IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 28, 2020, accepted November 16, 2020, date of publication November 24, 2020,
date of current version December 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3040275

Multi-Label Classification of Fundus Images With EfficientNet

JING WANG[1], LIU YANG[1], ZHANQIANG HUO[1], WEIFENG HE[2], AND JUNWEI LUO[1]
[1]College of Computer Science and Technology, Henan Polytechnic University, Jiaozuo 454003, China
[2]Henan Polytechnic University Hospital, Jiaozuo 454003, China

Corresponding author: Jing Wang (wjasmine@hpu.edu.cn)

This work was supported in part by the National Science Foundation of China under Grant 61872311, Grant 61972134, and Grant 61602156; in part by the Key Science and Technology Program of Henan Province under Grant 182102210053; and in part by the Excellent Young Teachers Program of Henan Polytechnic University under Grant 2019XQG-02.

ABSTRACT Convolutional neural network (CNN) has achieved remarkable success in the field of fundus images due to its powerful feature learning ability. Computer-aided diagnosis can obtain information with reference value for doctors in clinical diagnosis or screening through proper processing and analysis of fundus images. However, most of the previous studies have focused on the detection of a certain fundus disease, and the simultaneous diagnosis of multiple fundus diseases still faces great challenges. We propose a multi-label classification ensemble model of fundus images based on CNN to directly detect one or more fundus diseases in the retinal fundus images. Every single model consists of two parts. The first part is a feature extraction network based on EfficientNet, and the second part is a custom classification neural network for multi-label classification problems. Finally, the output probabilities of different models are fused as the final recognition result. And it was trained and tested on the data set provided by ODIR 2019 (Peking University International Competition on Ocular Disease Intelligent Recognition). The experimental results show that our model can be trained on fewer data sets and get good results.

INDEX TERMS CNN, deep learning, ensemble learning, fundus images, multi label classification, transfer learning.

I. INTRODUCTION

Fundus diseases can cause vision loss which are the primary cause of blindness [1]. At present, common fundus diseases that affect visual function include diabetic retinopathy (DR), age-related macular degeneration (AMD), cataract and so on. The development of fundus disease to the late stage often has a serious impact on the patient's visual function, and there is no specific treatment for such diseases. Diabetes patients are one of the largest disease groups in the world today. DR is the most common complication of diabetes. There are no obvious abnormal symptoms in the early stage of the disease, but it will eventually lead to blindness. It is one of the four major blindness diseases [2]. If it is found in the early stage, it is still treatable. If it is found in the late stage, even if the operation is successful, the visual prognosis is very poor, and the treatment cost is high [3]. With the aging of the population, AMD has become the first blinding eye disease in Western countries. Its incidence increases with age. Once there is a lack of timely and effective treatment, it will in the short term, the patient's vision will decline rapidly, which will cause irreversible visual impairment [4], [5]. Therefore, early detection and early treatment of fundus diseases are very important. Artificial intelligence technology can assist primary ophthalmologists in diagnosis based on comprehensive medical data, and provide new strategies for improving the level of eye disease diagnosis and treatment in primary hospitals. The combination of artificial intelligence and ophthalmology medical treatment is to meet the practical needs of a large number of patients with fundus diseases.

The feature information extracted by traditional image feature extraction methods needs to rely on the prior knowledge of researchers, which has great limitations. In recent years, deep learning models have developed rapidly in the field of computer vision, and their effects have also been greatly improved compared to traditional methods. An important model in this field is the convolutional neural network model, because its powerful representation ability makes up for the shortcomings of traditional feature extraction methods. Convolutional neural network (CNN) can automatically learn high-level feature information of images, and have achieved

The associate editor coordinating the review of this manuscript and approving it for publication was Alex Noel Joseph Raj.

good results in image classification, target detection and other fields.

However, there are still challenges in using CNN in fundus image research. First of all, multi-label fundus image classification is a more common and practical problem, because a real fundus image in the real world is likely to contain multiple fundus diseases. Second, it is difficult to obtain sufficient true fundus images, especially data on some rare fundus diseases. Third, under the condition of limited fundus image data and inevitable image noise, it is difficult to train a single model to effectively obtain high disease detection accuracy. Therefore, for the first problem, we use a method based on problem transformation to convert the multi-label classification problem of each image into a two-class classification problem for each label. For the second and third problems, we used two important strategies in deep learning: transfer learning [7], [8], [34] and ensemble learning [9], [10]. The main idea of transfer learning is to transfer big data to small data fields based on knowledge reuse, and solve the problem of scarcity of data and knowledge in small data fields. The main idea of ensemble learning is to integrate multiple weak classifiers to obtain a better and more comprehensive strong classifier. Ensemble models make the increased generalization performance better than each individual component. In the current work, we have developed an end-to-end deep learning multi-label classification framework to assist in the diagnosis of various fundus diseases.

The rest of this paper is organized as follows. The second part first introduces the related work of analysis and diagnosis of retinal eye diseases. The third part introduces the dataset used. The fourth part introduces the overall algorithm of this paper in detail. The experimental results are analyzed and explained in the fifth part. Finally, the sixth part gives discussion and conclusion.

