

# Large-Scale Screening of COVID-19 from Community Acquired Pneumonia using Infection Size-Aware Classification

Feng Shi, Liming Xia, Fei Shan, Dijia Wu, Ying Wei, Huan Yuan,  
Huiting Jiang, Yaozong Gao, He Sui, Dinggang Shen

proposed a machine learning approach for screening COVID-19 out of CAP(community acquired pneumonia )

- presented an infection size-aware method, and broke down the evaluation of COVID-19 classification task into multiple infection size ranges, respectively.
- proposed a location-specific feature extraction process, where COVID-19 features were collected and tailored according to current understanding of radiographic appearance.
- The proposed method was throughout evaluated on a large-scale dataset from multiple centers, covering patients with age ranging from 12 to 98 years.

They propose to utilize the disease characteristics, i.e., infection locations and spreading patterns, to extract handcrafted features. To do that, they automatically segmented infected lung regions and lung fields bilaterally. The infected lung regions were mainly related to manifestations of pneumonia, such as mosaic sign, ground-glass opacity (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 their in-house research portal software. Specifically, a deep learning based network called VB-Net was employed for image segmentation.

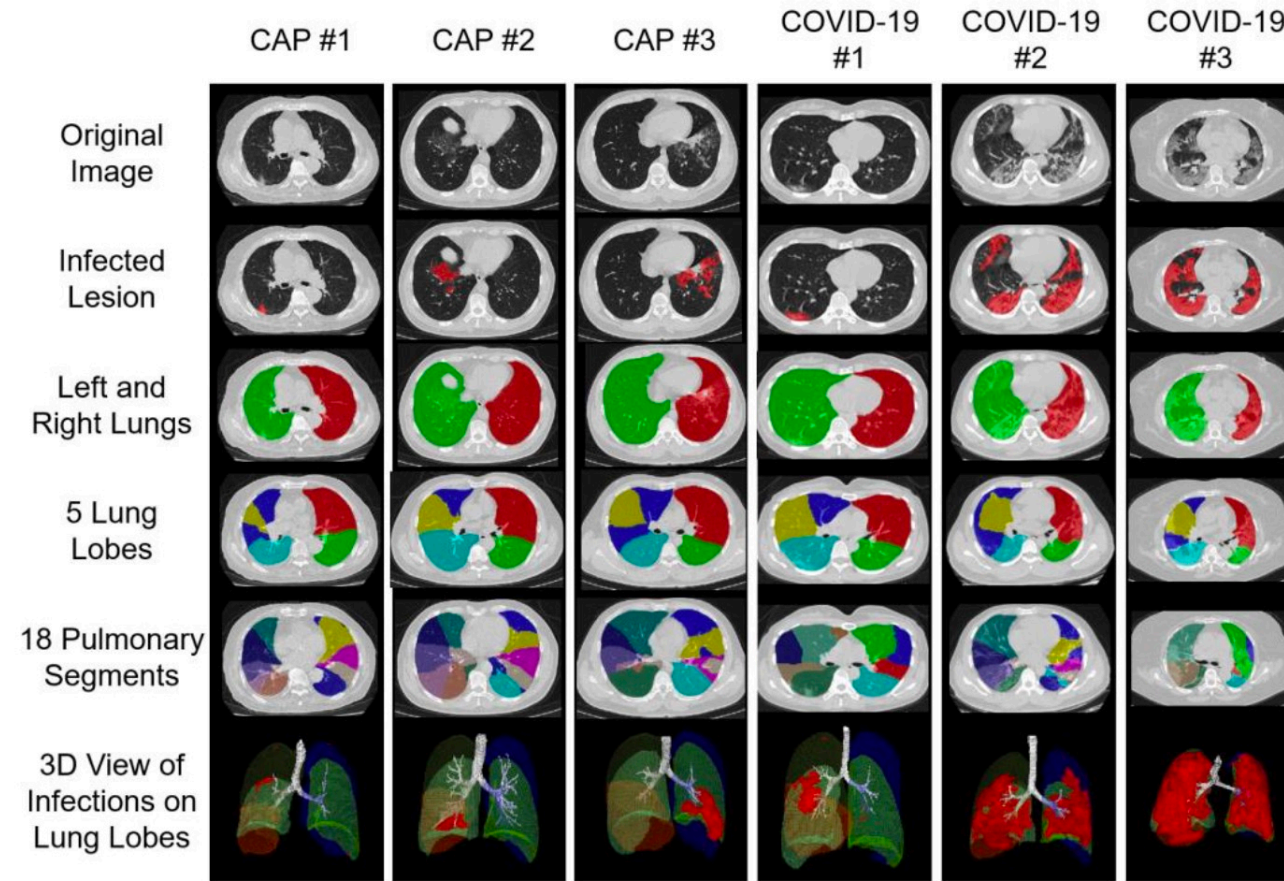


Fig.1. Illustration of lung images, preprocessed results of infected lesions and lung fields on 3 CAP (left 3 columns) and 3 COVID-19 patients (right 3 columns), respectively.

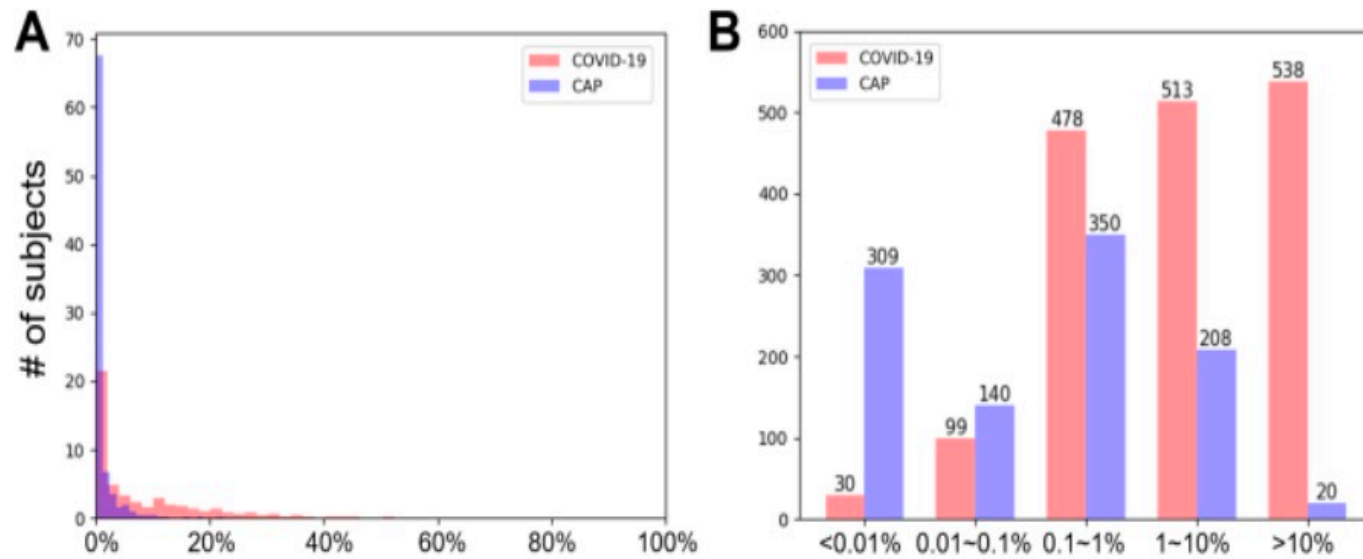


Fig.2. Illustration of size distribution in patients. (A) shows a highly skewed distribution for the number of subjects as a function of infection size. (B) separates the dataset into 5 groups with exponential size ranges.

They refer the infection size as the volume of infected regions against the volume of whole segmented lung.

Such 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 leads to a low performance for the classification of middle-size groups and does not reflect the real radiographic appearance differences between two types of pneumonia. In this work, they proposed a new classification strategy.

