IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 1 April 2023, accepted 16 April 2023, date of publication 21 April 2023, date of current version 1 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3269068

RESEARCH ARTICLE

Multi-Level Residual Feature Fusion Network for Thoracic Disease Classification in Chest X-Ray Images

QIANG LI[1], YU LAI[1], MOHAMMED JAJERE ADAMU[1], (Member, IEEE), LEI QU[2], JIE NIE[3], (Member, IEEE), AND WEIZHI NIE[4], (Member, IEEE)

[1]School of Microelectronics, Tianjin University, Tianjin 300072, China
[2]Hisense, Hisense Group Holdings Company Ltd., Qingdao 266071, China
[3]School of Electrical and Information Engineering, Ocean University of China, Qingdao 266100, China
[4]School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China

Corresponding author: Weizhi Nie (weizhinie@tju.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61471263 and Grant 62272337; in part by the Natural Science Foundation of Tianjin, China, under Grant 16JCZDJC31100; and in part by the Tianjin University Innovation Foundation under Grant 2021XZC-0024.

ABSTRACT Automated identification of thoracic diseases from chest X-ray images (CXR) is a significant area in computer-aided diagnosis. However, most existing methods have limited ability to extract multi-scale features and accurately capture the spatial location of lesions when dealing with thoracic diseases that exhibit concurrency and large variations in lesion size. Based on the above problems, we propose a multi-level residual feature fusion network (MLRFNet) for classifying thoracic diseases. Our approach can quickly capture receptive field information across different lesion sizes and enhance disease-specific features within the spatial domain on feature maps. The MLRFNet comprises two main components: a feature extractor that learns multi-scale semantic information from chest X-ray images and a multi-level residual feature classifier (MRFC) that refines disease-specific pathological features at spatial locations to reduce interference from irrelevant regions. Additionally, the ECA attention modules connect both components to enable flexible channel-wise focus on critical pathological information. We evaluated the performance of MLRFNet through a series of experiments on two datasets: ChestX-Ray14 and CheXpert. Our results show that MLRFNet achieves an average AUC of 0.853 on the ChestX-Ray14 dataset and 0.904 on the CheXpert dataset. The results of experiments demonstrate that our proposed method works better than the current state-of-the-art baselines. Future work will focus on investigating the interdependencies among labels for thoracic diseases and techniques for model compression.

INDEX TERMS Chest X-ray image classification, convolutional neural network, attention mechanism, residual feature vector, medical image processing.

I. INTRODUCTION

Clinical analysis of CXR images is one of the most critical methods for screening thoracic diseases such as pulmonary nodules, pneumonia, and pneumothorax. The hospital receives a large number of patients and creates considerable CXR image data every year. Traditional diagnostic methods rely on manual labeling by professional doctors, which is time-consuming and laborious. Computer-aided diagnosis (CAD) is significant because it makes doctors' jobs easier and reduces the number of misdiagnoses. Researchers have paid much attention to automatic medical image analysis technology based on deep learning in recent years. In particular, convolutional neural networks (CNN) have gradually become the preferred method in medical image segmentation [1], [2], [3], detection [4], [5], [6], and classification [7], [8], [9] tasks. It can learn potential pathological information from large datasets and automatically identify

The associate editor coordinating the review of this manuscript and approving it for publication was Kumaradevan Punithakumar.

thoracic diseases in CXR images to help doctors make clinical diagnoses. Many CNN-based classification methods for thoracic diseases have been proposed and can be broadly categorized into the following types: (1) Network structure optimization [10], [11], [12], [13], [14]; (2) Attention-guided methods [15], [16], [17], [18], [19]; (3) Correlation modeling methods [20], [21], [22], [23], [24].

Methods based on network structure optimization aim to improve feature extraction capabilities by refining the network structure to obtain richer feature representations from CXR images. Chen et al. [10] proposed a new double asymmetric feature learning network (DualCheXNet) to realize feature-level and decision-level fusion in structure. Rajpurkar et al. [12] proposed CheXNet for detecting thoracic diseases by fine-tuning the modified DenseNet121 [25]. However, the performance gains of these methods are limited due to their lack of attention to key pathological information. Attention-guided methods can direct the model's focus toward salient pathological features and have demonstrated notable success in the thoracic disease classification task. As a typical example, Wang et al. [16] proposed a triple attention network (A^3Net) for channel-level, element-level, and scale-level attention learning in feature extraction. However, this method is susceptible to the influence of incorrect labels and noisy regions, which may result in a decrease in model performance. Thoracic diseases exhibit comorbidity, and correlation modeling methods can effectively model the dependencies between diseases. Typically, Yao et al. [20] combined DenseNet and a long short-term memory network (LSTM) to learn the dependencies between target labels. Chen et al. [24] proposed a label co-occurrence learning framework based on graph convolutional neural networks (GNN) and CNN models to explore the correlations between pathological features. This method requires leveraging other model structures on top of CNNs, making it relatively complex.

As shown in Fig. 1, the thoracic diseases have multiple lesion areas with different lesion sizes, which is one of the main challenges in the automatic analysis of CXR images. However, the methods mentioned above share a common shortage that fails to fully utilize semantic information at different downsampling stages in the process of classifying thoracic diseases. Many shallow features, such as spatial location information and texture of many diseases, have been ignored, resulting in the model's inability to effectively capture the regions where disease lesions occur. Meanwhile, the multi-scale feature extraction capabilities of the backbone network used in many works are limited, making it difficult to adapt to the varying sizes of lesions. As we can see in Fig. 1, the "Nodule" are often small, but "Cardiomegaly" and "Infiltration" can cover a large area. Therefore, enhancing the ability of the network to capture multi-scale information is beneficial for increasing the accuracy of thoracic disease recognition. Furthermore, Khan et al. [26] reviewed the CXR image datasets available in deep learning methods and found that most of them suffer from sample imbalance.

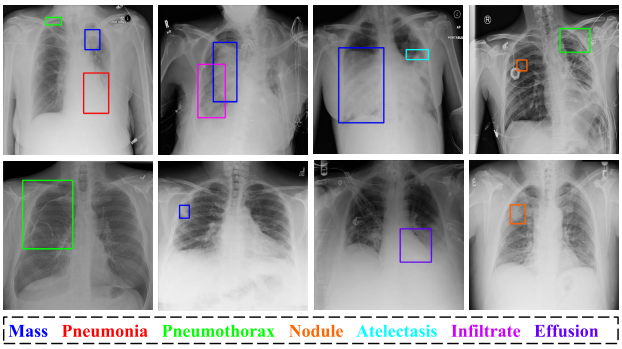


FIGURE 1. Example images of ChestX-Ray14. Thorax diseases may have multiple lesion areas and are variable in lesion size. The disease in the bounding box corresponds to the pathology name in the bottom row with the same color.

Imbalanced datasets can cause bias in the model's learning direction, resulting in poor classification performance.

To solve these issues, we propose a thoracic diseases classification network MLRFNet based on multi-level residual feature fusion. Firstly, to address the issue of complex morphological manifestations of different thoracic diseases in CXR images and the large differences in lesion sizes, Res2Net [27] is selected as the feature extraction network to obtain receptive field information from different lesion sizes and enhance the model's multi-scale feature extraction ability. Secondly, to mitigate the impact of variable lesion areas on the model's discriminative ability, a multi-level residual feature classifier (MRFC) is proposed to effectively capture the spatial location of lesions at different downsampling stages and convert them into residual feature vectors conducive to disease classification. Then, the efficient channel attention mechanism (ECA) [28] is introduced to enable the model to adaptively calibrate channel responses of feature maps and enhance key pathological features while suppressing the transmission of useless information. Finally, according to the characteristic that most disease categories in the dataset have more negative samples, a Biased Focal Loss (BFL) is proposed to increase the loss proportion of difficult-to-classify samples among negative samples and shift the optimization direction towards this part, thereby improving overall classification accuracy. Experiments show that MLRFNet achieves better results than the current state-of-the-art methods. We summarize the contributions of this work as follows:

- We propose a multi-level residual feature fusion network (MLRFNet) for the multi-label classification of thoracic diseases. Compared to the existing methods, our approach can quickly get receptive field information on different lesion sizes and improve disease-specific features at spatial locations to reduce the interference of irrelevant regions;
- We propose a novel multi-level residual feature classifier to generate classification vectors, which fully utilizes the

spatial location information of the disease in the feature map at different downsampling stages;
- We conduct a comprehensive experiment on the ChestX-Ray14 and CheXpert datasets. The results demonstrate that MLRFNet has better performance than the existing advanced models.

The rest of our document is organized as follows. Section II reviews the related works in the thoracic disease classification tasks. And Section III presents our approach in detail. We conduct comprehensive experiments in Section IV. A summary of the experimental results is also included in this Section. Section V discusses the purpose and achievements of this study, highlighting the limitations of the current work and suggesting potential future endeavors. In section VI, we draw a conclusion to this project.

II. RELATED WORK

The application of deep learning to medical image processing has made significant breakthroughs in recent years. The release of the public ChestX-Ray14 [29] and CheXpert [11] datasets makes more and more researchers pay attention to the automatic analysis technology of CXR images based on deep learning. This section will review state-of-the-art methods for classifying thoracic disease and some valuable works currently in this field.

