# CT-based Patient Triage of COVID-19: Radiomics Prediction of ICU Admission, Mechanical Ventilation, and Death of Patients

Xianghao Zhan[(1)], Yiheng Li[(2)]
(1) Department of Bioengineering, Stanford University
(2) Department of Biomedical Data Science, Stanford University

**Abstract**

This study used early on data of COVID-19 patients, including CT images, electronic health records, and lab results to build models for predicting high-risk patients that would have one of the three end points: ICU admission, mechanical ventilation, or potential death. With the attempts of combining each data type and the implementation of re-sampling and feature selection techniques, the optimal models were trained and selected based on a cohort of 1662 patients while the model performances for three tasks were tested on an unseen cohort of 700 patients. Bootstrapping were performed on each kind of data combination, and the results showed the combination of radiomics with clinical features and lab test data manifested the optimal results: (AUROC and 95% CI): ICU admission:0.911(0.887-0.935); Mechanical Ventilation: 0.921(0.905-0.944); Death: 0.834(0.786-0.880). Besides, we proved that in this problem, compared to radiologist findings (which are manual labeling of CT images by radiologists), AI-extracted radiomics features together with clinical and lab results showed a more robust ability of patient outcome prediction. This work could provide the medical practitioners with the risk information of each patient for better allocation of the limited medical resources. The code of this project is available on https://github.com/terryli710/COVID-19_prediction.

## Introduction

In order to achieve rapid stratification and timely intensive care of COVID-19 patients, as well as the optimization of medical resource allocation under this unprecedented public health emergency, in this study, prediction models, which take radiomics based features, are built to predict high-risk inpatients at the time of admissions.

## Materials and Methods

### Patient enrollment and population

In our study, 3522 NCP inpatients from 39 designated hospitals in China between December 27, 2019, and March 31, 2020, were preliminarily included according to the criteria as follows: (a) confirmed positive SARS-CoV-2 nucleic acid test; (b) thin-section CT examinations ( > 2.5 mm) and laboratory tests on the date of admission; (c) clear prognosis information was available (discharge, or adverse outcomes including in-hospital death, the admission to intensive care unit [ICU] and requiring mechanical ventilation support [MV]). Further, patients were filtered. Exclusion criteria included (a) Patients age < 18; (b)Patients transferred to other hospitals or remaining hospitalized without any adverse outcomes; (c) CT scans without a lung-related convolutional kernel; (d) CT scans lack serial information or with motion artifacts or significant resolution reductions. Figure 1 shows the procedure to enroll patients.

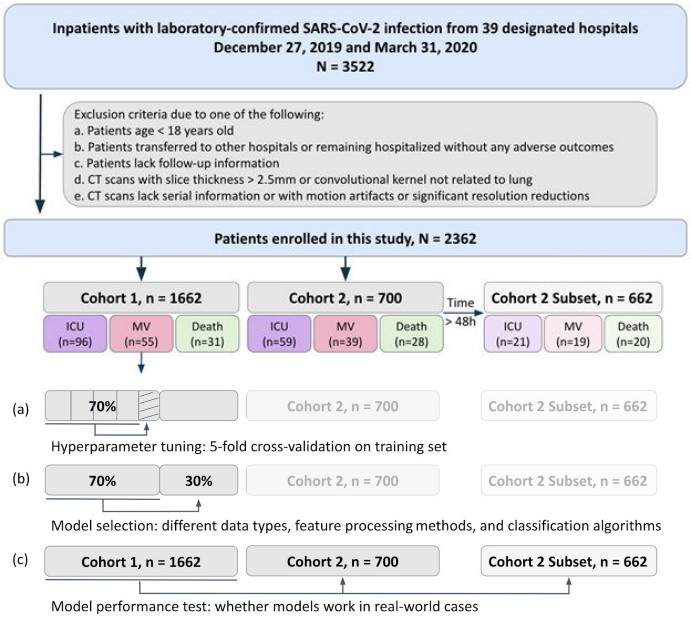


**Figure 1. Workflow of The Proposed Study.**

### Data Types

We reviewed clinical electronic medical records, laboratory results, radiomics features, and radiologist findings of all included COVID-19 patients. The following data were collected and analyzed: (i) Radiomics features (Radiom) (ii) Laboratory results (Lab) (iii) Clinical features (Clin) (iv) Radiologist findings.

