tvst

Artificial Intelligence

Automatic Multilabel Classification of Multiple Fundus Diseases Based on Convolutional Neural Network With Squeeze-and-Excitation Attention

Zhenzhen Lu[1,*], Jingpeng Miao[2,*], Jingran Dong[1], Shuyuan Zhu[1], Penghan Wu[3], Xiaobing Wang[4,5], and Jihong Feng[1]

[1] Department of Biomedical Engineering, Beijing International Science and Technology Cooperation Base for Intelligent Physiological Measurement and Clinical Transformation, Beijing University of Technology, Beijing, China
[2] Beijing Tongren Eye Center, Beijing Ophthalmology & Visual Sciences Key Lab, Beijing Tongren Hospital, Capital Medical University, Beijing, China
[3] Fan Gongxiu Honors College, Beijing University of Technology, Beijing, China
[4] Sports and Medicine Integrative Innovation Center, Capital University of Physical Education and Sports, Beijing, China
[5] Department of Ophthalmology, Beijing Boai Hospital, China Rehabilitation Research Center, School of Rehabilitation Medicine, Capital Medical University, Beijing, China

Correspondence: Jihong Feng, Department of Biomedical Engineering, Beijing International Science and Technology Cooperation Base for Intelligent Physiological Measurement and Clinical Transformation, Beijing University of Technology, 100 Pingleyuan, Chaoyang District, Beijing 100124, China. e-mail: jhfeng@bjut.edu.cn
Xiaobing Wang, Sports and Medicine Integrative Innovation Center, Capital University of Physical Education and Sports, 11 North Third Ring West Road, Beijing 100191, China. e-mail: little-bill@263.net

Received: May 5, 2022
Accepted: November 6, 2022
Published: January 20, 2023

Keywords: fundus images; multiple fundus diseases; multilabel classification; convolutional neural network; attention mechanism; deep learning

Citation: Lu Z, Miao J, Dong J, Zhu S, Wu P, Wang X, Feng J. Automatic multilabel classification of multiple fundus diseases based on convolutional neural network with squeeze-and-excitation attention. Transl Vis Sci Technol. 2023;12(1):22, https://doi.org/10.1167/tvst.12.1.22

Purpose: Automatic multilabel classification of multiple fundus diseases is of importance for ophthalmologists. This study aims to design an effective multilabel classification model that can automatically classify multiple fundus diseases based on color fundus images.

Methods: We proposed a multilabel fundus disease classification model based on a convolutional neural network to classify normal and seven categories of common fundus diseases. Specifically, an attention mechanism was introduced into the network to further extract information features from color fundus images. The fundus images with eight categories of labels were applied to train, validate, and test our model. We employed the validation accuracy, area under the receiver operating characteristic curve (AUC), and F1-score as performance metrics to evaluate our model.

Results: Our proposed model achieved better performance with a validation accuracy of 94.27%, an AUC of 85.80%, and an F1-score of 86.08%, compared to two state-of-the-art models. Most important, the number of training parameters has dramatically dropped by three and eight times compared to the two state-of-the-art models.

Conclusions: This model can automatically classify multiple fundus diseases with not only excellent accuracy, AUC, and F1-score but also significantly fewer training parameters and lower computational cost, providing a reliable assistant in clinical screening.

Translational Relevance: The proposed model can be widely applied in large-scale multiple fundus disease screening, helping to create more efficient diagnostics in primary care settings.

tvst.arvojournals.org | ISSN: 2164-2591

Introduction

Millions of people in the world are affected by fundus diseases such as diabetic retinopathy (DR),[1] age-related macular degeneration (AMD),[2] glaucoma,[3] cataract,[4] and hypertensive retinopathy.[5] Early detection and timely diagnosis may not be available with manual diagnosis due to the complexity of fundus diseases and the increasing number of patients. These diseases may lead to irreversible blurred vision and even blindness without accurate diagnosis and timely treatment. Therefore, the accurate diagnosis and treatment of fundus diseases are very important.

Convolutional neural networks (CNNs) can automatically learn the high-level information on features of images, which has demonstrated promising performance in fundus disease classification.[6–9] Several studies focused on the screening of DR based on fundus image classification.[10–13] Automatic single-label classification of multiclass retinal diseases has been reported based on color fundus images and optical coherence tomography images.[14–18] These studies ignored the fact that a fundus image in the real world is likely to contain multiple fundus diseases.

Recently, some works began to explore multilabel fundus disease classification.[19–21] Peking University launched a fundus image database called Ophthalmic Disease Intelligent Recognition (ODIR) for multilabel classification of multiple retinal diseases.[22] For example, He et al.[23] proposed a dense correlation network based on the ResNet network to classify normal and seven types of fundus diseases by using the spatial correlation between paired color fundus images. The network was composed of a feature extraction module, a spatial correlation module, and a classifier. The experiments showed that the network had better performance than the corresponding benchmark. Gour et al.[24] proposed two multilabel classification models based on CNN and transfer learning for fundus images of eight types of patients in the ODIR database. Two different input modes and two different optimization algorithms with stochastic gradient descent (SGD) and Adam were used. In the training process, pretrained ResNet, Inception V3, MoblieNet, and VGG16 network were used as the feature extractors, respectively. The results showed that the VGG16 network with the SGD optimizer and feature stitching had better classification performance. Jordi et al.[25] proposed a model to transform the multilabel retinal disease classification into a multiclass retinal disease classification. Three pretrained deep CNN networks (VGG16, GoogLeNet, and InceptionV3) were used to classify retinal images from the ODIR database. This model outperformed other methods, but it cannot detect multiple diseases in a fundus image at the same time. Wang et al.[26] proposed a multilabel classification ensemble model based on CNN to detect multiple diseases in the fundus images. The pretrained EfficientNet network was used as the feature extraction network. The color images and gray images after histogram equalization were input into the network to obtain two models, respectively. The output probabilities of the two models were averaged as the final prediction result, which achieved an area under the curve (AUC) of 0.74 and an F1-score of 0.89, but the network parameters were set at a high level. Lin et al.[27] proposed two classification networks for multilabel classification of fundus images with the ODIR database and 2529 collected fundus. One was based on graph convolutional networks, which were used to replace the fully connected (FC) layer as a classifier to capture the relevant information of multilabel fundus images. The other was based on graph convolutional networks and self-supervised learning, in which the self-supervised learning was used to learn unlabeled fundus images. The results showed the two networks had better performance, but the training may be unstable.

