Received: 27 July 2021 | Revised: 2 September 2021 | Accepted: 16 September 2021

DOI: 10.1002/jmv.27352

RESEARCH ARTICLE

JOURNAL OF MEDICAL VIROLOGY WILEY

Using different machine learning models to classify patients into mild and severe cases of COVID-19 based on multivariate blood testing

Rui-kun Zhang[1] | Qi Xiao[1] | Sheng-lang Zhu[2] | Hai-yan Lin[2] | Ming Tang[3]

[1]Health Science Center, Shenzhen University, Shenzhen, China

[2]Department of nephrology, Shenzhen Nanshan People's Hospital and The 6th Affiliated Hospital of Shenzhen University Health Science Center, Shenzhen, China

[3]Department of Critical Care Medicine, Shenzhen Third People's Hospital, The Second Hospital Affiliated to Southern University of Science and Technology, Shenzhen, China

Correspondence
Ming Tang, Department of Critical Care Medicine, Shenzhen Third People's Hospital, The Second Hospital Affiliated to Southern University of Science and Technology, No 29 Bulan Rd, Shenzhen 518000, Guangdong, China.
Email: 273453706@qq.com

Abstract

COVID-19 is a serious respiratory disease. The ever-increasing number of cases is causing heavier loads on the health service system. Using 38 blood test indicators on the first day of admission for the 422 patients diagnosed with COVID-19 (from January 2020 to June 2021) to construct different machine learning (ML) models to classify patients into either mild or severe cases of COVID-19. All models show good performance in the classification between COVID-19 patients into mild and severe disease. The area under the curve (AUC) of the random forest model is 0.89, the AUC of the naive Bayes model is 0.90, the AUC of the support vector machine model is 0.86, and the AUC of the KNN model is 0.78, the AUC of the Logistic regression model is 0.84, and the AUC of the artificial neural network model is 0.87, among which the naive Bayes model has the best performance. Different ML models can classify patients into mild and severe cases based on 38 blood test indicators taken on the first day of admission for patients diagnosed with COVID-19.

KEYWORDS
artificial intelligence, biostatistics & bioinformatics, coronavirus, infection, virus classification

1 | INTRODUCTION

Coronavirus disease 2019 (COVID-19) is an acute respiratory disease caused by the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2). COVID-19 pneumonia has gradually spread across the world since the first case was reported. As of July 12, 2021, 180 million people have been infected with COVID-19 and 4 million have perished. Compared with the severe acute respiratory syndrome (SARS) caused by the severe acute respiratory syndrome coronavirus (SARS-CoV), COVID-19 has a relatively lower mortality rate,[1] and most COVID-19 patients are mild cases. The clinical manifestations of patients with disease are mainly fever, cough, fatigue, and some other symptoms.[2,3] However, the condition of mild case patients may also deteriorate and develop into severe patients. A study from China shows that among the 1099 confirmed COVID-19 patients, the rate of severely ill patients is 5% and they need to be transferred to the ICU for continued treatment.[4] In Lombardy, Italy, COVID-19 patients admitted to the ICU account for about 16% of the confirmed cases.[5] The mortality rate of critically ill patients can reach 49%.[6] A meta-analysis has shown that hypertension, coronary heart disease, and diabetes are associated with the said higher risk of death. Studies have also shown that C-reactive protein (CRP), cardiac troponin (cTn), and interleukin 6 (IL-6) are lab indicators associated with high mortality.[7,8]

The occurrence of critical COVID-19 cases and other critical situations accompanying COVID-19 has caused a great burden on the medical care system and created a huge challenge to curb the spread of COVID-19. Therefore, how to identify critically ill patients as early as possible to implement early interventions and

Rui-kun Zhang and Qi Xiao contributed equally to this study and should be considered as co-first authors.

predicting the occurrence of COVID-19 critical illness is an important direction for the treatment of COVID-19. Machine learning (ML) algorithms have been widely used in the diagnosis, prediction, and other fields of COVID-19 due to their powerful complex data processing capabilities.[9,10] A meta-analysis showed that different ML models (RIDGE regression, random forest, and LASSO regression) can effectively differentiate COVID-19 patients and type A influenza patients,[11] the study of Di Castelnuovo et al used the Random Forest model to determine the predictive factors of COVID-19 patient death in hospital,[7] Yue et al used logistic regression, support vector machine (SVM), gradient boosting decision tree and neural network, the ensemble model composed of four ML methods can accurately predict the death risk of COVID-19.[12] In previous studies, the clinical variable indicators used to identify patients with mild and severe cases were determined.[13] Due to blood tests being efficient, simple, low-cost, and fast, they have become an alternative plan for the early identification of COVID-19 patients.[14–16] Currently, there are studies that apply ML models to blood test indicators to diagnose COVID-19 patients.[17] However, there has been no report on differentiating COVID-19 mild cases and severe cases solely based on the blood test indicators. At the same time, the condition of severe and critically ill patients with COVID-19 develops rapidly, with multiple organ failures including respiratory failure. It is difficult to treat with a high fatality rate, and generally comes with poor prognosis. This has had a great impact on medical staff and the healthcare system. How to early identify and actively intervene in severe and critically ill patients is one of the keys to reduce the mortality rate of COVID-19 patients and improve their prognoses. Therefore, this study examines 422 COVID-19 patients diagnosed at Shenzhen Third People's Hospital based on the blood indicators collected on the first day of their diagnosis and admission, different ML models are then built around these data. Internal verifications are then conducted to classify COVID-19 patients with mild and severe cases of COVID-19 to compare the accuracy of each model's prediction.

