# Automatic diagnosis of multiple fundus lesions based on depth graph neural network*

**JIANG Jiewei[1], GUO Liufei[1], LIU Wei[2], WU Chengchao[2], GONG Jiamin[1,2], and LI Zhongwen[3]****

*1. School of Electronic Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710071, China*

*2. School of Communication Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710071, China*

*3. Ningbo Eye Hospital, Wenzhou Medical University, Ningbo 315000, China*

(Received 1 December 2022; Revised 4 March 2023)



Fundus images are commonly used to capture changes in fundus structures and the severity of fundus lesions, and are the basis for detecting and treating ophthalmic diseases as well as other important diseases. This study proposes an automatic diagnosis method for multiple fundus lesions based on a deep graph neural network (GNN). 2 083 fundus images were collected and annotated to develop and evaluate the performance of the algorithm. First, high-level semantic features of fundus images are extracted using deep convolutional neural networks (CNNs). Then the features are input into the GNN to model the correlation between different lesions by mining and learning the correlation between lesions. Finally, the input and output features of the GNN are fused, and a multi-label classifier is used to complete the automatic diagnosis of fundus lesions. Experimental results show that the method proposed in this study can learn the correlations between lesions to improve the diagnostic performance of the algorithm, achieving better performance than the original ResNet and DenseNet models in both qualitative and quantitative evaluation.

**Document code:** A **Article ID:** 1673-1905(2023)05-0307-9

**DOI** https://doi.org/10.1007/s11801-023-2204-0

Fundus diseases such as diabetic retinopathy, age-related macular degeneration, and glaucoma have seriously threatened the visual acuity and health of human beings. Fundus lesions are the main markers for identifying fundus diseases[1]. For example, hemorrhage and exudation may appear on fundus images along with fundus diseases. Early and accurate identification of fundus lesions is the premise of grading fundus diseases and preventing their deterioration[2,3]. Clinically, by examining fundus images, ophthalmologists analyze the morphology and color of hemorrhage and exudative lesions to assess the severity of fundus diseases and the progression of related physical disorders[4,5]. However, limited by the scarcity and uneven distribution of ophthalmologists, the time-consuming, laborious and subjective manual diagnosis, and the inconspicuous early fundus lesions, it is easy to be missed diagnosis and misdiagnosis, and then leading to vision loss, amblyopia and even permanent blindness. With the development of artificial intelligence algorithms and the accumulation of medical images, the combination of deep learning algorithms and fundus images provides a new strategy for automatic diagnosis of fundus lesions, rapidly implementing large-scale screening of fundus lesions and assisting ophthalmologists to complete high-accuracy early diagnosis.

In recent years, various fundus lesion detection algorithms based on fundus images have been proposed and applied. According to the principles of the algorithms, they can be classified into two categories: traditional machine learning algorithms and deep learning. AGARWAL et al[6] used histogram equalization technology to construct an adaptive threshold to segment the optic disc and optic cup for automatic diagnosis of glaucoma, with an accuracy rate of 90%. The linear regression algorithm was applied to the detection of fundus exudates to automatically assist in the diagnosis of diabetic retina related diseases[7]. Morphological methods were employed to characterize exudates in fundus images for automatic detection of diabetic retinopathy[8,9]. JADHAV et al[10] proposed an automatic diagnosis method for fundus microaneurysms based on support vector machine. The above-mentioned traditional machine learning methods usually suffer from low accuracy rate, high false detection rate and high missed detection rate for fundus lesions. Compared with traditional machine learning, deep learning has attained remarkable performance in automatic diagnosis of keratitis[11], cataract[12,13], glaucoma[14], eye tumor[15], and diabetic retinopathy[16], and then extended

* This work has been supported by the National Natural Science Foundation of China (Nos.62276210, 82201148 and 61775180), the Natural Science Basic Research Program of Shaanxi Province (No.2022JM-380), the Natural Science Foundation of Zhejiang Province (No.LQ22H120002), the Medical Health Science and Technology Project of Zhejiang Province (Nos.2022RC069 and 2023KY1140), and the Xi'an University of Posts and Telecommunications Postgraduate Innovation and Entrepreneurship Fund Project (No.CXJJTL2021009).

** E-mail: li.zhw@qq.com

to fundus lesions detection[17-19], fundus image enhancement[20], and fundus image quality assessment[21,22]. WANG et al[23] performed multiple fundus lesions diagnosis based on convolutional neural network (CNN) ensemble learning, with accuracy and area under the curve (AUC) of 0.89 and 0.73, respectively. NEHA et al[24] employed transfer learning-based CNN to automatically diagnose multiple fundus lesions with accuracy, AUC, and $F1$ score of 0.896, 0.688, and 0.856, respectively. DING et al[25] proposed a single-group leap-frog-optimized CNN for multiple fundus lesions detection, and the accuracy of hemorrhages, microaneurysms, hard exudates, and soft exudates were 0.887, 0.872, 0.950, and 0.937, respectively. PAN et al[26] used deep learning to detect fundus lesions of non-perfusion area, microaneurysm, exudation, and laser scar with AUC values of 0.87, 0.94, 0.96, and 0.96, respectively. LI et al[27] proposed a densely connected CNN to analyze the correlation between the features of paired fundus images for fundus lesions detection.

