A Learning-based Model to Evaluate Hospitalization Priority in COVID-19 Pandemics

Yichao Zheng, Yinheng Zhu, Mengqi Ji, Rongpin Wang, Xinfeng Liu, Mudan Zhang, Choo Hui Qin, Lu Fang, Shaohua Ma

PII: S2666-3899(20)30120-3

DOI: https://doi.org/10.1016/j.patter.2020.100092

Reference: PATTER 100092

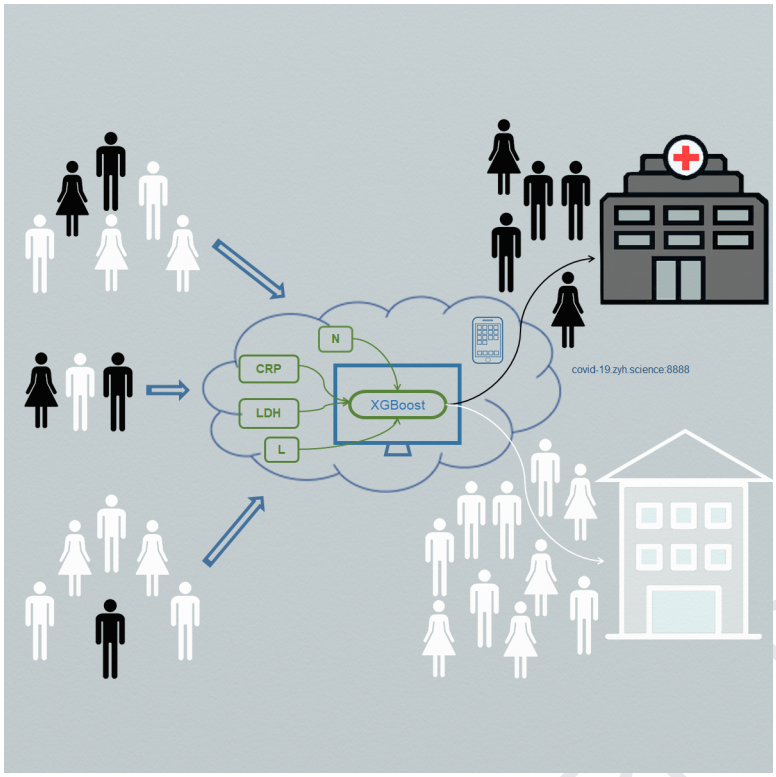
To appear in: *Patterns*

Received Date: 12 May 2020

Revised Date: 30 June 2020

Accepted Date: 29 July 2020

Please cite this article as: Zheng Y, Zhu Y, Ji M, Wang R, Liu X, Zhang M, Qin CH, Fang L, Ma S, A Learning-based Model to Evaluate Hospitalization Priority in COVID-19 Pandemics, *Patterns* (2020), doi: https://doi.org/10.1016/j.patter.2020.100092.



A Learning-based Model to Evaluate Hospitalization Priority in COVID-19 Pandemics

Yichao Zheng[1,4,5], Yinheng Zhu[1,4,5], Mengqi Ji[3,4], Rongpin Wang[2], Xinfeng Liu[2], Mudan Zhang[2], Choo Hui Qin[1,4], Lu Fang[1,4,*] and Shaohua Ma[1,4,6,*]

1. Tsinghua-Berkeley Shenzhen Institute (TBSI), Tsinghua University, Shenzhen, 518055, China
2. Department of Radiology, Guizhou Provincial People's Hospital, Guiyang, 550002, China
3. Department of Automation, Tsinghua University, Beijing, 100084, China
4. Shenzhen International Graduate School (SIGS), Tsinghua University, Shenzhen, 518055, China
5. These authors contribute equally
6. Lead contact: ma.shaohua@sz.tsinghua.edu.cn

[*]Correspondence: ma.shaohua@sz.tsinghua.edu.cn (S.M.); fanglu@sz.tsinghua.edu.cn (L.F.)

Summary

The emergence of novel coronavirus disease 2019 (COVID-19) is placing an increasing burden on the healthcare systems. Although the majority of infected patients have non-severe symptoms and can be managed at home, some individuals may develop severe disease and are demanding the hospital admission. Therefore, it becomes paramount to efficiently assess the severity of COVID-19 and identify hospitalization priority with precision. In this respect, a 4-variable assessment model, including lymphocyte, lactate dehydrogenase (LDH), C-reactive protein (CRP) and neutrophil, is established and validated using the XGBoost algorithm. This model is found effective to identify severe COVID-19 cases on admission, with a sensitivity of 84.6%, a specificity of 84.6%, and an accuracy of 100% to predict the disease progression toward rapid deterioration. It also suggests that a computation-derived formula of clinical measures is practically applicable for the healthcare administrators to distribute hospitalization resources to the most needed in epidemics and pandemics.

Introduction

The novel coronavirus disease 2019 (COVID-19) caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection was firstly reported in last December in China and rapidly spread across the world, affecting over 16 million people worldwide and killing more than a half million infected patients up till now [1-3]. Even worse, the global pandemic of COVID-19 is expected to continue growing, as no effective vaccines have been officially approved for prophylaxis of this disease [4]. Though the growth in detected infections has declined in East Asia and Europe, the number of infections in U.S., south America and African places are witnessed with continuous growth [3]. Moreover, the suspicion on a second generation pandemic outbreak still sustains [5].

In pandemic, a nation's healthcare system bears extraordinary burdens. However, a majority of patients infected with SARS-CoV-2 generally have non-severe disease progression and can be safely managed at home or self-quarantine, and recover under limited and basic medical care [6]. For the infections with severe syndromes or progression toward rapid deterioration, immediate admission to hospitals for close monitoring and intensive treatment has been proven effective to reduce the complications and mortality [7]. Therefore, identifying the COVID-19 patients at high risk for severe illness and prioritizing them for immediate admission to hospitals becomes urgently demanded, especially in nations and territories where the healthcare systems are insufficient to administrate all infections and suspicions.

Some studies have been reported to predict deterioration and mortality during hospitalization [8-15], and even predict the probability of SARS-CoV-2 infections that

enables the the timely quarantine of high rate infections and prevents their spreading [15-25]. But none of these studies aim to provide a solution for rational triage of patients in places where the medical resources are limited.

In light of this unmet need in efficient triage of COVID-19 cases, the study is sought to develop and validate a learning-based model that evaluates patients' priority of being admitted to hospital care due to their appearance or susceptibility toward severe COVID-19. The model, provided with a simple user interface, can efficiently assess the severity of COVID-19 and predict the disease progression, with high rates of accuracy. Our study is expected to have a prolonged social impact under the current circumstances, when the simple and practical model becomes accepted to assist clinicians in quick and efficient triage of COVID-19 patients. This study was approved by the Guizhou Provincial People's Hospital Ethics Committee.

