J. Meas. Sci. Instrum., 2025, 16(1): 1-10.
DOI: 10.62756/jmsi.1674-8042.2025001
http://jmsi.publish.founderss.cn/
jmsi@nuc.edu.cn

Multi-scale input mirror network for tuberculosis detection in CXR image

XING Guangxin[1], FAN Jingjing[2], ZHENG Yelong[1], ZHAO Meirong[1*]

1. State Key Laboratory of Precision Measuring Technology and Instruments, Tianjin University, Tianjin 300072, China;
2. Department of Pharmacy, Medical Supplies Center, PLA General Hospital, Beijing 100853, China

***Corresponding author:** ZHAO Meirong (meirongzhao_acad@163.com)

Received: July 23, 2024 **Revised:** August 20, 2024 **Accepted:** September 28, 2024

Abstract: Computer-aided diagnosis (CAD) can detect tuberculosis (TB) cases, providing radiologists with more accurate and efficient diagnostic solutions. Various noise information in TB chest X-ray (CXR) images is a major challenge in this classification task. This study aims to propose a model with high performance in TB CXR image detection named multi-scale input mirror network (MIM-Net) based on CXR image symmetry, which consists of a multi-scale input feature extraction network and mirror loss. The multi-scale image input can enhance feature extraction, while the mirror loss can improve the network performance through self-supervision. We used a publicly available TB CXR image classification dataset to evaluate our proposed method via 5-fold cross-validation, with accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and area under curve (AUC) of 99.67%, 100%, 99.60%, 99.80%, 100%, and 0.999 9, respectively. Compared to other models, MIM-Net performed best in all metrics. Therefore, the proposed MIM-Net can effectively help the network learn more features and can be used to detect TB in CXR images, thus assisting doctors in diagnosing.
Key words: computer-aided diagnosis (CAD); medical image classification; deep learning; feature symmetry; mirror loss

0 Introduction

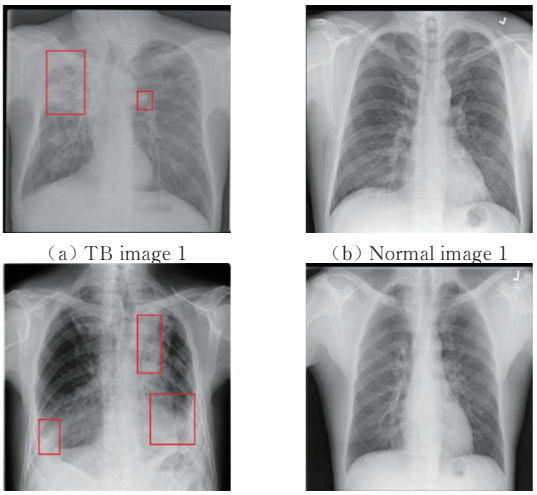
Tuberculosis (TB) is an infectious disease caused by Mycobacterium tuberculosis[1]. TB usually targets the lungs, and the bacteria can be spread through the air from an infected person. TB affects 10 million people globally, of which 56 percent are men, 32 percent are women, and 12 percent are children[2]. TB can be cured and prevented with proper diagnosis and appropriate medication[3].

Two tests are commonly used to diagnose TB infection: the Mantoux TB skin test and the interferon-gamma release assays (IGRA) blood test[4]. The skin test is sensitive but cannot distinguish between latent and active infections, while the IGRA is very accurate but expensive and difficult to perform in low and middle-income regions worldwide. Chest X-ray (CXR) is a commonly used medical imaging tool to detect lung abnormalities, especially TB. CXR is widely used for TB detection because it is fast, cost-effective, and easy to use in remote areas[5].

TB is a disease that occurs in the lung tissue, trachea, bronchi, and pleura[6]. It can take many forms, so diagnosing TB through CXR images is challenging and requires much expertise[7]. The importance of computer-aided diagnosis (CAD) systems is growing as technology and equipment become more widely available. CAD provides radiologists with more accurate and efficient diagnostic solutions without the interference of subjective factors[8]. The World Health Organization (WHO) recommends using CAD systems to replace radiologists in screening CXR images for pulmonary TB[9]. As shown in Fig.1, the first column shows a TB infection images, with red rectangles indicating lesion locations; and the second column shows a normal images (uninfected). Both lesions are pulmonary consolidations in the upper image of Fig.1(a) and (c). In the lower image of Fig.1(c), the lesion in the upper right is a pulmonary consolidation, with pleural effusions in the lower left and lower right. Due to various types of noise and interference information in CXR images, classifying TB CXR images is a challenging computer vision task.

With the development of computer science, many artificial intelligence (AI)-based studies have emerged, including solutions to CAD problems[10,11]. Deep learning based on convolutional neural networks (CNN) has been successfully applied to image detection. Many deep learning methods have been applied to medical image analysis, among which the research on deep learning-based TB detection methods has attracted much attention, and deep

learning TB detection methods have become one of the popular research areas in the interdisciplinary field of AI and medicine[12]. TB research can be categorized into traditional machine learning-based and deep learning-based methods, some of which have succeeded remarkably. Inspired by these studies and to further investigate the feasibility of deep CNNs in TB diagnosis, we proposed a novel model for TB CXR image detection.



(a) TB image 1 (b) Normal image 1

(c) TB image 2 (d) Normal image 2

Fig. 1 Different CXR images with red rectangles indicating the location of the lesion

This study aims to develop a high-performance neural network architecture named multi-scale input mirror network (MIM-Net) to detect TB better using CXR images, which uses multi-scale inputs as the backbone for image feature extraction, and mirror loss is the loss function. In this study, we used a 5-fold cross-validation approach to evaluate the experimental results so as to investigate the feasibility of the methodology of this study. In summary, the main contributions of this study can be summarized as follows:

1) This study proposes a novel network backbone for CXR image feature extraction. The network structure uses multi-scale images as input, while dense connections are incorporated into the network backbone to enhance feature extraction.