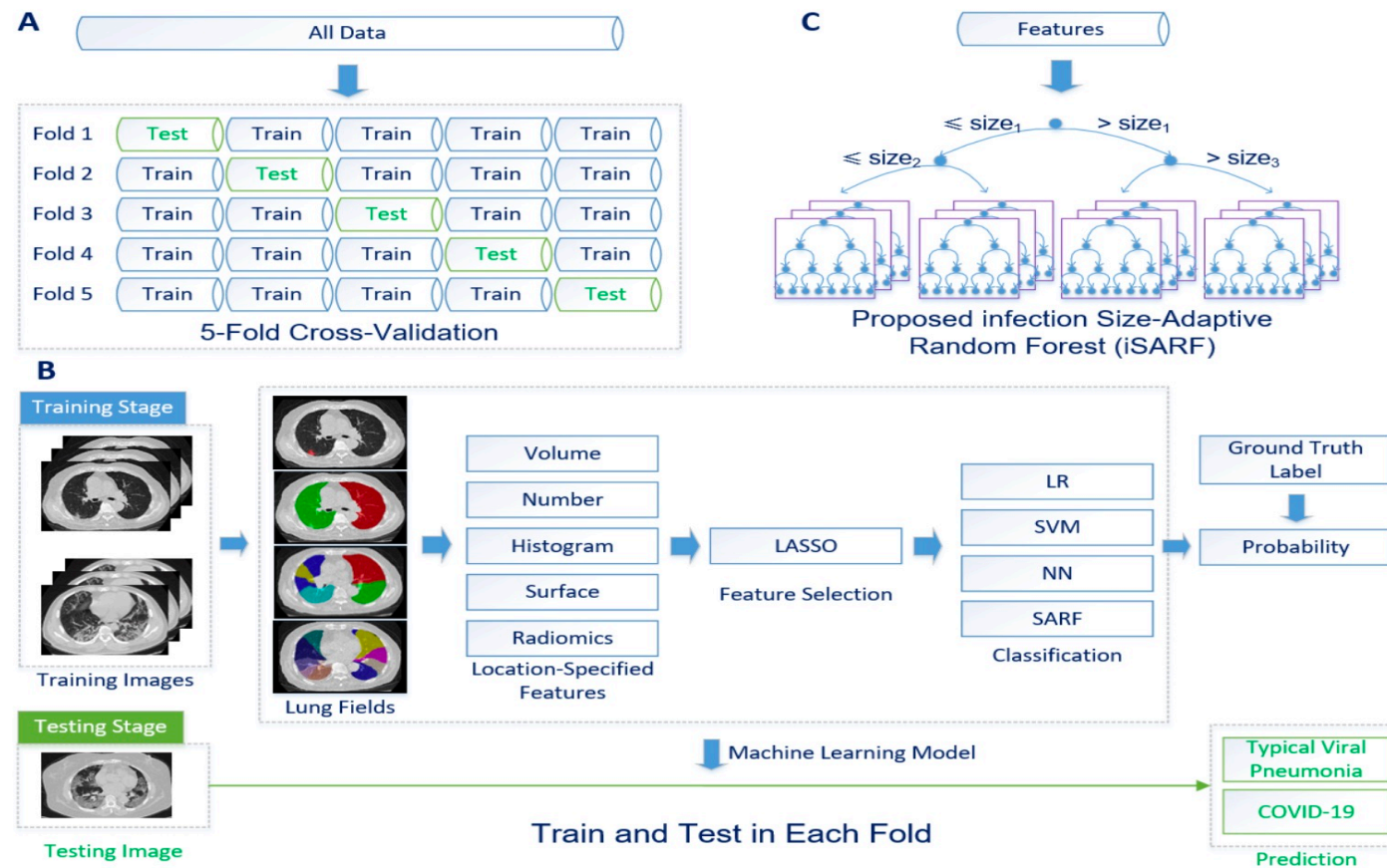



Fig. 3. Flow chart of (A) the 5-fold cross validation, (B) train and test process in each fold, and (C) the proposed size-aware method



In this study, they proposed a series of handcrafted features to be automatically extracted in CT images from infections and lung fields. These features are composed of 4 categories, including the volume, infected lesion number, histogram distribution, and surface area. The detailed feature distribution framework is shown in Fig. 4.