Results

Clinical characteristics of COVID-19 cases

The patients cohorts enrolled in this study were comprised of 134 COVID-19 cases retrieved from World Health Organization (WHO) COVID-19 database [26] (Figure S1, Table S1) and 467 COVID-19 cases recruited from a multi-center dataset in China. Amongst the 601 patients, 25.4% of patients had developed severe disease on admission and 6.5% of patients were presented with non-severe diseases on admission but progressed toward severe disease after admission. The minimal, medium and maximal time from hospital admission to severe disease progression were less than 1 day, 5 days and 12 days, respectively. The prevalence of underlying comorbidities was

24.8%. Hypertension (11%) was the most common comorbidity, follow by endocrine diseases (6.7%). The medium age was 48 years. Fever was the most common initial symptom (63.1%), followed by cough (50.1%), fatigue (21.3%) and dyspnea (9.3%). Table 1 shows the baseline laboratory results obtained on or soon after admission. All the patients were laboratory-confirmed COVID-19 cases and the severity of COVID-19 were stratified into severe and non-severe categories according to a criteria shown in Table 2.

Identification of Critical Variables for Model Establishment

The clinical variables of most patients were measured multiple times across different days during hospitalization to assess the prognosis. As this study was sought to identify the hospitalization priority according to the prehospital assessment of severe COVID-19 risk, only clinical data obtained on admission were used to evaluate the importance of clinical variables in identification of severe or potentially severe cases. Given the various missing data on different clinical variables and different patients, a strategy was adopted to set a threshold value alpha to remove these missing data, minimizing their impact on data analysis. It was found that as the threshold alpha increases, available variables decrease while available observations, i.e. the available COVID-19 cases, increase (Figure S2). A threshold alpha of 350 was selected to remove the missing data, and to obtain as many clinical variables and observations as possible. Hence, a total of 29 clinical variables and 214 patients with non-missing variable values were used for subsequent analysis.

Next, an univariate analysis was performed to investigate the difference in the 29 clinical variables between the severe and the non-severe groups. As shown in Table S2, a total of 12 clinical variables were significantly different between the two groups, including the age, fever, dyspnea, lymphocyte, neutrophil, C-reactive protein (CRP), lactic dehydrogenase (LDH), creatine kinase (CK), D-dimer, alanine aminotransferase (ALT), aspartate aminotransferase (AST) and albumin. These 12 clinical variables could be used to discriminate between the severe and non-severe COVID-19 cases.

Development of Risk Assessment Models

Extreme Gradient Boosting (XGBoost), which is a high-performance machine learning algorithm and works with a sequence of decision trees where the latter tree tries to minimize the net error from prior trees, was used to generate the risk assessment model. In addition to XGBoost classifier, other representative algorithms, including Linear Discriminant Analysis (LDA), Logistic Regression, Support Vector Machine (SVM), Random Forest and Decision Tree were benchmarked as the baselines.

A sample of 65 cases randomly chosen from the complete 214 cases served as a holdout testing set (Figure S3). The remaining 149 cases were used for training and cross validation. Then, the 12 significant variables in the previous univariant analysis were included to construct the risk assessment models based on XGBoost classifier as well as other classifiers. They were firstly trained in the training set of 149 cases and then evaluated in the holdout testing set of 65 cases, by comparing the values of accuracy, F1 score, sensitivity, specificity, and the area under curve (AUC) score of receiver operating characteristic (ROC) curve. The definition of these evaluation metrics

was showed in Table S3. The 12-variable XGBoost classifier proved an accuracy of 89.2% in discriminating severe COVID-19 cases from their non-severe counterparts (Table 3). Moreover, it outperformed other classifiers in the aforelisted evaluations, excepting the specificity (Table 3, and Figure 1). Our study was in agreement with a reported conclusion that the XGBoost algorithm had high discriminative performance [9], and thus could be used to assess the hospitalization priority with precision.

Next, the assembly of variables was minimized to ease the clinical use. For this purpose, a sequential variable selection approach was used to find the optimal variable set based on its assessment performance. Briefly, important variables ranked by XGBoost (Figure 2) were sequentially assembled in an individualized manner to investigate their incremental effects in terms of AUC scores by cross validation. The AUC scores ceased to grow when the count of assembled variables increased to 4 (Figure 3). Thus the previous 12-variable models shrinked to the selected 4-variable models, where XGBoost classifier achieved an accuracy of 84.6% in the identification of severe COVID-19 cases. Table 4 compares the performance of various classifiers in the 4-variable model. The AUC score of XGBoost was slightly decreased compared with others models, but XGBoost achieved the highest F1 score and accuracy among the classifiers (Table 4, Figure 4). An over 80% accuracy indicated that the 4-variable XGBoost model could play a crucial role in distinguishing the majority of cases that require immediate medical attention. Overall, the 4-variable XGBoost model was evaluated to be the most competitive and easy-to-use establishment throughout comparison with other prevalent choices.

Finally, the study was performed to investigate whether the 4-variable XGBoost model was effective to predict the risk of deterioration for patients who were presented with non-severe symptoms on admission. For this purpose, a total of 39 patients who had non-severe COVID-19 on admission but experienced deterioration during hospitalization were enrolled as an external testing set for analysis (Figure S3). The 4-variable XGBoost model achieved 100% accuracy in predicting the risk of rapid deterioration (Table 5). For 17 patients who had complete time course of exacerbation, the minimal, medium and maximal prediction horizon were less than 1 day , 5 days and 12 days, respectively, suggesting that our model could predict the risk of disease deterioration, for as long as 12 days earlier than its occurrence.

To test whether a clinical operable single-tree XGBoost classifier based on the lymphocyte count, CRP level and LDH level as reported by Li et al. [9] was able to accurately identify the risk of severe disease on admission, we performed the single-tree XGBoost in identification of severe COVID-19 cases as well as in prediction of risk of in-hospital deterioration from non-severe to severe disease. Table S4 shows the single-tree XGBoost had 80% accuracy in identification of severe COVID-19 cases on admission, but only 38.5% accuracy in prediction of risk of in-hospital deterioration. It suggested that a model established for other purposes or reported in other works does not fulfill our goal in this study, that is, identifying hospitalization priority for COVID-19 infections.

Collectively, the 4-variable XGBoost model is the first computation model established to assess hospitalization priority that enables rational triage of infected patients and prioritize hospitalization to the most needed.

166

167 **Model interpretation**

168 SHapley Additive exPlanations (SHAP), as a game-theoretic approach that interpreted
169 an impact of each input variable toward the model output, had been relied upon for the
170 model interpretation. In Figure S4, each dot corresponds to an individual case in the
171 study. Different colors encoded different values of input variables, while the SHAP value
172 represented the impact of each variable on the prediction outcome. As shown, the risk
173 of severe COVID-19 was found associated with a decrease in the lymphocyte count,
174 and an increase in the LDH level, the CRP level and the neutrophil count.
175 Next, the t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm as a technique
176 for dimensionality reduction was used to project the four-dimensional data (lymphocyte,
177 LDH, CRP, neutrophil) into a 3-dimension (3D) feature space for visualization [27]. It
178 enabled to visualize the difference in features among the three groups of patients,
179 including the severe cases, the non-severe cases and the progressed severe cases.
180 The progressed severe cases referred to patients with non-severe disease on
181 admission but developed severe disease afterward. There was a clear separation
182 between the non-severe (distributed in the core) and severe cases (distributed in the
183 periphery), whereas the progressed severe cases distributed in between the core and
184 the periphery (Figure S5, Video S1).

185

186 **Discussion**

187 The global epidemic of COVID-19 is increasingly stressing the healthcare systems in
188 many countries and territories as no effective vaccines have been approved to protect

the general population from this highly contagious disease [4,5]. Therefore, it requires the decision makers to efficiently distribute the medical resources by their needs in places where healthcare resources are rare.

