Large-scale screening to distinguish between COVID-19 and community-acquired pneumonia using infection size-aware classification

Feng Shi1,2, Liming Xia1,2, Fei Shan2,3, Bin Song4,5, Dijia Wu2, Ying Wei4, Huan Yuan1, Huiting Jiang1, Yichu He5, Yaozong Gao1, He Sui1 and Dinggang Shen1,6,7

1 Department of Research and Development, Shanghai United Imaging Intelligence Co., Ltd, Shanghai, People’s Republic of China
2 Department of Radiology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, People’s Republic of China
3 Department of Radiology, Shanghai Public Health Clinical Center, Fudan University, Shanghai, People’s Republic of China
4 Department of Radiology, Sichuan University West China Hospital, Chengdu, Sichuan Province, People’s Republic of China
5 Department of Radiology, China-Japan Union Hospital of Jilin University, Changchun, People’s Republic of China
6 School of Biomedical Engineering, ShanghaiTech University, Shanghai, People’s Republic of China
7 Department of Artificial Intelligence, Korea University, Seoul, Republic of Korea

E-mail: dinggang.shen@gmail.com

Keywords: COVID-19, pneumonia, decision tree, size-aware, random forest

Supplementary material for this article is available online

Abstract

The worldwide spread of coronavirus disease (COVID-19) has become a threat to global public health. It is of great importance to rapidly and accurately screen and distinguish patients with COVID-19 from those with community-acquired pneumonia (CAP). In this study, a total of 1,658 patients with COVID-19 and 1,027 CAP patients underwent thin-section CT and were enrolled. All images were preprocessed to obtain the segmentations of infections and lung fields. A set of handcrafted location-specific features was proposed to best capture the COVID-19 distribution pattern, in comparison to the conventional CT severity score (CT-SS) and radiomics features. An infection size-aware random forest method (iSARF) was proposed for discriminating COVID-19 from CAP. Experimental results show that the proposed method yielded its best performance when using the handcrafted features, with a sensitivity of 90.7%, a specificity of 87.2%, and an accuracy of 89.4% over state-of-the-art classifiers. Additional tests on 734 subjects, with thick slice images, demonstrates great generalizability. It is anticipated that our proposed framework could assist clinical decision making.

Introduction

Also known as COVID-19, coronavirus disease outbreaks in Wuhan, China, since December 2019 have rapidly spread nationwide and globally (Wang et al 2020a, Wu et al 2020). A form of pneumonia, the infection causes inflammation of air sacs in one or both lungs which then fill with fluid or pus, making it difficult for the patient to breathe. According to a recent report, its mortality rate reaches 3.6% (Shi et al 2020), second only to that of severe acute respiratory syndrome (SARS) (10%) and that of Middle East respiratory syndrome (MERS) (36%) (de Wit et al 2016, Sahin et al 2020). It can also transmit from person to person with a relatively high basic reproduction number ($R_0$) of 2.2 $\sim$ 3.28 (Li et al 2020, Liu et al 2020), and has had no efficient treatments and control strategies till now (Wang et al 2020b). Therefore, the screening of COVID-19 out of community-acquired pneumonia (CAP) is important for patient triage, treatment protocol design, and follow-up evaluation.

Currently, a patient showing respiratory symptoms, fever, cough, dyspnea, or pneumonia, would be referred for a test of real-time polymerase chain reaction (RT-PCR) for the final diagnosis of COVID-19 (Chan et al 2020). However, recent studies show that RT-PCR has a relatively low detection rate of around 30%-60%
(Fang et al 2020, Li and Xia 2020), and repeated tests are generally needed. Therefore, negative results of RT-PCR does not rule out the possibility of infection. A chest CT scan is a routine diagnostic tool for pneumonia, and is found to be very useful in detecting typical radiographic features of COVID-19, especially with thin slices (Wong et al 2003, Li and Xia 2020, Shen et al 2020). Basically, these features include bilateral and peripheral ground-glass and also consolidative pulmonary opacities. According to the time course of disease development, patients could be divided into mild, moderate, severe and critically ill stages (Zu et al 2020). In the mild stage, almost no pneumonia could be seen from the CT images. From the moderate to critically ill stages, one could observe ground-glass opacity (GGO), an increased crazy-paving pattern, and consolidation (Li and Xia 2020). If the symptoms of patient improves, gradual resolution of consolidation could be seen in the CT images. However, it remains a great challenge to screen COVID-19 from other pneumonia using CT images. First, their radiographic appearances may be similar, making it difficult to differentiate between them. Second, CT images, especially for thin slice acquisition, contain hundreds of slices, which is time-consuming for manual checking and small infections in the mild stage could be easily missed. Accordingly, an automated image analysis method to help screen COVID-19 from CAP is desired.