A. NETWORK STRUCTURE OPTIMIZATION

Optimizing or fine-tuning the structure of CNN models can help to obtain more comprehensive feature representations and was the initial research direction for researchers. Wang et al. [29] first released a large CXR image dataset ChestX-Ray14 and studied the performance of redesigned ImageNet model [30] for thoracic diseases classification, such as AlexNet [31], GoogleNet [32], and ResNet [33]. Subsequently, Rajpurkar et al. [12] detected 14 diseases in the ChestX-Ray14 dataset by fine-tuning the modified DenseNet121 [25] and calling it CheXNet. It is important to note that CheXNet was better at finding pneumonia than professional radiologists. Chen et al. [10] proposed a new double asymmetric feature learning network (DualCheXNet) based on ResNet and DenseNet for multi-label classification of thoracic diseases, which realized feature-level fusion and decision-level fusion in structure. Hashmi et al. [13] fine-tuned five classic CNN models using transfer learning and proposed a weighted classifier that combines the classification results of these CNN models, which achieved high accuracy in identifying pneumonia. Irvin et al. [11] presented the large dataset CheXpert and investigated different policies to process the uncertainty labels for training CNNs. Moreover, they reported the ensemble result from 30 models on the validation set. Jiang et al. [14] proposed a new variant of the Pyramid Vision Transformer (MXT) [34] for multi-label CXR image classification, which can capture visual information at short and long-range in CXR images through self-attention. However, the above methods only use mainstream deep learning networks to extract pathological features from CXR images and are easily affected by image noise and irrelevant regions. The MLRFNet proposed in this paper utilizes the ECA attention mechanism and the MRFC module to enhance the model's attention to critical features in channels and space, which is conducive to improving classification accuracy.

B. ATTENTION-GUIDED METHODS

As the attention mechanism has become more influential in computer vision, researchers have tried to introduce it into medical image processing. The attention mechanism can guide the deep learning model to focus on the lesion area in the CXR image. Wang et al. [16] proposed a triple attention network (A^3Net). Specifically, A^3Net utilizes pre-trained DenseNet121 as the backbone network for feature extraction and integrates three attention modules into a unified framework for channel-level, element-level, and scale-level attention learning. Guan et al. [18] proposed a network ConsultNet which uses a novel variational selective information bottleneck (VSIB) to focus on the disease-correlated regions. Chen et al. [15] proposed an attention-guided network LLAGnet to focus on the discriminative features from lesion location, which combines region-level attention (RLA) and channel-level attention (CLA). Zhu et al. [17] proposed a pixel classification and attention network (PCAN) to simultaneously perform disease classification and weakly supervised localization, providing interpretability for disease classification. Chen et al. [19] proposed a new network PCSANet for thoracic disease classification and COVID-19 detection based on pyramidal convolution and shuffle attention module. Although the above methods use attention mechanisms to guide the model's focus on key features, they rely solely on the output of the final feature map from a CNN during classification and lack the utilization of semantic information at different levels. This limits the improvement of classification accuracy.

C. CORRELATION MODELING METHODS

In multi-label thoracic disease classification tasks, modeling and analyzing dependencies between thoracic diseases can help to improve the model's recognition ability. The researchers also achieved good results in this area. Yao et al. [20] combined DenseNet and a long short-term memory network (LSTM) to learn the dependencies between target labels. Subsequently, Graph Neural Networks (GNNs) garnered considerable interest among researchers due to their robust ability to model relationships between node data. Chen et al. [24] proposed a label co-occurrence learning framework based on GNN and CNN models to explore the correlations between pathological features. To enable the model to leverage prior knowledge in diagnosing thoracic diseases like a professional radiologist, Chen et al. [21] introduced a Semantic Similarity Graph Embedding (SSGE) module designed to investigate the semantic similarity between images and optimize the feature extraction process.

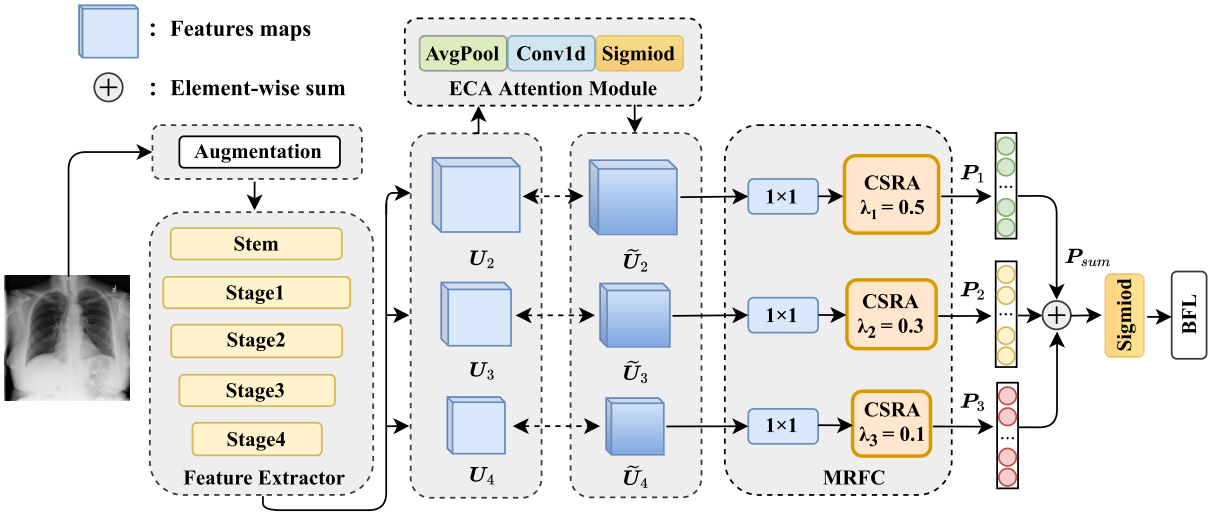


FIGURE 2. Overview of the proposed MLRFNet. The MLRFNet consists of a feature extractor, ECA attention modules, and a multi-level residual feature classifier (MRFC). First, we do data augmentation on a CXR image and then feed it into the feature extractor to get feature maps at different sampling stages. Then, we use attention modules to enhance disease potential critical features between channels adaptively. Finally, the MRFC utilizes feature maps of different sizes to generate residual feature vectors. The residual feature vectors are finally fused and added to achieve a joint decision.

Lee et al. [23] proposed a hybrid deep learning model (CheXGAT) based on CNN and graph convolution neural networks (GNN), which uses self-attention to aggregate domain features from graphical structures to enhance potential correlations between thoracic diseases. Jung et al. [22] proposed a novel framework called FGR-PCAM based on GNNs and CNNs that leverages graph structure to learn the relationships between lesion-specific features in localized regions. However, they still lack the utilization of semantic information at different downsampling stages.

D. OTHER VALUABLE WORKS

In addition to the aforementioned methods, there are other valuable research efforts focused on diagnosing thoracic diseases that warrant attention. For instance, some researchers have employed a combination of traditional machine learning and deep learning techniques to identify thoracic diseases. Rehman et al. [35] utilized CNN to extract deep features from CXR images and employed classical machine learning classifiers to process these features. This approach improves the accuracy of identifying COVID-19 and increases the predictability rates for other thoracic diseases. Khan et al. [36] employed VGG19 to extract deep features from CT images and concatenated them with handcrafted features to enhance the accuracy of pulmonary nodule identification. Furthermore, this method also employs image segmentation algorithms to extract lung nodule regions. A similar approach was used by Jaszcz et al. [37], who utilized a heuristic red fox algorithm to segment the lungs from CXR images before further processing. High-performance models generally require advanced hardware to run and may not be suitable for resource-constrained regions. Thus, the lightweight model is one of the areas of focus for researchers. Amirkhani et al. [38] trained a lightweight student model using multi-teacher distillation to improve the segmentation performance and robustness of the student model. Mahbub et al. [39] designed an easy-to-train and lightweight CNN model that achieved high accuracy in identifying COVID-19.

In contrast to the above methods, our proposed MLRFNet fully utilizes semantic information at different levels and

TABLE 1. The network structure of Res2Net50, where R2conv $k \times k$ block consists of a set of $k \times k$ convolution kernels connected in a hierarchical residual style as shown in Fig. 3. Conv $k \times k$ block stands for $k \times k$ convolution, Batch Normalization, and ReLU. Feature map represents the output tensor of different layers.

Layer	Output size	Operation	Feature map
Input	$224 \times 224 \times 3$		
Stem	$112 \times 112 \times 64$	Conv 7×7	U_0
Stage1	$56 \times 56 \times 64$	Max pool 3×3	U_1
	$56 \times 56 \times 256$	$\begin{bmatrix} \text{Conv } 1 \times 1 \\ \text{R2conv } 3 \times 3 \\ \text{Conv } 1 \times 1 \end{bmatrix} \times 3$	
Stage2	$28 \times 28 \times 512$	$\begin{bmatrix} \text{Conv } 1 \times 1 \\ \text{R2conv } 3 \times 3 \\ \text{Conv } 1 \times 1 \end{bmatrix} \times 4$	U_2
Stage3	$14 \times 14 \times 1024$	$\begin{bmatrix} \text{Conv } 1 \times 1 \\ \text{R2conv } 3 \times 3 \\ \text{Conv } 1 \times 1 \end{bmatrix} \times 6$	U_3
Stage4	$7 \times 7 \times 2048$	$\begin{bmatrix} \text{Conv } 1 \times 1 \\ \text{R2conv } 3 \times 3 \\ \text{Conv } 1 \times 1 \end{bmatrix} \times 3$	U_4

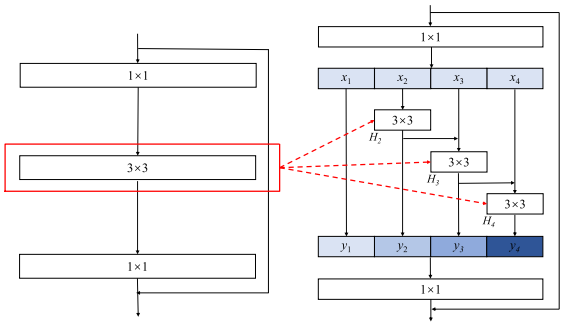


FIGURE 3. Comparison of ResNet block and Res2Net block. The left structure represents the ResNet block, and the right represents the Res2Net block. In the Res2Net block, feature splitting is processed multi-scale, which is conducive to extracting global and local information from CXR images.

