggregation strategies used in fusing the spatial and temporal features [40], *max* and *mean* feature aggregate operators are explored in current study. The *max* and *mean* operators compute the max and average value of the $\tilde{N}$ features at the same spatial location, respectively. They can be formulated as

$$\tilde{\boldsymbol{a}}_j^l = \max_{i \in [1, \tilde{N}_j]} \boldsymbol{a}_i^l \tag{3}$$

and

$$\tilde{\boldsymbol{a}}_j^l = \frac{1}{\tilde{N}_j} \sum_{i=1}^{\tilde{N}_j} \boldsymbol{a}_i^l, \tag{4}$$

where $\boldsymbol{a}_i^l$ denotes the features of image i, $\tilde{N}_j$ denotes the number of images of examination j, and $\tilde{\boldsymbol{a}}_j^l$ denotes the aggregated features.

For the problem of *where*, a CNNs based sub-network is first used to extract features from the variable number of images in an examination. To reduce the parameters to learn, the first sub-network is shared among the $\tilde{N}$ images in an examination. It has been proved that as the layers going deeper, higher level features are extracted in CNNs. Some previous studies [19], [20] have shown the deeper and wider networks contribute to better abstract ability of the CNNs. To explore the optimal abstract level to aggregate, different l are tested. Based on the aggregated features, a second sub-network is able to make the final prediction. The Inception-V2 network with feature aggregate operator in module 2 is presented in Fig. 6.

Training Method: CNNs with feature aggregate operators are trained with the back-propagation algorithm by minimizing the following cross-entropy cost function with respect to the parameters $\boldsymbol{\theta}$:

$$\mathcal{L} = -\frac{1}{M} \sum_{j=1}^{M} \boldsymbol{y}_j^\top ln(\boldsymbol{a}_j^L), \tag{5}$$

where $\boldsymbol{a}^L$ denotes the output of the network after applying the softmax function. The cross-entropy cost function represents the similarities between the true distributions of labels and the approximated distributions of the network. The Adadelta [41] algorithm is used to minimizing the cost function. Differ from the stochastic gradient descend (SGD) algorithm which has the fixed learning rate, Adadelta is an adaptive weight updates optimization method based on the first order information. The Adadelta algorithm contains two parameters: one is the initial learning rate and the other is decay rate used in the moving averages of the squared gradient.

IV. Experimental Setup and Results

In this section, the experimental setup is presented, including the chosen evaluation strategy and implementation of the proposed method. The experimental results are then presented in detail.

A. Experimental Setup

Configurations: The input of the model is RGB images in the size of $1600 \times 1200 \times 3$. To saving computational resources, the original images are resized to $320 \times 240 \times 3$ with the OpenCV library [42] through the bilinear interpolation.

TABLE III
THE OUTPUT SIZE OF EACH MODULE IN VGG-16, INCEPTION-V2, AND RESNET-50 NETWORKS. THE MODULE IN EACH NETWORK IS STACKED BY BLOCKS. THE INPUT IMAGES ARE DOWN-SAMPLED TO 40×30 AND 80×60 IN THE INCEPTION-V2 AND RESNET-50 RESPECTIVELY, THROUGH THE MAX POOLING AND CONVOLUTION WITH STRIDE 2. FURTHER DETAILS OF THE NETWORK ARCHITECTURE ARE PROVIDED IN SEVERAL PREVIOUS STUDIES [16], [18], [20]

Modules	VGG-16	Inception-V2	ResNet-50
module1	320×240	40×30	80×60
module2	160×120	20×15	40×30
module3	80×60	10×8	20×15
module4	40×30	\	10×8
module5	20×15	\	\

The resized images are then divided by 255, ensuring the pixel value is located into 0 and 1. The number of images per each examination is set to 12. For the examination with number of images less than 12, existing images are randomly chosen several times to round up 12. Otherwise, 12 images are randomly selected from the examination. The parameters of the Adadelta optimizer are set according to the suggested values where the initial learning rate and the decay rate are 1.0 and 0.95, respectively. The weight updates are performed in mini-batches where the number of examinations per batch is 5. The training process finished when up to 100 epochs.

