tvst

Special Issue

Ophthalmologist-Level Classification of Fundus Disease With Deep Neural Networks

Ping Jiang[1,2], Quansheng Dou[1,2], and Li Shi[3]

[1] School of Computer Science and Technology, Shandong Technology and Business University, Yantai, China
[2] Shandong Co-Innovation Center of Future Intelligent Computing, Yantai, China
[3] Hospital of Shandong Technology and Business University, Yantai, China

Correspondence: Quansheng Dou, Shandong Technology and Business University, Binhai Road NO.191, LaiShan District, Yantai City, Shandong Province, Yantai 264005, China. e-mail: 37222697@qq.com

Received: April 8, 2019
Accepted: May 31, 2020
Published: July 10, 2020

Keywords: ResNet CNN; end-to-end; automated classification; image augmentation

Citation: Jiang P, Dou Q, Shi L. Ophthalmologist-level classification of fundus disease with deep neural networks. Trans Vis Sci Tech. 2020;9(2):39, https://doi.org/10.1167/tvst.9.2.39

Purpose: To implement the classification of fundus diseases using deep convolutional neural networks (CNN), which is trained end-to-end from fundus images directly, the only input are pixels and disease labels, and the output is a probability distribution of a fundus image belonging to 18 fundus diseases.

Methods: Automated classification of fundus diseases using images is a challenging task owing to the fine-grained variability in the appearance of fundus lesions. Deep CNNs show potential for general and highly variable tasks across many fine-grained object categories. Deep CNNs need large amounts of labeled samples, yet the available fundus images, especially labeled samples, are limited, which cannot satisfy the training requirement. So image augmentations such as rotation, scaling, and noising are implemented to enlarge the training dataset. We fine-tune the ResNet CNN architecture with 120,100 fundus images consisting of 18 different diseases and use it to classify the fundus images into corresponding diseases.

Results: The performance is tested against two board-certified ophthalmologists. The CNN achieves performance on par with the experts for the classification accuracy.

Conclusions: Deep CNN is capable of predicting fundus diseases given fundus images as input, which can enhance the efficiency of diagnosis process and promote better visual outcomes. Outfitted with deep neural networks, mobile devices can potentially extend the reach of ophthalmologists outside of the clinic and provide low-cost universal access to vital diagnostic care.

Translational Relevance: This article implemented automatic prediction of fundus diseases that was done by ophthalmologists previously.

Introduction

Many diseases manifest in the retina that affect a large proportion of the population,[1,2] and if left untreated these diseases may cause poor patient outcomes such as permanent vision loss. The cost-effectiveness of regular retinal screenings has been well established,[3] but because of the lack of sufficient eye care practitioners trained in retinal images explanation, the screening efficiency is low, and the screening result is also subjective, so it is difficult to implement widespread retinal screenings. To solve these problems, we developed a computational method that may allow the practitioners to track fundus lesions earlier. By creating a novel fundus disease taxonomy and a training class–generation algorithm that maps fundus images into balanced training classes, a deep learning system for automated fundus disease classification can be built.

Because of insufficient data and standardized photographic equipment that generate highly standardized images, previous work in fundus image classification based on computer method[1–7] lacked the generalization capability of eye care practitioners. Photographic images (e.g., smartphone images) exhibit variability in factors such as zoom, angle, and lighting, making classification substantially more challenging.[8,9] We overcome this challenge by using a data-driven approach; the fundus images are augmented by