Although the above deep learning methods have studied the automatic diagnosis of multiple fundus lesions, they still lack consideration of the correlation between lesions. Clinically, varieties lesions co-occur on fundus images, and some correlation or regularity may be latent between them. In this study, three typical fundus lesions: retinal exudation (RE), retinal hemorrhage (RH) and glaucomatous optic neuropathy (GON) were taken as examples to explore the impact of the correlation between multiple lesions on the performance of the automatic diagnosis model. Fig.1(a) shows several typical images of fundus lesions, from which it is not difficult to find that different fundus lesions co-occur on the same fundus image. Fig.1(b) summarizes the co-occurrence probability of different lesions in the 2 083 fundus images included in this study. The co-occurrence probability of RE and RH is 49.9%, and the co-occurrence probability of GON and RE (or RH) is 6.7%, indicating the degree of correlation between different lesions is different. To utilize the correlation information between different lesions, this paper proposed a deep graph neural network (GNN) for automatic diagnosis of multiple fundus lesions (DGNNDS). First, the high-level features of fundus images were extracted from CNN, then the correlation between the high-level features was established with the help of the GNN, and finally the multi-label classifier was used for multiple fundus diagnosis. Compared with the native CNNs (eg, ResNet10[28] and DenseNet121[29]), the deep GNN model achieved better performance.

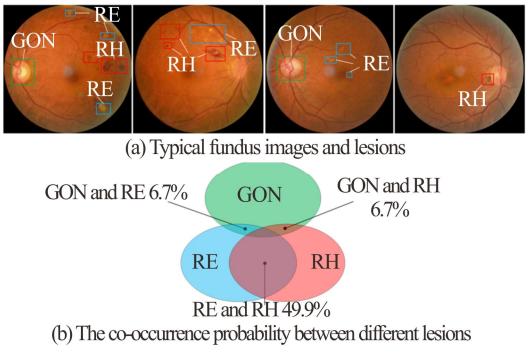



**Fig.1 Several typical fundus lesion images and the correlation between different lesion labels**

As shown in Fig.2, the DGNNDS framework primarily consists of four components: preprocessing, feature extraction, multiple lesions correlation learning, and multiple lesions classification. First, we performed denoising algorithms, data augmentation, and image standardization to filter noise, enhance the diversity of the dataset, and the generalization ability of the model. Two representative CNNs (such as ResNet101, DenseNet121) with transfer learning were employed to extract high-level features. The weights of the model trained on the ImageNet dataset were used as the pre-training weights of the feature extraction network, and the transfer learning fine-tuned the weights of the feature extraction network based on fundus images[30]. The GNN is suitable for modeling the correlation between different dataset. The features of the fundus image were extracted using the feature extraction network, and then input into a GNN to mine and learn the correlation between different lesions. Finally, considering the interaction between the deep image features and the label correlation features output by GNN, the two features are fused using multiplication fusion, and then the label classifier is used to complete the automatic diagnosis of multiple lesions.

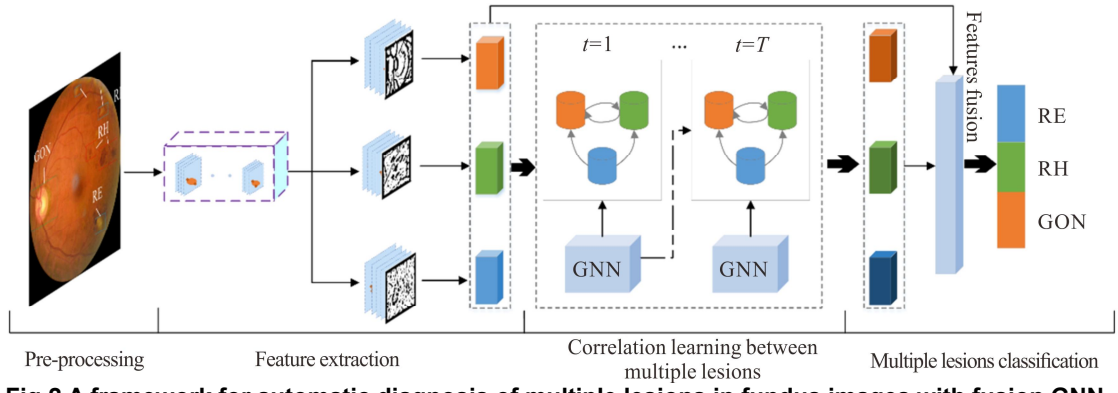



**Fig.2 A framework for automatic diagnosis of multiple lesions in fundus images with fusion GNN**

Except for fundus lesions, the redundant background is around the retina. If the entire image was directly input into the CNN, irrelevant noise would inevitably be extracted to affect the performance of the classifier. During preprocessing, Hough transformation and Canny detection were applied to segment the circle of retinal boundary, and the smallest circumscribed rectangle containing the retina was cropped to remove the surrounding noise. Image standardization technique was conducted prior to DGNNDS. The size of fundus images was resampled to a resolution of 224×224 pixels, where the value of each pixel was normalized to a range of 0—1. To increase the diversity of the Ningbo Eye Hospital (NEH) dataset and prevent overfitting and bias problems during training, data augmentation techniques, including random cropping, random rotations around the image center, and horizontal and vertical flips, were adopted to enlarge the original training dataset by 6 times. To avoid overfitting problem during the training process, not only the data augmentation and transfer learning techniques were applied, but also the batch normalization and early stopping strategies were used in this study to stabilize the model.

To obtain the optimal deep learning for identifying multiple fundus lesions, two state-of-the-art multi-label classification CNNs architectures (DenseNet121 and ResNet101) were investigated for comparison with our proposed DGNNDS framework. In this study, different CNNs can be selected as feature extraction module. For a fair comparison with the multi-label classification architectures, we also choose the representative DenseNet121 and ResNet101 CNNs for feature extraction in the DGNNDS framework. After combining GNN module, two DGNNDS models can be formed: DenseNet_GNN and ResNet_GNN. In this study, we quantitatively and qualitatively compare the performance of these four methods (DenseNet121, ResNet101, DenseNet_GNN, and ResNet_GNN) in detail to determine the optimal model.

