Neural Computing and Applications (2023) 35:17187–17199
https://doi.org/10.1007/s00521-023-08604-y

ORIGINAL ARTICLE

A new classification method for diagnosing COVID-19 pneumonia based on joint CNN features of chest X-ray images and parallel pyramid MLP-mixer module

Yiwen Liu[1] · Wenyu Xing[2] · Mingbo Zhao[1,3] · Mingquan Lin[3]

Received: 25 November 2022 / Accepted: 11 April 2023 / Published online: 28 April 2023

Abstract

During the past three years, the coronavirus disease 2019 (COVID-19) has swept the world. The rapid and accurate recognition of covid-19 pneumonia are ,therefore, of great importance. To handle this problem, we propose a new pipeline of deep learning framework for diagnosing COVID-19 pneumonia via chest X-ray images from normal, COVID-19, and other pneumonia patients. In detail, the self-trained YOLO-v4 network was first used to locate and segment the thoracic region, and the output images were scaled to the same size. Subsequently, the pre-trained convolutional neural network was adopted to extract the features of X-ray images from 13 convolutional layers, which were fused with the original image to form a 14-dimensional image matrix. It was then put into three parallel pyramid multi-layer perceptron (MLP)-Mixer modules for comprehensive feature extraction through spatial fusion and channel fusion based on different scales so as to grasp more extensive feature correlation. Finally, by combining all image features from the 14-channel output, the classification task was achieved using two fully connected layers as well as Softmax classifier for classification. Extensive simulations based on a total of 4099 chest X-ray images were conducted to verify the effectiveness of the proposed method. Experimental results indicated that our proposed method can achieve the best performance in almost all cases, which is good for auxiliary diagnosis of COVID-19 and has great clinical application potential.

Keywords Pneumonia · Auxiliary diagnosis · Object detection · Convolutional neural network · Parallel pyramid MLP-mixer

1 Introduction

Pneumonia is a common respiratory disease, which is a serious menace to human life and health. During the past three years, the incidence rate and mortality of pneumonia have shown an upward trend [1–3]. How to realize the rapid identification of the coronavirus disease 2019 (COVID-19) and other pneumonia from the source is therefore of great significance for rapid control, cutting off infection, and timely treatment in clinic [4–10]. At present, the clinical diagnosis of COVID-19 pneumonia mainly depends on nucleic acid and antigen kits detection. Although nucleic acid detection has high accuracy, its collection process has high requirements on clinicians or professional technicians causing inconsistency and inefficiency of diagnosis; on the other hand, though antigen kits detection is more convenient and faster than nucleic acid, it has low specificity and high probability of false positive so

✉ Mingbo Zhao
mzhao4@dhu.edu.cn

✉ Mingquan Lin
mqlin5-c@my.cityu.edu.hk

Yiwen Liu
ywliu@mail.dhu.edu.cn

Wenyu Xing
wyxing20@fudan.edu.cn

[1] College of Information Science and Technology, Donghua University, Shanghai, People's Republic of China

[2] School of Information Science and Technology, Fudan University, Shanghai, People's Republic of China

[3] Department of Electrical Engineering, City University of Hong Kong, Kowloon Tong, Hong Kong, People's Republic of China

that they cannot accurately distinguish different types of pneumonia.

During the past decade, medical imaging analysis (MIA) technologies, e.g., computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, and X-ray, have achieved rapid development due to their intuitiveness and visualization, so that they have become the important means of clinical disease diagnosis [11, 12]. Merited from the advantages of fast, low radiation and high resolution, Chest X-ray (CXR) been widely used in the diagnosis of patients with COVID-19 and other pneumonia. For example, Sitaula et al. [13] proposed an attention-based deep learning model by adding an attention module on VGG-16 to distinguish COVID-19 cases. Khan et al. [14] found that the integration of the clinical data features of patients and the features extracted from CXR images is helpful to improve the diagnostic accuracy of COVID-19 through EffientNetB7 model. A novel multi-feature convolutional neural network (CNN) architecture was designed in [15] which combined the original CXR data and enhanced CXR images by the local phase enhancement method. Lin et al. [16] proposed an adaptive attention network to distinguish COVID-19 patients, where a specially designed adaptive deformation ResNet is adapted to focus on the ROI on input image for feature extraction. Then, the attention-based encoder learned the context information of extracted features through the self-attention mechanism and the feedforward networks were used to predict the final results. Khuzani et al. [17] used the dimensionality reduction method to generate a set of optimal features of CXR images and built an efficient machine learning classifier, which can distinguish COVID-19 and non-COVID-19 cases. Babukarthik et al. [18] used more than 5000 CXR images to train the genetic deep learning CNN model from scratch to complete the classification of COVID-19. Shorfuzzaman et al. [19] proposed a collaborative approach, which integrated contrastive learning and fine-tuned pre-trained ConvNet encoder to capture unbiased feature representation, and distinguished COVID-19 cases using the Siamese network.

Meanwhile, the past decades witnessed a dramatic rise in the platforms, tools and applications based on neural networks. One important breakthrough technique is known as deep learning, which includes a large family of neural computing methods that employ deep architectures composed of multiple non-linear transformations to model high-level abstractions of raw data. Recent studies have shown that deep neural networks significantly improve the performance of MIA. These accomplishments are attributed to the ability of neural networks to understand rich mid-level data representations. There are also some researches that analyzed the performance of various neural network models for diagnosing COVID-19 based on CXR images. For instance, Li et al. [20] analyzed CXR images using 17 CNN models for distinguishing healthy individuals, bacterial pneumonia, and viral pneumonia, with high diagnostic accuracy. Salvatore et al. [21] trained the deep learning model with an ensemble of 10 CNN models, which was tested on two independent cohorts and had good experimental results. Gayathri et al. [22] adopted five pre-trained CNN models for diagnosing COVID-19 and non-COVID-19. Avola et al. [23] fine-tuned 12 well-known pre-trained models to discriminate healthy people and those showing pneumonia symptoms derived from either viral or bacterial source. Chetoui et al. [24] fine-tuned several vision transformers (ViT) models for investigating the efficacy of ViT for detecting COVID-19 in CXR images.