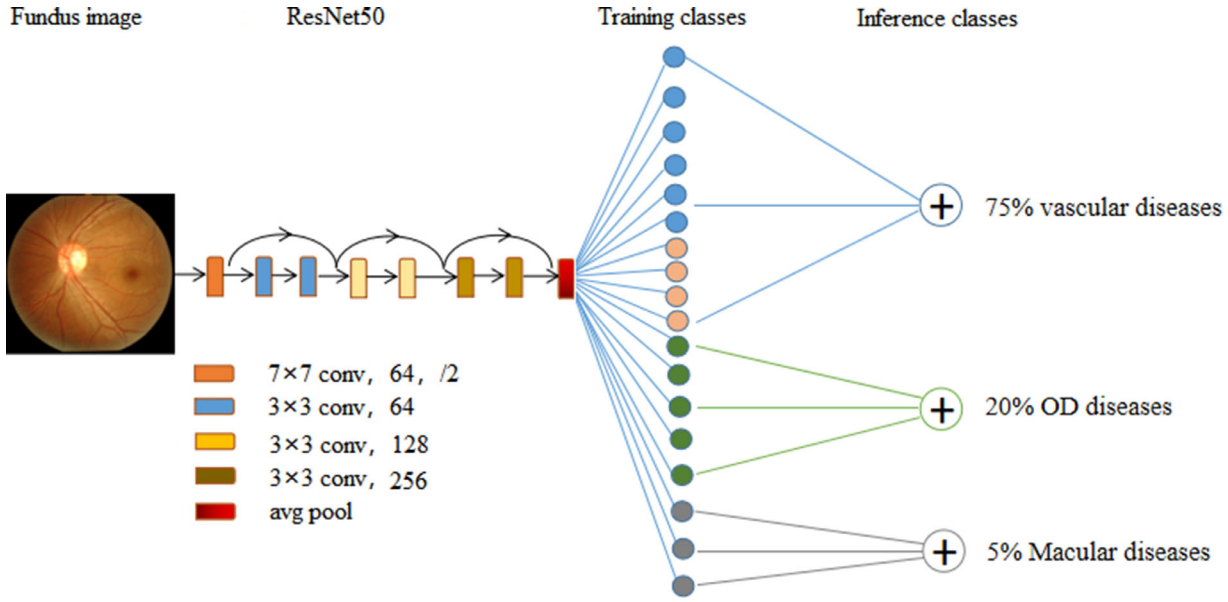

Figure 1. Deep ResNet CNN layout. Fundus image is sequentially warped into a probability distribution over clinical classes of fundus diseases using ResNet50 architecture fine-tuned on our own dataset of 120,100 images with 18 different diseases. Inference class is more general whose probability is calculated by summing the probabilities of the training classes according to taxonomy structure (see Methods).

rotating, zooming, noising, and shifting to generate large amount of pre-training and training images which makes classification robust to photographic variability. Many previous techniques require extensive preprocessing, lesion segmentation, and extraction of domain-specific visual features before classification. In contrast, the proposed system in this article does not require hand-crafted features, and it is trained end-to-end directly from image labels and raw pixels, which can accept both photographic and standardized images as input and give disease classification results as output. Previous literature about these kinds of work uses small datasets of typically less than a thousand images of fundus disease to train the network, which do not generalize well to new images. By data augmentation mentioned above, the labeled dataset for training has 120,100 clinical images, including 2174 standardized images, and our system demonstrate generalizable classification performance on these images.

Deep learning algorithms, powered by advances in computation and very large datasets,[10] have recently been shown to exceed human performance in visual tasks such as playing Atari games,[11] strategic board games like Go[12] and object recognition.[13] In this paper, based on transfer learning,[14] we fine-tune a ResNet convolutional neural networks (CNN) architecture which was pretrained on approximately 1.28 million images (1,000 object categories) from the 2015 ImageNet Large Scale Visual Recognition Challenge,[13] and train it with our fundus dataset to make the ResNet

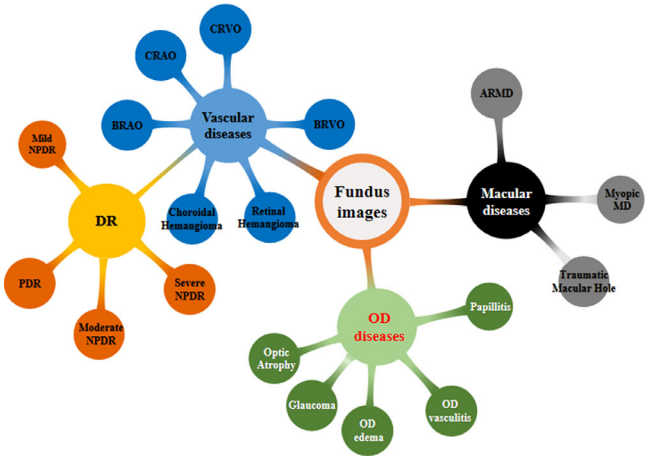

Figure 2. A schematic illustration of the taxonomy and examples of test set images. The tree-structured taxonomy of fundus disease contains 18 types of diseases. *Yellow* indicates DR (diabetic retinopathy) which belongs to vascular disease, *blue* indicates vascular diseases, *green* indicates OD diseases, and *black* indicates macular diseases.

suitable for the fundus disease classification. Figure 1 shows the working system. The CNN is trained using the 18 fundus disease classes from our datasets which are composed of ophthalmologist-labeled images organized in a tree-structured taxonomy, whose leaf node is the individual fundus disease. Figure 2 shows the full taxonomy. We split our dataset into 118,000 training and validation images and 2100 test images.

