Research on Risk Factor Prediction Model for COVID-19 Patients- Based on Machine Learning Methods

XiaoHui Zuo[1,2], Lingling Bai[1,2], Qin Ran[1,2], Xiang He[1,2], Guoping Li[1,2]

[1]Affiliated Hospital of Southwest Jiaotong University/Third People's Hospital of Chengdu City (610014)

[2]Laboratory of Allergy and Precision Medicine (610014)

Introduction. To establish a machine learning model for predicting the risk of adverse outcomes in COVID- 19 patients, evaluating the risk of disease progression and reducing the incidence of poor outcomes.

Methods. A retrospective analysis was conducted on 596 COVID- 19 patients who visited the Third People's Hospital of Chengdu City from December 2022 to February 2023. Feature selection algorithms such as Boruta and RFECV were used to construct ten machine learning models, including logistic regression, nearest neighbor algorithm, and decision tree. Shapley feature selection and one-way analysis of variance (ANOVA) were used to explore risk factors associated with combined fungal infection, hospitalization longer than 30 days,and death in COVID- 19 patients.

Results. In the baseline differential analysis, except for monocyte percentage, fever, and smoking quantity (cigarettes/day), there were no statistically significant differences in all other measured variables between the training and test sets ($p<0.05$). In predicting risk factors in COVID- 19 patients, quadratic discriminant analysis (QDA), logistic regression (LR), and support vector machine (SVC) models performed well. Monocyte percentage, CK-MB/CK, IL-6, fungal infection, immunotherapy, and antibiotic use were key clinical features influencing the output of the models.

Conclusion. Quadratic discriminant analysis (QDA), logistic regression (LR), and support vector machine (SVC) models performed well in predicting risk factors in COVID- 19 patients. Monocyte percentage, CK-MB/CK, IL-6, fungal infection, immunotherapy, and antibiotic use were significant risk factors for poor prognosis in COVID- 19 patients.

Keywords. machine learning; novel coronavirus pneumonia; fungal infection; hospitalization duration; risk factors

INTRODUCTION

The Coronavirus Disease 2019 (COVID- 19) is a severe acute respiratory infectious disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).As of October 22, 2023, there have been over 771 million confirmed cases of COVID- 19 and more than 6 million deaths reported globally[1].This pandemic has presented significant challenges to healthcare systems and clinical practitioners[2] .

Current research indicates that factors such as age, gender, and comorbidities can worsen the prognosis of COVID- 19[3].Yet, the factors associated with poor prognosis in COVID- 19 are multifaceted and varied.Consequently, swiftly identifying high-risk patients is vital for a prompt response to the pandemic and for the rational allocation of resources.Machine learning, a branch of artificial intelligence, excels in discerning patterns from vast data sets, thereby facilitating prediction and decision-making processes.Throughout the pandemic, the deployment of this technology has markedly enhanced the efficiency of information processing and offered effective and precise support for clinical decision-making[4,5] .

Although previous research has identified certain risk factors and assessed the risk of mortality, this study consolidates various data sources to comprehensively evaluate ten machine learning algorithms, aiming to improve the precision of disease prediction[6].

1 MATERIALS AND METHODS

1.1 Clinical Data Collection

Between December 2022 and February 2023, a comprehensive dataset comprising 596 clinical records of COVID- 19 patients was gathered. This dataset originated from various departments, including the Respiratory Medicine Department at the Third People's Hospital of Chengdu.All patient diagnoses and treatments conformed to the guidelines outlined in the "Diagnosis and Treatment Protocol for Novel Coronavirus Pneumonia (Trial Version 10)"[7] .

1.2 Inclusion and Exclusion Criteria

The inclusion criteria encompass: (1) clinical manifestations consistent with COVID- 19 infection, such as fever, dry cough, and fatigue;(2) one or more of the following etiological and serological test results:positive result for SARS-CoV-2 from a nucleic acid or antigen test;positive isolation and culture of SARS-CoV-2;a fourfold or greater rise in the titer of specific IgG antibodies against SARS-CoV-2 during the convalescent phase, in comparison to the acute phase.The exclusion criteria are as follows: (1) non-COVID- 19 respiratory infections, including those caused by other pathogens such as the common cold and influenza;(2) patients who are unable to cooperate with treatment or complete the diagnostic and treatment process due to mental or psychological factors;(3) patients whose clinical data are significantly incomplete.

1.3 Definitions and Clinical Characteristics

In this research, the diagnosis and treatment of COVID- 19 are based on China's "Diagnosis and Treatment Protocol for Novel Coronavirus Infection (Trial Version 10)." The diagnosis was established through a comprehensive evaluation of epidemiological history, clinical symptoms, and laboratory tests, with a positive nucleic acid test for SARS-CoV-2 serving as the primary diagnostic criterion.