2) This study proposes a mirror loss based on the symmetry features in CXR images, and the mirror loss allows the network to improve performance through self-supervision.

1 Related work

In medical image classification, TB detection research is divided into two categories: machine learning-based methods and deep learning-based methods.

Machine learning is a pattern recognition technique, and many studies have been done on diagnosing TB. Most traditional machine learning-based methods are based on manual extraction of features. Santosh et al.[13] analyzed the symmetry of the lungs by combining the shape, edges, and texture of the lungs to obtain vectors consisting of similarity scores. They then used Bayesian networks, multilayer perceptual neural networks, and random forests to detect anomalies in these vectors. Their experiments used three datasets with a final detection accuracy of 91% and an area under curve (AUC) of 0.96. Chandra et al.[14] used a two-step approach to extract features: the first step was to extract geometric features of the lungs, and the second step was to extract texture features and classify the extracted features into normal and abnormal. The accuracy of this method was 95.6% and 99.40% on the two experimental datasets. Soni et al.[15] used an Emboss differential filter to highlight the contours of the lung CT images, used the Gaussian spectrum to enhance the image contrast, and finally used a support vector machine (SVM) to classify the images. Their dataset contained 170 positive and 116 negative samples with a test accuracy of 96.5%. Although traditional image processing methods have received some attention, manually extracting image features requires identifying highly discriminative features to recognize target areas effectively. In addition, manual observation may be limited in image texture and geometry. As a result, image classification methods for traditional objectives are labor-intensive and often yield unsatisfactory results.

With the development of computer science, the application of deep learning methods in medical image analysis has achieved satisfactory results. The deep learning method for detecting TB garners attention. Deep learning-based methods mostly use CNN to automatically extract features from images without manually finding image features. Li et al.[16] combined CNN feature extraction with unsupervised features from autoencoder to construct a deep network model for TB chest image classification. They used a dataset of 3 600 CT images in their experiment and obtained 80.29% accuracy, 80.67% recall, and 80.42% $F1$-score. Ahsan et al.[17] proposed a VGG16-based transfer learning model for TB detection. In their experiments, they fine-tuned the model using 1 324 CXR images and their results were 80.0% accuracy without enhancement and 81.25% accuracy with enhancement. Huang et al.[18] proposed a deep transferred EfficientNet with SVM (DTE-SVM), which replaces the pre-trained EfficientNet classification layer with an SVM classifier. The DTE-SVM ran a 10-fold cross-validation approach on a dataset of 288 images, which resulted in a sensitivity of 93.89%±1.96%, a specificity of 95.35%±1.31%, a

precision of 95.30%±1.24%, an accuracy of 94.62%±1.00%, and an $F1$-score of 94.57%±1.05%. Mahbub et al.[19] proposed a lightweight deep neural network (DNN) for non-healthy CXR screening. On three diverse publicly accessible and fully categorized datasets, their proposed DNN achieved the following accuracies: 99.87% on COVID-19 versus healthy, 99.55% on Pneumonia versus healthy, and 99.76% on TB versus healthy datasets. Some scholars have applied Transformer to computer vision fields such as image classification[20]. Duong et al.[21] proposed a modified hybrid EfficientNet with vision Transformer (ViT) for detecting TB from CXR images. They evaluated the proposed method using datasets merged from various public datasets and showed that the maximum accuracy of the method was 97.72% with an AUC of 100%. Okolo et al.[22] proposed the input enhanced visual Transformer (IEViT) for chest X-ray image classification. Their experiments on four chest X-ray image datasets showed that the IEViT model outperformed ViT in all cases, with $F1$-score ranging between 96.39% to 100%, sensitivities ranging between 93.50% to 100%, and accuracies ranging between 97.96% to 100%. Deep learning-based TB detection methods are currently recognized as solutions that can help clinicians with diagnosis, especially in resource-limited areas.

2 Proposed methods

2.1 Feature extraction network

A CNN usually consists of convolutional, pooling, and fully connected layers. Convolutional and pooling layers generally do the feature extraction process of CNN. The convolutional layer is used to extract features from the input image, and the pooling layer reduces the size of the feature map and increases its receptive field. The fully connected layer is a classifier that uses the extracted features for classification. Convolution calculation can be defined as

$$x_{i+1}=\sigma\left(x_i \otimes w_i+b_i\right), \tag{1}$$

where x_i is the input of the layer, x_{i+1} is the output of the layer, w_i is the weight of the layer i, b_i is the bias of the layer i, σ is the activation function, and $\otimes$ is the convolution operator. The activation function improves the network fit by changing the nonlinear output of the convolutional layer.

Since the semantic information of CXR images is relatively homogeneous, we design a multi-scale image input and fusion as the network backbone for feature extraction. As shown in Fig.2, the network backbone of MIM-Net resizes the input image five times and then inputs the network at the same time, and the numbers in the figure represent the size of the feature map.