Table 1. Training Class–Generation Algorithm

Inputs
Taxonomy: the disease taxonomy
MaxClassSize(int): maximum images in a training class
Output
Training Class Set (list of Training Classes): organize the images into correct disease classes
Procedure CHILDREN_NODE(*node*)
Return {node} ∪ {CHILDREN_NODE(*child*) for *child* in *node.children*}
Procedure IMAGESIZE(*nodes*)
Return SUM(SIZE(*node.images*) for *node* in *nodes*)
Procedure ORGANIZE_IMAGES(*node*)
Class ← CHILDREN_NODE(*node*)
If IMAGESIZE(*class*)<*maxClassSize* then Add *class* to *Training Class Set*
Else
For *child* in *node.children* do
ORGANIZE_IMAGES (*child*)
Training Class Set ← []
ORGANIZE_IMAGES(*taxonomy.root*)
Return *Training Class Set*

To make use of fine-grained information contained within the taxonomy structure, an algorithm (Table 1) is created to partition the labeled fundus images into fine-grained training classes (e.g., mild nonproliferative diabetic retinopathy (NPDR), moderate NPDR, severe NPDR, and proliferative diabetic retinopathy (PDR)). The CNN is then fine-tuned with these training images, which outputs a probability distribution over these fine classes. To get the probability of the inference classes (e.g., vascular diseases, optic disc (OD) diseases, macular diseases, diabetic retinopathy (DR)), we sum the probabilities of their descendants (see Methods and Fig. 6 for more details).

The effectiveness of the method is validated in two ways by ninefold cross-validation as shown in Table 2. First, we validate the algorithm using a three-class disease classification—the first-level nodes of the taxonomy, which represent vascular diseases, OD diseases, macular diseases. For this inference class classification task, the CNN achieves 85.2% ± 0.7% (mean ± SD) overall accuracy (the average of individual inference class accuracies), and two comprehensive ophthalmologists get 78.56% and 76.3% accuracy on a subset of the validation set. Second, we validate the algorithm using a 18-class pathology classification—the second-level nodes of the tree. The CNN achieves 81.4% ± 1.7% overall accuracy, and the same two ophthalmologists get 75.2% and 73.2% accuracy, respectively. Figure 3 shows a few example images, demonstrating the difficulty in distinguishing between the fundus diseases because of the many similar visual features. The comparison metrics are sensitivity and specificity:

$$sensitivity = \frac{true\ positive}{positive}$$
$$specificity = \frac{true\ negative}{negative}$$

Where "true positive" is the number of correctly predicted fundus diseases, "positive" is the number of certain diseases the CNN predicted, including true-positive and false-positive results, "true-negative" is the number of correctly eliminated diseases, and "negative" is the number of diseases the CNN eliminated, including true-negative and false-negative results. When inputting an image into the CNN, we get a probability P of the 18 diseases it predicted as output. By setting a threshold probability t, we can finally determine one disease the image belongs to as $\hat{y}$for the image, whereas $\hat{y} = P \geq t$. The sensitivity and specificity of these probabilities can be computed, by changing t in the interval 0–1, a curve of sensitivities and specificities of the CNN about the classification effectiveness can be generated as shown in Figure 4. The area under the curve (AUC) measures the performance of the CNN, whose maximum value is 1. As shown by the AUC on Figure 4a, the deep learning CNN exhibits reliable fundus disease classification. Each red point on the plots represents the sensitivity and