To date, many studies have been reported to detect COVID-19 in patients with suspected infection [8-15]. For instance, Menni et al. reported the efficiency of self-reported symptoms in early prediction of potential COVID-19 [16]. It enables the medical providers to timely admit all potential COVID-19 cases to hospitals in countries and territories where the medical resources are relatively sufficient such as China [5]. But in places where the medical resources are limited for, all the infected patients, identifying patients at higher risk of severe COVID-19 and prioritizing them to hospitalization are paramount. Recent studies that assess the outcome of COVID-19 patients does not allow the distribution priority of high-risk infections to hospital admission. For instance, a machine learning model developed by Li et al. [9] and a big data analytics performed by Williamson et al. [8] focused on early identification of COVID-19 mortality during hospital admission, but rather the identification of severe diseases timely or in prediction. The poor performance of the mortality identification model in assessment of severity suggests that the previously reported prediction models are not suitable for identification of the hospitalization priority for the severe or potential severe COVID-19 cases. (Table S4).

More recently, a nomogram developed by Gong et al. can assist the early identification of severe COVID-19 cases with a sensitivity of 77.5% and a specificity of 78.4% [10], this study was prematurely accomplished before all the participants had fully experienced the outcome of event. Therefore, our study becomes the first to focus on the

development of an easy-to-use model that provides an insightful solution for doctors to triage patients, achieving a relatively high sensitivity and specificity, and meanwhile, without negatively impacting its performance by involving participants still in treatment. In this study, the clinical features of COVID-19 were screened to identify a total of 12 critical variables that were found to be associated with the risk of severe COVID-19 (Figure 2). To ease the clinical use, a model comprising 4 clinical variables was established. The 4-variable XGBoost model is effective in identification of nearly 85% of severe cases that require immediate medical attention (Table 4). Importantly, it precisely predicts the risk of progression toward rapid deterioration for as long as 12 days ahead of its occurrence (Table 5).

Our study is in line with the previous studies that showed the increased inflammatory responses, as evidenced by the increased neutrophil count and the raised CRP level, and impaired antiviral capacity, as evidenced by lymphopenia, were associated with severe diseases [11-13,28-31]. High LDH level associated with tissue breakdown in various diseases also indicates a severe disease [9].

This machine learning-based model is expected to be of clinical use in the context of this COVID-19 pandemic. The 4 indices included in the model are easily accessible even in rural and less developed places; evaluation can be rapidly proceeded by using the online program with a user-friendly interface (Figure S6; Link: covid-19.zyh.science:8888). The model has general applicability, as the training and testing datasets are retrieved from the datasets of different hospitals in different cities in China as well as the WHO database with participants experienced the event of interest and possessed enriched diversity and popularity.

There are some limitations in the present study. First, this is a respective study with a relative small sample size. Prospective study should be performed to validate our findings. Second, the datasets of different resources suffer various ratios of missing values that may impact the data analysis. The principle of cases and variables inclusion is a trade-off between the number of cases and the number of included variables (Figure S2). In this way, the impact of missing values on data analysis can be reduced. Third, there was no significant difference in the prevalence of comorbidies between the severe and the non-severe COVID-19 cases in our datasets (Table S2). This is probably because the number of patients with comorbidities is not enough to identify any difference in this factor between the two groups. It is unclear if the inclusion of comorbidy or other potential risk factors can improve the performance of model. But most recently, researches do not find the critical roles of comorbidy in improving the accuracy of model for prediction of COVID-19 prognosis [9,10]. But as more data become available, we can easily update our findings and generate a more accurate model with the same procedure.

In summary, we developed this machine learning-based model by holding the wish to contribute to the healthcare management on this pandemic occasion. It is practically applicable for the healthcare decision makers and professionals to efficiently distribute the infections and allocate inpatient cares to the most needed, and contribute this global battle against the spread of this coronavirus.

Experimental procedures

Resource Availability

Lead contact:

Shaohua Ma; ma.shaohua@sz.tsinghua.edu.cn.

Material Availability:

No new unique reagents were generated in the present study.

Data and Code Availability

The clinical data used in this study were obtained from the WHO COVID-19 database (https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov), Guizhou Provincial People's Hospital, Affiliated Hospital of Zunyi Medical University, Jiangjunshan Hospital of Guizhou Province, Zhongnan Hospital of Wuhan University, and Radiology Quality Control Center database of Hunan province. The code is available at GitHub (https://github.com/cow8/Covid-19_Severity). Additional Supplemental Items are available from Mendeley Data at http://dx.doi.org/10.17632/mmy4v3gkxb.3

Method Details

Patient enrollment

Data on COVID-19 cases were retrieved from WHO COVID-19 database [26], Radiology Quality Control Center database of Hunan province, and the datasets of four major hospitals (Guizhou Provincial People's Hospital, Affiliated Hospital of Zunyi Medical University, Jiangjunshan Hospital of Guizhou Province, Zhongnan Hospital of Wuhan University), respectively.

A strategy to retrieve studies that reported the individual data on clinical characteristics, clinical types and prognosis of COVID-19 cases is shown in the Figure S1. Briefly, the

publication list was first screened to exclude reviews, opinions, guidelines, corrections, the epidemiological and pharmacological studies, and other nonclinical or irrelevant investigations, and identify clinical studies reporting both the clinical characteristics and patient outcomes. To enroll COVID-19 cases from the multi-center dataset in China, the medical records were examined to identify patients who were admitted to hospitals for SARS-CoV-2 infection from January to March, 2020.

All of the patients recruited to this study were laboratory-confirmed COVID-19 cases. Patients who were pregnant or younger than 18 were excluded from this study. The demographic information, clinical characteristics, laboratory findings, and prognosis of patients were extracted from these datasets. By the time of data collection, all patients have experienced the outcome of event (e.g. death, recovery or discharge).

Disease Assessment

All the patients enrolled in this study were tested for SARS-COV-2 before or after being admitted to the hospitals. Oropharyngeal swab, nasopharyngeal swab, sputum, serum, faeces, endotracheal aspirate or bronchoalveolar lavage were collected from each patient to detect the SARS-CoV-2 ribonucleic acid (RNA) or anti-SARS-COV-2 IgM/IgG if feasible. The COVID-19 cases were confirmed upon the detection of unique sequences of SARS-CoV-2 RNA by real-time reverse transcription polymerase chain reaction (RT-PCR) assay or viral genome sequencing, or anti-SARS-COV-2 IgM/IgG in the paired serum specimens, following the clinical guidelines [28,32]. All the patients enrolled to the study were also assessed for disease severity on admission and were repeatedly evaluated for progression during hospitalization. The severity of COVID-19

was stratified into the non-severe and severe categories using the criteria which were published by WHO and National Health Commission of China [28,31], with minor modifications (Table 2). Specifically, a severe case of COVID-19 was defined by the presence of any of the following conditions in the quiescent state, such as an increased respiratory rate of ≥ 30 breaths/minute, decreased oxygenation index ≤ 300 mmHg or declined SpO2 of ≤ 93%. Moreover, patients who developed shock, multiple organ failure (MOF) that were required to be admitted to Intensive Care Unit (ICU), or respiratory failure that warranted mechanical ventilation were stratified into the severe category in the present study. Finally, patients with pulmonary lesions that showed rapid progression of over 50% within 24-48 hours were considered to have severe disease.