Modern CNNs are typically constructed using blocks (e.g Inception or Residual block) and divide into several modules. A module is composed of several blocks, and the feature maps in a module have identical sizes. To explore the effect of network architecture on the ROP recognition task, the VGG-16, Inception-V2, and ResNet-50 networks are tested. For VGG-16, the last two fully-connected layers are replaced by the global average pooling [43] which average the feature maps along spatial dimension to reduce overfitting. Table III shows the output size of each module in the networks. For each module in a specific network, the feature maps of the last block are aggregated.

Implementation: The proposed method is implemented by TensorFlow [44]. All experiments are performed using a server with Linux OS and hardware of CPU Intel Xeon E5-2620 @ 2.4GHz, GPU NVIDIA Tesla K40m, and 64 GB of RAM.

Evaluation Metrics: The training process is carried out on the training set, while the validation set is used to fine-tune the model. The overall performance of each model is assessed on the test set. Different metrics are calculated, including Accuracy, Sensitivity, Specificity, Precision, and F_1-score. The Receiver Operating Characteristic (ROC) and the Area Under Curve (AUC) are calculated to compare performances between models.

B. Results

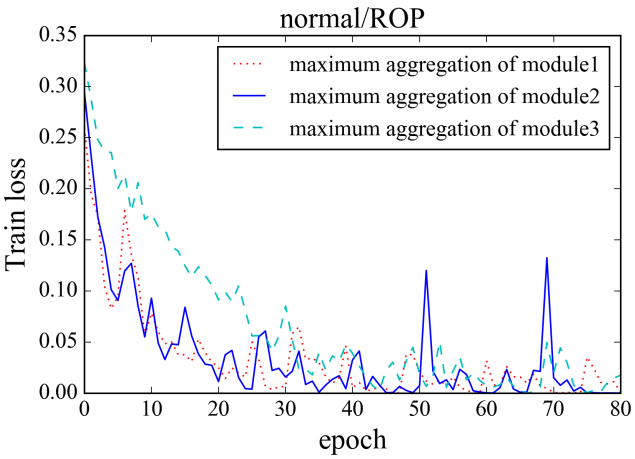
1) Feature Aggregate Operators: The performance of the proposed feature aggregate operators is described below. Table IV shows the test accuracy of the proposed method for classification of normal/ROP and mild/severe type of ROP. Performance between different networks is first compared. As seen in the table, Inception-V2 exhibit superior performance compared with VGG-16 and ResNet-50 in the two tasks. The high classification accuracy of Inception-V2 mainly benefited from the multiple operations in the Inception block (e.g., convolutional kernel size of 1×1, 3×3, and 5×5). These operations can extract variable features in multiple scales. For VGG-16 and ResNet-50, although the two CNNs are designed to be deep enough to extract high-level features of the inputs, their performance oscillated more than that of Inception-V2 when aggregate in different modules. This is mainly caused by the fixed size of convolutional kernel (e.g., 3×3).

TABLE IV
TEST ACCURACY OF THE PROPOSED METHOD IN DIFFERENT NETWORK ARCHITECTURES FOR CLASSIFICATION OF NORMAL/ROP AND MILD/SEVERE ROP

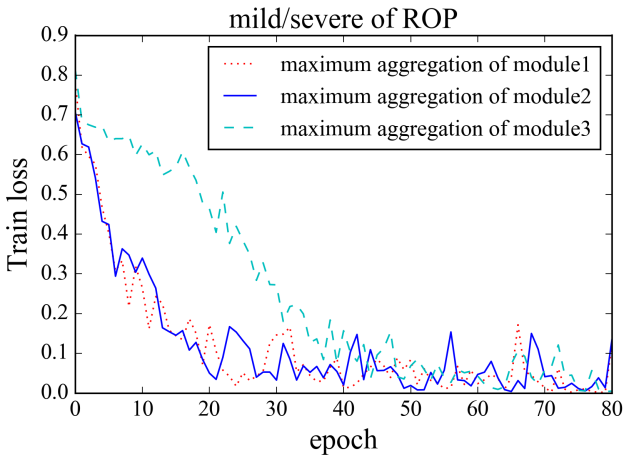
Networks	Modules	Normal and ROP		Mild and Severe	
		max	mean	max	mean
VGG-16	module1	0.853	0.903	0.670	0.650
	module2	0.903	0.910	0.720	0.710
	module3	0.920	0.943	0.660	0.710
	module4	0.910	0.950	0.670	0.810
	module5	0.900	0.923	0.770	0.700
Inception-V2	module1	0.963	0.953	**0.840**	0.690
	module2	**0.970**	0.967	**0.840**	0.670
	module3	0.946	0.956	0.790	0.750
ResNet-50	module1	0.893	0.886	0.720	0.680
	module2	0.920	0.900	0.820	0.620
	module3	0.930	0.880	0.820	0.680
	module4	0.933	0.880	0.810	0.800