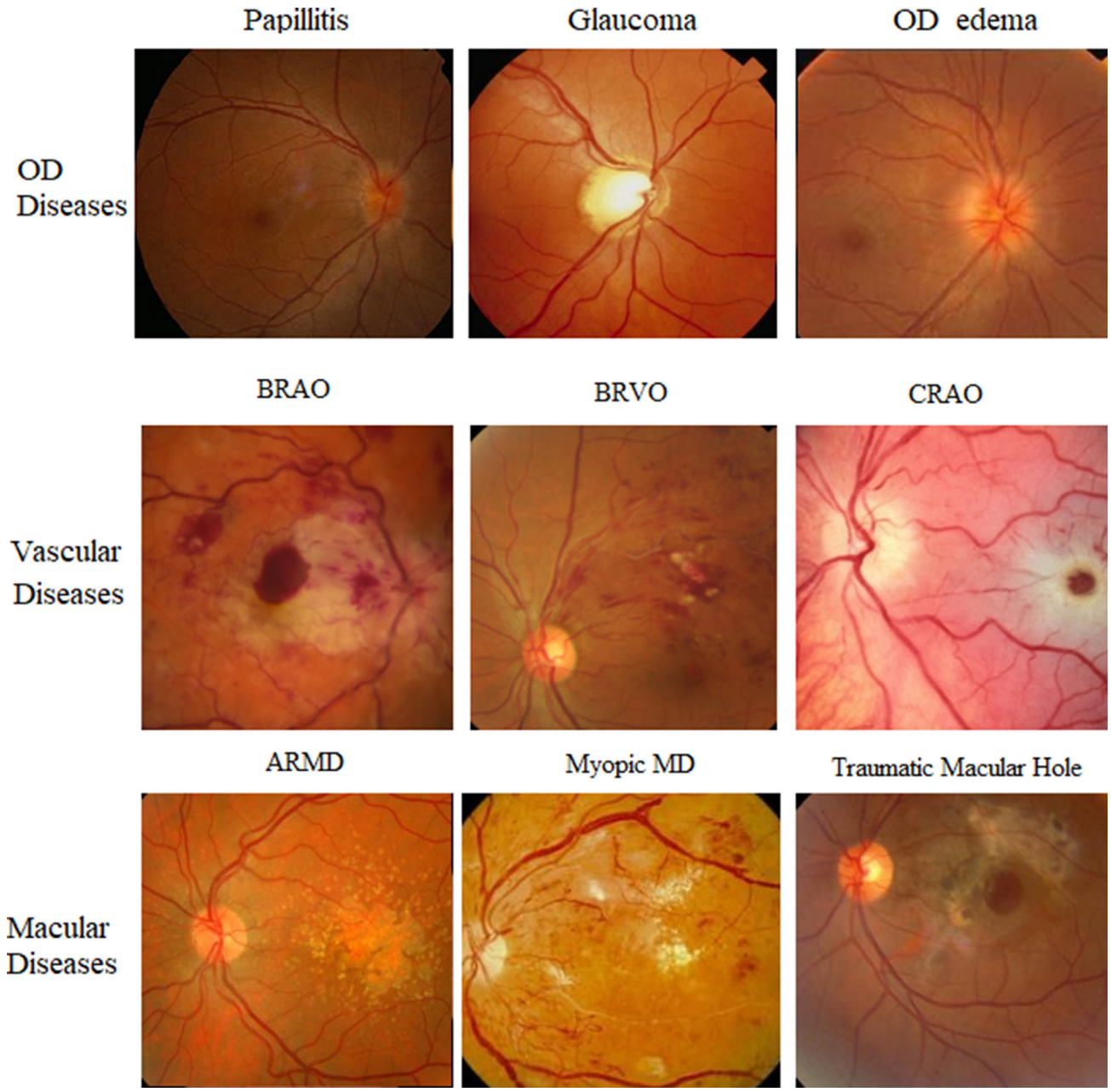

Figure 3. Example images. These test images highlight the difficulty for the disease classification tasks.

specificity of a single ophthalmologist. We can see that the CNN's performance is superior to the two ophthalmologists because the red points are under the blue curve of the CNN. When tested on a larger dataset (macular diseases: 800 images, vascular diseases: 720 images, OD diseases: 808 images; Fig. 4b), we found the tiny changes in the AUC compared with the small dataset, which show the robust and reliable classification performance on larger dataset.

Using t-distributed stochastic neighbor embedding (t-SNE),[15] the internal features learned by the CNN can be examined as shown in Figure 5. Each point represents a fundus image projected from the last hidden layer of the CNN into two-dimensional space. Most of the same pathology images are in the same cluster, whereas some points are mixed in different clusters, which indicates the wrong classification made by the CNN.