Critical Variables Inclusion

Clinical variables that were widely available in clinic and were previously demonstrated to be closely associated with the severity of COVID-19 were obtained for data analysis in this study. They include but are not limited to the age, gender, underlying comorbidity, symptoms (i.e., dyspnea), peripheral capillary oxygen saturation, the hematological and biochemical parameters of lymphocyte, neutrophil, CRP, procalcitonin (PCT), D-dimer, erythrocyte sedimentation rate, ALT, AST, bilirubin, albumin, LDH, serum creatinine, blood urea nitrogen (BUN), prothrombin time (PT), lactic acid, CK, and SpO2 [10-13,28-31,33-35]. All these hematological and biochemical variables were detected by using standard automated laboratory methods or commercially available kits following manufacturers' protocols. All the clinical variables of patietns were determined on admission or soon after hospitalization. The critical variables chosen to establish the model were easily

accessible measures by practitioners from a majority of countries and territories, and were common to all databases across different institutions.

Statically Analysis

Data analysis and visualization were implemented using Python.

Categorical variables were described as number and frequency, while the continuous variables were expressed as mean, standard deviation (SD), interquartile range (IQR) where appropriate.

As there were various missing data in the datasets, a threshold alpha was set to remove to these missing data (Figure S2). First, the clinical variables with missing data which exceeded the threshold alpha were removed (bottom figure). Second, the observations (that were COVID-19 cases) with missing data for any of the resulting clinical variables were removed. An optimal threshold alpha should be selected to obtain as many clinical variables and observations as possible.

Next, COVID-19 cases with non-missing variable values were grouped into the severe and non-severe categories, according to the severity of disease. The difference in the clinical variables between the two groups was identified via the univariate descriptive statistics (Table S2). A p-value of less than 0.05 was considered statistically significant and was used as a threshold to identify key clinical variables for model development. These key clinical variables were assembled to generate the risk assessment models based on the XGBoost classifier as well as other classifiers, such as the LDA, Logistic Regression, SVM, Random Forest, and Decision Tree.

Then, the COVID-19 cases were randomly split into the training set and holdout testing set at a ratio of 7:3. The different models were trained in the training set and evaluated in the holdout testing set by comparing the values of accuracy, F1 score, sensitivity, specificity, and AUC score of ROC curve (Table S3).

Subsequently, the assembly of key clinical variables was minimized to generate the simplified models. For this purpose, all the key clinical variables were ranked according to the importance calculated by XGBoost (Figure 2), followed by the sequential variable selection approach (Figure 3). This was to minimize the variable set while optimize the model performance. The simplified models based on the minimized variable set were trained and evaluated in accordance with a method mentioned above. To assess the effectiveness of models in early prediction of severe progressions, patients who were presented with non-severe symptom on admission but developed severe disease during hospitalization were enrolled as an external testing set for analysis. The performance of models was reflected by accuracy (Table 5). As for comparison, the performance of a previously validated single-tree XGBoost model [9] in identification of severe COVID-19 risk was assessed in our datasets (Table S4).

Finally, the SHAP values were plotted to visualize the impact of each input variable toward the model output and the t-SNE algorithm was used for dimensionality reduction that enabled to visualize the difference in features among different groups of patients.

Acknowledgements

We thank the above-mentioned cooperating hospitals for kindly sharing the data with us, in accordance with the Declaration of Helsinki. The work was supported by the national

natural science foundation of China (Grant Number: 61722209, 61971255), the fund from the Shenzhen Science and Technology Innovation Committee (Grant Number: KQJSCX20180327143623167), and the Guizhou Science and Technology Project (Grant Number: QKHZC[2020]4Y002).

Author Contributions

S.M. and L.F. conceived this project. YC.Z, YH.Z., and M.J. designed and performed the experiments and data analysis. YC.Z and YH.Z innovated and implemented the algorithm. R.W., X.L., and M.Z. contributed the data. S.M., L.F., and R.W critically reviewed manuscript. All co-authors contributed to the manuscript writing.

Competing Interests Statement

The authors declare no competing interests.

References

1 Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R. et al. (2020). A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med 382, 727-733, doi:10.1056/NEJMoa2001017.

2 Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., Hu, Y., Tao, Z. W., Tian, J. H., Pei, Y. Y. et al. (2020). A new coronavirus associated with human respiratory disease in China. Nature 579, 265-269, doi:10.1038/s41586-020-2008-3.

3 Coronavirus disease 2019 (COVID-19) Situation Report – 189, (2020) <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports>.

4 Yang, P. & Wang, X. (2020). COVID-19: a new challenge for human beings. Cell Mol Immunol 17, 555-557, doi:10.1038/s41423-020-0407-x.

5 Li, Z., Chen, Q., Feng, L., Rodewald, L., Xia, Y., Yu, H., Zhang, R., An, Z., Yin, W., Chen, W. et al. (2020). Active case finding with case management: the key to tackling the COVID-19 pandemic. Lancet, doi:10.1016/S0140-6736(20)31278-2.

6 Infection prevention and control in the household management of people with suspected or confirmed coronavirus disease (COVID-19), (2020). <https://www.ecdc.europa.eu/sites/default/files/documents/Home-care-of-COVID-19-patients-2020-03-31.pdf>.

7 Sun, Q., Qiu, H., Huang, M. & Yang, Y. (2020). Lower mortality of COVID-19 by early recognition and intervention: experience from Jiangsu Province. Ann Intensive Care 10, 33, doi:10.1186/s13613-020-00650-2.

8 Williamson, E., Walker, A. J., Bhaskaran, K. J., Bacon, S., Bates, C., Morton, C. E., Curtis, H. J., Mehrkar, A., Evans, D., Inglesby, P. et al. (2020). OpenSAFELY: factors associated with COVID-19-related hospital death in the linked electronic health records of 17 million adult NHS patients. medRxiv, 2020.2005.2006.20092999, doi:10.1101/2020.05.06.20092999.

9 Yan, L., Zhang, H.-T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., Sun, C., Tang, X., Jing, L., Zhang, M. et al. (2020). An interpretable mortality prediction model for COVID-19 patients. Nature Machine Intelligence 2, 283-288, doi:10.1038/s42256-020-0180-7.

10 Gong, J., Ou, J., Qiu, X., Jie, Y., Chen, Y., Yuan, L., Cao, J., Tan, M., Xu, W., Zheng, F. et al. (2020). A Tool to Early Predict Severe Corona Virus Disease 2019 (COVID-19) : A Multicenter Study using the Risk Nomogram in Wuhan and Guangdong, China. Clin Infect Dis, doi:10.1093/cid/ciaa443.

11 Tan, L., Wang, Q., Zhang, D., Ding, J., Huang, Q., Tang, Y. Q., Wang, Q. & Miao, H. (2020). Lymphopenia predicts disease severity of COVID-19: a descriptive and predictive study. Signal Transduct Target Ther 5, 33, doi:10.1038/s41392-020-0148-4.

12 Shi, Y., Yu, X., Zhao, H., Wang, H., Zhao, R. & Sheng, J. (2020). Host susceptibility to severe COVID-19 and establishment of a host risk score: findings of 487 cases outside Wuhan. Crit Care 24, 108, doi:10.1186/s13054-020-2833-7.