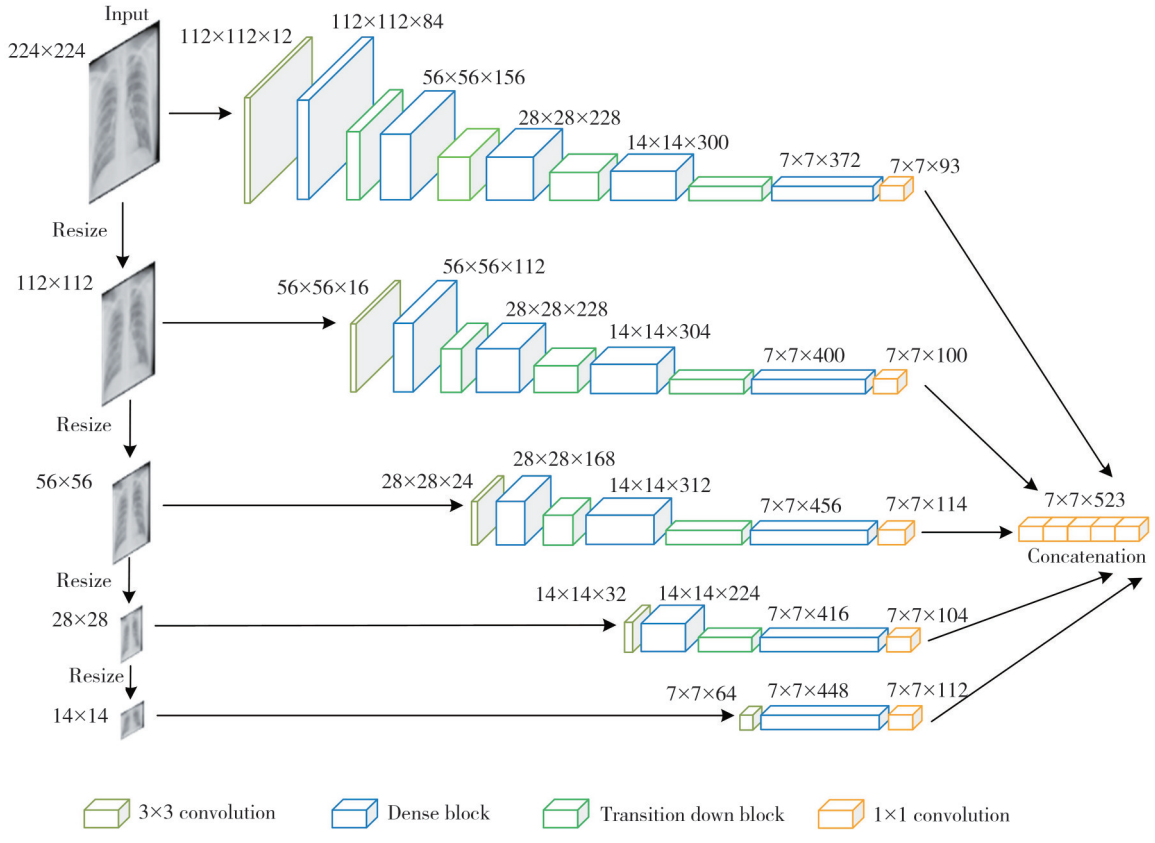


Fig. 2 Backbone network of MIM-Net

The network backbone of MIM-Net has five branches, and the input image is resized five times, each time reducing the width and height of the image to half of the previous one. The input image is fed by five branches simultaneously for feature extraction: the first branch performs four downsampling operations, the second branch performs three downsampling operations, the third branch performs two downsampling operations, the fourth branch performs one downsampling operation, and the fifth branch does not perform any downsampling operation. Downsampling is performed by transition down block, which consists of batch normalization (BN) [23], ReLU activate function, 1×1 convolution with stride 1, Dropout[24], and 2×2 max pooling. The number of branches is equal to the maximum number of downsampling between branches; therefore, there are five branches in MIM-Net for feature extraction of the input image. The first layer of each branch is a 3×3 convolution with a step size of 2. Finally, using 1×1 convolution with a stride of 1, the downsampled feature maps of each branch are adjusted to 1/4 of the previous channel numbers before merging. This combined feature map is then fed into the subsequent network for classification.

Dense connectivity allows more efficient use of features through feature reuse. Dense connectivity allows the network to maintain higher accuracy with fewer parameters. Dense connectivity connects each layer of the network directly to its front layer. Therefore, the layer i can directly use the feature maps of all previous layers

$$x_{i+1}=H_i([x_1,x_2,\cdots,x_i]), \tag{2}$$

where $[x_1,x_2,\cdots,x_i]$ represents the concatenation of feature mapping from layer 1 to layer i, and $H_i(\cdot)$ consists of BN, ReLU activate function, a convolution, and Dropout. The output of layer i is k feature maps, and k is the growth rate, which is usually set to a smaller value. The structure of the dense block is shown in Fig.3, the dense block designed in this study contains 12 convolutions. The layer in the dense block outputs feature maps of the same colour, the superposition of different coloured feature maps represents the concatenation, and the final feature map is a superposition of 12 layers of feature maps. As shown in Fig.2, in the network backbone of MIM-Net, the growth rates of the five branches from top to bottom are set to 6, 8, 12, 16, and 32, respectively.

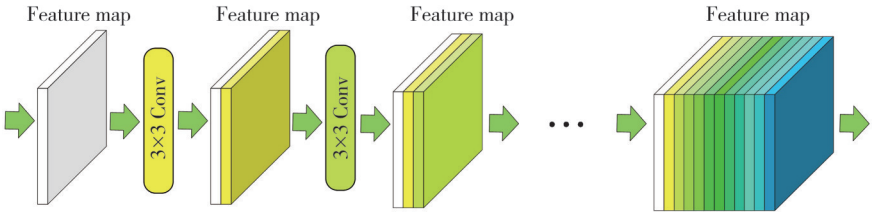


Fig. 3 Structure of dense block

2.2 Loss function

We propose mirror loss during feature extraction to better utilize the CXR image's features. Due to the feature symmetry of the lungs, we mirror-flip the last convolutional layer of MIM-Net, calculate its pixel-level image difference from itself, and add it to the final loss function, as shown in Fig.4. Mirror loss compares the mirror output of the convolutional layer with itself based on the features of the left lung and right lung in the image, thereby improving data classification performance through self-supervision of the network.

