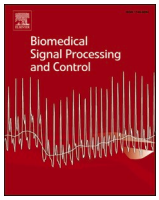
Biomedical Signal Processing and Control 96 (2024) 106621

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc



Cnn-trans model: A parallel dual-branch network for fundus image classification

Shuxian Liu [a,*], Wei Wang [b], Le Deng [a], Huan Xu [a]

[a] School of Information Science and Engineering, Xinjiang University, Urumqi 830017, China
[b] School of Information Science and Technology, Xinjiang Teacher's College, Urumqi 830043, China

ARTICLE INFO

Keywords:
Fundus image classification
Parallel dual branch
Attention mechanism
Feature fusion
CNN

ABSTRACT

The existence of fundus diseases not only endangers people's vision, but also brings serious economic burden to the society. Fundus images are an objective and standard basis for the diagnosis of fundus diseases. With the continuous advancement of computer science, deep learning methods dominated by convolutional neural networks (CNN) have been widely used in fundus image classification. However, the current CNN-based fundus image classification research still has a lot of room for improvement: CNN cannot effectively avoid the interference of repeated background information and has limited ability to model the whole world. In response to the above findings, this paper proposes the CNN-Trans model. The CNN-Trans model is a parallel dual-branch network, which is the two branches of CNN-LSTM and Vision Transform (ViT). The CNN-LSTM branch uses Xception after transfer learning. As the original feature extractor, LSTM is responsible for dealing with the gradient disappearance problem in neural network iterations before the classification head, and then introduces a new type of lightweight attention mechanism between Xception and LSTM: Coordinate Attention, so as to emphasize the key information related to classification and suppress the less useful repeated background information; while the self-attention mechanism in the ViT branch is not limited by local interactions, it can establish long-distance dependence on the target and extract global features. Finally, the concatenation (Concat) operation is used to fuse the features of the two branches. The local features extracted by the CNN-LSTM branch and the global features extracted by the ViT branch form complementary advantages. After feature fusion, more comprehensive image feature information is sent to the to the classification layer. Finally, after a large number of experimental tests and comparisons, the results show that: the CNN-Trans model achieved an accuracy of 80.68% on the fundus image classification task, and the CNN-Trans model has a classification that is comparable to the state-of-the-art methods. performance..

1. Introduction

With the continuous development of modern lifestyle, computers, mobile phones and other digital screens have become more and more popular. However, staring at electronic screens for a long time can cause eye fatigue, dryness and discomfort. This phenomenon is very common. The main reason is that we When using electronic devices, the frequency of blinking decreases and the eyes are not adequately lubricated. In addition, factors such as light radiation, unhygienic eye use, age decline, etc. can also cause eye damage. For individuals, eye health problems can lead to less efficient learning, as poor vision can affect work performance and academic performance. For society, people with vision problems may need more frequent eye treatments and glasses, which increases the demand for medical resources and is a significant economic burden.

There are many kinds of fundus diseases, which seriously affect human visual health and are one of the main causes of blindness worldwide. Cataracts, diabetic retinopathy, age-related macular degeneration and high myopia are all common blinding diseases. Although there are many types of fundus diseases, many fundus diseases will produce certain signs in the fundus. However, since most patients have low awareness of these signs, they may ignore eye changes and delay treatment. Therefore, early detection of fundus diseases is of great significance for subsequent treatment.

Fundus examination uses a professional fundus camera or scanning instrument to capture fundus images for analysis. These images show the structures inside the fundus of the eye, including the retina, optic nerve,

* Corresponding author.
E-mail addresses: liushuxian@xju.edu.cn (S. Liu), 2292590189@qq.com (W. Wang), sc_dengle@stu.xju.edu.cn (L. Deng), xhuan@stu.xju.edu.cn (H. Xu).

https://doi.org/10.1016/j.bspc.2024.106621
Received 21 May 2023; Received in revised form 14 February 2024; Accepted 25 June 2024
Available online 13 July 2024

blood vessels, and other important tissues. By observing fundus images, doctors can detect whether the fundus blood vessels are normal and whether there are abnormal expansions, lesions, or bleeding. Fundus imaging has many advantages and is a non-invasive method that does not require any surgery or irritating procedures. Secondly, fundus images provide objective visual information, reducing the impact of subjective factors on diagnosis. Doctors can assess the status of systemic blood circulation, thereby early detecting and managing systemic health-related problems.

Therefore, fundus examination can not only be used to detect eye diseases, such as vitreous, retina, choroid and optic nerve diseases, but also a monitoring window for many systemic diseases [1]. However, clinical diagnosis of fundus diseases relies on professional ophthalmologists. As the number of patients continues to increase, fundus image data is also becoming larger and larger. The diagnosis of fundus data will consume a lot of energy and time of ophthalmologists. And limited professional medical resources should be used in more valuable directions.

In order to assist ophthalmologists to make accurate diagnosis based on fundus images, it is necessary to develop a computer-aided diagnosis system (CAD for short) for fundus image classification as soon as possible. At present, in the field of computer vision, the challenges of fundus image classification are mainly reflected in the following aspects [2,3]:

(1) Publicly available fundus image datasets are generally small in size. There are two main reasons:

a. Fundus images contain sensitive personal information, such as retinal structure, etc. When processing and sharing fundus image data, appropriate privacy protection measures must be taken to ensure that patient privacy and data security are not violated. To protect patient privacy, fundus image data collected by hospitals are usually kept confidential.

b. In order to accurately classify fundus images, manual annotation and verification of the data set is the first step, but this requires highly specialized knowledge and experience to ensure the accuracy of diagnostic results, therefore, mobilizing a large number of professional ophthalmologists Participating in the creation and annotation process of data sets is a difficult problem for researchers.

(2)Most fundus image datasets suffer from severe class imbalance problem, which becomes more prominent when the number of classes increases. The direct reason why category imbalance occurs is that the probability of developing fundus diseases is quite different. However, there are various factors that affect the probability of developing fundus diseases[4–6], such as:

a. Geographical factors: The incidence of trachoma is usually higher in warm, dry and poor areas, which are more likely to become hotbeds for the spread of infectious diseases. For example, sub-Saharan Africa, parts of Asia, and some countries in Oceania have long been high-incidence areas for trachoma.

b. Hygiene factors. In areas with poor sanitary conditions, such as places lacking clean drinking water and sanitation facilities, people are more likely to be exposed to sources of infection, thereby increasing the risk of infection from bacterial or viral fundus diseases.

c. Genetic factors, there are racial differences in the occurrence and development of myopia. Studies on the incidence of myopia in different races have found that in rapidly developing economies in East Asia such as China, its prevalence is growing at an alarming rate.

d. Age factor. Cataracts are common in middle-aged and elderly people over 50 years old. This is because the organs in the body gradually age and the lens will also undergo degenerative changes, such as turbidity.

(3) Different models of equipment used to obtain fundus images in hospitals may have differences in image resolution, lighting conditions, imaging quality, etc.. For example, some images may be produced by old equipment that has not been updated in time, and there are cases of low pixels and resolutions. It may also be taken in a dimly lit environment, resulting in uneven brightness and darkness in various areas of the fundus image. The low-quality images caused by these factors greatly increase the difficulty of classification.

In recent years, with the rapid development of deep learning technology in the field of computer vision, more and more researchers have begun to use deep learning algorithms to analyze and process fundus images. Deep learning is a machine learning method based on multilayer neural networks that can automatically learn and extract features from large amounts of data to achieve highly accurate image analysis and recognition. In fundus image analysis, deep learning technology can extract features in fundus images by training a large amount of fundus image data, and perform tasks such as classification, positioning and segmentation.

In 2019, Imran Qureshi et al. [7] compared related CAD systems based on statistical parameters for quantitative evaluation. The comparison results showed that accurate development of CAD systems is still needed to assist clinical diagnosis of diabetic retinopathy. In 2021, Chea et al. [8] found that existing computer-aided systems cannot simultaneously detect multiple major eye diseases. To better understand the multi-category classification of fundus images, they used an optimal residual deep neural processing technology. In 2023, Wen Jingyi et al. [9] believed that the symptoms of chronic kidney disease, hypertension, type 2 diabetes and other diseases could be found based on the specific performance of retinal images, and pointed out that the latest development of AI technology is the use of retinal images to diagnose kidney diseases. Rapid large-scale screening and prognosis prediction bring great potential. These research results will help establish an accurate computer-aided diagnosis (CAD) system to provide doctors with more accurate diagnosis results and treatment suggestions, thereby improving patients' diagnosis and treatment experience and treatment effects.

However, in the face of the urgent needs of clinical fundus disease screening, diagnosis and other auxiliary medical care, Existing CAD systems generally can only detect a few major eye diseases. There are few related studies on detecting multiple eye diseases and their performance is average. And most studies directly transfer the features learned by the last convolutional layer of CNN to the classification layer. A problem: (1) It is impossible to effectively avoid the interference of repeated fundus image background information, and ignores the subtle changes in the fundus image that may represent the diseased tissue; (2) The core operation of CNN is the convolution kernel, which has local sensitivity. However, in the fundus image The lesion area may be discontinuous, so it is difficult for CNN to grasp the global features and extract them.

This study contributes a new solution to the two limitations of the above-mentioned fundus image classification research: through a series of algorithm improvements to the convolutional neural network (CNN), the CNN-Trans model is proposed, which is an attention mechanism and feature fusion. The detailed architecture and principles of the model are explained in Section 3.2 of this article., considering the problem that the existing CAD system detects a small number of fundus diseases, Our model also conducted related 7 classification experiments on the fundus image data set: age-related macular degeneration (Age_degeneration), cataract (Cataract), diabetic retinopathy (Diabetes), glaucoma (Glaucoma), hypertension complications (Hypertension), pathological myopia (Myopia) and normal (Normal), the classification accuracy reached 80.68 %, which achieved better performance than other similar studies.

2. Related work

The research of image classification on retinal fundus images has always been favored by researchers. With the continuous development of image imaging technology, more and more scholars have carried out a series of research on fundus images of retinal fundus diseases and achieved many results.

In 2017, Joon Yul Choi et al. [10] performed multi-category

classification of fundus images based on random forest transfer learning of VGG-19 architecture, including normal and 9 abnormal retinal fundus diseases, and proved that: with the increase of the number of categories, the performance of the deep learning model gradually decreases.