13 Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y., Song, B., Gu, X. et al. (2020). Clinical course and risk factors for mortality of adult

inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet 395, 1054-1062, doi:10.1016/S0140-6736(20)30566-3.

14 Yuan, M., Yin, W., Tao, Z., Tan, W. & Hu, Y. (2020). Association of radiologic findings with mortality of patients infected with 2019 novel coronavirus in Wuhan, China. PLoS One 15, e0230548, doi:10.1371/journal.pone.0230548.

15 Wynants, L., Van Calster, B., Bonten, M. M. J., Collins, G. S., Debray, T. P. A., De Vos, M., Haller, M. C., Heinze, G., Moons, K. G. M., Riley, R. D. et al. (2020). Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. BMJ 369, m1328, doi:10.1136/bmj.m1328.

16 Menni, C., Valdes, A. M., Freidin, M. B., Sudre, C. H., Nguyen, L. H., Drew, D. A., Ganesh, S., Varsavsky, T., Cardoso, M. J., El-Sayed Moustafa, J. S. et al. (2020). Real-time tracking of self-reported symptoms to predict potential COVID-19. Nat Med, doi:10.1038/s41591-020-0916-2.

17 Xie, X., Zhong, Z., Zhao, W., Zheng, C., Wang, F. & Liu, J. (2020). Chest CT for Typical 2019-nCoV Pneumonia: Relationship to Negative RT-PCR Testing. Radiology, 200343, doi:10.1148/radiol.2020200343.

18 Diaz-Quijano, F. A., Silva, J. M. N. d., Ganem, F., Oliveira, S., Vesga-Varela, A. L. & Croda, J. (2020). A model to predict SARS-CoV-2 infection based on the first three-month surveillance data in Brazil. medRxiv, 2020.2004.2005.20047944, doi:10.1101/2020.04.05.20047944.

19 Chen, J., Wu, L., Zhang, J., Zhang, L., Gong, D., Zhao, Y., Hu, S., Wang, Y., Hu, X., Zheng, B. et al. (2020). Deep learning-based model for detecting 2019 novel

coronavirus pneumonia on high-resolution computed tomography: a prospective study. medRxiv, 2020.2002.2025.20021568, doi:10.1101/2020.02.25.20021568.

20 Meng, Z., Wang, M., Song, H., Guo, S., Zhou, Y., Li, W., Zhou, Y., Li, M., Song, X., Zhou, Y. et al. (2020). Development and utilization of an intelligent application for aiding COVID-19 diagnosis. medRxiv, 2020.2003.2018.20035816, doi:10.1101/2020.03.18.20035816.

21 Song, C.-Y., Xu, J., He, J.-Q. & Lu, Y.-Q. (2020). COVID-19 early warning score: a multi-parameter screening tool to identify highly suspected patients. medRxiv, 2020.2003.2005.20031906, doi:10.1101/2020.03.05.20031906.

22 Martin, A., Nateqi, J., Gruarin, S., Munsch, N., Abdarahmane, I. & Knapp, B. (2020). An artificial intelligence-based first-line defence against COVID-19: digitally screening citizens for risks via a chatbot. bioRxiv, 2020.2003.2025.008805, doi:10.1101/2020.03.25.008805.

23 Wang, Z., Weng, J., Li, Z., Hou, R., Zhou, L., Ye, H., Chen, Y., Yang, T., Chen, D., Wang, L. et al. (2020). Development and Validation of a Diagnostic Nomogram to Predict COVID-19 Pneumonia. medRxiv, 2020.2004.2003.20052068, doi:10.1101/2020.04.03.20052068.

24 Wu, J., Zhang, P., Zhang, L., Meng, W., Li, J., Tong, C., Li, Y., Cai, J., Yang, Z., Zhu, J. et al. (2020). Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results. medRxiv, 2020.2004.2002.20051136, doi:10.1101/2020.04.02.20051136.

25 Mao, X., Liu, X.-P., Xiong, M., Yang, X., Jin, X., Li, Z., Zhou, S. & Chang, H. (2020). Development and validation of chest CT-based imaging biomarkers for

early stage COVID-19 screening. medRxiv, 2020.2005.2015.20103473, doi:10.1101/2020.05.15.20103473.

26 Global research on coronavirus disease (COVID-19), (2020) <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov>.

27 Maaten, L. v. d. & Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research 9, 2579-2605.

28 Diagnosis and treatment of novel coronavirus pneumonia (trial version 7), (2020). <http://www.gov.cn/zhengce/zhengceku/2020-03/04/content_5486705.htm>.

29 Guan, W. J., Ni, Z. Y., Hu, Y., Liang, W. H., Ou, C. Q., He, J. X., Liu, L., Shan, H., Lei, C. L., Hui, D. S. C. et al. (2020). Clinical Characteristics of Coronavirus Disease 2019 in China. N Engl J Med, doi:10.1056/NEJMoa2002032.

30 Ai, J., Li, Y., Zhou, X. & Zhang, W. (2020). COVID-19: treating and managing severe cases. Cell Res, doi:10.1038/s41422-020-0329-2.

31 Clinical management of severe acute respiratory infection when COVID-19 is suspected: Interim guidance V 1.2, (2020). <https://www.who.int/publications-detail/clinical-management-of-severe-acute-respiratory-infection-when-novel-coronavirus-(ncov)-infection-is-suspected>.

32 Laboratory testing for coronavirus disease 2019 (COVID-19) in suspected human cases: interim guidance, (2020). <https://apps.who.int/iris/handle/10665/331329>.

33 Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., Wang, B., Xiang, H., Cheng, Z., Xiong, Y. et al. (2020). Clinical Characteristics of 138 Hospitalized Patients

With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. JAMA, doi:10.1001/jama.2020.1585.

34 Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X. et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395, 497-506, doi:10.1016/S0140-6736(20)30183-5.

35 Cao, W. & Li, T. (2020). COVID-19: towards understanding of pathogenesis. Cell Res, doi:10.1038/s41422-020-0327-4.

Figure titles and legends

Figure 1. Receiver operating characteristic (ROC) curve for the performance of 12-variable models in discriminating the severe COVID-19 cases.

The 12 variables included age, fever, dyspnea, lymphocyte, neutrophil, C-reactive protein, lactic dehydrogenase, creatine kinase, D-dimer, alanine aminotransferase, aspartate aminotransferase and albumin.

Figure 2. Top key clinical variables that are ranked according to their importance in the Multi-tree XGBoost algorithm.

Abbreviations: L, lymphocyte; LDH, lactic dehydrogenase; CRP, C-reactive protein; N, neutrophil; ALT, alanine aminotransferase; D, D-dimer; ALB, albumin; CK, creatine kinase; AST, aspartate aminotransferase.

Figure 3. Important variables ranked by XGBoost algorithm were sequentially aseembled to investigate their incremental effects on the model performance.

Abbreviation: AUC, area under the receiver operating characteristic curve.

Figure 4. Receiver operating characteristic (ROC) curve for the performance of 4-variable models in discriminating the severe COVID-19 cases.

The 4 variables were lymphocyte, lactic dehydrogenase, C-reactive protein and neutrophil.