For the max and mean aggregation operators in Inception-V2, it can be seen from Table IV that the max operator exhibit higher performance, except for the module 3 in classification of normal and ROP cases. Comparison of the performance of different modules in Inception-V2 in classification of normal and ROP cases reveals the optimal feature for aggregation is in the middle of the network (e.g., module 2 with max aggregation operator yields the highest test accuracy of 0.97). Both module 1 and 2 with the max aggregation operator achieved accuracy of 0.84 in classifying mild and severe of ROP cases. It should be noted that test accuracy in the classification of normal and ROP is much higher than that of mild and severe ROP cases, indicating the latter task is substantially more difficult for the CNNs. This is consistent with the clinical diagnosis. Recognition of the severity of ROP is an intractable problem even for the experienced ophthalmologists.

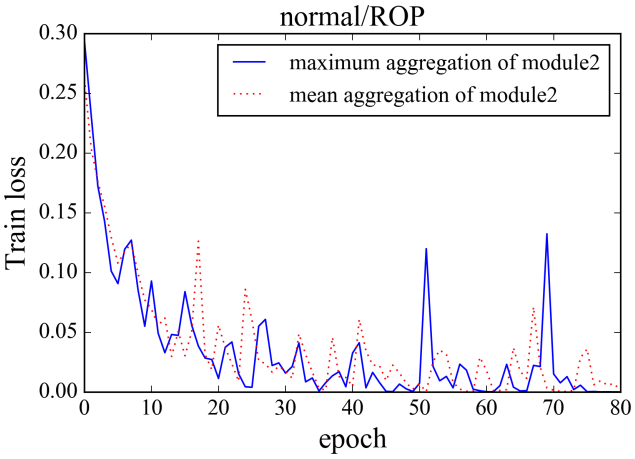
Fig. 7 shows the training loss in Inception-V2 between the max and mean aggregation operators. The different aggregation modules in the max operator are also compared. Fig. 7a and Fig. 7b show that the convergence speed of module 1 and module 2 are higher than those of module 3 in the



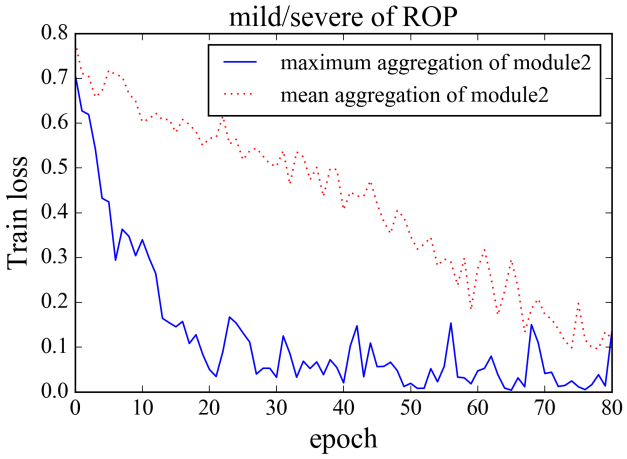
(a) Training loss of Inception-V2 with max aggregation in classification of normal and ROP cases.



(b) Training loss of Inception-V2 with max aggregation in classification of mild and severe of ROP cases.



(c) Training loss of Inception-V2 with max and mean aggregation in classification of normal and ROP cases.



(d) Training loss of Inception-V2 with max and mean aggregation in classification of mild and severe of ROP cases.