The pretrained network based on the ImageNet data set without structural optimization is mostly used for multilabel classification of fundus diseases, which limits the classification accuracy. In addition, it is difficult to deploy the existing multilabel classification model for various types of fundus diseases to clinical diagnosis with more training parameters and high computational cost. In this study, we proposed a multilabel classification model based on CNN and the squeeze-and-excitation (SE) module that can automatically classify normal and seven types of common fundus diseases. Our proposed model was evaluated with the public ODIR database, achieving better performance with a smaller number of training parameters and computing load compared to the two state-of-the-art models.

Data Set

In this study, a public database was used, provided by Peking University through a global grand challenge named “International Competition on Ocular Disease Intelligent Recognition (ODIR)” (https://odir2019.grand-challenge.org). The data set is publicly available in Li et al.[22] This data set collects fundus images from the left and right eyes of 5000 patients and diagnostic keywords from doctors at different hospitals and

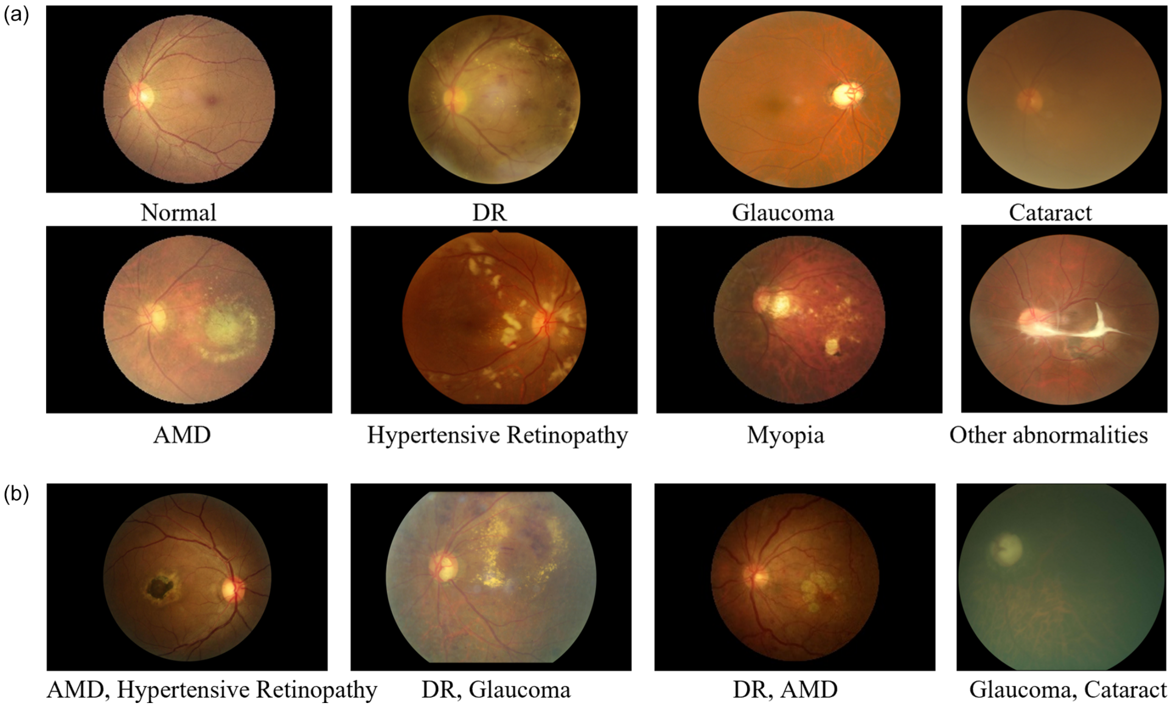

Figure 1. Representative fundus images in ODIR database. (a) Fundus images with single label. (b) Fundus images with multiple labels.

Table 1. Demographics Characteristics of the ODIR Data Set

Characteristics	Training	On-Site Testing	Off-Site Testing
No. of patients	3500	1000	500
No. of eyes	6999	2000	1000
Right	3500	1000	500
Left	3499	1000	500
No. of images	7000	2000	1000
Age, mean ± SD, y	57.85 ± 11.72	57.76 ± 12.45	58.15 ± 11.99
Gender, No.			
Male	1885	537	269
Female	1615	463	231

medical centers in China. Demographic data for each patient, including age and sex, are also provided in the data set. Eight categories of disease labels are provided for each patient, which refer to normal, DR, glaucoma, cataract, AMD, hypertension, myopia, and other abnormal diseases/abnormalities, as shown in Figure 1. A fundus image is marked by either a single label or multiple labels in eight categories, as shown in Figure 1a and Figure 1b, respectively. The ODIR database is divided into three parts: the training set, the

Table 2. Proportion of Images per Category in the ODIR Data Set[22]

Class	Training, No.	On-Site Testing, No.	Off-Site Testing, No.	Total
Normal	1138	324	162	1624
Diabetes	1130	327	163	1620
Glaucoma	215	58	32	305
Cataract	212	65	31	308
AMD	164	49	25	238
Hypertension	103	30	16	149
Myopia	174	46	23	243
Other diseases	982	275	136	1393

on-site testing set, and the off-site testing set, consisting of 3500, 1000, and 500 pairs of fundus images, respectively. Table 1 reveals the details of demographics characteristics of the ODIR data set. Table 2 shows the image distribution with respect to eight categories of labels in the ODIR database. The ODIR database has an unequal quantity of photos for each label, resulting in a class imbalance problem where normal, DR, and other abnormality categories have enough images, while glaucoma, cataract, AMD, hypertensive retinopathy, and myopia have significantly less fundus images.

Method

Image Processing and Labeling

Before the network training, we processed the fundus images in the ODIR database to confirm that the images were of sufficient quality for the experiment. First, we removed the poor-quality images for the left or right eye. According to the diagnostic keywords of the left eye and right eye, exclusion criteria for images followed the rules: (1) images with lens dust, (2) images with an optic disk photographically invisible, (3) images with low image quality, (4) images with image offset, or (5) images with only the label of other abnormal. By manually browsing the fundus images, we throw away the completely invisible images. A total of 2824 images were abandoned from the total of 10,000 fundus images in the ODIR database, leaving 7176 images. Second, the fundus images with a large area of black background were cropped to remove the black background areas by automatically reading the upper, bottom, left, and right boundaries of the images. As fundus images captured by various cameras on the market resulted in varied sizes, we extracted the retina region to 299 × 299 pixels based on the detected retina circle. Here, the Hough Circles transformation was used to detect the circle of the retina.[28] To reduce the influence of nonuniform illumination on images, we then performed contrast-limited adaptive histogram equalization on retinal fundus images. Two typical original and processed images are shown in Figure 2.