Video S1. t-SNE visualization

530 Table 1. Baselince characteristics of patients enrolled to the study

531

Clinical features	Overall
Age, mean ± std, years	47.63 ± 15.6
Gender, n(%)	
Male	319 (53%)
Female	280 (47%)
Underlying comorbidities, n(%)	149 (24.8%)
Hypertension	67(11%)
Endocrine disease	40(6.7%)
Cardiovascular disease	18(3%)
Chronic Lung disease	9(1.5%)
Digestive Disease	20(3.3%)
Renal disease	7(1.2%)
Tumor	3(0.5%)
Cerebrovascular/Nervous Disease	7(1.2%)
Immune disorder	8(1.3%)
Others	35(5.8%)
Signs or symptoms, n(%)	
Fever	379 (63.1%)
Cough	301 (50.1%)
Expectoration	6 (1.0%)
Hemoptysis	2 (0.3%)
Dyspnea	56 (9.3%)
Catarrh	9 (1.5%)
Fatigue	128 (21.3%)
Anorexia	3 (0.5%)
Nausea/Emesis	14 (2.3%)
Myalgia	47 (7.8%)
Dizziness/Headache	37 (6.2%)
Pharyngalgia	9 (1.5%)
Abdominal pain/diarrhea	7 (0.2%)
Laboratory findings, mean± std	
White blood cell count, 10^9/L	6.02 ± 15.98
Lymphocyte count, 10^9/L	1.25 ± 1.22
Neutrophil count, 10^9/L	3.64 ± 2.73
Erythrocyte sedimentation rate, mm/h	43.75 ± 28.86
C-reactive protein, mg/L	25.24 ± 28.92
Procalcitonin, ng/ml	0.91 ± 7.58
D-dimer, ug/ml	68.26 ± 515.42
Alanine aminotransferase, U/L	25.54 ± 15.47
Aspartate aminotransferase, U/L	29.5 ± 16.15
Total bilirubin, umol/l	13.56 ± 8.09
Albumin, g/L	39.75 ± 9.28
Lactate dehydrogenase, U/L	229.5 ± 118.41
Blood urea nitrogen, mmol/L	5.29 ± 3.34
Serum creatinine, μmol/L	62.52 ± 34.39
Prothrombin time, s	13.27 ± 3.34
Lactic acid	271.37 ± 359.03
Creatine kinase, U/L	110.23 ± 118.38
SpO_2, %	93.93 ± 14.69

532 Table 2. Classification of the COVID-19 severity.

Classifications	Definitions
Non-severe COVID-19	Patients have non-specific symptoms such as fever, cough, fatigue, myalgia, pharyngalgia, but have no signs of dehydration, sepsis or shortness of breath. The radiological examination showes no signs of severe pneumonia.
Severe COVID-19	Adult cases meeting any of the following criteria: (1) Respiratory rate $\geqq$ 30 breaths/ min; (2) Oxygen saturation $\leqslant$ 93% at rest; (3) FiO2 $\leqq$ 300mmHg. (4) Pulmonary lesion progression exceeds 50% in 24-48 hours (5) Respiratory failure that requires mechanical ventilation; (6) Shock; (7) Organ failure that requires to be managed in Intensive Care Unit

533 Table 3. The performance of 12-variable models for identification of severe COVID-19

534 on admission

	LDA	Logistic Regression	Random Forest	Decision Tree	SVM	XGBoost
AUC macro	0.929	0.917	0.903	0.676	-	0.953
F1 weighted	0.891	0.854	0.848	0.769	0.848	0.896
Accuracy	0.892	0.862	0.862	0.800	0.862	0.892
Sensitivity	0.692	0.538	0.462	0.231	0.462	0.846
Specifity	0.942	0.942	0.962	0.942	0.962	0.904

535
536

537 Table 4. The performance of 4-variable models for identification of severe COVID-19 on
538 admission

	LDA	Logistic Regression	Random Forest	Decision Tree	SVM	XGBoost
AUC macro	0.876	0.879	0.864	0.680	-	0.859
F1 weighted	0.815	0.802	0.815	0.769	0.815	0.856
Accuracy	0.831	0.815	0.831	0.800	0.831	0.846
Sensitivity	0.385	0.385	0.385	0.231	0.385	0.846
Specifity	0.942	0.923	0.942	0.942	0.942	0.846