The GNN is commonly employed to model graph-structured dataset, which provides a feasible solution for graph-structured prediction tasks based on nodes and edges. The feature vectors of different lesion categories extracted by the CNN are correlated using the co-occurrence probability matrix of the GNN in DGNNDS framework to explore and learn the interaction between different lesions. For a clear explanation, a graph structure $G=\{V, A\}$ is introduced here, in which nodes $V$ and edges $A$ denote the categories and the co-occurrence probability between corresponding categories, respectively[31]. Specifically, suppose that the dataset includes $C$ categories, $V$ can be represented as $\{v_0, v_1,\ldots,v_i,\ldots,v_{c-1}\}$ with element $v_i$ denoting the category $i$ and $A$ can be represented as $\{a_{00}, a_{01},\ldots, a_{ij},\ldots, a_{(c-1)(c-1)}\}$, with element $a_{ij}$ denoting the probability of the existence of object belonging to category $j$ in the presence of object belonging to category $i$. In this study, the GNN adopts a gated recurrent update mechanism to propagate message and learns contextualized node-level features. Specifically, for each category node $v_i \in V$, it has a hidden state $\boldsymbol{h}_i^t$ at timestep $t$. The initial hidden state $\boldsymbol{h}_i^0$ at $t$=0 is set as the feature vector extracted from the CNN, which can be formalized as $\boldsymbol{h}_i^0 = \boldsymbol{f}_i$. Therefore, the GNN can be employed to explore the interactions among the semantic specific features and aggregate message from its neighbor nodes at timestep $t$. The aggregated feature vector $\boldsymbol{a}_i^t$ is formally expressed as

$$\boldsymbol{a}_i^t = \left[ \sum_j \left(a_{ij}\right) \boldsymbol{h}_i^{t-1}, \sum_j \left(a_{ji}\right) \boldsymbol{h}_i^{t-1} \right]. \tag{1}$$

In this way, message propagation is promoted if node $i$ has a high correlation with node $j$, otherwise, it is suppressed. Therefore, it can propagate message through the graph and explore node interactions under the guidance of the prior knowledge of statistical label co-occurrence. Then, the hidden state $\boldsymbol{h}_i^t$ is updated based on the aggregated feature vector $\boldsymbol{a}_i^t$ and its hidden state at previous timestep $\boldsymbol{h}_i^{t-1}$ via a gated mechanism similar to the gated recurrent unit[32], formulated as

$$\begin{cases} \boldsymbol{z}_c^t = \sigma\left(\boldsymbol{W}^z \boldsymbol{a}_c^t + \boldsymbol{U}^z \boldsymbol{h}_c^{t-1}\right) \\ \boldsymbol{r}_c^t = \sigma\left(\boldsymbol{W}^r \boldsymbol{a}_c^t + \boldsymbol{U}^r \boldsymbol{h}_c^{t-1}\right) \\ \tilde{\boldsymbol{h}}_c^t = \tanh\left(\boldsymbol{W}\boldsymbol{a}_c^t + \boldsymbol{U}\left(\boldsymbol{r}_c^t \odot \boldsymbol{h}_c^{t-1}\right)\right) \\ \boldsymbol{h}_c^t = \left(1 - \boldsymbol{z}_c^t\right) \odot \boldsymbol{h}_c^{t-1} + \boldsymbol{z}_c^t \odot \tilde{\boldsymbol{h}}_c^t \end{cases}, \tag{2}$$

where $\sigma()$, tanh() and $\odot$ are the logistic sigmoid function, the hyperbolic tangent function, and the element-wise multiplication operation, respectively. $\boldsymbol{W}$ and $\boldsymbol{U}$ are the learnable weight matrices during the training process. $\boldsymbol{z}_c^t$ and $\boldsymbol{r}_c^t$ denote the controllable forgetting message and update message, respectively. $\tilde{\boldsymbol{h}}_c^t$ represents the newly generated message. $\boldsymbol{h}_c^t$ is the final updated message, in which $\left(1-\boldsymbol{z}_c^t\right) \odot \boldsymbol{h}_c^{t-1}$ and $\boldsymbol{z}_c^t \odot \tilde{\boldsymbol{h}}_c^t$ represent selectively forgetting past message and remembering newly generated message, respectively. Therefore, each node can aggregate message from other nodes and simultaneously transfer the message through the graph, enabling interactions among all feature vectors of all categories. This process is repeated $T$ times, and the final hidden states are updated, i.e., $\{\boldsymbol{h}_0^{\mathrm{T}}, \boldsymbol{h}_1^{\mathrm{T}}, ..., \boldsymbol{h}_{c-1}^{\mathrm{T}}\}$. Here, the hidden state of each node $\boldsymbol{h}_i^{\mathrm{T}}$ not only encodes features of category $i$, but also carries contextualized message from other categories. Finally, $\boldsymbol{h}_i^{\mathrm{T}}$ and the input feature vector $\boldsymbol{h}_i^0$ were concatenated to predict the confidence score of the presence of category $i$. It can be formally expressed as Eq.(3), where $f_{\mathrm{o}}(\cdot)$ is an output function that maps the concatenation of $\boldsymbol{h}_i^{\mathrm{T}}$ and $\boldsymbol{h}_i^0$ into an output vector, and the $f_i(\cdot)$ is multi-label classification function that predicts a score indicating the probability of category $i$. In the same way, all categories were performed to obtain a score vector $\boldsymbol{s}=\{s_0, s_1,\ldots, s_{c-1}\}$.

$$s_i = f_i\left(f_{\mathrm{o}}\left(\boldsymbol{h}_i^{\mathrm{T}}, \boldsymbol{h}_i^0\right)\right). \tag{3}$$

To verify the reasonability of the DGNNDS framework,

the gradient-weighted class activation mapping (Grad-CAM) visualization technique was adopted to generate the heatmaps for highlighting the multiple lesion-related regions on which the diagnosis model focused most. The Grad-CAM is an explainable method for CNN-based models, which utilized the gradients of any target concept flowing into last convolutional layer to produce a localization map highlighting remarkable regions in an image for predicting the concept.