Although there is research summarizing typical CT radiographic signs in COVID-19 patients for guiding clinical practice, studies of automated machine learning assisted disease screening is still limited. For example, Wang et al proposed a deep learning method to classify the patches of infected lesions into COVID-19 or typical viral pneumonia of 99 subjects, using manually labeled infections (Wang et al 2021). Xu et al studied the early screening of COVID-19 from influenza-A viral pneumonia and healthy cases with a total of 618 CT samples (Xu et al 2020), where lesion patches were extracted and combined with relative distance-from-edge features for diagnosis. Wang et al proposed a deep learning model with convolutional layers to extract features from lung segmentations, and provide diagnosis results in a dataset of 313 COVID-19 and 229 non-COVID-19 patients (Wang et al 2020c).

Another difficulty in the screening of COVID-19 patients is the different distributions of infections in the data. Generally, patients with COVID-19 tend to have more widespread infections than CAP patients due to the emergent outbreak of COVID-19 with no effective treatments available. This leads to a tendency for machine learning classifiers to assert that small infections found in CT images most likely belong to CAP, while widespread infections are COVID-19. This situation means that the effect of the size of infection should be considered and properly evaluated in the classification experiments.

In this paper, we propose a machine learning approach to distinguish COVID-19 from CAP. Our contributions are three-fold: (1) we propose a set of handcrafted location-specific features dedicated to diagnosis according to the current understanding of COVID-19 clinicoradiological appearance; (2) we propose a size-aware strategy for the classification task and evaluate in multiple infection size ranges; (3) the proposed method is evaluated throughout on a large-scale dataset, as well as an independent dataset, with thick slice acquisition to test its generalizability. Details are provided in the following sections.

Materials and methods

Participants and image acquisition
CT images of 2,685 participants were retrospectively collected. In this dataset, 1,658 cases were confirmed COVID-19 cases diagnosed by positive nucleic acid testing with conformation by the national CDC. The other 1,027 cases were CAP patients. All patients underwent chest CT scans with thin sections. Reconstructed CT thicknesses ranged from 0.625 to 2 mm. A five-fold cross-validation strategy was employed to evaluate the classification performance, where 20% data were considered as testing and the rest was used as training in each fold. It is worth noting that to increase the stability of the results, the five-fold cross-validation was repeated ten times with random splits of training and testing datasets, and the mean result from all runs was finally reported. Besides the above main dataset, to further test the generalizability of the proposed method on other imaging settings, an independent testing dataset was also acquired with thick-slice CT images (slice thickness = 5 mm) on 400 COVID-19 patients and 334 CAP patients.

Specifically, for the above main and independent datasets, the related CT scanners include uCT 780 from UIH, Optima CT520, Discovery CT750, LightSpeed 16 from GE, Aquilion ONE from Toshiba, SOMATOM Force from Siemens, and SCENARIA from Hitachi. CT protocol includes: 120 kV, with breath hold at full inspiration. Images were provided from collaborating hospitals, including Tongji Hospital of Huazhong University of Science and Technology, Shanghai Public Health Clinical Center of Fudan University, China-Japan Union Hospital of Jilin University, and Sichuan University West China Hospital. Images of COVID-19 patients were acquired from 9 January 2020 to 14 February 2020, and images of CAP patients were obtained from 30 July 2018 to 22 February 2020. All images were de-identified before sending for analysis. The study was approved by the Institutional Review Board of participating institutes. Written informed consent was waived.
due to the retrospective nature of the study. A mediastinal window (with window width 350 HU and window level 40 HU) and/or lung window (with window width 1,200 HU and window level -600 HU) were used for reading.

**Preprocess of lung infections and fields**

We propose to utilize the clinically observed disease characteristics, i.e., infection locations and spreading patterns, to extract handcrafted features. To do that, we automatically segmented infected lung regions and lung fields bilaterally. The infected lung regions were mainly related to manifestations of pneumonia, such as mosaic sign, GGO, lesion-related signs (air bronchogram), and interlobular septal thickening. The resulting lung fields include left and right lungs, five lung lobes, and eighteen pulmonary segments.