In 2018, Sourya Sengupta et al. [3] published the first review article on the application of deep learning in ophthalmological diagnosis of retinal fundus images. They first review various retinal image datasets available for deep learning purposes. Then, the application of deep learning to optic disc, blood vessel and retinal layer segmentation and lesion detection is reviewed. Finally, the latest deep learning models for age-related macular degeneration, glaucoma, diabetic macular edema, and diabetic retinopathy were also reported.

In 2019, Andres Diaz-Pinto et al. [11] used five different ImageNet training models (VGG16, VGG19, InceptionV3, ResNet50 and Xception) for automated glaucoma assessment using fundus images. The experiment process used five public databases (1707 images), and finally found that: after using the Xception architecture, the average AUC was 0.9605, the average specificity was 0.8580, and the average sensitivity was 0.9346, and its performance was significantly higher than other models.

In 2020, Jothi J. Balaji et al. [12] compared diabetic retinopathy, high myopia, and normal Foveal Avascular Zone (FAZ) in fundus images. Their approach is to first iteratively preprocess the image with a DOG filter, followed by Prewitt edge detection and repeated image upscaling at different angles, applying image closure, removing noise and small objects, resulting in segmentation boundaries. Final average FAZ diameter (mm) derived from new automated techniques, manual segmentation (ground truth) and built-in instrument algorithms. The results of the study showed that there was a significant difference in the FAZ area ($p = 0.003$) of diabetic retinopathy compared with myopia ($p = 0.016$) and normal subjects.

In 2021, Masum Shah Junayed et al. [13] proposed the cataract network, which is proposed for automatic cataract detection in fundus images. The network is trained using small kernels, fewer training parameters and layers, and its computational cost and average the running time is significantly reduced compared to other pre-trained convolutional neural network (CNN) models. In the binary classification task of 1130 cataract and non-cataract fundus images, an accuracy rate of 99.13 % was achieved.

In 2022 Muhammad Mohsin Butt et al. [14] presented a hybrid technique for detecting and classifying diabetic retinopathy in fundus images: transfer learning (TL) is used on a pre-trained convolutional neural network (CNN) model to extract features that are combined to generate a hybrid feature vector. This feature vector is passed to various classifiers for binary and multi-class classification of fundus images. In the same year, Thisara Shyamalee et al. [15] used fundus images to classify glaucoma subjects, using three different convolutional neural network (CNN) architectures, and not only used a variety of data preprocessing and enhancement techniques to avoid over-simulation. Merging to achieve high accuracy, the performance obtained by different configurations of CNN architecture and hyperparameter tuning is also analyzed comparatively.

In 2023, A review published by Ademola E. Ilesanmi a et al. [16] conducted a comprehensive review of CNN algorithms for retinal fundus image segmentation and classification. The review included a total of 62 studies and analyzed the use of databases and the methods used. Strengths and weaknesses, etc., provide valuable insights, limitations, observations, and future directions in the field. Despite certain limitations, the research results show that the CNN algorithm is always able to achieve high accuracy. In the same year, Zhenzhen Lu et al. [17] introduced a squeeze-stimulated attention module into the constructed network.By modeling the interdependence between channels, the SE module recalibrates the feature map in the channel dimension, automatically obtains the importance values of different channels in the feature map, and makes the model emphasize the lesion features extracted from the fundus image, and enhanced the representation ability of custom classification network.

Through the investigation of fundus image classification literature in recent years, most of the studies on fundus image classification have chosen binary classification to solve the problem of "with or without a certain fundus disease". Their models are not good for various abnormal fundus. It has the ability to recognize, but it cannot guarantee good performance in detecting various fundus lesions, and some methods do not have model visualization interpretation, so the credibility of the model cannot be guaranteed.

This article aims to build an automatic screening system for detecting a variety of common blinding diseases of the fundus, and to make a visual interpretation of the results predicted by the model, which has more application value in clinical practice, to reduce the workload of doctors examining fundus diseases, and to facilitate patients themselves Screen fundus diseases to solve the problem of lack of experienced doctors in remote areas.

3. Materials and methods

In this chapter, the datasets used for training and testing and the deep learning model used in this study will be discussed. Datasets are discussed further in Section 3.1. Likewise, Section 3.2 discusses the proposed classification method for fundus images.

3.1. Dataset

In this section, we collect two fundus image datasets: Dataset-1 (D1) and Dataset-2 (D2) from different public sources.

D1 is ODIR-5 K ("real" patient information set collected by Shanggong Medical Technology Co., Ltd. from different hospitals/medical centers in China, updated in December 2019), 10,000 images of left and right eyes from 5,000 clinical patients Fundus images, some of the images in this dataset have multiple labels, and its release follows Chinese ethics and privacy rules. In this paper, a modified version of the original data set is used. After removing the multi-label fundus images, in order to reduce the interference of outlier data on the model, all the data marked as O are deleted, and then the data set is sorted according to the quality factors of the image. Age-related macular degeneration (A), cataract (C), diabetic retinopathy (D), glaucoma (G), hypertension complications (H), pathological myopia (M), normal (N), a total of 5684 images. The sample images of these seven types of eye diseases are shown in Fig. 1.

D2 is a cataract data set publicly available on kaggle. The data set consists of normal (N), diabetic retinopathy (D), cataract (C) and glaucoma retina (G) images. Each type of image has about 1000 images of D1 and D2. The specific composition is shown in Table 1.

It can be seen from this that the composition of analogies in D1 is unbalanced. The number of categories with the largest number in D1 is more than 20 times that of the category with the smallest number. Generally, more than 10 times is considered an unbalanced data set. We use the ImageDataGenerator class of keras to expand the data of the D1 dataset and save it in the original folder. The main methods include rotation by 45 degrees, the brightness change interval is [0.5,0.9], up and down flip and horizontal flip. The category composition of the modified datasets D1 and D2 is shown in Table 2.

In order to minimize the unnecessary interference of the extra noise brought by the black area of the fundus image to the feature extraction process, the redundant black area is cropped. In this paper, the OpenCV library is used to load the image into the form of a pixel vector, and the edge position coordinates of the retinal region of the fundus image are obtained by equation (1):

$$\begin{cases} (x_0, y_0) = Min(Coordinates(Mask)) \\ (x_1, y_1) = Max(Coordinates(Mask)) \end{cases} whereMask = image > 0 \quad (1)$$

it is cut along the minimum bounding box of Mask to obtain the

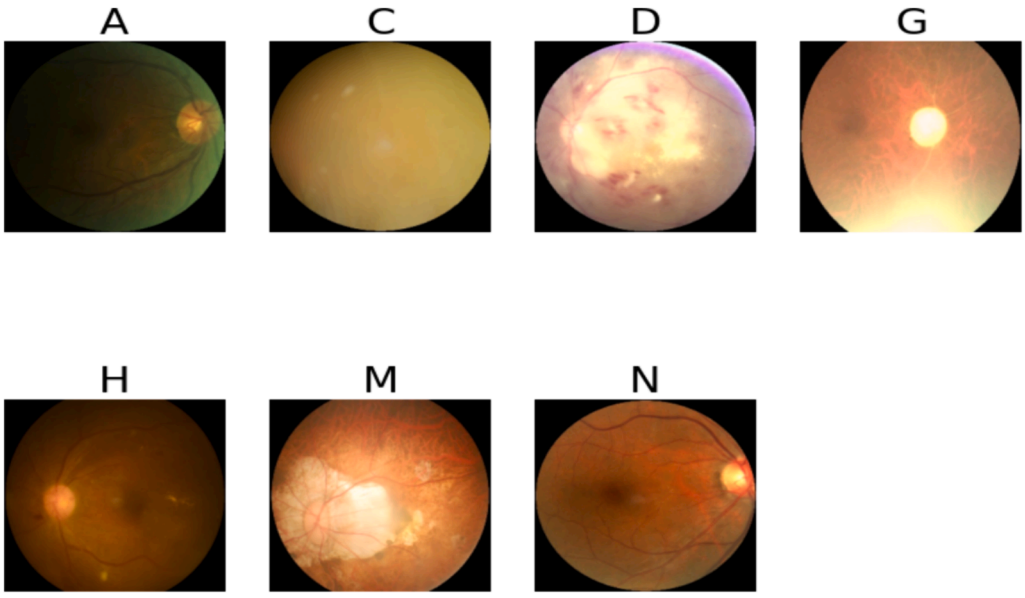


Fig. 1. Samples diagram of seven types of eye diseases in D1.

Table 1
The specific composition of D1 and D2.

Dataset	Source Address	Label							Total
		A	C	D	G	H	M	N	
D1(ODIR-5 K)	https://odir2019.grand-challenge.org/(accessed on April 30, 2023)	266	293	1608	284	128	232	2873	5684
D2(Caratact dataset)	https://www.kaggle.com/datasets/gunavenkatdoddi/eye-diseases-classification (accessed on April 30, 2023)	–	1038	1098	1007	–	–	1074	4217

Table 2
Category composition of D1 and D2 after modification.

Dataset	Label							Total
	A	C	D.	G	H	M	N	
D1(ODIR-5 K)	1064	1172	1608	1136	896	1160	2873	9909
D2(Caratact dataset)	–	1038	1098	1007	–	–	1074	4217

fundus image after removing the black area, as shown in Fig. 2. After that, the image is further resized to 224 × 224 size. Finally, normalize the original image.

3.2. Proposed method

The proposed fundus image classification model based on attention mechanism and feature fusion which is named CNN-Trans model in this paper. Its architecture is shown in Fig. 3. The model is a parallel dual-branch network.

Firstly, the patient's fundus image is cropped to remove the black area, and the size is reset to 224 × 224 × 3, and then normalized. After the above preprocessing steps are completed, it is ready to be sent to the next model.