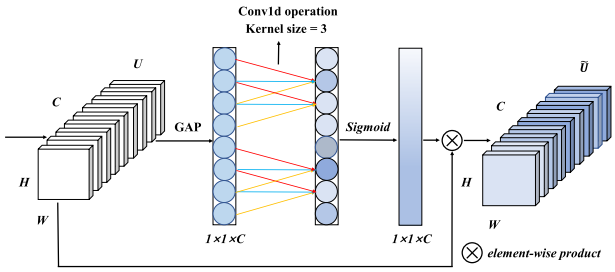


FIGURE 4. Diagram of the ECA attention mechanism. The aggregated information can be obtained by global average pooling. ECA learns the weights of channels by a fast 1D convolution, and the 1D convolution kernel size is set to 3 in the MLRFNet.

enhances the model's attention to specific disease features in spatial locations. This further improves the model's ability to identify thoracic diseases.

III. METHOD

This section introduces the details of the proposed multi-level residual feature fusion network (MLRFNet) for the multi-label CXR image classification. Fig. 2 shows the architecture of MLRFNet. We first describe the feature extractor in Sec. III-A and then introduce the ECA attention module and multi-level residual feature classifier in Sec. III-B and Sec. III-C, respectively. Finally, we present the optimization of the loss function in Sec. III-D.

A. FEATURE EXTRACTOR

Feature extractor aims to extract feature maps in different downsampling stages. CXR images show a significant difference between the lesion size of different diseases. Thus the neural network is required to have solid multi-scale feature extraction capability. We utilize Res2Net50 as the feature extractor, which can represent multi-scale features at a finer level of granularity. Fig. 3 shows the difference between the traditional bottleneck block in ResNet and the Res2Net block. Res2Net50 replaces a set of 3×3 convolution kernels with smaller convolution kernel groups while connecting different convolution kernel groups in a hierarchical residual style.

As shown in Table 1, the feature extractor consists of a 7×7 convolutional layer, a 3×3 max pooling layer, and four consecutive stages containing different numbers of Res2Net50 blocks. A given CXR image is preprocessed first and then fed into the feature extractor. After a series of convolution operations, the $H \times W \times C$ feature tensor from different downsampling stages can be obtained. The feature maps $\boldsymbol{U}_2$, $\boldsymbol{U}_3$, and $\boldsymbol{U}_4$ from Stage2, Stage3, and Stage4 are used as input of the ECA attention module.

B. ECA ATTENTION MODULE

During the training phase, the network should pay more attention to the relevant channel feature of the lesion. The attention mechanism can model dependencies between channels and calibrate the response characteristics for each channel. In this work, the ECA attention module is adapted to enhance the critical information for each class, which is an efficient and lightweight component. As shown in Fig. 4, the global average pooling is used to gather the global information of the feature map. Then, the ECA attention takes an efficient approach to learn the dependencies between channels. This strategy can be implemented by a 1D convolutional kernel with the size of k, which allows channels to share learning parameters. Finally, the sigmoid function is utilized to generate the weighting factors. The above process is implemented as follows:

$$\widetilde{\boldsymbol{U}} = \boldsymbol{U} \odot \text{sigmoid}[\text{C1D}_{\text{k}}(\text{GAP}(\boldsymbol{U})], \widetilde{\boldsymbol{U}} \in \mathbb{R}^{H \times W \times C}, \quad (1)$$

where GAP represents the global average pooling layer, and C1D_{k} stands for 1D convolutional operation. The kernel size k is set to 3 in this work. $\odot$ denotes the channel-wise multiplication. After the above operation, each channel of the input feature map $\boldsymbol{U}$ is assigned a different weight, which enhances relevant pathological information and suppresses irrelevant information transfer.

C. MULTI-LEVEL RESIDUAL FEATURE CLASSIFIER (MRFC)

On a CXR image, there is often more than one lesion area, so one or more pathologies are assigned based on what they mean. We hope that the model not only focuses on the global information of the image but also on the spatial location of the lesions in the feature map. At different stages of CNN's feature extraction, MRFC can make the network pay more attention to the critical positions of certain diseases and pay less attention to the irrelevant parts.

MRFC consists of three CSRA modules [40] with different λ. CSRA has achieved satisfactory results in the multi-label image classification task of natural images. It uses the spatial pooling operation to make a simple spatial attention score map that describes the spatial feature information of a certain category. Then the global average pooling, independent of the specific category, combines itself with the spatial pooling to generate the output vector. The working principle of CSRA is shown in Fig. 5.

In MRFC, the feature vector $\widetilde{U}$ from the ECA attention module undergoes dimensionality reduction via a 1×1 convolution operation to a size of $H \times W \times N$, where N represents the number of disease categories under consideration. Subsequently, the resulting feature map is fed into the CSRA module, where it is decoupled into characteristic tensors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3, \ldots, \boldsymbol{\mu}_{HW} (\boldsymbol{\mu}_j \in \mathbb{R}^{1\times1\times N})$. Then, after the softmax operation, the corresponding probability of locations with large pixel values in the feature map is further increased, and we can get the spatial attention scores of specific categories in the j_{th} position. Using the attention score as the weight value, we can get the results of spatial pooling, which is also called a residual classification vector. On the other hand, the CSRA module uses the global average pooling vector as the primary classification vector. Finally, the prediction score $\boldsymbol{P} \in \mathbb{R}^{1\times1\times N}$ of the diseases can be obtained:

$$\boldsymbol{P} = \frac{1}{HW}\sum_{j=1}^{HW}\boldsymbol{\mu}_j + \lambda\sum_{j=1}^{HW}\text{softmax}(T\boldsymbol{\mu}_j)\boldsymbol{\mu}_j, \tag{2}$$

where T is called the temperature coefficient, which controls the sharpness of a single position score. λ is the hyperparameter that controls the weight of the residual classification vector. To effectively enhance the valuable information on the spatial location of thorax disease, T is set to a typical value of 99 [40].

During the downsampling of CXR images by CNN, the feature map with higher resolution corresponds to a smaller receptive field. It retains more low-level features, such as the spatial position and texture of lesions. The loss of this original information is not conducive to the network's ability to recognize diseases.

To make better use of the spatial position information in the feature map, MLRFNet adopts a multi-level residual feature fusion method to predict 14 types of pathologies. As shown in Table 1 and Fig. 2, the shape of feature map output by Stage2, Stage3, and Stage4 is $28 \times 28 \times 512$, $14 \times 14 \times 1024$, and $7 \times 7 \times 2048$, respectively. They first pass through the ECA module to enhance the critical features between channels and then are sent to the CSRA modules corresponding to $\lambda_1 = 0.5$, $\lambda_2 = 0.3$, $\lambda_3 = 0.1$. The hyperparameter λ controls the weight of the residual features in the prediction vector. Finally, the prediction vectors $\boldsymbol{P}_1, \boldsymbol{P}_2$, and $\boldsymbol{P}_3$ output by three CSRA modules are added for a joint judgment of thoracic diseases: $\boldsymbol{P}_{\text{sum}} = \boldsymbol{P}_1 + \boldsymbol{P}_2 + \boldsymbol{P}_3$.

D. OPTIMIZATION OF LOSS FUNCTION

The uneven distribution of samples usually hinders the improvement of the accuracy of multi-label classification tasks. In the ChestX-Ray14, some pathologies like "Hernia" and "Pneumonia" tend to have fewer positive samples and more negative samples. Unbalanced distribution leads to serious classification difficulty with fewer positive samples, and CNN needs to learn more pathological information. To alleviate this problem, we optimized the focal loss [41] and called

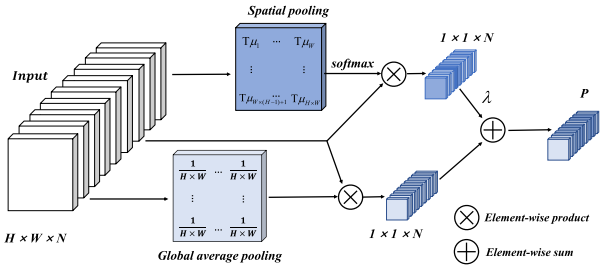


FIGURE 5. Diagram of class-specific residual attention module. The residual feature can be obtained by spatial pooling, which enhances the larger value among all spatial locations for specific diseases.

it the biased focal loss (BFL):

$$\begin{aligned}\text{BFL} = &-\alpha\left(1-\hat{\boldsymbol{P}}_{\text{sum}}\right)^{\beta}\boldsymbol{y}\log\hat{\boldsymbol{P}}_{\text{sum}} \\ &-(1-\alpha)(\hat{\boldsymbol{P}}^{s}_{\text{sum}})^{\beta}\,(1-\boldsymbol{y})\log\left(1-\hat{\boldsymbol{P}}_{\text{sum}}\right),\end{aligned} \tag{3}$$

where $\hat{\boldsymbol{P}}_{\text{sum}}$ is the predicted confidence score normalized by a sigmoid layer. The labels for each CXR are expressed as a one-shot vector $\boldsymbol{y} = [y_1, \ldots, y_i, \ldots, y_N]$, where N is 14 in ChestX-Ray14 and CheXpert. $y_i \in \{0, 1\}$ represents the ground truth of class i. $\beta > 0$ denotes the focus parameter. Compared to the focal loss, we further set the weight factor α to increase the contribution of negative samples to the loss function. In the experiment, α is set to 0.4. In addition, inspired by [42], probability shift factor s is set to make the network focus more on the hard-to-classify parts of negative samples:

$$\hat{\boldsymbol{P}}^{s}_{\text{sum}} = \max(\hat{\boldsymbol{P}}_{\text{sum}} - s, 0), \tag{4}$$

when the predicted score of a negative sample is between 0 and s, it is considered an easy-to-classify negative sample and is thresholded to 0 after the probability shift operation. In BFL, the shift probability s is 0.2.