The ODIR database only provides the disease labels at the patient level, left eye diagnostic keywords, and right eye diagnostic keywords. According to the diagnostic keywords of the left eye and right eye, eight categories of labels were assigned to each fundus image to reduce the complexity of network construction, which enabled the model to detect the fundus disease of an individual eye. After image processing and labeling, the remaining 7176 images were used for training and testing. Among these 7176 images, 966 fundus images from the off-site testing set were used as the testing set, and 6210 images remained. In Table 1, there is an extremely severe class imbalance problem in the ODIR database. When there is one of the underrepresented minority classes, oversampling

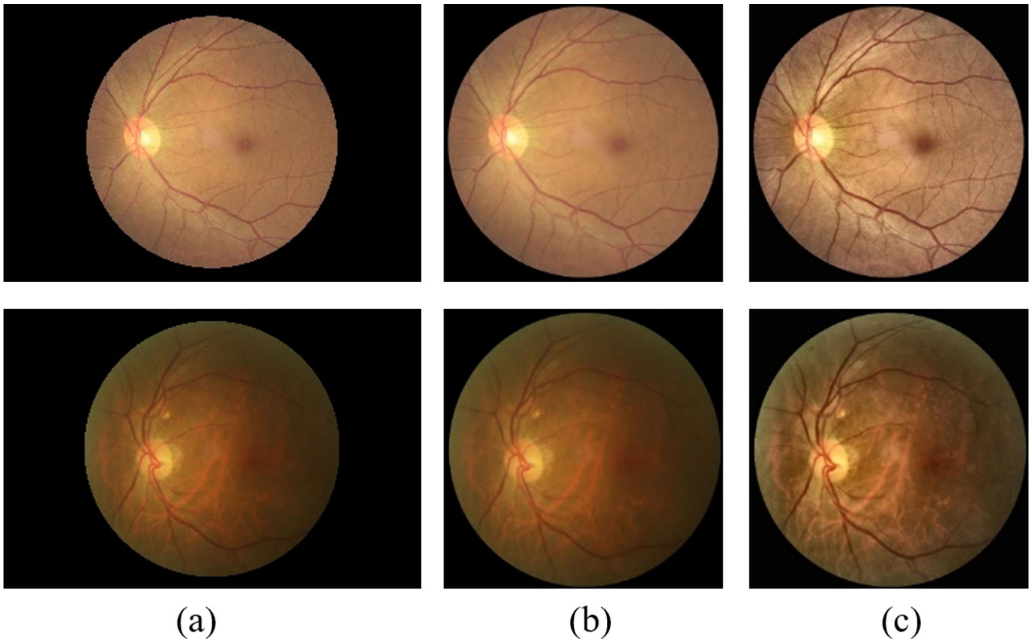

Figure 2. Representative original and processed images. (a) Original images. (b) Images with 299 × 299 pixels after cropping black background. (c) Processed images.

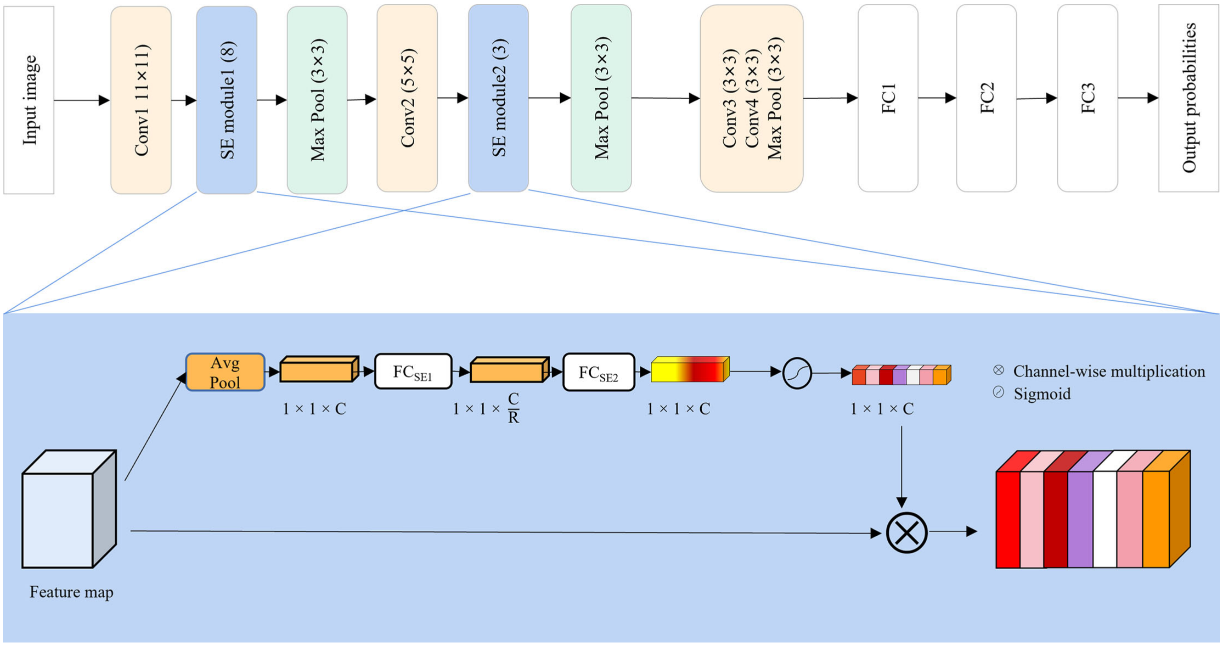

Figure 3. Our multilabel classification network.

techniques duplicate minority class samples to reduce the risk of overfitting.[29–31] In addition to the normal, DR, and other abnormal diseases of 4513 images, the other categories of 1697 images were only duplicated four times to obtain 6788 images. The normal, DR, and other abnormal diseases of 4513 images integrating the augmented other categories of 6788 images resulted in 11,301 fundus images for network training and validation with a ratio of 9:1, or 10,171:1130.