Then, on the one hand, the preprocessed pictures enter the CNN-LSTM model branch. Firstly, a feature map with a size of 7 × 7 × 2048 is extracted through pre-trained Xception, and then the coordinated attention module decomposes channel attention into two one-dimensional feature encoding processes, and aggregates features along two spatial directions to obtain a feature map. Then, the feature maps are encoded as a pair of orientation-aware pairs and a location-sensitive attention map to augment the representation of objects of interest, these steps can effectively improve the processing efficiency of the input feature maps. After this, the feature map of the coordinated attention

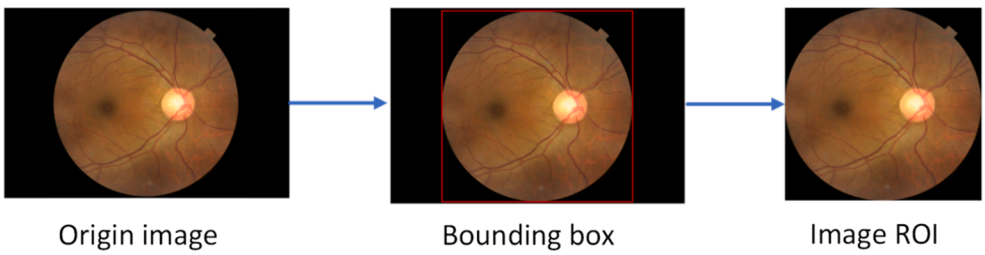


Fig. 2. process of removing black areas.

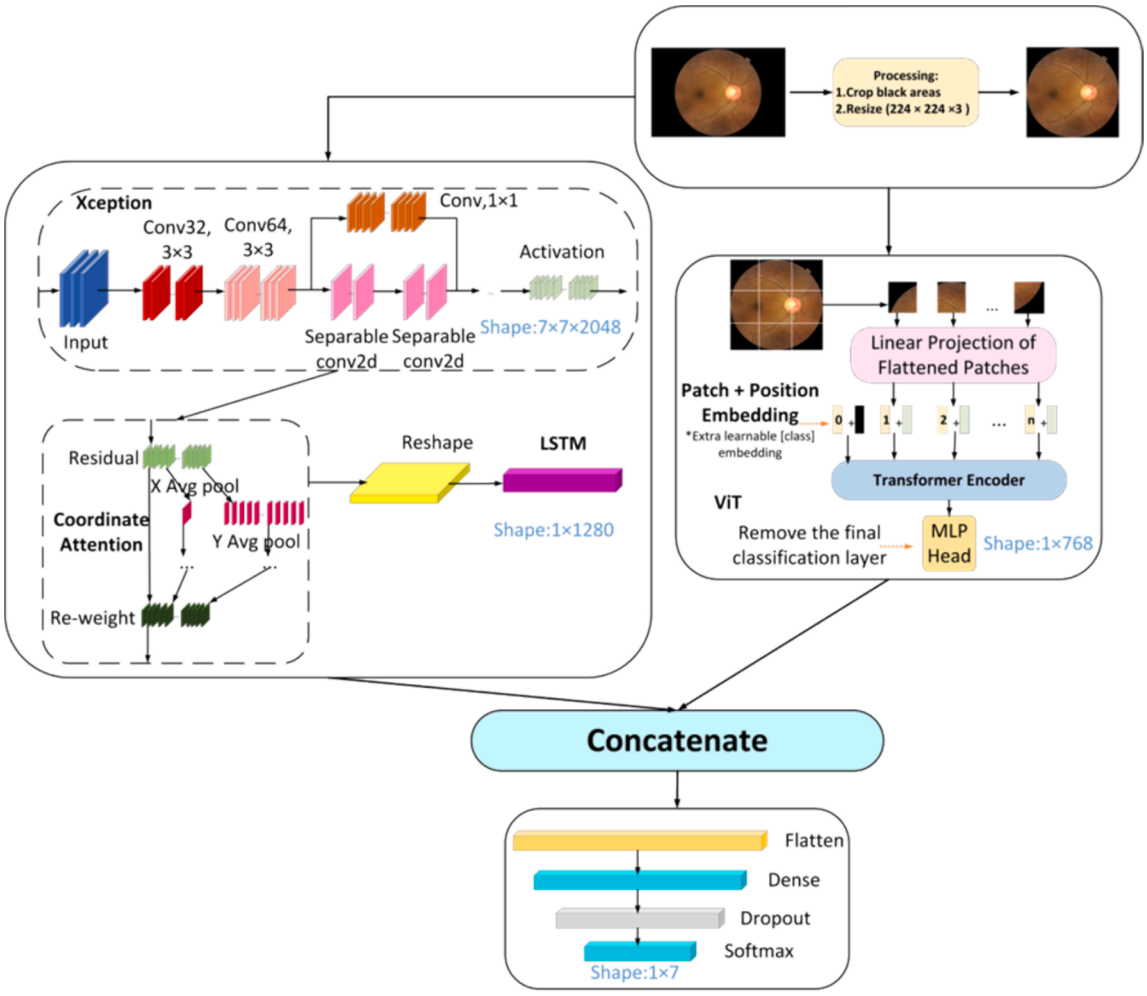


Fig. 3. The overall architecture of the CNN-Trans model.

module is input to the reshape layer, which changes the shape from (h, w, c) to (h * w, c) input to LSTM.

On the other hand, the preprocessed image enters the ViT branch. First, the input image is divided into image blocks and serialized, position encoding is added, then adds learnable embedding vectors, and inputs them into the encoder for encoding, and finally outputs Learnable embedding vectors are used for subsequent classification.

Finally, the features extracted by the Xception branch and the features extracted by the ViT branch are concat spliced and input to the reconstructed classification head, that is, a Flatten layer, a Dense layer with an L2 regularization parameter of 0.01 and 512 output units, and a Dropout layer with a dropout rate of 0.5, a Dense layer with a SoftMax function and an output unit of 7.

3.2.1. CNN-LSTM branch

The CNN-LSTM branch mainly consists of three parts: Xception, LSTM and coordinated attention mechanism, which is also part of our previous research work [18].

(1) Xception

The limited available data discourages the use of large CNN models, and very large models such as DenseNet-121 may negatively impact generalization ability [19]. Therefore, when we choose the model, we consider the lightweight and simple model as much as possible, and use the model with better initial classification effect as the feature extractor through simple experiments. In the end, we chose Xception as the original feature extractor:

Xception [20] is an improvement to Inception-v3 proposed by Google after Inception in 2016. The structure is shown in Fig. 4. Its full name is "Extreme Inception", that is, Limit Inception.

Xception uses depthwise separable convolution (Depthwise Separable Convolution) instead of conventional convolution operations, which is different from Inception. Depth-separable convolution divides the standard convolution operation into two steps: depthwise convolution and pointwise convolution. The former focuses on the spatial features within the input channel, while the latter focuses on the relationship between different channels. In this way, the number of parameters and calculations in the model can be greatly reduced, the computational efficiency of the model can be improved and the risk of overfitting can be reduced. In addition, Xception also introduces a special extreme Inception block, which replaces the standard convolution operation in the branch convolutional layer of the Inception block with a depthwise separable convolution. The role of this extreme Inception block is to further reduce the number of parameters and calculations in the model and improve the accuracy of the model.

Xception has achieved very good performance in many computer vision fields, such as image classification, target detection, semantic segmentation and other tasks. In the ImageNet image classification challenge, Xception achieved an accuracy of 0.790 and 0.945 on the Top-1 and Top-5 error rates, respectively, surpassing the most advanced ResNet model at the time.

(2) LSTM.

Long Short-Term Memory (LSTM) is a common Recurrent Neural Network (RNN) architecture. The emergence of LSTM is mainly due to the problem of gradient disappearance or explosion in traditional RNN

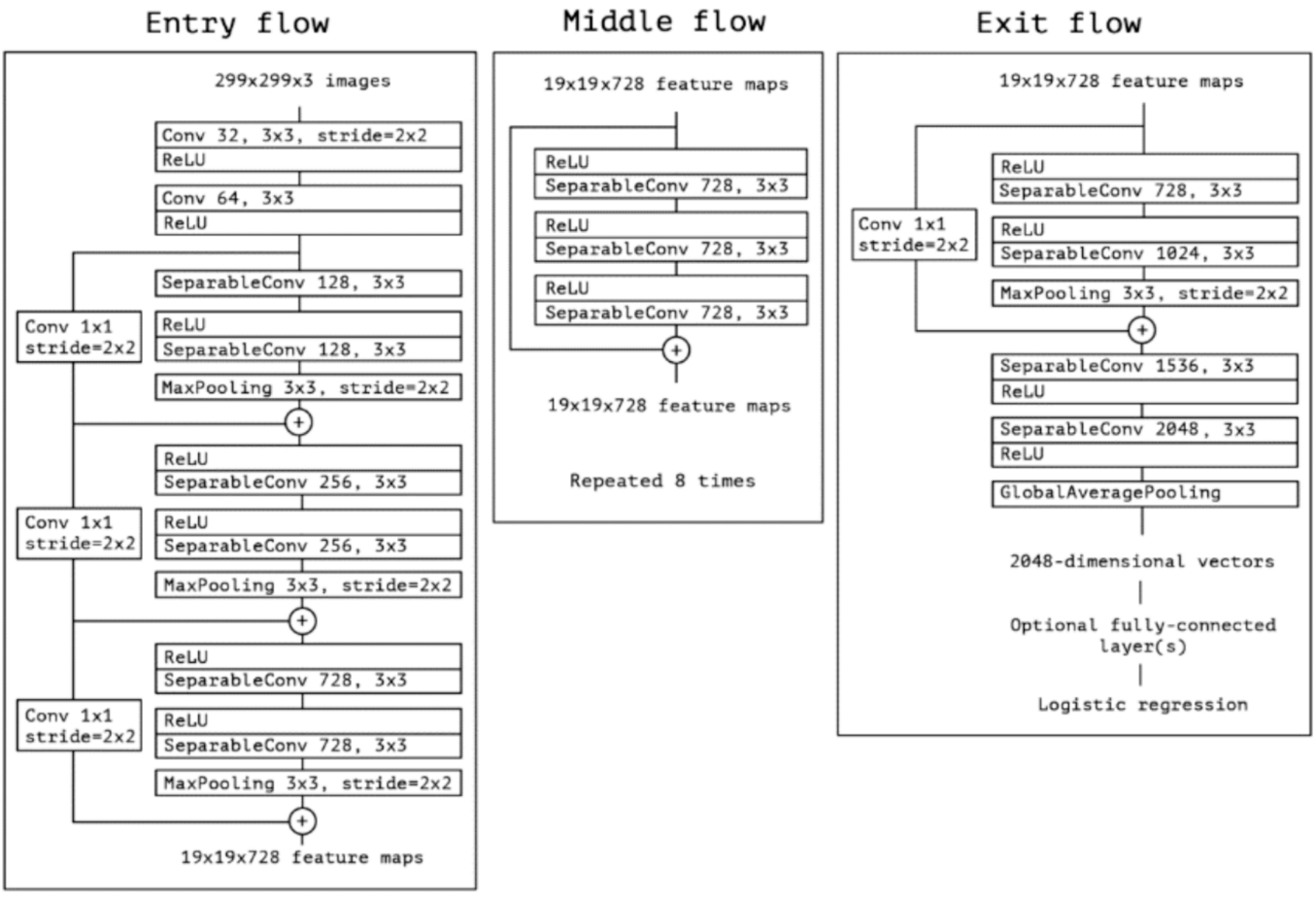


Fig. 4. Xception 's network structure.