The above research proves the effectiveness of deep learning in the diagnosis of COVID-19. It can be concluded that the improved model generally comes from the following aspects: input diversity, feature diversity, model structure improvement and multiple model fusion. Increasing the diversity of input data types, or inputting extracted features for transfer learning, which lacks further analysis of input. The way to add information processing module or fuse multiple model increases the calculation difficulty and parameter quantity while improving the accuracy. To handle these problems, visual Transformer (ViT) and its extensive work have remove the module of convolution layer and adopt the self-attention strategy by involving several standard Transformer blocks, where hand-crafted manipulation can be avoided and the inductive biases can be reduced. As a result, it can achieve comparable or even higher accuracy than CNN. More recently, researchers take a step forward, where they develop a multi-layer perceptron (MLP)-based method, namely, MLP-Mixer, with only MLP layers involved for reducing even more inductive bias than ViT. The advantage of this pioneering work is that it could enable the intersection of communications between channels and patches based on the channel-mixing MLP layer and token-mixing MLP layer, respectively so as to establish long-range dependencies of the input patches but solely rely on pure multi-layer perceptrons. It can also achieve competitive results compared with ViT. Despite the simplicity of MLP-mixer, there still exists a gap in accuracy between MLP-mixer and ViT, and there is seldom work developed for diagnosing COVID-19 by utilizing MLP-mixer. Motivated by this end, we proposed a novel deep learning-based method for diagnosing COVID-19, other pneumonia, and normal cases of different patients. The contributions are summarized as follows:

(1) We have proposed a new pipeline of deep learning framework for diagnosing COVID-19 pneumonia,

where the proposed method can automatically detect the thoracic region as well as distinguish Normal, COVID-19 pneumonia and other pneumonia cases.

(2) To grasp more extensive feature correlation, the parallel pyramid MLP-Mixer module was designed to simultaneously analyze multi-channel input, each of which can extract multi-scale spatial and channel features.

(3) Extensive simulations have been conducted. Simulation results verify the effectiveness of the proposed work, which is good for auxiliary diagnosis of COVID-19 and has great clinical application potential.

The rest of this paper is organized as follows. In Sect. 2, we give a detailed description of the proposed image classification method; In Sect. 3 both the experimental results and analysis are described for evaluations; In Sect. 4, discussions are provided, and the final conclusions are drawn in Sect. 5.

2 Methodology

2.1 Image processing

Due to the influence of the patient's body differences and subjective and objective factors in the acquisition process, the size and area of CXR images are inconsistent. Meanwhile, pneumonia is mainly reflected in the change of the parenchyma region, for focusing on the changes of the region of interest (ROI) and eliminating the disturbance of other regions, we preprocessed original CXR images, of which the ROI were located and segmented by used the self-trained YOLO-v4 network [25]. Segmented images were fine-tuned and filled to the same size of 256×256. In addition, the number of CXR images of patients with COVID-19 is relatively limited. In order to balance the number of different groups, a small angle rotation was used for data augmentation.

2.2 Design of diagnosis model

Aiming at the diagnosis task of normal cases without pneumonia, patients with COVID-19, or other pneumonia, we excogitated a novel diagnostic strategy, which is mainly composed of three blocks, including a multi-layer CNN-based depth feature map transformation block, self-designed parallel pyramid MLP-Mixer-based multi-channel feature analysis block, and SoftMax-based classifier block. The flow chart of our method is shown in Fig. 1.

2.2.1 Abstract transformation of original images

Multi-layer CNN (Table 1) pre-trained by ImageNet data set was used to obtain the abstract transformation maps of original images, which can extract deeper image representation information with various receptive fields. The transformation map of the strongest active channel in 13 recursive convolution layers was adopted and scaled to the same size of the original image. Subsequently, by fusing the original image and 13 extracted transformation maps, the 14-dimensional (14D) matrixes were applied to represent CXR images and used for the multi-channel input of parallel pyramid MLP-Mixer modules, which not only describes the abstract deep information but also supplements the concrete information of the CXR image, enriching the diversity and comprehensiveness of the input.

2.2.2 Parallel pyramid MLP-mixer module

MLP-Mixer was proposed by Google Brain Team in 2021 [26] which is fully based on MLP. It has been applied in many fields [27, 28]. Compared with the traditional convolutional model to process features, MLP-Mixer can consider the channel and spatial information of input at the same time, and its structure is relatively simple and effective. Inspired by it, we designed a parallel pyramid MLP-Mixer module to process multidimensional data, as shown in the feature analysis block in Fig. 1. It consists of 14 parallel pyramid MLP-Mixer processing channels, each of which is composed of three layers of deformable MLP-Mixer layers and a 1 * 1 convolution layer.

The deformable MLP-Mixer layer is shown in Fig. 2. It takes the sequence combination of patches of pictures as input, and each layer of deformable MLP-Mixer layers mixes the information of channel and spatial position through two MLP layers with adjusted parameters to realize comprehensive image information extraction. Each MLP layer is composed of two fully-connected layers and a Rectified Linear Unit (ReLU) activation function.

When the input enters the processing channel, the input information of the picture block is firstly processed into a matrix with dimension $(n, d) = patch \times channel$. Then the token-mixing MLP on each column of the matrix was acted to mix the information of different patches in the way of cross-location. Subsequently, the channel-mixing MLP on each row acted to mix the information of different channels per per-location as the final output of this mixer layer. The principle of the mixer module in our model can be described as Eq. (1).

$$\begin{aligned} U_{*,i} &= W_2\sigma(W_1 LayerNorm(X)_{*,i}),\ for\ i = 1\ldots C, \\ Y_{j,*} &= W_4\sigma(W_3 LayerNorm(U)_{j,*}),\ for\ j = 1\ldots S. \end{aligned} \tag{1}$$