The segmentation process was done through a pretrained in-house research portal software (Shan et al 2020). Specifically, a deep learning-based network called VB-Net was employed for image segmentation. VB-Net is a modified network that combines V-Net with bottleneck layers to reduce and combine feature map channels (Milletari et al 2016). The network includes a contracting path to extract global image features and an expansive path to integrate fine-grained image features. A bottleneck structure is integrated in the network to reduce the number of feature map channels and thus speed up the spatial convolution. Given proper annotations as training data, V-Net is capable of segmenting infected lesions as well as lung fields. This software has been evaluated in the segmentation of infections and lung fields and achieved a Dice similarity coefficient of 92% between automated and manual infection segmentations. In this work, all images were first resampled into 1.5 mm isotropic resolution. We then directly used this pretrained software for the segmentation of infected lesions and lung fields.

An example of images and preprocessed results is shown in figure 1. A total of three CAP and three COVID-19 patients are included in six columns, respectively. Segmentations of infected lesions and lung fields are also shown.

**Infection size distribution and proposed size-aware method**

We refer to the infection size as the volume of infected regions divided by the volume of the whole segmented lung. Figure 2(A) shows the distribution of infection size in patients. As the relationship between size and disease progress is not linear, we also tried to separate patients into four groups, based on exponential intervals. The results show that figure 2(B) can better separate the patients. We can observe that COVID-19 patients tend to have relatively larger infections than that of CAP patients. For example, they account for 96% (538 versus 20) in the size group of >10%. In contrast, CAP patients occupy 91% (309 versus 30) in the size group of <0.01%.

As mentioned above, such a polarized patient distribution is not ideal for conventional classification models. In these models, size would be chosen as a major feature as it could easily separate both groups, although this

---

*Figure 1.* Illustration of lung images, preprocessed results of infected lesions and lung fields on three CAP (left three columns) and three COVID-19 patients (right three columns), respectively.
leads to a low performance for the classification of middle-size groups and does not reflect the real radiographic appearance differences between the two types of pneumonia. In this work, we propose a classification strategy and its flow chart is shown in figure 3. In particular, for cross-validation, all data were first randomly partitioned into five equally-sized subsamples. In each fold, four subsamples were used for training data to construct the classification model, and the remaining one subsample was used as testing data. This process was repeated until all subsamples were tested. Figure 3(A) shows the schematic diagram, where in each fold all images were preprocessed, and machine learning models were trained including feature extraction, selection, and disease classification. In the testing stage, an unseen testing image would undergo the preprocessing steps, and then the model predicted its probability of being COVID-19 or CAP.

Herein, we propose an infection size-aware random forest (iSARF) method to deal with the issue of infection distribution imbalance (figure 3(B)). Out of all the features, the infection size was used as the only feature in a decision tree to separate the data into four size groups. Then a group of random forests were constructed for each size group in the training process, where in each group the classifier could focus on the characteristics of this specific data subset and its best discriminative features could be selected to optimize the classification performance. This strategy would constrain the classifiers to specifically deal with the cases with similar size infections in each group and thus improve the classification accuracy. In the testing stage, the testing data would be sent to a proper size group through the decision tree and then classified by the following (respective) random forests for the final diagnosis. Details of feature extraction and classification are provided in the following subsections.

**Extraction of location-specific features**

In this study, we propose a series of dedicated handcrafted features to be automatically extracted in CT images from infections and lung fields. These features are composed of four categories, including the volume, infected lesion number, histogram distribution, and surface area. The detailed feature distribution framework is shown in figure 4.
**Volume features**
We extracted the total volume of the infected region, and calculated the percentage of the infected region of the whole lung. By using the lung field masks, we further extract the volume and percentage, respectively, in each lobe and pulmonary segment. As there is evidence that COVID-19 is more likely occur in right lungs (Shi et al 2020), we calculate the infected lesion difference as well as the percentage difference between left and right lungs.

**Infected lesion number**
Another image difference between COVID-19 and CAP is that most COVID-19 infections encompassed bilateral lungs with multifocal involvement (Chung et al 2020, Li and Xia 2020), and COVID-19 generally also has concentrated infected lesions while CAP shows up as being small in volume and patchy in distribution (Hansell et al 2008). Therefore, we calculate the characteristics of the total number of infected regions in the bilateral lungs, lung lobes, and pulmonary segments, respectively.