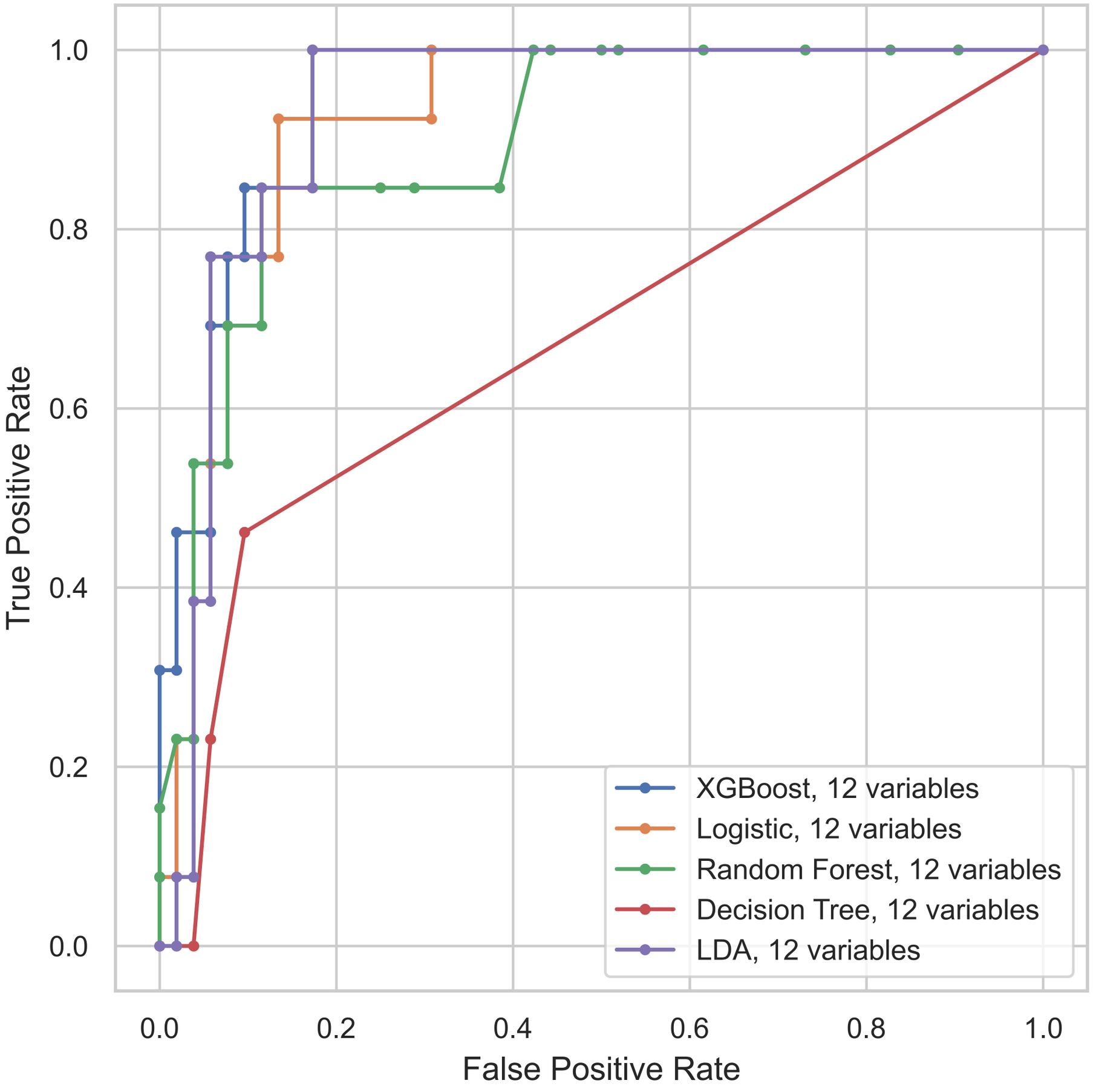
539
540

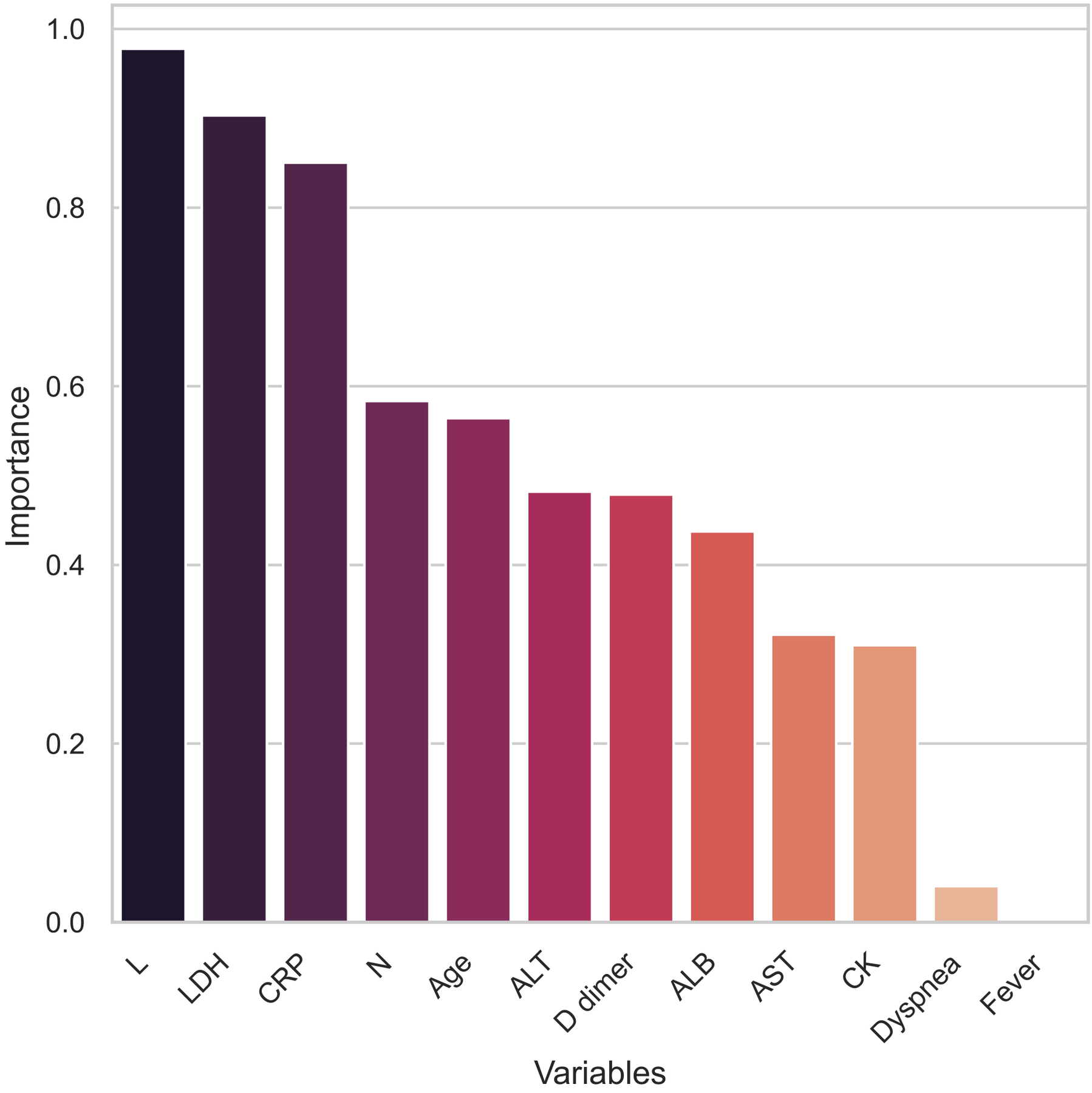
541 Table 5. The performance of the 4-variable XGBoost model for prediction of COVID-19

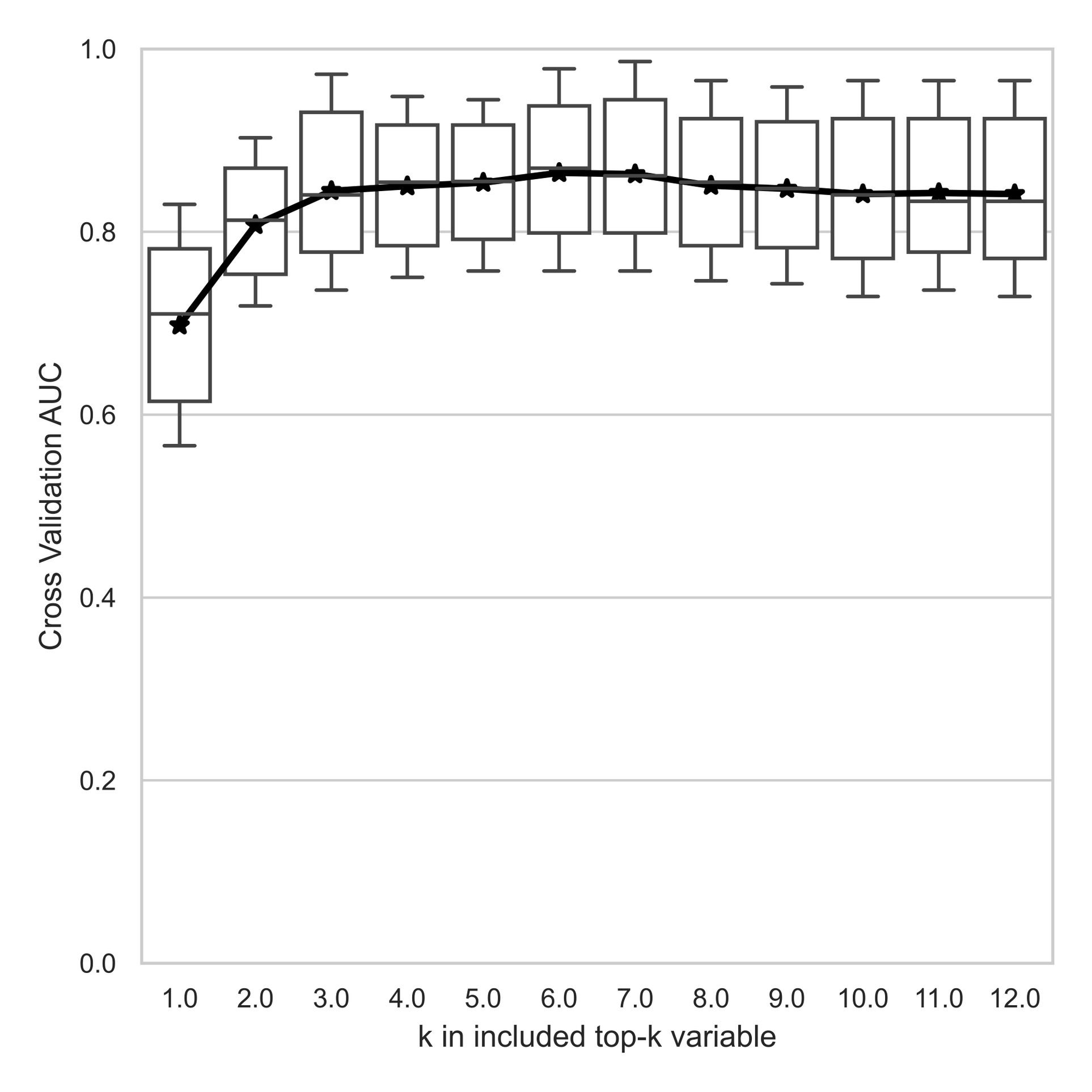
	LDA	Logistic Regression	Random Forest	Decision Tree	SVM	XGBoost
Accuracy	1.000	1.000	0.974	0.974	1.000	1.000