II. RELATED WORKS

A. TRADITIONAL METHODS

Before the advent of deep learning, computer-aided diagnosis (CAD) has been applied to the intelligent diagnosis of fundus diseases. From a macro point of view, the composition of the CAD system is very similar to the process of early clinical diagnosis and screening of fundus diseases by ophthalmologists by observing fundus images. If the whole process of early screening methods is compared to an assembly line, then at the front end of the pipeline, CAD can automatically perceive the distortion of the fundus image [13] and perform corresponding image enhancement [14], [36] and restoration [15], [35]; In the intermediate stage, CAD can segment the lesion and extract many features; At the back end, CAD can transform complex diagnostic logic into classification or clustering problems in machine logic, and classification or clustering problems can be solved by machine learning. Like CAD, image preprocessing and then image segmentation are performed [16], [17], then feature extraction, and finally the process of intelligent diagnosis through machine learning is actually imitating the doctor's diagnostic thinking. That is, initial observation of abnormalities in fundus images location (preprocessing), carefully observe the lesion location (segmentation), understand its shape and texture characteristics (feature extraction), and finally compare with the diagnostic criteria, and make judgments based on previous diagnostic experience (pattern recognition).

Image processing and visual perception technology based on the human visual system (HSV) represents a current trend, because the main body of image processing and visual perception is the human eye, which has many unique properties, such as multi-channel characteristics, point spread function (PSF), color sense consistency and so on. These properties make the computer have the characteristics of HVS and assist the human eye to complete the visual perception of images, and provide new ideas for the early detection and diagnosis of fundus diseases. Zhao *et al.* proceeded from the perception principle of color consistency, first extracted part of the low-contrast vascular structure, and then used the local phase enhancement technology to enhance the blood vessel, which is of great help to the diagnosis of retinal vascular disease [20].

B. DEEP LEARNING METHODS

In December 2016, the Google research team published a paper in the well-known journal JAMA. By allowing artificial intelligence algorithms to learn from a large number of patient fundus photos, they trained a deep neural network to detect DR, which can detect DR lesions in fundus images. And its accuracy is not inferior to professional ophthalmologists [21]. Subsequently, Theodore Leng and others of Stanford University published a paper on the DR intelligent detection system based on deep learning in the authoritative ophthalmology journal Ophthalmology [22]. For a time, the computer-assisted diagnosis of fundus diseases based on deep learning has attracted widespread attention. The CC-Cmsien, a deep learning based congenital cataract diagnosis and treatment platform developed and launched by Sun Yat-sen University in China, has a correct diagnosis rate of 99% for cataract patients, and has proposed correct treatment plans for more than 97% of patients. Its accuracy is comparable to that of an ophthalmologist. The research results were published in Nature Biomedical Engineering [23].

Arun Govindaiah *et al.* studied the effectiveness of using deep convolutional neural networks to screen age-related macular degeneration (AMD) individuals[24]. The experiment used an improved VGG16 neural network with batch normalization and training dataset is an age-related eye

disease study (AREDS) dataset containing more than 150,000 images. The experimental results of the research team proved that when there is sufficient training data, fully training a deep neural network is better than using pre-trained network classification, especially in AMD detection and screening.

Xiaogang Li *et al.* used transfer learning to classify the fundus images of diabetic retinopathy[25]. The experimental results on the two public data sets of DR1 and MESSIDOR,

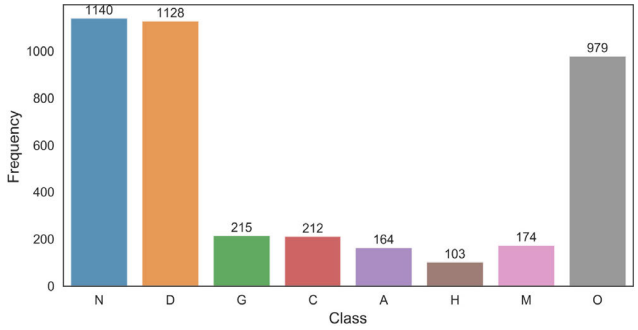


FIGURE 1. The proportion of images in each category in the training set data.

show that the knowledge learned in other large data sets (source domain) can be used to obtain better classification in small data sets (target domain) through transfer learning. In 2019, Wenai Song *et al.* proposed an improved semi-supervised learning method for the damage effects of cataract diseases [26], which obtained additional information from unlabeled cataract fundus images and improved the accuracy of the basic model that only trained labeled images.

S. J. Pan *et al.* first proposed to find a good feature representation across domains through a new learning method, transfer component analysis (TCA), for domain adaptation [37]. R. Polikar reviewed conditions under which ensemble based systems may be more beneficial than their single classifier counterparts, algorithms for generating individual components of the ensemble systems, and various procedures through which the individual classifiers can be combined [38]. In 2019, Sha Yuan *et al.* used innovative diagrams to clarify several important concepts of ensemble learning, and found that ensemble models with several specific single models can further boost the performance [39].

III. DATASET

The dataset comes from the ''International Competition on Ocular Disease Intelligent Recognition'' sponsored byPeking University. The dataset is ''real'' patient data collected by Shanggong Medical Technology Co. Ltd. from different hospitals/medical centers in China. The training set is a structured ophthalmic database of 3500 patients with age, color fundus photographs from left and right eyes and doctors' diagnostic keywords from doctors. The testing set is the color fundus photos of 500 patients, excluding age and gender. Fundus images are captured by various cameras on the market, such as Canon, Zeiss and Kowa, thus resulting in various image resolutions. These data classify patient into eight labels, as shown in Fig.2 (a), including normal(N), diabetic retinopathy(D), glaucoma(G), cataract(C), AMD(A), hypertensive retinopathy (H), myopia (M) and other diseases/abnormalities (O). It should be noted that one patient may contains one or multiple labels, as shown in Fig.2(b).

It can be seen from Fig.1 that the data distribution is extremely uneven. There are about 1,000 images in the three categories of normal (N), diabetic retinopathy (D), and other diseases (O). To this end, we use a weighted loss function to alleviate the problem of data imbalance.