In summary, BFL makes the model pay more attention to the hard-to-classify parts of negative samples according to the sample distribution in the chest X-ray datasets and dynamically adjusts the learning direction, thereby improving the overall classification ability of the network.

IV. EXPERIMENTS

A. DATASET

In this section, we employ two publicly available datasets as evaluation benchmarks: ChestX-Ray14 released by the National Institutes of Health (NIH), and CheXpert released by researchers at Stanford University. Detailed information regarding dataset split criteria and the sample distribution for both datasets are presented in Table 2 and Table 3, respectively. It should be noted that only the training and validation sets of CheXpert are shown in Table 2 as its test set has not been made publicly available.

ChestX-Ray14 [29] contains 112,120 frontal-view X-ray images with 14 pathologies. Except for the 60,361 images labeled "No Finding", each image is assigned one or more of

TABLE 2. Dataset split criteria for ChestX-Ray14 and CheXpert.

Name	ChestX-ray 14			CheXpert	
Image / Patient Number (#)	112,120/30,805			224,316/65,240	
Abnormality Number	14			14	
Image Size	1024×1024			320×320	
Split	train	valid	test	train	valid
Image Number (%)	70	10	20	–	–
Image Number (#)	78,468	11,219	22,433	223,414	200

the 14 pathologies, and 880 images have been annotated with 984 labeled bounding boxes for eight types of pathologies. We evaluate performance on all 14 labels. For fairness, the dataset split in the comparative experiments strictly follows the official splitting standards of the dataset published by Wang et al. [29].

CheXpert [11] contains 224,316 X-ray scans of 65,240 patients, with 14 observations extracted from the medical reports. Each observation is assigned a positive (1), negative (0), or uncertain (-1). The validation set of CheXpert consists of 200 chest radiographic studies manually annotated by three board-certified radiologists. Same as Irvin et al. [11], we evaluate performance with five observations: ''Atelectasis'', ''Cardiomegaly'', ''Consolidation'', ''Edema'', and ''Pleural Effusion'' on the validation set. ''U-Ones'' and ''U-Zeros'' are policies to handle the uncertainty mentioned in Irvin et al. [11]. Specifically, ''U-Ones'' treats the uncertain labels as positive, while ''U-Zeros'' treats uncertain labels as negative.

B. EVALUATION METRICS

The receiver operating characteristics (ROC) curve is used to represent the algorithm's ability to identify each pathology, and the area under the ROC curve (AUC) value is calculated for quantitative analysis and comparison in this paper. In the ROC curve, FPR shows the percentage of negative classes that were wrongly thought to be positive classes in all negative classes. TPR shows how many of the positive classes were correctly identified out of all of the positive classes. FPR and TPR are precisely calculated as follows:

$$TPR = \frac{TP}{TP + FN},$$
$$FPR = \frac{FP}{FP + TN}. \tag{5}$$

Meanwhile, the average Accuracy, Sensitivity, Specificity, and F1-score are considered additional evaluation metrics, which can be expressed as:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN},$$
$$Sensitivity = \frac{TP}{TP + FN},$$
$$Specificity = \frac{TN}{FP + TN},$$
$$F1 - score = \frac{2 \times TP}{2 \times TP + FP + FN}, \tag{6}$$

TABLE 3. The detailed sample distribution for pathologies in the ChestX-Ray14 (Xray14) and CheXpert (Xpert) datasets. In this table, Pos., Neg., and Unc. represent the number of positive, negative, and uncertain samples.

Xray14	Pos.	Neg.	Xpert	Pos.	Unc.	Neg.
Atel	11,559	100,561	NoFi	16,627	0	207,689
Card	2,776	109,344	EnCa	9,020	10,148	205,148
Effu	13,317	98,809	Card	23,002	6,597	194,717
Infi	19,894	92,226	Lesi	6,856	1,071	216,389
Mass	5,782	106,338	Opac	92,669	4,341	127,306
Nodu	6,331	105,789	Edem	48,905	11,571	163,840
Pneu1	1,432	110,688	Cons	12,730	23,976	187,610
Pneu2	5,302	106,818	Pneu1	4,576	15,658	204,082
Cons	4,667	107,453	Atel	29,333	29,377	165,606
Edem	2,303	109,817	Pneu2	17,313	2,663	204,340
Emph	2,516	109,604	Effu	75,696	9,419	139,201
Fibr	1,686	110,434	Other	2,441	1,771	220,104
P_T	3,385	108,735	Frac	7,270	484	216,562
Hern	227	111,893	Devi	105,831	898	117,587

* The 14 pathologies in ChestX-Ray14 are Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening, and Hernia, respectively.
* The 14 observations in CheXpert are No Finding, Enlarged Cardiomediastinum, Cardiomegaly, Lung Lesion, Lung Opacity, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture, and Support Devices, respectively.

where FP, TN, TP, and FN represent false positives, true negatives, true negatives, and false negatives, respectively.

Moreover, we also calculated the floating-point operations (FLOPs) and testing time required for the model to process a single CXR image during the testing phase to measure the implementation cost.

C. IMPLEMENTATION DETAILS

A reasonable experimental setup and data augmentation strategy can effectively improve the final classification performance of the model. The implementation details during the training process will be introduced in this part.

1) EXPERIMENTAL SETUP

As shown in Table 4, the experiment is implemented on the Pytorch [43]. For training, we optimize the network by Adam [44] optimizer with a batch size of 32 and train for 20 epochs. The initial learning rate is 0.0001, and the learning rate is multiplied by 0.9 every two epochs. To improve the convergence speed as well as the learning ability of the model, the backbone networks in the experiments will be pre-trained models on ImageNet. Training will be stopped when the loss on the validation set no longer decreases or starts to increase.

2) DATA AUGMENTATION

During the data preprocessing stage, we perform data augmentation on the input CXR images. The detailed methods and steps are as follows:

TABLE 4. Experimental setup details in training phrases.

Implementation	Value
Framework	Pytorch
Batch Size	32
Initial Learning Rate	1×10^{-4}
Optimizer	Adam
Number of Epochs	20 (Early stop)
Weight Initialization	Pretrained

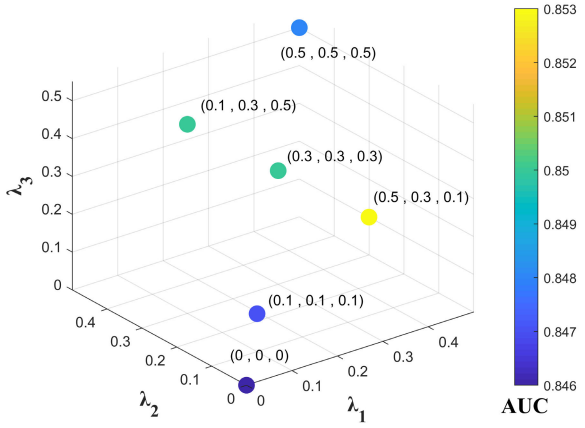


FIGURE 6. Three-dimensional scatter plots of different λ combinations. The depth of the color corresponds to the size of the AUC value, representing the impact of different hyperparameter combinations on the classification accuracy of the ChestX-Ray14 test set.

(1) Resizing the image to 256×256;
(2) Randomly crop the image to 224×224;
(3) Flip the image horizontally with a probability of 0.5;
(4) Randomly rotate the image within the range of $[-5, 5]$ degrees;
(5) Set the contrast, saturation, and hue of the image to vary between 90% and 110%;
(6) Totensor and Normalization.

D. HYPER-PARAMETRIC ANALYSIS

In this section, we will analyze the hyperparameters from two aspects: the impact of λ in MRFC on model classification accuracy and the impact of different learning rates during training on loss optimization. Since the ChestX-Ray14 dataset provides a complete and publicly available test set, we conducted our hyper-parametric experiment on it.

1) EFFECTS OF λ COMBINATION

In the MRFC, the hyperparameter λ controls the weight of residual features in the classification vector. The pathological information will vary with the resolution of the feature map. Therefore, the class-specific residual classification block should set different λ according to the size of the feature map. In order to verify the impact of the λ on the classification results, we set different value combinations to train the network while keeping the other settings unchanged. The experimental results are shown in Fig. 6.

We selected 0.1, 0.3, and 0.5 as λ candidate values, and λ_1, λ_2 and λ_3 correspond to the three CSRA modules processing feature maps of $28 \times 28 \times 512$, $14 \times 14 \times 1024$ and $7 \times 7 \times 2048$ sizes, respectively. As seen from Fig. 6, different combinations of λ greatly impact the classification results. Among them, the best classification results are achieved when $\lambda_1 = 0.5$, $\lambda_2 = 0.3$, and $\lambda_3 = 0.1$. We find that the high-resolution feature map has some low-level semantic information and the spatial location information of the lesion is not lost due to continuous downsampling. A larger value of λ ensures that more residual features are incorporated into the output classification vector. These residual features have a lot of information about where they are in spatial location, which helps the network pay more attention to certain types of diseases in the right places.