2 | METHODS

2.1 | Data gathering

This study is a retrospective analysis. The subjects of the study are patients who were diagnosed with COVID-19, who were admitted, and hospitalized in Shenzhen Third People's Hospital from January 2020 to June 2021. All patients were examined due to either having close contact with patients or having fevers and other respiratory symptoms that are related to COVID-19. The standard of diagnosis is for real-time fluorescent reverse transcription-polymerase chain reaction (RT-PCR) to detect the positive nucleic acid of COVID-19, that is, the PCR method is used to detect the nucleic acid fragment of the coronavirus. A positive test is defined as the viral load being higher than the lower threshold of the COVID-19 nucleic acid detection reagent. It is confirmed that the patient is diagnosed with COVID-19 after a second review at the Shenzhen Centers for Disease Control and Prevention still turns out to be positive. All patient information is recorded in the electronic medical record system of Shenzhen Third People's Hospital. To protect patients' privacy, all patients have their personal identification information removed, and each patient only retains a unique randomly generated digital identity tag. This study extracted information on all COVID-19 cases recorded in the electronic medical journal system as of June 2021. The exclusion criterion are: less than 18 years old of age, admission time for less than 2 days, patients that are missing more than 75% of their data. We collected blood test indicators of patients on the first day of admission. All data were extracted from electronic medical records. A total of 38 indicator variables (lactic acid, potassium ion, sodium ion, white blood cell, neutrophil ratio, neutrophil count, Lymphocyte ratio, lymphocyte count, hematocrit, hemoglobin, platelets, alanine aminotransferase, aspartate aminotransferase, total bilirubin, direct bilirubin, albumin, globulin, lactate dehydrogenase, urea, creatinine, β−2M, cystostatin C, creatine kinase, creatine kinase isoenzyme, B-type natriuretic peptide, troponin I, killer T cells, helper T cells, suppressor T cells, erythrocyte sedimentation rate, C Reactive protein, procalcitonin, interleukin 6, prothrombin time, activated partial thromboplastin time, fibrin, D dimer, and PaO_2/FiO_2) were collected. This study was approved by the Ethics Committee of Shenzhen Third People Hospital.

2.2 | Data pre-processing

We combined the cases and all variables into a matrix and imported it into R (version 4.0.3); normalized and standardized continuous variables, and binarized the outcome variables into Booleans. After scaling, the missing values were then reduced to 0, to compensate for missing values, the method of interpolation was used. Commonly used interpolation methods include mean value interpolation, hot-deck interpolation, and multiple interpolation. Mean value interpolation is the simplest interpolation method and has been successfully applied For missing values in large data sets,[18] this study uses the k-nearest neighbor (KNN) mean interpolation algorithm to process the missing values. This method selects the K closest data values and calculates the average of the K data values as the estimate of the missing values. So, the Impute module was used to supplement the missing values of the data using the KNN algorithm.

2.3 | Machine learning analysis

This study chose seven ML models: random forest, naïve Bayes, SVM, KNN, logic regression, and artificial neural network. The data were randomly separated into two parts. We randomly divided the data into 80% training data set and 20% test data set. In the training data set, 38 blood indicator variables were used as input to construct the corresponding model. The model is used to classify patients with mild and severe COVID-19. The models are then put through a fivefold cross-validation process and grid

searches automatically optimize to determine the best hyperparameters. The hyperparameters that need to be adjusted include but is not limited to the max depth and *n*_estimators for the random forest model and learning rate for the artificial neural network model. Next, the model trained in the training set is used in the test data set for internal verification to evaluate the performance of the model in classifying patients with mild and severe COVID-19. Training and testing sets' baseline demographics are shown in Table S1. All ML is done through Python using numpy (v1.19), scikit-learn (v0.24), and tensorflow (v2.5). The process is shown in Figure 1.

2.4 | Model performance

We evaluated the accuracy of the model by drawing the receiver operating characteristic (ROC) curve and the precision–recall (PR) curve in the test data set, and used the following equations to calculate the sensitivity and specificity to evaluate the performance of the model:

$$Sensitivity = \frac{TP}{TP + FN} * 100$$

$$Specificty = \frac{TN}{TN + FP} * 100$$

Finally, the area under the ROC curve (AUC) of different models is compared. The 95% confidence interval (CI) of the AUC is then obtained by repeatedly resampling with replacement 1000 times in the test data set using bootstrapping procedure.

3 | RESULTS

3.1 | Variable characteristics of patients with mild and severe cases of the illness

Between January 2020 and June 2021, 493 patients were diagnosed with RT-PCR at Shenzhen Third People's Hospital. Amongst which, 71 patients were excluded due to missing data exceeding 75%. In the end, a total of 320 mild patients and 102 severe patients were included in this study. All patients were hospitalized at Shenzhen Third People's Hospital. Patients who meet any of the following criteria were defined as severely ill[19]: (1) shortness of breath, more than 30 breaths/min; (2) oxygen saturation during rest is less than 93%; (3) arterial oxygen partial pressure/inhaled oxygen fraction (FiO_2) ≤ 300 mmHg. The blood test indicators characteristics of all patients are shown in Table 1.