IV. METHODS

The purpose of this study is to establish a framework for automatic identification of multi-label fundus diseases, and to achieve it by designing a corresponding ensemble model. As shown in Fig.3, several components form a corresponding ensemble model, and each component is a neural network composed of two parts. The first part is a feature extractor, including a pre-trained model initialized by transfer learning without top layer. The second part is a multi-label classifier that makes predictions based on the above features. This is a customized neural network implemented through training from scratch. Next, the specific implementation steps of the framework will be introduced in detail.

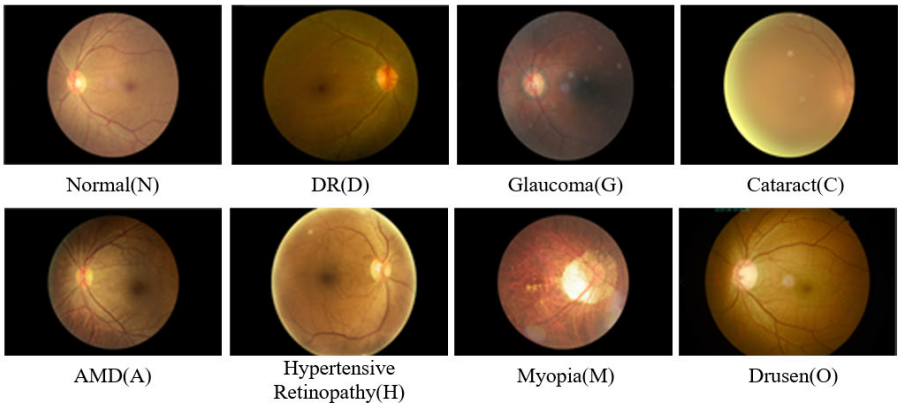
A. DATA PREPROCESSING

First, the original data is divided into training set and validation set according to the ratio of 9:1, then the data is cropped, the aspect ratio is cropped to 1:1, and finally the image is resized to 299 $\times$ 299. As shown in Fig.2(c), we have used data enhancement techniques, such as random rotation of 45, 90 degrees, translation, etc. to expand the size of the data set, retaining the main features of the original image, not a complete copy. Then the histogram equalization operation is performed on the original image and the gray image respectively, so that the gray value distribution of the image is more uniform, the contrast is higher, and the detail characteristics are more vivid.

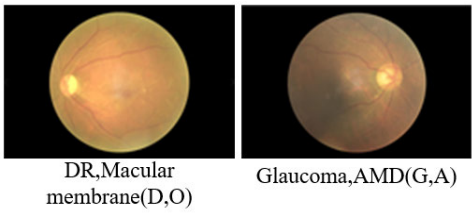
B. ENSEMBLE STRATEGY

If a better performance classification model is used as the basic classification model, the performance of the overall classification model will be improved[32]. At the same time, the predictive ability of the classifier is closely related to its ability to extract high quality features. Therefore, a high-performance CNN must be selected as the feature extractor.

Due to its high capacity/flexibility, deep neural networks usually have the characteristics of high variance and low deviation. The variance of models can be dramatically reduced via averaging if the models are independent [27], [30]. In the current work, we average the sigmoid probability values of all models to solve this problem, because these probabilities from different models may have varying output magnitudes. As shown in Fig.3, we decided to combine the two weak classification models here in order to obtain a better strong classification model. The integration strategy is to first perform histogram equalization on the original images and grayscale images after data preprocessing. The operation is used for two sets of equal training data, and then the same CNN is independently trained on these two sets of equal data sets as two weak classification models, and finally the sigmoid output probabilities of the two models are averaged as the final output value.



(a) Sample images for each label



(b) Sample images with multiple labels

(c) Data preprocessing

FIGURE 2. Sample images.

C. FEATURE EXTRACTOR

In order to train a strong feature extractor, we need to choose an excellent CNN as the first part to perform feature extraction on each fundus image and generate a compact feature vector representation. The current development of CNN is generally based on fixed resources for development. If the computing power is sufficient, the network will continue to deepen. EfficientNet [31] used simple and efficient composite coefficients to uniformly scale all dimensions of depth, width and resolution. And EfficientNet surpassed state-of-the-art accuracy with an order of magnitude fewer parameters and floatingpoint operations per second (FLOPS), on both ImageNet and five commonly used transfer learning datasets. Due to the high efficiency of EfficientNet, we proposed a multi-label classification model of fundus images using EfficientNet as the feature extractor. At the same time, we also used some excellent CNNs as feature extractors for comparative research.

We first used the classic VggNet [27]. In order to improve the computational speed of the model and its adaptability to scale, we used ResNet50 [12] and InceptionV3 [6]. Then used Xception [11] (Improved CNN of InceptionV3), DenseNet [19] and MobileNet [18]. A major feature of DenseNet is that it realizes the reuse of features, and MobileNet is a lightweight CNN. In this task, considering the connection with the second part classifier, we remove all the top layers of CNNs.

In this paper we studied two different initialization strategies. First, we initialize the network parameters with random values [28] to train the model. Second, we initialize the network with pre-trained parameters to transfer information from different domains and tasks. We use fine-tuning (FT) for training in transfer learning tasks. The CNN network pre-trained on ImageNet [29] is used as a feature extractor, parameters are initialized, and then the entire network is retrained with samples from the new domain.