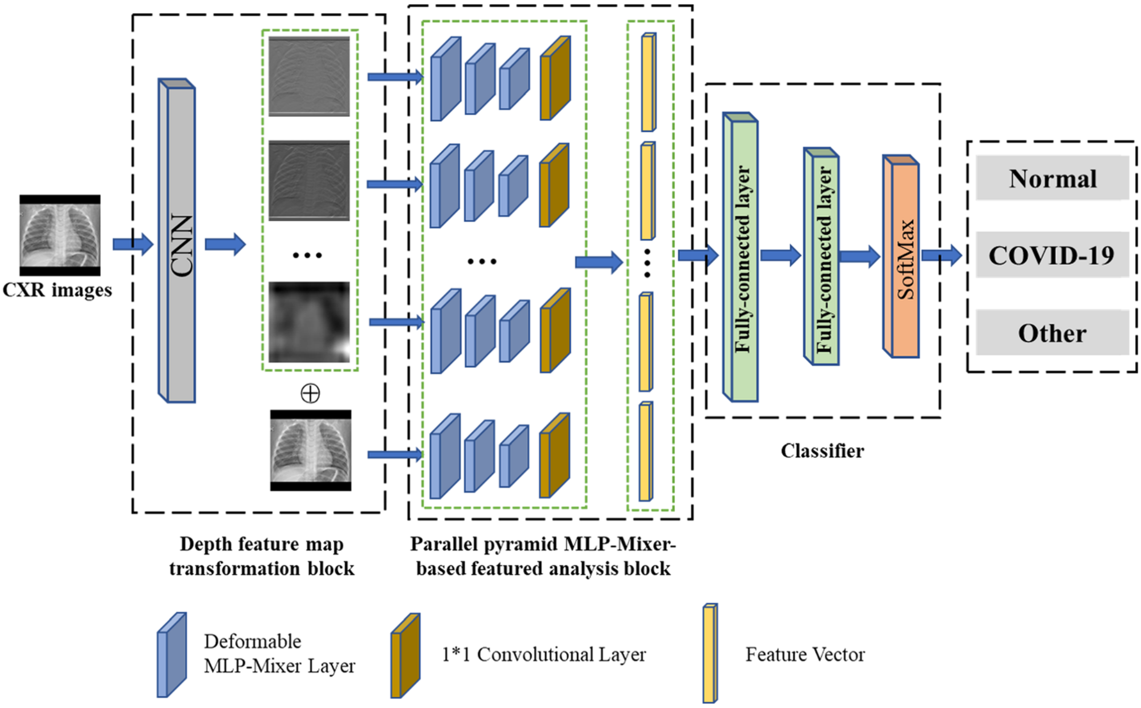


Fig. 1 Flow chart of the proposed method

Table 1 Model parameters of pre-trained convolutional neural network

Layers	Kernel size	Feature map
Conv 1-1 &1-2	3 × 3 × 1	256 × 256 × 64
Max-Pooling	2 × 2 × 2	128 × 128 × 64
Conv 2-1 &2-2	3 × 3 × 1	128 × 128 × 128
Max-Pooling	2 × 2 × 2	64 × 64 × 128
Conv 3-1 &3-2 &3-3	3 × 3 × 1	64 × 64 × 256
Max-Pooling	2 × 2 × 2	64 × 64 × 256
Conv 3-1 &3-2 &3-3	3 × 3 × 1	32 × 32 × 512
Max-Pooling	2 × 2 × 2	16 × 16 × 512
Conv 4-1 &4-2 &4-3	3 × 3 × 1	8 × 8 × 512

Here, $X = [x_1, ..., x_S] \in \mathbb{R}^{C \times S}$ represents the input of the model, where C and S denote channels and patches respectively. σ is an element-wise nonlinearity. Layer normalization is denoted by *LayerNorm*(). U and Y indicate the output after the token-mixing MLP layer and Channel-mixing MLP layer, respectively. W_1, W_2, W_3, W_4 represent the weight of the fully-connected layer, respectively.

In our proposed model, three deformable MLP-Mixer layers with different scales were stacked to form a pyramid feature extraction module for analyzing the input image matrix. Thus, size changes of out patch in each MLP-Mixer layer realized the multi-scale analysis of each single-channel input image matrix. What is more important, the parallel and repeated use of the deformable MLP-Mixer module made it possible to analyze the input 14D image matrix at the same time, realizing the function of multi-channel and multi-scale feature extraction.

Finally, the outputs of 14 parallel pyramid feature extraction modules were processed by convolutional layer with kernel size of 1 × 1 and flattened to form a high-dimensional feature vector and describe input CXR image, which was used for subsequent classification block. Assuming the size of the input feature map was $M \times M \times 1$, the output sizes of three deformable MLP-Mixer layers are shown in Table 2.

2.2.3 Classifier and loss function

The classification block was composed of two fully connected layers and one SoftMax layer. Considering the convergence rate and actual effect, we use L_2 function to train our classification model, as shown in Eq. (2). here, y_i and $f(x_i)$ represent the ground truth and the predicted value of the model, respectively.

$$D_{L2} = \sum_{i=1}^{n} (y_i - f(x_i))^2. \tag{2}$$

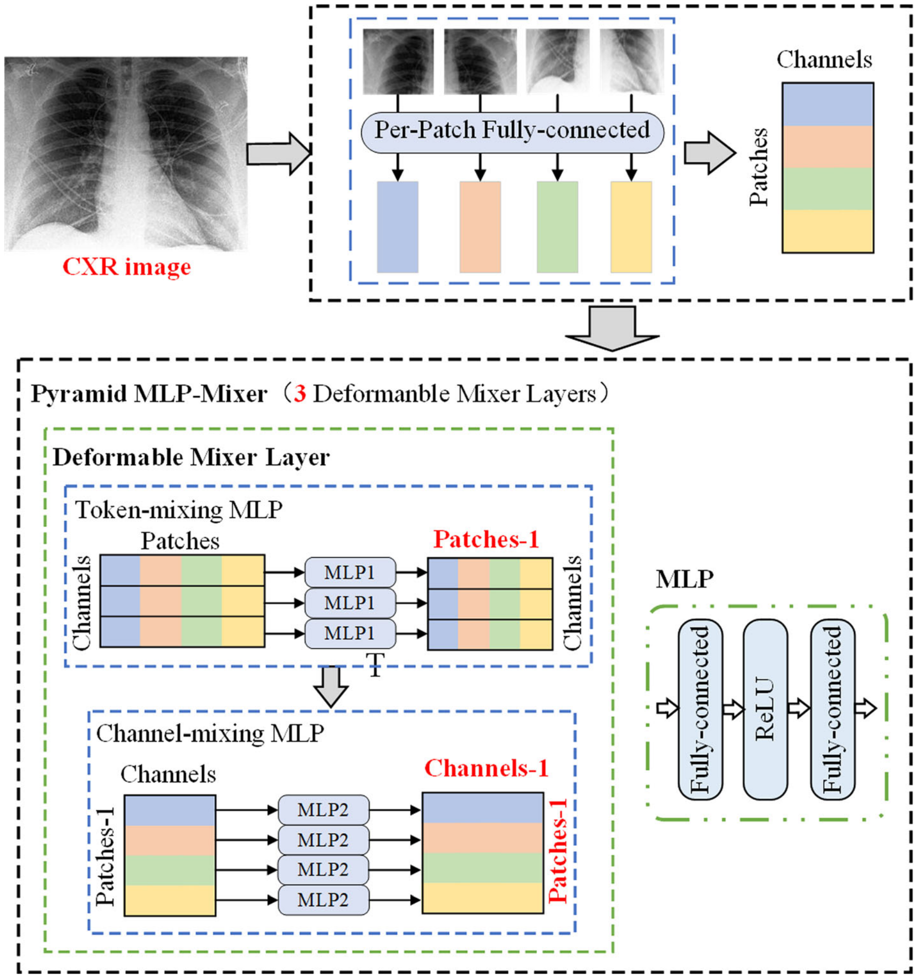


Fig. 2 Structure diagram of Pyramid MLP-Mixer module

Table 2 Parameters of three deformable MLP-Mixer layers

Index	Output size	Number of channels
Deformable MLP-Mixer 1	$M/2 \times M/2$	16
Deformable MLP-Mixer 2	$M/4 \times M/4$	32
Deformable MLP-Mixer 3	$M/8 \times M/8$	64

2.3 Statistics

Statistical analysis was performed using SPSS 22.0 for the Windows system (SPSS Inc. Chicago, IL, USA). Experimental results were summarized as mean value±standard deviation. The paired student T-test was used to analyze whether there is a statistical difference between the two different models. If the P value is less than 0.05, there is a statistical difference; otherwise, there is no difference. Meanwhile, the receiver operating characteristic (ROC) curve and 95% confidence interval (CI) was adopted to evaluate model performance.