3.2 | Building a machine learning model to classify patients and evaluate the accuracy of said models

First, the 38 blood test indicators of the training data set are used as input variables to construct an ML model. In the random forest

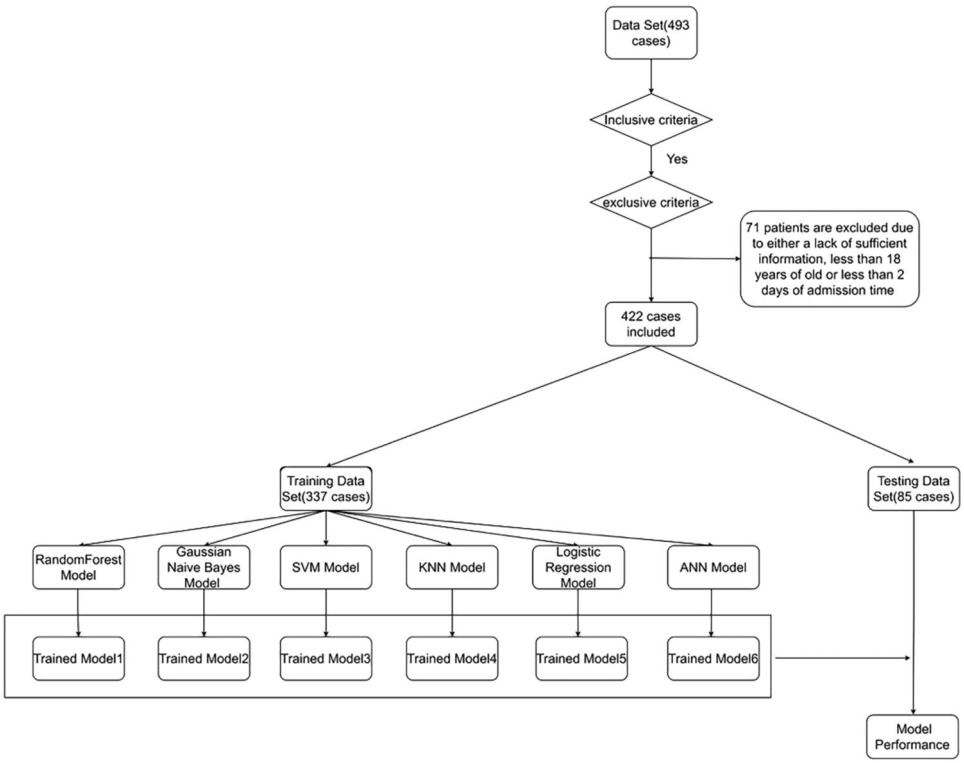

FIGURE 1 Characteristics of blood test indicators in mild patients and severe patients on the first day of admission

TABLE 1 Characteristics of blood test indicators in mild patients and severe patients on the first day of admission

Feature	Mild (mean)	Severe (mean)	*p*
Lactate (LACT), mmol/L	1.487698 ± 0.719862	1.382157 ± 0.477990	0.090545
K (K), mmol/L	4.032940 ± 1.267171	4.024118 ± 1.760442	0.962667
Na (Na), mmol/L	138.398843 ± 2.339505	136.723529 ± 3.779705	0.000045
Leukocyte (WBC), 10^9/L	5.039468 ± 1.931214	5.244020 ± 2.085185	0.381285
Neutrophil (N), %	56.504468 ± 12.17181	65.886275 ± 13.824229	<0.000001
Neutrophil count (NEUT), 10^9/L	2.874781 ± 1.393810	3.609216 ± 2.017959	0.000819
lymphocyte ratio (L), %	31.565937 ± 10.99588	23.089216 ± 10.805113	<0.000001
lymphocyte count (LYMPH), 10^9/L	1.609968 ± 0.814858	1.138039 ± 0.707512	<0.000001
Hematocrit (HCT), %	40.364062 ± 4.82476	40.251961 ± 4.404784	0.827197
Hemoglobin (HGB), g/L	137.215625 ± 15.545	138.666667 ± 15.635915	0.414905
Platelet count (PLT), 10^9/L	206.325000 ± 66.24752	158.715686 ± 45.764869	<0.000001
Alanine transaminase (ALT), U/L	24.441562 ± 20.02796	30.665686 ± 16.525764	0.001945
Aspartateaminotransferase (AST), U/L	28.140937 ± 14.73527	38.408824 ± 18.470598	<0.000001
TBil (TB), μmol/L	12.963406 ± 7.042354	14.093137 ± 9.530329	0.271125
DBil (DB), μmol/L	5.09875 ± 8.240408	5.543137 ± 6.186003	0.562603
Albumin (ALB), g/L	43.599593 ± 3.715105	40.823529 ± 4.155472	<0.000001
Globulin (GLO), g/L	25.555250 ± 3.63909	27.038627 ± 4.592339	0.003408
Lactate dehydrogenase (LDH), U/L	264.567968 ± 137.9644	390.563725 ± 249.702195	<0.000001
Urea (Urea), mmol/L	3.878843 ± 1.262233	5.075392 ± 2.351022	0.000002
Creatinine (Cr), μmol/L	62.058968 ± 17.383806	76.032353 ± 27.438619	0.000003
β−2M (β−2M), mg/L	2.854742 ± 2.661013	3.583978 ± 1.493213	0.000579
Cystatin C (CysC), mg/L	1.038077 ± 8.959017	1.266290 ± 0.349494	0.649709
Creatine kinase (CK), U/L	81.034111 ± 94.940552	122.968083 ± 143.923992	0.006660
Creatine kinase-MB (CKMB), U/L	2.294972 ± 3.557806	1.519831 ± 2.520650	0.015882
B-type natriuretic peptide (BNP), pg/ml	23.281879 ± 36.55976	30.183150 ± 40.815456	0.129560
Troponin I (TnI), ng/ml	0.007076 ± 0.00760	0.128589 ± 1.124593	0.277762
Killer T cell count (Tc.Count), /μl	1229.319513 ± 570.852511	671.892157 ± 380.036447	<0.000001
Helper T cell count (Th.Count), /μl	702.036770 ± 509.670617	393.436819 ± 222.245064	<0.000001
Suppressor T cell count (Ts.Count), /μl	460.958263 ± 247.096219	235.814815 ± 157.923257	<0.000001
Erythrocyte sedimentation rate (ESR), mm/h	29.016440 ± 22.258282	42.936275 ± 24.009478	<0.000001
C-reactive protein (CRP), mg/dl	12.691781 ± 16.477197	38.723784 ± 42.984750	<0.000001
Procalcitonin (Pct), ng/ml	0.051649 ± 0.166472	0.202475 ± 0.960864	0.117684
Interleukin-6 (IL6), pg/ml	10.120406 ± 11.545838	26.866078 ± 26.475956	<0.000001
Prothrombin time (PT), s	11.976697 ± 1.517530	12.652941 ± 3.216193	0.042429
Activated partial thromboplastin time (APTT), s	36.279267 ± 23.764997	37.253922 ± 5.708854	0.500011
Fibrinogen (FIB), G/L	4.015114 ± 2.308259	4.466461 ± 1.285479	0.013287
D-Dimer (D-Di), μg/ml	0.473766 ± 0.859517	1.038431 ± 2.308311	0.017260
PaO_2/FiO_2 (PF), mmHg	436.308088 ± 125.378358	343.205696 ± 107.754954	<0.000001