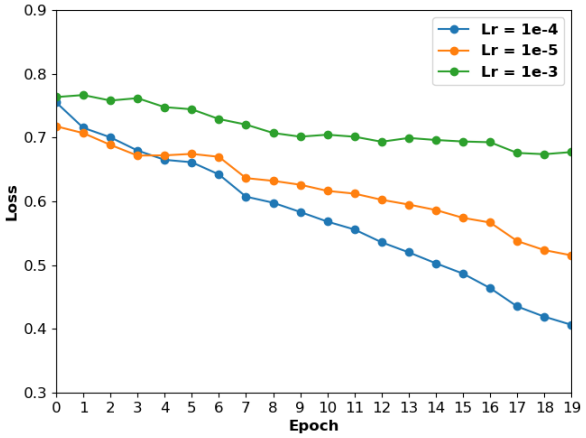


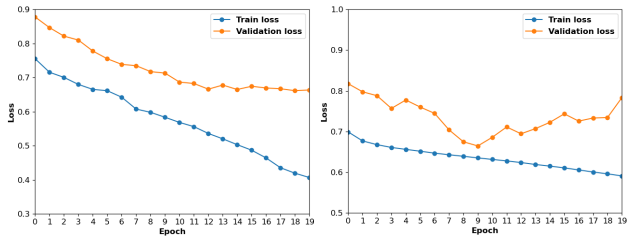
FIGURE 7. Relationship between learning rate and loss optimization in the ChestX-Ray14 training set.

2) EFFECTS OF LEARNING RATE

The learning rate is a crucial hyperparameter that determines the speed and outcome of model weight optimization. In our training process, we experimented with initial learning rates of 1×10^{-3}, 1×10^{-4}, and 1×10^{-5} to observe the changes in loss values over training epochs. As shown in Fig. 7, our experimental results indicate that when the learning rate is set to 1×10^{-3}, the loss curve ceases to decline from the 9th epoch and maintains a high loss value. Our analysis suggests that this is due to an excessively large learning rate causing the model to oscillate around a local optimum. On the other hand, when the learning rate is set to 1×10^{-5}, the model's loss curve steadily converges but at a relatively slow pace. However, with a learning rate of 1×10^{-4}, we observed the fastest convergence of the model. As such, we ultimately chose an initial learning rate of 1×10^{-4}.

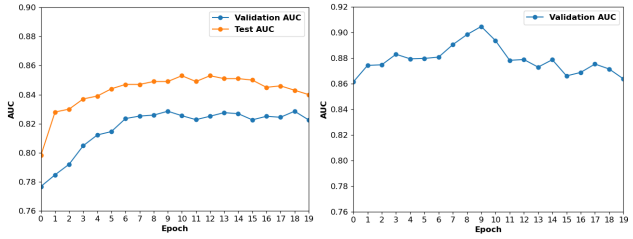
E. COMPARISON WITH SOTA METHODS

We compared the proposed MLRFNet classification network with the current state of the arts on the ChestX-Ray14



(a) ChestX-Ray14 (b) CheXpert

FIGURE 8. Illustration of the training and validation loss curves on the ChestX-Ray14 (a) and CheXpert (b) datasets.



(a) ChestX-Ray14 (b) CheXpert

FIGURE 9. Illustration of the validation and test AUC curves on the ChestX-Ray14 (a) and CheXpert (b) datasets.

and CheXpert datasets. The AUC score of each pathology and the average AUC scores across all classes are computed. Specifically, we selected the SOTA methods listed in Section II as baseline models for comparison, including (1) Network structure optimization, i.e. DCNN [29], Ensemble [11], and MXT [14]. (2) Attention-guided methods, i.e. LLAGNet [15], A^{3}Net [16], PCAN [17], ConsultNet [18], and PCSANet [19]. (3) Correlation modeling methods, i.e. SSGE [21], CheXGCN [24], CheXGAT [23], and F-PCAM [22].

1) RESULTS ON ChestX-Ray14

We first evaluated the performance of our model on the test set of the ChestX-Ray14 dataset. The ROC curves for each pathology are drawn in Fig. 10(a) to visually represent the classification performance of the proposed method. Table 5 presents the comparative results obtained by our MLRFNet and other SOTA baselines. Moreover, the loss and AUC curves during training and validation are given in Fig. 8(a) and Fig. 9(b), respectively.

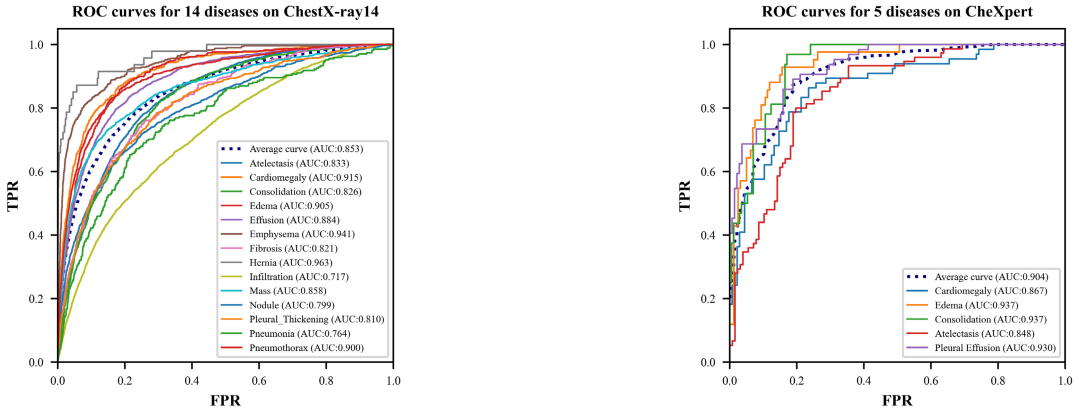
Compared with these SOTA methods, the MLRFNet proposed in this paper achieves better classification results: the average AUC score is 0.853 across the 14 thoracic diseases. The ROC curves for each pathology are located in the upper left corner, which indicates that the overall classification performance of MLRFNet is excellent. As can be seen from Fig. 8(a), the training set loss has been decreasing, while the validation set loss gradually stabilizes after the 10th epoch and no longer optimizes. Fig. 9(a) shows that the AUC value of the model on the test set reaches its maximum at the 10th epoch.

Based on the comparison results in Table 5, we can draw the following conclusions and analyses: (1) Our MLRFNet achieves the most advanced level for 10 of the 14 thoracic diseases in the classification experiments. The other two diseases, "Emphysema" (0.941) and "Fibrosis" (0.821), are also almost close to the best. It is worth mentioning that the accuracy of the proposed method for identifying "Consolidation" (0.826), "Atelectasis" (0.833), and "Hernia" (0.963) exceeds that of the best results of the baseline models above by 3.0%, 3.2%, and 2.1%, respectively, representing a significant improvement. Compared to the baseline models, our MLRFNet fully utilizes the residual features at different sampling stages and enhances the pathological features of diseases in spatial locations, thus achieving good results in the identification of most diseases; (2) Like most methods, our model's ability to identify "Nodule" (0.799) and "Infiltration" (0.717) needs improvement. "Infiltration" appear patchy with blurred edges on imaging and their diagnosis relies on subtle texture changes. On the other hand, pulmonary nodules are small lesions that are easily influenced by irrelevant features. Therefore, the identification of both can be relatively difficult; (3) Compared to attention-guided methods, the correlation modeling methods, such as SSGE and CheXGAT, demonstrate strong advantages in handling dependencies between labels. In particular, SSGE achieves the highest accuracy in identifying "Nodules" (0.812) and "Emphysema" (0.948). Among the network structure optimization, MXT achieves the highest average AUC value and demonstrates the best recognition accuracy for "Infiltrations" (0.719) and "Fibrosis" (0.847), showcasing the powerful modeling capabilities of the transformer architecture; (4) Despite our work outperforming the compared methods, our MLRFNet's ability to handle label dependencies is suboptimal. This is a current limitation that our MLRFNet faces and will be a focus of future work.

2) RESULTS ON CheXpert

We evaluated the performance of our model on the validation set of the CheXpert dataset. The ROC curves for each pathology are drawn in Fig. 10(b), while the comparative results obtained by our MLRFNet and other SOTA baselines are presented in Table 6. Moreover, the loss and AUC curves during training and validation are given in Fig. 8(b) and Fig. 9(b), respectively.

In the comparison, we focus on the performance achieved by a single architecture. We quote the ensemble result of Irvine et al. [11] as the single checkpoint performance is not given. We first verified the classification performance of the model on the CheXpert validation set under the "U-Ones" and "U-Zeros" policies. Secondly, like Jung et al. [22], we adopted a mixed policy "U-O&U-Z" to handle uncertain labels. Specifically, "Atelectasis", and "Edema" were trained with the "U-Ones" policy. "Cardiomegaly", "Consolidation", and "Pleural Effusion" were trained with the "U-Zeros" policy.



(a) ChestX-Ray14 (b) CheXpert

FIGURE 10. ROC curves of thoracic diseases on the ChestX-Ray14 (a) and CheXpert (b) datasets, respectively. The corresponding AUC scores are given in Table 5 and Table 6.

TABLE 5. Comparison results of previous SOTA methods on ChestX-Ray14 dataset measured by AUC score of the test set. In each column, the best result is highlighted in bold.