The mirror loss consists of two parts: the cross-entropy loss and the mean squared error (MSE)-based loss. First, in the binary classification task, the commonly used cross-entropy loss function $L_{\text{cross-entropy}}$ can be defined as

$$L_{\text{cross}-\text{entropy}}=-\sum_{i=1}^{n}[\,y_i\log x_i+(1-y_i)\log(1-x_i)\,], \tag{3}$$

where y_i is the input label with 1 for positive and 0 for negative, and x_i is the probability that the input is predicted to be positive. The MSE-based loss function $L_{\text{MSE-based}}$ is defined as

$$L_{\text{MSE}-\text{based}}=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}y_i(p_i-q_i)^2. \tag{4}$$

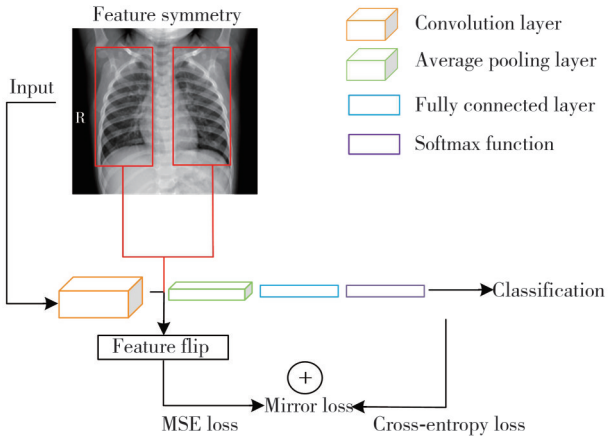


Fig. 4 Structure of mirror loss

Similarly, y_i is the input label with 1 for positive and 0 for negative. p_i is the pixel's value in the last convolutional layer feature map and q_i is the mirror flip

value of its corresponding pixel, and n is the number of pixels on the feature map. The final loss function L_{mirror} can be expressed as

$$L_{\text{mirror}} = L_{\text{cross-entropy}} + \alpha L_{\text{MSE-based}}, \quad (5)$$

where α is the scale factor, which is set to 1/5 here since the ratio of TB images to normal images is 1∶5.

3 Experiments and discussion

3.1 Materials and evaluation methods

The TB CXR dataset used for training, validation, and testing in this study is a publicly available TB CXR image classification dataset[25]. The dataset comprises 700 TB and 3 500 normal images, totaling 4 200 images. This study used a 5-fold cross-validation approach for data preparation and model evaluation. The ratio of TB images to normal images is about 1∶5, so we used a stratified cross-validation approach for model evaluation. We randomly divided the TB and normal CXR images from the original dataset into six groups, five groups of 720 images for cross-validation and one group of 600 images for testing. We ensured that images in each group did not cross over during the experiment while maintaining the ratio of the two types of images in each group.

Specifically, the cross-validation approach is shown in Fig.5, with four folds for training, one fold for validation, one fold for testing, and the numbers indicate the images in each fold. Each fold in the cross-validation contains 120 TB images and 600 normal images, and the test set includes 100 TB images and 500 normal images. The training set is used to set the model's parameters, the validation set is used to adjust the model's hyperparameters, and the test set is used to evaluate the model's performance.

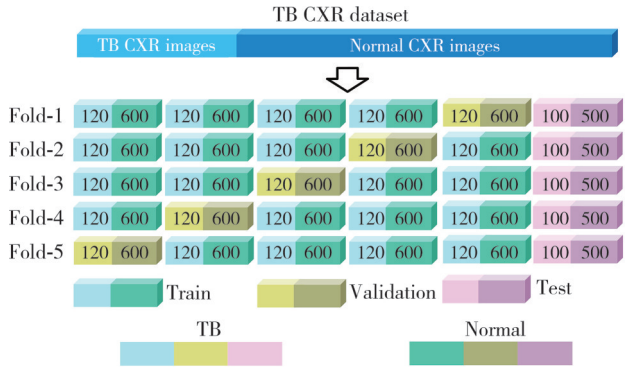


Fig. 5 Schematic representation of 5-fold cross-validation

In the classification phase, the model performance can be evaluated by six performance indicators: accuracy (*acc*), sensitivity (*sen*), specificity(*spec*), positive predictive value (V_{PP}), negative predictive value (V_{NP}), and AUC. Accuracy is the proportion of correctly predicted samples to all samples, sensitivity is the proportion of correctly predicted positive samples to all positive samples, specificity is the proportion of correctly predicted negative samples to all negative samples, V_{PP} is the proportion of correctly predicted positive samples to all predicted positive samples, and V_{NP} is the proportion of correctly predicted negative samples to all predicted negative samples. The formulas are defined as

$$Acc = (TP + TN)/(TP + TN + FP + FN), \quad (6)$$

$$Sen = TP/(TP + FN), \quad (7)$$

$$Spec = TN/(TN + FP), \quad (8)$$

$$V_{\text{PP}} = TP/(TP + FP), \quad (9)$$

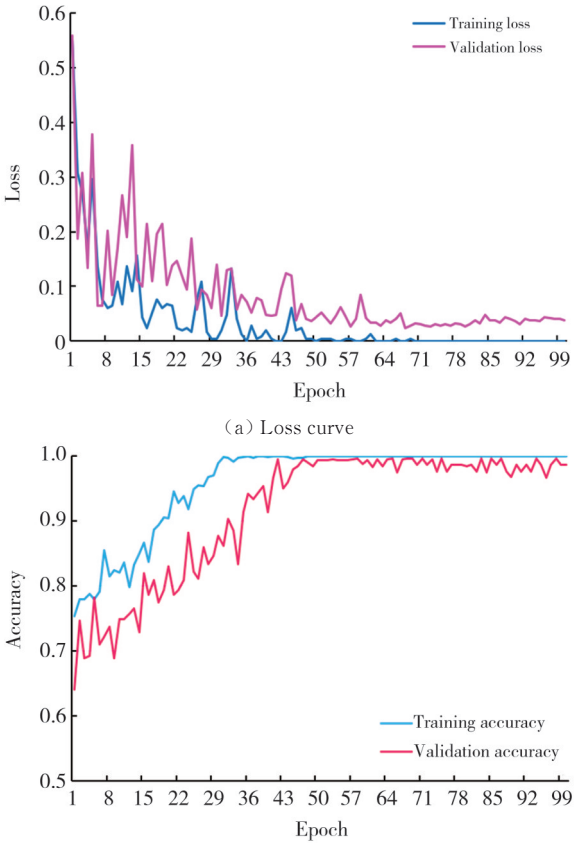
$$V_{\text{NP}} = TN/(TN + FN), \quad (10)$$

where TP is true positive, representing a correctly predicted positive result; TN is true negative, representing a correctly predicted negative result; FP is false positive, representing an incorrectly predicted negative result; and FN is false negative, representing an incorrectly predicted negative result.