Designing a Multilabel Classification Network

Existing classification methods for multitype retinal diseases based on CNN were computationally intensive, restricting memory potency and training, which affected the optimization of hyperparameters.[32] Thus, we designed a shallow CNN to reduce computing load, memory requirements, and hyperparameter scale as the backbone of our multilabel classification network, which can learn multilabel lesion features for multitype retinal disease classification. The shallow CNN consisted of four convolution layers, three max-pooling layers, and three FC layers. Besides, to refine feature representations more effectively, we further introduced a channel attention mechanism with the SE module[33] to the shallow CNN. The architecture of our proposed multilabel classification network is presented in Figure 3. The proposed multilabel classification network was composed as follows: input layer, convolution layer, namely Conv1, SE module1, max-pooling layer, Conv2, SE module2, max-pooling layer, Conv3, Conv4, max-pooling layer, FC1, FC2, FC3 and output layer. The ReLU nonlinearity was applied to the output of every convolution and FC layer except for the last FC layer.

The SE module performed the recalibration of feature maps in the channel dimension to automatically obtain the importance values of different channels in the feature maps. With these importance values, the proposed network can selectively enhance the informative features useful for the current classification task and suppress less useful ones. The principle of the SE module is described as follows.[33] First, the original feature maps produced by convolutional operation are defined as $U \in R^{H \times W \times C}$ and the original feature maps can be written as $U = [u_1, u_2, \ldots u_C]$. The original feature maps U are first passed through a squeeze operation, which can incorporate the global spatial information by generating channel-wise statistics. Specifically, the $H \times W$ spatial dimensions of the original feature maps are shrunk to generate the global spatial feature $Z \in R^C$. The c channel element of Z is computed by

$$z_c = \frac{1}{H \times W} \sum_{i-1}^{H} \sum_{j=1}^{W} u_c\,(i, j), \qquad (1)$$

where the spatial dimension of the original feature maps is $H \times W$, and $u_c \in R^{H \times W}$. Second, to make use of the information aggregated in the squeeze operation,

Table 3. Structure and Parameters of Our Proposed Network

Layer (Type)	Output Shape	Filter Size	Stride	Padding
Input layer	(None, 299, 299, 3)			
Zero padding	(None, 302, 302, 3)			
Convolution	(None, 76, 76, 24)	11 × 11	4 × 4	Same
SE module	(None, 76, 76, 24)			
Max pooling	(None, 38, 38, 24)	3 × 3	2 × 2	Same
Convolution	(None, 38, 38, 64)	5 × 5	1 × 1	Same
SE module	(None, 38, 38, 64)			
Max pooling	(None, 19, 19, 64)	3 × 3	2 × 2	Same
Convolution	(None, 19, 19, 64)	3 × 3	1 × 1	Same
Convolution	(None, 19, 19, 64)	3 × 3	1 × 1	Same
Max pooling	(None, 9, 9, 64)	3 × 3	2 × 2	Valid
Flatten	(None, 5184)			
Dropout (0.5)	(None, 5184)			
FC (dense)	(None, 512)			
Dropout (0.6)	(None, 512)			
FC (dense)	(None, 512)			
FC (dense)	(None, 8)			

the aggregation is followed by an excitation operation that aims to fully capture channel-wise dependencies. For global spatial feature Z, the channel dimension of C is reduced to C/R by the first FC layer and then is activated by the ReLU function. The channel dimension of C/R is returned to the channel dimension of the original feature maps by the second FC layer. Subsequently, a series of per-channel modulation weights between 0 and 1 is produced by the sigmoid activation function. The global spatial feature Z is forwarded to two FC layers to finally generate the channel attention map $S \in R^C$, encoding which channel to emphasize or suppress. This process is called feature recalibration, namely, the gating mechanism. A simple gating mechanism is employed to achieve this objective:

$$S = \sigma \left(W_2 \delta \left(W_1 Z \right) \right), \tag{2}$$

where δ denotes the ReLU function, σ refers to the sigmoid function, $W_1 \in R^{\frac{C}{R} \times C}$ and $W_2 \in R^{C \times \frac{C}{R}}$ are the weights of the two FC layers, respectively. R is the reduction ratio used to reduce the channel dimension of the first FC layer in the SE module. Finally, the output of the SE module is obtained by rescaling the feature maps U with the channel attention map S:

$$X = U \otimes S, x_c = F_{scale} \left(u_c,\ s_c \right) = u_c s_c, \tag{3}$$

where $X = [x_1, x_2, \ldots x_C]$ and $\otimes$ denotes channel-wise multiplication. F_{scale} (u_c, s_c) refers to the channel-wise multiplication between the scalar s_c and the feature map $u_c \in R^{H \times W}$.

The reduction ratios of the first SE module and second SE module in the proposed multilabel classification network were 8 and 3, respectively.

Training and Optimizing

We trained the proposed network from scratch with the training set consisting of 10,171 images, as described in "Image Processing and Labeling." Each image referred to one fundus image of the left or right eye of a patient. Fundus images and the corresponding ground truths stored in the CSV file were input into the proposed network. The input images with 299 × 299 pixels passed through RGB channels. We adopted the Zero Padding operation to fill one, two, one, and two layers of zeros in the upper, lower, left, and right of the input images, respectively. The last FC layer of the network was used as a classifier, and a sigmoid was utilized as the last FC layer's activation function. Binary cross-entropy was adopted as the loss function. Dropout was applied to prevent overfitting. At the end, the proposed network outputs eight probability values, corresponding to the eight categories of labels. Compared with setting multiple independent classifiers in the multilabel classification model,[22,34] our setting reduced the risk of overfitting. Table 3 shows the structure and parameters of our proposed multilabel fundus disease classification network.

SGD and Adam optimizers were adopted to study the classification model of performance. Under the same conditions, we carried out comparative

Table 4. Configuration of Hyperparameter

Configuration	Value
Optimizer	Adam
Epoch	200
Batch size	32
Learning rate	5.00E-4
ReduceLROnPlateau	monitor='val_loss', factor=0.1, patience=10, min_lr=1.00E-07
EarlyStopping	monitor='val_acc', patience=20
ModelCheckpoint	monitor='val_acc', save_best_only=True, save_weights_only=True

experiments. It was found that the Adam optimizer was significantly better than the SGD optimizer in terms of shortening training time and convergence. Therefore, we selected Adam as the optimizer for our proposed network. We trained the network for 200 epochs, and the batch size was set to 32. To avoid overfitting and save training time, we implemented an early stop trick. If no progress was made on the accuracy of the validation data set in 20 successive epochs, the entire training process would be terminated early. After 80 epochs, the training was stopped due to the absence of further improvement in both validation loss and accuracy. The best model was selected based on validation accuracy for the validation phase. The learning rate was initially set to 0.0005, and the learning rate decay strategy was the learning rate multiplied by 0.1 when the validation accuracy plateaued within 10 epochs. The configuration of hyperparameters in our model or the multilabel fundus disease classification model is shown in Table 4.