D. DESIGN OF CLASSIFIER

According to the specific data distribution and difficulties of this task, a customized neural network is used as a multi-label

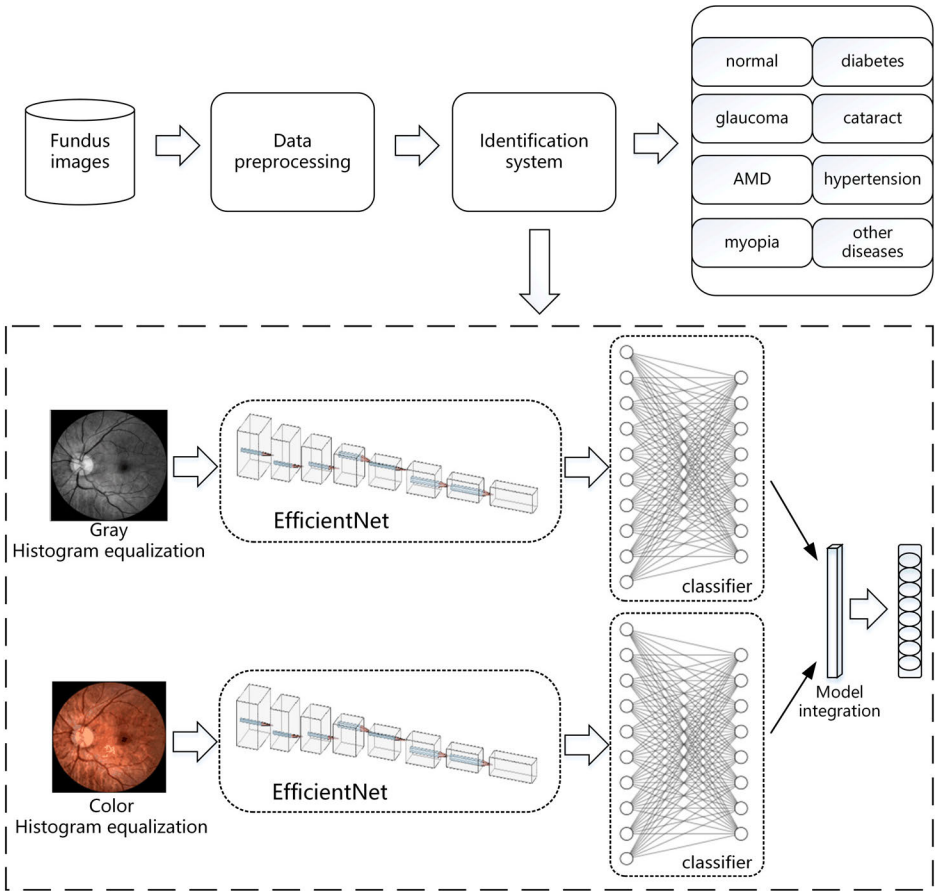


FIGURE 3. Automatic recognition framework for multi-label fundus diseases.

classifier, and the output of the feature extractor is the input of the multi-label classifier. For different feature extractors, the classifiers have the same network depth but different parameters. This slight difference affects important changes in predictive performance.

In this task, after multiple sets of experiments and observations, the following custom neural network is designed as a classifier. Adding or deleting a layer will reduce the learning ability of the network. First, the feature maps extracted from CNN are in high-dimensional space, but they can be down-sampled by average pooling (AP). Therefore, we designed a gap layer as the first layer of the classifier in order to connect to average pooling for down sampling. Secondly, add a dropout layer. This method can effectively omit a large number of hidden neurons during the training process to ensure the validity of the data; if the network relies on some nodes too much s in a certain layer, it can also reduce or prevent data overfitting. We update each layer of nodes with a probability of p = 0.5. Finally, it is a fully connected output layer with 8 hidden neurons, and the activation function is sigmoid. The complete model structure when the feature extractor is EfficientNet is shown in Table 1.

E. MULTI-LABEL LOSS FUNCTION

Our task is a multi-label classification problem, so we cannot use the traditional loss function to train the model. All images $X = \{x_1, x_2 \ldots x_n\}$, x_i are associated with the real label y_i, we seek a classification function F: X-Y that can minimize the loss function L. we use N sets of labeled training data $(x_i, y_i)i = 1...n$ and apply one-hot method to encode each y_i, and each y has 8 labels. We transform the multi-label classification problem into a two-class classification problem on each label, calculate the loss value of each label of each sample, and then take the average. After studying the weighting loss functions (such as positive/negative balance and class balance), we decided to use the following weighted binary cross entropy as the loss function, where the loss_weight = (1,1.2,1.5,1.5,1.5,1.5, 1.5, 1.2).

$$L = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\left(p(y_i)\right) + (1 - y_i)\log\left(1 - p(y_i)\right) \quad (1)$$

where N indicates the number of samples, y_i is the label of sample i. The positive class is 1, and the negative class is 0. And $p(y_i)$ is the probability that sample i is predicted to be positive.

After having the loss function, we need to optimize the learning parameters by optimization function. Different optimizers can have different effects on parameter training, so we closely focused on the effects of SGD and Adam on model performance. Under the same conditions, we conducted multiple comparison experiments. It is found that Adam is significantly better than SGD in terms of convergence and shortening training time. It may be because the gradient of

TABLE 1. Model structure.