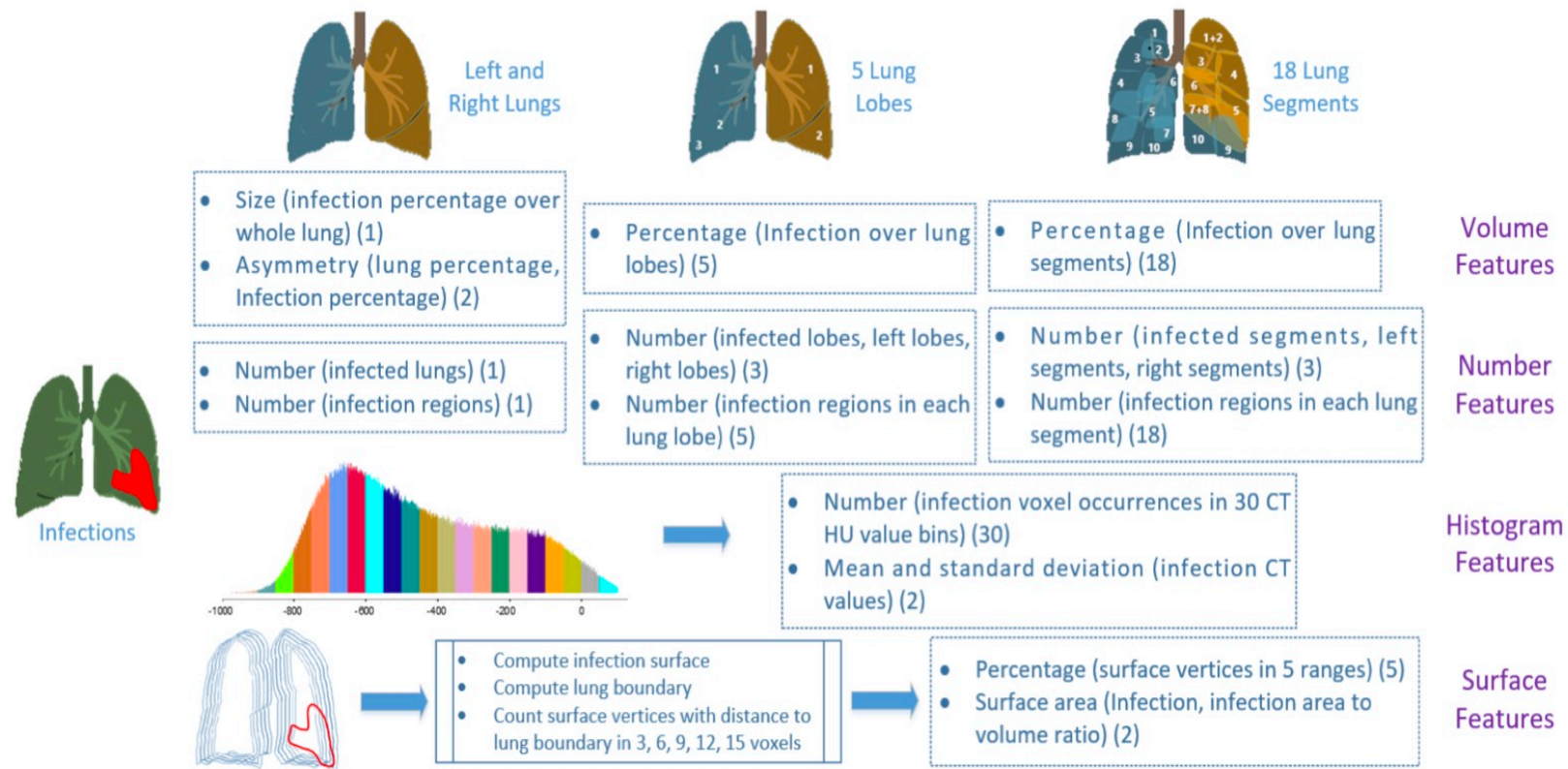


Fig. 4. Illustration of the extraction of handcrafted image features. There are totally 96 features, including 26 volume features, 31 number features, 32 histogram features, and 7 surface features.