Method	Atel	Card	Effu	Infi	Mass	Nodu	Pne1	Pne2	Cons	Edem	Emph	Fibr	PT	Hern	Mean
DCNN [29]	0.700	0.810	0.759	0.661	0.693	0.669	0.658	0.799	0.703	0.805	0.833	0.786	0.684	0.872	0.745
SSGE [21]	0.792	0.892	0.840	0.714	0.848	**0.812**	0.733	0.885	0.753	0.848	**0.948**	0.827	0.795	0.932	0.830
LLAGNet [15]	0.783	0.885	0.834	0.703	0.841	0.790	0.729	0.877	0.754	0.851	0.939	0.832	0.798	0.916	0.824
CheXGCN [24]	0.786	0.893	0.832	0.699	0.840	0.800	0.739	0.876	0.751	0.850	0.944	0.834	0.795	0.929	0.826
A^3Net [16]	0.779	0.895	0.836	0.710	0.834	0.777	0.737	0.878	0.759	0.855	0.933	0.838	0.791	0.938	0.826
ConsultNet [18]	0.785	0.899	0.835	0.699	0.838	0.775	0.738	0.871	0.763	0.850	0.924	0.831	0.776	0.922	0.822
CheXGAT [23]	0.787	0.879	0.837	0.699	0.839	0.793	0.741	0.879	0.755	0.851	0.945	0.842	0.794	0.931	0.827
MXT [14]	0.798	0.896	0.842	**0.719**	0.856	0.809	0.758	0.879	0.759	0.849	0.906	**0.847**	0.800	0.913	0.830
PCAN [17]	0.785	0.897	0.837	0.706	0.834	0.786	0.730	0.871	0.763	0.854	0.921	0.817	0.791	0.943	0.824
F-PCAM [22]	0.777	0.890	0.836	0.703	0.833	0.796	0.732	0.876	0.745	0.847	0.933	0.824	0.793	0.905	0.821
PCSANet [19]	0.807	0.910	0.879	0.698	0.824	0.750	0.750	0.850	0.802	0.888	0.890	0.812	0.768	0.915	0.825
Our MLRFNet	**0.833**	**0.915**	**0.884**	0.717	**0.858**	0.799	**0.760**	**0.900**	**0.826**	**0.905**	0.941	0.821	**0.810**	**0.963**	**0.853**

TABLE 6. Comparison results of previous SOTA methods on the CheXpert dataset measured by the AUC score of the validation set. The "U-Ones" and "U-Zeros" are different settings for uncertain labels. "U-O&U-Z" is a mixed approach to handling uncertain labels using both policies. "CT" and "LSR" stands for conditional training and label smoothing in [45]. In each column, the best result is highlighted in bold.

Method	Policy	Atel	Card	Cons	Edem	Effu	Mean
Ensemble [11]	U-Ones	0.858	0.832	0.899	**0.941**	0.934	0.893
ConsultNet [18]	U-Ones	0.847	0.868	0.923	0.924	0.926	**0.898**
PCAN [17]	U-Ones	0.848	0.865	0.908	0.912	**0.940**	0.895
DCNN [45]	U-O&CT&LSR	0.825	0.855	**0.937**	0.930	0.923	0.894
Our MLRFNet	U-Ones	**0.861**	**0.935**	0.930	0.835	0.923	0.897
Ensemble [11]	U-Zeros	0.811	0.840	0.932	**0.929**	**0.931**	0.889
ConsultNet [18]	U-Zeros	0.804	0.874	**0.940**	0.894	0.923	0.889
DCNN [45]	U-Zeros	0.745	0.813	0.882	0.921	0.930	0.858
Our MLRFNet	U-Zeros	**0.851**	**0.950**	0.914	0.824	0.922	**0.892**
F-PCAM [22]	U-O&U-Z	**0.954**	0.857	0.924	**0.940**	0.930	0.901
Our MLRFNet	U-O&U-Z	0.848	**0.867**	**0.937**	0.937	0.930	**0.904**

The average AUC score for the five pathologies in the validation set is 0.904 for the proposed MLRFNet when adopting a mixed policy to handle uncertain labels. This is better than the other SOTA methods. As can be seen from Fig. 8(b), the training loss gradually decreases as the number of training epochs increases, while the validation loss first decreases and then begins to rise after the 9th epoch. This is due to the small number of samples in the validation set and the large difference in sample distribution between the validation and training sets. Fig. 9(b) shows that the AUC value of the model on the validation set reaches its maximum at the 9th epoch.

Based on the comparison results in Table 5, we can draw the following conclusions and analyses: (1) When handling uncertain labels with the mixed policy "U-O&U-Z" or "U-Zeros", our MLRFNet outperforms the baseline for comparison and achieves average AUC scores of 0.904 and 0.892, respectively; (2) Our method shows its superiority for some diseases with different label policies. When treating the uncertain labels as "Zero", the performance of "Atelectasis" (0.851) and "Cardiomegaly" (0.950) improved significantly

TABLE 7. Comparison of other evaluation metrics on the ChestX-Ray14 dataset.

Method	Accuracy	Sensitivity	Specificity	F1
LLAGNet [15]	76.7	74.3	76.9	73.7
CheXGCN [24]	76.8	74.6	77.0	73.9
SSGE [21]	77.2	74.9	77.3	74.1
PCAN [17]	77.5	75.3	77.6	74.3
PCSANet [19]	77.0	74.5	77.1	73.8
Our MLRFNet	79.3	76.2	79.6	75.9

by 4.9% and 8.1%, respectively, compared to the best results in the baselines. When using the ''U-O&U-Z'' policy, the performance of ''Cardiomegaly'' (0.867) and ''Consolidation'' (0.937) improved slightly by 1.6% and 1.4%, respectively, compared to the best results in the baselines; (3) For all methods in Table 6, the classification accuracy on the validation set is higher when using the ''U-Ones'' policy than when using the ''U-Zeros'' policy. This may be due to the fact that setting uncertain labels to ''Ones'' allows the model to learn more about the pathological features of the five diseases present in the validation set; (4) Although our MLRFNet outperforms the best results of baselines under the ''U-Zeros'' and ''U-O&U-Z'' label policy, the ability to recognize ''Edema'' still needs to be improved when using either the ''U-Ones'' or ''U-Zeros'' policy. This is a limitation of our method.

F. COMPARISON OF OTHER EVALUATION METRICS

In this section, we utilize the average Sensitivity, Specificity, Accuracy, and F1-score to further measure the performance of our MLRFNet on ChestX-Ray14. The comparative results of the above evaluation metrics are summarized in Table 7. Meanwhile, the FLOPs and test time consumed by a single image during the test phase are illustrated in Table 8 to measure the computational complexity of the model.

1) CLASSIFICATION METRICS ANALYSIS

Based on the comparison results in Table 7, we can draw the following conclusions and analyses: (1) Our MLRFNet shows improvements of 2.3%, 1.2%, 2.6%, and 2.2% respectively over the best results of the baseline model in terms of Accuracy, Sensitivity, Specificity, and F1-score. It indicates that our model has significantly improved overall performance. (2) Typically, the performance of medical diagnostic systems is primarily measured by their Specificity and Sensitivity. Improvements in both Sensitivity and Specificity mean that our MLRFNet can diagnose more patients with thoracic diseases while reducing the rate of misdiagnosis.

2) COMPUTATIONAL COMPLEXITY ANALYSIS

As can be seen from Table 8, our MLRFNet processes single CXR images quickly during the testing phase. However, its floating-point computation is relatively high compared to the latest models, such as PCAN and PCSANet, which is a limitation of our method. Nevertheless, considering the average AUC score, our model achieves a significant improvement in classification accuracy with only a slight increase in computation. It remains highly competitive in thoracic disease diagnosis algorithms.

G. ABLATION STUDY

In this section, we will verify the effectiveness of the components in MLRFNet and the optimized loss function BFL through a series of ablation experiments on ChestX-Ray14 and CheXpert. It is worth noting that we utilize the previously mentioned ''U-O&U-Z'' policy to handle uncertain labels in the CheXpert dataset.

1) MODULE ABLATION EXPERIMENT

An ablation experiment was set up on the CheXpert and ChestX-Ray14 to investigate the effectiveness of the ECA attention module and multi-level residual feature classifier (MRFC). MLRFNet is our baseline, and we remove the corresponding modules to explore their impact on classification accuracy. Table 9 shows the results of the ablation experiment. First, removing the ECA attention module reduces the baseline by 0.4% and 0.6% on CheXpert and ChestX-Ray14, respectively. Second, replacing the MRFC with the most commonly used classifier (global average pooling with fully connected layer), the AUC of the network dropped by approximately 0.9% and 2.4% on the two datasets, respectively. Experimental results show that the simultaneous use of the ECA attention mechanism and MFRC is of great help in enhancing features between channels as well as pathological features at specific locations in space.

2) BACKBONE NETWORK ABLATION EXPERIMENT

The backbone network of the MLRFNet in this paper is Res2Net50, and its 3×3 convolution adopts a hierarchical connection method, which has more robust multi-scale extraction capabilities. To verify the influence of the backbone network on results, we replace Res2Net50 with ResNet50 or DenseNet121 as the feature extractor while keeping the other settings unchanged. To improve the convergence speed and learning ability of the model, the backbone network in the experiment will use the pre-trained model on the ImageNet dataset. The experimental results in Table 10 show that Res2Net50 achieves the best classification results on the CheXpert and ChestX-Ray14, which are 1.3% and 0.8% higher than the second-ranked DenseNet121, respectively.

3) LOSS FUNCTIONS ABLATION EXPERIMENT

We conducted comparative experiments to explore the impact of different loss functions on classification accuracy. The experimental result is shown in Table 11. The average AUC achieved by the MLRFNet using the biased focal loss or the focal loss is better than that obtained using the binary cross-entropy loss on the CheXpert and ChestX-Ray14 datasets. The focus loss enhances the model's attention to

TABLE 8. Comparison of computational consumption by a single image during the test phase on the ChestX-Ray14 dataset.

Method	SSGE [21]	CheXGCN [24]	LLAGNet [15]	PCAN [17]	PCSANet [19]	Our MLRFNet
FLOPs (G)	17.74	17.86	34.96	2.86	3.92	4.29
Time (s)	0.059	0.061	0.094	0.054	0.072	0.046
AUC	0.830	0.826	0.824	0.830	0.825	0.853

TABLE 9. Ablation experiment on ECA and MRFC module.

ECA	MRFC	The average AUC	
		CheXpert	ChestX-Ray14
✓	✓	0.904	0.853
✓	×	0.882	0.845
×	✓	0.900	0.848

TABLE 10. The influence of backbone networks on the results.

