J. Shanghai Jiao Tong Univ. (Sci.), 2022, 27(1): 70-80
https://doi.org/10.1007/s12204-021-2392-3

Multi-Model Ensemble Deep Learning Method to Diagnose COVID-19 Using Chest Computed Tomography Images

WANG Zhiming[1] (王志明), *DONG Jingjing*[2,3*] (董静静), *ZHANG Junpeng*[1*] (张军鹏)
(1. College of Electrical Engineering, Sichuan University, Chengdu 610056, China; 2. Key Laboratory of Aerospace Medicine of Ministry of Education, Air Force Medical University, Xi'an 710032, China; 3. Lintong Rehabilitation and Recuperation Center, PLA Joint Logistic Support Force, Xi'an 710600, China)

© Shanghai Jiao Tong University and Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract: Deep learning based analyses of computed tomography (CT) images contribute to automated diagnosis of COVID-19, and ensemble learning may commonly provide a better solution. Here, we proposed an ensemble learning method that integrates several component neural networks to jointly diagnose COVID-19. Two ensemble strategies are considered: the output scores of all component models that are combined with the weights adjusted adaptively by cost function back propagation; voting strategy. A database containing 8 347 CT slices of COVID-19, common pneumonia and normal subjects was used as training and testing sets. Results show that the novel method can reach a high accuracy of 99.37% (recall: 0.998 1; precision: 0.989 3), with an increase of about 7% in comparison to single-component models. And the average test accuracy is 95.62% (recall: 0.958 7; precision: 0.955 9), with a corresponding increase of 5.2%. Compared with several latest deep learning models on the identical test set, our method made an accuracy improvement up to 10.88%. The proposed method may be a promising solution for the diagnosis of COVID-19.
Key words: COVID-19, deep learning, computed tomography (CT) images, ensemble model, convolutional neural network
CLC number: TP 183, R 445 **Document code:** A

0 Introduction

COVID-19 has caused a large number of infections and deaths, since it broke out, and it is spreading globally. Real-time polymerase chain reaction (RT-PCR) is considered to be the gold standard for diagnosing COVID-19, but its high false positive and unsatisfactory detection efficiency hinder its rapid detection of suspicious cases[1-2]. COVID-19 mainly causes lung infection; pathological features, such as ground-glass opacity (GGO) and lung consolidation, can be found on computed tomography (CT) scans[3-4]. An experienced doctor can capture these characteristics and make judgments; however, visual inspection is time consuming and prolonged concentration can easily lead to misjudgment. In contrast, computer-aided diagnosis (CAD) can make up for the lack of professional physicians and improve inspection efficiency, and deep learning is a promising approach for intelligent assisted diagnosis.

Recent studies have shown the outstanding performance of deep learning in the diagnosis of COVID-19[5-6]. Popular convolutional neural networks (CNNs) in the field of image recognition can distinguish between COVID-19 CT slices and others[7-9], even reaching or surpassing human experts in some aspects[10]. Recently, some novel neural networks with attention or auxiliary enhancement mechanisms have been proposed, enhancing the robustness of deep learning to complex samples. Shi et al.[11] proposed an attention transfer deep neural network (DNN) that used a variable attention module to enhance the response of infected areas. Li et al.[12] proposed a comparative multi-task convolutional neural network (CMT-CNN), and their research showed that simple auxiliary tasks could strongly enhance generalization ability. In addition, some smarter multi-task networks that can diagnose and locate COVID-19 at the same time also aroused interest[13-14].

As a kind of combinational optimization learning method, ensemble learning can efficiently solve practical application problems[15]. Related studies have shown that simply training several neural networks and

Received: 2021-01-07 **Accepted:** 2021-06-07
Foundation item: the Sichuan Science and Technology Department Research and Development Key Project (No. 21ZDYF3607), the Weining Cloud Hospital Based AI Medical Software System Service and Demo Project (No. 2019K0JTS0159), and the China Postdoctoral Science Foundation (No. 2020T130137ZX)
∗**E-mail:** teamedicine@163.com; junpeng.zhang@gmail.com

integrating their prediction can significantly improve the performance of neural networks[16-17]. However, ensemble learning was seldom applied to diagnose lung-related diseases.

In this study, ensemble models composed of different component models were built to diagnose COVID-19, and they were compared with single component models on the same test data. Transfer learning was applied to initialize the parameters of all component models before training. The proposed method was compared with some latest ones, and its competitive performance was proved on the same test data.

1 Materials and Methods

1.1 Dataset

Our database contains 8 347 chest CT images (Fig. 1) from more than 400 patients and 32 healthy subjects. These images come from three freely publicly available databases, COVIDx-CT, CC-CCII and COVID-CT, among which there are 2 849, 2 897 and 2 601 samples of COVID-19, common pneumonia (CP) and normal samples, respectively. Figure 1 shows three types of image examples, in which red marks indicate some obvious infection characteristics.

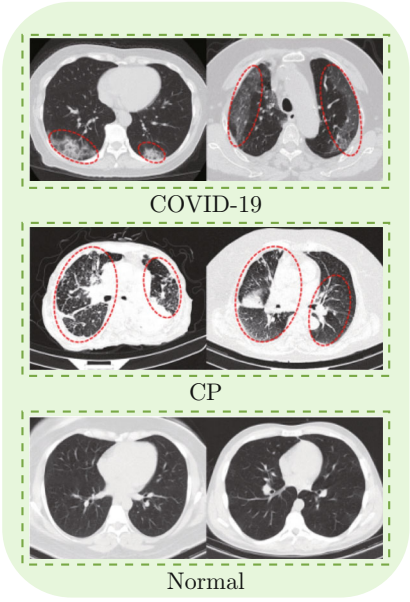


Fig. 1 CT images of COVID-19, CP and the normal

COVID-CT database contains 349 COVID-19 positive CT images and 397 negative ones[18]. All of them were collected from 216 patients from COVID-19-related papers from preprints (available via: https://github.com/UCSD-AI4H/COVID-CT). COVIDx-CT database is a large CT scan database that contains 104 009 CT slices from 1 489 patients[19]; this database is derived from CT database from CNCB[14], and includes three types of chest CT images: COVID-19, and CP and normal control. CT slices in CC-CCII database come from the China Consortium of Chest CT Image Investigation; this database contains a total of 617 775 CT slices from 6 752 CT scans of 4 154 patients. We manually selected 8 347 CT images from the three databases as data sets. The source and related information of these CT images are summarized in Table 1.

