# Research on Medical Image Data Privacy Protection and Collaborative Diagnosis Based on Federated Learning

Haolun Di *

School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou, Jiangxi, China

* Corresponding author Email: 15832263317@163.com

**Abstract.** Deep learning has shown strong capabilities in handling visual tasks such as medical imaging. However, the privacy protection characteristics of medical data have led to an extreme scarcity of data available for training deep models, which limits their further development. Based on the federated learning framework, this paper proposes a medical image classification algorithm FL-RA that integrates the attention mechanism, aiming to solve the problems of medical data privacy protection and data silos. The experiments used four public medical image datasets (COVID-19, PneumoniaMNIST, OrganSMNIST, and OCTMNIST). The results show that the performance of the model varies across different datasets, but in most cases, the recognition accuracy for positive samples is high. The results of the ablation experiment show that after introducing the attention mechanism, the classification accuracy of the model is improved by 1%-2%, and the accuracy of the model trained jointly using federated learning is 4%-8% higher than that of local training alone. This study provides an effective technical solution for privacy protection and collaborative diagnosis in medical image diagnosis, and has important practical significance and broad application prospects.

**Keywords:** Medical Image Diagnosis; Federated Learning; FL-RA; ResNet-50; Attention Mechanism.

## 1. Introduction

Medical image diagnosis technology refers to the method of diagnosing diseases using different types of medical imaging technologies. Using technologies such as deep learning, medical images can be automatically analyzed and diagnosed, assisting doctors in disease analysis and lesion diagnosis, reducing the diagnostic pressure on doctors, improving the accuracy and efficiency of diagnosis, and reducing the possibility of misdiagnosis and missed diagnosis. It has good development and application prospects and practical significance in the field of medical and health care[1].

Deep learning has shown strong capabilities in handling tasks such as medical imaging, but the privacy and confidentiality of medical data have led to an extreme lack of training data for deep models, restricting the development of deep learning in medical image diagnosis tasks. On the one hand, in reality, data often exists in the form of "silos", and these "islands" cannot communicate with each other[2]. For medical image diagnosis tasks, the data from a single medical institution is difficult to train a highly accurate and robust image diagnosis model, thereby reducing the effectiveness and authority of diagnosis[3]. On the other hand, preventing and punishing acts that infringe on personal information rights and interests is an important law in the field of personal information protection. It can be seen that under the background of data silos and data privacy security, traditional deep learning under centralized data conditions faces enormous difficulties and challenges.

Faced with these two major problems, federated learning, as a distributed learning method that can protect data privacy and enable joint training of data information, provides a feasible solution to solve the problems of privacy protection and data silos[1]. Federated learning can establish connections between scattered data "silos" under the premise of ensuring that private data is not leaked, and use data information from multiple medical institutions for joint training of deep learning models.

In summary, this paper proposes a medical image classification algorithm FL-RA based on the federated learning framework, and designs a fusion mechanism to effectively combine the attention mechanism with ResNet-50. The algorithm enables ResNet-50 to focus more on valuable feature

information through the attention mechanism, expands the receptive field of the channel, and allows the model to capture subtle lesion features when processing complex and variable medical images. The method used in this paper enhances the performance of the model in medical image classification tasks and provides strong technical support for improving the accuracy and reliability of medical image diagnosis.

## 2. Research Methods and Model Construction

### 2.1 Dataset

This paper conducts experiments based on four public medical image datasets, and the basic information of the datasets is shown in Table.1. The medical MNIST datasets used in this paper are all from the medical image dataset MedMNIST[4].

**Table 1.** Introduction to datasets

| Dataset Name | Data Modality | Task Category | Number of Samples | Training/Validation/Testing |
|---|---|---|---|---|
| COVID-19 | Chest CT | Binary Classification | 6700 | 5360/670/670 |
| PneumoniaMNIST | Chest X - ray | Binary Classification | 5856 | 4708/524/624 |
| OrganSMNIST | Abdominal CT | Multi - classification/11 | 25221 | 13940/2454/8829 |
| OCTMNIST | Retinal OCT | Multi - classification/4 | 109309 | 97477/10832/1000 |