In this study, all models were trained using Pytorch (Torch1.7.1 and Torchvision0.8.2) based on four Nvidia Titan RTX graphics processing units (24G). The adaptive moment estimation (Adam) optimizer with initial learning rate of 0.001, $\beta_1$ of 0.9, $\beta_2$ of 0.999, and weight decay of $10^{-4}$ was adopted. The batch-size was set 128 for one iteration training and calculated the average value of these images to update the trainable parameters. Each model was trained for 80 epochs. During the training process, validation loss was calculated on the validation dataset in each epoch and used as a reference for model selection. If the validation loss decreased in one epoch, the model state and corresponding weight parameters were saved. The model state with the lowest validation loss was saved as the final state of the model for use on the test dataset.

To evaluate the performance of the deep learning models for the automatic diagnosis of RE, RH and GON, we calculated quantitative evaluation indicators with 95% confidence interval (*CI*) using the Wilson score approach, including the average per-class precision (*CP*), average per-class recall (*CR*), average *F*1 score per-class (*CF*1), average overall precision (*OP*), average overall recall (*OR*), average overall *F*1 score (*OF*1), and the mean average precision (*mAP*), as shown in Eq.(4) to Eq.(9). The *F*1 score and the receiver operating characteristic curve (ROC) comprehensive evaluation indicators used to compare the performance of the diagnostic systems. The *F*1 score was calculated using *OP* and *OR* indicators, and the ROC curve was plotted using the ratio of true positive cases (sensitivity) and the ratio of false-positive cases (1-specificity). The area under the receiver operating characteristic curve was calculated using Empirical Bootstrap with 1 000 replicates. A larger AUC value indicated better performance.

$$OP = \frac{\sum_i TP^i}{\sum_i \left(TP^i + FP^i\right)}, i \in [1,\ldots,N], \tag{4}$$

$$OR = \frac{\sum_i TP^i}{\sum_i \left(TP^i + FN^i\right)}, i \in [1,\ldots,N], \tag{5}$$

$$OF1 = 2 * \frac{OP * OR}{OP + OR}, \tag{6}$$

$$CP = \sum_i \frac{TP^i}{TP^i + FP^i} / N, i \in [1,\ldots,N], \tag{7}$$

$$CR = \sum_i \frac{TP^i}{TP^i + FN^i} / N, i \in [1,\ldots,N], \tag{8}$$

$$CF1 = 2 * \frac{CP * CR}{CP + CR}, \tag{9}$$

where $N$ is the number of lesion categories, $N$=3 in this study, including RE, RH and GON. $TP^i$, $FP^i$, $TN^i$, and $FN^i$ denote the numbers of true positives, false positives, true negatives, and false negatives for lesion category $i$, respectively.

In this study, a total of 2 083 fundus images consecutively collected from NEH between May 2018 and October 2021 were leveraged to develop and evaluate the DGNNDS. Three ophthalmologist specialists, each with over five years of clinical experience, separately annotated all fundus lesions. Annotated labels include RE, RH and GON, and each image may contain one or more lesion labels. The dataset was randomly divided into three independent datasets: 70% in a training dataset (1 459 images), 15% in a validation dataset (312 images), and the remaining 15% in a test dataset (312 images). The detailed information on the distribution of lesions in the training set, validation set, and test set is described in Tab.1. According to the occurrence times of different lesions in the statistical dataset, the co-occurrence probability between lesions can be constructed to form a co-occurrence probability matrix, as shown in Tab.2. The co-occurrence probability was used to measure the statistical correlation between different lesions, from which it could be concluded that retinal exudation and hemorrhage had a strong correlation with a co-occurrence probability of 0.499. The co-occurrence probability matrix, as the prior knowledge of the GNN, was employed to guide the parameter learning of the multiple lesions diagnosis model during the training process.

**Tab.1 Distributions of fundus images**

| Class | Training | Val | Test | Total |
|---|---|---|---|---|
| RE | 600 | 124 | 126 | 850 |
| RH | 587 | 142 | 130 | 859 |
| GON | 628 | 128 | 136 | 892 |

**Tab.2 Co-occurrence probability matrix between different lesions in fundus images**

| Class | RE | RH | GON |
|---|---|---|---|
| RE | 0 | 0.499 | 0.067 |
| RH | 0.499 | 0 | 0.067 |
| GON | 0.067 | 0.067 | 0 |