**Histogram distribution**
The manifestations of COVID-19 have their own characteristics, which are different from other types of pneumonia, such as influenza-A viral pneumonia (Xu et al 2020). The predominant chest CT findings show that bilateral and peripheral GGO and consolidation are radiologic hallmarks of COVID-19 (Bernheim et al 2020, Li et al 2020, Wang et al 2020b). GGO is a pattern of hazy increased lung opacity with preservation of bronchial and vascular margins, whereas consolidation is characterized by a homogeneous increase in lung parenchymal attenuation that obscures the margins of vessels and airway walls on the CT image (Hansell et al 2008). In order to extract the CT intensity distribution of the infected region, we calculated the histogram features of the infected region. According to the window width of the lung window of 1,500 and the window level of $-600$, the intensity value range of the lung region image obtained by inference tends to be between $-1,350$ and 150. We thus divided this interval into 30 equal bins, and counted the frequency of intensity level in the infected region at each bin to obtain the frequency distribution histogram features. Meanwhile, the mean and standard deviation of HU values were calculated, representing the density of the infected region.

**Surface area**
Compared with CAP, it has been found that COVID-19 has a predominate distribution in the posterior and peripheral lung (Song et al 2020), and the abnormalities of lung parenchyma eventually spread to the central area and bilateral upper lobes (Li and Xia 2020). Therefore, we constructed the infection surface as well as the lung boundary surface. We further calculated the distance of each infection surface vertex to the nearest lung boundary surface, and categorized them into five ranges, as 3, 6, 9, 12, and 15 voxels (voxel spacing is 1.5 mm). For features, the number of infection surface vertices within each range of distances to the lung wall were calculated. Furthermore, the percentage of infection surface vertex number against the number of whole infection surface vertices in each range were also obtained.
Feature selection and prediction

After generating the 96 handcrafted features, we applied machine-learning methods to select a proper subset of features, and built classifiers to predict COVID-19 patients and CAP patients. The feature selection and prediction process of each fold are detailed as follows.

Feature selection is one of the most important steps that could improve the performance of predictive modeling. The benefits of feature selection include that the model would be easier to interpret after removing redundant variables, the algorithm would work faster since the feature number is reduced, and the potential overfitting problem would be largely alleviated. The least absolute shrinkage and selection operator (LASSO) method is widely used as it can perform both variable selection and regularization. Briefly, it uses a shrinkage parameter where data are shrunk to a certain central point and conducts variable selection by forcing the coefficients of ‘not-so-significant’ variables to become zero through a penalty (Kim et al 2007). LASSO has proven to be a versatile tool for relevant feature selection and has been widely used in many research areas such as biomedical imaging, genomes studies, and finance (Wang et al 2010, Colombani et al 2013, Vasquez et al 2016). In this work, we employ LASSO to explore the optimal subset of radiological features for the classification in the training data, and the selected results are directly used in the testing data. We use the LASSO implementation from the scikit-learn package (sklearn.linear_model.Lasso) with default parameters in training data at each fold. After LASSO, the selected features were then fed into the proposed classifier to construct the disease diagnosis model. In detail, random forests were implemented by using LightGBM, an open-source implementation of a gradient boosting framework that uses a sequence of trees to solve classification tasks. This classifier is newly recognized as an efficient method with high performance, in comparison to conventional ones. It has three major merits. First, the best split gain node could be found in LightGBM by using a traditional leaf-wise strategy with limited depth of the trees. This ensures high efficiency and prevents the occurrence of the overfitting problem in the case of very deep tree structure. Second, the optimal hyperparameters could be automatically tuned for the best performance in the training dataset by using Bayesian sequential model-based optimization, instead of predefining hyperparameters as in conventional classifiers. Third, tasks could be performed in parallel, which significantly improves the computational speed. For lightGBM, a total of 100 trees are used for each size group, the maximum depth of trees is seven, and the Bayes function is used to optimize parameters. In the testing stage, these trained models are then applied to a new test image to predict its probability of being COVID-19 rather than CAP.

Comparison of features

For comparison, we first involve two widely-used methods in the field: (1) the CT severity score (CT-SS). CT-SS is widely used by radiologists to evaluate lung abnormalities. Following previous works (Ooi et al 2004, Yang et al 2020), we assign each of 18 pulmonary segments to a score based on the following: score 0, 0% involvement; score 1, less than 25% involvement; score 2, 25% to less than 50% involvement; score 3, 50% to less than 75% involvement; and score 4, 75% or greater involvement. The summation of scores was considered as an additional feature that provides overall lung involvement (ranging from 0 to 72); (2) radiomics features. The radiomics features were extracted directly from infected lesions, including 19 first-order intensity statistics (e.g., average gray level intensity, range of gray values), and 74 texture features (e.g., gray level co-occurrence matrix, gray-level run-length matrix, gray-level size-zone matrix, and neighborhood gray-tone difference matrix). Similarly, the LASSO method was also used to select the most useful features before classification.