Stage i	Operator Fi	Resolution Hi x Wi	Channels Ci	Layers Li
1	Conv3x3	299x299	40	1
2	MBConv1,k3x3	150x150	24	2
3	MBConv6,k3x3	75x75	32	3
4	MBConv6,k5x5	38x38	48	3
5	MBConv6,k3x3	19x19	96	5
6	MBConv6,k5x5	19x19	136	5
7	MBConv6,k5x5	10x10	232	6
8	MBConv6,k3x3	10x10	384	2
9	Conv1x1	10x10	1536	1
10	AveragePooling &dropout	2x2	1536	1
11	Dense	1x1	8	1
12	sigmoid	1x1	8	1

each sample is updated every time when we used SGD as optimizer, which increases the noise. Each iteration is not toward the overall optimization direction, and may converge to a local minimum, resulting in a decrease in accuracy. And Adam not only stores the exponential decay average of the square of the past gradient like RMSprop, but also maintains the exponential decay average of the past gradient like momentum. Makes it has a significant improvement in training speed and stability. SGD and Adam are defined in equations (2) and (3).

$$\theta_{t+1} = \theta_t - \eta \bullet \nabla_\theta J(\theta_t; x^{(i)}; y^{(i)}) \tag{2}$$

$$\begin{cases} m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \\ v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \\ \theta_{t+1} = \theta_t - \dfrac{\eta}{\sqrt{\hat{v}_t + \varepsilon}} \hat{m}_t \end{cases} \tag{3}$$

In formula (2), $\nabla_\theta J(\theta_t; x^{(i)}; y^{(i)})$ is the gradient of the objective function, and η is the learning rate. In formula (3), m_t and v_t are the first moment estimation and second moment estimation corresponding to the gradient g_t. $\hat{m}_t$ and $\hat{v}_t$ are the biases-corrected of m_t and v_t.

The above formula illustrates the similarities and differences between these two functions. By adding first-order momentum and second-order momentum, Adam makes the learning rate more adaptive while updating parameters.

V. EXPERIMENTS

A. METRICS

Accuracy is the most primitive evaluation index in classification problems, it refers to the proportion of samples that are classified correctly. Precision refers to the probability of the sample that is actually positive among all the samples that are predicted to be positive. The recall rate is for the original sample, and refers to the probability of being predicted as a positive sample in the actual positive sample. AUC is the area under the ROC curve. The closer to 1, the better the classification performance of the model. It is often used to measure the stability of the model. The kappa coefficient is a method used in statistics to assess consistency, and it can also be used to measure classification accuracy. Fβ_score is the harmonic average of precision and recall. When β =1, it is the common f1_score, and f1_score close to 1 indicates good performance. Final_score refers to the average value of F1_score, Kappa and Auc. These evaluation indicators are as follows:

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

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

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

$$\text{F1\_score} = \frac{2TP}{2TP + FP + FN} \tag{7}$$

$$\text{F}\beta\text{\_score} = \frac{(1+\beta^2) \times Precision \times recall}{\beta^2 \times Precision + recall} \tag{8}$$

$$\text{Kappa} = \frac{p_0 - p_e}{1 - p_e} \tag{9}$$

$$\text{Final\_score} = \frac{F1\_score + Kappa + Auc}{3} \tag{10}$$

B. CONFIGURATION

All experiments are carried out on a dedicated server, the CPU is i9 9900K, 8 cores and 16 threads, the GPU is NVIDIA RTX2080Ti, and the memory is 64gb. The data set provided by the competition is divided into 90% for training, 10% for verification, and an additional 1,000 data sets for testing. The configuration of hyper-parameter are shown in Table 2.

In order to better verify the effectiveness of our method, we designed multiple sets of comparative experiments.

C. COMPARISON EXPERIMENT OF TRAINING STRATEGY

It can be seen from Table 3 that the effect of training CNN from scratch is not very good. Due to the limitation of the data set, the results of the model on the validation set are significantly higher than the results on the testing set, it seems that the training process may be over-fitting. In order to solve

FIGURE 4. Fine tuning.

TABLE 2. Hyper-parameter configuration.

Configuration	Value
Optimisation function	Adam
Epoch	30(Complete training) 10(fine tuning)
BatchSize	15
Learning rate	1.00E-03
Batch Normalization	True
Drop out	5.00E-01
ReduceLROnPlateau	Monitor='final_score', factor=0.2, Patience=4 ,min_delta=0.001
EarlyStopping	Monitor = 'val_loss', patience=5
ModelCheckpoint	Monitor = 'final_score', mode='Max', save_best_only=True

this problem, we tried to use different strategies, We tried to add Batch Normalization [33]. At each layer of the network input, a normalization layer was inserted, which is a learnable network layer with parameters. The training speed of the network was improved, but the classification results were not significantly improved. Then we decided to fine-tune the pre-trained model trained on 'Imagenet'. From Fig. 4, it can be seen that the results on the testing set have been significantly improved. It may be because the convolutional neural network has a good hierarchical structure. The usual convolutional neural network has a hierarchical structure similar to convolution-pooling-convolution-pooling-full connection.

When the depth is sufficient, the features of each level of the image can be extracted. Although the fundus image is not exactly same with the data on 'Imagenet', its features are same with some low-level features such as edge features, colors, textures, etc. Therefore, the features extracted by the pre-training model are more accurate and comprehensive.

D. COMPARISON EXPERIMENT OF ENSEMBLE MODELS

In the course of the experiment, we tried to use multiple different CNN pre-training models to independently train the same dataset and then integrate their output values, but we did not get good experimental results. Therefore, in the next experiment, we will apply two identical models to different data sets according to the method introduced in part IV, and then integrate the output. We integrated the output of the previous model and didn't train a new model. Therefore, Table 4 doesn't include the values on the validation set, only the results on the test set. From the results in Table 4, it can be seen that the final_score of some models has improved slightly, among which EfficientNetB3 and InceptionResNetV2 have improved obviously. We speculate that the difference in the performance of the basic classifier affects the result of the final ensemble model. the integration of weak classifiers can bring some improvement to the results.