The receiver operating characteristic (ROC) curve reflects the relationship between sensitivity and specificity and is used to assess the classification ability of a model at different probability thresholds. The ROC curve has false positive rate (R_{FP}) as the horizontal coordinate and sensitivity as the vertical coordinate, where $R_{\text{FP}}=1-spec$. However, the performance of the ROC is not sufficiently intuitive. AUC is the area under the ROC curve, which visually shows the classification ability of the model represented by the ROC curve. The AUC value is between 0 and 1; the larger the value, the better the model's classification performance.

3.2 Results and discussion

During training, the CXR image was resized to 224×224 and underwent random flipping and normalization. We conducted all experiments with a high-level Python wrapper of the Pytorch library. We used the Tesla V100 Graphics Card for training. During training, we set the batch size to 10, used the Adam optimizer[26] for optimization, and used cosine annealing[27] to adjust the learning rate. Fig. 6 shows the training and validation curves and the corresponding accuracy curves of MIM-Net in one cross-validation. The horizontal coordinate in both Fig. 6(a) and Fig. 6(b) are epochs, the vertical coordinate in Fig.6(a) is the loss value, and the vertical coordinate in Fig.6(b) is the accuracy value. As seen in Fig.6, the model converged after epoch 50 and overfitted after epoch 60. The model of epoch 50 was used as the final test model in this experiment.



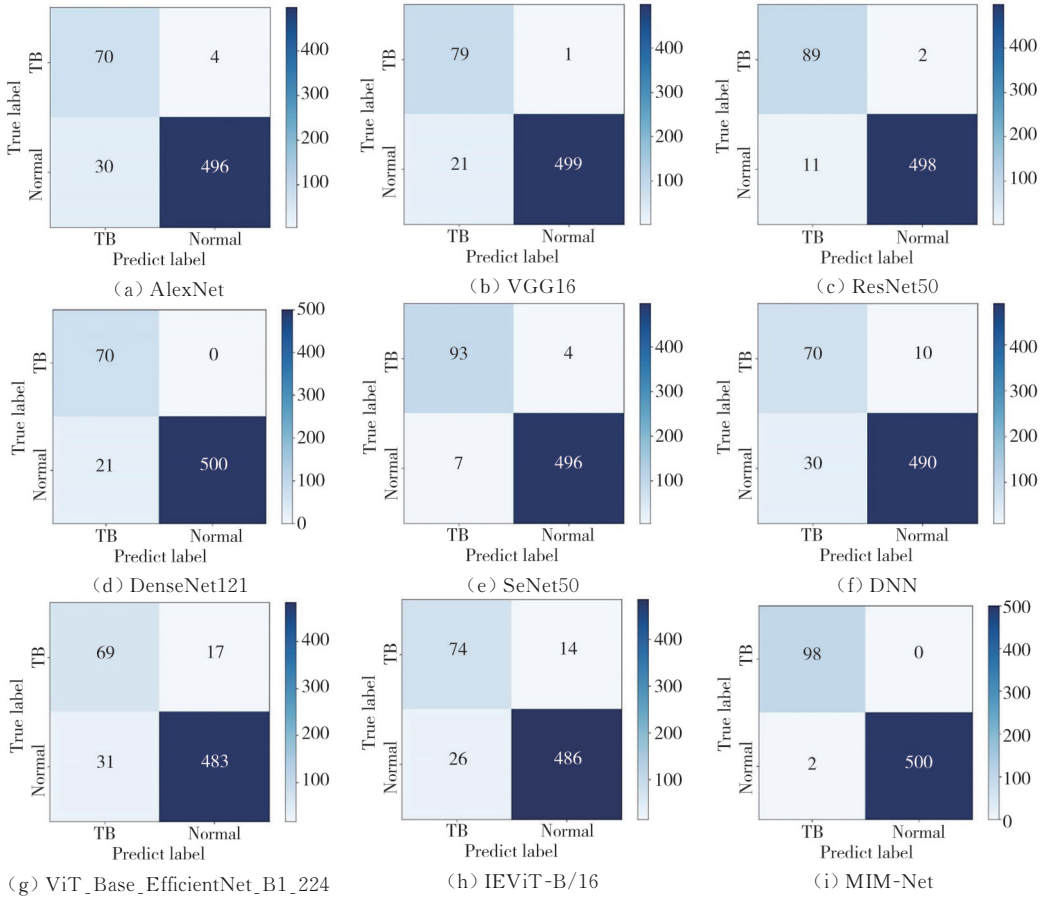
(a) Loss curve

(b) Accuracy curve

Fig. 6 Variation of the curve of MIM-Net in the experiment

3.2.1 Comparison experiments

We compared the performance of different classification models on the TB CXR dataset, including AlexNet[28], VGG16[29], ResNet50[30], DenseNet121[31], SeNet50[32], DNN[19], ViT_Base_EfficientNet_B1_224[21], IEViT-B/16[22], and MIM-Net. For a fair comparison, we performed a 5-fold cross-validation approach on all the different classification models. We set the different classification models with the same hyperparameters, and all models were not pre-trained. All models run 100 epochs per training. We trained the convolution-based models using a single graphics card, with AlexNet and DNN taking about 1 h per training, VGG16 and ResNet50 taking about 2 h per training, and DenseNet121 and SeNet50 taking about 3 h per training. We trained the two ViT-based models simultaneously on 4 graphics cards, each taking about 10 h per training. We selected the best validation model for each classification model in the 5-fold cross-validation approach and tested it on the test set. Fig. 7 shows the confusion matrix of the test results of different classification models, including *TP*, *TN*, *FP*, and *FN*. In the confusion matrix, we horizontally define the true labels of positive and negative samples and vertically define the predicted results of positive and negative samples.