Comparison of classifiers

We employ support vector machine (SVM), logistic regression (LR), neural network (NN), and random forest (RF) to evaluate the performance of our proposed method. For SVM, a radial basis function (RBF) kernel is used, together with a kernel coefficient of 0.05. For the LR model, an l2-norm penalty with penalty weight 1 is applied. For NN, the classification network is set to have one hidden layer with 100 neurons, and is trained using an Adam optimizer for 500 iterations with a learning rate of 0.005 and a batch size of 200. Here the RF method is an ablation study of the proposed method without the infection size decision tree modification. It was also implemented using LightGBM with the same parameter settings, such as the number of trees as 100, maximum depth of seven and the use of Bayes optimization. Other parameters were tuned due to validation scores in training data.

To evaluate the performance, models were performed in the testing data at each fold. Receiver operating characteristic (ROC) curve analysis was performed. Sensitivity evaluates the ratio of correctly identified positive cases versus all positive cases. Specificity measures the ratio of negatives correctly found in all true negatives. The area under the ROC curve (AUC) demonstrates the ability of the classifier considering both sensitivity and specificity. Differences between various ROC curves were evaluated using the Delong test (Sun and Xu 2014) and the statistical threshold was set at $p < 0.05$. Statistical analysis was conducted with R software (version 4.0.2; https://www.r-project.org). We further investigated the metrics of results with respect to the infected lesion
size. Meanwhile, to promote the related research, these extracted features will be made available to the scientific community at https://github.com/simonsf/covid-feature-data.

Results

Demographic characteristics of participants

The main dataset includes thin-section CT images (slice thickness < 2 mm) of 1,658 patients with COVID-19 and 1,027 patients with CAP. The age and gender distribution of all subjects are shown in figure 5. The age of the COVID-19 subjects is 49 ± 14 years, which is significantly younger than that of CAP subjects, which is 56 ± 14 years (p < 0.001 with two-sample t-test). There are 856 males and 802 females in the COVID-19 group, which shows a slightly higher male to female ratio than the CAP group, with 488 males and 539 females (p = 0.038 through the Chi-square test).

Evaluation of handcrafted features

The results are shown in figure 6. In particular, for the proposed method, the accuracy is relatively low at 63.8% when using CT-SS, and largely improved to 86.9% when using radiomics. The performance of handcrafted features is highest with 89.4% in accuracy. Results show that CT-SS is still a coarse measure with relatively low
performance. The radiomics method has a similar performance with the proposed handcrafted method, while the latter has better interpretability with a greater potential to be used in clinical practice.

**Evaluation of classifiers**

We then compared the proposed iSARF method to other machine learning classifiers including LR, SVM, NN, and RF. As can be seen, the proposed method demonstrates a consistently superior ROC curve with roughly 2% higher accuracy over other methods. When using the handcrafted features, best performance was achieved at a sensitivity of 90.7%, specificity of 87.2%, and accuracy of 89.4%. The reason lightGBM outperforms other methods may be that lightGBM employs an ensembling strategy where results from 100 trees are summarized for final results, and the Bayesian function is used to efficiently tune the optimal hyperparameters.

**Effects of infection sizes**

Next, we further broke down the whole dataset into five size groups to better evaluate the performance of methods. Results reveal that, although overall accuracy seems high in the size group of $<0.01\%$, the method actually shows low sensitivity, which means the COVID-19 patients with small infections (most likely in the early stage) are difficult to detect. Similarly, it shows low specificity in the size group of $>10\%$, which suggests the subjects with large infections would be mainly treated as COVID-19. This suggests that the studies that report only overall performance (e.g., figure 6(B)) would easily be biased due to the imbalanced distribution of infected lesion sizes. For the remaining sizes, the overall screening performance increases gradually as the infected lesion size becomes larger. For example, accuracy increases from 73.7%, 84.6% to 93.3% in the groups of 0.01%–0.1%, 0.1%–1%, and 1%–10%.

**Ablation study of proposed handcrafted features**

An ablation study on the handcrafted features was performed (supplementary material S1 table (available online at stacks.iop.org/PMB/66/065031/mmedia)). The results show that the volume and surface features have a relatively better performance with an ACC of 82.5%. By combining volume and surface, the ACC improves to 87.1%, and the addition of the histogram further increases the ACC to 88.4%. The use of all features increases it by 1% more as ACC is 89.8%.