In this study, ResNet101 and DenseNet121 were chosen as baseline models to verify the indispensability of the GNN module in the automatic diagnosis of multiple lesions. After utilizing the GNN module, two feature processing strategies were proposed and compared, as shown in Tab.3. The first strategy only utilized the output feature of GNN for the multi-label classifier. The second strategy combined the output features of CNN

and GNN for the multi-label classifier, in which the output feature of CNN is also the input feature of GNN. According to the different fusion methods of CNN and GNN features, the second strategy can be divided into three categories: additive fusion, concatenation fusion, and multiplication fusion[33]. Detailed ablation experiments were conducted in this study to compare the performance of two feature processing strategies and baseline CNNs to obtain the optimal model. Tab.3 showed the test results of different models with the 95% *CI*. From ablation experiments, two meaningful conclusions were obtained. First, the performance of two feature processing strategies using the GNN module was superior to those of the two baseline models. Second, the multiplication fusion of the second strategy using output features of CNN and GNN outperformed the first strategy using the only feature of GNN. Consistent conclusions were obtained on ResNet101 and Dense-Net121. Specifically, compared with the ResNet101, the *CP*, *CR*, *CF*1, *OP*, *OR*, and *OF*1 of ResNet_GNN with multiplication fusion were improved by −0.78%, 8.48%, 3.78%, −0.78%, 8.48%, and 3.78%, respectively. Compared with the DenseNet121, the *CP*, *CR*, *CF*1, *OP*, *OR*, and *OF*1 of DenseNet_GNN with multiplication fusion were improved by 0.89%, 3.90%, 2.41%, 1.00%, 4.01%, and 2.65%, respectively. The above ablation experiments and comparative analysis verified that the second strategy with the output features of CNN and GNN achieved the optimal performance in multi-lesion diagnosis. This is probably due to the fact that the correlation of different lesions was learned by the GNN module. This learned correlation is an effective complement to the feature of CNN, which is beneficial to improve the performance of automatic diagnosis models.

**Tab.3 Performance comparison of two features processing strategies and baseline models on the diagnosis of multiple lesions**

| Metric | DenseNet121 | ResNet101 | Only output feature of GNN | | Multiplication fusion of output features of CNN and GNN | |
|---|---|---|---|---|---|---|
| | | | DenseNet_GNN | ResNet_GNN | DenseNet_GNN | ResNet_GNN |
| *CP* (95% *CI*) | 0.895 (0.861—0.929) | 0.898 (0.864—0.931) | 0.873 (0.836—0.910) | 0.896 (0.863—0.930) | 0.903 (0.870—0.936) | 0.891 (0.857—0.926) |
| *CR* (95% *CI*) | 0.847 (0.807—0.887) | 0.849 (0.810—0.889) | 0.880 (0.844—0.916) | 0.868 (0.831—0.906) | 0.880 (0.844—0.916) | 0.921 (0.891—0.951) |
| *CF*1 (95% *CI*) | 0.870 (0.833—0.908) | 0.873 (0.836—0.910) | 0.877 (0.840—0.913) | 0.882 (0.847—0.918) | 0.891 (0.857—0.926) | 0.906 (0.873—0.938) |
| *OP* (95% *CI*) | 0.894 (0.833—0.928) | 0.897 (0.836—0.931) | 0.873 (0.837—0.910) | 0.897 (0.864—0.931) | 0.903 (0.870—0.936) | 0.890 (0.873—0.925) |
| *OR* (95% *CI*) | 0.846 (0.806—0.886) | 0.849 (0.809—0.889) | 0.880 (0.844—0.916) | 0.870 (0.833—0.907) | 0.880 (0.844—0.916) | 0.920 (0.890—0.950) |
| *OF*1 (95% *CI*) | 0.869 (0.832—0.907) | 0.872 (0.835—0.909) | 0.877 (0.840—0.913) | 0.883 (0.848—0.919) | 0.892 (0.857—0.926) | 0.905 (0.872—0.938) |

Furthermore, to confirm the optimal features fusion method in the second strategy for the diagnosis of multiple lesions, three feature fusion methods, including additive fusion, concatenation fusion, and multi-plication fusion, were investigated and compared. Taking the ResNet_GNN model as an example, Fig.3 showed the performance difference of different feature fusion methods. It is not difficult to obtain the following two conclusions: the performance of additive fusion and concatenation fusion was comparable, and the multiplication fusion method was superior to the other two methods. This indicated that the multiplication fusion method was more suitable for the fusion of features of CNN and GNN.

Tab.4 summarizes the average precision (*AP*) of different models on the three lesion labels. Comparative analysis of the two different structures of the CNN after the fusion of the GNN has improved the average accuracy of various labels. The *AP* of DenseNet121 and DenseNet_GNN in RE, RH, and GON lesions increased by 1.66%, 0.84%, and 1.47%, respectively. The *AP* of ResNet101 and ResNet_GNN in RE, RH, and GON lesions increased by 2.72%, 2.76%, and 0.42%, respectively. The correlation between RE and RH obtained from the prior knowledge of co-occurrence probability is relatively high, and the accuracy of the two labels of RE and RH has been greatly improved in this experiment. This further indicates that the GNN has learned the strong correlation property between these two types of lesions during the training process.

In order to more clearly compare the detection performance of different models for the three lesions, the *AP* of the three lesions in each batch size of the four models in the test dataset were counted, and the data were visualized using boxplots, as shown in Fig.4. The detection performance of DenseNet121 for GON is the worst, ResNet101 is slightly better than DenseNet121, and ResNet_GNN and DenseNet_GNN have better performance. DenseNet121 performs the worst for RE detection and ResNet_GNN performs the best. ResNet101 performed the worst for RH detection, and ResNet_GNN performed the best.