The experiments ran on a workstation equipped with a NVIDIA GeForce GTX 1060 GPU, Intel Core i7-8700, and 8 GB memory, and the running operating system was Windows 10. The development platform was based on pycham2019.1.3 Community Edition, in which the designed implementation of the model was based on Tensorflow1.7.0 and Kares2.6.0 framework.

Evaluating the Model

We employed the testing set to evaluate the performance of our multilabel fundus disease classification model. When the fundus images without corresponding labels were input into the trained multilabel classification model, the model output eight probabilities between 0 and 1 to predict the eight categories of diseases, stored in a CSV file. Then, we applied three evaluation metrics, such as validation accuracy, F1-score, and AUC, to evaluate the performance of our proposed model. Accuracy is used for classification tasks, which corresponds to the proportion of correctly classified images with identical label sets of prediction and ground truth in all images. The accuracy of the training set and validation set is used to observe the risk of overfitting. The F1-score is the harmonic average of precision and recall, which ranges from 0 to 1. AUC represents the area under the receiver operating characteristics curve, which is a trade-off parameter between sensitivity and specificity parameters, usually measuring the stability of the model. The threshold is set at 0.5. All these metrics are calculated by the sklearn package. These evaluation metrics are given as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1\_score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

$$TPR{=}\frac{TP}{TP + FN}, FPR = \frac{FP}{FP + TN} \tag{8}$$

$$\mathrm{AUC} = \int_{x=0}^{1} TPR\left(FPR^{-1}(x)\right)dx \tag{9}$$

where TP, FP, TN, and FN refer to true-positive predictions, false-positive predictions, true-negative predictions, and false-negative predictions; TPR and FPR are true-positive rate and false-positive rate.

Results

Tensorboard (https://tensorflow.google.cn/tensorboard), a visual tool of TensorFlow, was used to observe the convergence of our multilabel fundus disease classification model. In our model, the accuracy and loss curves of the training set and validation set are shown in Figure 4. After 80 epochs, the training was stopped due to the absence of further improvement in both accuracy and loss. In the loss curves, our model converged rapidly, indicating that the parameters of this model were suitable for this multilabel classification task of fundus diseases. In the processed training set and validation set, our model obtained an accuracy of 98.64% and 94.27%, respectively. These results showed that our proposed model achieved

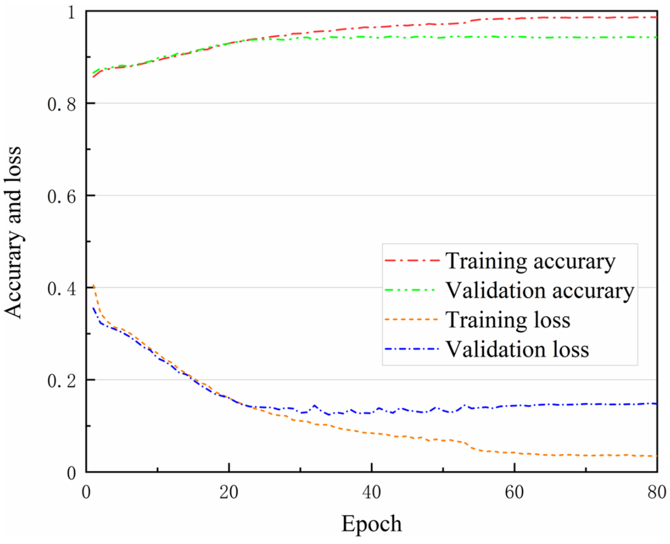

Figure 4. Accuracy and loss curves of training set and validation set.

high accuracy in multidisease classification without overfitting, even with a small number of images in the training set.

Class-wise performance analysis is important in the case of class imbalance for better insight into the overall performance of a model. There is a class imbalance since the training set had a class imbalance problem, so we evaluated the class-wise performance of our model on the images of the testing set. The class-wise performance of our model is summarized in Table 5 for accuracy, precision, sensitivity, and specificity parameters. The model showed high specificity and accuracy toward disease classes having good representation in the data set. The high precision of the cataract and myopia classes may be because these two types of image features are more obvious. The high accuracy and specificity for the minority class are because for that particular class, the number of negative samples is much higher than positive samples. However, the model failed to recognize the minority class images well.

On the same data set, the performance of our proposed model was compared with the results mentioned in methods[26,27] on F1-score, AUC, and total number of training parameters. As summarized in Table 6, our model was effectively enhanced in performance compared to model I proposed in Wang et al.[26] The validation accuracy of our model was improved by about 2.27% without overfitting. Our model achieved an AUC of 85.80%, or a 11.8% improvement, compared to model I. The F1-score was 86.08% in our proposed model, or 2.92% less than that of model I. At the same time, the total number of training parameters of our proposed model was three times less than that of model I. Compared to model II proposed in Lin et al.,[27] our model was more stable and performed better. In our model, the AUC and F1-score obtained a 7.64% improvement and a 3.58% deterioration compared to those of model II. Furthermore, the number of training parameters in model II is nearly eight times higher than that in our model. These results showed that our proposed model

Table 5. Class-Wise Performance of Our Model for Accuracy, Precision, Sensitivity, and Specificity

Class	Accuracy	Precision	Sensitivity	Specificity
Normal	0.63	0.57	0.45	0.76
Diabetes	0.69	0.43	0.43	0.80
Glaucoma	0.91	0.28	0.20	0.98
Cataract	0.94	0.63	0.89	0.98
AMD	0.96	0.57	0.18	0.99
Hypertension	0.93	0.13	0.03	0.99
Myopia	0.95	0.71	0.90	0.98
Other diseases	0.73	0.29	0.32	0.84

Table 6. Comparison of the Performance of Our Model with the State-of-the-Art Models

Model	Train_Accurary (%)	Val_Accurary (%)	AUC (%)	F1-Score (%)	Training Parameters
Model I[26]	\	92.00	74.00	89.00	>8.90M
Model II[27]	\	\	78.16	89.66	>25.50M
SENet50[33]	98.14	92.97	81.12	85.11	>28.09M
Our model	**98.64**	**94.27**	**85.80**	86.08	**3.05M**