Table 1 Dataset information

Dataset source	COVID-19	CP	Normal	Patient information
COVIDx-CT	2 500	2 500	–	Yes
CC-CCII	–	–	2 601	Yes
COVID-CT	349	397	–	Partly

Note: "yes" means detailed subject information is provided, and "partly" means only part of that is provided.

1.2 Data Preprocessing and Augmentation

Images with excessive text or markings and too small lung lobe area (less than 30% of the intact lung lobe area) were removed. All samples were visually inspected to ensure those that failed to meet the requirements were removed. After screening, 2 485 COVID-19 samples, 2 548 CP samples and 2 475 normal samples were obtained to perform model training and testing.

Data augmentation strategies include image scaling, center cropping, random flip, and color jitter. Center cropping and image scaling can remove irrelevant content around the image and ensure that the image of the appropriate size is obtained. Application of random cropping and color jitter (pixel value changed within 10%) reduces the influence of both position and color change during training.

Independent test set was used to verify the actual performance of the trained model, in which approximately 20% of all samples were involved, and the rest were randomly divided into training set, fine-tuning set and validation set according to a ratio of 0.64 : 0.16 : 0.2 (training set: 3 246 CT samples; fine-tuning set: 811; validation set: 1 014; test set: 1 543).

1.3 Models and Methods

1.3.1 *Used Component Models*

In this study, three popular CNNs (VGG-19, ResNet-18, DenseNet-121) have been applied, and all codes were written using PyTorch (1.5.0) based on python (3.7.4). Three types of CNNs will be briefly explained, and the detailed structures can be found in Tables A1, A2 and A3 in Appendix.

(1) VGG-19 is a CNN with 19 convolutional layers[20], which uses successive small kernels for feature extraction, and this network structure has been proven to have good extraction capabilities for images, including medical images.

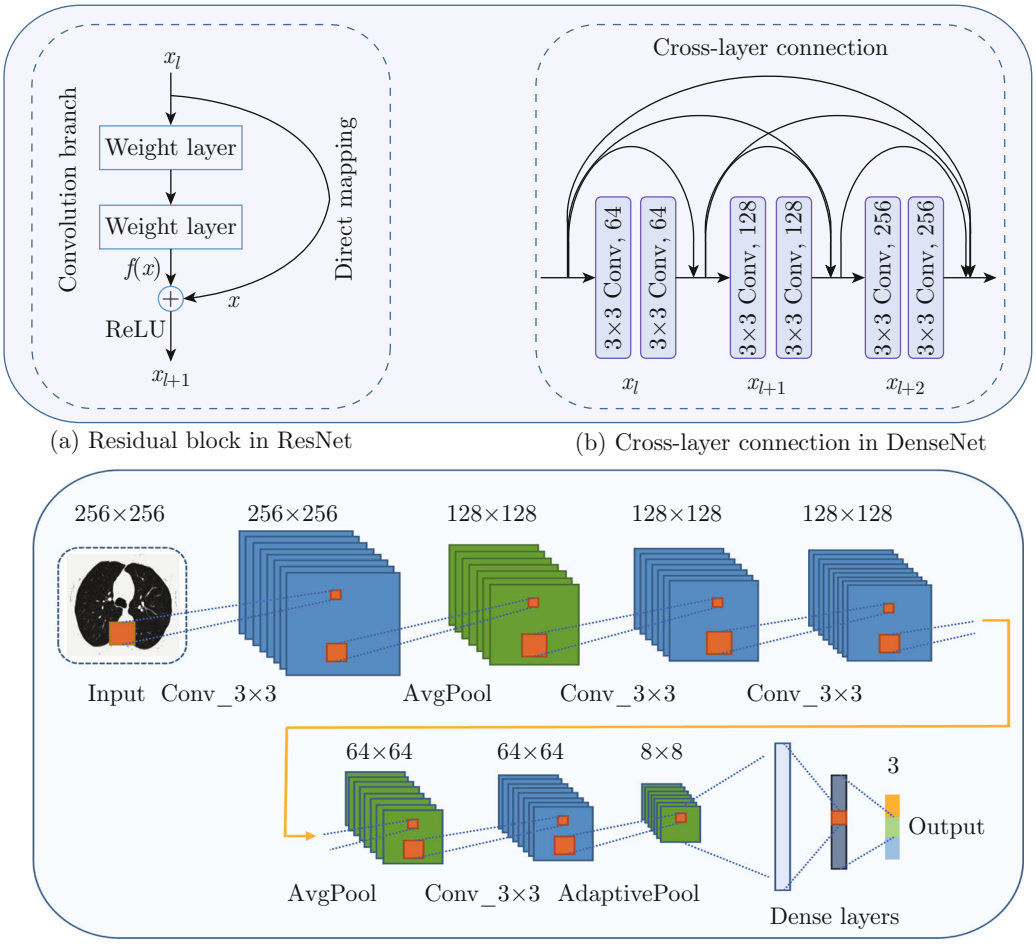
(2) Residual network is a neural network composed

of a residual structure[21]. This structure can reduce the gradient dispersion in the network learning process and thus reduce the loss of information. For complex medical images, this structure can build a deeper network and help deal with more complex feature patterns (Fig. 2(a)).

(3) DenseNet is a DNN constructed using a cross-layer connection method[22]. Its core idea is to establish the connection relationship between different layers (Fig. 2(b)). Owing to the limited medical image data, the cross-layer connection structure can significantly reduce the number of network parameters.

A residual block can be expressed as $x_{l+1} = x_l + f(x_l)$, and has been proved to be beneficial to the optimization of back-propagation neural network. In Fig. 2(a), the residual block is the basic structure of ResNet, composed of a direct mapping x that is the input of residual block and a residual part $f(x)$ that is the output of the convolution branch. In Fig. 2(b), cross-layer connection strengthens the transfer of features by creating leaping connections between layers and helps network training to a certain extent, while also reducing the size and calculation of the model.