This research involves four medical image datasets: The COVID-19 dataset is derived from the COVID-CTset collected by the Negin Medical Center in Sari, Iran[5], containing CT images of 95 patients and 282 normal individuals with a resolution of 512×512. After screening and division, the training set has 5000 images, the test set has 1000 images, and the validation set has 1000 images, which are cropped to 224×224. PneumoniaMNIST[6] integrates 5856 children's lung X-ray images, divided into training set and validation set at a ratio of 9:1, and then the validation set is used as the test set. The images are cropped and adjusted to 1×28×28. The numbers of images in the training set, validation set, and test set are 4708, 524, and 624 respectively. OrgansMNIST[6] is obtained from 3D computed tomography images of the Liver Tumor Segmentation Benchmark (LiTS), adjusted to a size of 1×28×28, covering 11 types of human organ images, totaling 25221 images. The numbers of images in the training set, validation set, and test set are 13940, 2452, and 8829 respectively. OCTMNIST is built based on a prior set[7], collecting 109309 OCT images for the treatment of retinal diseases, including 4 diagnostic categories. The numbers of images in the training set, validation set, and test set are 97477, 10832, and 1000 respectively. These datasets provide data support for research on medical image-related algorithms.

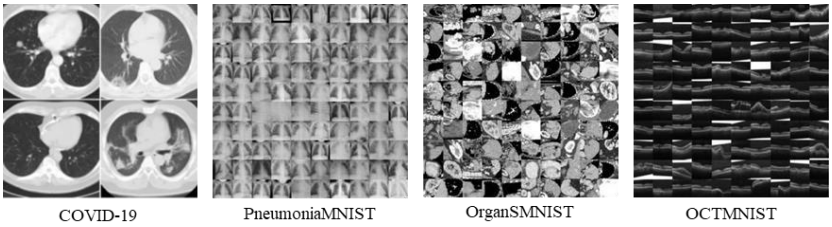


**Figure 1.** Partial examples of datasets

### 2.2 Federated Learning-Residual Network with Attention (FL-RA)

This paper aims to solve the problem of jointly training and obtaining a classification model with high generalization performance by multiple medical institutions using their respective data under the premise of ensuring privacy and security. It is assumed that there are K medical institutions as participants conducting joint modeling in a horizontal federated learning system. Let $D_k$ be the data

set owned by the k-th medical institution, where each item is $(x, y), x \in R^d$ is the input feature, y is the corresponding label, $P_k$ is the index set corresponding to the data of the k-th participant, $n_k = |P_k|$ is the size of the data set of the k-th participant, $\omega_i$ is the weight parameter of the local model of each participant, and $f(\omega_i, x_k)$ is the crossentropy loss function of the model[3]. The optimization objective function for each participant is:

$$\min_{e_i} L_i(\omega_i) = -\frac{1}{|P_k|}\sum_{k}^{|p_k|} y_k \log\left(f(\omega_i, x_k)\right)$$

The specific workflow of horizontal federated learning modeling is as follows:

(1) Update local model parameters: If it is the first round of joint training, the central server will initialize the model weight parameters and send them to each medical institution; if it is not the first round of training, the latest model after joint modeling will be sent to each participant.

(2) Select local users: Before the first global joint modeling, the central server will select some medical institution units that meet the conditions for participating in joint training. In subsequent training rounds, only participants who have completed the previous round of training and are in an idle state can participate in the next round of joint learning training.

(3) Update local user parameters: After receiving the globally aggregated weight parameter model, each medical institution conducts local model training and parameter updates, and optimizes the local model using stochastic gradient descent.

(4) Upload local models: After each medical institution reaches the local iteration rounds and completes the local model update, it uploads the latest weight parameters to the central server. During the above model transmission process, the data of each participant remains local, which will not cause data leakage.

(5) Global model aggregation: After receiving the local model weight parameters of each medical institution in the new round, the central server uses the federated averaging algorithm to perform global model aggregation to obtain the latest global model parameters. The above steps (1) to (5) constitute one global iteration round. The joint learning process terminates until the local model converges or the specified number of global iterations is reached.