E. COMPARISON EXPERIMENT OF DIFFERENT IMAGE SIZES

In this experiment, we changed the image size to 448 × 448, and still used those CNN pre-training models in the previous experiment for feature extraction, and then integrate the results. MobileNet cannot be used due to limited image size. From the result of TABLE 5, it can be seen that increasing the image size has different degrees of improvement for each model. We speculate that the larger the image size, the more texture and information it contains, and better features can be captured. However, according to our experiments, when the image size reaches a certain level, the classification performance will not increase, but may decrease. Moreover, the computational cost will increase accordingly.

F. VERIFY THE VALIDITY OF THE MODEL

In order to verify the validity of our proposed model, We collected 40 fundus images of real patients from https://drive.grand-challenge.org/. The photographs for this database were obtained from a diabetic retinopathy screening program in The Netherlands. It contains diagnostic keywords marked by the doctor, and several images have two labels. We use the trained EfficientNetB3 ensemble model to predict these 40 fundus images, and the results are shown in TABLE 6. The model we proposed can get good recognition

TABLE 3. Complete training.

Model	Val_Accuracy	Val_Precision	Val_Recall	Val_Fβ_score	Auc	Kappa	F1_score	Final_score
Vgg16	0.86	0.62	0.39	0.87	0.64	0.35	0.83	0.61
Vgg19	0.86	0.61	0.37	0.86	0.63	0.34	0.82	0.60
Xception	0.90	0.65	0.49	0.89	0.61	0.40	0.85	0.62
InceptionV3	0.87	0.65	0.48	0.86	0.65	0.40	0.85	0.63
EfficientNetB3	**0.90**	**0.66**	**0.50**	**0.89**	**0.67**	**0.43**	**0.85**	**0.65**
ResNet50	0.86	0.64	0.39	0.85	0.63	0.34	0.81	0.59
MobileNet	0.89	0.58	0.41	0.88	0.63	0.36	0.83	0.61
InceptionResNetV2	0.89	0.61	0.36	0.89	0.66	0.40	0.83	0.63
DenseNet	0.88	0.62	0.41	0.89	0.63	0.39	0.83	0.62

TABLE 4. Ensemble results.

Model	Auc	Kappa	F1_score	Final_score
Vgg16	0.69	0.45	0.85	0.66
Vgg19	0.69	0.45	0.86	0.66
Xception	0.68	0.43	0.85	0.65
InceptionV3	0.68	0.47	0.87	0.67
EfficientNetB3	**0.73**	**0.50**	**0.88**	**0.70**
ResNet50	0.66	0.42	0.85	0.64
MobileNet	0.67	0.39	0.87	0.64
InceptionResNetV2	**0.72**	**0.47**	**0.87**	**0.69**
DenseNet	0.70	0.44	0.85	0.66

TABLE 5. Results of image size 448 × 448.

Model	Val_Accuracy	Val_Precision	Val_Recall	Val_Fβ_score	Auc	Kappa	F1_score	Final_score
Vgg16	0.91	0.68	0.58	0.91	0.72	0.50	0.89	0.70
Vgg19	0.91	0.70	0.57	0.91	0.70	0.48	0.88	0.69
Xception	0.92	0.70	0.64	0.92	**0.73**	**0.51**	**0.89**	**0.71**
InceptionV3	0.91	0.68	0.64	0.91	0.72	0.46	0.87	0.68
EfficientNetB3	0.92	**0.71**	**0.66**	0.92	**0.74**	**0.52**	**0.89**	**0.72**
ResNet50	0.89	0.65	0.56	0.90	0.67	0.45	0.84	0.65
InceptionResNetV2	0.91	0.71	0.64	0.91	0.72	0.49	0.88	0.70
DenseNet	0.91	0.69	0.60	0.91	0.70	0.45	0.87	0.67

TABLE 6. Results of 40 fundus images.

Model	Accuracy	Precision	Recall	Fβ_score	Auc	Kappa	F1_score	Final_score
EfficientNetB3	0.89	0.63	0.58	0.89	0.73	0.49	0.89	0.70

results in the situation of multiple fundus diseases and some patients with two eye diseases.

G. EXPERIMENT SUMMARY

In this paper, in order to solve the multi-label classification task of fundus images, we adopt a variety of different training strategies. According to the results, the best training method is shown as follows: First, we resize the preprocessed data to 448 × 448. Second, we use the parameters of the pre-trained CNN model EfficientNet into the transferred neural network, where the pre-trained EfficientNet is trained on ImageNet. Then, fundus images are used to fine-tune the transferred neural network. Finally, we use the averaging method to integrate two independently trained weak classifiers to get

the final result. In the process of integration, the quality of each individual weak classifier will have a great impact on the performance of the final integration. When the training data is very limited and the difference between the source domain and the target domain is small, fine-tuning the pre-trained CNN with a small data set can bring good results.

VI. DISCUSSION AND CONCLUSION

We have developed an automated end-to-end framework for detecting multi-label fundus diseases, and achieved well results on the public data set ODIR-2019. According to the obtained results, as more and more public data sets are available, the training of deep neural networks in the medical field is a feasible choice, but the practical application of deep learning in clinical practice is still an open problem. First, for the ODIR-2019 data set, one of the labels 'O' (other diseases) contains a variety of uncommon fundus diseases. The amount of data for some diseases is very limited, which makes it very difficult to improve the performance of a network. Another basic limitation comes from the black box of the nature of deep networks. The network automatically learns features from images, but the specific features learned are unknown.