In addition to the above component models, a simple CNN model with only four convolutional layers was built (Fig. 2(c)) as a component to build ensemble models (see Table A4 in Appendix for detailed network architecture). This simple model uses four convolutional layers for feature extraction, and features are transferred to the dense layer for classification.



(a) Residual block in ResNet

(b) Cross-layer connection in DenseNet

(c) Simple CNN structure

Fig. 2 Structures of different component models

1.3.2 *Single Component Model Training Strategy*

Component models were trained on training set. To prevent overfitting, a dropout layer was added before the classification layer for each component model. It should be noted that snapshot ensembles were applied to optimize the parameters of each component model[23], and the advantage of this method is that different models with different parameters could be obtained in one training process through this method. This method can not only reduce the training time, but also obtain good training performance.

Transfer learning was used to obtain the pre-trained

parameters, which are used to initialize the component model. For each component model, four sets of optimal model parameters were obtained using above mentioned snapshot ensemble strategy in one trial. During training, learning rates of convolution layers and classification layers were set to 1×10^{-5} and 1×10^{-4}, respectively and cross-entropy cost function was applied to evaluate model performance. The mini-batch size was set to 16 and Adam optimizer was used. The training of component models was repeated 10 times.

1.3.3 *Model Integration Strategy*

In each training, four different models were obtained. An important question is how we should combine four component models to form an ensemble model. Here, two simple strategies were proposed: score fusion and prediction voting.

Score Fusion The output scores of each component model are given a weighting coefficient $\beta\,(0 \leqslant \beta < 1)$, and the sum of the weighted scores of each category forms a final score of the ensemble model (Fig. 3(a)), i.e.,

$$S_i = \sum_k \beta_k S_{ki}, \tag{1}$$

where S_i is the output score of the ith category predicted by the ensemble model, S_{ki} is the output score of the ith category predicted by the kth component model, and β_k is the weight coefficient of the output score of the kth component model for each category (adaptively adjusted through back-propagation in the fine-tuning process).

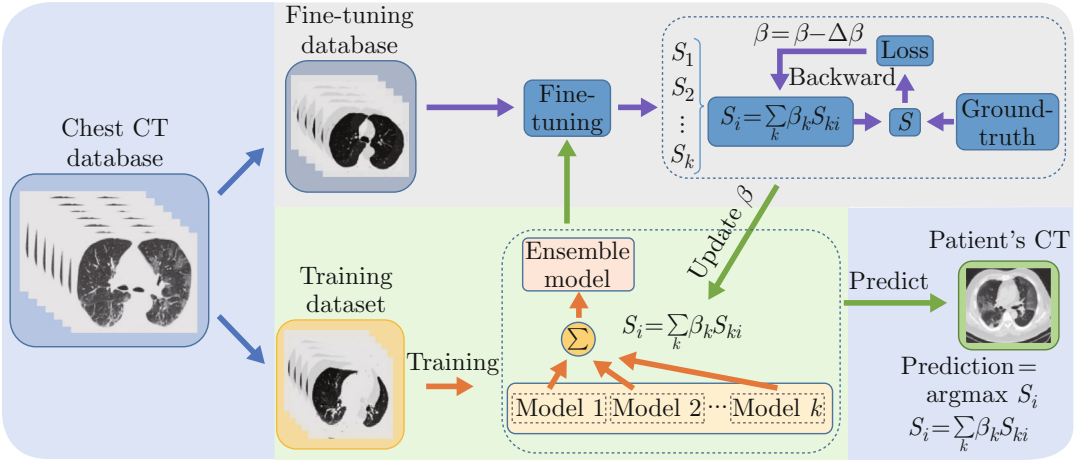
The training of ensemble model was divided into two stages. In the first stage, the data set is randomly divided into two parts (training data set and fine-tuning data set). After k (here $k = 4$) component models are obtained through snapshot ensembles on the training data set, an ensemble model with weight parameter β is constructed from these trained component models. In the second stage, the parameter β of the ensemble model is optimized on the fine-tuning data set through back-propagation. It is worth noting that the entire training process is done adaptively.

Prediction Voting Prediction voting is to select the one with the largest number of categories predicted by all component models according to the principle of majority winning (Fig. 3(b)), i.e.,

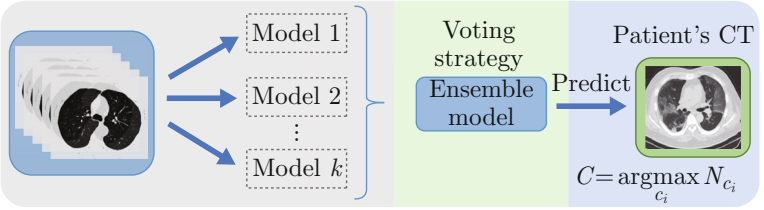
$$C = \arg\max_{c_i} N_{c_i}, \tag{2}$$

where C represents the category predicted by the ensemble model, c_i represents the ith predicted category, and N_{c_i} represents the number of component models whose prediction is the ith category.

We used four component models to form an ensemble model, and evaluated the ensemble model. The overall framework of the model is shown in Fig. 4. Overall experimental scheme design is illustrated in Fig. 5.