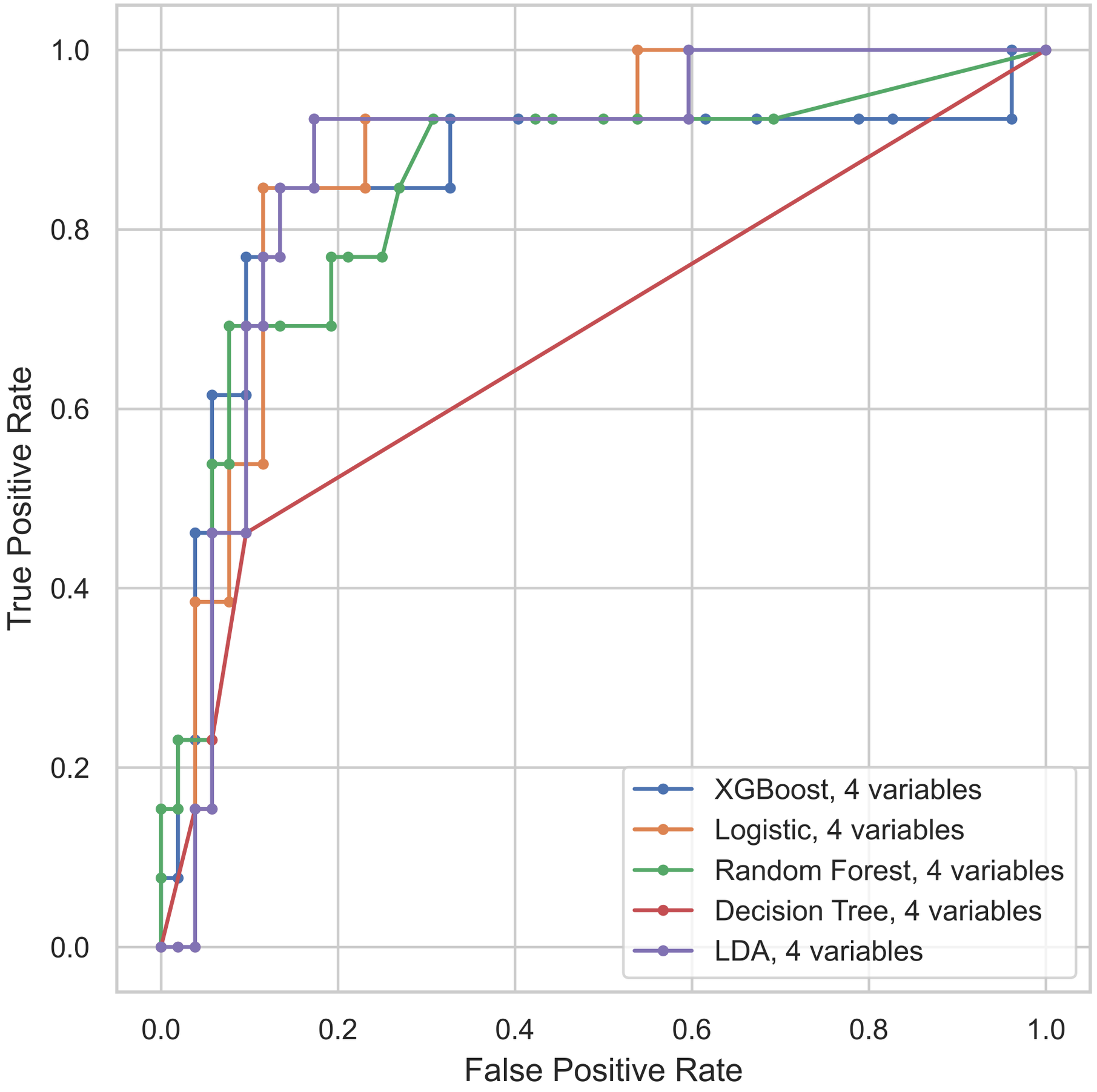
542 deterioration

543


True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
XGBoost, 12 variables
Logistic, 12 variables
Random Forest, 12 variables
Decision Tree, 12 variables
LDA, 12 variables


1.0
0.8
0.6
0.4
0.2
0.0
Importance
L
LDH
CRP
N
Age
ALT
D dimer
ALB
AST
CK
Dyspnea
Fever
Variables


1.0
0.8
0.6
0.4
0.2
0.0
Cross Validation AUC
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
k in included top-k variable


True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
XGBoost, 4 variables
Logistic, 4 variables
Random Forest, 4 variables
Decision Tree, 4 variables
LDA, 4 variables

Highlights:

1 A model was developed to evaluate hospitalization priority in COVID-19 pandemics.
2 This model used easily accessible biomarkers to evaluate the risk of severe COVID-19.
3 The evaluation can be rapidly proceeded using an online program.
4 Performance of different algorithms in evaluation of COVID-19 severity was explored.

eTOC blurb:

The authors proposed a learning-based model to assist clinicians in quick and efficient triage of patients in places where the medical resources are limited in COVID-19 pandemics. This model used four easily accessible biomarkers to assess the severity of COVID-19, and was found effective to identify the risk of severe COVID-19. It will enable the healthcare administrators to distribute hospitalization resources to the most needed.

Bigger Picture:

COVID-19 pandemic is threatening millions of lives and stressing the medical systems worldwide. Though the infection growth in some areas has creased, the risk of second wave of outbreak is under threatening. So, a sustainable strategy to defend the pandemic using current limited but effective healthcare resources is in high demand. Our study is deemed to find a solution that triages patients to hospitalization by identifying their severity progression. In this study, a model that used four easily accessible biomarkers to assess the risk of severe COVID-19 was successfully developed. This model is easy to use and it eliminates the dependencies on exquisite equipment to make a decision. It was found effective to identify the risk of severe COVID-19. So, it is practically applicable for general practitioners to effectively distribute the infections and allocate in-patient cares to the most needed. Our study is expected to have a prolonged social impact under the current circumstances.