In the future work, we need to collect large training datasets containing tens of thousands of abnormal cases from other hospitals through different types of cameras. It will provide more different features to help improve accuracy and generalization. At the same time, for the detection of fundus diseases, in addition to the image data itself, other factors such as age, gender, and family history can also be considered for potential fundus diseases. In order to improve the clinical acceptance of deep learning models, understanding various aspects of deep neural networks and visualization is also a very important research field.

ACKNOWLEDGEMENT

The fundus images used in this research is ''real-life'' set of patient information collected by Shanggong Medical Technology Co., Ltd. from different hospitals/medical centers in China.

REFERENCES

[1] J. B. Jonas, R. R. A. Bourne, R. A. White, S. R. Flaxman, J. Keeffe, J. Leasher, K. Naidoo, K. Pesudovs, H. Price, T. Y. Wong, S. Resnikoff, and H. R. Taylor, ''Visual impairment and blindness due to macular diseases globally: A systematic review and meta-analysis,'' *Amer. J. Ophthalmol.*, vol. 158, no. 4, pp. 808–815, Oct. 2014.

[2] J. L. Leasher, R. R. A. Bourne, S. R. Flaxman, J. B. Jonas, J. Keeffe, K. Naidoo, K. Pesudovs, H. Price, R. A. White, T. Y. Wong, S. Resnikoff, and H. R. Taylor, ''Global estimates on the number of people blind or visually impaired by diabetic retinopathy: A meta-analysis from 1990 to 2010,'' *Diabetes Care*, vol. 39, pp. 1643–1649, Sep. 2016, doi: 10.2337/dc15-2171.

[3] O. B. Walton, R. B. Garoon, C. Y. Weng, J. Gross, A. K. Young, K. A. Camero, H. Jin, P. E. Carvounis, R. E. Coffee, and Y. I. Chu, ''Evaluation of automated teleretinal screening program for diabetic retinopathy,'' *JAMA Ophthalmol.*, vol. 134, no. 2, pp. 204–209, Feb. 2016, doi: 10.1001/jamaophthalmol.2015.5083.

[4] N. M. Bressler, ''Age-related macular degeneration is the leading cause of blindness,'' *JAMA*, vol. 291, no. 15, pp. 1900–1901, Apr. 2004, doi: 10.1001/jama.291.15.1900.

[5] H. Ye, Q. Zhang, X. Liu, X. Cai, W. Yu, S. Yu, T. Wang, W. Lu, X. Li, H. Jin, Y. Hu, X. Kang, and P. Zhao, ''Prevalence of age-related macular degeneration in an elderly urban Chinese population in China: The Jiangning eye study,'' *Investigative Ophthalmol. Vis. Sci.*, vol. 55, no. 10, pp. 6374–6380, Sep. 2014, doi: 10.1167/iovs.14-14899.

[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, ''Going deeper with convolutions,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.

[7] S. J. Pan and Q. Yang, ''A survey on transfer learning,'' *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010, doi: 10.1109/TKDE.2009.191.

[8] R. S. Michalski, ''A theory and methodology of inductive learning,'' in *Machine Learning*. Berlin, Germany: Springer, 1983, pp. 83–134.

[9] L. Rokach, ''Ensemble-based classifiers,'' *Artif. Intell. Rev.*, vol. 33, nos. 1–2, pp. 1–39, 2010.

[10] D. Opitz and R. Maclin, ''Popular ensemble methods: An empirical study,'' *J. Artif. Intell. Res.*, vol. 11, pp. 169–198, Aug. 1999.

[11] F. Chollet, ''Xception: Deep learning with depthwise separable convolutions,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1800–1807, doi: 10.1109/CVPR.2017.195.

[12] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[13] J. M. P. Dias, C. M. Oliveira, and L. A. da Silva Cruz, ''Retinal image quality assessment using generic image quality indicators,'' *Inf. Fusion*, vol. 19, pp. 73–90, Sep. 2014.

[14] H. Cui, S. Shen, W. Gao, H. Liu, and Z. Wang, ''Efficient and robust large-scale structure-from-motion via track selection and camera prioritization,'' *ISPRS J. Photogramm. Remote Sens.*, vol. 156, pp. 202–214, Oct. 2019.

[15] E. Peli and T. Peli, ''Restoration of retinal images obtained through cataracts,'' *IEEE Trans. Med. Imag.*, vol. 8, no. 4, pp. 401–406, Dec. 1989, doi: 10.1109/42.41493.

[16] J. Cheng, J. Liu, Y. Xu, F. Yin, D. W. K. Wong, N.-M. Tan, D. Tao, C.-Y. Cheng, T. Aung, and T. Y. Wong, ''Superpixel classification based optic disc and optic cup segmentation for glaucoma screening,'' *IEEE Trans. Med. Imag.*, vol. 32, no. 6, pp. 1019–1032, Jun. 2013, doi: 10.1109/TMI.2013.2247770.

[17] C. Köse, U. Şevik, C. İkibaş, and H. Erdöl, ''Simple methods for segmentation and measurement of diabetic retinopathy lesions in retinal fundus images,'' *Comput. Methods Programs Biomed.*, vol. 107, no. 2, pp. 274–293, Aug. 2012.

[18] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, ''MobileNets: Efficient convolutional neural networks for mobile vision applications,'' 2017, *arXiv:1704.04861*. [Online]. Available: http://arxiv.org/abs/1704.04861

[19] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, ''Densely Connected Convolutional Networks,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.