(a) Score fusion



(b) Prediction voting

Fig. 3 The proposed ensemble learning strategies

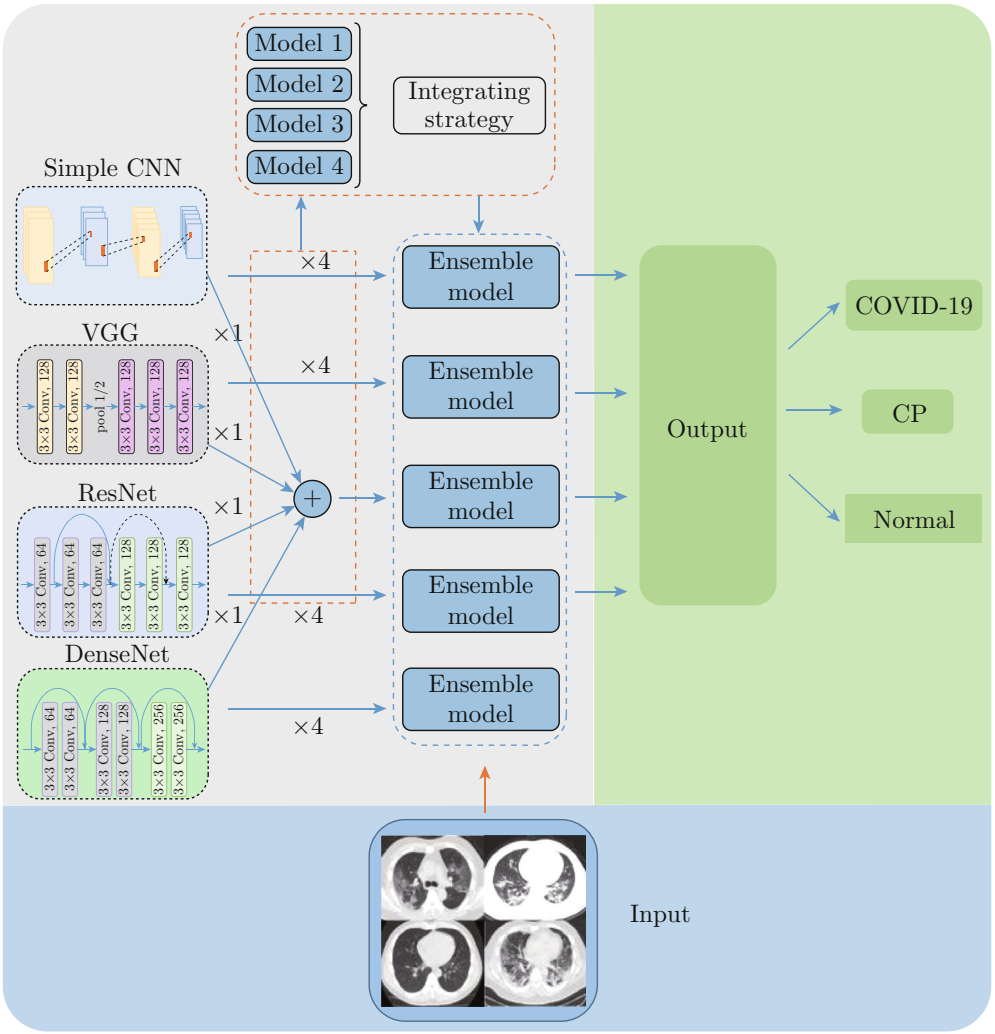


Fig. 4 Model framework diagram

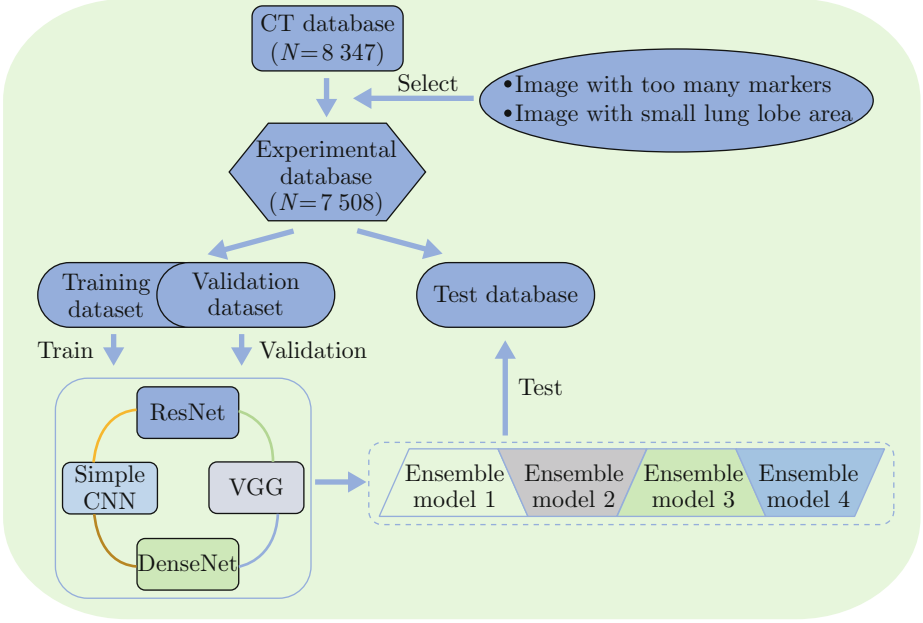


Fig. 5 Diagram of the overall experimental design

Different well trained component models compose different ensemble models according to integration strategies. The same or different component models are considered in each ensemble model (note: "×4" represents four specific component models are used to form an ensemble model). The categories of the input CT images are finally predicted by ensemble models.

In the experiment, data screening and preprocessing were first performed. All component models were trained and validated in training data set and validation data set, respectively. Well-trained models were saved to form different ensemble models according to different integration strategies, and the performance of these ensemble models was evaluated on the independent test set. In particular, the test performance of each component model is recorded at the same time for comparison. The workflow is shown in Fig. 5.

1.4 Evaluation Metrics

An independent test data set was used to evaluate the performance of both component and ensemble models. Accuracy, F1 score, recall rate and precision rate were separately calculated to show the test performance. Accuracy was used to measure the proportion of samples that are correctly classified. F1 score can be regarded as a harmonic average of the model's precision rate and recall rate that are used to determine the test performance of the model on positive and negative samples:

$$\mathrm{F1} = 2\frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}, \tag{3}$$

with

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \quad \mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{4}$$

where TP represents true positive and FN represents false negative, FP stands for false positive, and TN stands for true negative.

2 Results

2.1 Performance of Component Models

Figure 6 shows training curves of some component models. All models can converge quickly, benefiting from transfer learning. In addition, these curves demonstrate that the data set is sufficient and data augmentation is effective to avoid over-fitting. Table 2 lists evaluation metrics of component models on the test set. It can be seen that three popular CNNs have achieved an average accuracy of more than 83%. ResNet-18 obtained an average accuracy of 90.42% (highest) and F1 score of 0.894 3 (highest), followed by DenseNet-121 with an average accuracy of 88.38% and F1 score of 0.885 2. Besides, simple CNN obtained an average accuracy of 64.75%, which is significantly higher than a random probability of 33.3%.