**Data Split**

We split the cohort into two subsets based on the date of admission: cohort 1 (n = 1662) for model development and cohort 2 (n = 700) for the validation and comparison of models (Figure 1). We constructed radiomics-based machine learning models for three prediction tasks, including admission to ICU (positive cases in cohort 1, n = 96; cohort 2, n = 60), requiring MV (positive cases in cohort 1, n = 55; cohort 2, n = 39), and in-hospital death (within 28 days) (positive cases in cohort 1, n = 32; cohort 2, n = 29). In cohort 1 (n=1662), we further splitted the data into a training set and a test set with a ratio of 7:3.

**Feature Engineering**

To build a data fusion model, in addition to using the raw data to select model and do prediction, while modeling, we applied specific feature processing methods: 1) SMOTEENN [1] 2) PCA [2-4] 3) LASSO feature selection [5] 4) GUS (generic univariate selection) 5) FPR (false positive rate test).

**Prediction Models**

The following multivariable models were developed and compared with the sklearn package [6]: (a)Radiom; (b)RadiomClin; (c)RadiomClinLab. The algorithm used in this study included: 1) Logistic Regression (LR) [7]; 2) Random Forest (RF) [8]; 3)Support Vector Machine (SVM) [9-10]; 4) Multilayer Perceptron (MLP) [11]; 5) LightGBM [12]. The hyperparameter tuning, model selection and model performance test pipeline could be shown in Fig. 1.

**Results**

Firstly, upon finding the best set of hyperparameters on the cross-validation on the 70% training samples of Cohort 1, we tested all combinations of data, feature engineering methods and classification algorithms and reported the performances on the 30% test samples of Cohort 1 in table 1. According to the results, the AUROC and AUPRC indicated that the optimal models selected based on the test samples in Cohort 1 could classify the patients based on each of the data types while RadioClinLab and RadioClin showed generally better classification performances. The best ICU admission prediction models were: 1) Radiom: SMOTEENN, Lasso feature selection with C=50, multilayer perceptron; 2) RadioClin: SMOTEENN, Lasso feature selection with C=0.5, lightGBM; 3) RadioClinLab: SMOTEENN, Lasso feature selection with C=1; 4) ClinLab: SMOTEENN, Lasso feature selection with C=1, logistic regression; 5) R Score: logistic regression. The best mechanical ventilation prediction models were: 1) Radiom: SMOTEENN 1:3, lightGBM; 2) RadioClin: SMOTEENN 1:3, lightGBM; 3) RadioClinLab: SMOTEENN, Lasso feature selection with C=1, lightGBM; 4) ClinLab: SMOTEENN 1:3,multilayer perceptron; 5) R Score: logistic regression. The best death prediction models were: 1) Radiom: SMOTEENN, Lasso feature selection with C=1; 2) RadioClin: SMOTEENN, FPR feature selection with criterion F-classif; 3) RadioClinLab: SMOTEENN, Lasso feature selection with C=1, support vector machine; 4) ClinLab: SMOTEENN, Lasso feature selection with C=30, multilayer perceptron; 5) R Score: logistic regression.

| | **ICU** | | | **MV** | | | **Death** | | |
|---|---|---|---|---|---|---|---|---|---|
| Data | AUROC | ACC | AUPRC | AUROC | ACC | AUPRC | AUROC | ACC | AUPRC |
| Radiom | 0.732 | 0.780 | 0.261 | 0.823 | 0.954 | 0.307 | 0.881 | 0.970 | 0.300 |
| RadioClin | 0.836 | 0.826 | 0.383 | 0.836 | 0.826 | 0.383 | 0.948 | 0.960 | 0.395 |
| RadioClinLab | 0.837 | 0.824 | 0.307 | 0.850 | 0.970 | 0.420 | 0.826 | 0.984 | 0.417 |
| ClinLab | 0.876 | 0.784 | 0.335 | 0.876 | 0.784 | 0.335 | 0.838 | 0.968 | 0.121 |
| R Score | 0.600 | 0.939 | 0.096 | 0.607 | 0.967 | 0.065 | 0.704 | 0.979 | 0.056 |

**Table 1. The performances of optimal models on the test set of Cohort 1.**

Then with the optimals model selected based on the 30% test samples of Cohort 1, we validated the model performances on unseen data in Cohort 2 (N=700). The receiver operating characteristic curves and precision recall curves of different data types in three prediction tasks were shown in Fig. 2. According to the results, it

was clear that with a larger test set, there was a general improvement in AUROC and AUPRC of the optimal models. The models with the most diverse types of data (RadioClinLab) outperformed the other four types of models with the highest AUROC and AUPRC.

