An end-to-end framework combining multiple deep learning modules for diagnosing COVID-19 pneumonia

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2023

Abstract

To fully extract the feature information of lung parenchyma in Chest X-ray images and realize the auxiliary diagnosis of COVID-19 pneumonia, this paper proposed an end-to-end deep learning model, which is mainly composed of object detection, depth feature generation, and multi-channel fusion classification. Firstly, the convolutional neural network (CNN) and region proposal network (RPN)-based object detection module was adopted to detect chest cavity region of interest (ROI). Then, according to the obtained coordinate information of ROI and the convolution feature map of original image, the new convolution feature maps of ROI were obtained with number of 13. By screening 4 representative feature maps form 4 convolution layers with different receptive fields and combining with original ROI image, the 5-dimensional (5D) feature maps were generated as the multi-channel input of classification module. Moreover, in each channel of classification module, three pyramidal recursive MLPs were employed to achieve cross-scale and cross-channel feature analysis. Finally, the correlation analysis of multi-channel output was realized by bi-directional long short memory (Bi-LSTM) module, and the auxiliary diagnosis of pneumonia disease was realized through fully connected layer and SoftMax function. Experimental results show that the proposed model has better classification performance and diagnosis effect than previous methods, with great clinical application potential.

Keywords: Pneumonia, end-to-end framework, multi-task learning, MLP-Bi-LSTM, deep learning.

1. Introduction

COVID-19 was discovered at the end of 2019 and began to break out in a large scale in 2020. So far, it is still a powerful and threatening infectious disease that plagues the lives and health of people all over the world (Frid-Adar et al., 2021). At present, the detection of COVID-19 diagnosis mainly depends on antigen and nucleic acid detection(Ji et al., 2020). They all have some limitation, such as high false positive or low efficiency. Due to the timeliness and visibility of medical imaging, the combination with artificial intelligence has become an important means for detecting COVID-19 in clinic.

With the advantages of fast, low radiation, and high resolution, Chest X-ray (CXR) has been widely used in the diagnosis of patients with COVID-19 and other pneumonia. At present, there are many researches based on CXR images and artificial intelligence, and most of them used deep learning as the main diagnosis framework. For example, Monday et al. (Monday et al., 2022) proposed a super-resolution-based Siamese wavelet multiresolution CNN model for COVID-19 diagnosis. Fan et al. (Fan et al., 2021) designed a multi-kernel-size spatial-channel attention deep learning model to detect COVID-19 pneumonia. Nikolaou et al. (Nikolaou et al., 2021) proposed a new CNN framework by adding

© 2023 Anonymous.

a dense layer to the pre-training baseline CNN (EfficientNet-B0) to achieve high-precision classification of COVID-19. Priyatharshini et al. (Priyatharshini et al., 2021) used U-Net to segment the CXR image, and then used Inception v3 model to train the classification model to distinguish COVID–19. Li et al. (Li et al., 2021)proposed a capsule network model with multi-head attention routing algorithm (MHA-CoroCapsule) for accurate diagnostics of COVID-19. Agarwal et al. (Agarwal et al., 2021) proposed a two-stage deep learning diagnosis framework, CoroNet, to automatically perform the diagnosis task of COVID-19. Park et al. (Park et al., 2021) proposed a multi-scale class residual attention (MCRA) network based on CXR images, which can effectively detect COVID-19.

Although above researches have achieved good effect in the diagnosis of COVID-19, there are still some limitations that limit their clinical application. For example, for those end-to-end diagnostic models composed of a single depth learning module, the CXR images are directly processed, which cannot avoid the influence of other tissues in chest region. Moreover, the feature analysis using convolution alone cannot achieve a comprehensive analysis, and the extracted features have not been filtered, so there is some redundancy for subsequent feature analysis and calculation. Therefore, in order to address those problems, this paper proposed an end-to-end deep learning model based on the combination of convolution, multilayer perceptron (MLP) and bidirectional long short memory (Bi-LSTM) modules to achieve the auxiliary diagnosis of COVID-19 and other pneumonia. The main contributions are summarized as follows.

- The end-to-end model can achieve the tasks of object detection, feature extraction, classification and recognition synchronously. Compared with the multi-stage cascade model, it has better operational efficiency and can better meet the clinical needs.
- By embedding RPN-based object detection module in the end-to-end model, it realizes the precise positioning of the chest region and reduces the interference of other regions and the calculation of subsequent features.
- The convolution module with different receptive fields and the multi-scale MLP module achieve more sufficient local and global information extraction of CXR images. Bi-LSTM also increases the correlation of the output of different channel, improving the accuracy of auxiliary diagnosis.