The features picked mostly by the proposed method are shown in figure 7. Note that they are sorted by the number of times they were picked in five folds. The results show that the volume and number features are mostly from the lung segments, and several features are from the lung lobes and overall asymmetry. Histogram features are largely selected, including many bins, and surface features focus on the short range of three and six voxels.

**Additional evaluation of independent thick slice data**

An independent thick-slice testing dataset was employed to test the generalizability of the proposed method (sample images could be seen in the supplementary material S1 figure). This dataset includes CT images with thick slice thickness (5 mm) from 400 COVID-19 patients and 334 CAP patients. From the main dataset (figure 6), we constructed a single model and applied this testing dataset. Results show the proposed method achieved a sensitivity of 83.0%, a specificity of 87.1%, and an accuracy of 84.9% (supplementary material S2 figure). Similarly, the performance of handcrafted features consistently outperforms the radiomics as well as the CT-SS, and the proposed method shows significantly increased AUC over other classifiers, including SVM, NN and RF when using handcrafted features ($p < 0.05$ with the Delong test).

**Discussion**

Positive results in nucleic acid-based laboratory testing are still the diagnosis standard for COVID-19. Nevertheless, laboratory testing was also reported to be time-consuming and having a high false-negative rate, and thus CT imaging is suggested as a complementary tool or even a recommended approach to COVID-19 screening in China. A recent study compared RT-PCR results with chest CT scans in 1014 subjects, and found that chest CT scans achieved a high sensitivity of 97%, a low specificity of 25%, and an accuracy of 68%. However, the performance could potentially be improved if the chest CT reading could leverage advanced image analysis techniques.

To date, there have been some reported studies on machine-learning-based COVID-19 CT image analysis. Wang et al reported a disease screening framework using a deep learning method with a sensitivity of 74% and a specificity of 67%, in a total of 44 COVID-19 patients and 55 typical viral pneumonia patients (Wang et al 2021). Xu et al studied the early screening of COVID-19 from influenza-A viral pneumonia and healthy cases with a total of 618 CT samples (Xu et al 2020). Wang et al performed a study on screening among 313 COVID-19 and 229 non-COVID-19 patients (Wang et al 2020c). Shi et al presented a machine learning model to distinguish severe patients from a nonsevere group, where 45 severe cases and 151 nonsevere were involved (Shi et al 2021).
In this study, we propose a machine learning method with COVID-19 specific features to explore the potential application of CT-based COVID-19 screening. The results show high performance with 90.7% sensitivity, 87.2% specificity, and 89.4% accuracy. Furthermore, although thin-section CT is preferred as it has the most detailed information for the best clinical diagnosis, there exist CT images with thick slice thickness in clinical practice, for many reasons, such as the considerations of lower radiation doses or limited capacity of CT scanners. To the best of our knowledge, this is the first study that has evaluated the model generalizability of performance on thick slice CT images, and the results are promising, with a sensitivity of 83.0%, a specificity of 87.1%, and an accuracy of 84.9% (supplementary material S2 figure). These findings suggest that the proposed method has great generalizability and might largely contribute and help clinicians in distinguishing COVID-19 from CAP in both thin and thick CT images.

One of the compelling findings of the current study is the unbalanced distribution of infection size between COVID-19 and CAP patients. In accordance with previous reports (Bernheim et al 2020, Huang et al 2020, Li and Xia 2020), the observation regarding the high prevalence of bilateral and peripheral lung opacities, multiple lobular and subsegmental areas infections, and multifocal involvement is the most typical findings on CT images of COVID-19 patients. For example, COVID-19 pneumonia tends to manifest on CT images as bilateral and with a slight predominance in the right lower lobes, with the mean involved lung segment number of 10.5 (Shi et al 2020). Multifocal involvement was more common than unifocal involvement in COVID-19 patients (Li and Xia 2020). In the study of Ouyang et al (2020), a 3D CNN network was used to discriminate COVID-19 from CAP in chest CT images, in which the imbalanced distribution of infection size between COVID-19 and CAP patients was considered, and a dual-sampling strategy was developed to alleviate the imbalanced problem. Equipped with a balanced infection size sampling strategy, a discrimination performance with 87.5% accuracy was achieved. In the present study, we proposed a size-aware strategy that separated the cohorts into different size groups through the decision tree, thus reducing the bias of infection size. From five-fold experiments, the obtained decision trees show that, when separating the dataset into five subgroups (<0.01%, 0.01%–0.1%, 0.1%–1%, 1%–10% and >10%), consistent optimal classification results of accuracy could be obtained figure (6). Meanwhile, we also compared the current handcrafted features in this study with CT-SS and radiomics features. The results show that the handcrafted features have a superior performance, largely over CT-SS at 25.6% and also on radiomics features at 2.5% in accuracy. Similar observations could be made on other classifiers (figure 6). This suggests that the proposed handcrafted features summarized the disease related information well.