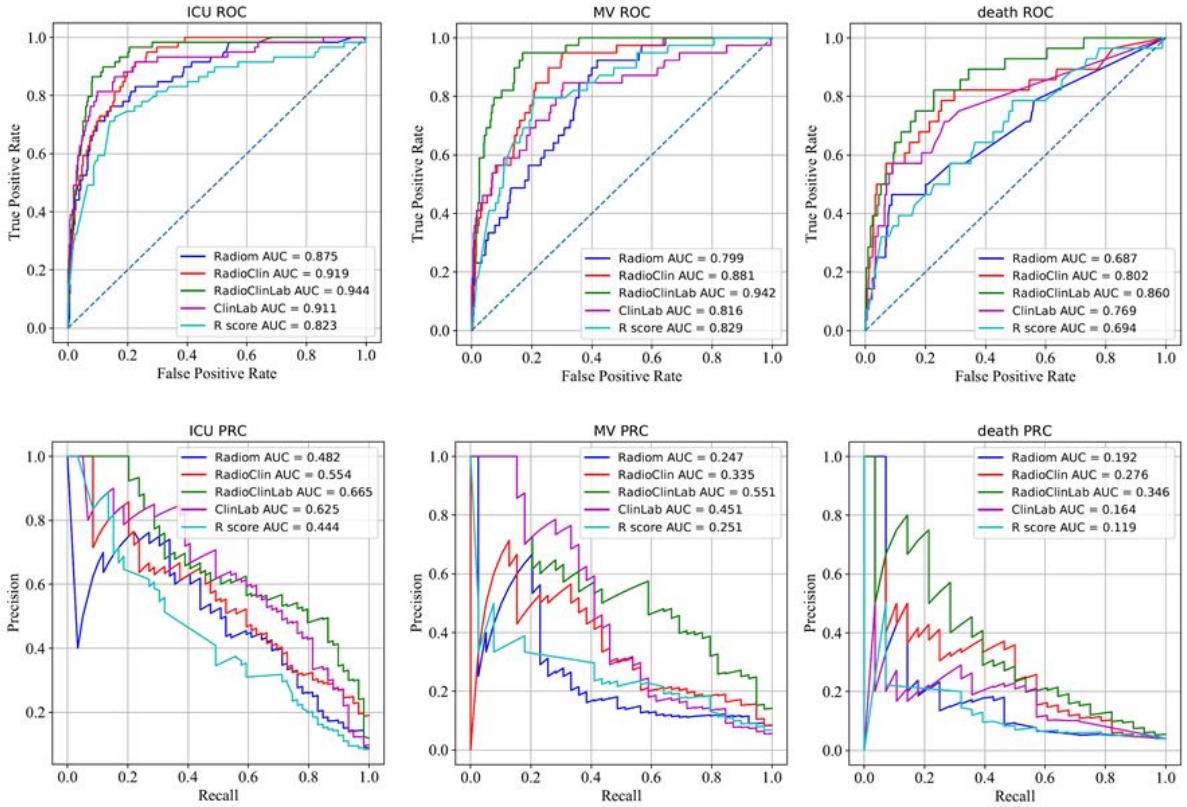


**Fig. 2 The receiver operating characteristic curves and the precision recall curves of the optimal models on predicting samples in Cohort 2.**

To test the model robustness and verify the statistical significance in model comparison, we did the bootstrapping experiments for thirty times and reported the results in the box plot in Fig. 3. Meanwhile, paired one-sided t-test was done to compare specific models in terms of AUPRC and AUROC. According to the results, the Radiom models outperformed radiological score models in AUROC in ICU prediction and in AUPRC in all three cases with statistical significance ($p\_value<0.05$) while there were not statistically significant differences of AUROC in MV prediction and death prediction. RadioClin models outperformed Radiom models in AUROC and AUPRC in all three cases ($p\_value<0.01$). RadioClinLab models outperformed ClinLab models in AUROC in three prediction tasks and in AUPRC in MV and death prediction ($p\_value<0.05$) while there was no statistically significant difference in ICU prediction. Generally, the RadioClinLab models were the models with the best performances when compared with other models in all of the three outcome prediction tasks.