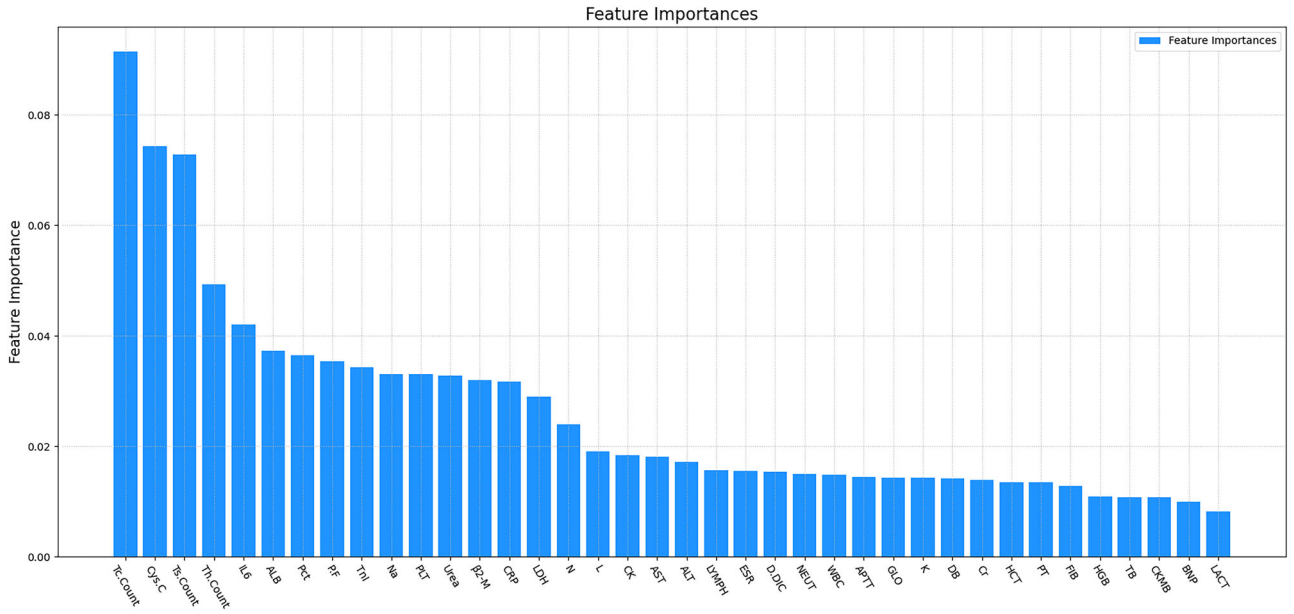

FIGURE 2 Importance of feature variables in Random Forest model

model, the max depth of hyperparameters is determined by grid search to be 10, and *n*_estimators being 1000. The AUC of the model in the test set is 0.89 (95% CI = 0.86–0.93), sensitivity is 83.8%, specificity is 81.2%, using feature importance analysis, with 0.04 as the threshold. Tc.Count, Cys.C, Ts.Count, Th.Count, and IL-6 were determined as important variables (Figure 2). The AUC of the naive Bayes model in the test set is 0.90 (95% CI = 0.84–0.92), the sensitivity is 85.9%, and the specificity is 75.0%. In the SVM model (this study uses support based on linear functions), the AUC of the model in the test set is 0.86 (95% CI = 0.82–0.90), the sensitivity is 87.1%, and the specificity is 72.7%. For the KNN model (with *K* value being 3), the AUC in the test set is 0.78 (95% CI = 0.76–0.85), sensitivity is 76.0%, and specificity is 66.7%. The AUC of the logistic regression model in the test set was 0.84 (95% CI = 0.81–0.89), sensitivity was 82.1%, and specificity was 71.4%. In the artificial neural network model, this study set up a three-layer artificial neural network (Figure 3), with four neurons in the input layer, five neurons in each hidden layer, and two neurons in the output layer. Using the backpropagation algorithm, the activation function of the hidden layer is the Sigmoid function, the output function is the SoftMax function, the learning rate is 0.001, and the iteration is 100 times. Finally, the AUC of the model in the test set is 0.87 (95% CI = 0.83–0.94), the sensitivity is 78.9%, and the specificity is 70.2%. The results show that the models built by different ML methods have relatively good effects on the prediction and classification of patients with mild and severe diseases. Among them, the naive Bayes model has the highest performance, with an AUC of 0.90. The ROC curve, recall curve, and performance comparison of each model of all models are shown in Figure 4, Figure 5, and Table 2.

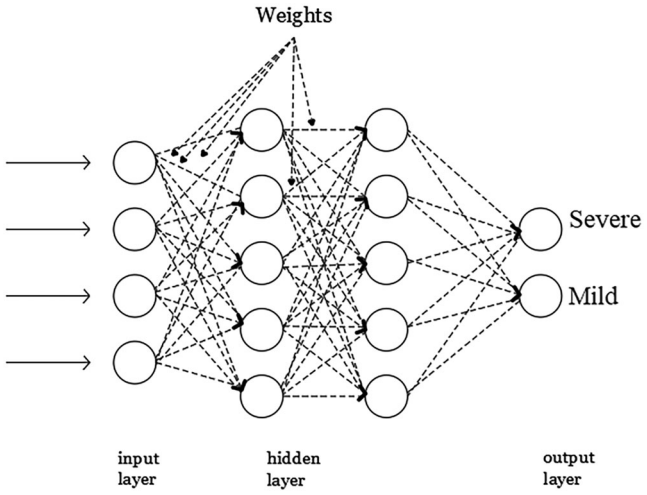

FIGURE 3 The construction of the artificial neural network

4 | DISCUSSION

For the first time in this study, 38 blood test indicator variables from patients diagnosed with COVID-19 on the first day of admission were used to construct different ML models, including random forest, naive Bayes, SVM, KNN, logistic regression, and artificial neural network. The models were used to classify patients with mild and severe cases of COVID-19. Each model shows a good classification effect. The AUC of the random forest model is 0.89. The AUC of the naive Bayes model is 0.90. The AUC of the SVM model is 0.86. The AUC of the KNN model is 0.78. The AUC of the logistic regression model is 0.84. The