3 Experimental results and analysis

In this paper, a total of 4099 CXR images collected from public data set [3, 29] were used to verify our designed model, including 1341 images from normal cases, 1464 images containing rotated images from COVID-19 cases, and 1294 images from other common pneumonia cases. In order to evaluate the universality and robustness of our model, we adopted the 10-fold cross-validation method to conduct the experiment and repeated it five times by changing the order of the dataset. The proportion of the training set, validation set and the testing set was divided into 8:1:1.

Our approach was implemented in Python framework. We ran the experiments on a workstation with Intel Xeon Gold 6248R CPU (3 GHz), 256 G RAM and Tesla V100 GPU.

3.1 Experimental results

1384 CXR images were randomly selected to generate one-to-one object detection labels for training the YOLO-v4 model with the backbone of pre-trained CSPDarknet, which was used to locate and segment the thoracic region of all CXR images. Detection results of ROI, which has been verified and confirmed by the clinician are shown in Fig. 3, which has a good detection effect for different types of CXR images. Then, the segmented ROI images were uniformly scaled to the same size of 256 × 256 and put into the designed classification model. After the convolution neural network in the first stage, 13 abstract transformation maps were obtained, which were combined with the original image to compose 14D matrixes as the input of multi-channel parallel pyramid MLP-Mixer modules, as shown in Fig. 4.

Training parameters of the classification model were set as follows. Learning rate: 0.0005, batch size: 4, optimizer: AdamW, data encoding method: one-hot encoding, and loss function: L_2 loss. After 30 epochs, the change of loss function in the training process of 5-fold cross-validation experiments is shown in Fig. 5a, with the mean loss value of 0.00894, which demonstrates that the proposed classification model achieved superior convergence effect. Additionality, the results of 5 repeated experiments and the confusion matrix of the testing set are shown in Table 3 and Fig. 5b. The average accuracy of the training set, validation set, and testing set reached 99.74%, 98.34%, and 98.25%, respectively. The experimental results demonstrate that the model has great stability and classification performance in the diagnosis of each type of three cases.

3.2 Model evaluation

In order to evaluate model performance, we made quantitative analysis of accuracy, precision, sensitivity, specificity, and F1-score (Eq. (3)) according to confusion matrix of each experiment. They are all positively correlated with the model performance, ranging from [0, 1]. The results of 5 repeated experiments are displayed in the form of mean value±standard deviation, as shown in Table 4. The data in the table indicates that the five evaluation indexes of this model had basically reached 98% or above, which verifies that our designed model has outstanding performance.

$$
\begin{aligned}
\text{Accuracy} &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}, \\
\text{Percision} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FR}}, \\
\text{Sensitivity} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \\
\text{Specificity} &= \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}, \\
\mathrm{F1}-\text{score} &= \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FN}+\mathrm{FP}}.
\end{aligned}
\tag{3}
$$

where TP, FP, TN, and FN mean true positive, false positive, true negative, and false negative, respectively.

3.3 Verification of chest region detection

Clinical diagnosis is mainly based on the left and right lung fields. In our proposed method, the target detection model was used to process the original image to obtain the lung field and reduced the influence of the surrounding area on the classification model. To illustrate the advantages of the trained classification model of the detected image, we used the original image to carry out the same feature analysis and classification experiments with the same training parameters, and the results are shown in Table 5.

Obviously, compared with using the original image directly, the image after target detection proposed in this paper had increased 1.85%, 1.70%, 0.92%, 1.82% and 1.81% respectively in accuracy, precision, specificity, recall and f1-score. It can be concluded that the image after target detection can eliminate the interference of

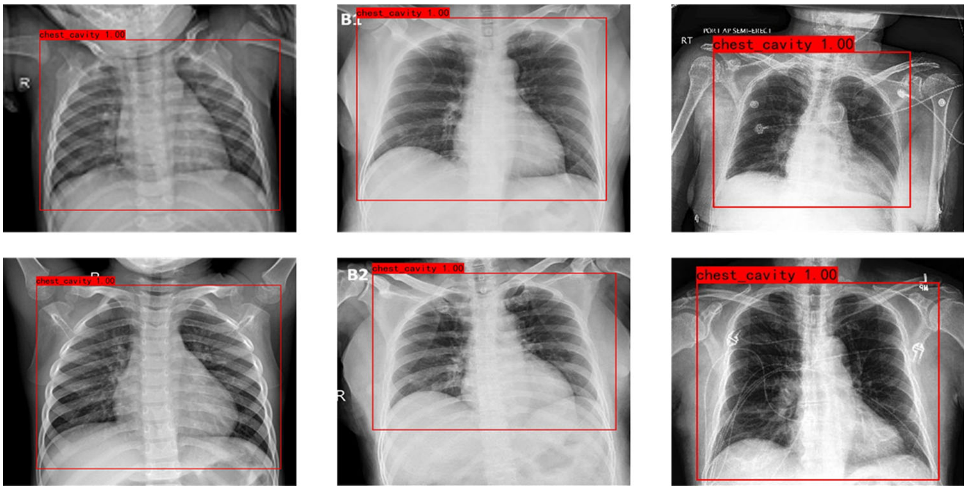
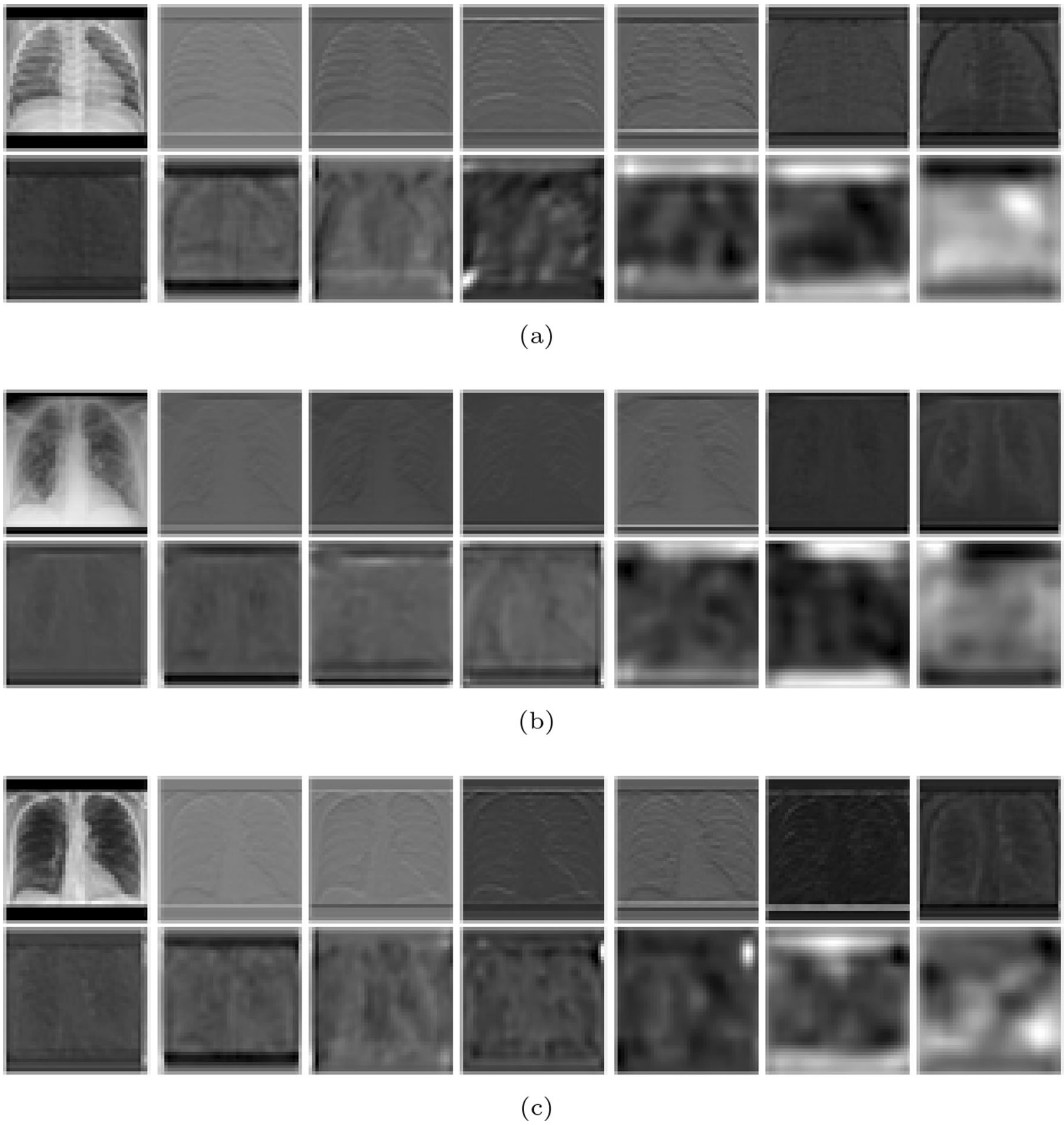