Fig. 7. Comparison of training loss in Inception-V2 with different configurations.

two classification tasks, suggesting that the classification sub-network in Fig. 2 is important for learning. Although the Inception-V2 with max aggregation operator in module 2 achieves the highest test accuracy of 0.97, greater oscillation can be observed after the 50th epoch. From Fig. 7c, it can be seen that the convergence speed of the max operator is slightly higher than the mean in the classification of ROP and normal cases. The difference is magnified in the classification of mild and severe of ROP cases in Fig. 7d, where the training loss of max aggregation declined rapidly compared with the mean.

2) Comparison With the State of the Art: Table V shows a comparison of the proposed model with the first automated ROP detection system [14], which achieved better performance than traditional hand-craft features. To implement the method proposed by Worrall et al., a per-image classifier is established by assuming the images in an examination shared the same type of label. The classifier is based on an ImageNet pre-trained GoogLeNet and optimized by the RMSProp [45] algorithm, as in Worrall et al.'s study. The model exhibiting the highest accuracy with the validation set is used in the per-examination classifier. For classification of normal and ROP cases, the proposed model exhibited superior performance on all metrics. The test accuracy of the proposed model is 3% higher than that of Worrall et al.'s method, and the F1 score is also considerably higher. For classification of mild and severe of ROP cases, the proposed model outperform Worrall et al.' method on most metrics except sensitivity, where the two models exhibit the same level of performance. The sensitivity and specificity of Worrall et al.'s method are 0.82 and 0.64 respectively, suggesting the model is preferable for predicting

TABLE V
COMPARISON OF THE PROPOSED MODEL WITH WORRALL ET AL.' METHOD

Metrics	Normal and ROP		Mild and Severe of ROP	
	Worrall et al.	Proposed	Worrall et al.	Proposed
Raw Acc	0.940	**0.970**	0.730	**0.840**
Sensitivity	0.926	**0.960**	**0.820**	**0.820**
Specificity	0.953	**0.980**	0.640	**0.860**
Precision	0.952	**0.979**	0.694	**0.854**
F1	0.939	**0.969**	0.752	**0.836**

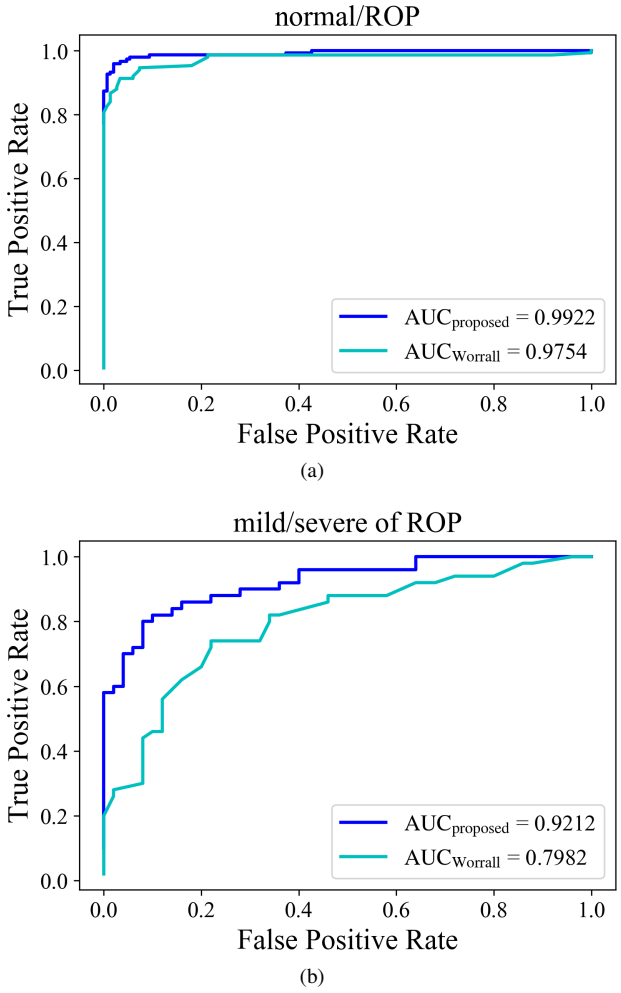


Fig. 8. ROC analysis for the proposed model and Worrall et al's in the classification of normal/ROP and mild/severe of ROP.