2. Methodology

2.1. Overall framework

This paper proposed a novel end-to-end classification model for diagnozing COVID-19 and other pneumonia, which is mainly composed of a multilayer CNN-based object detection and feature map generation module and MLP-Bi-LSTM-based classification module, as shown in Figure 1.

2.2. Multi-task learning module for object detection and feature map generation

Due to the influence of patients' differences, operating habits, and types of instruments, the quality of CXR obtained from patients is different, and the proportion of chest ROI in

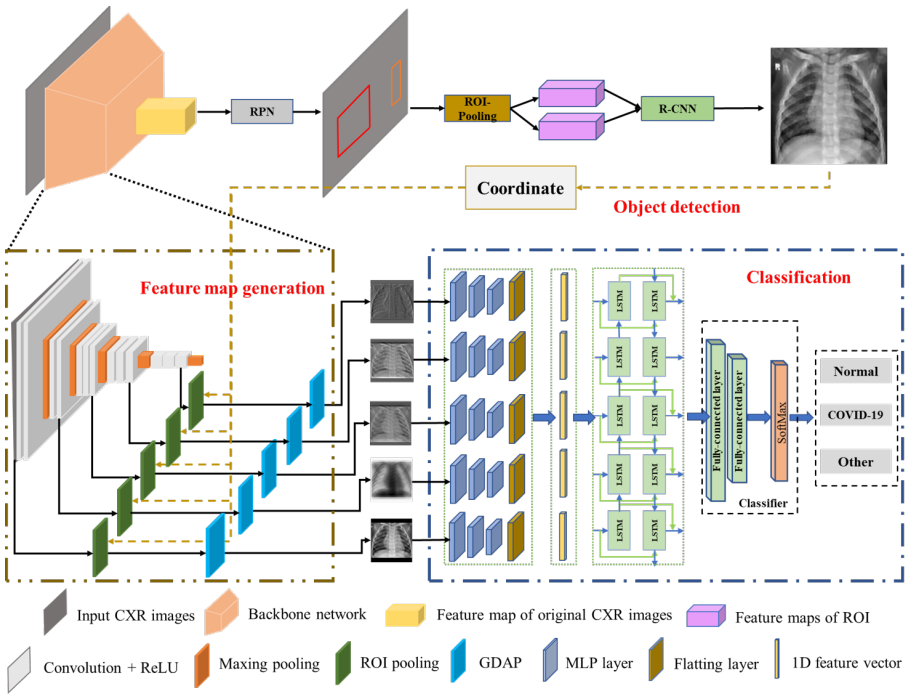


Figure 1: Schematic diagram of end-to-end classification framework

CXR are different. In order to reduce the impact of these factors, follow-up research needs to be carried out on the basis of accurate positioning of the thoracic cavity. Therefore, this paper designed a multi-task learning module that integrates object detection and feature generation.

For object detection of thoracic cavity ROI, feature map of original image was firstly generated through CNN backbone module, which was put into Region Proposal Network (RPN) to obtain the proposal box of object area. Then, according to the combination of proposal box and feature map, ROI pooling was employed to re-extract a fixed size feature map for each proposal box. Next, according to these feature maps, the objects in the proposed box were classified (i.e., background or chest) through fully connected layer and SoftMax function. Finally, the boundary information of ROI was further adjusted according to the predicted category to obtain accurate coordinate information.

For feature map generation, according to the obtained coordinate information and the convolution feature map of original image, the new feature maps of ROI were obtained. The CNN backbone model used in this paper contains 13 convolution layers. The convolution layers of four different receptive fields was screened for selecting the feature map of the last layer of each receptive field, where a Global Depth Average Pooling (GDAP) layer was designed to change the feature map with the size of (c, h, w) to a single channel feature map with the size of (h, w). Finally, the ROI feature maps of four convolution layer with different receptive field were combined with original ROI image to generate a new 5D feature input for MLP-Bi-LSTM classification model.

2.3. MLP-Bi-LSTM classification module

MLP-Mixer was proposed by Google Brain Team in 2021 (Tolstikhin et al., 2021) which is fully based on MLP. It has been applied in many fields, which put the token of input images into token-mixer layer and channel-mixer layer in turn to mix the information of both feature channels and spatial locations.