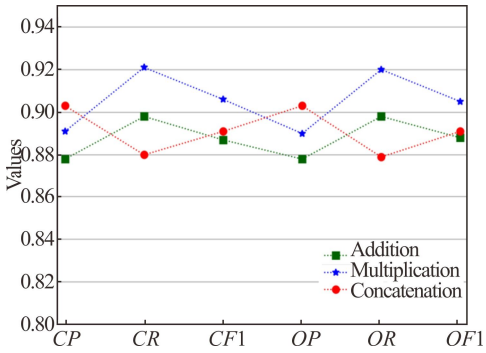



**Fig.3 Performance comparison of different fusion methods on the ResNet_GNN model**

**Tab.4 Comparison of *AP* values of different deep learning models on multiple lesions in the fundus**

| Model | RE *AP* | RH *AP* | GON *AP* | *mAP* |
|---|---|---|---|---|
| DenseNet121 | 0.899 | 0.955 | 0.955 | 0.937 |
| ResNet101 | 0.920 | 0.943 | 0.959 | 0.941 |
| DenseNet_GNN | 0.914 | 0.963 | 0.969 | 0.949 |
| ResNet_GNN | 0.945 | 0.969 | 0.963 | 0.959 |

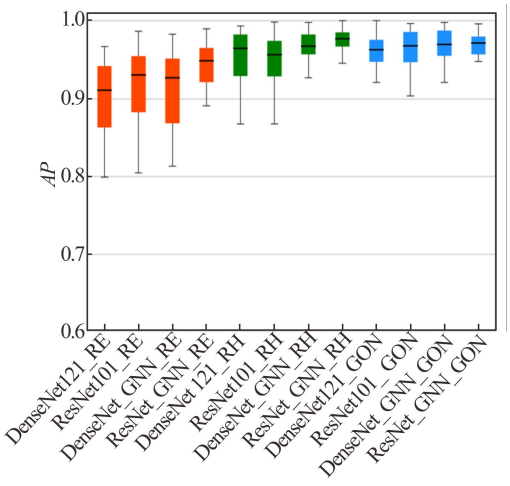



**Fig.4 *AP* boxplot of different deep learning models on a batch size of the test dataset**

To investigate the impact of the depth of the GNN on model performance, we calculated the variation trend of ResNet_GNN performance as the number of layers of the GNN increased, as shown in Fig.5. It can be seen that with the continuous increase of the number of layers of the GNN, the performance of the model first increases and then decreases, and when the number of GNN layers reaches 3, the model obtains the best performance. When the number of network layers is greater than or equal to 3, the performance of the model shows a declining trend. The reason may be that when the number of GNN layers is deep, a large number of parameters lead to the disappearance of the backpropagation gradient and overfitting, and the *AP* does not increase but decreases.

Furthermore, the ROC curves and AUC values of different models on the three lesions were compared to comprehensively analyze the detection ability of the models for different fundus lesions, and the comparison chart shown in Fig.6 was obtained. Fig.6(a) shows the ROC curve and AUC values of DenseNet121 and DenseNet_GNN on multiple fundus lesions. The AUC values of the DenseNet_GNN model in RE, RH, and GON are 0.956, 0.978, and 0.971, respectively. Compared with DenseNet121, the AUC values of DenseNet_GNN in the three lesions were increased by 2.4%, 0.3%, and 1.7%, respectively. Fig.6(b) shows the ROC curve and AUC values of ResNet101 and ResNet_GNN on multiple fundus lesions. The AUC values of the ResNet_GNN model in RE, RH, and GON are 0.963, 0.986, and 0.975, respectively. Compared with ResNet101, the AUC values of ResNet_GNN in the three lesions increased by 1.8%, 2.0%, and 2.1%, respectively. The ROC curve can intuitively be used to compare the detection performance of different models for different lesions. The closer the ROC curve is to the upper left corner, the higher the accuracy of the model is. Therefore, it can be seen from the figure that the ROC curve of the fusion GNN model is better than the original model's ROC curve.

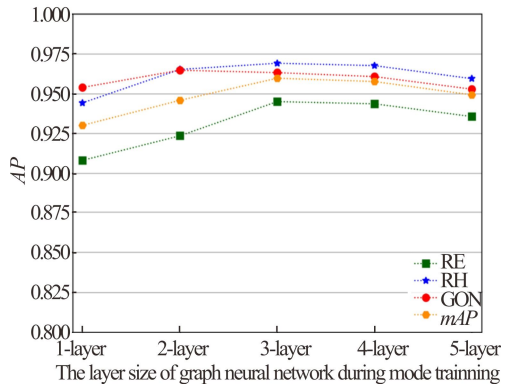
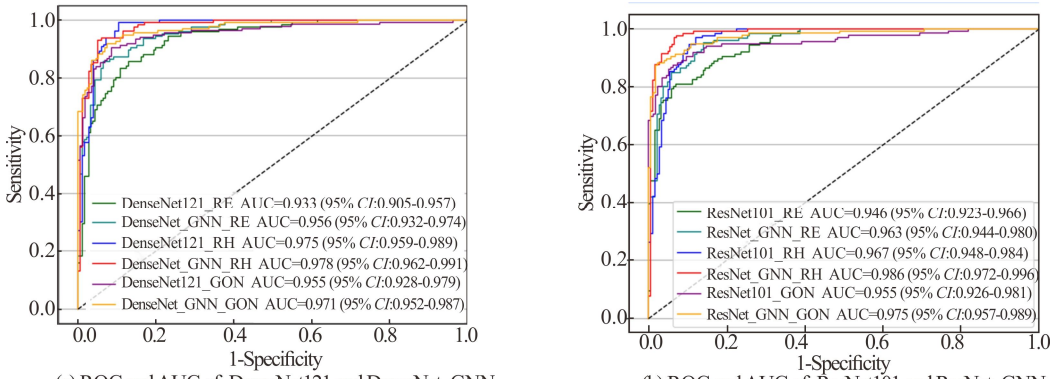