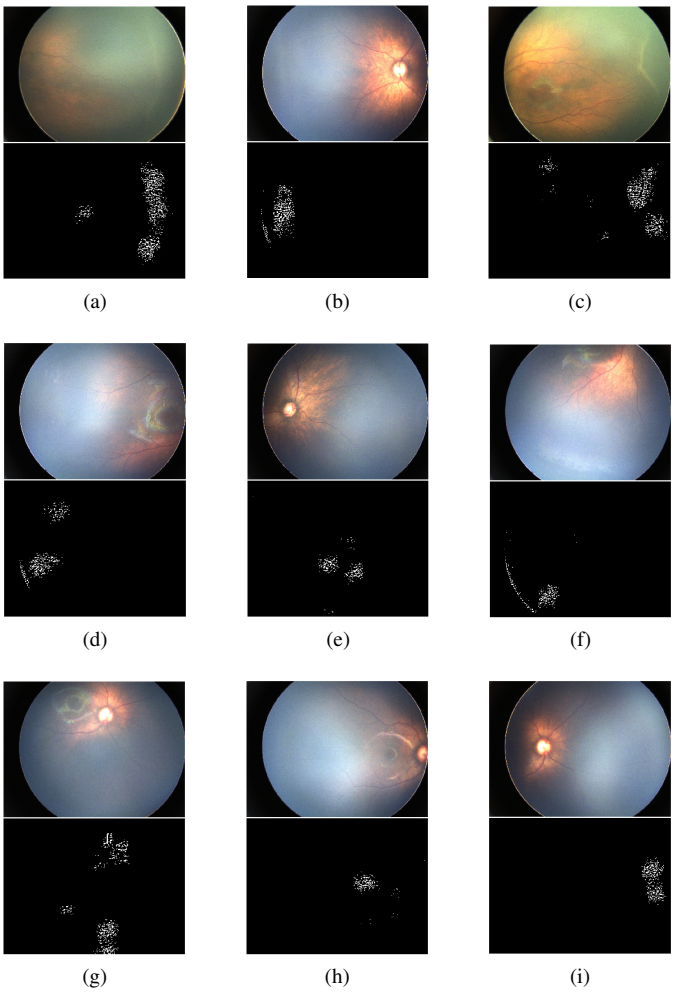


Fig. 9. Visualization of the input images in the test dataset that most activate the softmax layer with the guided backpropagation algorithm. The three rows represent the images of the true positive, false negative, and false positive examinations.

severe ROP. In the terms of F1 score, the performance of the proposed model is almost 11% higher than that of Worrall et al.'s method.

For a more detailed comparison at different operating points, ROC analysis is performed. Fig. 8 shows the ROC curves and the AUC values for the proposed model and Worrall et al.'s method. In the classification of normal and ROP cases, the proposed model achieves superior performance. The ROC values of the proposed model and Worrall et al.'s method are 0.9922 and 0.9754, respectively. The superiority of the proposed model is more clearly observed in the classification of mild and severe of ROP cases, in which the AUC value of the proposed model is 15% higher than that in Worrall et al.'s method.

There are several possible explanations accounting for the superior performance of the proposed model compared with the method reported by Worrall et al.: (i) In the per-image classifier, Worrall et al. assumed that the images from an examination have the same type of label because the labels are based on the examination instead of the image. This assumption can't be fully verified, because the images are taken from different views and with different artifacts. In some cases, it is difficult to determine whether an image exhibits disease-related characteristics or not from visual (e.g., 3a, 3b, 3c, and 3e in Fig.3). (ii) In Worrall et al.'s model, recognition of ROP in an examination is based on the Beta distribution, which is determined by statistics based on the number of images classified as healthy and diseased in the training data. The parameters of Beta distribution should be explicit set up and can't be tuned by the gradient-based optimization method. In the current paper, the proposed model is trained to learn the disease-related characteristics from the data. Because the proposed method does not require prior knowledge about the distribution of the dataset, it may be more generalizable.