[21], It is a RNN based on a gating mechanism. The structure of LSTM consists of a group of special neuronal units that can selectively forget, retain or add information to achieve long-term memory.

The basic units of LSTM include forget gate, input gate and output gate, which fuse the input data and the information of the previous state through nonlinear transformation, and decide whether to retain or

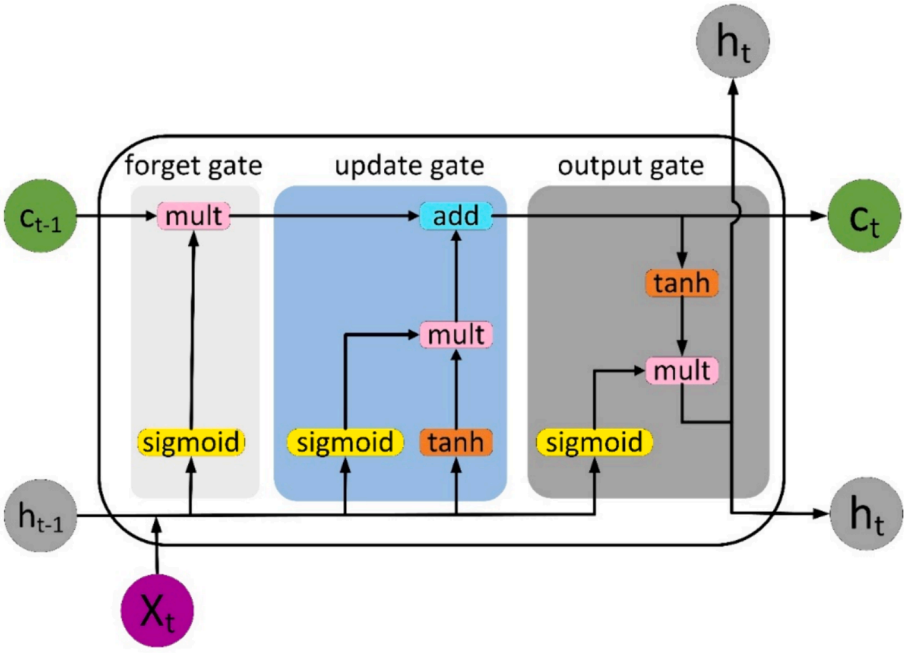


Fig. 5. The internal structure of LSTM.

forget certain information through the gating mechanism. The forget gate controls the forgetting of the past state, the input gate controls the input of new information, and the output gate controls the output of the current state. During the training process of LSTM, these gates are trained to automatically adapt to the pattern of input data and the corresponding context information. The value range of the gate is (0,1), controlled by the Sigmoid function, where X_t refers to the current input, c_t and c_{t-1} denote the new and previous cell states, respectively, h_t and h_{t-1} are the current and previous outputs, respectively. The internal structure of LSTM is shown in Fig. 5.

LSTM adds a cell state to save the long-term state, which is its main difference from RNN, which can remember the previous information and connect it to the currently obtained data. LSTM is mainly used to process sequence data, which can capture the long-term dependence in the sequence, while CNN is mainly used to process image data, which can capture the spatial local features in the image. Combining the two and making full use of their respective characteristics can effectively improve the training speed and efficiency of the model, and maintain the state information of the features encountered in the previous generation of image classification.

(3) Coordinate Attention

Coordinate Attention (CoordAtt for short) is a plug-in proposed by Qibin Hou et al. [22] inspired by the Squeeze-Excitation (SE) attention module in 2021. A ready-to-use mobile network attention mechanism, coordinated attention is different from channel attention, it uses two one-dimensional feature encoding processes to aggregate features in different spatial directions to capture long-range dependencies and precise location information at the same time, and ultimately achieve enhanced object of interest expressed purpose. Its structure is shown in Fig. 6.

Qibin Hou et al. found that compared with the most popular SE attention module in mobile networks, the coordinated attention mechanism is simple, can be flexibly inserted into the network, and has almost no computational overhead. While taking advantage of the modularity, it is also able to capture the long-term dependencies of precise position information.

3.2.2. ViT branch

Vision Transformer (ViT for short) is a method of using the Transformer model to process image data [23], proposed by Alexey Dosovitskiy et al. in 2020. The core idea of ViT is to divide the image into several small blocks, then convert these small blocks into vectors, and input these vectors into the Transformer model for processing.

The ViT architecture is shown in Fig. 7, which mainly consists of a module that splits the input image into blocks, a layer that embeds blocks and positions, a Transformer encoder, and an MLP classification head.

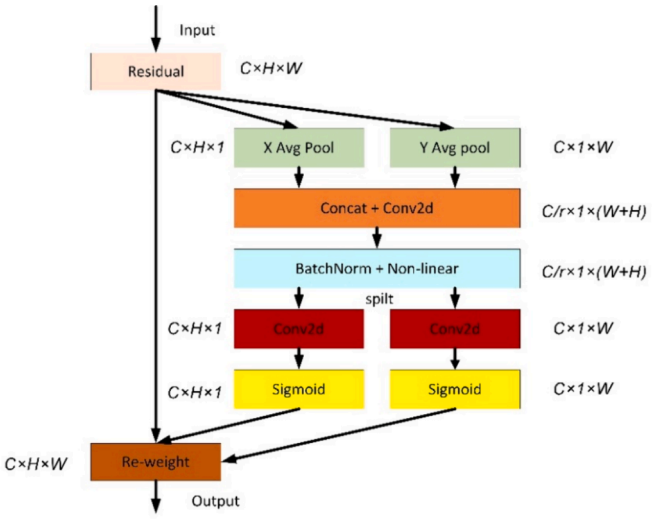


Fig. 6. The structure of coordinated attention.

Specifically, ViT first divides the image into some small blocks of the same size, and then maps each small block into a vector through a fully connected layer. These vectors are treated as tokens in the sequence and fed into the Transformer encoder for processing. In the Transformer encoder, each token is encoded through a multi-layer self-attention mechanism and a feed-forward neural network, and a context-dependent vector representation is obtained. Finally, these vectors are passed through a fully connected layer for classification or regression. In this way, ViT does not need to use traditional convolutional neural networks, and can perform tasks such as classification and detection on images using only Transformer, and has achieved good results on some standard computer vision datasets. When training on large amounts of data in the early stage, the performance of ViT is also comparable to the performance of CNN on small or medium-sized datasets [24].

3.2.3. Feature fusion

In the feature fusion of image processing, the two commonly used feature fusion (FLF) methods are concatand add. addThe method increases the information content of the image features, but does not increase the dimensionality of the image. On the other hand, concatthe method merges the number of channels to increase the number of channels describing the image, thereby keeping the relevant information of each feature unchanged. If xthe dimension of the sum of the two input features is p, ythen qthe zdimension of the output feature is p + q, as shown in Fig. 8.

Equations (2) and (3) are related mathematical expressions. In this study, since the number of channels for extracting features from the two parallel branches of ViT and Xception is different, we use concat to fuse the features extracted by the two. as the picture shows. Fusion(x_y) The number of channels is Feature(x) the sum of the number of channels of and Feature(y).

$$F_{flf} = \text{concat}(f_v, f_x) \tag{2}$$

$$F_{flf} = \text{add}(f_v, f_x) \tag{3}$$

Among them, f_v the features extracted f_x by ViT are the features extracted by F_{flf} Xception, which is the fused feature set.

Concat fusion in the CNN-Trans model is: in the branch CNN-LSTM and another branch ViT, they can both perform feature extraction on the input medical image data and output a vector representation. These two vector representations can represent different understandings of input data by CNN-LSTM and ViT respectively. In order to fuse the different understandings extracted by the two, that is, the features of medical images, and use the fused results for final classification. At this point, this article uses the concat operation to splice these two vectors along a certain dimension, and the resulting new vector will be a tensor of shape (N, c1 + c2), where N represents the number of samples, and c1 and c2 are respectively Represents the vector length of CNN-LSTM and ViT output. This new fused tensor contains the features of CNN-LSTM and ViT, is a richer vector representation, and is used for the final classification task.

It should be noted that the concat fusion method needs to normalize or scale the feature representation output by each model before splicing to avoid large deviations in the spliced results. At the same time, the concat fusion method also has some shortcomings, such as increasing the amount of calculation and storage of the model, so it needs to be selected and optimized according to specific tasks and data conditions.

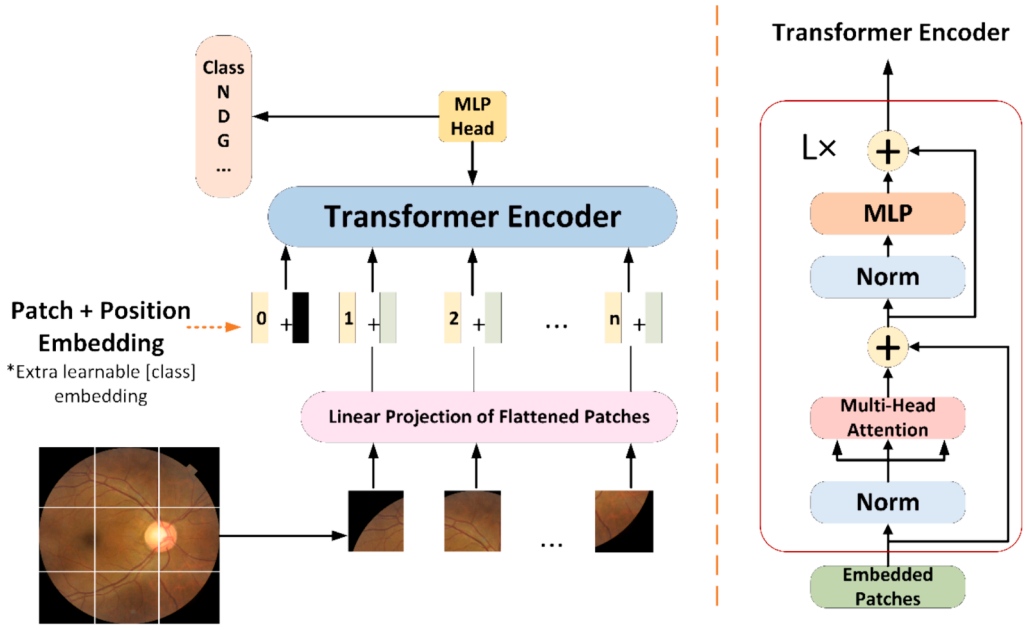


Fig. 7. The architecture of ViT.