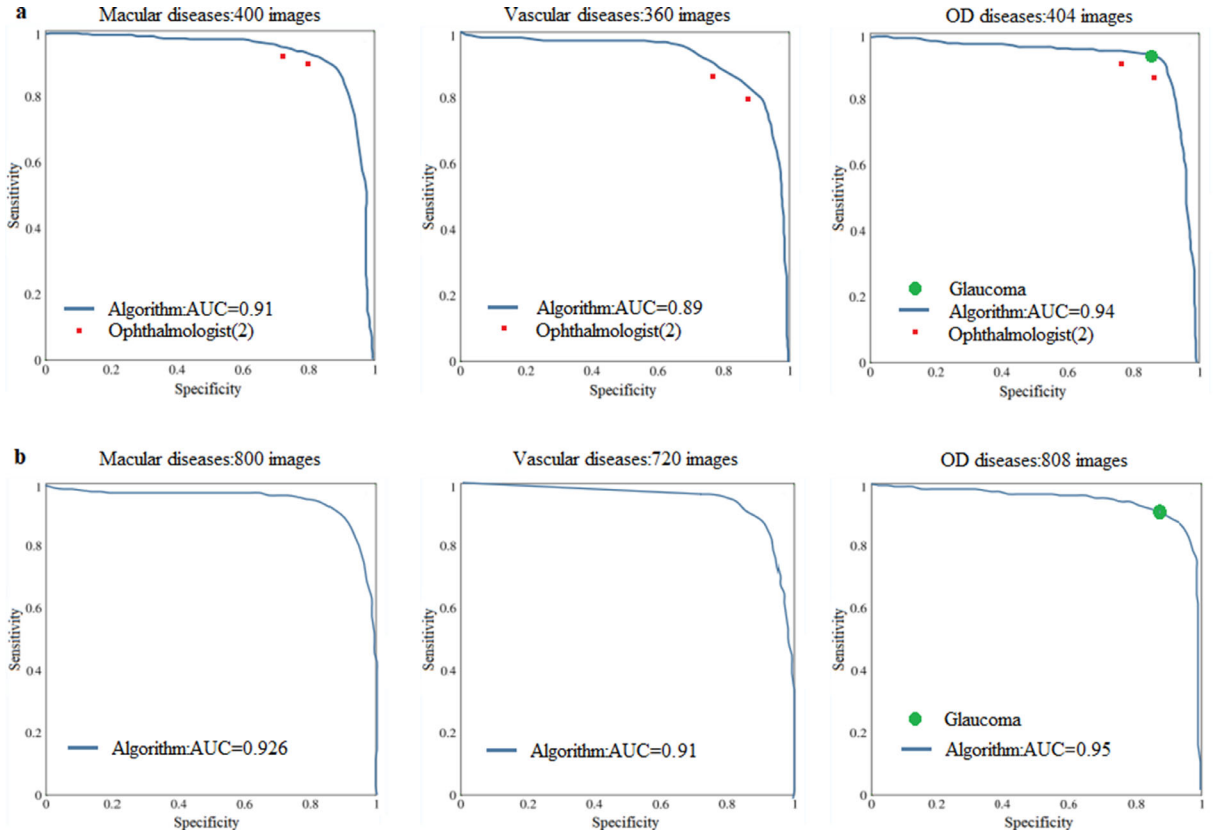

Figure 4. Fundus disease classification performance of the CNN with receiver operating characteristic (ROC) curve of sensitivity and specificity.