Inspired by this, we proposed a pyramid MLP-Bi-LSTM (PMBL) classification module with multi-channel fusion under multidimensional input to achieve cross-channel and cross-space feature analysis and classification of input, as shown in Figure 1. The 5-D feature input generated by the multi-task module was input into three serial MLP layers with different scales in parallel to analyze the channel and spatial information of the input. Then the output of the MLP layers was flattened, and five parallel outputs were spliced to generate sequence information, which is transmitted to Bi-LSTM to explore the correlation between its channels. Finally, the classification detection of COVID-19 patients was completed through two full connection layers and SoftMax layer.

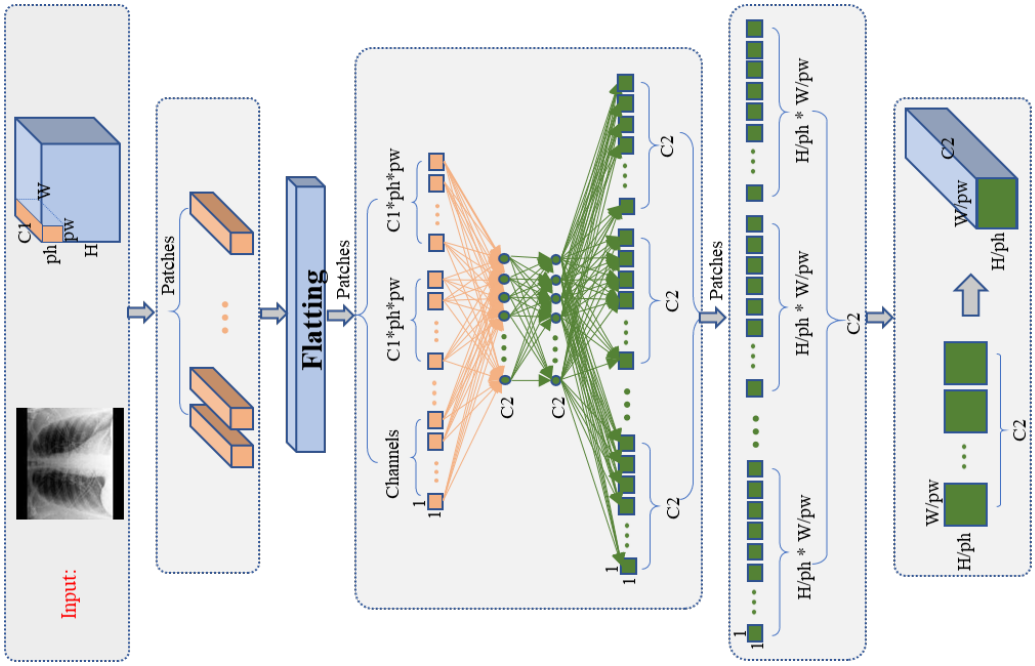


Figure 2: Schematic diagram of the structure of the MLP layer

The structure of the MLP layer is shown in Figure 2. Assuming the size of the input feature is $(b, C1, H, W)$ in each MLP layer. Firstly, the dimension of the feature map was reduced. The input was divided into patches with the size of (ph, pw), and then the patches were expanded into one-dimensional vectors The input was converted from $(b, C1, H, W)$ to $(b, H/ph \times W/pw, ph \times pw \times C2)$ to realize the dimension reduction operation. Then the feature extraction of one-dimensional vector was realized through the MLP layer composed of two layers of fully connected layers with $C2$ neurons. Finally, through the dimension elevation layer, we completed the spatial information reorganization of the feature by restor-

ing $(b, H/ph \times W/pw, C2)$ to the data form consistent with the original input dimension $(b, C2, H/ph, W/pw)$, as the input of the next MLP layer. Here, $ph = pw = 2$, and the input patch sizes of the three recursive MLP layers are $n1 \times n1$, $n2 \times n2$, $n3 \times n3$ respectively, where $n3 = 2 \times n2 = 4 \times n1$. Assume the size of input feature map was $M \times M \times 1$, the output sizes of three pyramidal recursive MLPs are $M/2 \times M/2 \times 16$, $M/4 \times M/4 \times 32$, and $M/8 \times M/8 \times 64$.