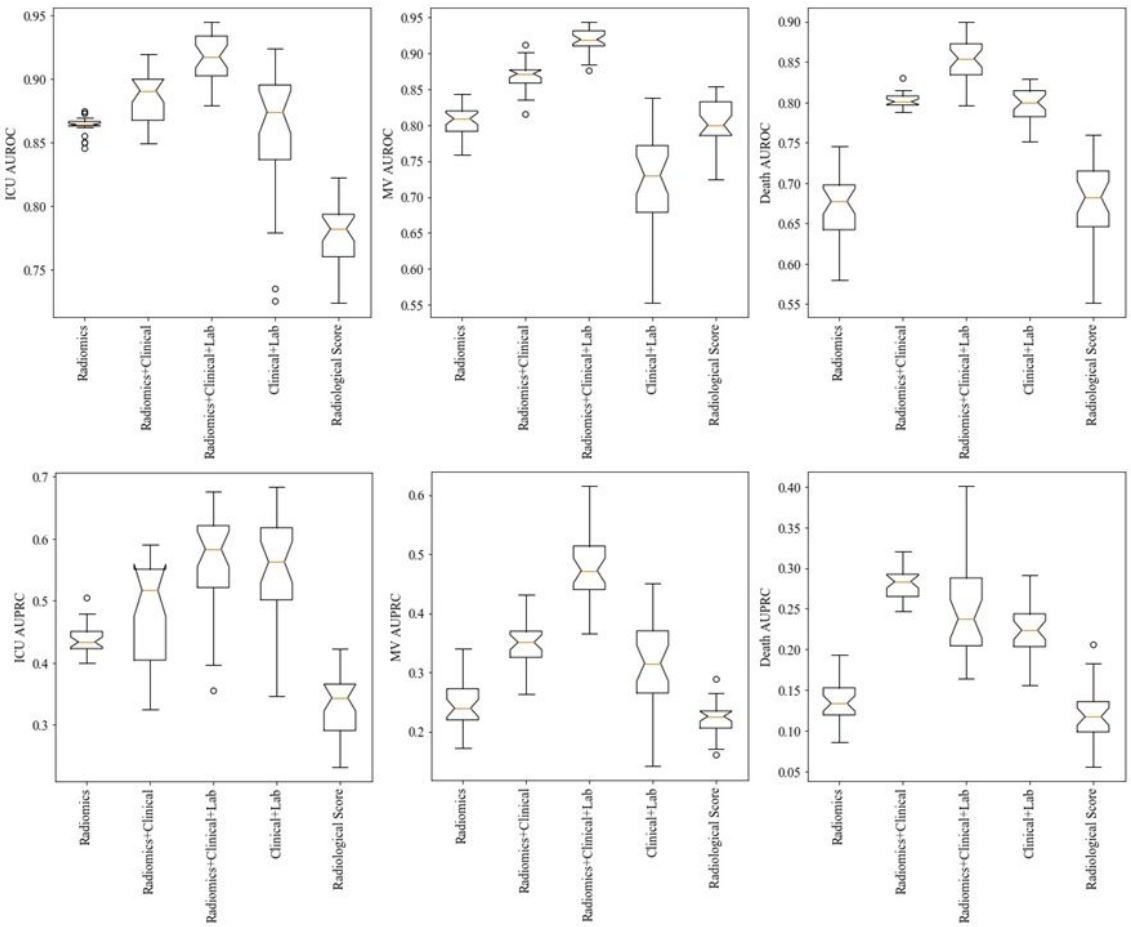


**Fig. 3 Box plots of the AUROC and AUPRC in thirty bootstrapping experiments on Cohort 2.**

To interpret the models, we extracted the most important features based on the RadioClinLab models in the bootstrapping experiments by normalizing the feature importance in each experiment and averaging the values over all thirty bootstrapping experiments. The ten most important features for the three outcome prediction tasks were shown in table S1. The features clearly showed the contribution of different types of data and the important risk factors such as the dyspnea and age.

**Conclusion**

In this work, a machine learning-based patient triaging method was proposed to predict the outcomes of patients in terms of ICU admission, the requirement of mechanical ventilation and potential death based on the CT features and electronic health records upon hospital admission. Models with five different types of data were trained: radiomics data; radiomics data with demographics, clinical symptoms and comorbidity information; radiomics data with demographics, clinical symptoms, comorbidity information and lab tests; radiological scores. The models were trained on a cohort of 1662 COVID-19 patients and the performances were validated on an unseen cohort of 700 COVID-19 patients with high AUROC and AUPRC. The model with radiomics, demographics, clinical symptoms, comorbidity and lab tests showed the best performances on three prediction tasks while various models based on different types of data could also enable the users to flexibly choose based on the data available. This work enables the risk evaluation of COVID-19 patients and facilitates the resource allocation to those with high risks.