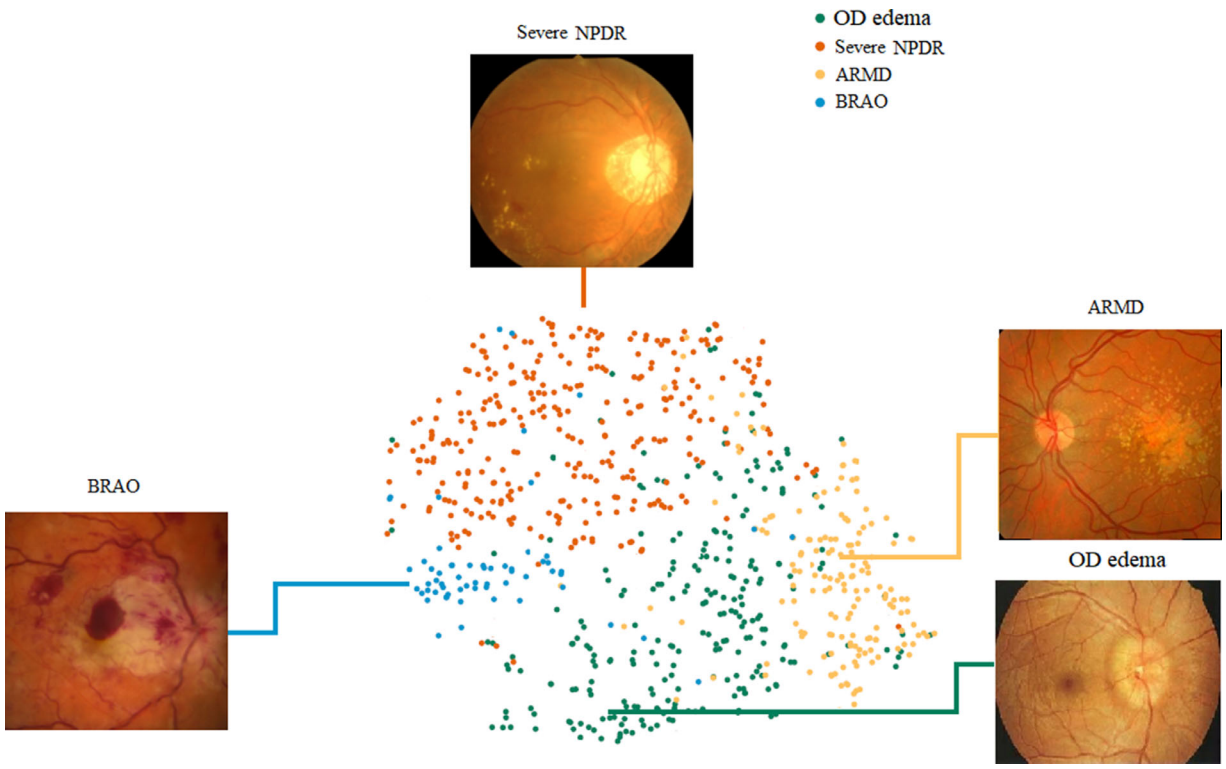

Figure 5. The t-SNE visualization of the last hidden layer representations in the CNN for the four disease classes. Different color clusters represent the different fundus diseases, showing how the algorithm clusters the diseases.

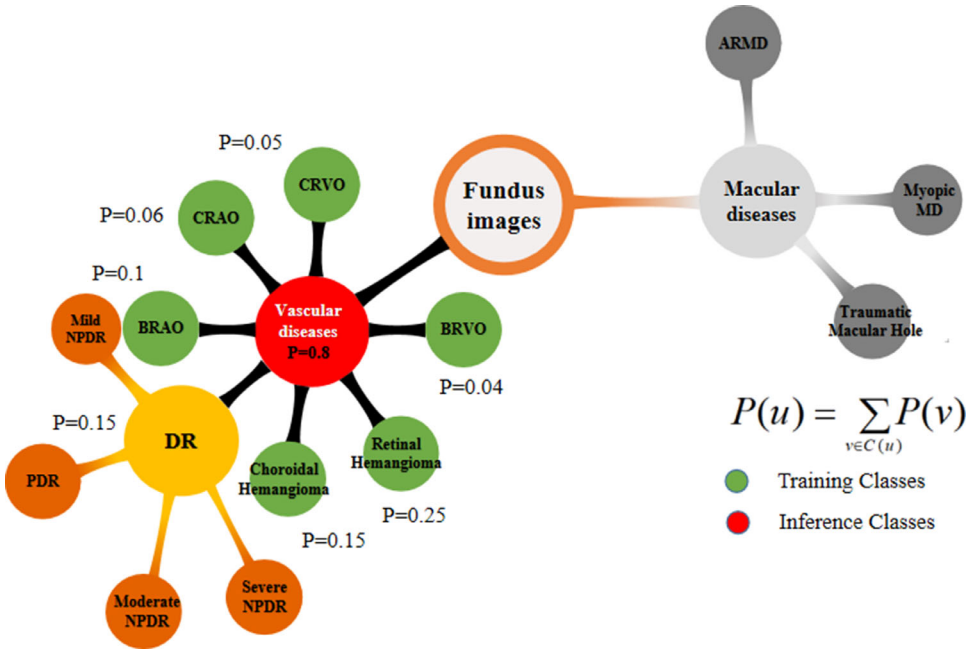

Figure 6. Illustration of calculating inference class probabilities from training class probabilities. Inference class (e.g., vascular diseases) is the red node while training classes (e.g., central retinal vein occlusion (CRVO), central retinal artery occlusion (CRAO), branch retinal artery occlusion (BRAO), branch retinal vein occlusion (BRVO)) are the green nodes in the tree. The probability of the parent equals to the sum of the child probabilities. As shown in the example, the probability of inference class of vessel disease: $P_{vessel\ disease} = 0.8 = 0.05 + 0.06 + 0.1 + 0.15 + 0.15 + 0.25 + 0.04$.

Methods

Taxonomy

Our taxonomy represents 18 individual diseases arranged in a tree structure with three root nodes representing general disease classes: (1) Macular disease, (2) vascular diseases and (3) OD diseases (Fig. 2a). It was derived using a bottom-up procedure by ophthalmologists : individual diseases were initialized as leaf nodes, based on clinical and visual similarity, the leaf nodes were merged until the entire tree was formed. The taxonomy is helpful in generating training classes that are suitable for machine learning classifiers. The first-level nodes are used in the first validation strategy and represent the most general partition. The child nodes of the root are used in the training of the ResNet to fine-tune the weights to make it suitable for the fundus disease classification.