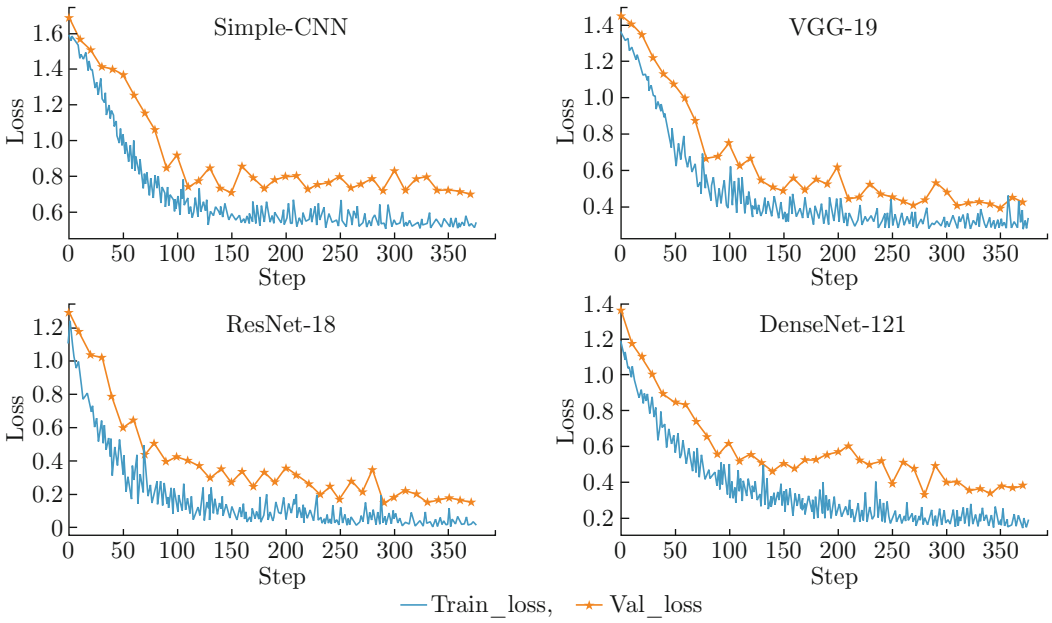


Fig. 6 Training and validation loss curves of some component models

It is easy to find that compared with simple CNN, other models with better performance have a deeper network structure and more reasonable design of feature processing. These advantages may have a large impact on the improvement of their performance.

2.2 Performance of Ensemble Models

Ensemble models were evaluated on the same test set. Table 3 shows different evaluation metrics. Confusion matrices on test set are shown in Figs. A1—A5 in Appendix. Among these ensemble models, the one

Table 2 Performance of component models on test set

Model	Accuracy	F1 score	Recall	Precision
Simple CNN	0.647 5 ± 0.023	0.635 4 ± 0.027	0.630 4 ± 0.021	0.624 5 ± 0.023
VGG-19	0.835 0 ± 0.014	0.826 5 ± 0.024	0.833 1 ± 0.025	0.789 5 ± 0.021
ResNet-18	**0.904 2 ± 0.015**	**0.894 3 ± 0.018**	**0.917 1 ± 0.021**	0.882 4 ± 0.016
DenseNet-121	0.883 8 ± 0.013	0.885 2 ± 0.018	0.876 7 ± 0.024	**0.903 7 ± 0.019**

Note: all data are reported as mean ± standard deviation, and the numbers in bold indicate the item with the highest score under each metric (similarly hereinafter).

Table 3 Performance of different ensemble methods on independent test set

Metrics		Simple-CNN × 4	VGG-19 × 4	ResNet-18 × 4	DenseNet-121 × 4	Hybrid model
Score fusion	Accuracy	0.654 8 ± 0.025	0.852 1 ± 0.020	**0.923 1 ± 0.018**	0.910 1 ± 0.024	0.854 9 ± 0.024
	F1 score	0.658 1 ± 0.021	0.861 5 ± 0.017	**0.936 8 ± 0.015**	0.907 1 ± 0.020	0.847 4 ± 0.019
	Recall	0.676 9 ± 0.019	0.862 8 ± 0.021	**0.935 6 ± 0.017**	0.904 5 ± 0.027	0.849 4 ± 0.027
	Precision	0.641 6 ± 0.023	0.861 4 ± 0.029	**0.938 2 ± 0.015**	0.910 7 ± 0.018	0.845 8 ± 0.023
Prediction voting	Accuracy	0.680 5 ± 0.049	0.881 3 ± 0.029	**0.956 2 ± 0.027**	0.931 5 ± 0.032	0.873 6 ± 0.029
	F1 score	0.727 6 ± 0.023	0.896 1 ± 0.018	**0.958 8 ± 0.021**	0.957 3 ± 0.026	0.851 4 ± 0.019
	Recall	0.712 1 ± 0.019	0.894 1 ± 0.024	**0.958 7 ± 0.019**	0.956 0 ± 0.019	0.837 4 ± 0.025
	Precision	0.746 0 ± 0.021	0.898 5 ± 0.027	0.955 9 ± 0.023	**0.958 4 ± 0.018**	0.859 7 ± 0.023

composed of four ResNet-18 models with voting strategy showed an average accuracy of 95.62% (highest) with an average F1 score of 0.958 8.

Figures 7 and 8 show the comparison between single component models and ensemble models of different integrated strategies. The average accuracy of ensemble models, especially with voting strategy, is significantly higher than that of single component models that compose them (Fig. 7). In Fig. 7, "fusion_ensemble" represents ensemble model with score fusion strategy, "voting_ensemble" represents ensemble model with voting strategy, and "hybrid_model" represents ensemble model composed of four different component models.

Figure 8 shows the distribution of the highest accuracy of single component models versus that of ensemble models in each trial (10 repeat trials in total for each ensemble or component model). It is easy to find that the accuracy of ensemble models is significantly higher than that of single component models in most trials. Accuracy of ensemble models based on ResNet-18 adopting voting strategy is increased by about 5% on average, and the highest up to about 7%, and similar phenomena can also be found in other ensemble models. In addition, it seems that the voting strategy can bring a more significant improvement in accuracy, compared with the score fusion strategy (Fig. 8).