Backbone	The average AUC	
	CheXpert	ChestX-Ray14
Res2Net50	0.904	0.853
ResNet50	0.885	0.844
DenseNet121	0.892	0.846

TABLE 11. The influence of loss function on the results.

Loss function	The average AUC	
	CheXpert	ChestX-Ray14
Biased Focal Loss	0.904	0.853
Binary Cross-entropy Loss	0.884	0.846
Focal Loss	0.894	0.849

hard-to-classify samples through the focus factor. However, the dataset has more negative samples for most thoracic diseases. The biased loss function further sets the probability shift factor and weight factor to enhance the model's ability to identify the hard-to-classify parts of negative samples, thus achieving the best classification results.

H. VISUALIZATION ANALYSIS

In this section, we generate some heatmaps of the focal areas using Grad-CAM [46] in Fig. 11 and visualize classification scores in Fig. 12, respectively. Grad-CAM can back-propagate the model and calculate the feature maps' gradient information. By weighting the feature maps' channel with gradient information, a heat map is made to show the area of interest for a category in the CXR image. To validate the accuracy of model recognition, we compared the generated heat map to the lesion-labeled maps provided by professional doctors in the ChestX-Ray14.

As shown in Fig. 11, it can be seen that the lesion areas of the corresponding diseases are activated, and the red highlighted parts represent the most concerned parts of the model, which are also the main parts of the diagnostic basis. It can be seen visually that the activated regions in CXR images are consistent with the regions labeled by professional physicians, and the highlighted red areas are very concentrated, allowing the naked eye to locate the area where the lesion occurs quickly. Even though the size of the lesion and the abnormal areas look different, the network can still do accurate identification. Fig. 12 shows that the network can give a higher prediction score for most diseases, and the top-scoring diseases are consistent with the labeled actual value. Especially for "Hernia", the network prediction score is quite higher than other diseases, indicating that the network has a strong predictive ability for this disease.

V. DISCUSSION

When dealing with thoracic diseases that display concurrency and significant variations in lesion size, most existing methods struggle to extract features at multiple scales and accurately capture the spatial location of lesions. This makes it difficult to increase classification accuracy due to uncertain pathological features. Moreover, in the ChestX-Ray14 and CheXpert datasets, there is an imbalance between the positive and negative samples of some specific pathologies, leading to serious classification difficulties with fewer positive samples. To address the aforementioned issues, we propose MLRFNet, a thoracic disease classification network based on multi-level residual feature fusion, and optimize the loss function during training.

The difference between the previous works and our approach for generating prediction vectors is that MLRFNet fully uses the residual feature vectors of different downsampling stages. Specifically, the MRFC module applies spatial pooling to the feature maps to achieve pixel-level attention. Based on the global average pooling, the residual feature vector is used as an extra part to help make the prediction scores. With the above method, the network can pay more attention to specific disease-related locations without adding more model parameters, which is a lightweight way to do it.

We conducted comprehensive experiments on the ChestX-Ray14 and CheXpert datasets. The results in Table 5 and Table 6 show that MLRFNet achieves better performance compared with the current SOTA methods. In the ablation study, we remove the ECA attention module and the MRFC module, respectively, resulting in a drop in overall classification accuracy. During the training phase, we replace the BFL with two other common loss functions for multi-label classification while keeping the original network structure unchanged, but the classification performance of the network also degrades. In the visualization analysis, Fig. 11 shows

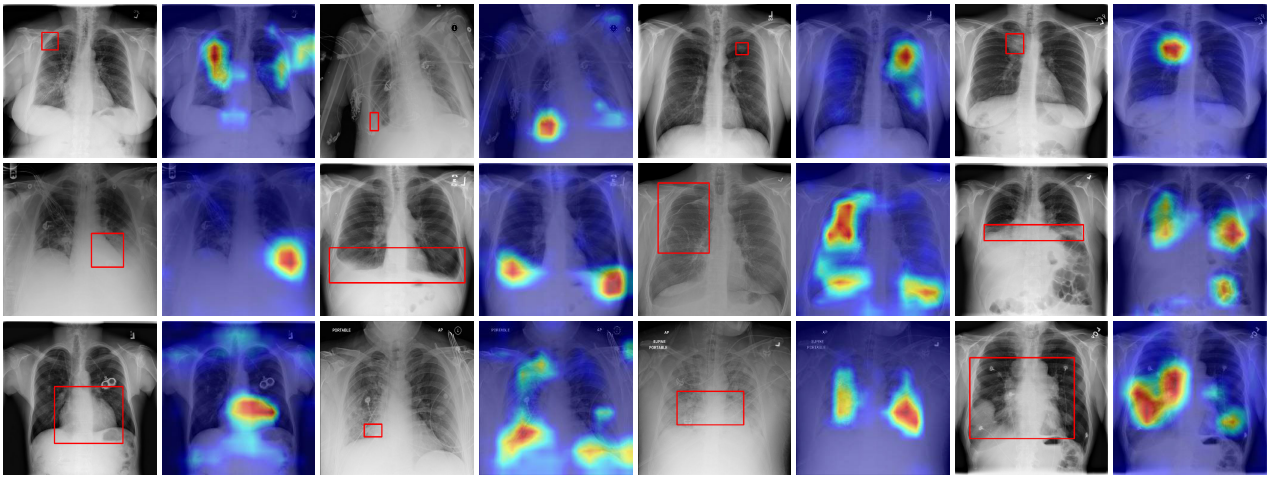


FIGURE 11. Examples of heatmaps generated from MLRFNet. Manual annotations from professional doctors labeled with the ground truth bounding boxes provided by [29]. Noting that the network is very accurate in identifying small lesions.

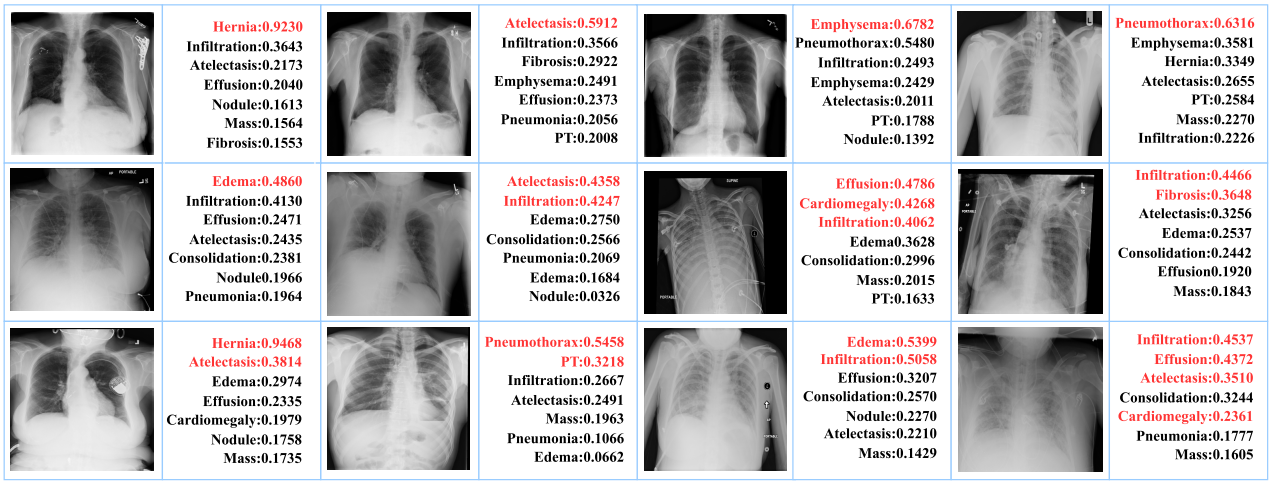


FIGURE 12. Examples of classification results on ChestX-Ray14. The top-7 predicted categories and corresponding probability scores are presented. The ground-truth pathologies are highlighted in red. A higher prediction score indicates a higher probability of the disease corresponding to that score. Noting that MLRFNet can accurately predict ground-truth pathologies in the single Chest X-ray image.

that MLRFNet can accurately locate lesions of different sizes and has a good effect on identifying small lesions. This is because Res2Net has excellent multi-scale feature extraction capabilities. In addition, the red activation area in the heat map is very concentrated so that the relevant lesion area can be noticed intuitively. It is mainly because the spatial pooling in MRFC enhances pathological features on feature maps while features unrelated to the disease are suppressed.

Despite achieving high classification accuracy on the ChestX-Ray14 and CheXpert datasets, the method proposed in this paper still has some limitations. In particular, when exploring potential relationships between labels, our model relies solely on CNNs to learn the semantic information contained in images and lacks explicit modeling of the relationships between different categories of labels. This is not conducive to exploring the intrinsic connections of thoracic diseases. Additionally, while our model is fast in inference speed, it has a slight increase in floating-point computations compared to the lightest comparison method. To address these limitations, future work will introduce additional neural network structures to further explore the dependencies between labels in CXR images and improve the classification accuracy of the model. Additionally, using model compression methods such as knowledge distillation is also one of the directions for future work. This will help the model maintain its lightweight while achieving high classification performance.

VI. CONCLUSION

In this paper, we propose the MLRFNet for multi-label chest X-ray image classification. The proposed network can automatically learn pathological features end-to-end to recognize

common thoracic diseases. MLRFNet achieves the average AUC of 0.853 and 0.904 for ChestX-Ray14 and CheXpert datasets, respectively, which perform excellently in classifying chest X-ray images. Our network has the following advantages: (1) The network has solid multi-scale feature extraction capability, which can quickly get receptive field information on different lesion sizes. (2) The network adopts a novel multi-level residual feature fusion method to generate classification vectors so that the prediction of the network considers the spatial location information of specific diseases. (3) Benefiting from the biased focal loss function we proposed, the network can learn more about the difficult-to-classify parts of negative samples, improving the overall classification accuracy. A single chest X-ray image often contains multiple thoracic diseases, and related diseases appear simultaneously. In future work, we will focus on exploring the dependencies between labels and techniques for model compression to further improve the overall performance of the model.