Fig. 3 Detection results of ROI in CXR images



(a)

(b)

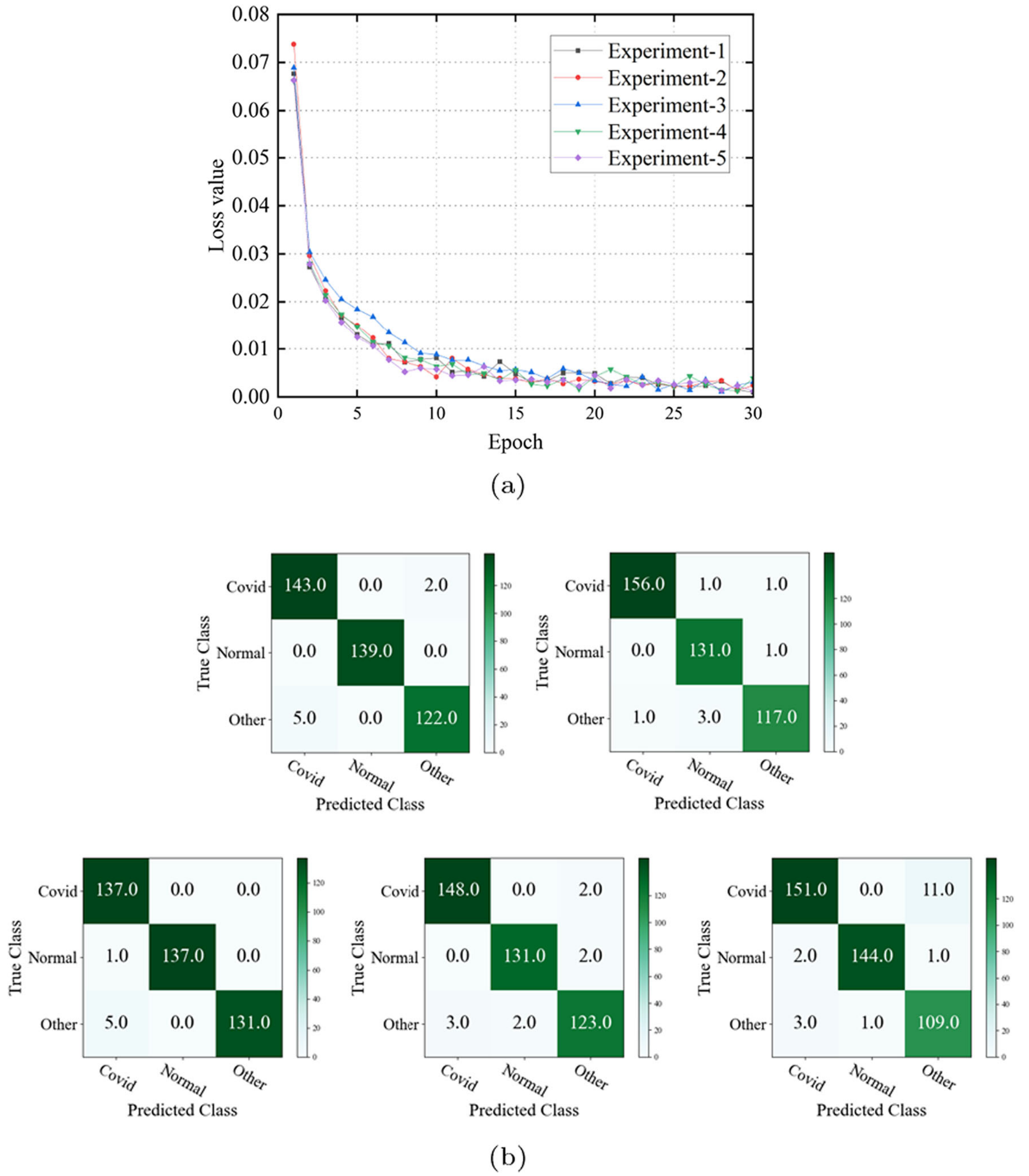
(c)

Fig. 4 14D transformation maps of **a** Normal, **b** COVID-19, **c** Other pneumonia

surrounding tissues and make the model more focused on the ROI area. In addition, the detection of the chest region, which only accounts for about 2/3 of the original image area, can save a certain amount of computing power and training time.