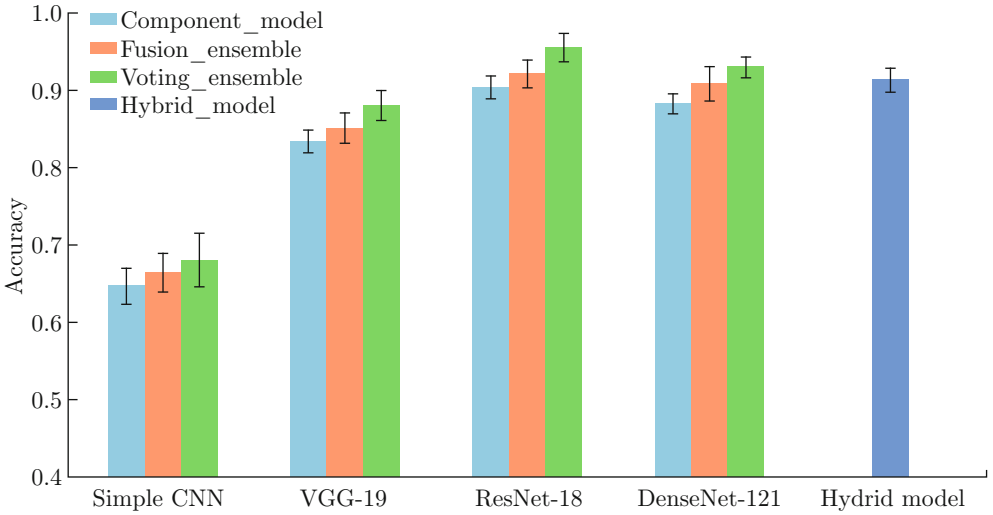


Fig. 7 Average accuracy of component models and ensemble models with different integration strategies on independent test set

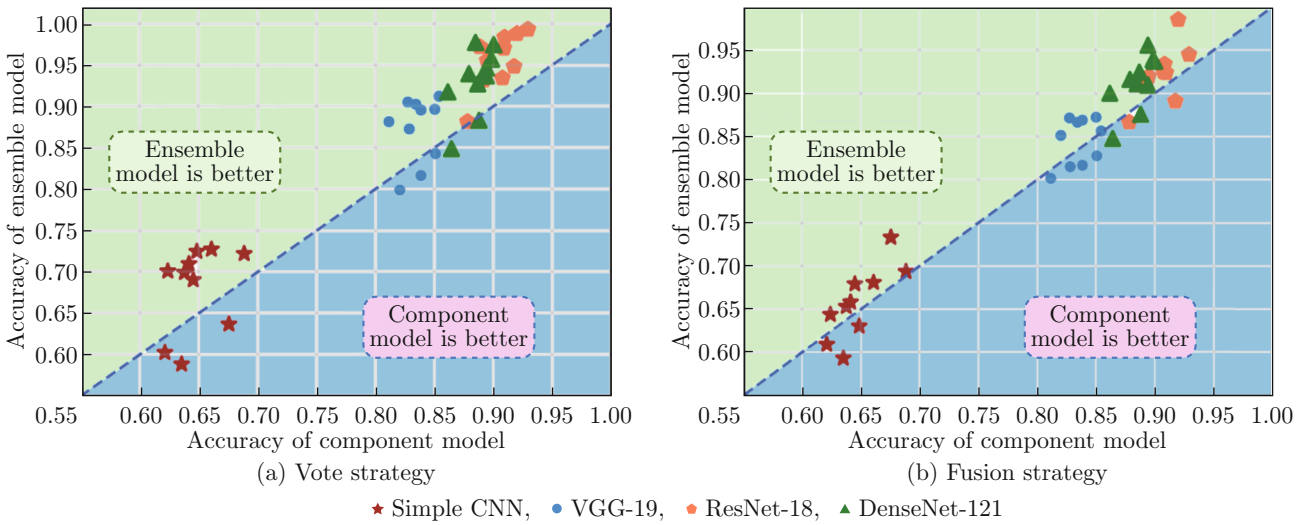


Fig. 8 Distribution diagram of the highest test accuracy of component models versus test accuracy of ensemble models in each trial

2.3 Comparison with Current Methods

We collected some latest COVID-19 diagnostic models based on deep learning and compared them with the proposed method. The performance of all models is obtained on the same test set. Table 4 shows the comparative results. The settings of the training and testing process were the same as before. It can be seen that the best accuracy of other methods on our test set is 93.61%, which is slightly lower than our average accuracy of 95.62% and significantly lower than our best accuracy of 99.37%. Compared with these methods, the proposed method has an accuracy improvement up to about 15%, showing the significant advantage of the proposed ensemble method.

Table 4 Comparison of current methods

Method	Accuracy
DRE-Net[9]	0.847 4
COVIDNet-CT[19]	0.904 8
Proposed ensemble model with vote strategy	**0.956 2**

3 Conclusion

CAD can greatly reduce the workload of professional physicians and bring new hope to the automated diagnose of COVID-19. Deep learning can distinguish between COVID-19 CT images and others. However, most studies only focus on the improvement of the ability of a single model, and ignore the effect of ensemble intelligence. To improve the accuracy of neural network recognition of COVID-19 CT samples, we thus proposed ensemble learning methods combining multiple component models in order to obtain a more powerful classifier.

In our research, transfer learning and snapshot ensembles were adopted to train component models. In the performance, ensemble models were compared with both their component models and some current popular methods, and their accuracy has been significantly improved (up to 10.88%). In addition, two different ensemble learning methods were compared, showing that different ensemble learning methods also have a significant impact on the performance of ensemble models. Our research is of reference significance for AI-based assisted diagnosis systems.

References

[1] AI T, YANG Z L, HOU H Y, et al. Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in China: A report of 1014 cases [J]. *Radiology*, 2020, **296**(2): E32-E40.

[2] ZHANG N R, WANG L L, DENG X Q, et al. Recent advances in the detection of respiratory virus infection in humans [J]. *Journal of Medical Virology*, 2020, **92**(4): 408-417.

[3] HUANG C L, WANG Y M, LI X W, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China [J]. *The Lancet*, 2020, **395**(10223): 497-506.

[4] CHUNG M, BERNHEIM A, MEI X Y, et al. CT imaging features of 2019 novel coronavirus (2019-nCoV) [J]. *Radiology*, 2020, **295**(1): 202-207.