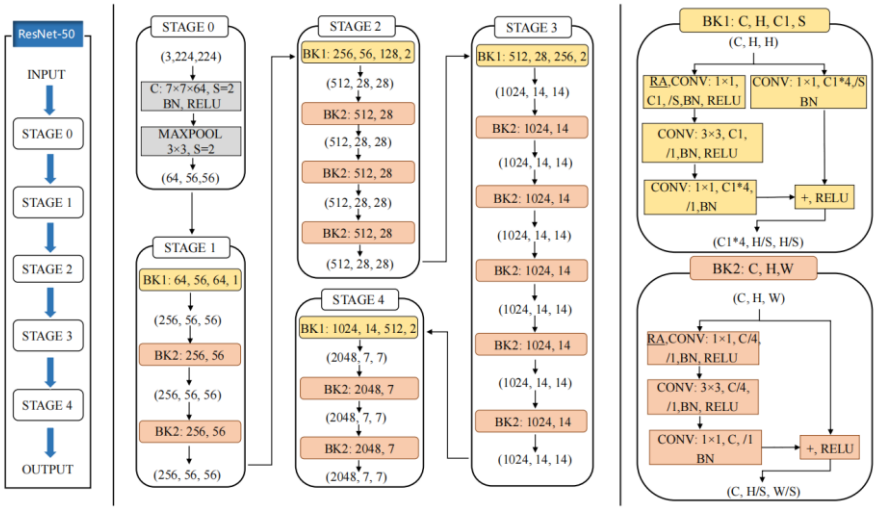


**Figure 2.** ResNet-50 backbone network structure

As shown in Figure 2, STAGE1-STAGE4 are residual operations[7,8]. BN is the abbreviation of Batch Normalization layer, and BK is the abbreviation of Bottleneck. In the BK1 structure, the number of channels of the feature map before and after input and output is different, while in BK2, the number of channels of the feature map before and after input and output is the same. Among them, C is the number of channels of the input feature map, C1 is the number of channels of the left

convolution layer in the BK structure, H and W correspond to the height and width of the input feature map respectively, and S is the stride of the convolution kernel. To enhance the attention of the neural network to important features, this paper introduces a channel attention mechanism into the BK structure, where RA represents the part combining channel attention and convolutional neural network.

For the attention mechanism part, as shown in Figure 3, first, three $1 \times 1$ convolution kernels are used to convolve the feature map to obtain query, key, and value in the attention mechanism. At the same time, the size of the feature map obtained through the convolution operation is set to $N \times H \times W \times C/N$, where the first $N$ is used to simulate multiple heads of the attention mechanism[9]. Then, the similarity matching weight of attention is obtained by combining each head, and the N weights are connected to output the similarity matching weight with a size of $H \times W \times C$, denoted as $W_1$. For the convolution part, this paper still uses $1 \times 1$ convolution kernels to perform convolution operations on the feature map, and the convolution output result is denoted as $W_2$. The output result of the module is obtained by linearly combining the results of the attention part and the convolution part, which is $\alpha W_1 + \beta W_2$. In this paper, the $H \times W \times C$ feature map input into Bottleneck1 and Bottleneck2 is divided into convolution and multi-head attention modules, and their results are weighted and summed.

This paper first performs maximum pooling downsampling on key and value, and the remaining steps follow the traditional multi-head self-attention module. The advantages of this method are as follows: (1) Reduce the feature dimension of the attention module, reduce the number of parameters that the network needs to learn and the computational complexity; (2) Expand the receptive field and further extract effective features.

Let the input feature $F \in \mathbb{R}^{C_{in} \times H \times W}$, the output feature $G \in \mathbb{R}^{C_{in} \times H \times W}$, and $C_{in}$ and $C_{out}$ are the input and output channel sizes. $f_{ij} \in R^{C_m}$ and $g_{ij} \in R^{Cout}$ represent the tensors corresponding to the pixel (i, j) of F and G . The output of the attention module is:

$$g_{ij} = \|_{l-1}^{N} \left( \sum_{a,b \in N_k(i,j)} A\left(W_q^{(l)} f_{ij}, W_k^{(l)} f_{ab}\right) W_v^{(l)} f_{ab} \right)$$

Among them, $\|$ is the standard self-attention module for N heads, $W_q^{(l)}, W_q^{(l)}$, and $W_q^{(l)}$ correspond to queries, keys, and values respectively, representing the local area of the pixel. Its spatial range $k$ is centered on $(i, j)$, and $A\left(W_q^{(i)} f_i, W_k^{(i)} f_{ab}\right)$ is the corresponding attention weight for features within $N_k(i, j)$.