Table 7. Results of Testing on the External Validation Set

Model	Precision	Recall	AUC	F1-Score
Model I[26]	0.46	0.45	0.82	0.86
Model II[27]	0.46	0.46	0.69	0.86
SENet50[33]	0.42	0.40	0.73	0.85
Our model	0.54	0.47	0.85	0.88

achieved better performance with a validation accuracy of 94.27%, or a 2.27% improvement, and an AUC of 85.80%, or a 7.64% improvement, compared to the two state-of-the-art models. Most important, the number of parameters has dramatically dropped by three and eight times in our model compared to model I and model II, respectively. To highlight the advantages of introducing the SE module into our classification network, we compared our model with SENet50.[33] In Table 6, the AUC and F1-score of SENet50 are 81.12% and 85.11%, respectively. Our model has the highest AUC and F1-score with a relatively fast classification speed, but its number of training parameters is only 3.05M, which is only about 10.86% of that of SENet50. The advantage of our model over SENet50 is the implementation of multilabel retinal disease classification with fewer training parameters and a relatively fast classification speed. The proposed model in our work is a lightweight architecture, which has an advantage over the models mentioned above in our experiments.

In order to verify the validity of our model, we collected 70 fundus images to build an external validation set from the collected data set in Lin et al.[27] The external validation set contained 10 images of normal, DR, glaucoma, cataract, AMD, hypertension, and myopia, respectively. The external validation set is used for testing the generalization performance of the models mentioned in methods,[26,27] SENet50, and our model. Table 7 lists the experimental results of the models on the external validation set in terms of the precision, recall, AUC, and F1-score. As illustrated in Table 7, our model achieves the highest precision of 54.10%, the highest recall of 47.14%, the highest AUC of 85.35%, and the highest F1-score of 88.39% with relatively fewer training parameters. Its precision is about 8%, 8%, and 11% higher than that of model I, model II, and SENet50, respectively. Its recall is about 2%, 1%, and 5% higher than that of model I, model II, and SENet50, respectively. Its AUC is about 2%, 16%, and 5% higher than that of model I, model II, and SENet50, respectively. Its F1-score is about 2%, 2%, and 3% higher than that of model I, model II, and SENet50, respectively. The result shows that our model significantly improves the accuracy and robustness of classification for multiple fundus diseases.

To demonstrate that the SE module can emphasize lesion features extracted from fundus images and enhance the representational power, we adopted the gradient class activation map technique[35] to show the activation maps of the proposed model with and without the SE module, respectively.

The gradient class activation map technique was performed to identify the areas contributing the most to our model's classification of the predicted diagnosis. The activation maps of three representative images are shown in Figure 5, where the important feature areas in the image responsible for the classification are highlighted in red and yellow. The three original images with the pathologic features labeled by the ophthalmologist are shown in Figure 5a, d, and g; the class activation maps without the SE module are shown in Figure 5b, e, and h; and the class activation maps with the SE module are shown in Figure 5f, and i. In Figure 5c, arteriostenosis, cotton wool spot, superficial retinal hemorrhage, and papilloedema are extracted correctly in the class activation map, which is most connected to the diagnosis of hypertensive retinopathy. In Figure 5f, the loss of the optic disc rim is identified correctly in the class activation map with the SE module, which represents the classic damage of glaucoma other than that without the SE module, while the diffuse optical media opacity probably originates from corneal edema secondary to intraocular hypertension, causing poor imaging quality. In Figure 5i, the peripapillary atrophy and macular degeneration are picked up correctly in the class activation map, which is most connected to the diagnosis of pathologic myopia. These results were compatible with the judgment of the ophthalmologist. It is found that our proposed model with the SE module can better capture and emphasize the pathologic features in the three representative images.

Discussion

Currently, the classification models based on CNN mostly focus on the detection of a single retinal disease. In this article, we proposed a multilabel fundus disease classification model to automatically classify normal and seven types of retinal diseases. With modeling interdependencies between channels, the SE module made our model emphasize lesion features extracted from fundus images and enhanced the representational power of the custom classification network. Our experiment results demonstrated that our model