[5] ISMAEL A M, ŞENGÜR A. Deep learning approaches for COVID-19 detection based on chest X-ray images [J]. *Expert Systems With Applications*, 2021, **164**: 114054.

[6] OH Y, PARK S, YE J C. Deep learning COVID-19 features on CXR using limited training data sets [J]. *IEEE Transactions on Medical Imaging*, 2020, **39**(8): 2688-2700.

[7] LI L, QIN L, XU Z, et al. Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: Evaluation of the diagnostic accuracy [J]. *Radiology*, 2020, **296**(2): E65-E71.

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

[9] SONG Y, ZHENG S J, LI L, et al. Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images [J]. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5361, **PP**(99): 1.

[10] BAI H X, WANG R, XIONG Z, et al. Artificial intelligence augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other origin at chest CT [J]. *Radiology*, 2021, **299**(1): E225.

[11] SHI W Q, TONG L, ZHU Y D, et al. COVID-19 automatic diagnosis with radiographic imaging: Explainable attention transfer deep neural networks [J]. *IEEE Journal of Biomedical and Health Informatics*, 2021, **25**(7): 2376-2387.

[12] LI J P, ZHAO G M, TAO Y L, et al. Multi-task contrastive learning for automatic CT and X-ray diagnosis of COVID-19 [J]. *Pattern Recognition*, 2021, **114**: 107848.

[13] QIAN X L, FU H Z, SHI W Y, et al. M3 Lung-Sys: A deep learning system for multi-class lung pneumonia screening from CT imaging [J]. *IEEE Journal of Biomedical and Health Informatics*, 2020, **24**(12): 3539-3550.

[14] ZHANG K, LIU X, SHEN J, et al. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography [J]. *Cell*, 2020, **181**(6): 1423-1433.

[15] POLIKAR R. Ensemble based systems in decision making [J]. *IEEE Circuits and Systems Magazine*, 2006, **6**(3): 21-45.

[16] FOLINO F, FOLINO G, GUARASCIO M, et al. On learning effective ensembles of deep neural networks for intrusion detection [J]. *Information Fusion*, 2021, **72**: 48-69.

[17] HANSEN L K, SALAMON P. Neural network ensembles [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1990, **12**(10): 993-1001.

[18] GUNRAJ H, WANG L, WONG A. COVIDNeT-Ct: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest CT images [J]. *Frontiers in Medicine*, 2020, **7**: 608525.

[19] ZHAO J Y, HE X H, YANG X Y, et al. COVID-CT-dataset: A CT scan dataset about COVID-19 [EB/OL]. [2021-01-07]. https://arxiv.org/abs/2003.13865.

[20] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2021-01-07]. https://arxiv.org/abs/1409.1556.

[21] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]//*2016 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). Las Vegas, NV: IEEE, 2016: 770-778.

[22] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks [C]//*2017 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). Honolulu, HI: IEEE, 2017: 2261-2269.

[23] HUANG G, LI Y X, PLEISS G, et al. Snapshot Ensembles: Train 1, get M for free [EB/OL]. [2021-01-07]. https://arxiv.org/abs/1704.00109.

Appendix

Table A1 VGG-19 network architecture

Layer	Network configuration	Output size
Input	–	224 × 224
Features	Conv3-64	
	Conv3-64	
	Maxpool2-64	112 × 112
	Conv3-128	
	Conv3-128	
	Maxpool2-128	56 × 56
	Conv3-256	
	Conv3-256	
	Conv3-256	
	Conv3-256	
	Maxpool2-256	28 × 28
	Conv3-512	
	Conv3-512	
	Conv3-512	
	Conv3-512	
	Maxpool2-512	14 × 14
	Conv3-512	
	Conv3-512	
	Conv3-512	
	Conv3-512	
	Dropout	
	Maxpool2-512	7 × 7
Classifier	Flatten	25 088
	Dense-1 000	
	Dropout	
	Dense-2	2
	Softmax	2

Note: all dropout rates are set at 0.5; the convolution layer parameter is denoted as Conv(receptive field size)-(number of output channels), and the dense layer parameter are denoted as Dense-(output dim).

Table A2 ResNet-18 network architectures

Layer	Configuration	Output size
Input	–	224 × 224
Conv1	Conv7-64	
	Maxpool3-64	112 × 112
Conv2_x	[Conv3-64; Conv3-64] × 2	56 × 56
Conv3_x	[Conv3-128; Conv3-128] × 2	28 × 28
Conv4_x	[Conv3-256; Conv3-256] × 2	14 × 14
Conv5_x	[Conv3-512; Conv3-512] × 2	7 × 7
Pool	Dropout	
	Averagepool	1 × 1
Classifier	Flatten	
	Dense-512	
	Dense-2	
	Softmax	2

Table A3 DenseNet-121 network architectures

Layer	Configuration	Output size
Input	–	224 × 224
Conv1	Conv7-64	112 × 112
	Maxpool3-64	56 × 56
Dense block 1	[Conv1-128; Conv3-32] × 6	56 × 56
Transition layer 1	Conv1-128	
	Averagepool	28×28
Dense block 2	[Conv1-128; Conv3-32] × 12	28 × 28
Transition layer 2	Conv1-256	
	Averagepool	14 × 14
Dense block 3	[Conv1-128; Conv3-32] × 24	14 × 14
Transition layer 3	Conv1-512	
	Averagepool	7 × 7
Dense block 4	[Conv1-128; Conv3-32] × 24	7 × 7
Classifier	Dropout	
	Averagepool	1 × 1
	Dense-256	
	Dense-2	
	Softmax	2

Table A4 Simple CNN network architecture

Layer	Network configuration	Output size
Input	–	256 × 256
Features	Conv_ReLU 3-32	
	BatchNorm	256 × 256
	Maxpool 2-32	128 × 128
	Conv_ReLU 3-64	
	Conv_ReLU 3-64	
	BatchNorm	64 × 64
	AdaptiveAvgpool	8 × 8
Classifier	Flatten	4 096
	Dropout	
	Dense-256	256
	Dense-3	
	Softmax	3

Note: all dropout rates are set at 0.5; the convolution-ReLU layers and pool layers are denoted as Conv_ReLU (receptive field size)-(number of output channels), and the dense layer parameter is denoted as Dense-(output dim).