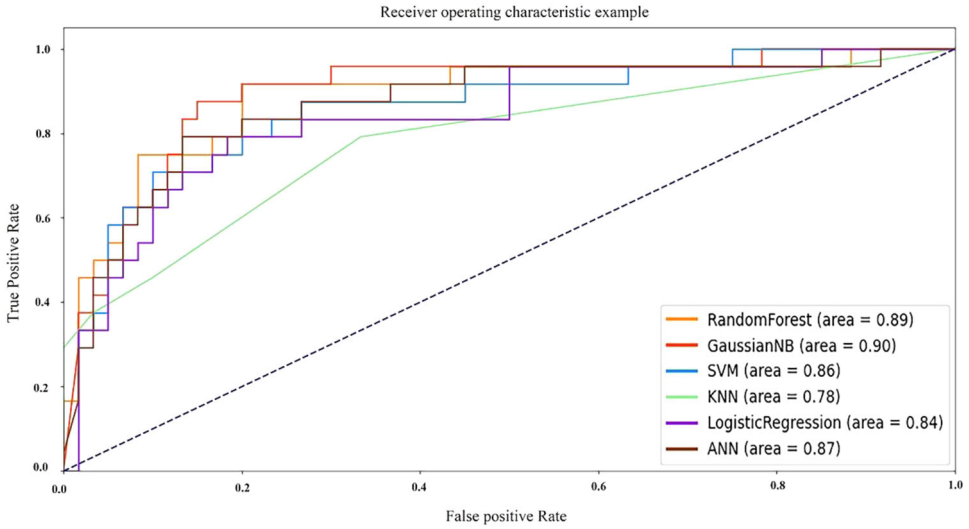

FIGURE 4 The ROC curve of the various models

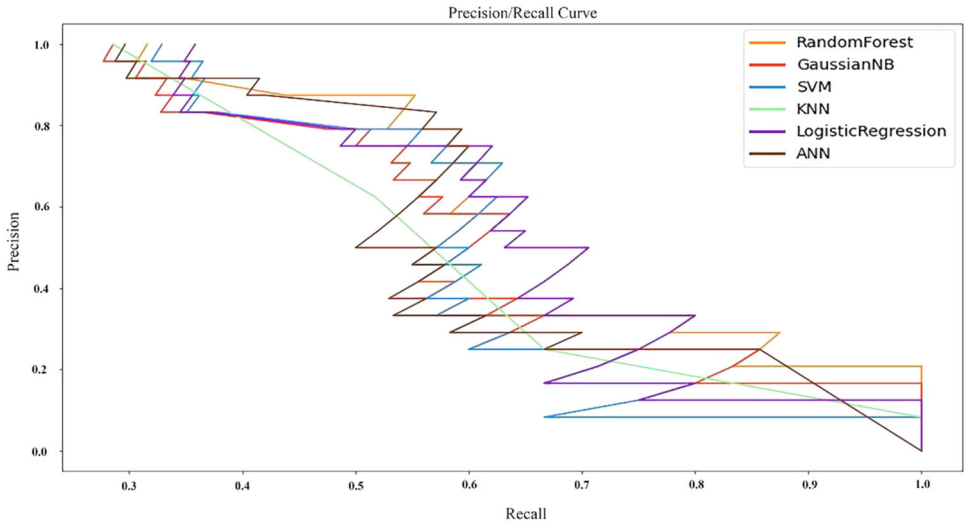

FIGURE 5 The precision–recall curve of the various models

ML model	Data set	Sensitivity (%)	Specificity (%)	AUC	AUC 95%CI
Random Forest	Testing data set	83.8	81.2	0.89	0.86–0.93
Naïve Bayes	Testing data set	85.9	75.0	0.90	0.84–0.92
SVM	Testing data set	87.1	72.7	0.86	0.82–0.90
KNN	Testing data set	76.0	66.7	0.78	0.76–0.85
Logistic regression	Testing data set	82.1	71.4	0.84	0.81–0.89
ANN	Testing data set	78.9	70.2	0.87	0.83–0.94

TABLE 2 Performance of the ML models on the classification of mild and severe cases of COVID-19

Abbreviations: AUC, area under the curve; CI, confidence interval; KNN, k-nearest neighbor; ML, machine learning; SVM, support vector machine.

AUC of the artificial neural network model is 0.87. Among these models, the Naive Bayes model has the best performance.

Previous studies have focused more on screening out several important factors related to COVID-19 critical illness,[13,20] but simply considering individual variables as classification and predictive indicators has a certain degree of one-sidedness, and COVID-19 can be present in various systems throughout the body, causing damage.[21] Therefore, COVID-19 patients need to be fully evaluated when they

are admitted to the hospital. Routine blood tests are indispensable for COVID-19 patients when they are admitted to the hospital. The 38 blood test indicators included in this study can initially assess the patient's liver and kidney functions, blood coagulation functions, heart functions, and understand the patient's internal environment, immune system, and inflammation situation. In response to the situation, this study also shows that the 38 blood indicators on the first day of admission for patients diagnosed with COVID-19 can be used as detection indicators to classify patients with mild and severe cases of the disease.

At the same time, in the feature importance analysis of the random forest model, several important indicators such as Killer T cell count, Cystatin C, Suppressor T cell count, Helper T cell count, Interleukin-6 and outcome variables can be determined. Among them, the serum concentration of Cys.C is mainly determined by the glomerular filtration rate, which is an important indicator of the glomerular filtration rate. Renal failure can lead to an increase in the concentration of Cys.C, therefore, Cystatin C is an earlier and more sensitive marker of kidney malfunction.[22] The function of cystatin C is closely related to the function of its target enzyme. It exerts a variety of immunomodulatory functions by controlling the activity of cysteine proteases and other mechanisms. including the regulation of innate immune cell phagocytosis, and the major histocompatibility complex- II (MHC-II) mediated antigen presentation, and so on. In SARS-CoV-2, it plays a key role in the infection and inflammation caused by the virus,[23] and is important for its diagnosis. Research reports have shown that Cystatin C can be used as a potential biomarker to predict the adverse outcomes of SARS-CoV-2 infection.