3) Visualization: Visualization of the input images that most activate the softmax layer is presented in Fig.9. The results is based on the guided backpropagation algorithm [46] in the classification of normal and ROP cases. The guided backpropagation algorithm computes the gradient of the activation of the specific neuron with regard to the inputs. The negative gradients which have inhibitory impact on the target neuron are masked out. Note that to improve visualization performance, the generated gradient images are binarized. The images in the first row are from true positive examinations. By comparing

the input images with the corresponding visualization results, it is clear that the output of the softmax layer is highly correlated with the ridge area in the input images. This is consistent with the guidelines for clinicians, in which the ridge in the fundus is important for the diagnosis of ROP. The visualization results demonstrate that the proposed model learned to extract the essential features for the diagnosis of ROP, despite the shapes and orientations of the ridge. For example, the model recognized the ridge in Fig. 9a and Fig. 9b, locating at the left and right parts of images, respectively. Even when the ridge has a different shape, the model is still able to recognize it (e.g., Fig. 9c).

The second row in Fig. 9 shows images from a false negative examination. The Fig. 9d shows that the ridge located in the top-left part of the image is successfully identified, along with the down-left area which has no obvious disease-related features from the visual. The fundus image in Fig. 9f shows apparent ROP-indicative features. However, the model failed to accurately recognize the all of the lesions, but only the left part of it. The images belong to a false positive examination can be seen at the third row of Fig. 9. As shown in Fig. 9g, the model associate the prediction result with the optic disc and the lower area in the fundus image. From the Fig. 9h and Fig. 9i, it can be seen that the model falsely recognize the reflection of light as the ridge of ROP, leading to an incorrect prediction result.

4) Analytic Experiments: In this sub-section, two experiments are performed to validate the proposed model. One is to test the impact of the number of ROP-related images on the model, and the other is to examine the model's classification performance of examinations from premature infants.

The model's performance change according to the number of ROP-related images in test dataset is illustrated in Table VI. The second row in Table VI represents the number of examinations annotated as ROP case. The third row represents the number of misclassified examinations. It can be seen from the table that there are two and three examinations with one and two ROP-related images misclassified, respectively. The model misclassified one examination with four ROP-related images. Examinations with more than five ROP-related images are all correctly predicted as ROP.

To further validate the proposed model, 406 examinations collected in January 2018 from 195 premature infants (birth weight ≤ 2500 grams or gestational age ≤ 28 weeks) are used as test dataset after the annotation phases. The data annotation phases are kept the same as described in Section III. These premature infants have not appeared in the dataset used to construct the model. Fig. 10a and Fig. 10b show the confusion matrix of the normal/ROP and mild/severe ROP classification tasks, respectively. It can be seen from Fig. 10a that the model accurately recognized most normal examinations (352 out of 356), but misclassified five ROP examinations. This is mainly due to the diversity of the characteristics of ROP. In classification of mild/severe of ROP, the model only misclassified one examination in each class. The results further demonstrate the effectiveness of the proposed model in diagnosing ROP.

TABLE VI
STATISTICS OF THE NUMBER OF ROP-RELATED IMAGES IN TEST DATASET.

Number of ROP-related images	1	2	3	4	5	6	≥7
Number of examinations annotated as ROP	35	38	20	24	10	12	11
Number of wrong predicted examinations	2	3	0	1	0	0	0

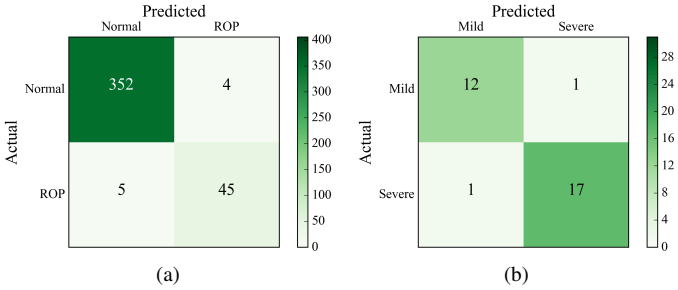


Fig. 10. The confusion matrix of the two classification tasks. (a) and (b) represent the normal/ROP and mild/severe of ROP classification tasks, respectively.