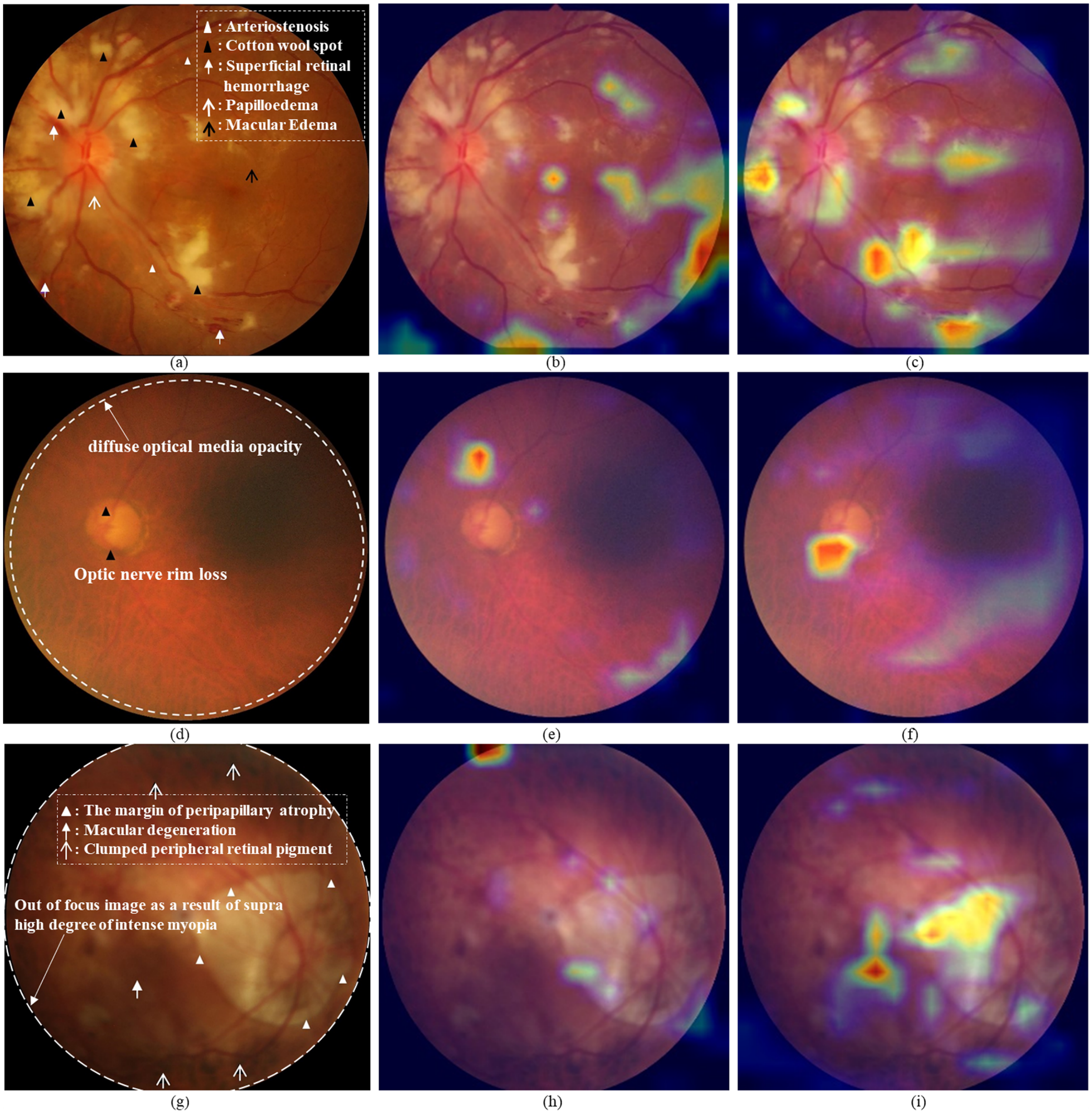
▲ : Arteriostenosis
▲ : Cotton wool spot
↑ : Superficial retinal hemorrhage
↑ : Papilloedema
↑ : Macular Edema
(a)
(b)
(c)
diffuse optical media opacity
Optic nerve rim loss
(d)
(e)
(f)
▲ : The margin of peripapillary atrophy
↑ : Macular degeneration
↑ : Clumped peripheral retinal pigment
Out of focus image as a result of supra high degree of intense myopia
(g)
(h)
(i)

Figure 5. Three representative images with the different pathologic features and the class activation maps without and with the SE module. (a) The original image with the pathologic features of hypertensive retinopathy labeled by the ophthalmologist. (b) Class activation map for hypertensive retinopathy without the SE module. (c) Class activation map for hypertensive retinopathy with the SE module. (d) The original image with the pathologic features of glaucoma labeled by the ophthalmologist. The optic nerve rim loss (*black arrowheads*) and the diffuse optical media opacity that probably originated from cataract or corneal edema secondary to intraocular hypertension (*white dotted circle*) are labeled. (e) Class activation map for glaucoma without the SE module. (f) Class activation map for glaucoma with the SE module. (g) The original image with the pathologic features of myopia labeled by the ophthalmologist. The out-of-focus image as a result of a supra high degree of intense myopia (*white dotted circle*) is labeled. (h) Class activation map for pathologic myopia without the SE module. (i) Class activation map for pathologic myopia with the SE module.

achieved better accuracy and AUC for multilabel classification compared to the two state-of-the-art models without overfitting, even with a small number of images and the problem of imbalanced classes in the public ODIR database. Our model contained about 3 million parameters, replacing about 8.9 million and about 25.5 million network parameters in the two state-of-the-art models. This indicated that our model was more compatible with small data set sizes, requiring fewer computational resources and time, thereby promoting the deployment of clinical applications. The F1-score of 86.08% decreased 2.92% and 3.58% compared to the two state-of-the-art models. Color fundus images and gray images were used as the inputs for model I during the training process, which indirectly increased the number of training images. The images from other databases were used as additional training sets in model II.

Although our model achieved better performance for the detection of normal and seven types of fundus diseases, its practical application in clinical trials is still a great challenge. In the ODIR database, there is more ambiguous information about other abnormal diseases, and the number of some diseases is very small, making it difficult to further improve the performance of our model. The ODIR database only provides patient-level ocular disease category labels. This may reduce the number of images with two or more retinal diseases, limiting the general performance of the model. From Table 5 and Table 7, the recall and precision of most diseases are low, and this issue mainly reflects some challenges and deficiencies. First of all, due to the broad source of the images, there is a wealth of intraclass diversity. Although they are labeled as the same category, there are large differences in color, lighting, and shooting conditions. Second, the classification of images with multiple diseases is affected and confused by various fundus abnormalities. Cataracts prevent the model from identifying hard exudate. Hard exudation and drusen are difficult to distinguish from each other, which makes most misclassified images difficult to extract valid regional features from the model and is also the reason why the sensitivity of cataract is relatively high in Table 5. In addition, the ODIR database has serious class imbalances, as shown in Table 2, where the low sensitivity in the case of high specificity is because, for that particular class, the number of negative images is much higher than positive images. Just as in Jordi et al.,[25] the sensitivity is higher than that of our model for normal and other disease images. The sensitivity of our model is higher for cataract and myopia with obvious pathologic features, and the class-wise performance of our model in terms of specificity is higher than that of the method mentioned in Jordi et al.[25] Third, local features are not obvious. The determination of glaucoma requires an accurate ratio of the optic cup and disc, and AMD requires more detailed features of the macular area.

In the future, we will improve the performance of the model on disease identification in the clinic scenario by balancing the number of images of different diseases types and adding more images with varying disease severity to modify the existing data set or reconstruct our data set. On the other hand, we intend to introduce depth-wise separable convolution to the network to greatly reduce the cost of computing and maintain accuracy simultaneously.

Conclusion

We proposed a multilabel fundus disease classification model with high accuracy, achieving better performance in validation accuracy, AUC, and F1-score compared to two state-of-the-art models. Most important, the number of parameters has dramatically dropped by three and eight times compared to the two state-of-the-art models. The proposed model is effective and reliable in the classification of normal fundus and seven major fundus diseases, which will push deep learning in clinical application.

Acknowledgments

Disclosure: **Z. Lu**, None; **J. Miao**, None; **J. Dong**, None; **S. Zhu**, None; **P. Wu**, None; **X. Wang**, None; **J. Feng**, None

[*] ZL and JM contributed equally to this work.