After pre-processing the data through filtering, cleaning, imputation, and standardization, we successfully narrowed down the number of clinical characteristics for patients to 73.Thereafter, we analyzed the differences in baseline characteristics between the training and testing datasets (Figure 1 outlines the division process).The clinical characteristics encompassed laboratory test results, imaging data, and patient history, including:

antibiotic use, antiviral therapy, steroid therapy, mechanical ventilation, hospital stay duration, outcome, white blood cell count, neutrophil count, neutrophil percentage, lymphocyte count, lymphocyte percentage, monocyte count, monocyte percentage, red blood cell count, hemoglobin level, platelet count, procalcitonin, interleukin-6, B-type natriuretic peptide, creatine kinase, creatine kinase MB isoenzyme, CK-MB/CK ratio, myoglobin concentration, high-sensitivity cardiac troponin T, D-dimer quantitative test, fibrinogen degradation product determination, total bilirubin, total protein, albumin, globulin, alanine aminotransferase, aspartate aminotransferase, AST/ALT ratio, lactate dehydrogenase, urea, creatinine, uric acid, fever, cough, sputum production, dyspnea, chest pain, throat swelling, hemoptysis, chest tightness, palpitations, fatigue, neurological symptoms, digestive system symptoms, convulsions, nausea, vomiting, smoking history, current smoking status, smoking duration, daily smoking amount (cigarettes/day), number of vaccine doses, coronary heart disease, hypertension, pulmonary disease, diabetes, kidney disease, Parkinson's disease, liver disease, tumor, hematological disease, immunodeficiency, other systemic disease history, respiratory rate (breaths per minute), oxygen saturation (SpO2), CT staging.

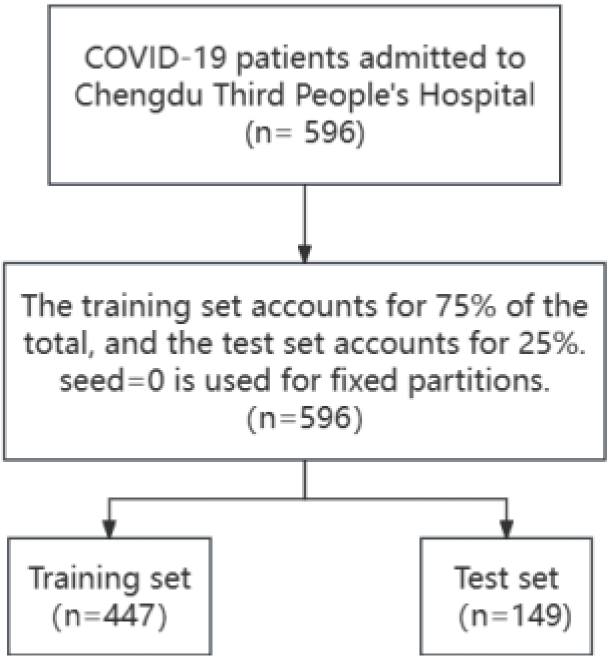

COVID-19 patients admitted to Chengdu Third People's Hospital (n= 596)
The training set accounts for 75% of the total, and the test set accounts for 25%. seed=0 is used for fixed partitions. (n=596)
Training set (n=447)
Test set (n=149)

Figure 1: Dividing the training set into the test set

1.4 Data Checking and Exploratory Data Analysis

Utilizing Python for Data Inspection and Exploratory Data Analysis (EDA), the focus is on missing value analysis, outlier detection, data type review, and consistency assessment. EDA is predominantly employed for descriptive statistical analysis, data visualization, and correlation analysis[8]. In this study, third-party libraries such as Pandas are utilized for meticulous missing value analysis, with the aid of Numpy and Pandas for calculating descriptive statistical metrics. Additionally, Pandas and the Scikit-learn library are employed for correlation analysis. Moreover, the K-Nearest Neighbors (KNN) algorithm is adopted to impute missing data, thereby reducing the extent of data absence.

1.5 Feature Selection

Employing three algorithms that include Boruta feature selection based on the Random Forest model, Recursive Feature Elimination with Cross-Validation (RFECV), and Linear Support Vector Machine[9]. After the feature selection process is completed using individual methods, the plotly library is utilized for visualization analysis to identify key features[10] .

1.6 Model Building and Tuning

The efficacy of the following ten machine learning models is evaluated based on metrics such as the Area Under the Receiver Operating Characteristic (ROC) curve (AUC), accuracy, recall, precision, and F1 score[11] :

1. Logistic Regression (LR);
2. Nearest Neighbors algorithm;
3. Support Vector Machine (SVM);
4. Decision Tree;
5. Random Forest;
6. AdaBoost;
7. Gradient Boosting;
8. Naive Bayes (NB);
9. Linear Discriminant Analysis (LDA);
10. Quadratic Discriminant Analysis (QDA)[12] .