After generating the features, they apply machine-learning methods to select proper features, and predict COVID-19 patients from CAP patients.

In the training stage, they employed least absolute shrinkage and selection operator (LASSO) to explore the optimal subset of clinico-radiological features for the classification, due to its ability to provide variable importance and interpretability. After that, the selected features were fed into LR, SVM, NN, and the proposed method, respectively, to determine their hyperparameters for disease diagnosis. For LR, they used with default parameters of penalty as L2 (the norm used in the penalization) and regularization C as 1. For SVM, they used radial basis function (RBF) kernel with parameters of regularization C and gamma being determined through an inner 5-fold grid search. For NN, they used MLP Classifier with one hidden layer of 100 nodes and with max iterations of 500. For the proposed method, 100 trees of random forest were used for each size group, maximum depth of tree is 10, and Gini impurity is employed to measure the quality of a split. These trained models were then applied to new test images to predict their probability of being COVID-19 against CAP in the testing stage.

The performance of the proposed method was evaluated through a 5-fold cross-validation. Comparison methods include logistic regression (LR), support vector machine (SVM), and neural network (NN). The mean ROC was averaged from the 5 folds.

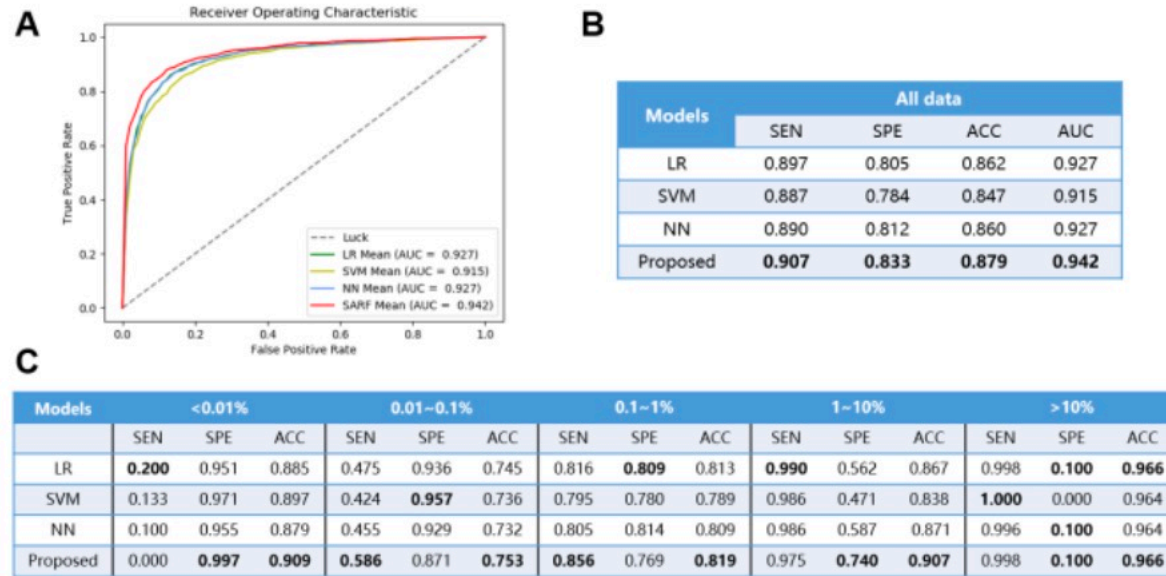


Fig. 5. Illustration of (A) the mean ROC curves from 5 folds in different classifiers, (B) overall performance, and (C) performance after dividing into 5 groups. SEN means sensitivity, SPE means specificity, ACC means accuracy, and AUC means area under curve.