V. CONCLUSION

In this paper, a novel architecure of CNNs is proposed to recognize the existence and severity of ROP. The architecture is composed of a feature extract sub-network, followed by a feature aggregate operator to bind features from variable images in an examination. The prediction is accomplished using a second sub-network with the aggregated features as inputs. Max and mean aggregate operators are explored based on the architecture. Several ImageNet pretrained networks are tested in the study, including VGG-16, Inception-V2, and ResNet50. The proposed architecure is verified with a large dataset of 2668 examinations of the fundus in infants. The experimental results demonstrate that the Inception-V2 with the max aggregate operator in module 2 is a proper network architecture for the recognition of the existence and severity of ROP. Compared with the mean aggregate operator, the max has better classification accuracy and convergence speed. Meanwhile, a patient's multiple examinations in train, validation, and test datasets has little impact on model's performance, mainly because the characteristics of the eyes of the premature infants are varied over time.

The visualization results demonstrate that the proposed architecture learned the clinical characteristics of ROP, despite the location and shape of the ridge in ROP. However, the reflection of light in the image may impact the recognition result of the model. The proposed model also outperformed the state-of-the-art, verifying the effectiveness of our proposed architecture. In future studies, we will extend the method to diagnose the plus disease in ROP and integrate the recognition of stage and plus disease to aid ophthalmologist in clinical diagnosis.

REFERENCES

[1] W. Tasman, A. Patz, J. A. Mcnamara, R. S. Kaiser, M. T. Trese, and B. T. Smith, "Retinopathy of prematurity: The life of a lifetime disease," *American Journal of Ophthalmology*, vol. 141, no. 1, pp. 167–174, 2006.

[2] T. L. Terry, "Extreme prematurity and fibroblastic overgrowth of persistent vascular sheath behind each crystalline lens from the massachusetts eye and ear infirmary," *American Journal of Ophthalmology*, vol. 25, no. 2, pp. 203–204, 1942.

[3] A. M. Roth, "Retinal vascular development in premature infants," *American Journal of Ophthalmology*, vol. 84, no. 5, pp. 636–640, 1977.

[4] J. Chen and L. E. Smith, "Retinopathy of prematurity," *Angiogenesis*, vol. 10, no. 2, pp. 133–140, 2007.

[5] Committee for the Classification of Retinopathy of Prematurity, "An international classification of retinopathy of prematurity," *Pediatrics*, vol. 74, no. 1, pp. 127–133, 1984.

[6] ICROP Committee for Classification of Late Stages ROP, "An international classification of retinopathy of prematurity," *Pediatrics*, vol. 82, no. 1, pp. 37–43, 1988.

[7] W. V. Good, "Final results of the early treatment for retinopathy of prematurity (etrop) randomized trial," *Transactions of the American Ophthalmological Society*, vol. 102, pp. 233–250, 2004.

[8] J. P. Campbell, J. Kalpathycramer, D. Erdogmus, P. Tian, D. Kedarisetti, C. Moleta, J. D. Reynolds, K. Hutcheson, M. J. Shapiro, and M. X. Repka, "Plus disease in retinopathy of prematurity: A continuous spectrum of vascular abnormality as a basis of diagnostic variability," *Ophthalmology*, vol. 123, no. 11, pp. 2338–2344, 2016.

[9] G. S. Yau, J. W. Lee, V. T. Tam, S. Yip, E. Cheng, C. C. Liu, B. C. Chu, and I. Y. Wong, "Incidence and risk factors for retinopathy of prematurity in multiple gestations: A chinese population study," *Medicine*, vol. 94, no. 18, pp. 185–191, 2015.

[10] P. K. Shah, V. Prabhu, S. S. Karandikar, R. Ranjan, V. Narendran, and N. Kalpana, "Retinopathy of prematurity: Past, present and future," *World Journal of Clinical Pediatrics*, vol. 5, no. 1, pp. 35–46, 2016.

[11] T. Aslam, B. Fleck, N. Patton, M. Trucco, and H. Azegrouz, "Digital image analysis of plus disease in retinopathy of prematurity," *Acta Ophthalmologica*, vol. 87, no. 4, pp. 368–377, 2009.

[12] A. Gschlieer, E. Stifter, T. Neumayer, E. Moser, A. Papp, N. Pircher, G. Dorner, S. Egger, N. Vukojevic, and I. Oberacher-Velten, "Inter-expert and intra-expert agreement on the diagnosis and treatment of retinopathy of prematurity," *American Journal of Ophthalmology*, vol. 160, no. 3, pp. 553–560, 2015.