The widely used self-attention module calculates the attention weight as follows:

$$A\left(W_q^{(t)} f_y, W_k^{(t)} f_{ab}\right) = \mathrm{softmax}_{N_t(i,j)} \left( \frac{\left(W_q^{(t)} f_y\right)^T \left(W_k^{(t)} f_{ab}\right)}{\sqrt{d}} \right)$$

Among them, d is the feature size of $W_q^{(l)} f_{ij}$. In summary, the multi-head channel self-attention mechanism can be expressed in two stages [9]:

$$\text{Stage I: } q_{ij}^{(l)} = W_q^{(l)} f_{ij}, k_{ij}^{(l)} = W_K^{(l)} f_{ij}, v_{ij}^{(l)} = W_v^{(l)} f_{ij}$$

$$\text{Stage II: } g_{ij} = \prod_{l-1}^{N} \left( \sum_{a,b \in N_k(i,j)} A\left(q_{ij}^{(l)}, k_{ab}^{(l)}\right) v_{ab}^{(l)} \right)$$

## 3. Results

The experimental parameter settings in this paper are shown in Table.2, where the global iteration number T represents the number of global model aggregations in joint modeling, the local update

number S represents the number of iterations for medical institution participants to train the classification model using local data, and the coefficient K represents the number of clients actually participating in joint modeling, with a value range of [0,1] and a default value of 1.

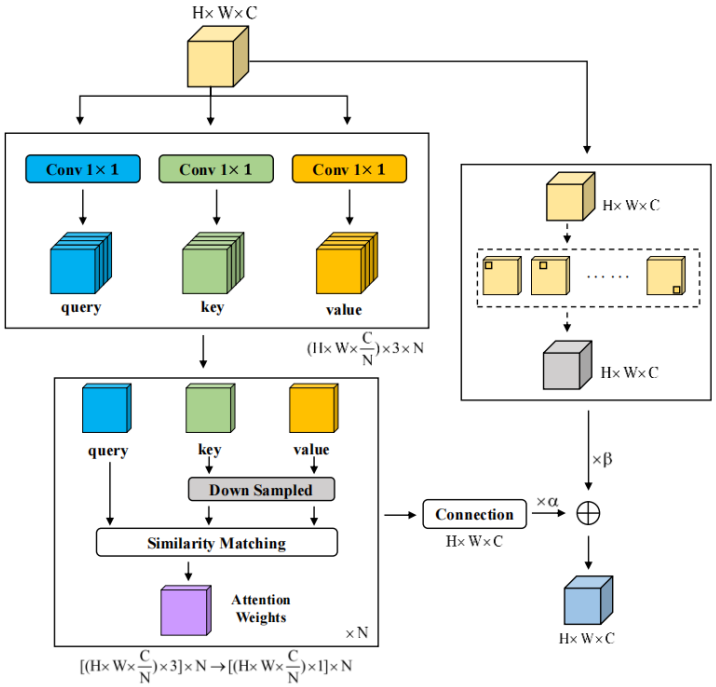


**Figure 3.** Attention fusion mechanism in FL-RA algorithm

**Table 2.** Comparison of results of each model (including new/old user variables)

| Parameter Symbol | Parameter Description | COVID-19 | PneumoniaMNIST | OrganSMNIST | OCTMNIST |
|---|---|---|---|---|---|
| T | Global iteration rounds | 20 | 20 | 20 | 20 |
| S | Local update rounds | 10 | 10 | 10 | 10 |
| M | Local batch size | 16 | 16 | 16 | 16 |
| lr | Learning rate | 0.01 | 0.01 | 0.01 | 0.01 |
| K | Proportion of participants in training | 1 | 1 | 1 | 1 |
| N | Total number of participants | 5 | 10 | 20 | 20 |

As shown in Figure 4, this chapter compares the accuracy of classification models under different training conditions on COVID-19, PneumoniaMNIST, OrganSMNIST, and OCTMNIST datasets, including the accuracy of the participant's local model without combining the channel attention mechanism, the accuracy under joint modeling and centralized data conditions, and the corresponding model accuracy when combining the Attention mechanism.

Among them, local, fed, and assemble respectively represent the accuracy of medical image classification of the local training model of medical institution participants when not combining the channel attention mechanism, under joint modeling and centralized data conditions; FL-RA-loc, FL-RA-fed, and RA-assemble respectively represent the accuracy of the medical image classification model of the FL-RA algorithm under the local, joint modeling, and centralized data conditions of participants. It can be seen from the bar chart of experimental results that whether it is the local model of participants, joint modeling, or centralized data conditions, the classification accuracy of the classification model introducing the channel attention mechanism is increased by 1%-2%. This indicates that after introducing the Attention mechanism, the receptive field of the neural network is increased, and more attention is paid to global features when extracting features. For classification

tasks, the features extracted by the model network are more representative. At the same time, in the case of joint training, the model jointly established by participants using local data has higher generalization performance. On the basis of separate training by medical institution participants, the accuracy of the classification model for joint modeling is increased by 4%-8%, which indicates that federated learning can effectively break data barriers and integrate the information value in data "islands".