3.4 Varidation of number of deformable MLP-Mixer layer

After the deformable MLP-Mixer layer, the input was processed, with different sizes for the output. The more layers passed, the smaller the input to the next layer of the Mixer, which will affect the ability of the network for capturing features. Therefore, a certain number of experiments was conducted to verify the influence of the number of the deformable MLP-Mixer layer. Considering the size of the actual input image, we discuss the cases when the number of deformable MLP-Mixer layers is 2, 3, 4 and 5, respectively. ROC curves of models with different deformable MLP-Mixer layers are shown in Fig. 6, which shows that when the number was set as 3, there is the highest AUC value of 0.999 with 95% CI of 0.9989 to 0.9995. In addition, other five indices (i.e., accuracy, precision, specificity, recall and F1-Score) were adopted to evaluate classification preformation. Quantitative evaluation results of models with different numbers of deformable MLP-Mixer layers (two layers (2 L), three layers (3 L), four layers (4 L), and five layers (5 L)) are shown in



(a)

(b)

Fig. 5 Training process **a** and confusion matrix **b** in testing set

Table 6, and the student *t*-test between three layers and other cases is shown in Table 7.

It can be clearly concluded from Fig. 6 that compared with the model performance of other layers, 3 is the best choice for the number of deformable MLP-Mixer layers. On the premise of avoiding the structural redundancy and computational waste caused by too many layers (4 L and 5 L) and too small input, it also avoids the insufficient information extraction caused by fewer layers (2 L) and larger input. Meanwhile, we can notice from the data in Tables 5 and 6 that the classification performance of our proposed number of layers (3 L) has significantly improved on five indicators ($P<0.01$) compared with other layers. Although there is no significant improvement compared with 4 L ($P>0.05$), 3 L still has certain advantages in terms of time required and computational complexity. These results explain the rationality of our proposed model architecture.

Table 3 Experimental results of proposed models of 5 repeated experiments

Order	Accuracy/%		
	Train	Validation	Test
1	99.85	98.78	98.54
2	99.63	98.29	98.30
3	99.82	97.80	98.30
4	99.54	98.29	98.30
5	99.85	98.53	97.81
Average	**99.74**	**98.34**	**98.25**

Bold values show the best results of prediction

3.5 Comparison with other methods

For verifying the superiority of the classification model proposed in this paper, we compared our methods with seven prevalent deep learning models using the same data set and training parameters, including Visual attention network (VAN) [30], ConvNet [31], ResNet-50 [32], DenseNet [33], MobilenetV3 [34], Vision Transformer (ViT) [35], VGG-16 [36], and MLP-Mixer [26] ROC curves are show in Fig. 7, which shows that the method proposed in this paper has optimal classification effect, with average AUC gain of 0.094. Meanwhile, the results of quantitative evaluation using five indices (i.e., accuracy, precision, specificity, recall and F1-score) and paired t-test are shown in Tables 8 and 9, which proves that our method is superior to other network models in this classification task and has a significant improvement on five indices ($P < 0.001$), with an average gain of 11.51%, 11.97%, 6.22%, 12.19% and 12.82%, respectively.

4 Discussion

For eliminating the obstacles of existing models: manually designed or extracted features, insufficient feature extraction, and wasting computational power of integrated models, we proposed a novel classification method based on joint CNN features of CXR images and parallel pyramid MLP-Mixer module to distinguish the type of lung pneumonia, including normal cases, COVID-19, and other common pneumonia.

Our methods employed Yolo-V4 to segment the region of the thoracic cavity first, and segmented ROI images were put into a CNN model containing 13 convolution layers. The output maps of all convolution layers and segmented ROI images were combined to conform 14D image matrix which was used as the input of the following parallel pyramid MLP-Mixer feature extraction modules and classifier module. By the way of the 14D input, the expression of original CXR images was increased, which is

Table 4 Evaluation Results of proposed models of 5 repeated experiments

Model	Index (%)				
	Accuracy	Precision	Specificity	Recall	F1-score
1	98.54	98.60	99.28	98.53	98.53
2	98.30	98.24	99.16	98.22	98.22
3	98.30	98.34	99.17	98.27	98.27
4	98.30	98.40	99.17	98.14	98.25
5	97.81	97.79	98.91	97.75	97.77
Average	**98.25±0.27**	**98.27±0.30**	**99.14±0.14**	**98.17±0.28**	**98.21±0.28**

Bold values show the best results of prediction

Table 5 Evaluation Results of proposed model of 14D input without detection

Model	Index (%)				
	Accuracy	Precision	Specificity	Recall	F1-score
1	96.84	96.94	98.38	96.89	96.91
2	96.59	96.63	98.28	96.50	96.56
3	95.62	95.90	97.85	95.50	95.58
4	96.59	96.86	98.39	96.41	96.53
5	96.35	96.53	98.18	96.45	96.43
Average	96.40±0.47	96.57±0.41	98.22±0.22	96.35±0.51	96.40±0.49
Ours	**98.25±0.27**	**98.27±0.30**	**99.14±0.14**	**98.17±0.28**	**98.21±0.28**