(a) AlexNet (b) VGG16 (c) ResNet50

(d) DenseNet121 (e) SeNet50 (f) DNN

(g) ViT_Base_EfficientNet_B1_224 (h) IEViT-B/16 (i) MIM-Net

Fig. 7 Confusion matrix for different classification models

Table 1 shows the test results of each model on the TB CXR dataset, and Fig.8 shows the ROC curves for each model test result.

In Table 1, GPU refers to the forward pass time on a single GPU (Graphics Processing Unit, NVIDIA Tesla V100), where the forward pass time is the average test result of 600 all test images in milliseconds. CPU refers to the forward pass time on the CPU (Central Processing Unit, Intel Core i5-8400@4.00 GHz), and the forward pass time is the average test result of 100 test images in seconds. *Params* are the number of parameters of different models.

Table 1 Test results of different classification models

Model	*Acc*/%	*Sen*/%	*Spec*/%	V_{PP}/%	V_{NP}/%	AUC/%	GPU/ms	CPU/s	*Params*/($\times 10^6$)
AlexNet[29]	94.33	94.59	94.30	70.00	99.20	97.48	4.14	0.26	2.4
VGG16[30]	96.33	98.75	95.96	79.00	99.80	97.61	10.93	0.68	14.7
ResNet50[31]	97.83	97.80	97.84	89.00	99.60	98.64	10.69	0.71	23.5
DenseNet121[32]	96.50	100	95.97	79.00	100	99.14	14.23	0.85	7.0
SeNet50[33]	98.17	95.88	98.61	93.00	99.20	99.31	15.83	1.01	36.5
DNN[20]	98.33	100	98.04	90.00	100	97.34	5.96	0.38	3.0
ViT_Base_EfficientNet_B1_224[22]	92.00	80.23	93.97	69.00	96.60	95.85	30.65	2.15	94.8
IEVit-B/16[23]	93.33	84.09	94.92	74.00	97.20	96.17	43.37	3.03	90.8
MIM-Net	99.67	100	99.60	98.00	100	99.99	15.13	0.95	4.6

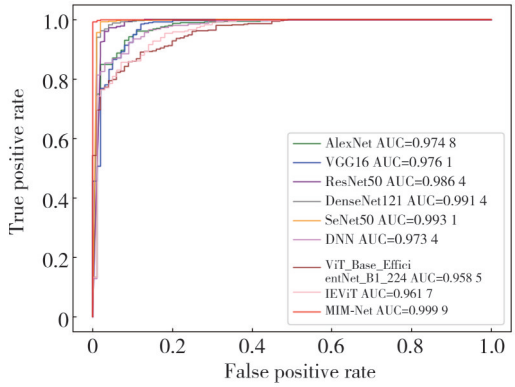


Fig. 8 ROC curves for different classification models

Experimental results show that MIM-Net performs best among all classification models, with the highest accuracy, sensitivity, specificity, V_{PP}, V_{NP}, and AUC. As seen from the first three rows in Table 1, the deeper network provides a better image feature extractor; the comparison between ResNet50 and DenseNet121 shows that some are overfitting in the network with too much depth. Thus, we use convolution with different growth rates on dense blocks on each branch of the MIM-Net as the network's backbone for image feature extraction. In addition, the ViT structure utilizes the global receptive field of the image, whereas the convolution utilizes the local receptive field of the image. Convolutional-based models can focus more on the lesions in a medical image, so ViT-based models do not perform as well as convolution-based models in this task. Besides, MIM-Net has a sensitivity of 100 percent for TB image detection, which is important because we want to detect as many TB cases as possible because they are contagious.

On the GPU, MIM-Net's forward pass time is 15.13 ms, or 66 frames per second (FPS), which is almost real-time performance; on the CPU, MIM-Net's forward pass time is 0.95 s, which is also acceptable in clinical diagnostics. From Table 1, it can be seen that the model inference time is not proportional to the number of parameters, and the actual efficiency of the model is not linearly related to the number of parameters. Other factors, such as memory access cost and model structure, also affect model inference time.

In addition to quantitative metrics, we evaluated our method using Grad-CAM[34] to visualize the testing process results. Fig.9 shows the Grad-CAM results of different networks on the input images during the test. In Fig.9, the first column is the input image, and the second, third, and fourth columns are the Grad-CAM results for VGG16, ResNet50, Densenet121, and MIM-Net, respectively. The Grad-CAM highlights are the main focus and judgment of the network. It can be seen that the Grad-CAM highlights of MIM-Net are more concentrated in the lungs of the CXR image, which indicates the network focuses more on the location of the lungs. This coincides with the practice of human doctors who focus on the site of the lesion during the diagnostic process.

3.2.2 Ablation experiments

Here, we conducted an ablation analysis of the MIM-Net to better understand the relative importance of its multiple aspects. Similarly, all networks performed 5-fold cross-validation on the TB CXR dataset, and the best results were selected for testing. The test results on the TB CXR dataset are shown in Table 2, and as in Table 1, the evaluation metrics used are accuracy, sensitivity, specificity, V_{PP}, V_{NP}, and AUC.

From the first row of Table 2, it can be seen that the performance of MIM-Net decreases most significantly without multi-scale input, which indicates that using different scale inputs and high-dimensional fusion is an effective feature extraction method for medical images

that are not rich in semantic information. It is worth noting that the MIM-Net with only one branch input and performing 5 downsampling is structurally DenseNet101, and comparing it with DenseNet121 in Table 1, it is clear that the classification performance of DenseNet101 is better, which also suggests that too deep networks may be overfitted when the training data is limited. The second row of Table 2 shows that replacing the mirror loss of the MIM-Net with cross-entropy also leads to a decrease in network performance. This indicates that the mirror loss allows the network to extract features better during the same training process without increasing the parameters. The last row of Table 2 shows that dense connections also benefit network performance but do not impact the network as much as the other two factors.