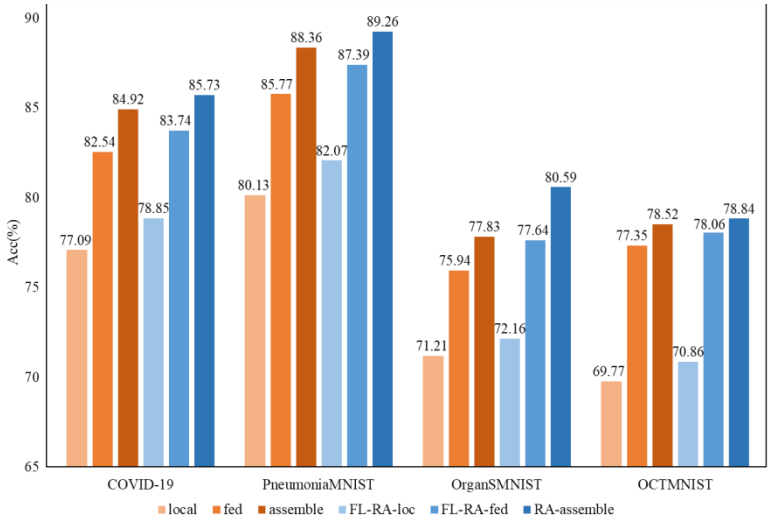


**Figure 4.** Ablation experiments of FL-RA algorithm under various conditions

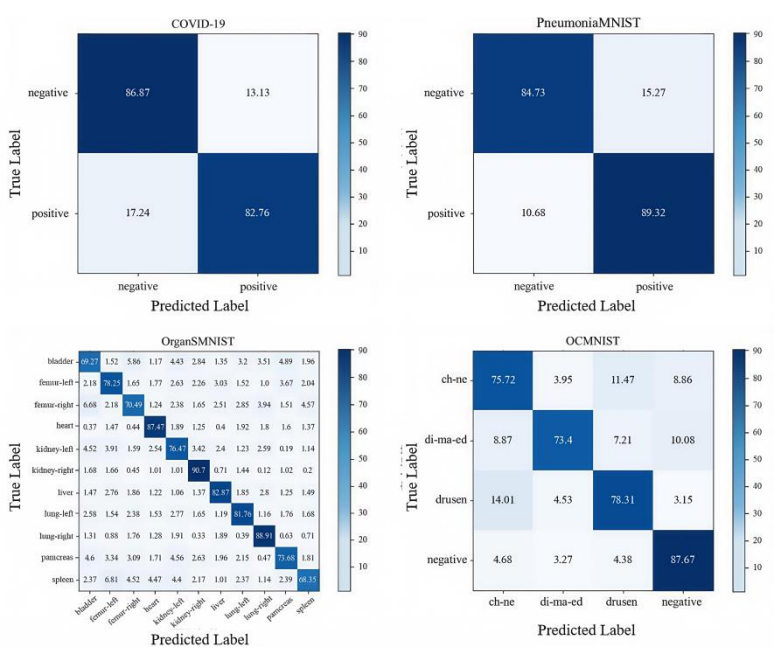


**Figure 5.** Confusion matrix based on FL-RA algorithm

Figure 5 shows the confusion matrix of classification under the framework of joint training for each dataset. For binary classification tasks, this paper denotes the categories as positive and negative. Among them, the classification effect of the negative category of COVID-19 is better than that of the positive category; the classification effect of the two categories of PneumoniaMNIST is opposite to that of COVID-19, that is, the classification effect of the predicted positive category is better; for the OrganSMNIST dataset, the classification effects of heart, kidney-right, and lung-right are better than those of other categories; in OCTMNIST, the classification effect of the negative category is the best, followed by drusen, which is better than the other two categories.