[20] Y. Zhao, Y. Liu, X. Wu, S. P. Harding, and Y. Zheng, ''Retinal vessel segmentation: An efficient graph cut approach with retinex and local phase,'' *PLoS ONE*, vol. 10, no. 4, Apr. 2015, Art. no. e0122332.

[21] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, ''Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs,'' *JAMA*, vol. 316, no. 22, pp. 2402–2410, Dec. 2016.

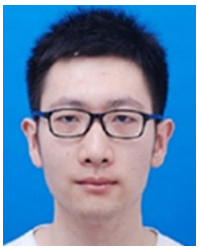
[22] R. Gargeya and T. Leng, ''Automated identification of diabetic retinopathy using deep learning,'' *Ophthalmology*, vol. 124, no. 7, pp. 962–969, Jul. 2017.

[23] E. Long, H. Lin, Z. Liu, X. Wu, L. Wang, J. Jiang, Y. An, Z. Lin, X. Li, J. Chen, J. Li, Q. Cao, D. Wang, X. Liu, W. Chen, and Y. Liu, ''An artificial intelligence platform for the multihospital collaborative management of congenital cataracts,'' *Nature Biomed. Eng.*, vol. 1, no. 2, pp. 1–8, Jan. 2017.

[24] A. Govindaiah, M. A. Hussain, R. T. Smith, and A. Bhuiyan, ''Deep convolutional neural network based screening and assessment of age-related macular degeneration from fundus images,'' in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Washington, DC, USA, Apr. 2018, pp. 1525–1528, doi: 10.1109/ISBI.2018.8363863.

[25] X. Li, T. Pang, B. Xiong, W. Liu, P. Liang, and T. Wang, "Convolutional neural networks based transfer learning for diabetic retinopathy fundus image classification," in *Proc. 10th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Shanghai, China, Oct. 2017, pp. 1–11, doi: 10.1109/CISP-BMEI.2017.8301998.
[26] W. Song, Y. Cao, Z. Qiao, Q. Wang, and J.-J. Yang, "An improved semi-supervised learning method on cataract fundus image classification," in *Proc. IEEE 43rd Annu. Comput. Softw. Appl. Conf. (COMPSAC)*, Milwaukee, WI, USA, Jul. 2019, pp. 362–367, doi: 10.1109/COMPSAC.2019.10233.
[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.
[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Apr. 2015.
[30] C. Ju, A. Bibaut, and M. van der Laan, "The relative performance of ensemble methods with deep convolutional neural networks for image classification," *J. Appl. Statist.*, vol. 45, no. 15, pp. 2800–2818, Feb. 2018.
[31] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
[32] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*. Berlin, Germany: Springer, 2000, pp. 1–15.
[33] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. ICML*, 2015, pp. 1–11.
[34] L. Yan, B. Fan, H. Liu, C. Huo, S. Xiang, and C. Pan, "Triplet adversarial domain adaptation for pixel-level classification of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3558–3573, May 2020.
[35] H. Liu, X. Tang, and S. Shen, "Depth-map completion for large indoor scene reconstruction," *Pattern Recognit.*, vol. 99, Mar. 2020, Art. no. 107112, doi: 10.1016/j.patcog.2019.107112.
[36] G. Molodij, E. N. Ribak, M. Glanc, and G. Chenegros, "Enhancing retinal images by extracting structural information," *Opt. Commun.*, vol. 313, pp. 321–328, Feb. 2014.
[37] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011, doi: 10.1109/TNN.2010.2091281.
[38] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits Syst. Mag.*, vol. 6, no. 3, pp. 21–45, 3rd Quart., 2006, doi: 10.1109/MCAS.2006.1688199.
[39] S. Yuan, Y. Zhang, J. Tang, W. Hall, and J. B. Cabotà, "Expert finding in community question answering: A review," *Artif. Intell. Rev.*, vol. 53, no. 2, pp. 843–874, Feb. 2020, doi: 10.1007/s10462-018-09680-6.

JING WANG received the B.S. degree from the Henan University of Science and Technology, China, in 2006, and the Ph.D. degree from the College of Computing and Communication Engineering, Graduate University of Chinese Academy of Science, Beijing, China, in 2012. She is currently an Associate Professor with the School of Computer Science and Technique, Henan Polytechnic University, Jiaozuo, China. Her research interests include image processing, computer vision, and machine learning.



LIU YANG received the B.E. degree in electronic information engineering from the Hubei University of Economics, in 2017. He is currently pursuing the M.S. degree in software engineering with Henan Polytechnic University. His research interests include image processing, machine learning, computer vision, pattern recognition, and deep learning.

ZHANQIANG HUO received the B.Sc. degree in mathematics and applied mathematics from the Hebei Normal University of Science and Technology, China, in 2003, and the M.Sc. degree in computer software and theory and the Ph.D. degree in circuit and system from Yanshan University, China, in 2006 and 2009, respectively. He is currently an Associate Professor with the College of Computer Science and Technology, Henan Polytechnic University, China. His research interests include computer vision and machine learning.

WEIFENG HE received the M.S. degree in medical imaging diagnosis and digital gastrointestinal examination of the digestive system from Henan University. He currently works at Henan Polytechnic University Hospital and has been engaged in medical imaging diagnosis for eight years. He has rich experience in first-line diagnosis.

JUNWEI LUO received the Ph.D. degree in computer science from Central South University, Changsha, China. He is currently an Associate Professor with Henan Polytechnic University, Jiaozuo, China. His current research interests include machine learning, bioinformatics, and data mining.