By analyzing these metrics, the aim is to select the optimal model algorithm, with the specific process depicted in Figure 2.

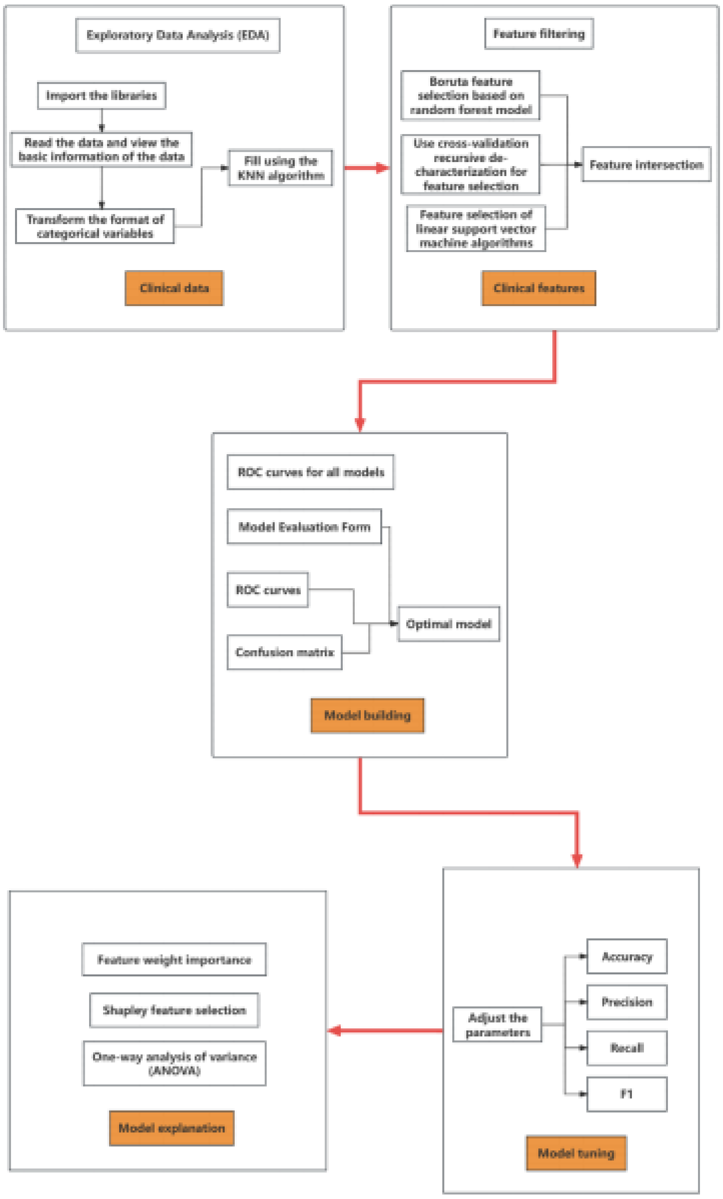

Exploratory Data Analysis (EDA)
Import the libraries
Read the data and view the basic information of the data
Fill using the KNN algorithm
Transform the format of categorical variables
Clinical data
Feature filtering
Boruta feature selection based on random forest model
Use cross-validation recursive de-characterization for feature selection
Feature intersection
Feature selection of linear support vector machine algorithms
Clinical features
ROC curves for all models
Model Evaluation Form
ROC curves
Optimal model
Confusion matrix
Model building
Feature weight importance
Shapley feature selection
One-way analysis of variance (ANOVA)
Model explanation
Adjust the parameters
Accuracy
Precision
Recall
F1
Model tuning

Figure 2: Modeling Process

The provided flowchart illustrates the interconnection of various concepts within the modeling process. The K-Nearest Neighbors (KNN) algorithm is utilized to estimate the missing value for a data point that contains missing values by leveraging the k most similar neighbors to that data point. The term ROC curve refers to the Receiver Operating Characteristic curve.

1.7 Model Interpretation and Statistical Analysis

The methods applied in model interpretation include feature weight importance, Shapley feature selection, and Analysis of Variance (ANOVA)[13] .

The dataset was divided into training and testing sets using a random sampling method. The `descrTable` function from the `compareGroups` package was employed for baseline difference analysis. Variables post preprocessing were analyzed one by one using the Python `Dataprep` library. Pearson's correlation coefficient, Spearman's rank correlation coefficient, and Kendall's rank correlation coefficient were utilized to measure the correlation between variables. The statistical significance level for two-tailed tests was set at $p < 0.05$. All of these analyses were conducted in the R (version 4.2.3) and Python (version 3.8) environments to ensure the accuracy of the results.