[24] It is also an independent risk factor predicting the death of COVID-19's critically ill patients.[22] In addition, it is generally believed that SARS-CoV-2 can induce immune disorders and excessive inflammation,[25,26] which is mainly manifested in the depletion of peripheral blood lymphocytes and cytokine cascades. SARS-CoV-2 can cause a specific T cell response and activated Th1 (T helper cells) pass NF-kB.[27] The signaling pathway produces a pro-inflammatory response and causes a downstream cytokine storm leading to diffuse lung tissue damage,[28] which can even lead to death of the patient.[29] Lymphopenia is the most consistent laboratory abnormality in patients infected with COVID-19 and it can be observed in up to 72%–85% of severe cases,[28] and a decrease in the total number of T lymphocytes can be observed in severe COVID-19 patients.[30] SARS-CoV-2 can drive T cell failure in COVID-19 patients,[30] and the release of T cell-dependent cytokines and direct cytotoxicity can also cause tissue inflammation and toxicity, and accelerate mortality.[31] In this study, the 3 T-lymphocyte subsets (Tc.Count, Ts.Count, Th.Count) of severe COVID-19 patients are generally decreased. When compared with mild patients, the numbers are significantly reduced (Tc.Count: 671.892157 vs. 1229.319513/μl; $p < 0.000001$, Ts.Count: 235.814815 vs. 460.958263/μl; $p < 0.000001$, 393.436819 vs. 702.036770/μl; $p < 0.000001$; Table 1). The level of inflammatory cytokines was significantly increased (CRP: 38.723784 vs. 12.691781 mg/dl; $p < 0.000001$, IL6: 26.866078 vs. 10.120406 pg/ml; $p < 0.000001$; Table 1). The decrease of lymphocytes can lead to unfavorable changes for COVID-19 patients. There are also research reports showing that lymphocyte subsets can be used for early screening of severe COVID-19.[32] In addition, SARS-CoV-2 infection can also trigger a cytokine storm through the JAK/STAT pathway and produce various inflammatory markers through hosts such as macrophages, monocytes, and lymphocytes,[33] among which IL-6 is an important inflammatory biomarker, mainly produced by macrophages.[34] The overproduction of IL-6 can trigger the biological effects of organ damage, so it is the main pro-inflammatory mediator that induces the acute phase response,[35] and can cause extensive local and systemic changes. IL-6 can also inhibit T cell activation and cause lymphopenia.[36] Some studies have shown that IL-6 increases with the severity of the disease; it is the severity of the disease A predictor of degree,[37,38] increasing IL-6 levels can be observed in more than 50% of COVID-19 patients.[39] In this study, it can be observed that severe case patients of COVID-19 have a significantly higher IL6 level compared to mild case patients (26.866078 vs. 10.120406 pg/ml; $p < 0.000001$), indicating that IL6 can be used as an important marker for the severity of disease in patients with COVID-19. Therefore, in testing for COVID-19 patients diagnosed in hospital, in addition to 38 blood variable indicators, the level of T lymphocyte subsets and IL-6 can also be focused on to identify changes in the patients' conditions as soon as possible.