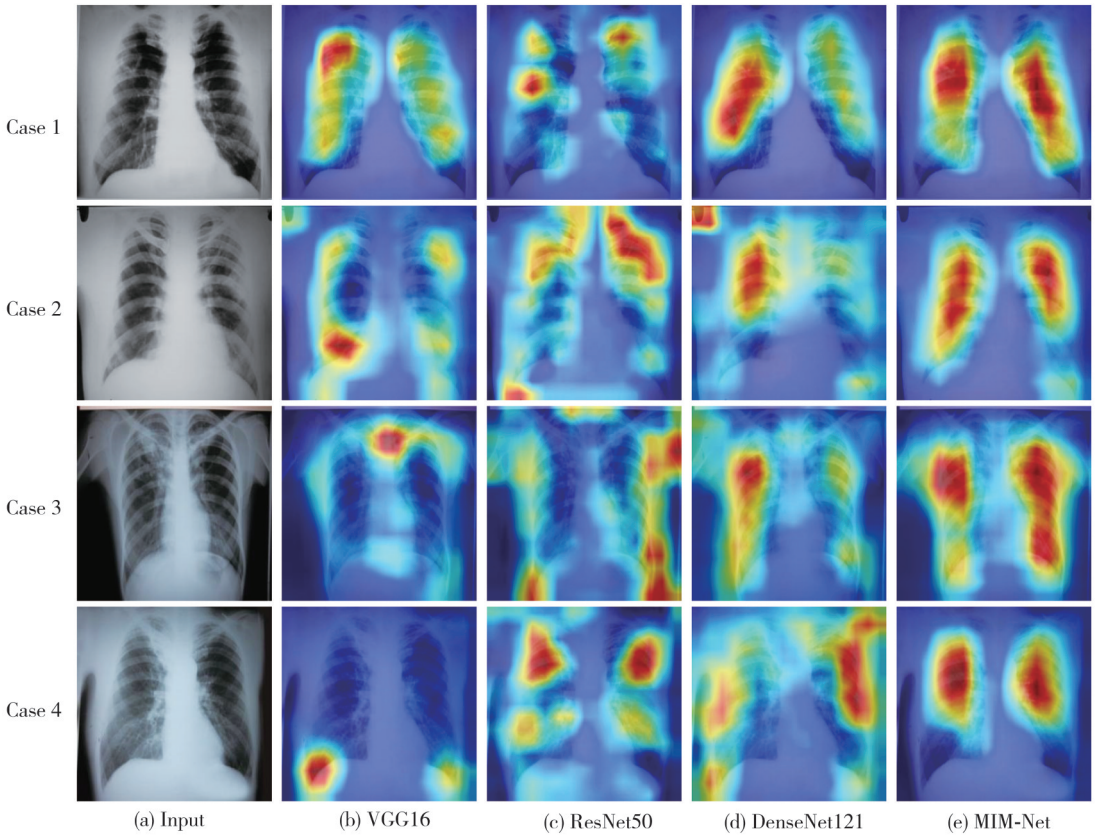


Fig. 9 Grad-CAM comparison for different classification networks

Table 2 Test results of MIM-Net ablation network

Network	*Acc*/%	*Sen*/%	*Spec*/%	V_{PP}/%	V_{NP}/%	AUC/%
Only one branch input	96.83	100	96.34	81.00	100	99.06
With cross-entropy loss	98.33	100	98.04	90.00	100	99.27
No dense connections	98.83	98.95	98.81	94.00	99.80	99.75

4 Conclusions

In this study, we proposed a deep learning method named MIM-Net for TB detection. Compared to other classification models, our proposed MIM-Net showed the best performance. The multi-scale input in MIM-Net enhances the network's feature extraction capability. The mirror loss utilizes the symmetric features in the image for self-supervised training to improve the network's performance. However, our CXR dataset comes from publicly available sources, and image quality will vary, which may lead to differences between study results and practical applications. Therefore, we will expand the dataset by adding more images in subsequent studies, and a broader dataset will better reflect the ability of the model representation. However, the high cost of data labeling remains a challenge. Therefore, in future work, we will use semi-supervised learning methods to create models and keep improving them to classify images better.

Acknowledgement

This work was supported by the Joint Fund of the Ministry of Education for Equipment Pre-research (No.8091B0203), and National Key Research and Development Program of China (No.2020YFC2008700).

Declaration of conflicting interests

The authors have no conflict of interest related to this

publication.

References

[1] TASCI E. Pre-processing effects of the tuberculosis chest X-ray images on pre-trained CNNs: an investigation// Artificial Intelligence and Applied Mathematics in Engineering Problems, April 20-22, 2019, Antalya, Turkey. New York: Springer International Publishing, 2020: 589-596.

[2] CHATTOPADHYAY S, KUNDU R, SINGH P K, et al. Pneumonia detection from lung X-ray images using local search aided sine cosine algorithm based deep feature selection method. International Journal of Intelligent Systems, 2022, 37(7): 3777-3814.

[3] ZENG J B, LIU Z, SHEN G L, et al. MRI evaluation of pulmonary lesions and lung tissue changes induced by tuberculosis. International Journal of Infectious Diseases, 2019, 82: 138-146.

[4] JAIN S K, ANDRONIKOU S, GOUSSARD P, et al. Advanced imaging tools for childhood tuberculosis: potential applications and research needs. The Lancet Infectious Diseases, 2020, 20(11): e289-e297.

[5] DHOOT R, HUMPHREY J M, O′MEARA P, et al. Implementing a mobile diagnostic unit to increase access to imaging and laboratory services in western Kenya. BMJ Global Health, 2018, 3(5): e000947.

[6] LANGE C, MORI T. Advances in the diagnosis of tuberculosis. Respirology, 2010, 15(2): 220-240.

[7] QIN C L, YAO D M, SHI Y H, et al. Computer-aided detection in chest radiography based on artificial intelligence: a survey. Biomedical Engineering Online, 2018, 17(1): 113.

[8] KIM T K, YI P H, WEI J C, et al. Deep learning method for automated classification of anteroposterior and posteroanterior chest radiographs. Journal of Digital Imaging, 2019, 32(6): 925-930.