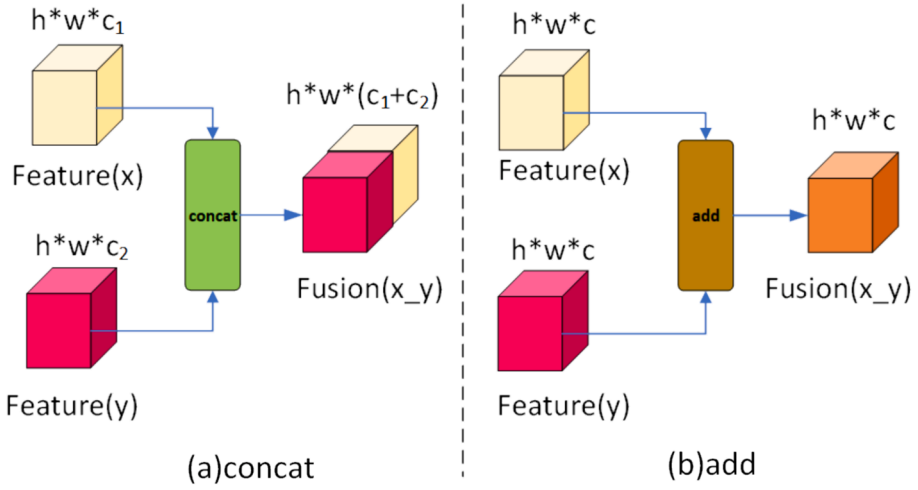


Fig. 8. Two of the most commonly used feature fusion methods.

4. Experiments

4.1. Evaluation indicators

For each image, if the predicted value of a certain label is greater than 0.5, it will be identified as a hard label in the experiment. According to the predicted label and the real label, the confusion matrix of each category can be obtained. Performance indicators based on confusion matrix calculations, such as Precision, Sensitivity (Recall), Specificity, F1, and Accuracy are still essential:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Sensitivity}(Recall) = \frac{TP}{TP + FN} \tag{5}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{6}$$

$$\text{F1} - \text{score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP} \tag{8}$$

TP, TN, FP, and FN are elements of the confusion matrix (see Table 3).

Table 3
Correlation calculation of confusion matrix.

Based on the Gold Standard	Disease Present	Disease Absent	Total
Predicted Model Positive	True positive (TP)	False positive (FP)	TP + FP
Predicted Model Negative	False negative (FN)	True negative (TN)	FN + TN
Total	TP + FN	FP + TN	TP + FP + FN + TN

In addition, this paper also added the area under the receiver operating characteristic curve (Receiver Operating Characteristic Curve, referred to as ROC) (Area Under the Curve, referred to as AUC) as a measurement index, the larger the area, the better the performance. The abscissa of the ROC curve is the false positive rate (False positive rate, FPR for short), and the ordinate is the true positive rate (True positive rate, TPR for short), as shown in Fig. 9, the drawing process is as follows:

a. Sort the prediction results according to the predicted positive class probability value;
b. the threshold from 1, and predict the samples as positive examples one by one in this order, and calculate the current FPR and TPR values each time;
c. The image is drawn with TPR as the ordinate and FPR as the abscissa.

TPR (True Positive Rate) and FPR (False Positive Rate) are calculated as follows:

$$TPR = \frac{TP}{TP + FN} \tag{9}$$

$$FPR = \frac{FP}{FP + TN} \tag{10}$$

4.2. Experimental setup

The model is implemented in Keras based on Tensorflow 2.5, and Xception is pre-trained on the ImageNet dataset. First sample 10 % of the data from the training set as the validation set, and then use the Adam optimizer on the prepared data set. For end-to-end retraining, the learning rate is 2×10^{-4}, the batch size is 32, the epoch value is 50, and the loss selects categorical_crossentropy. For training, data shuffling is enabled, which includes shuffling the data before each epoch. All experiments and training are carried out on a Linux server equipped with a 3090 graphics card, and the CPU model is AMD EPYC 7601, 16 cores, 64G.

Considering that the fundus disease recognition task is a multi-category problem with highly imbalanced categories, it is not suitable to train the model with traditional loss functions. We can use $X = x_1, x_2 \cdots x_n$, which is related x_i to the real label y_i. We wish to find a classification function $F : X \rightarrow Y$ that minimizes the loss function L, using N sets of labeled training data (x_i, y_i), where $i = 1, \cdots N$, and y_i applying a one-hot method to each encoding, each y is one of 7 labels. After studying weighted loss functions such as sample balance and category balance, it is finally decided to use the multi-classification cross-entropy function in Equation 4–3 as the loss function, and use the compute_class_weight function of the sklearn library to calculate the proportion of the loss of different categories of the data set Weight, get loss_weight ={' age_degeneration ': 1.3, ' cataract': 1.2, 'diabetes': 0.9, 'glaucoma': 1.2, ' hypertension': 1.6, 'myopia': 1.2, 'normal': 0.5} , in During model training, set the class_weight of model.fit to loss_weight to achieve the effect of weighting various losses.

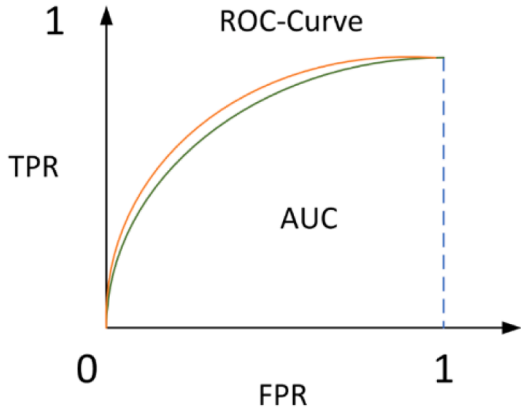


Fig. 9. ROC curve.

$$CE(x) = -\sum_{i=1}^{C} y_i \log f_i(x) \tag{11}$$

Where x is the input, C is the number of classification categories, y_i and is the true label of the i- th category.

The above hyperparameter settings are summarized in Table 4.

4.3. Ablation experiment

In order to determine the effect of each improvement, this paper conducts ablation studies, as shown in Table 5, the experiments show the results on the test set. First, the backbone network ViT is used for classification and an accuracy rate of 70.65 % is obtained. ViT achieved an accuracy rate of 74.16 % after feature fusion with Xception. Next, after adding the coordinated attention mechanism and LSTM to Xception in turn, the accuracy rate is 77.21 %. Finally, after we replaced the classifier with the new one, the accuracy rate was 80.68 %. It can be seen that every improvement has an impact on the CNN-Trans model, especially the feature fusion operation.

4.4. Seven classification results

We record the F1 score, test accuracy and AUC value of the CNN-Trans model on the test set of the dataset D1, which are obtained by averaging the results of 5-fold cross-validation. Table 6 shows the performance of the proposed architecture in each trade-off, and Fig. 10 shows the confusion matrix of the CNN-Trans model on the first fold (D1, multi-classification).

4.5. Comparison with other algorithms

In order to verify the superiority of the CNN-Trans model in this chapter, this paper conducts a large number of experiments on the CNN-Trans model on the data sets D1 and D2, and compares it with the classic network VGG16, ResNet50 and DenseNet121. It should be noted that, considering that the current mainstream research focuses on two-to-five classification tasks, and the more classification categories, the worse the classification effect, in order to facilitate subsequent comparison with other literatures, in addition to the seven-category comparison in the D1 dataset in the experiment, the D2 data set was also introduced to conduct a four-category comparison experiment.

In the seven-category task, in addition to modifying the output unit of the last classification layer to 7, in the four-classification task, the

Table 4
hyperparameter settings.

hyperparameters	value
(train:validation): test	(90 %: 10 %): 20 %
optimizer	Adam
learning rate	2×10^{-4}
batch size	32
epoch	50
shuffle	True
loss function	categorical_crossentropy
class_weight	{' age_degeneration ': 1.3, 'cataract': 1.2, 'diabetes': 0.9, 'glaucoma': 1.2, 'hypertension': 1.6, 'myopia': 1.2, 'normal': 0.5}
graphics card	3090
server	Linux
CPU	AMD EPYC 7601, 16 cores, 64G

Table 5
Ablation study, evaluation index: Accuracy (%).

Bone network	Feature fusion	CoordAtt	LSTM	New classification header	Accuracy
√	–	–	–	–	70.65
√	√	–	–	–	74.16
√	√	√	–	–	76.73
√	√	√	√	–	77.21
√	√	√	√	√	80.68

Table 6
The performance of the proposed architecture in each trade-off (D1, multi-classification).