Data Preparation

Blurry and far-away images were used in training the CNN but were removed from the test and validation sets. Yet the training images are still insufficient, in order to enlarge the training datasets, the images are augmented by rotating randomly between 0° and 359, and then scaling by a factor of ±0.2, some of which were randomly added noise. Due to the fine-grained variation of the fundus diseases, stretching is not used for image augmentation to avoid the false generation of fundus lesions. No overlap (that is, same lesion, multiple viewpoints) exists between the test sets and the training/validation data.

Training Class–Generation Algorithm

After image augmentation, the training datasets get enlarged but with unbalanced distribution, i.e., different fundus diseases have different sizes which may effect the training performance of the CNN. So we designed a algorithm to partition the individual diseases into training classes as outlined in Table 1. It is a recursive algorithm that takes the taxonomy as input and generates training classes whose individual diseases are clinically and visually similar. The algorithm also has another input parameter, maxClassSize, which forces the average generated training class size to be slightly less than its only maxClassSize. The algorithm keeps a balance between (1) generating training classes that are overly fine grained but do not have sufficient data to be learned properly; (2) generating training classes that are too coarse, too many data and may affect the

Table 2. General Validation Results

a. Disease classes: three-way classification	
Vascular diseases	
OD diseases	
Macular diseases	
b. Diseases classes: 18-way classification	
CRAO	
CRVO	
BRAO	
BRVO	
Choroidal hemangioma	
Retinal hemangioma	
Papillitis	
Glaucoma	
Optic atrophy	
OD edema	
OD vasculitis	
ARMD	
Myopic MD	
Traumatic macular hole	
Mild NPDR	
Moderate NPDR	
PDR	
Severe NPDR	
c. Classifier	three-way accuracy
Ophthalmologist1	78.56%
Ophthalmologist2	76.3%
CNN	85.2 ± 0.7%
CNN-TGA	87.3 ± 0.9%
d. Classifier	18-way accuracy
Ophthalmologist1	75.2%
Ophthalmologist2	73.2%
CNN	81.4 ± 1.7%
CNN-TGA	85.4 ± 0.8%

CRAO, •••; CRVO, •••; BRAO, •••; BRVO, •••; ARMD, •••; MD, •••; PDR, •••.

algorithm toward them. With maxClassSize = 3000 this algorithm yields a disease partition of 18 classes. All training classes are descendants of inference classes.

Training Algorithm

We use ResNet CNN architecture pretrained to achieve 3.57% error on the 1000 object classes (1.28 million images) of the ILSVRC 2015. Remove the final classification layer from the network and add a layer with 18 nodes, which represents our 18 diseases, then retrain it with our dataset and the parameters are fine-tuned across all layers. During training each image was resized to 224 × 224 pixels to make it compatible with the original dimensions of the ResNet network architecture while leveraging the natural-image features learned by the ImageNet pretrained network. This procedure, known as transfer learning, is optimal given the amount of data available. The CNN is trained using backpropagation. All layers of the network are fine-tuned using the global learning rate of 0.002 and a decay factor of 15 every 30 epochs. We use RMSProp with a decay of 0.85, momentum of 0.9 and epsilon of 0.1 to update the parameters of each layer.

Inference Algorithm

According to the tree convention, each node contains its children. Each node in the taxonomy represents a training class. Each inference class is a parent node of its descendent training nodes. As shown in Figure 1, the red node is inference class, and the green nodes are training classes. When inputting a fundus image through the ResNet CNN, it outputs a probability distribution over the training nodes. Probabilities over the taxonomy are computed by the following equation:

$$P(u) = \sum_{v \in C(u)} P(v),$$

where u is a parent node, $P(u)$ is the probability of u, and $C(u)$ are the child nodes of u. Therefore, to get the probability of any inference node, we just add all the probabilities of its descendant nodes together.

This algorithm takes the taxonomy tree as input and organizes all the fundus images into fine-grained balanced training classes. The experiments show that training on these finer classes helps to improve the classification accuracy of the corresponding inference classes. The algorithm begins with the root node of the taxonomy tree and recursively descend, if the amount of images in a node does not exceed a specified threshold, then the node are turned into a training class. During the organization procedure, the recursive property keeps the tree structure of the taxonomy, while make sure that the clinical similar fundus images are grouped into the same training class. The data restriction (maxClassSize) property helps to make that training data fairly evenly distributed among the leaf nodes. The algorithm generates training classes that leverage the fine-grained information contained in the taxonomy structure while keeping a balance between generating classes that are overly fine-grained and but don't have enough images to train the CNN properly, and classes that are too coarse with too many

images and so as to prevent the algorithm from generating small size training classes.