However, this study also has certain limitations, The potential limitations of this study may include the moderate amount of samples when constructing the model, relatively small size of samples being used to verify the sample. The data is completely from China, which means the model may be of limited applicability in other parts of the world. Therefore, later follow-up studies will require more samples to verify the accuracy of the model, including the verification in other parts of China, other nations, and ethnicities. At the same time, while the blood test indicators are holistic in their approach and can be used for the overall status of the patients when they are initially hospitalized, but due to the different economic situations for different nations and regions and the different examination methods for these nations and regions, this may limit the promotion of the study's results in other parts of the world.

5 | CONCLUSION

In this study, different ML models were constructed based on 38 blood indicators variables of patients diagnosed with COVID-19 on the first day of admission, which can more accurately classify patients into mild and severe cases of COVID-19.

ACKNOWLEDGMENT

Yi-jin Yang at Shenzhen University provided English editing assistance.

ETHICS STATEMENT

All patients who participated also consented to the publication of this manuscript. This study has been approved by the Ethics Committee

of the Shenzhen Third People's Hospital. All subjects signed informed consent and all methods are done according to the Declaration of Helsinki together with relevant rules and standards.

AUTHOR CONTRIBUTIONS

Rui-kun Zhang and Qi Xiao conceived the study design and determined appropriate surveys and survey items for the study. Rui-kun Zhang and Qi Xiao collected the data and performed the data analysis. Rui-kun Zhang, Qi Xiao, Sheng-lang Zhu, Hai-yan Lin, and Ming Tang interpreted analysis results and participated in drafts and revisions of this manuscript. All authors reviewed and provided approval for the final version of the manuscript.

DATA AVAILABILITY STATEMENT

Due to institutional restrictions on data sharing and privacy issues, the data set generated and/or analyzed during the current research period has not been publicly available. However, the data and code can be obtained from the corresponding author upon reasonable requests.