Metric	Class	Fold 1st fold	2nd fold	3rd fold	4th fold	5th fold	Average
F1	A	85.29	83.53	84.26	84.81	84.27	84.43
	C	96.44	96.82	96.23	96.35	96.34	96.44
	D	59.10	61.24	61.10	60.53	60.16	60.43
	G	85.53	84.23	85.46	85.17	84.25	84.93
	H	87.94	86.19	85.77	86.29	86.38	86.51
	M	98.12	97.25	97.89	97.68	98.54	97.90
	N	75.63	76.88	75.86	75.94	76.25	76.11
Accuarcy		81.24	80.16	80.45	80.49	81.06	80.68
AUC		96.14	95.75	95.80	95.82	95.98	95.90

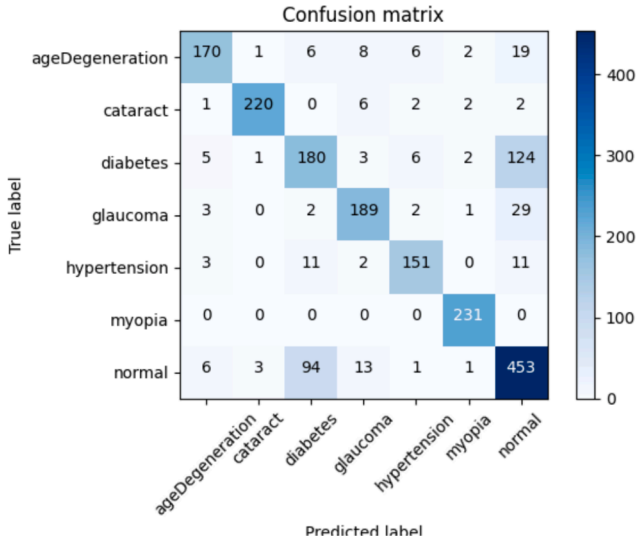


Fig. 10. Confusion matrix of CNN-Trans Net on the first fold (D3, multi-classification).

output unit of the last classification layer is modified to 4, and the activation function is softmax. Other experimental settings are the same as those in Section 4.2. same.

First, we compared the two most commonly used indicators to measure the complexity and computation of the model. Table 7 shows the model parameters and Floating Point of operations (FlOPs for short). This article uses The get_flops function of the keras_flops module counts the calculation amount of the model, and the model.summary interface counts the amount of model parameters.

Secondly, on the D1 data set, we used five times, that is, five times cross-validation method to train all models. On the D2 data set dedicated to four categories, a five-fold cross-validation method is also used. We conducted two sets of experiments: the first group (multiple classifications on D3), and the second group (four classifications on D4, taking cataract as an example).

Table 8 and Table 9 show the performance of the CNN-Trans model on different tasks. Compared with other traditional pre-trained models, the proposed model achieves a high accuracy of 80.68 % in the first set, and achieves the highest accuracy in the second set of experiments. It shows excellent performance with an accuracy as high as 94.72 %.

Finally, other research results on fundus image classification and the details of their respective datasets are shown in Table 10. It should be noted that, considering that the codes of most literatures are not public, and the experimental data sets are not public or have specific modifications, our comparison is a rough comparison method.

Table 10 show that most of the current research on fundus images focuses on two-category, four-category and five-category. The proposed CNN– TransNet has carried out seven-category experiments on fundus images, and its classification on four-category tasks The effect is also comparable to the latest other deep learning research algorithms.

5. Visualization

Visualize the features extracted by the two branches of CNN-Trans respectively.

(1) CNN-LSTM branch.

Visual analysis of neural networks is of great importance in both research and practical applications. For the model developed in research, not only can automatic classification be achieved, but also the principles behind its behavior should be familiarized so that clinicians can trust and use it without worry. Visualizations can more accurately distinguish classes, better reveal the trustworthiness of classifiers, and help identify bias in datasets.

For the CNN-LSTM branch, Fig. 11 shows the four-category results of the CNN-LSTM model (CNN-LSTM branch + new classification head) on some sample images of the test set in the dataset D2. Three images are randomly selected for each category, and a total of 12 images, where the font is green to indicate that the predicted label of the model is consistent with the real label, which is correct, and the font is red to indicate that the predicted label of the model is inconsistent with the real label, which is an error.

Gradient weighted class activation map (Grad-CAM) can be applied to various CNN models, including CNN with fully connected layers (such as VGG), for structured output (such as subtitles), for multiple Modal input (such as visual question answering) or CNNs for reinforcement learning tasks without architectural changes or retraining [32],

Table 7
Model parameters and FlOPs.

	Vgg16	Resnet50	Densenet121	CNN-Trans model
# of Parameters	15 million	24 million	7 million	85 million
FLOPs	1.97×10^6 m*	4.96×10^5 m	3.65×10^5 m	2.84×10^6 m

*m means calculations per million floating points.

Table 8
Group (1) (D1, seven categories), unit: %.

Metric	Model class	Vgg16	Resnet50	Densenet121	CNN-Trans model
F1	A	70.36	78.84	75.46	84.43
	C	90.15	93.22	94.28	96.44
	D.	59.24	53.65	54.73	60.43
	G	64.16	76.24	75.22	84.93
	h	69.85	78.18	80.16	86.51
	m	88.42	92.36	91.84	97.90
	N	64.79	63.49	65.27	76.11
Accuarcy		69.26	74.25	75.53	80.68
AUC		91.21	92.50	92.94	95.90

Table 9
Group (2) (D2, four categories), unit: %.

Metric	Model			
	Vgg16	Resnet50	Densenet121	CNN-Trans model
Caratact_Sensitivity_	78.26	83.33	89.04	97.86
Caratact_specificity_	92.52	94.27	96.18	99.20
Accuracy	84.36	86.61	89.62	94.72

therefore, Grad-CAM is widely used in CNN visualization work. The correlation analysis of Grad-CAM helps us understand the inner working principle of the CNN model. It mainly shows the important regions of the input image considered by the model through heat maps, which can verify the model and highlight the regions of interest required for classification tasks. Fig. 12 shows the different fundus image areas that CNN-LSTM focuses on for specific categories when classifying the above 12 images.

(2) ViT branch.

In recent years, Vision Transformer networks have become the main tool for traditional computer vision tasks, such as object detection [33] and image recognition[34,35]. The importance of the Vision Transformer network creates an urgent need to visualize its decision-making process. This visualization can help debug models, help verify that models are fair and unbiased, and enable downstream tasks.

However, there is still very little literature on Transformer visualization for image classification, and several factors prevent its use in visualization applications developed for other forms of neural networks compared to CNNs: this includes inactive activation functions. use, frequent use of skip connections, and the challenge of modeling multiplication used in self-attention.

the Vision Transformer network is the self-attention layer [36,37], which assigns a pairwise attention value between every two tokens. In NLP, a token is usually a word or part of a word. Visually, each token can be associated with a patch, and "self-attention" combines information from participating embeddings into the focal embedding representation of the next layer. Therefore, information from different tokens is increasingly mixed in various layers of Vision Transformer.

In order to visualize the part of the image that leads to some classification, a common practice is to rely on the attention map obtained by visualizing the Vision Transformer. This paper draws the relevant attention map through the visualize.attention_map function of the vit-keras module. For the CNN-Trans model, Fig. 13 shows the four-category results of some sample images in the test set in the dataset D2. Three images are randomly selected for each category, a total of 12 images, where the font is green to indicate the predicted label of the model Consistent with the real label is correct, and the font is red to indicate that the predicted label of the model is inconsistent with the real label, which is an error.

Fig. 14 presents some qualitative results of ViT attention-weighted activation maps. We observed that the lesion area judged by the model was basically consistent with that judged by clinicians. For example, the clinical diagnosis of cataract mainly depends on whether the fundus image is overall blurred, and the light and dark areas of the attention map are relatively average, while the diagnosis of glaucoma mainly focuses on the optic disc area.

By visualizing the features extracted by the two branches of CNN-Trans, we found that: CNN-Trans can not only extract representative local features by the coordinated attention mechanism of CNN-LSTM branch, but also by the self-attention of ViT branch. The mechanism extracts global features complementary to local information, and finally, local features are concatenated with global features and fused to generate a more comprehensive feature representation, which in turn improves the final classification accuracy.

Table 10
Comparison of CNN-Trans model with state-of-the-art systems.

Years	Author	Category	Image type	Method	Accuracy
2021	Pratik Joshi et al. [25]	2 Class: (normal: 3503, abnormal: 3341)	fundus image	EfficientNetV2	95.30 %
2021	Ali Raza et al.[26]	4Class: (normal: 900, cataract:300, glaucoma: 303 retinal diseases: 300)	fundus image	Inception v4	96.66 %
2022	Chi-Ju Lai et al. [27]	2 Class: (cataract: 4514, non-cataract: 5154)	digital camera images	CNNDCI	98.50 %
2022	A.Smitha ·P.Jidesh[28]	5Class: (normal: 2096 age_degeneration:179, diabetes: 1356, glaucoma: 180, other: 1364)	fundus image	Semi-supervised GAN	87.00 %
2023	Yuhang Pan et al. [29]	(3Class: normal:364 macular degeneration:329 Tessellated fundus:339)	fundus image	ResNet-50	93.81 %
2023	Ahlam Shamsan et al.[30]	4Class: (cataract:1038 diabetes_retinopathy:1098 glaucoma:1007 normal:1074)	fundus image	Integration of MobileNet and DenseNet121	98.5 %
2024	Md. Aiyub Ali a et al.[31]	4Class: (normal:500 diabetes:500 age-associated macular degeneration (AMD):500 cataract:500)	fundus image	AMDNet23	96.50 %
2024 (ours)	Proposed model	7Class: (age_degeneration: 1064, cataract: 1172, diabetes: 1608, glaucoma: 1136, hypertension: 896, myopia: 1160, normal: 2873)	fundus image	CNN-TransNet	80.68 %
		4Class: (cataract: 1038, diabetes: 1098, glaucoma: 1007, normal: 1074)			94.72 %