The above tables show the classification accuracy with 120,100 images organized in two different strategies: three-way classification and 18-way classification. The reported values are the mean and standard deviation of the accuracy. (a) Disease classes used for the three-way classification represent highly general disease classes. (b) Disease classes for 18-way classification represent the 18 fine-grained fundus diseases. (c) Three-way classification accuracy of our algorithm and two ophthalmologists. The three classes are the first-level nodes of our taxonomy tree. A CNN trained directly on these three classes achieves inferior performance to the one trained with our training class–generation algorithm (TGA). (d) An 18-way classification accuracy of our algorithm and two ophthalmologists. The 18 classes are the second-level nodes of our taxonomy. A CNN trained directly on these 18 classes also achieves inferior performance to one trained with our TGA.

Discussion

This article demonstrates the effectiveness of deep CNN in fundus disease classification. By using a single convolutional neural network trained on fundus images, we match the performance of two board-certified ophthalmologists tested across three critical diagnostic tasks: vascular diseases, OD diseases, and macular diseases classification. Our method is fast and scalable, which can be deployed on mobile devices and holds the potential for substantial clinical impact, including broadening the scope of primary care practice and augmenting clinical decision-making for ophthalmology specialists.

Yet, as we know, the retinal vascular diseases are more complex in nature and often present overlapping structural changes, sometimes even in the best scenario, the diseases can be misdiagnosed by CNN. So the systemic history information is needed to help our CNN to get a more accurate medical diagnosis. Currently the results from CNN can be used to help the ophthalmologists screening the retinal disease at the primary level, which is half of the work done.

In the future, we will combine the systemic history information, as well as genome information with the CNN, and improve the model to deal with the overlapping diseases, data bias, and more, and evaluate its performance in a real-world, clinical setting, to validate this technique across the full disease distribution. The major constraints of this method is lack of labeled data; if sufficient training images exist, the CNN can work under many visual conditions. Deep learning is independent of image type it is applied on, and so it could be adapted to other specialties, including dermatology, otolaryngology, radiology, and pathology.

Acknowledgments

Supported by National Natural Science Foundation of China, under grant nos. 61373079, 61272244, 61175023, 61175053, and 61272430, PhD foundation NO.BS201804.

Disclosure: **P. Jiang,** None; **Q. Dou,** None; **L. Shi,** None

References

1. Sisodia DS, Nair S, Khobragade P. Diabetic retinal fundus images: preprocessing and feature extraction for early detection of diabetic retinopathy. *Biomed Pharmacol J*. 2017;10:615–626.
2. Kiran SM, Chandrappa DN. Automatic detection of glaucoma using 2-D DWT. *Int Res J Eng Technol*. 2016;3:201–205.
3. Ganesh Babu TR, Sathishkumar R, Rengarajvenkatesh . Segmentation of optic nerve head for glaucoma detection using fundus images. *Biomed Pharmacol J*. 2014;7:697–705.
4. Annu N, Judith J. Automated classification of glaucoma images by wavelet energy features. *Int J Eng Technol*. 2013;5:1716–1721.
5. Kim, PY, Iftekharuddin, KM, Davey PG, et al. Novel fractal feature-based multiclass glaucoma detection and progression prediction. *IEEE J Biomed Health Inform*. 2013;17: 269–276.
6. Oh E, Yoo TK, Park E-C. Diabetic retinopathy risk prediction for fundus examination using sparse learning: a cross-sectional study. *BMC Med Inform Decis Mak*. 2013;13:106.
7. Mookiah MRK, Acharya UR, Chua CK, Lim CM, Ng EYK, Laude A. Computer-aided diagnosis of diabetic retinopathy: a review. *Comput Biol Med*. 2013;43:2136–2155.
8. Ramlakhan K, Shang Y. A mobile automated skin lesion classification system. In: *23rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, 2011; pp. 138–141..

9. Ballerini L, Fisher RB, Aldridge B, Rees J. A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions. In: Celebi ME, Schaefer G, eds. *Color Medical Image Analysis*. Dordrecht: Springer. 2013; pp. 63–86.
10. Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2009; pp. 248–255.
11. Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. *Nature*. 2015;518:529–533.
12. Silver D, Huang A, Maddison CJ, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*. 2016;529:484–489.
13. Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge. *Int J Comput Vis*. 2015;115:211–252.
14. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng*. 2010;22:1345–1359.
15. Van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;9:2579–2605.