To explore the effectiveness of federated learning, as shown in Table.3, K = 1 is set here. When the data volume of the four datasets is fixed, as the number of medical institution participants N increases, the local data volume of medical institutions decreases, and the accuracy of the model also decreases. The accuracy of the classification model obtained by joint training has an improvement of 1%-7%. This indicates that the less data information the participants have, the greater the information entropy of the local model, and the more significant the improvement effect brought by federated learning.

**Table 3.** Impact of the number of participants on the accuracy of the classification model

| Dataset | Number of Participants | Local | Federal | Dataset | Number of Participants | Local | Federal |
|---|---|---|---|---|---|---|---|
| COVID-19 | 5 | 78.85 | 83.46 | OrganSMNIST | 5 | 77.92 | 79.64 |
| | 10 | 75.54 | 78.82 | | 10 | 75.85 | 78.89 |
| | 20 | 69.73 | 75.57 | | 20 | 72.03 | 76.25 |
| | 50 | 62.17 | 70.29 | | 50 | 63.87 | 71.52 |
| PneumoniaMNIST | 5 | 86.21 | 88.34 | OCTMNIST | 5 | 78.84 | 79.52 |
| | 10 | 82.07 | 87.11 | | 10 | 78.25 | 79.68 |
| | 20 | 77.26 | 84.27 | | 20 | 75.26 | 78.44 |
| | 50 | 70.69 | 78.51 | | 50 | 68.34 | 74.83 |

## 4. Conclusion

This research addresses the issues of medical image data privacy protection and collaborative diagnosis, and proposes the FL-RA algorithm based on the federated learning framework, which integrates the attention mechanism with ResNet-50. The main conclusions are as follows:

The experimental results verify the effectiveness of the proposed method. On the four datasets of COVID-19, PneumoniaMNIST, OrganSMNIST, and OCTMNIST, after introducing the channel attention mechanism, the classification accuracy of the model in local training, joint modeling, and centralized data scenarios all increased by 1%-2%. This indicates that the attention mechanism can help the network expand the receptive field, more effectively capture representative features, and improve the ability to identify subtle lesions in complex medical images.

Federated learning has a significant effect in breaking data silos. Compared with the local training model of a single medical institution, the classification accuracy of multi-institution joint modeling is increased by 4%-8%, which proves that federated learning can integrate data information from various institutions without leaking data privacy, improve the generalization performance of the model, and effectively solve the problem of data silos. Research on the impact of the number of participants shows that when the increase in the number of participants leads to a reduction in the local data volume of each institution, the accuracy of the joint training model can still be improved by 1%-7%. This indicates that federated learning is still robust under the condition of uneven data distribution, providing a practical solution for multi-institutions to carry out collaborative diagnosis under the constraint of privacy protection.

In summary, the FL-RA algorithm proposed in this paper effectively combines federated learning with the attention mechanism, enhancing the performance of the medical image classification model. In the future, the fusion mechanism of the attention module can be further optimized to expand its application in more complex medical image diagnosis tasks.

## References

[1] Guan H, Yap P T, Bozoki A, et al. Federated learning for medical image analysis: A survey[J]. Pattern Recognition, 2024: 110424.

[2] Huang S, Diao S, Wan Y, et al. Research on multi-agency collaboration medical images analysis and classification system based on federated learning[C]//Proceedings of the 2024 International Conference on Biomedicine and Intelligent Technology. 2024: 40-44.

[3] Haripriya R, Khare N, Pandey M. Privacy-preserving federated learning for collaborative medical data mining in multi-institutional settings[J]. Scientific Reports, 2025, 15(1): 12482.

[4] PAN X, GE C, LU R, et al. On the integration of self-attention and convolution[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 815-825.

[5] YANG J, SHI R, NI B. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis[C]//2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). IEEE, 2021: 191-195.

[6] RAHIMZADEH M, ATTAR A, SAKHAEI S M. A fully automated deep learning-based network for detecting COVID-19 from a new and large lung CT scan dataset[J]. Biomedical Signal Processing and Control, 2021, 68: 102588.

[7] YANG J, SHI R, WEI D, et al. v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification[J]. arXiv preprint arXiv:2110.14795, 2021, 3(4).

[8] QIN Z, ZHANG P, WU F, et al. Fcanet: Frequency channel attention networks[C]//Proceedings of the IEEE/CVF international conference on computer vision. 021: 783-792.

[9] NIU Z, ZHONG G, YU H. A review on the attention mechanism of deep learning[J]. Neurocomputing, 2021, 452: 48-62.