ORCID

Rui-kun Zhang http://orcid.org/0000-0001-6455-4617

REFERENCES

1. Alsharif W, Qurashi A. Effectiveness of COVID-19 diagnosis and management tools: a review. *Radiography*. Vol 27. London, England: 1995;2021:682-687.
2. Pascarella G, Strumia A, Piliego C, et al. COVID-19 diagnosis and management: a comprehensive review. *J Intern Med*. 2020;288(2):192-206.
3. Wu D, Wu T, Liu Q, Yang Z. The SARS-CoV-2 outbreak: what we know. *Int J Infect Dis*. 2020;94:44-48.
4. Guan WJ, Ni ZY, Hu Y, et al. Clinical characteristics of Coronavirus disease 2019 in China. *N Engl J Med*. 2020;382(18):1708-1720.
5. Grasselli G, Pesenti A, Cecconi M. Critical care utilization for the COVID-19 outbreak in Lombardy, Italy: early experience and forecast during an emergency response. *JAMA*. 2020;323(16):1545-1546.
6. Wu Z, McGoogan JM. Characteristics of and important lessons from the Coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese center for disease control and prevention. *JAMA*. 2020;323(13):1239-1242.
7. Di Castelnuovo A, Bonaccio M, Costanzo S, et al. Common cardiovascular risk factors and in-hospital mortality in 3,894 patients with COVID-19: survival analysis and machine learning-based findings from the multicentre Italian CORIST Study. *Nutr Metab Cardiovasc Dis*. 2020;30(11):1899-1913.
8. Tian W, Jiang W, Yao J, et al. Predictors of mortality in hospitalized COVID-19 patients: a systematic review and meta-analysis. *J Med Virol*. 2020;92(10):1875-1883.
9. Mei X, Lee HC, Diao KY, et al. Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. *Nature Med*. 2020;26(8):1224-1228.
10. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ (Clin Res Ed)*. 2020;369:m1328.
11. Li WT, Ma J, Shende N, et al. Using machine learning of clinical data to diagnose COVID-19: a systematic review and meta-analysis. *BMC Med Inform Decis Mak*. 2020;20(1):247.
12. Gao Y, Cai GY, Fang W, et al. Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. *Nat Commun*. 2020;11(1):5033.
13. Tang M, Yu XX, Huang J, et al. Clinical diagnosis of severe COVID-19: a derivation and validation of a prediction rule. *World J Clin Cases*. 2021;9(13):2994-3007.
14. Fan BE, Chong VCL, Chan SSW, et al. Hematologic parameters in patients with COVID-19 infection. *Am J Hematol*. 2020;95(6):E131-E134.
15. Ferrari D, Motta A, Strollo M, Banfi G, Locatelli M. Routine blood tests as a potential diagnostic tool for COVID-19. *Clin Chem Lab Med*. 2020;58(7):1095-1099.
16. Formica V, Minieri M, Bernardini S, et al. Complete blood count might help to identify subjects with high probability of testing positive to SARS-CoV-2. *Clin Med*. 2020;20(4):e114-e119.
17. Cabitza F, Campagner A, Ferrari D, et al. Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests. *Clin Chem Lab Med*. 2020;59(2):421-431.
18. Ibrahim JG, Chen MH, Lipsitz SR, Herring AH. Missing-data methods for generalized linear models: a comparative review. *J Am Stat Assoc*. 2005;100(469):332-346.
19. Wei PF. Diagnosis and treatment protocol for novel coronavirus pneumonia (trial version 7). *Zhonghua yixue zazhi*. 2020;133:1087-1095.
20. Liang W, Liang H, Ou L, et al. Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients With COVID-19. *JAMA Intern Med*. 2020;180(8):1081-1089.
21. Wang D, Hu B, Hu C, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel Coronavirus-infected Pneumonia in Wuhan, China. *JAMA*. 2020;323(11):1061-1069.
22. Li Y, Yang S, Peng D, et al. Predictive value of serum cystatin C for risk of mortality in severe and critically ill patients with COVID-19. *World J Clin Cases*. 2020;8(20):4726-4734.
23. Zi M, Xu Y. Involvement of cystatin C in immunity and apoptosis. *Immunol Lett*. 2018;196:80-90.
24. Chen D, Sun W, Li J, et al. Serum Cystatin C and Coronavirus disease 2019: a potential inflammatory biomarker in predicting critical illness and mortality for adult patients. *Mediators Inflamm*. 2020;2020:3764515.
25. Vardhana SA, Wolchok JD. The many faces of the anti-COVID immune response. *J Exp Med*. 2020;217(6).
26. Yao Z, Zheng Z, Wu K, Junhua Z. Immune environment modulation in pneumonia patients caused by coronavirus: SARS-CoV, MERS-CoV and SARS-CoV-2. *Aging*. 2020;12(9):7639-7651.
27. Veldhuizen RAW, Zuo YY, Petersen NO, Lewis JF, Possmayer F. The COVID-19 pandemic: a target for surfactant therapy? *Expert Rev Respir Med*. 2021;15(5):597-608.
28. Bordallo B, Bellas M, Cortez AF, Vieira M, Pinheiro M. Severe COVID-19: what have we learned with the immunopathogenesis? *Advs Rheumatol*. 2020;60(1):50.
29. Sacchi A, Grassi G, Bordoni V, et al. Early expansion of myeloid-derived suppressor cells inhibits SARS-CoV-2 specific T-cell response and may predict fatal COVID-19 outcome. *Cell Death Dis*. 2020;11(10):921.
30. Diao B, Wang C, Tan Y, et al. Reduction and functional exhaustion of T cells in patients with Coronavirus disease 2019 (COVID-19). *Front Immunol*. 2020;11:827.
31. Vardhana SA, Wolchok JD. The many faces of the anti-COVID immune response. *J Exp Med*. 2020;217(6).
32. Qin C, Zhou L, Hu Z, et al. Dysregulation of immune response in patients with Coronavirus 2019 (COVID-19) in Wuhan, China. *Clin Infect Dis*. 2020;71(15):762-768.
33. Satarker S, Tom AA, Shaji RA, Alosious A, Luvis M, Nampoothiri M. JAK-STAT pathway inhibition and their implications in COVID-19 therapy. *Postgrad Med*. 2021;133(5):489-507.
34. Paces J, Strizova Z, Smrz D, Cerny J. COVID-19 and the immune system. *Physiol Res*. 2020;69(3):379-388.

35. Fattori E, Cappelletti M, Costa P, et al. Defective inflammatory response in interleukin 6-deficient mice. *J Exp Med*. 1994;180(4):1243-1250.
36. Domingo P, Mur I, Pomar V, Corominas H, Casademont J, de Benito N. The four horsemen of a viral apocalypse: the pathogenesis of SARS-CoV-2 infection (COVID-19). *EBioMedicine*. 2020;58:102887.
37. Han H, Ma Q, Li C, et al. Profiling serum cytokines in COVID-19 patients reveals IL-6 and IL-10 are disease severity predictors. *Emerg Microbes Infect*. 2020;9(1):1123-1130.
38. Thevarajan I, Buising KL, Cowie BC. Clinical presentation and management of COVID-19. *Med J Aust*. 2020;213(3):134-139.
39. Prompetchara E, Ketloy C, Palaga T. Immune responses in COVID-19 and potential vaccines: lessons learned from SARS and MERS epidemic. *Asian Pac J Allergy Immunol*. 2020;38(1):1-9.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Zhang R-k, Xiao Q, Zhu S-l, Lin H-y, Tang M. Using different machine learning models to classify patients into mild and severe cases of COVID-19 based on multivariate blood testing. *J Med Virol*. 2022;94:357-365. https://doi.org/10.1002/jmv.27352