Bold values show the best results of prediction

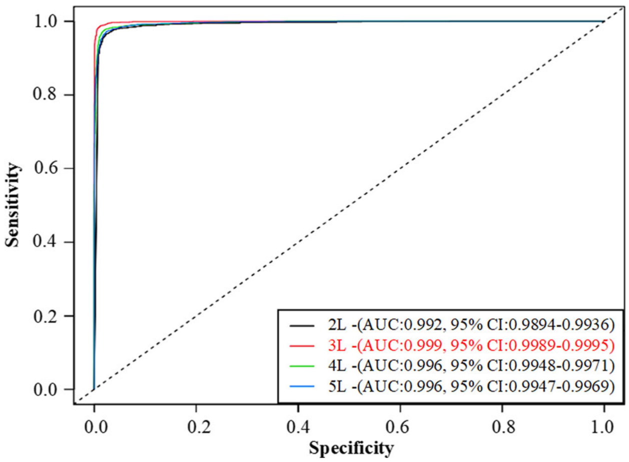


Fig. 6 ROC curves of models with different numbers deformable MLP-Mixer layers

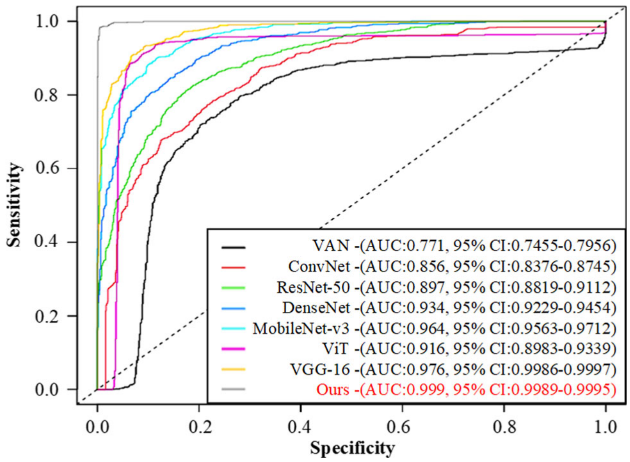


Fig. 7 ROC curves of models with different deep learning models

helpful to the training of classification models. The parallel pyramid MLP-Mixer feature extraction module was composed of three stacked deformable MLP-Mixer layers, which draw lessons from the concept of the pyramid and set different sizes for input and output, as shown in Table 2. As a result, it can more comprehensively extract the information of receptive fields of different sizes, thus enriching the extraction of feature information. Although MLP-Mixer did not perform position embeddings on the patch as in transformers, the token-mixing MLPs have considered the location information, which is more flexible. It also takes the channel mixing into account, which analyzed the correlation between different channels of the same patch. What's more, compared with some mainstream network frameworks, the MLP-based architecture is simpler, and MLP-Mixer also avoids the calculation of key, query and attention matrix.

Simultaneously, we also verified the necessity of target detection. Table 5 shows that effective ROI detection can help the model to further focus on the lesion area. Meanwhile, the number of MLP-Mixer layers was also discussed in this paper, different numbers may extract different features, and obviously different experimental results will be obtained. In other words, too many layers may lead to an excessive reduction of information and reduce the ability to capture information, and too few layers may result in incomplete feature processing, which is insufficient to support subsequent classification. Figure 6 and Table 6 proved that three MLP-Mixer layers could obtain optimal classification performance, which provides a reference for the application of this model to other classification tasks.

What's more, we compared our method with the mainstream methods in the experiment and made a quantitative evaluation of performance. All the experimental results in Fig. 7 and Table 8 show that the method proposed in this paper has better ability than other networks. It also explains the important value of the multi-model combination and

Table 7 Statistical difference between models with different numbers of deformable MLP-Mixer layers

Layer	T-test				
	Accuracy	Precision	Specificity	Recall	F1-score
2 L	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
3 L	–	–	–	–	–
4 L	0.077	0.080	0.071	0.086	0.091
5 L	0.006	0.005	0.005	0.006	0.005

Table 6 Performance comparison of models with different numbers of deformable MLP-Mixer layers

Layer	Index (%)				
	Accuracy	Precision	Specificity	Recall	F1-score
2 L	95.86	96.07	97.89	95.96	95.98
3 L	**98.25**	**98.27**	**99.14**	**98.17**	**98.21**
4 L	96.89	96.98	98.41	96.97	96.90
5 L	96.35	96.40	98.15	96.37	96.35
Average	98.25±0.27	98.27±0.30	99.14±0.14	98.17±0.28	98.21±0.28

Bold values show the best results of prediction

Table 8 Performance comparison of various deep learning models

Model	Index (%)				
	Accuracy	Precision	Specificity	Recall	F1-score
VAN	76.28	78.84	88.83	77.20	75.71
ConvNet	78.00	78.65	89.20	78.67	77.52
ResNet-50	82.89	82.69	91.26	83.00	82.73
DenseNet	88.51	80.06	89.76	79.62	78.65
MobileNetV3	89.24	89.91	94.51	89.32	89.46
ViT	91.44	91.97	95.83	91.84	91.44
VGG-16	91.81	92.46	96.07	92.23	91.86
MLP-Mixer	95.72	95.82	97.88	95.94	95.74
Ours	**98.25**	**98.27**	**99.14**	**98.17**	**98.21**

Bold values show the best results of prediction

Table 9 Statistical difference between our method and previous methods

Model	T-test				
	Accuracy	Precision	Specificity	Recall	F1-score
VAN	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
ConvNet	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
ResNet-50	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
DenseNet	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
MobileNetV3	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
ViT	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
VGG-16	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
MLP-Mixer	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Ours	–	–	–	–	–

multi-information fusion in auxiliary diagnosis research, and our proposed model is progressiveness and potential in the clinical auxiliary diagnosis of pneumonia.

Although our proposed method has obtained good performance on the auxiliary diagnosis of different types of pneumonia, there are some limitations. Firstly, our method realized full extraction of image features by integrating the channel information of the same patch and the spatial information of different patches. However, non-overlapping patch sampling may cause the loss of edge information between adjacent patches. Secondly, in order to exclude the influence of areas outside the lung parenchyma, it was necessary to segment the lung parenchyma in advance. However, due to the poor quality of lung parenchyma imaging in CT images of severe patients, this paper only adapts the YOLO-V4 model to cut the approximate area of the chest, rather than segment the lung parenchyma area. Thus, there were still some interference areas. Thirdly, we used public data sets to verify the model, which may have some errors in the actual clinical application.

5 Conclusions

In this paper, we proposed a novel deep learning-based method to predict COVID-19 and other pneumonia on CXR images. This method adopted CNN to obtain more representations of the original images for enriching image information. And through the self-designed parallel pyramid MLP-Mixer module, the comprehensive feature analysis of the multi-channel input image matrix was analyzed to achieve the accurate classification of pneumonia diseases. Experimental results showed that our method has good classification performance and is better than these mainstream methods. To conclude, the method proposed in this paper could be regarded as an important diagnostic tool for COVID-19 pneumonia, with great application potential in clinics.

In the future, we will introduce the attention module into our model to pay attention to those important features. Meanwhile, we will study the deep learning-based unsupervised segmentation model to obtain the accurate lung parenchyma region, thus, the pre-processed images can be better used for model training and obtain a better model performance. Additionally, we will also collect more clinical data to train and verify our model for serving clinical diagnosis.

Acknowledgements This work is supported by the National Natural Science Foundation of China (61971121).

Data availability statements The datasets analyzed during the current study are available in the following Kaggle repository and COVID-XRay-5K Dataset, respectively:https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia and https://github.com/shervinmin/DeepCovid/tree/master/data.

Declarations

Conflict of interest we certify that there is no actual or potential conflict of interest in relation to this article.