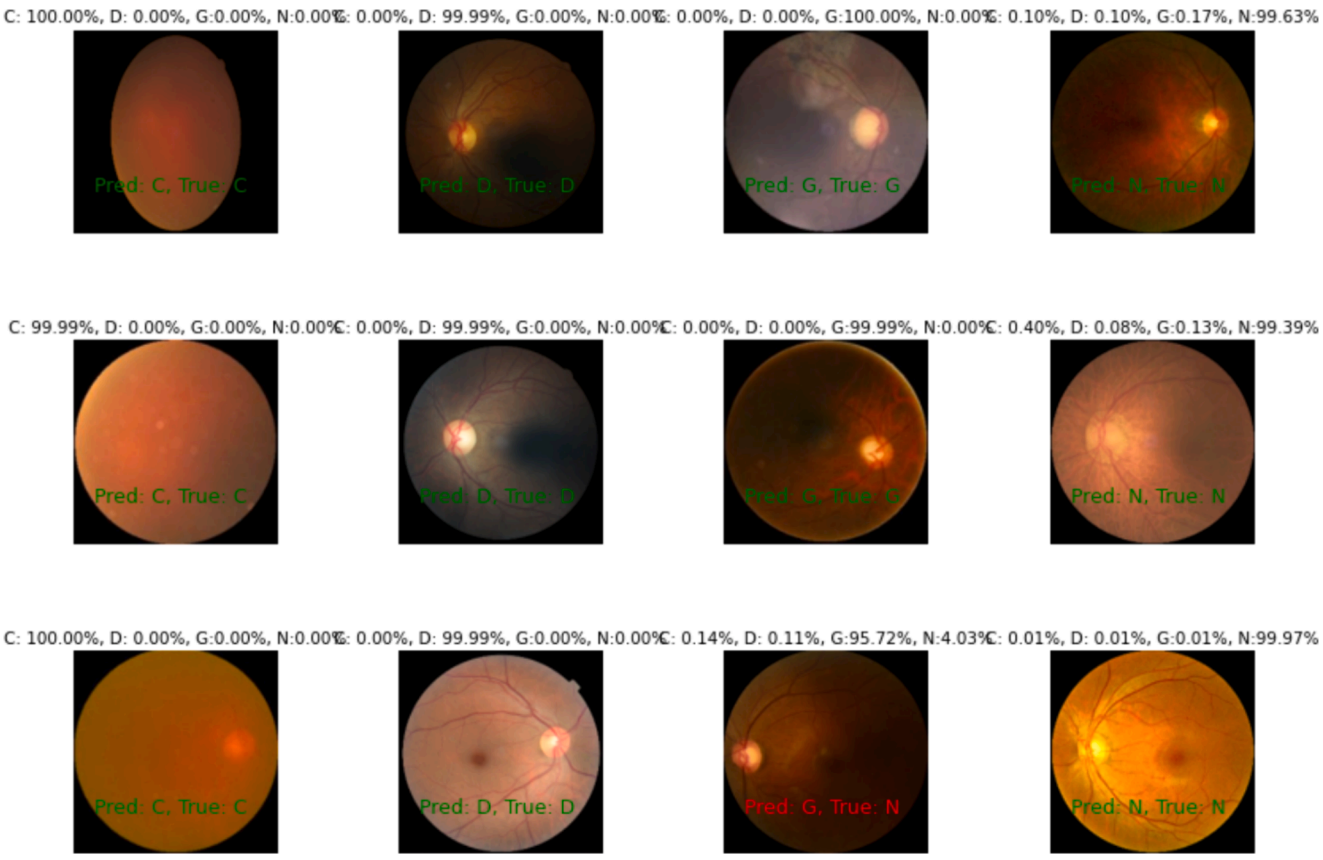


Fig. 11. CNN-LSTM four-category results of some sample images of the test set in the dataset D2.

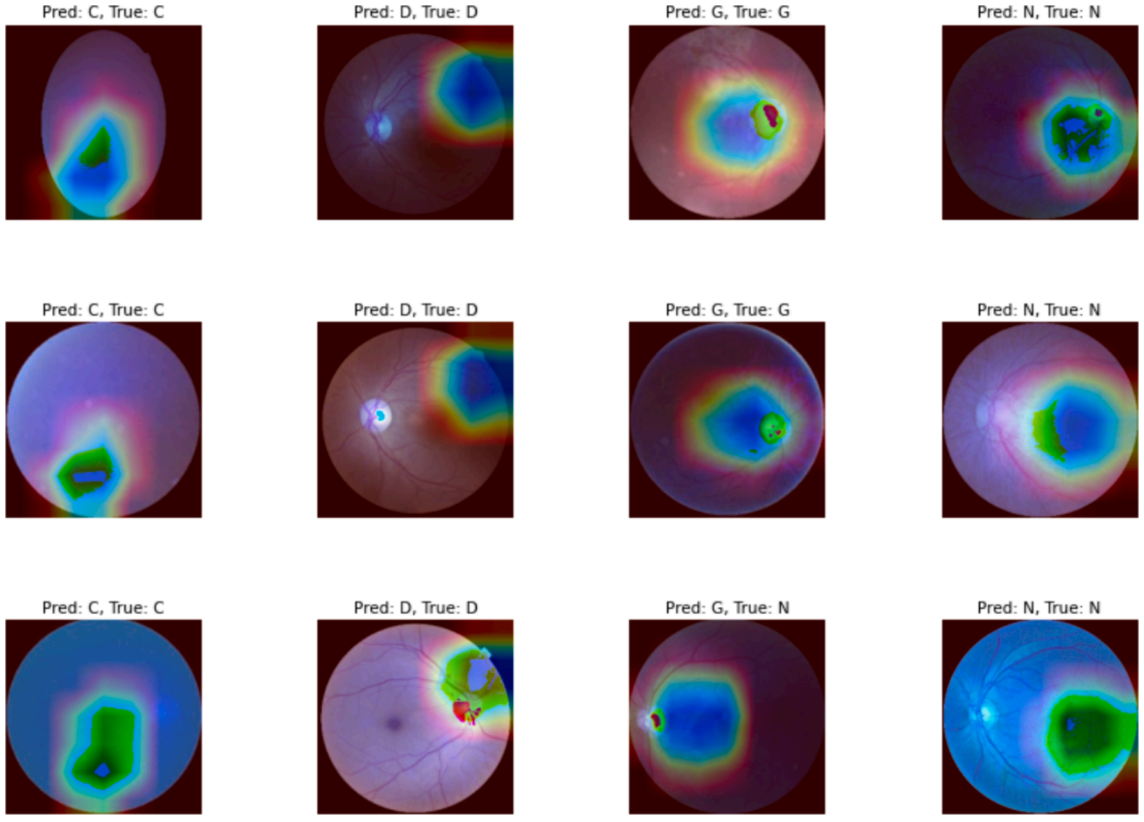


Fig. 12. Grad-CAM visualization of CNN-LSTM.

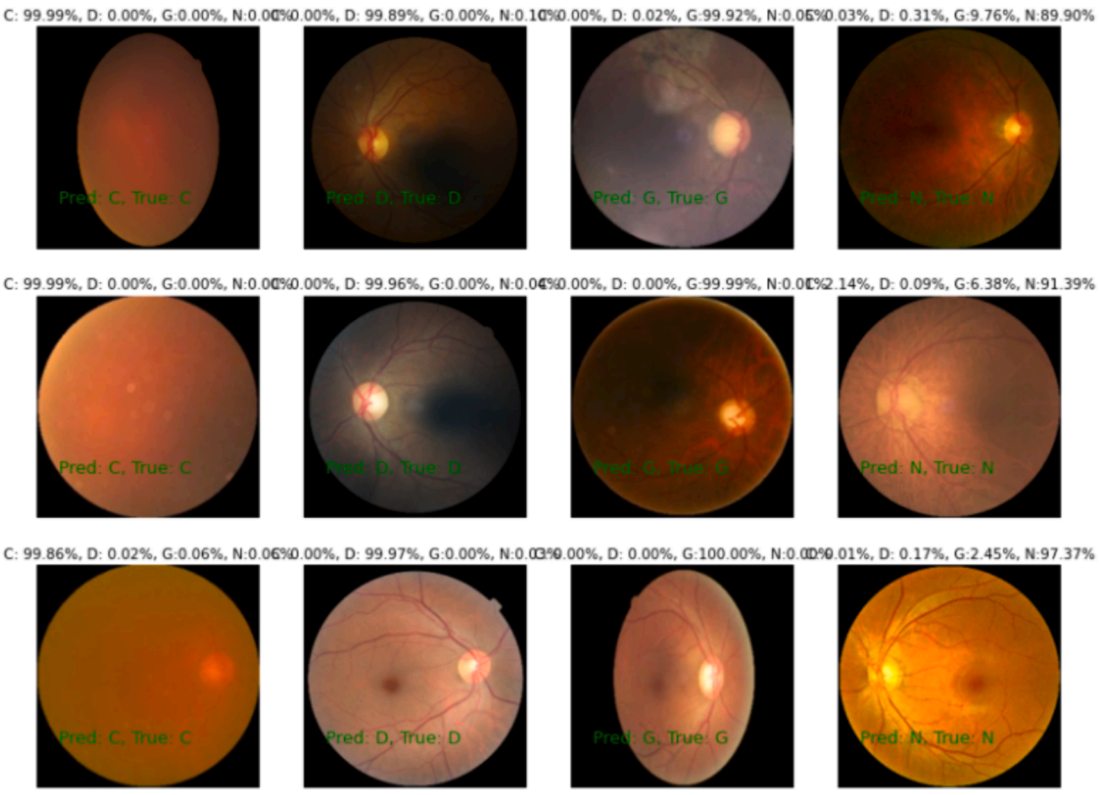


Fig. 13. ViT four-category results of some sample images of the test set in the dataset D2.