3. Experiments and Results

Experimental dataset used in this paper includes 1341 cases of normal, 183 cases of COVID-19 patients, and 1294 cases of other pneumonia patients. To balance the number of samples in different categories, the images of COVID-19 cases were expanded to 1464 by random angle rotation. As a result, a total of 4099 images (Normal: 32.71%, COVID-19: 35.72%, Other: 31.57%) were used in this experiment. Meanwhile, 10-fold cross validation method was used in experiments, which was repeated 5 times, with the mean value as final results.

3.1. Object detection and feature map generation

Through the multi-task learning module combining the pre-trained 13-layer convolutional neural network and RPN, the 13D depth feature map of thoracic cavity ROI can be generated. Meanwhile, in order to reduce the computational complexity, the characteristics of four different receptive fields were screened in experiments. Then through the GDAP layer, four fixed-size outputs of different depths were obtained, which were combined with original ROI image to form a 5D feature input, as shown in Figure 3.

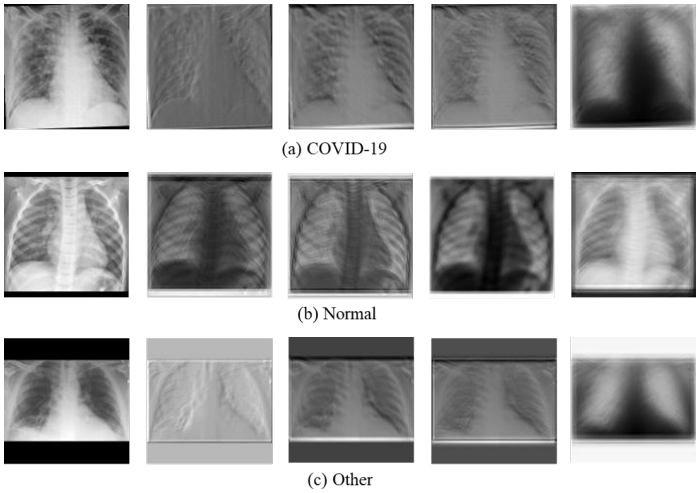
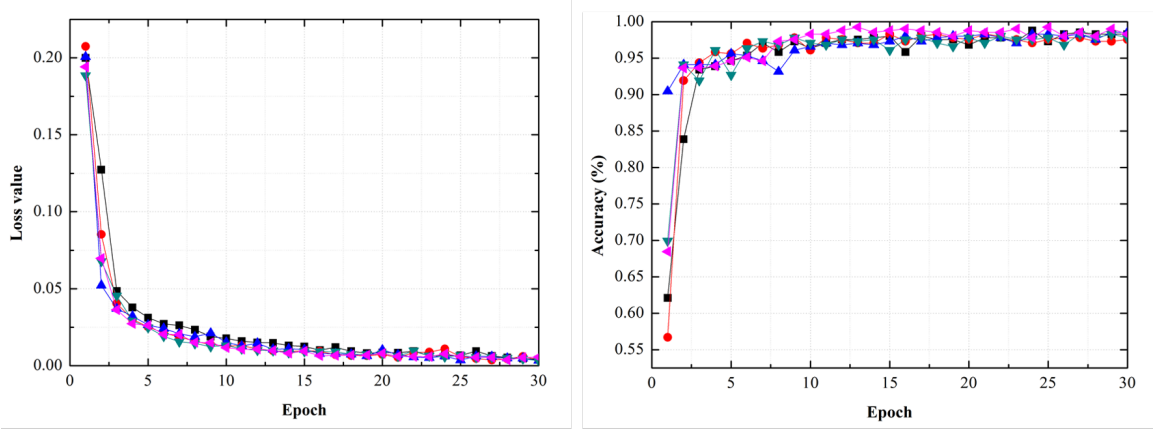


Figure 3: Display of 5D feature maps.

3.2. Display of classification results

The above 5D features were input into the multi-channel recursive mlp-bilstm classification module, with training hyperparameter of epoch, learning rate, batch size, and optimizer

being 30, 0.0005, 32, and AdamW, respectively. The data encoding method adapt one-hot encoding for calculating cross entropy loss function value to obtain the classification results for normal, COVID-19 and other pneumonia. The change of loss value and accuracy during the training process is shown in Figure 4. The experimental results show that the proposed classification model can converge well and has great robustness, with mean accuracy value reaching 98.10%.