References

1. Xing W, He C, Li J, Qin W, Yang M, Li G, Li Q, Ta D, Wei G, Li W et al (2022) Automated lung ultrasound scoring for evaluation of coronavirus disease 2019 pneumonia using two-stage cascaded deep learning model. Biomed Sig Process Cont 75:103561
2. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R, et al (2020) A novel coronavirus from patients with pneumonia in china, 2019. New England J Med
3. Minaee S, Kafieh R, Sonka M, Yazdani S, Soufi GJ (2020) Deep-covid: Predicting covid-19 from chest x-ray images using deep transfer learning. Med Img Anal 65:101794
4. Rahimi I, Chen F, Gandomi AH (2021) A review on covid-19 forecasting models. Neural Comput Appl 4:1–11
5. Tan L, Yu K, Bashir AK, Cheng X, Ming F, Zhao L, Zhou X (2021) Toward real-time and efficient cardiovascular monitoring for covid-19 patients by 5g-enabled wearable medical devices: a deep learning approach. Neural Comput Appl. 1–14
6. Canayaz M, Şehribanoğlu S, Özdağ R, Demir M (2022) Covid-19 diagnosis on ct images with bayes optimization-based deep neural networks and machine learning algorithms. Neural Comput Appl 34(7):5349–5365
7. Garain A, Basu A, Giampaolo F, Velasquez JD, Sarkar R (2021) Detection of covid-19 from ct scan images: A spiking neural network-based approach. Neural Comput Appl 33(19):12591–12604
8. D'Angelo G, Palmieri F (2021) Enhancing covid-19 tracking apps with human activity recognition using a deep convolutional neural network and har-images. Neural Comput Appl, 1–17
9. Cong R, Zhang Y, Yang N, Li H, Zhang X, Li R, Chen Z, Zhao Y, Kwong S (2022) Boundary guided semantic learning for real-time covid-19 lung infection segmentation system. IEEE Trans Consumer Electron 68(4):376–386
10. Hassantabar S, Stefano N, Ghanakota V, Ferrari A, Nicola GN, Bruno R, Marino IR, Hamidouche K, Jha NK (2021) Coviddeep: Sars-cov-2/covid-19 test based on wearable medical sensors and efficient neural networks. IEEE Transact Consumer Electron 67(4):244–256
11. Zhou Z, Chen X, Zhang Y, Hu D, Qiao L, Yu R, Yap P-T, Pan G, Zhang H, Shen D (2020) A toolbox for brain network construction and classification (brainnetclass). Hum Brain Mapp 41(10):2808–2826
12. Lin M, Jiang M, Zhao M, Ukwatta E, White JA, Chiu B (2022) Cascaded triplanar autoencoder m-net for fully automatic segmentation of left ventricle myocardial scar from three-dimensional late gadolinium-enhanced mr images. IEEE J Biomed Health Inform 26(6):2582–2593
13. Sitaula C, Hossain MB (2021) Attention-based vgg-16 model for covid-19 chest x-ray image classification. Appl Intell 51(5):2850–2863
14. Khan IU, Aslam N, Anwar T, Alsaif HS, Chrouf SMB, Alzahrani NA, Alamoudi FA, Kamaleldin MMA, Awary KB (2022) Using a deep learning model to explore the impact of clinical data on covid-19 diagnosis using chest x-ray. Sensors 22(2):669
15. Qi X, Brown LG, Foran DJ, Nosher J, Hacihaliloglu I (2021) Chest x-ray image phase features for improved diagnosis of covid-19 using convolutional neural network. Int J Comput Assisted Radiol Surgery 16(2):197–206
16. Lin Z, He Z, Xie S, Wang X, Tan J, Lu J, Tan B (2021) Aanet: adaptive attention network for covid-19 detection from chest x-ray images. IEEE Transact Neural Netw Learn Syst 32(11):4781–4792
17. Zargari Khuzani A, Heidari M, Shariati SA (2021) Covid-classifier: An automated machine learning model to assist in the diagnosis of covid-19 infection in chest x-ray images. Sci Rep 11(1):1–6
18. Babukarthik R, Adiga VAK, Sambasivam G, Chandramohan D, Amudhavel J (2020) Prediction of covid-19 using genetic deep learning convolutional neural network (gdcnn). IEEE Access 8:177647–177666
19. Shorfuzzaman M, Hossain MS (2021) Metacovid: A siamese neural network framework with contrastive loss for n-shot diagnosis of covid-19 patients. Pattern Recognit 113:107700
20. Li D, Li S (2022) An artificial intelligence deep learning platform achieves high diagnostic accuracy for covid-19 pneumonia by reading chest x-ray images. Iscience 25(4):104031
21. Salvatore C, Interlenghi M, Monti CB, Ippolito D, Capra D, Cozzi A, Schiaffino S, Polidori A, Gandola D, Alì M et al (2021) Artificial intelligence applied to chest x-ray for differential diagnosis of covid-19 pneumonia. Diagnostics 11(3):530
22. Gayathri J, Abraham B, Sujarani M, Nair MS (2022) A computer-aided diagnosis system for the classification of covid-19 and non-covid-19 pneumonia on chest x-ray images by integrating cnn with sparse autoencoder and feed forward neural network. Comput Biol Med 141:105134
23. Avola D, Bacciu A, Cinque L, Fagioli A, Marini MR, Taiello R (2022) Study on transfer learning capabilities for pneumonia classification in chest-x-rays images. Comput Methods Programs Biomed 221:106833
24. Chetoui M, Akhloufi MA (2022) Explainable vision transformers and radiomics for covid-19 detection in chest x-rays. J Clin Med 11(11):3013
25. Bochkovskiy A, Wang C-Y, Liao H-YM (2020) Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934
26. Tolstikhin IO, Houlsby N, Kolesnikov A, Beyer L, Zhai X, Unterthiner T, Yung J, Steiner A, Keysers D, Uszkoreit J et al (2021) Mlp-mixer: an all-mlp architecture for vision. Adv Neural Inform Process Syst 34:24261–24272
27. Xing W, Zhu Z, Hou D, Yue Y, Dai F, Li Y, Tong L, Song Y, Ta D (2022) Cm-segnet: a deep learning-based automatic segmentation approach for medical images by combining convolution and multilayer perceptron. Comput Biol Med 147:105797
28. Liu R, Li Y, Tao L, Liang D, Zheng H-T (2022) Are we ready for a new paradigm shift? a survey on visual deep mlp. Patterns 3(7):100520
29. Tahir AM, Chowdhury ME, Khandakar A, Rahman T, Qiblawey Y, Khurshid U, Kiranyaz S, Ibtehaz N, Rahman MS, Al-Maadeed S et al (2021) Covid-19 infection localization and severity grading from chest x-ray images. Comput Biol Med 139:105002

30. Guo M-H, Lu C-Z, Liu Z-N, Cheng M-M, Hu S-M (2022) Visual attention network. arXiv preprint arXiv:2202.09741
31. Liu Z, Mao H, Wu C-Y, Feichtenhofer C, Darrell T, Xie S (2022) A convnet for the 2020s. In: 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 11966–11976
32. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), pp 770–778
33. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 2261–2269
34. Howard A, Sandler M, Chen B, Wang W, Chen L-C, Tan M, Chu G, Vasudevan V, Zhu Y, Pang R, Adam H, Le Q (2019) Searching for mobilenetv3. In: 2019 IEEE/CVF international conference on computer vision (ICCV), pp 1314–1324
35. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al (2020) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929
36. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.