[9] LI X K, ZHOU Y K, DU P, et al. A deep learning system that generates quantitative CT reports for diagnosing pulmonary Tuberculosis. Applied Intelligence, 2021, 51(6): 4082-4093.

[10] LAKHANI P, SUNDARAM B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology, 2017, 284(2): 574-582.

[11] YI P H, KIM T K, WEI J C, et al. Automated semantic labeling of pediatric musculoskeletal radiographs using deep learning. Pediatric Radiology, 2019, 49(8): 1066-1070.

[12] KANT S, SRIVASTAVA M M. Towards automated tuberculosis detection using deep learning//2018 IEEE Symposium Series on Computational Intelligence, November 18-21, 2018, Bangalore, India. New York: IEEE, 2018: 1250-1253.

[13] SANTOSH K C, ANTANI S. Automated chest X-ray screening: can lung region symmetry help detect pulmonary abnormalities? IEEE Transactions on Medical Imaging, 2018, 37(5): 1168-1177.

[14] CHANDRA T B, VERMA K, SINGH B K, et al. Automatic detection of tuberculosis related abnormalities in Chest X-ray images using hierarchical feature extraction scheme. Expert Systems with Applications, 2020, 158: 113514. New York: IEEE, 2021: 408-413.

[15] SONI A, RAI A, AHIRWAR S K. Mycobacterium tuberculosis detection using support vector machine classification approach//2021 10th IEEE International Conference on Communication Systems and Network Technologies, June 18-19, 2021, Bhopal, India. New York: IEEE, 2021: 408-413.

[16] LI L J, HUANG H Y, JIN X Y. AE-CNN classification of pulmonary tuberculosis based on CT images//2018 9th International Conference on Information Technology in Medicine and Education, October 19-21, 2018, Hangzhou, China. New York: IEEE, 2018: 39-42.

[17] AHSAN M, GOMES R, DENTON A. Application of a Convolutional Neural Network using transfer learning for tuberculosis detection//2019 IEEE International Conference on Electro Information Technology, May 20-22, 2019. Brookings, SD, USA. New York: IEEE, 2019: 427-433.

[18] HUANG C X, WANG W, ZHANG X, et al. Tuberculosis diagnosis using deep transferred EfficientNet. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2023, 20(5): 2639-2646.

[19] MAHBUB M K, BISWAS M, GAUR L, et al. Deep features to detect pulmonary abnormalities in chest X-rays due to infectious diseaseX: Covid-19, pneumonia, and tuberculosis. Information Sciences, 2022, 592: 389-401.

[20] YANG J, JIN Y X, LIU Y B, et al. Research on fabric material recognition method based on improved transformer. Journal of North University of China (Natural Science Edition), 2023, 44(2): 138-145.

[21] DUONG L T, LE N H, TRAN T B, et al. Detection of tuberculosis from chest X-ray images: Boosting the performance with vision transformer and transfer learning. Expert Systems with Applications, 2021, 184: 115519.

[22] OKOLO G I, KATSIGIANNIS S, RAMZAN N. IEViT: An enhanced vision transformer architecture for chest X-ray image classification. Computer Methods and Programs in Biomedicine, 2022, 226: 107141.

[23] IOFFE S, SZEGEDY C, PARANHOS L, et al. Batch normalization: accelerating deep network training by reducing internal covariate shift. 2015: 1502.03167. https://arxiv.org/abs/1502.03167v3.

[24] MURPHY K, SCHÖLKOPF B, SRIVASTAVA N, et al. Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[25] RAHMAN T, KHANDAKAR A, KADIR M A, et al. Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization. IEEE Access, 2020, 8: 191586-191601.

[26] KINGMA D P, BA J. Adam: A method for stochastic optimization. ArXiv preprint:1412.6980, 2014.
[27] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization. ArXiv preprint: 1711.05101, 2017.
LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization. ArXiv e-Prints, 2017: arXiv: 1711.05101.
[28] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84-90.
[29] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition. 2014: 1409.1556. https://arxiv.org/abs/1409.1556v6.
[30] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.
[31] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 2261-2269.
[32] HU J, SHEN L, SUN G. Squeeze-and-excitation networks//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 7132-7141.
[33] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks *via* gradient-based localization. International Journal of Computer Vision, 2020, 128(2): 336-359.

用于CXR图像中结核病检测的多尺度输入镜像网络

邢广鑫[1], 樊晶晶[2], 郑叶龙[1], 赵美蓉[1*]

1. 天津大学 精密测试技术及仪器全国重点实验室, 天津 300072;
2. 中国人民解放军总医院 医疗保障中心药剂科, 北京 100853

摘　要： 计算机辅助诊断可用于检测结核病例，为放射科医生提供更准确、更高效的诊断解决方案。结核病胸部X光(Chest X-ray, CXR)图像中的各种干扰噪声是这一分类任务的主要挑战。本研究旨在提出一种结核病CXR图像检测的高性能模型，即基于CXR图像对称性的多尺度输入镜像网络(Multi-scale input mirror network, MIM-Net)，它由多尺度输入特征提取网络和镜像损失组成：多尺度图像输入可增强特征提取，而镜像损失则通过自监督提高网络性能。该模型在一个公开的结核CXR图像分类数据集上通过5倍交叉验证进行了评估，其准确率、灵敏度、特异性、阳性预测值、阴性预测值和曲线下面积(Area under curve, AUC)分别达到了99.67%、100%、99.60%、98.00%、100%和0.999 9。与其他模型相比，MIM-Net在所有指标上都表现最佳。因此，我们提出的MIM-Net可以有效帮助网络学习更多特征，并可用于检测CXR图像中的结核病，从而帮助医生做出诊断。

关键词： 计算机辅助诊断；医学图像分类；深度学习；特征对称；镜像损失函数

引用格式： XING Guangxin, FAN Jingjing, ZHENG Yelong, et al. Multi-scale input mirror network for tuberculosis detection in CXR image. Journal of Measurement Science and Instrumentation, 2025, 16(1): 1-10. DOI: 10.62756/jmsi.1674-8042.2025001