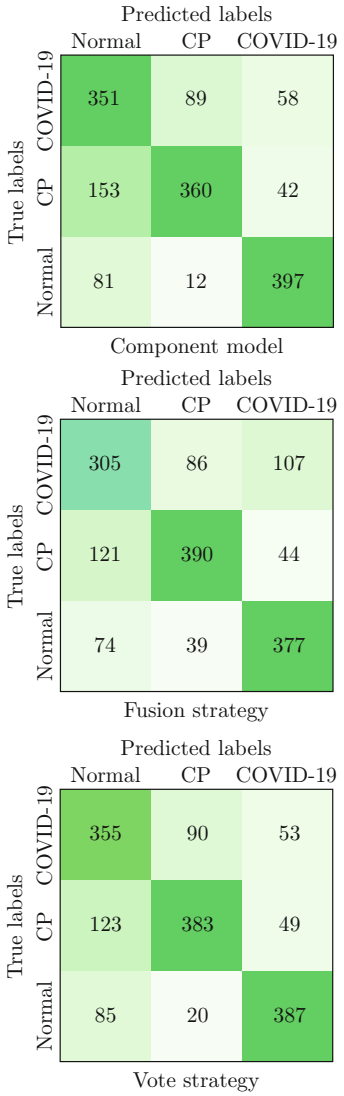


Fig. A1 Confusion matrices of simple CNN on independent test set

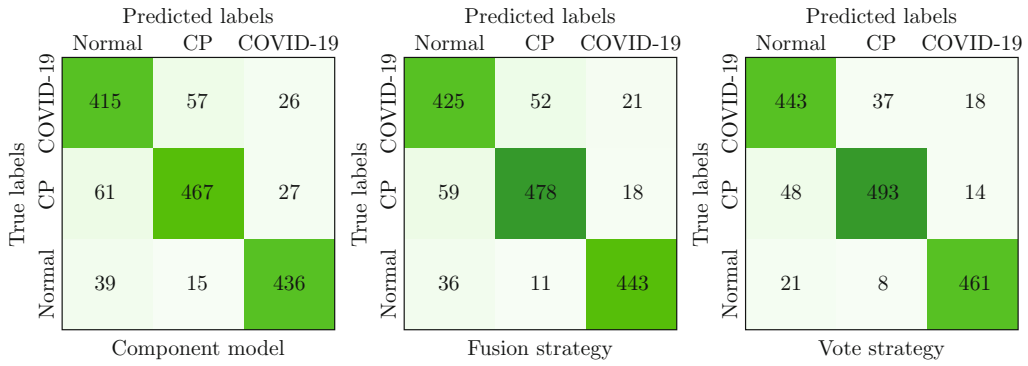


Fig. A2 Confusion matrices of VGG-19 on independent test set

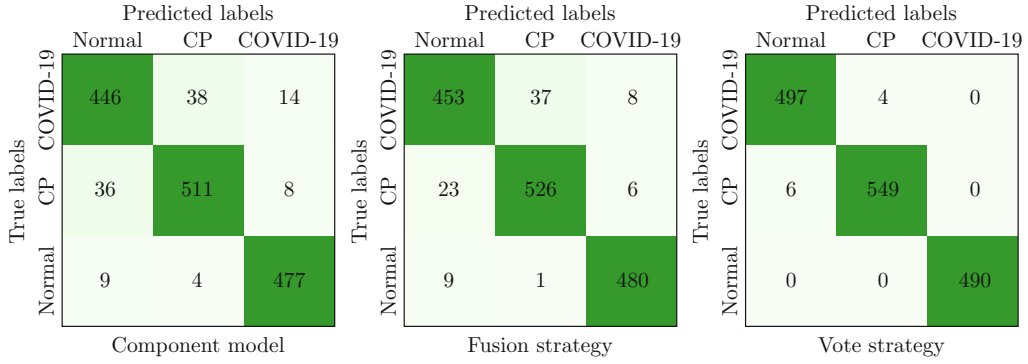


Fig. A3 Confusion matrices of ResNet-18 on independent test set

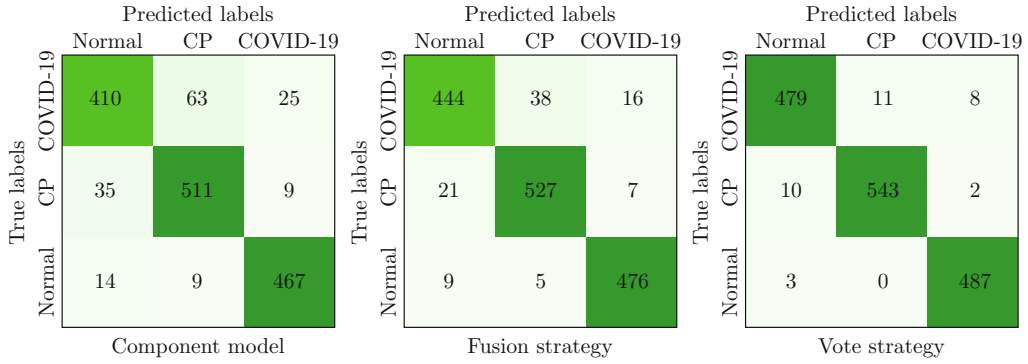


Fig. A4 Confusion matrices of DenseNet-121 on independent test set

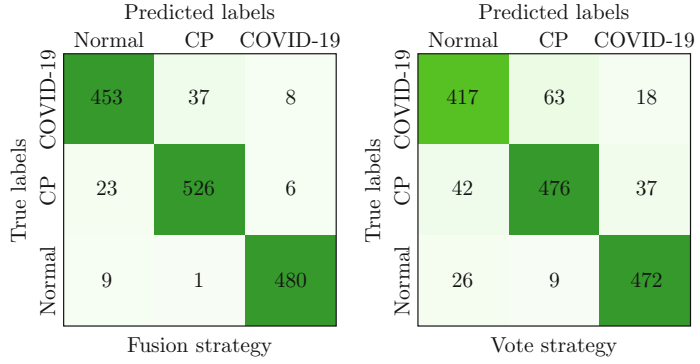


Fig. A5 Confusion matrices of hybrid model of simple CNN, VGG-19, ResNet-18 and DenseNet-121 on independent test set