(a) Variation of loss value (b) Variation of accuracy

Figure 4: Training of 5-D feature maps input

After five repeated experiments, mean value of five indicators (i.e., Accuracy, Precision, Specificity, Recall, and F1-socre) were obtained to quantitatively evaluate the classification performance. Experimental results are shown in Table 1, which show that the proposed model has superior classification performance, with five indicators being 98.10±0.68, 98.13±0.67, 99.03±0.35, 98.30±0.83, and 98.12±0.67, respectively.

Table 1: Quantitative evaluation results of the proposed model.

Order	Indicator value (%)				
	Accuracy	Precision	Specificity	Recall	F1-score
1	98.54	98.57	99.23	99.35	98.61
2	97.81	97.69	98.88	97.90	97.75
3	97.81	98.00	98.91	97.85	97.89
4	99.03	99.01	99.52	98.99	99.00
5	97.32	97.36	98.61	97.40	97.36
Avreage±STD	98.10±0.68	98.13±0.67	99.03±0.35	98.30±0.38	98.12±0.67

3.3. Analysis of model ablation experiment

3.3.1. Feasibility of 5D feature selection strategy

In order to verify the feasibility of 5D input composed of the 4D feature map selected from four convolutional layers with different receptive fields and original image, we compared 5D input with original 14D input (Figure 5), which was composed 13D feature map output from CNN backbone and original image. The performance comparison of the classification model trained through 14D and our proposed model was shown in Table 2, with gain of 1.02%, 1.03%, 0.48%, 1.29%, and 1.09% respectively in accuracy, precision, specificity, recall, and F1-score, which proved that the 5D feature used in the proposed model can replace the 14D feature, which can not only reduce calculation complexity, but also improve the classification accuracy.

3.3.2. Effectivity of Bi-LSTM module

In the classification model, a feature extraction module based on recursive MLP was designed to carry out more in-depth feature analysis of multi-channel input images, and Bi-LSTM was adopted to enhance the correlation between different input channels. In order to validate the effectivity of Bi-LSTM module, this paper conducted the same experiment using 5D input on the classification model without Bi-LSTM module. The performance comparison with our proposed model are shown in Table 3, with gain of 3.80%, 3.66%, 1.91%, 4.08%, and 3.85% respectively in accuracy, precision, specificity, recall, and F1-score, which proved that the Bi-LSTM module can capture the correlation between different input channels and greatly improves the model's classification performance.

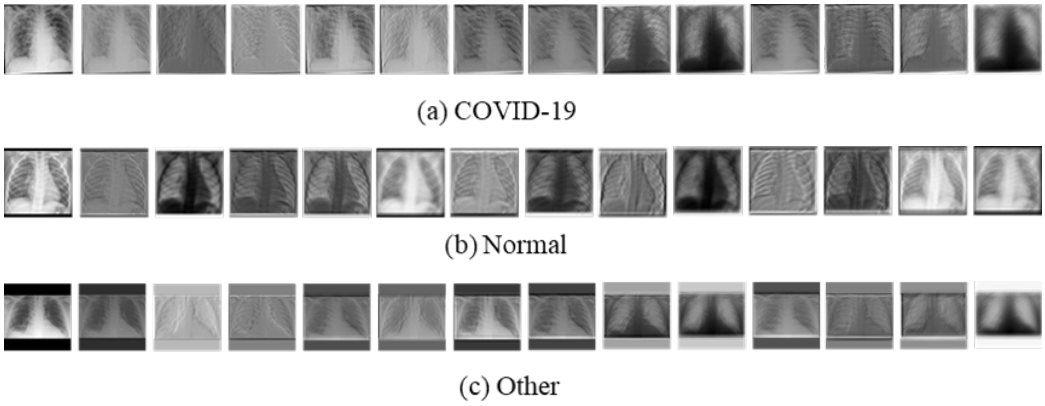


Figure 5: Display of 14D feature maps.

3.4. Comparison with other deep learning models

In order to verify the superiority of the method proposed in this paper, it was compared with previous classical deep learning models, such as Visual attention network (VAN) (Guo et al., 2022), ConvNet (Liu et al., 2022), MobileNet-V3 (Howard et al., 2019), DenseNet (Huang et al., 2017), MLP-Mixer (Tolstikhin et al., 2021), and Vision Transformer (ViT)

(Dosovitskiy et al., 2020). Experimental results were shown in Table 4, which proved that the proposed model has significantly improved performance and can more accurately realize the auxiliary diagnosis of COVID-19, with great clinical application potential.

Table 2: Comparison of 14-D feature input and our model.