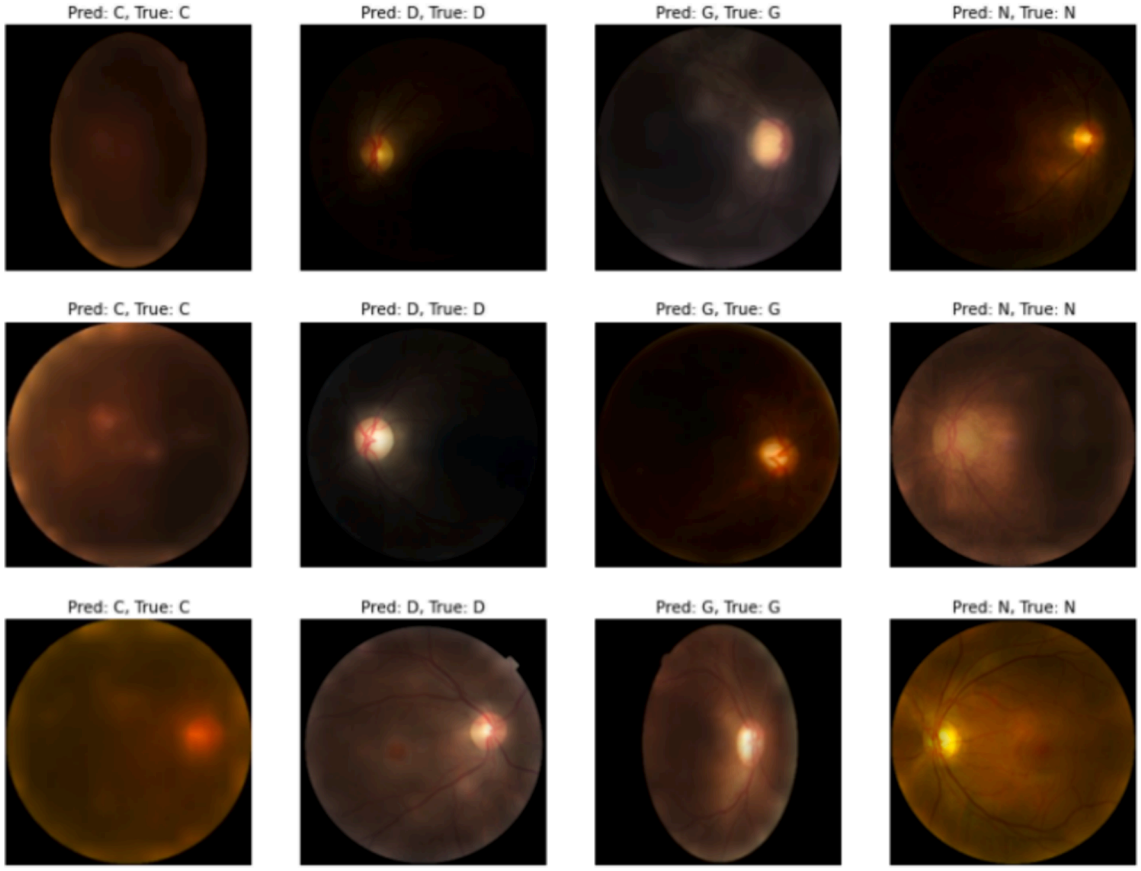


Fig. 14. ViT 's Attention Map Visualization.

6. Conclusion

This paper first introduces the background and significance of fundus image classification research, then points out why fundus image classification is more challenging than other natural image classification, and sorts out the research status of fundus image classification in recent years. Inspired by the wide application of attention mechanism and feature fusion in image classification and recognition in deep learning, this paper constructs the CNN-Trans model, and conducts a large number of related experiments on fundus images, proving the proposed model's effectiveness.

Although the algorithm in this paper can complete the fundus image classification task well, there are still some limitations. This paper has the following outlook for future work:

(1) In the experiment, we found that the addition of LSTM only slightly improved the final result. But the main purpose of using LSTM is to enhance the model image classification by maintaining the state information of the features encountered in the previous generation. This feature can significantly speed up the convergence speed of the model. This finding prompted us to keep the addition of LSTM. The next step is to find or build functionally similar but more efficient components to replace LSTMs.

(2) The proposed method misclassifies some test data samples. Therefore, incorporating uncertainty [38] into model predictions, which can be provided with error bars to allow medical staff to manually review erroneous predictions, is a challenge for current models.

(3) Although feature fusion can reduce the model's dependence on a single feature and provide diversified features, this feature will also increase the risk of overfitting. If it is not controlled, overfitting may cause the model to fail in the test data. poor performance on. Developing new feature fusion modules to prevent overfitting problems is one of the future research directions.

Funding

This research was funded by National Natural Science Foundation of China, grant number: 61762085, General Program of Natural Science Foundation of Xinjiang Uygur Autonomous Region, grant number: 2019D01C081.

CRediT authorship contribution statement

Shuxian Liu: Writing – review & editing, Writing – original draft, Validation, Resources, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Wei Wang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Le Deng:** Validation, Investigation, Formal analysis, Data curation. **Huan Xu:** Visualization, Supervision, Investigation, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the link in the main text.

References

[1] Yu. Lun, W. Lifang, P. Lin, Research progress of fundus image registration technology, J. Biomed. Eng. 28 (05) (2011) 1043–1047.
[2] Signal processing and machine learning for biomedical big data[M]. CRC press, 2018.
[3] Sengupta S, Singh A, Leopold H A, et al. Application of Deep Learning in Fundus Image Processing for Ophthalmic Diagnosis–A Review. arXiv preprint arXiv: 1812.07101, 2018.
[4] D. Socia, C.J. Brady, S.K. West, et al., Detection of trachoma using machine learning approaches, PLoS Negl. Trop. Dis. 16 (12) (2022) e0010943.
[5] P.J. Foster, Y. Jiang, Epidemiology of myopia, Eye 28 (2) (2014) 202–208.
[6] S. Faizal, C.A. Rajput, R. Tripathi, et al., Automated cataract disease detection on anterior segment eye images using adaptive thresholding and fine tuned inception-v3 model, Biomed. Signal Process. Control 82 (2023) 104550.
[7] I. Qureshi, J. Ma, Q. Abbas, Recent development on detection methods for the diagnosis of diabetic retinopathy, Symmetry 11 (6) (2019) 749.
[8] Chea N, Nam Y. Classification of fundus images based on deep learning for detecting eye diseases. 2021.
[9] J. Wen, D. Liu, Q. Wu, et al., Retinal image-based artificial intelligence in detecting and predicting kidney diseases: Current advances and future perspectives, View (2023) 20220070.
[10] J.Y. Choi, T.K. Yoo, J.G. Seo, et al., Multi-categorical deep learning neural network to classify retinal images: A pilot study employing small database, PLoS One 12 (11) (2017) e0187336.
[11] A. Diaz-Pinto, S. Morales, V. Naranjo, et al., CNNs for automatic glaucoma assessment using fundus images: an extensive validation, Biomed. Eng. Online 18 (2019) 1–19.
[12] J.J. Balaji, A. Agarwal, R. Raman, et al., Comparison of foveal avascular zone in diabetic retinopathy, high myopia, and normal fundus images[C]//Ophthalmic Technologies XXX, SPIE 11218 (2020) 86–97.
[13] M.S. Junayed, M.B. Islam, A. Sadeghzadeh, et al., CataractNet: An automated cataract detection system using deep learning for fundus images, IEEE Access 9 (2021) 128799–128808.
[14] M.M. Butt, D.N.F.A. Iskandar, S.E. Abdelhamid, et al., Diabetic Retinopathy Detection from Fundus Images of the Eye Using Hybrid Deep Learning Features, Diagnostics 12 (7) (2022) 1607.
[15] Shyamalee T, Meedeniya D. CNN based fundus images classification for glaucoma identification[C]//2022 2nd International Conference on Advanced Research in Computing (ICARC). IEEE, 2022: 200-205.
[16] A.E. Ilesanmi, T. Ilesanmi, A.G. Gbotoso, A systematic review of retinal fundus image segmentation and classification methods using convolutional neural networks, Healthcare Analytics 100261 (2023).
[17] Z. Lu, J. Miao, J. Dong, et al., Automatic Multilabel Classification of Multiple Fundus Diseases Based on Convolutional Neural Network With Squeeze-and-Excitation Attention, Transl. Vis. Sci. Technol. 12 (1) (2023) 22.
[18] W. Wang, S. Liu, H. Xu, et al., COVIDX-LwNet: A Lightweight Network Ensemble Model for the Detection of COVID-19 Based on Chest X-ray Images, Sensors 22 (21) (2022) 8578.
[19] E. Tartaglione, C.A. Barbano, C. Berzovini, et al., Unveiling covid-19 from chest x-ray with deep learning: a hurdles race with small data, Int. J. Environ. Res. Public Health 17 (18) (2020) 6933.
[20] C.F. Xception, Deep Learning with Depthwise Separable Convolutions[c]// proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017:) 1251–1258.
[21] K. Greff, R.K. Srivastava, J. Koutník, et al., LSTM: A search space odyssey, IEEE Trans. Neural Networks Learn. Syst. 28 (10) (2016) 2222–2232.
[22] Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design. arXiv 2021. arXiv preprint arXiv:2103.02907, 2021.
[23] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.
[24] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. Springer International Publishing, 2020: 213-229.
[25] Joshi P, Masilamani V. An Efficient Transfer Learning Based Approach for Detecting the Abnormal Fundus Images[C]//2021 5th Conference on Information and Communication Technology (CICT). IEEE, 2021: 1-5.
[26] Raza A, Khan M U, Saeed Z, et al. Classification of eye diseases and detection of cataract using digital fundus imaging (DFI) and inception-V4 deep learning model [C]//2021 International Conference on Frontiers of Information Technology (FIT). IEEE, 2021: 137-142.
[27] C.J. Lai, P.F. Pai, M. Marvin, et al., The Use of Convolutional Neural Networks and Digital Camera Images in Cataract Detection, Electronics 11 (6) (2022) 887.
[28] A. Smitha, P. Jidesh, Classification of multiple retinal disorders from enhanced fundus images using semi-supervised GAN, SN Computer Science 3 (2022) 1–11.
[29] Y. Pan, J. Liu, Y. Cai, et al., Fundus image classification using Inception V3 and ResNet-50 for the early diagnostics of fundus diseases, Front. Physiol. 14 (2023) 160.
[30] A. Shamsan, E.M. Senan, H.S.A. Shatnawi, Automatic Classification of Colour Fundus Images for Prediction Eye Disease Types Based on Hybrid Features, Diagnostics 13 (10) (2023) 1706.
[31] M.A. Ali, M.S. Hossain, M.K. Hossain, et al., AMDNet23: Hybrid CNN-LSTM Deep Learning Approach with Enhanced Preprocessing for Age-Related Macular Degeneration (AMD) Detection, Intelligent Systems with Applications 200334 (2024).
[32] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy 22–29 (October 2017) 618–626.

[33] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. Endto-end object detection with transformers. arXiv preprint arXiv:2005.12872, 2020.
[34] Mark Chen, Alec Radford, Rewon Child, Jeff Wu, Heewoo Jun, Prafulla Dhariwal, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In Proceedings of the 37th International Conference on Machine Learning, volume 1, 2020.
[35] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.
[36] A. Parikh, Oscar Tackstr ¨œm, Dipanjan Das, and Jakob ¨Uszkoreit. A decomposable attention model for natural language inference, in: In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 2249–2255.
[37] J. Cheng, L.i. Dong, M. Lapata, Long shortterm memory-networks for machine reading, in: In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 551–561.
[38] Y. Gal, Z. Ghahramani, Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning 20–22 (2016) 1050–1059.