REFERENCES

[1] S. Zhou, D. Nie, E. Adeli, J. Yin, J. Lian, and D. Shen, "High-resolution Encoder–Decoder networks for low-contrast medical image segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 461–475, 2020.

[2] W. Weng and X. Zhu, "INet: Convolutional networks for biomedical image segmentation," *IEEE Access*, vol. 9, pp. 16591–16603, 2021.

[3] Y. Yuan, M. Chao, and Y.-C. Lo, "Automatic skin lesion segmentation using deep fully convolutional networks with Jaccard distance," *IEEE Trans. Med. Imag.*, vol. 36, no. 9, pp. 1876–1886, Sep. 2017.

[4] D. E. Worrall, C. M. Wilson, and G. J. Brostow, "Automated retinopathy of prematurity case detection with convolutional neural networks," in *Proc. LABELS/DLMIA*. Athens, Greece: Springer, Oct. 2016, pp. 68–76.

[5] T. Padma, C. U. Kumari, D. Yamini, K. Pravallika, K. Bhargavi, and M. Nithya, "Image segmentation using mask R-CNN for tumor detection from medical images," in *Proc. Int. Conf. Electron. Renew. Syst. (ICEARS)*, Mar. 2022, pp. 1015–1021.

[6] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.

[7] P. Kumar, M. Grewal, and M. M. Srivastava, "Boosted cascaded ConvNets for multilabel classification of thoracic diseases in chest radiographs," in *Proc. Int. Conf. Image Anal. Recognit. (ICIAR)*. Cham, Switzerland: Springer, Jun. 2018, pp. 546–552.

[8] X. Wei, W. Li, M. Zhang, and Q. Li, "Medical hyperspectral image classification based on end-to-end fusion deep neural network," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 11, pp. 4481–4492, Nov. 2019.

[9] P. Nardelli, D. Jimenez-Carretero, D. Bermejo-Pelaez, G. R. Washko, F. N. Rahaghi, M. J. Ledesma-Carbayo, and R. S. J. Estépar, "Pulmonary artery–vein classification in CT images using deep learning," *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2428–2440, Nov. 2018.

[10] B. Chen, J. Li, X. Guo, and G. Lu, "DualCheXNet: Dual asymmetric feature learning for thoracic disease classification in chest X-rays," *Biomed. Signal Process. Control*, vol. 53, Aug. 2019, Art. no. 101554.

[11] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 590–597.

[12] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," 2017, *arXiv:1711.05225*.

[13] M. F. Hashmi, S. Katiyar, A. G. Keskar, N. D. Bokde, and Z. W. Geem, "Efficient pneumonia detection in chest Xray images using deep transfer learning," *Diagnostics*, vol. 10, no. 6, p. 417, Jun. 2020.

[14] X. Jiang, Y. Zhu, G. Cai, B. Zheng, and D. Yang, "MXT: A new variant of pyramid vision transformer for multi-label chest X-ray image classification," *Cogn. Comput.*, vol. 14, no. 4, pp. 1362–1377, Jul. 2022.

[15] B. Chen, J. Li, G. Lu, and D. Zhang, "Lesion location attention guided network for multi-label thoracic disease classification in chest X-rays," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 7, pp. 2016–2027, Jul. 2020.

[16] H. Wang, S. Wang, Z. Qin, Y. Zhang, R. Li, and Y. Xia, "Triple attention learning for classification of 14 thoracic diseases using chest radiography," *Med. Image Anal.*, vol. 67, Jan. 2021, Art. no. 101846.

[17] X. Zhu, S. Pang, X. Zhang, J. Huang, L. Zhao, K. Tang, and Q. Feng, "PCAN: Pixel-wise classification and attention network for thoracic disease classification and weakly supervised localization," *Comput. Med. Imag. Graph.*, vol. 102, Dec. 2022, Art. no. 102137.

[18] Q. Guan, Y. Huang, Y. Luo, P. Liu, M. Xu, and Y. Yang, "Discriminative feature learning for thorax disease classification in chest X-ray images," *IEEE Trans. Image Process.*, vol. 30, pp. 2476–2487, 2021.

[19] K. Chen, X. Wang, and S. Zhang, "Thorax disease classification based on pyramidal convolution shuffle attention neural network," *IEEE Access*, vol. 10, pp. 85571–85581, 2022.

[20] L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Bernard, and K. Lyman, "Learning to diagnose from scratch by exploiting dependencies among labels," 2017, *arXiv:1710.10501*.

[21] B. Chen, Z. Zhang, Y. Li, G. Lu, and D. Zhang, "Multi-label chest X-ray image classification via semantic similarity graph embedding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2455–2468, Apr. 2022.

[22] B. Jung, L. Gu, and T. Harada, "Graph interaction for automated diagnosis of thoracic disease using X-ray images," *Proc. SPIE*, vol. 12032, pp. 135–147, Apr. 2021.

[23] Y.-W. Lee, S.-K. Huang, and R.-F. Chang, "CheXGAT: A disease correlation-aware network for thorax disease diagnosis from chest X-ray images," *Artif. Intell. Med.*, vol. 132, Oct. 2022, Art. no. 102382.

[24] B. Chen, J. Li, G. Lu, H. Yu, and D. Zhang, "Label co-occurrence learning with graph convolutional networks for multi-label chest X-ray image classification," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 8, pp. 2292–2302, Aug. 2020.

[25] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2261–2269.

[26] W. Khan, N. Zaki, and L. Ali, "Intelligent pneumonia identification from chest X-rays: A systematic literature review," *IEEE Access*, vol. 9, pp. 51747–51771, 2021.

[27] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.

[28] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 11531–11539.

[29] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 3462–3471.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Miami, FL, USA, Jun. 2009, pp. 248–255.

[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 2, pp. 84–90, 2017.

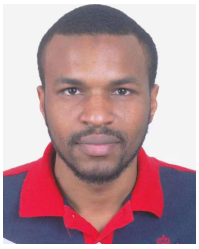
[32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9.

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

[34] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 548–558.
[35] N.-U. Rehman, M. S. Zia, T. Meraj, H. T. Rauf, R. Damaševičius, A. M. El-Sherbeeny, and M. A. El-Meligy, "A self-activated CNN approach for multi-class chest-related COVID-19 detection," *Appl. Sci.*, vol. 11, no. 19, p. 9023, Sep. 2021.
[36] M. A. Khan, V. Rajinikanth, S. C. Satapathy, D. Taniar, J. R. Mohanty, U. Tariq, and R. Damaševičius, "VGG19 network assisted joint segmentation and classification of lung nodules in CT images," *Diagnostics*, vol. 11, no. 12, p. 2208, Nov. 2021.
[37] A. Jaszcz, D. Polap, and R. Damasevicius, "Lung X-ray image segmentation using heuristic red fox optimization algorithm," *Sci. Program.*, vol. 2022, Jul. 2022, Art. no. 4494139.
[38] A. Amirkhani, A. Khosravian, M. Masih-Tehrani, and H. Kashiani, "Robust semantic segmentation with multi-teacher knowledge distillation," *IEEE Access*, vol. 9, pp. 119049–119066, 2021.
[39] M. K. Mahbub, M. Biswas, L. Gaur, F. Alenezi, and K. Santosh, "Deep features to detect pulmonary abnormalities in chest X-rays due to infectious diseaseX: COVID-19, pneumonia, and tuberculosis," *Inf. Sci.*, vol. 592, pp. 389–401, May 2022.
[40] K. Zhu and J. Wu, "Residual attention: A simple but effective method for multi-label recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 184–193.
[41] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2999–3007.
[42] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor, "Asymmetric loss for multi-label classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 82–91.
[43] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," 2019, *arXiv:1912.01703*.
[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[45] H. H. Pham, T. T. Le, D. Q. Tran, D. T. Ngo, and H. Q. Nguyen, "Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels," *Neurocomputing*, vol. 437, pp. 186–194, May 2021.
[46] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 618–626.

QIANG LI received the B.E. and M.E. degrees from the School of Information Engineering, Taiyuan University of Technology, Taiyuan, China, in 1997 and 2000, respectively, and the Ph.D. degree from the School of Electronic Information Engineering, Tianjin University, Tianjin, China, in 2003. He is currently a Professor with the School of Microelectronics, Tianjin University. His research interests include intelligent signal processing and AI system design.

YU LAI received the B.E. degree in communication engineering from Zhengzhou University, China, in 2020. He is currently pursuing the master's degree with Tianjin University, China. His current research interests include deep learning and medical image processing.



MOHAMMED JAJERE ADAMU (Member, IEEE) received the M.Eng. degree in signal and information processing from the Tianjin University of Technology and Education, China, in 2017. He is currently pursuing the Ph.D. degree with Tianjin University, China. His current research interests include signal and medical image processing and wearable antenna design and analysis for medical devices.

LEI QU is currently with Hisense Group Holdings Company Ltd., China. His current research interests include social media and multimedia content analysis.

JIE NIE (Member, IEEE) received the Ph.D. degree in computer science from the Ocean University of China, Qingdao, China, in 2011. From September 2009 to September 2010, she was a Visiting Scholar with the School of Electrical Engineering, University of Pittsburgh, Pittsburgh, PA, USA. From 2015 to 2017, she was a Postdoctoral Researcher with Tsinghua University, Beijing, China. She is currently with the Ocean University of China. Her current research interests include social media and multimedia content analysis.

WEIZHI NIE (Member, IEEE) received the Ph.D. degree in computer science from Tianjin University, China, in 2015. He is currently with Tianjin University. His current research interests include social media and multimedia content analysis.