Model	Indicator value (%)				
	Accuracy	Precision	Specificity	Recall	F1-score
14D	97.08	97.10	98.55	97.01	97.03
Ours	98.10	98.13	99.03	98.30	98.12
Gain	1.02	1.03	0.48	1.29	1.09

Table 3: Quantitative evaluation results of the 5D feature inputs without Bi-LSTM.

Model	Indicator value (%)				
	Accuracy	Precision	Specificity	Recall	F1-score
Without Bi-LSTM	94.30	94.47	97.12	94.22	94.27
Ours	98.10	98.13	99.03	98.30	98.12
Gain	3.80	3.66	1.91	4.08	3.85

Table 4: Comparative results of various models.

Model	Indicator value (%)				
	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
MobileNet-V3	89.24	89.91	94.51	89.32	89.46
DenseNet	85.21	86.5	92.74	85.66	85.24
MLP-Mixer	95.23	95.34	97.64	95.5	95.25
ViT	90.46	91.35	95.53	91.02	90.37
Ours	98.10	98.13	99.03	98.30	98.12

4. Conclusion

In this paper, we proposed a novel end-to-end deep learning framework to achieve rapid and accurate diagnosis of COVID-19. This framework realized the integration of object detection and feature generation, which can effectively eliminate the interference of other regions and generate multi-source feature maps. Meanwhile, in the pyramid MLP-Bi-LSTM classification module with multi-channel input, not only the deep abstract feature extraction of the multi-channel input was completed, but also the correlation between multi-channel input was combined, which further enhanced the information analysis of proposed model on CXR image, and realized the high-precision classification diagnosis of COVID-19, with great clinical application potential.

References

Chirag Agarwal, Shahin Khobahi, Dan Schonfeld, and Mojtaba Soltanalian. Coronet: a deep network architecture for enhanced identification of covid-19 from chest x-ray images. In *Medical Imaging 2021: Computer-Aided Diagnosis*, volume 11597, pages 484–490. SPIE, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Yuqi Fan, Jiahao Liu, Ruixuan Yao, and Xiaohui Yuan. Covid-19 detection from x-ray images using multi-kernel-size spatial-channel attention network. *Pattern Recognition*, 119:108055, 2021.

Maayan Frid-Adar, Rula Amer, Ophir Gozes, Jannette Nassar, and Hayit Greenspan. Covid-19 in cxr: From detection and severity scoring to patient disease monitoring. *IEEE journal of biomedical and health informatics*, 25(6):1892–1903, 2021.

Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu. Visual attention network. *arXiv preprint arXiv:2202.09741*, 2022.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Tianxing Ji, Zhenwei Liu, GuoQiang Wang, Xuguang Guo, Changchun Lai, Haoyu Chen, Shiwen Huang, Shaomei Xia, Bo Chen, Hongyun Jia, et al. Detection of covid-19: A review of the current literature and future perspectives. *Biosensors and Bioelectronics*, 166:112455, 2020.

Fudong Li, Xingyu Lu, and Jianjun Yuan. Mha-corocapsule: Multi-head attention routing-based capsule network for covid-19 chest x-ray image classification. *IEEE Transactions on Medical Imaging*, 41(5):1208–1218, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

Happy Nkanta Monday, Jianping Li, Grace Ugochi Nneji, Saifun Nahar, Md Altab Hossin, Jehoiada Jackson, and Chukwuebuka Joseph Ejiyi. Covid-19 diagnosis from chest x-ray images using a robust multi-resolution analysis siamese neural network with super-resolution convolutional neural network. *Diagnostics*, 12(3):741, 2022.

Vasilis Nikolaou, Sebastiano Massaro, Masoud Fakhimi, Lampros Stergioulas, and Wolfgang Garn. Covid-19 diagnosis from chest x-rays: developing a simple, fast, and accurate neural network. *Health information science and systems*, 9(1):1–11, 2021.

Junghoon Park, Il-Youp Kwak, and Changwon Lim. A deep learning model with self-supervised learning and attention mechanism for covid-19 diagnosis using chest x-ray images. *Electronics*, 10(16):1996, 2021.

R Priyatharshini, Ram AS Aswath, MN Sreenidhi, Samyuktha S Joshi, and Reshmika Dhandapani. An efficient approach for automatic detection of covid-19 using transfer learning from chest x-ray images. In *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*, pages 741–746. IEEE, 2021.

Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 34:24261–24272, 2021.