A few limitations still exist in this study. First, the current data collection did not include sufficient clinical characteristics to examine the correlation between symptoms and radiologic findings. For example, the
CORADS scoring system is widely used in clinical facilities. A future study might explore the description of the radiologic reading of COVID-19 pneumonia and its correlation with scoring systems. Second, only baseline CT findings of COVID-19 patients that clinicians and radiologists first encountered were included, rather than findings from follow-up CT scans. Therefore, a follow-up study of the disease progression is needed in future work. Third, the symptom severity of COVID-19 and differential diagnosis of pneumonia subtypes were not included in this study, which warrants further investigation. Finally, there exist differences in subject demographics (e.g., age, gender, acquisition time) between COVID-19 and CAP patients in data collection, where future studies could be better designed to involve matched participants.

It is worth noting that the proposed method has been integrated into a cloud platform as an online service and is available to over 50 clinical facilities in China to date. Meanwhile, our research portals that include computer hardware such as GPUs and software have been deployed and used in over 90 hospitals, of which over 20 are in Wuhan, China.

Conclusion

The early screening and control of COVID-19 is vital for public health. Therefore, automated techniques with extensive and reliable imaging clues for COVID-19 disease could greatly assist in the clinical diagnosis practice. In this paper, we have developed a machine learning framework for distinguishing COVID-19 from CAP in CT images. The CT images were first segmented into infections and lung fields with a deep learning model, and a set of handcrafted location-specific features were then designed with prediction models generated. The proposed method produces superior performance over the models that were trained with CT-SS and radiomics features as well as other classifiers, and has a great generalizability both in thin- and thick-section CT images, demonstrating the effectiveness and clinical significance of our approach. In conclusion, CT imaging demonstrates high accuracy through the machine learning technique, and thus could be an efficient tool for COVID-19 screening. The infection size bias needs to be considered in both the method development and the result evaluation process. Although our model is built upon the specific application of pneumonia classification, the infection size-aware strategy can be applied to other size bias classification tasks to further improve the classification performance.

Funding

This work was supported in part by the National Key Research and Development Program of China (2018YFC0116400), Wuhan Science and Technology Program (2018060401011326), Hubei Provincial Novel Pneumonia Emergency Science and Technology Project (2020FCA021), Huazhong University of Science and Technology Novel Coronavirus Pneumonia Emergency Science and Technology Project (2020kfXGYJ014), and Novel Coronavirus Special Research Foundation of the Shanghai Municipal Science and Technology Commission (20441900600).

Competing interests

The funder provided support in the form of salaries for authors FS, DW, YW, HY, HJ, YH, YG, and DS, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the ‘author contributions’ section.

Author contributions

Conceptualization: Feng Shi, Dinggang Shen
Data curation: Liming Xia, Fei Shan, Bin Song
Formal analysis: Feng Shi, Ying Wei, Yichu He
Funding acquisition: Liming Xia, Fei Shan, Dinggang Shen
Investigation: Dijia Wu, Huiting Jiang, Yaozong Gao, He Sui
Methodology: Yaozong Gao, Huiting Jiang
Supervision: Dinggang Shen
Validation: Ying Wei, Yichu He, Liming Xia, Fei Shan, Bin Song
Visualization: Ying Wei, Huiting Jiang, Huan Yuan
Writing—original draft: Feng Shi, Ying Wei
Writing—review and editing: Dijia Wu, Dinggang Shen
References