[13] M. F. Chiang, L. Jiang, R. Gelman, Y. E. Du, and J. T. Flynn, "Interexpert agreement of plus disease diagnosis in retinopathy of prematurity," *Archives of Ophthalmology*, vol. 125, no. 7, pp. 875–880, 2007.

[14] D. E. Worrall, C. M. Wilson, and G. J. Brostow, "Automated retinopathy of prematurity case detection with convolutional neural networks," in *International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, 2016, pp. 68–76.

[15] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[18] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[20] K. He, X. Zhang, and S. Ren, "Identity mappings in deep residual networks," in *European Conference on Computer Vision*, 2016, pp. 630–645.

[21] D. K. Wallace, Z. Zhao, and S. F. Freedman, "A pilot study using "roptool" to quantify plus disease in retinopathy of prematurity," *Journal of American Association for Pediatric Ophthalmology and Strabismus*, vol. 11, no. 4, pp. 381–387, 2007.

[22] S. R. Aylward and E. Bullitt, "Initialization, noise, singularities, and scale in height ridge traversal for tubular object centerline extraction," *IEEE Transactions on Medical Imaging*, vol. 21, no. 2, pp. 61–75, 2002.

[23] E. Ataer-Cansizoglu, V. Bolon-Canedo, J. P. Campbell, A. Bozkurt, D. Erdogmus, J. Kalpathy-Cramer, S. Patel, K. Jonas, R. P. Chan, S. Ostmo *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," *Translational Vision Science Technology*, vol. 4, no. 6, p. 5, 2015.

[24] E. Bas, E. Ataer-Cansizoglu, D. Erdogmus, and J. Kalpathy-Cramer, "Retinal vasculature segmentation using principal spanning forests," in *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on*, 2012, pp. 1792–1795.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[26] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, p. 533, 1986.

[27] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[28] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural computation*, vol. 1, no. 2, pp. 270–280, 1989.

[29] Z. Yi and K. K. Tan, *Convergence analysis of recurrent neural networks*. Norwell, MA, USA: Kluwer Academic Publishers, 2004.

[30] Z. Yi, "Foundations of implementing the competitive layer model by lotka–volterra recurrent neural networks," *IEEE Transactions on Neural Networks*, vol. 21, no. 3, pp. 494–507, 2010.

[31] J. Wang, L. Zhang, Y. Chen, and Z. Yi, "A new delay connection for long short-term memory networks," *International Journal of Neural Systems*, 2017.

[32] L. Zhang, Z. Yi, and S. I. Amari, "Theoretical study of oscillator neurons in recurrent neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. PP, no. 99, pp. 1–7, 2018.

[33] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F. F. Li, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[34] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab, "Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1313–1321, 2016.

[35] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[36] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1207–1216, 2016.

[37] G. Carneiro, J. Nascimento, and A. P. Bradley, "Automated analysis of unregistered multi-view mammograms with deep learning," *IEEE Transactions on Medical Imaging*, vol. 36, no. 11, pp. 2355–2365, 2017.

[38] V. Bolón-Canedo, E. Ataer-Cansizoglu, D. Erdogmus, J. Kalpathy-Cramer, O. Fontenla-Romero, A. Alonso-Betanzos, and M. F. Chiang, "Dealing with inter-expert variability in retinopathy of prematurity: a machine learning approach," *Computer Methods and Programs in Biomedicine*, vol. 122, no. 1, pp. 1–15, 2015.

[39] C. A. Ricard, D. Cel, and O. Dammann, "Screening tool for early postnatal prediction of retinopathy of prematurity in preterm newborns (step-rop)." *Neonatology*, vol. 112, no. 2, pp. 130–136, 2017.

[40] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," *arXiv preprint arXiv:1604.06573*, 2016.

[41] D. Kingma and J. Ba, "Adadelta: An adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[42] G. Bradski, "The opencv library," *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, vol. 25, no. 11, pp. 120–123, 2000.

[43] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[44] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, and M. Devin, "Tensorflow: large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[45] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural Networks for Machine Learning*, vol. 4, no. 2, pp. 26–31, 2012.

[46] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: the all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.