**Fig.5 Performance trend of the ResNet_GNN with the number of layers of the GNN**

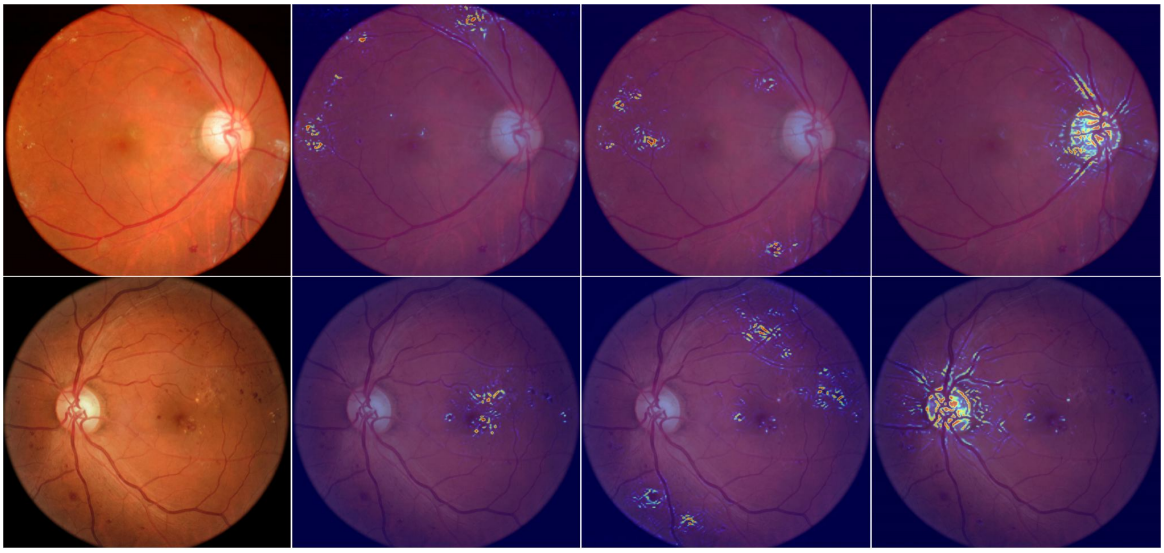
To visualize the areas contributing most to the DGNNDS, we generated a heatmap that superimposed a visualization layer at the end of the CNN. For images of RE and RH lesions, the heatmap can accurately locate the corresponding lesion area. For images of GON, the heatmap can effectively highlight the optic disc area. Typical examples of the heatmaps for RE, RH, and GON images are presented in Fig.7. Fig.7(a) shows typical images of fundus lesions, and Fig.7(b)—(d) are the three lesions classified as RE, RH, and GON generated by guided backpropagation in Fig.7(a) as a visualized heatmap. In the images of RE and RH, the highlighted area can be accurately located at the area where the lesion occurs, and in the images of GON, the highlighted area is located around the optic disc of the fundus. This interpretability feature of the DGNNDS may promote its application in real-world settings because photographers can understand how a final classification is made by the DGNNDS.




(a) ROC and AUC of DenseNet121 and DenseNet_GNN models in fundus multiple lesions

(b) ROC and AUC of ResNet101 and ResNet_GNN models in fundus multiple lesions

**Fig.6 ROC and AUC of different deep learning models on fundus multiple lesions**



(a) Original image (b) RE (c) RH (d) GON

**Fig.7 Visual heatmaps of ResNet_GNN model on three lesions: (a) Typical fundus lesion images; (b) (c) (d) Generated visualization heatmaps for RE, RH, and GON, respectively**

In the diagnosis of multiple lesions in fundus images, making full use of the relationship between fundus lesions is an important way to improve detection performance. In this paper, we propose a fusion GNN approach for multiple lesions diagnosis of fundus images, which extracts lesion features in fundus images by the CNN, then inputs the extracted features to the GNN to learn the correlation between different lesions. Finally, the input and output features of the GNN are fused and performed the diagnosis using a multi-label classifier. The experimental results show that the models of fused GNN, ResNet_GNN and DenseNet_GNN, achieve better results in terms of evaluation metrics compared to other models. In the next step, we will investigate in depth the dependencies of category semantic features and category-related image features, as well as the imbalanced data problem of different categories.

## Source code and data sharing statement

The code and data used in this study can be accessed at GitHub (https://github.com/jiangjiewei/DGNNDS_MFL).

## Statements and Declarations

The authors declare that there are no conflicts of interest related to this article.

## References

[1] YE H, ZHANG Q, LIU X, et al. Prevalence of age-related macular degeneration in an elderly urban chinese population in China：the Jiangning eye study[J]. Investigative ophthalmology & visual science, 2014, 55(10)：6374-6380.

[2] SONG L, LIN J, WANG Z J, et al. An end-to-end multi-task deep learning framework for skin lesion analysis[J]. IEEE journal of biomedical and health informatics, 2020, 24(10)：2912-2921.

[3] OGURTSOVA K, DA ROCHA FERNANDES J, HUANG Y, et al. Idf diabetes atlas：global estimates for the prevalence of diabetes for 2015 and 2040[J]. Diabetes research and clinical practice, 2017, 128：40-50.

[4] MUKAIDA M, OKAMI Y, SUETAKE N, et al. Contour

line extraction of vein with use of special tracking strategy and fuzzy inference for arteriosclerosis diagnosis of retinal blood vessel[C]//IECON 2020 the 46th Annual Conference of the IEEE Industrial Electronics Society, October 18-21, 2020, Singapore. New York：IEEE, 2020：435-440.

[5] LAI X, LI X, QIAN R, et al. Four models for automatic recognition of left and right eye in fundus images[C]//MultiMedia Modeling：25th International Conference, January 8-11, 2019, Thessaloniki, Greece. Berlin：Springer, 2019：507-517.