**Contribution**

Xianghao Zhan collaborated on the data agreement with Jinling Hospital and did the data analysis with Yiheng Li with a focus on the following feature engineering methods: raw data, SMOTEENN, SMOTEENN Lasso (C=0.2/0.5/1). He also did hyperparameter tuning for those models and bootstrapping for results of those models and visualized results into ROC and PRC curves and bootstrapped box plots. Yiheng Li trained and tested models for GUS data, FPR data and part of lasso selected data. He also did hyperparameter tuning for those models and bootstrapping for results of those models. He coded the pipeline for merging data types, data preprocessing and model selection, as well as model training and testing. This work is also supported by Qinmei Xu for her support on data collection and radiomics feature extraction, and Dr. Olivier Gevaert for his advice and supervision throughout the study.

The codes are available on https://github.com/terryli710/COVID-19_prediction.

**Reference**


1. Batista, Gustavo EAPA, Ana LC Bazzan, and Maria Carolina Monard. "Balancing Training Data for Automated Annotation of Keywords: a Case Study." WOB. 2003.
2. Pasini, Giorgia. "Principal component analysis for stock portfolio management." International Journal of Pure and Applied Mathematics 115.1 (2017): 153-167.
3. Yang, Libin. "An application of principal component analysis to stock portfolio management." (2015).
4. Zhan, Xianghao, et al. "Discrimination between Alternative Herbal Medicines from Different Categories with the Electronic Nose." Sensors 18.9 (2018): 2936.
5. Tibshirani, Robert. "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society: Series B (Methodological) 58.1 (1996): 267-288.
6. Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." the Journal of machine Learning research 12 (2011): 2825-2830.
7. Cramer, J.S.. (2002). The Origins of Logistic Regression. Tinbergen Institute, Tinbergen Institute Discussion Papers. 10.2139/ssrn.360300.
8. Ho, T. K. (1995, August). Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE.
9. Asa Ben-Hur, David Horn, Hava T. Siegelmann, and Vladimir Vapnik. 2002. Support vector clustering. J. Mach. Learn. Res. 2 (March 2002), 125–137.
10. Zhan, Xianghao, et al. "Feature Engineering in Discrimination of Herbal Medicines from Different Geographical Origins with Electronic Nose." 2019 IEEE 7th International Conference on Bioinformatics and Computational Biology (ICBCB). IEEE, 2019.
11. R. Collobert and S. Bengio (2004). Links between Perceptrons, MLPs and SVMs. Proc. Int'l Conf. on Machine Learning (ICML).
12. Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree." Advances in neural information processing systems. 2017.


**Supplementary Material**

| Tasks | Important Features |
|---|---|
| ICU | Dyspnea (Clinical Symptom), Age (Demographics), Lactate Dehydrogenase (Lab test), wavelet-LHH_glszm_LargeAreaHighGrayLevelEmphasis_std (Radiomics), White Blood Cell Count (Lab test), original_glszm_SmallAreaLowGrayLevelEmphasis_std (Radiomics), lymphocyte (Lab test), C-reactive protein (Lab test), Hypertension (Comorbidity), Neutrophil (Lab test) |
| MV | Dyspnea (Clinical Symptom), Neutrophil (Lab test), Age (Demographics),Lactate dehydrogenase (Lab test), C-reactive protein (Lab test), Hydroxybutyrate Dehydrogenase (Lab test), Lymphocyte (Lab test), Potassium (Lab test), White Blood Cell Count (Lab test), wavelet-LHH_glszm_LargeAreaHighGrayLevelEmphasis_std (Radiomics) |
| Death | Dyspnea (Clinical Symptoms), Lactate Dehydrogenase (Lab test), Age (Demographics), White Blood Cell Count (Lab test), wavelet-HLH_glcm_InverseVariance_75 (Radiomics), original_firstorder_Minimum_75 (Radiomics), wavelet-LHL_firstorder_Skewness_medium (Radiomics), Neutrophil (Lab test), wavelet-HLL_glszm_LargeAreaLowGrayLevelEmphasis_std (Radiomics), D-dimer (Lab test) |

**Table S1. The 10 most important features in the predictions of three outcomes.**