References

1. Gulshan V, Peng LH, 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):2402–2410.
2. Wong W, Su X, Li X, et al. Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. *Lancet Glob Health*. 2014;2(2):e106–e1162.
3. Tham Y, Li X, Wong TY, et al. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology.* 2014;121(2):2081–2090.
4. Liu Y-C, Wilkins M, Kim T, Malyugin B, Mehta JS. Cataracts. *Lancet.* 2017;390(10096):600–612.
5. Harjasouliha A, Raiji VR, García González JM. Review of hypertensive retinopathy. *Dis Mon.* 2017;63(3):63–69.

6. Fatima A, Badar M, Haris M. Application of deep learning for retinal image analysis: a review. *Comput Sci Rev*. 2020;35:100203.
7. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med.* 2018;24(9):1342–1350.
8. Kar SS, Maity SP. Automatic detection of retinal lesions for screening of diabetic retinopathy. *IEEE Trans Biomed Eng*. 2018;65(3):608–618.
9. Ting DS, Pasquale LR, Peng LH, et al. Artificial intelligence and deep learning in ophthalmology. *IEEE Trans Biomed Eng*. 2019;103(2): 167–175.
10. Abbas Q, Fondon I, Sarmiento A, Jiménez S, Alemany P. Automatic recognition of severity level for diagnosis of diabetic retinopathy using deep visual features. *Med Biol Eng Comput*. 2017;55(11):1959–1974.
11. Burlina P, Paul W, Mathew PA, Joshi NJ, Pacheco KD, Bressler NM. Low-shot deep learning of diabetic retinopathy with potential applications to address artificial intelligence bias in retinal diagnostics and rare ophthalmic diseases. *JAMA Ophthalmol.* 2020;138(10):1070–1077.
12. Hemanth DJ, Deperlioglu O, Kose U. An enhanced diabetic retinopathy detection and classification approach using deep convolutional neural network. *Neural Comput Appl.* 2020;32(3):707–721.
13. Riaz H, Park J, Choi H, Kim H, Kim J. Deep and densely connected networks for classification of diabetic retinopathy. *Diagnostics (Basel)*. 2020;10(1):24.
14. Chea N, Nam Y. Classification of fundus images based on deep learning for detecting eye diseases. *Comput Mat Contin*. 2021;67(1):411–426.
15. Kermany DS, Goldbaum MH, Cai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell.* 2018;172(5):1122–1131.
16. Lee J, Lee J, Cho S, et al. Development of decision support software for deep learning-based automated retinal disease screening using relatively limited fundus photograph data. *Electronics.* 2021;10(2):163.
17. Ramya J, Rajakumar MP, Maheswari BU. HPWO-LS-based deep learning approach with S-ROA-optimized optic cup segmentation for fundus image classification. *Neural Comput Appl.* 2021;33(15):9677–9690.
18. Shankar KA, Sait R, Gupta D, Lakshmanaprabu SK, Khanna A, Pandey HM. Automated detection and classification of fundus diabetic retinopathy images using synergic deep learning model. *Pattern Recognit Lett*. 2020;133:210–216.
19. Pan X, Jin K, Cao J, et al. Multi-label classification of retinal lesions in diabetic retinopathy for automatic analysis of fundus fluorescein angiography based on deep learning. *Graefes Arch Clin Exp Ophthalmol*. 2020;258(4):779–785.
20. Abdelmaksoud E, El-Sappagh S, Barakat S, Abuhmed T, Elmogy M. Automatic diabetic retinopathy grading system based on detecting multiple retinal lesions. *IEEE Access.* 2021;9: 15939–15960.
21. AbdelMaksoud E, Barakat S, Elmogy M. A comprehensive diagnosis system for early signs and different diabetic retinopathy grades using fundus retinal images based on pathological changes detection. *Comput Biol Med.* 2020;126: 104039.
22. Li N, Li T, Hu C, Wang K, Kang H. A benchmark of ocular disease intelligent recognition: one shot for multi-disease detection. In: Wolf F, Gao W, eds. *Proceedings of the International Symposium on Benchmarking, Measuring and Optimization*. Cham, Switzerland: Springer; 2020:177–193.
23. He J, Li C, Ye J, Qiao Y, Gu L. Multi-label ocular disease classification with a dense correlation deep neural network. *Biomed Signal Process Control*. 2021;63:102167.
24. Gour N, Khanna P. Multi-class multi-label ophthalmological disease detection using transfer learning based convolutional neural network. *Biomed Signal Process Control*. 2021;66:102329.
25. Jordi CC, Joan Manuel NDR, Carles VR. *Ocular Disease Intelligent Recognition through Deep Learning Architectures*. Barcelona, Spain: Universitat Oberta de Catalunya; 2019.
26. Wang J, Yang L, Huo Z, He W, Luo J. Multi-label classification of fundus images with EfficientNet. *IEEE Access.* 2020;8: 212499–212508.
27. Lin J, Cai Q, Lin M. Multi-label classification of fundus images with graph convolutional network and self-supervised learning. *IEEE Signal Process Lett*. 2021;28:454–458.
28. Cen LP, Ji J, Lin JW, et al. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nat Commun*. 2021;12(1):4828.
29. Abdi L, Hashemi S. To combat multi-class imbalanced problems by means of over-sampling and boosting techniques. *IEEE Trans Knowl Data Eng*. 2015;28(1):238–251.
30. Barandela R, Valdovinos RM, Sánchez JS, et al. The imbalance training sample problem: under or over sampling? In: Fred A, Caelli TM, Duin RPW,

Campilho AC, de Ridder D, eds. *Proceedings of the Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*. Berlin, Germany: Springer; 2004:806–814.

31. Engstrom L, Tran B, Tsipras D, et al. A rotation and a translation suffice: fooling CNNs with simple transformations. arXiv.org e-prints December 01, 2017, https://ui.adsabs.harvard.edu/abs/2017arXiv171202779E.
32. Sunija AP, Gopi VP, Palanisamy P. Redundancy reduced depthwise separable convolution for glaucoma classification using OCT images. *Biomed Signal Process Control*. 2022;71:103192.
33. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. *IEEE Trans Pattern Anal Mach Intell*. 2020;42(8):2011–2023.
34. Son J, Shin JY, Kim HD, Jung KH, Park KH, Park SJ. Development and validation of deep learning models for screening multiple abnormal findings in retinal fundus images. *Ophthalmology.* 2020;127(1):85–94.
35. Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int J Comput Vis*. 2020;128(2):336–359.