[6] AGARWAL A, GULIA S, CHAUDHARY S, et al. Automatic glaucoma detection using adaptive threshold based technique in fundus image[C]//2015 38th International Conference on Telecommunications and Signal Processing (TSP), July 09-11, 2015, Prague, Czech Republic. New York：IEEE, 2015：416-420.

[7] RATHORE S, ASWAL A, SARANYA P. Bright lesion detection in retinal fundus images for diabetic retinopathy detection using machine learning approach[J]. Annals of the Romanian Society for Cell Biology, 2021, 25(5)：4360-4367.

[8] COLOMER A, IGUAL J, NARANJO V. Detection of early signs of diabetic retinopathy based on textural and morphological information in fundus images[J]. Sensors, 2020, 20(4)：1005.

[9] ALAGUSELVI R, MURUGAN K. Performance analysis of automated lesion detection of diabetic retinopathy using morphological operation[J]. Signal, Image and Video Processing, 2021, 15(4)：797-805.

[10] JADHAV M, SHAIKH M, SARDAR V. Automated microaneurysms detection in fundus images for early diagnosis of diabetic retinopathy[C]//2020 the 3rd International Conference on Intelligent Control and Computing, January 10-12, 2020, Xiamen, China. Singapore：Springer, 2021：87-95.

[11] LI Z, JIANG J, CHEN K, et al. Preventing corneal blindness caused by keratitis using artificial intelligence[J]. Nature communications, 2021, 12(1)：1-12.

[12] JIANG J, WANG L, FU H, et al. Automatic classification of heterogeneous slit-illumination images using an ensemble of cost-sensitive convolutional neural networks[J]. Annals of translational medicine, 2021, 9(7)：550.

[13] IMRAN A, LI J, PEI Y, et al. Fundus image-based cataract classification using a hybrid convolutional and recurrent neural network[J]. The visual computer, 2021, 37(8)：2407-2417.

[14] LI Z, GUO C, LIN D, et al. Deep learning for automated glaucomatous optic neuropathy detection from ultra-widefield fundus images[J]. British journal of ophthalmology, 2021, 105(11)：1548-1554.

[15] LI Z, QIANG W, CHEN H, et al. Artificial intelligence to detect malignant eyelid tumors from photographic images[J]. NPJ digital medicine, 2022, 5(1)：1-9.

[16] GULSHAN V, PENG L, CORAM M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs[J]. Jama, 2016, 316(22)：2402-2410.

[17] OH K, KANG H M, LEEM D, et al. Early detection of diabetic retinopathy based on deep learning and ultra-wide-field fundus images[J]. Scientific reports, 2021, 11(1)：1-9.

[18] TANG S, YU F. Construction and verification of retinal vessel segmentation algorithm for color fundus image under bp neural network model[J]. The journal of supercomputing, 2021, 77(4)：3870-3884.

[19] DAS S, KHARBANDA K, SUCHETHA M, et al. Deep learning architecture based on segmented fundus image features for classification of diabetic retinopathy[J]. Biomedical signal processing and control, 2021, 68：102600.

[20] RAJ A, SHAH N A, TIWARI A K. A novel approach for fundus image enhancement[J]. Biomedical signal processing and control, 2022, 71：103208.

[21] LI Z, JIANG J, ZHOU H, et al. Development of a deep learning-based image eligibility verification system for detecting and filtering out ineligible fundus images：a multicentre study[J]. International journal of medical informatics, 2021, 147：104363.

[22] SHEN Y, SHENG B, FANG R, et al. Domain-invariant interpretable fundus image quality assessment[J]. Medical image analysis, 2020, 61：101654.

[23] WANG J, YANG L, HUO Z, et al. Multi-label classification of fundus images with efficientnet[J]. IEEE access, 2020, 8：212499-212508.

[24] GOUR N, KHANNA P. Multi-class multi-label ophthalmological disease detection using transfer learning based convolutional neural network[J]. Biomedical signal processing and control, 2021, 66：102329.

[25] DING W, SUN Y, REN L, et al. Multiple lesions detection of fundus images based on convolution neural network algorithm with improved sfla[J]. IEEE access, 2020, 8：97618-97631.

[26] PAN X, JIN K, CAO J, et al. Multi-label classification of retinal lesions in diabetic retinopathy for automatic analysis of fundus fluorescein angiography based on deep learning[J]. Graefe's archive for clinical and experimental ophthalmology, 2020, 258(4)：779-785.

[27] LI C, YE J, HE J, et al. Dense correlation network for automated multi-label ocular disease detection with paired color fundus photographs[C]//2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), April 3-7, 2020, Iowa City, IA, USA. New York：IEEE, 2020：1-4.

[28] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 26-July 1, 2016, Las Vegas, USA. New York：IEEE, 2016：770-778.

[29] HUANG G, LIU Z, VAN DER MAATEN L, et al.

Densely connected convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, July 22-25, 2017, Hawaii, USA. New York：IEEE, 2017：4700-4708.

[30] MORID M A, BORJALI A, DEL FIOL G. A scoping review of transfer learning research on medical image analysis using imagenet[J]. Computers in biology and medicine, 2021, 128：104115.

[31] CHEN T, XU M, HUI X, et al. Learning semantic-specific graph representation for multi-label image recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, October 27-November 2, 2019, Seoul, Korea. New York：IEEE, 2019：522-531.

[32] DEY R, SALEM F M. Gate-variants of gated recurrent unit (GRU) neural networks[C]//2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), August 6-9, 2017, Boston, MA, USA. New York：IEEE, 2017：1597-1600.

[33] LI Z Q, SUN J, WU X J, et al. Multiplication fusion of sparse and collaborative-competitive representation for image classification[J]. International journal of machine learning and cybernetics, 2020, 11：2357-2369.