Bernheim A et al, 2020 Chest CT findings in coronavirus disease-19 (COVID-19): relationship to duration of infection Radiology 295 685–91
Chan J F et al, 2020 A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster Lancet 395 519–23
Chung M et al, 2020 CT imaging features of 2019 novel coronavirus (2019-nCoV) Radiology 295 202–7
Colombani C et al, 2013 Application of Bayesian least absolute shrinkage and selection operator (LASSO) and BayesCpi methods for genomic selection in French Holstein and Montbeliarde breeds J. Dairy Sci. 96 575–91
Fang Y et al, 2020 Sensitivity of chest CT for COVID-19: comparison to RT-PCR Radiology 296 E115–7
Hansel D M, Bankier A A, MacMahon H, McLoud T C, Muller N L and Remy J 2008 Fleischner society: glossary of terms for thoracic imaging Radiology 246 697–722
Huang C et al, 2020 Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China Lancet 395 497–506
Kim K K S-J, Lustig M, Boyd S and Gorinevsky D 2007 An interior-point method for large-scale 1-regularized least squares IEEE J. Sel. Top. Sign. Proc. 1 606–17
Li Q et al, 2020 Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia New Engl. J. Med. 382 1199–207
Li X, Zeng X, Liu B and Yu Y 2020 COVID-19 infection presenting with CT halo sign Radiol.: Cardiothorac. Im. 2 e200026
Li Y and Xia L 2020 Coronavirus disease 2019 (COVID-19): role of chest CT in diagnosis and management AJR Am. J. Roentgenol. 214 1280–6
Liu Y, Gayle A A, Wilder-Smith A and Rocklov J 2020 The reproductive number of COVID-19 is higher compared to SARS coronavirus J. Travel Med. 27 taaa021
Milletari F, Navab N and Ahmadi S A 2016 V-net: fully convolutional neural networks for volumetric medical image segmentation 2016 Fourth Int. Conf. on 3D Vision (3DV) pp 565–71(IEEE)
Oei G C et al, 2004 Severe acute respiratory syndrome: temporal lung changes at thin-section CT in 30 patients Radiology 230 836–44
Ouyang X, Huo J, Xia L, Shan F and Shen D 2020 Dual-sampling attention network for diagnosis of COVID-19 from community acquired pneumonia IEEE Trans. Med. Imaging 39 2595–605
Sahin A R et al, 2019 Novel coronavirus (COVID-19) outbreak: a review of the current literature EJMO 4 1–7
Shan F et al, 2020 Abnormal lung quantification in chest CT Images of COVID-19 patients with deep learning and its application to severity prediction Med. Phys. (https://doi.org/10.1002/mp.14609)
Shen D, Gao Y, Munoz-Barrutia A, Debbc D C and Percannella G 2020 Guest editorial: special issue on imaging-based diagnosis of COVID-19 IEEE Trans. Med. Imaging 39 2569–71
Shi H et al, 2020 Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study Lancet Infect. Dis. 20 425–34
Shi W et al, 2021 A deep learning-based quantitative computed tomography model for predicting the severity of COVID-19: a retrospective study of 196 patients Annals Translational Med. 9 216
Song F et al, 2020 Emerging 2019 novel coronavirus (2019-nCoV) pneumonia Radiology 295 210–7
Sun X and Wu X 2014 Fast implementation of Delong’s algorithm for comparing the areas under correlated receiver operating characteristic curves IEEE Signal Process Lett. 21 1389–93
Vasquez M M, Hu C, Roe D J, Chen Z, Halonen M and Guerra S 2016 Least absolute shrinkage and selection operator type methods for the identification of serum biomarkers of overweight and obesity: simulation and application BMC Med. Res. Methodol. 16 154
Wang C, Daniels M J, Scharfstein D O and Land S 2010 A bayesian shrinkage model for incomplete longitudinal binary data with application to the breast cancer prevention trial J. Am. Stat. Assoc. 105 1333–46
Wang C, Horby P W, Hayden F G and Gao G F 2020a A novel coronavirus outbreak of global health concern Lancet 395 470–3
Wang D et al, 2020b Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China Jama 323 1061–9
Wang X et al, 2020a A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT IEEE Trans. Med. Imag. 39 2615–25
Wang S et al, 2021 medRxiv: A deep learning algorithm using CT images to screen for coronavirus disease (COVID-19) Eur. Radiol. (https://doi.org/10.1007/s00330-021-04775-1)
de Wit E, van Doremalen N, Falzarano D and Munster V J 2016 SARS and MERS: recent insights into emerging coronaviruses Nat. Rev. Microbiol. 14 523–34
Wong K et al, 2003 Thin-section CT of severe acute respiratory syndrome: evaluation of 73 patients exposed to or with the disease Radiology 228 395–400
Wu J T, Leung K and Leung G M 2020 Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study Lancet 395 669–79
Xu X et al, 2020 A deep learning system to screen novel Coronavirus disease 2019 pneumonia J. Eng. 6 1122–9
Yang R et al, 2020 Chest CT severity score: an imaging tool for assessing severe COVID-19 Radiol.: Cardiothorac. Im. 2 e200047
Zu Z Y et al, 2020 Coronavirus disease 2019 (COVID-19): a perspective from China Radiology 296 E15–25