2 RESULTS

2.1 Patient Characteristics

Atotal of 596 clinical data records of patients with COVID- 19 were included in this study, with all data being divided into training and testing sets in a ratio of 3:1.

The baseline characteristics of the training and testing sets are analyzed and presented in Table 1. Among them, 22% of the patients had fungal infections, with the incidence rates of fungal infections in the training and testing sets being 22.1% and 21.5%, respectively. The mortality rate was 5.9%, with the mortality rates in the training and testing sets being 6.04% and 5.37%, respectively. The overall average length of hospital stay was 23.68 days, with the average hospital stay for patients in the training set being 23.8 days (standard deviation of 17 days), and for patients in the testing set being 23.4 days (standard deviation of 13.7 days).

Apart from the percentage of monocytes, fever, and the amount of smoking (cigarettes/day), there were no statistically significant differences ($p<0.05$) in all other measured variables between patients in the training and testing sets[4] .

Table 1: Analysis of the difference between the baseline characteristics of the training set and the test set (see the attached table for the full table)

Feature	Test N=149	Train N=447	p.overall
Gender:			0.346
Female	47 (31.5%)	162 (36.2%)	
Male	102 (68.5%)	285 (63.8%)	
Age	73.6 (13.1)	74.3 (13.8)	0.558
BMI	22.9 (4.31)	22.8 (3.73)	0.807
Immunotherapy:			0.319
No	112 (75.2%)	315 (70.5%)	
Yes	37 (24.8%)	132 (29.5%)	
Antifungal:			0.864
No	115 (77.2%)	350 (78.3%)	
Yes	34 (22.8%)	97 (21.7%)	
Antibiotics:			0.962
No	87 (58.4%)	264 (59.1%)	
Yes	62 (41.6%)	183 (40.9%)	
Antiviral:			0.730

No	115 (77.2%)	353 (79.0%)	
Yes	34 (22.8%)	94 (21.0%)	
Hormones:			0.881
No	52 (34.9%)	151 (33.8%)	
Yes	97 (65.1%)	296 (66.2%)	

Data arepresented as mean ± standard deviation or n%.

2.2 Feature Selection

Feature selection utilizing three distinct algorithms was performed to identify key features closely associated with three outcome variables through intersection. The following key features were found to be closely related to the outcome variable "presence or absence of fungal infection": the use of antibiotics; length of hospital stay; presence or absence of immunotherapy; percentage of monocytes; and myoglobin concentration.

Key features associated with the outcome variable "hospital stay exceeding 30 days" include: alanine aminotransferase (ALT); use of antibiotics; presence or absence of fungal infection; presence or absence of immunotherapy; percentage of monocytes; myoglobin concentration; total serum protein; and uric acid concentration.

The key features for the outcome variable "mortality" are: age; creatine kinase MB isoenzyme to total creatine kinase (CK-MB/CK) ratio; interleukin 6; and myoglobin concentration.

2.3 Model Evaluation and Selection

The present study conducted a comprehensive analysis of ten algorithms, including the ROC curve, model evaluation tables, and confusion matrices, to determine the optimal predictive model for each outcome variable[14].

As depicted in Figure 3A, for the outcome variable "presence or absence of fungal infection," the QDA, LR, LDA, and NB models all achieved the highest AUC value of 0.8. Concurrently, the QDA model also exhibited the highest accuracy, precision, and F1 score (as shown in Table 2). Therefore, the QDA model demonstrates a superior performance in the judgment and prediction of fungal infections.

The optimal predictive model for the outcome variable "hospital stay exceeding 30 days" is the LR model (as shown in Figure 4A). The AUC value of the LR model reached the highest value of 0.77, and it also performed well in terms of accuracy, precision, and F1 score (as shown in Table 3).

Using the aforementioned analytical methods, it was determined that the optimal predictive model for the outcome variable "mortality" is the SVC model (as shown in Table 4 and Figure 5).

Figures 3, 4, and 5, panels B and C, respectively, illustrate the ROC curves and confusion matrices of the best predictive models for the three outcome variables after parameter tuning. Following parameter tuning and optimization, the models' accuracy, AUC values, and precision were improved to varying degrees, thereby enhancing model performance.

Table 2: Model Evaluation Table for the Outcome Variable "Presence or Absence of Fungal Infection"

Classifier	Accuracy	ROC_AUC	Recall	Precision	F1
Quadratic DA	83.22	0.8	0.37	0.65	0.47
Logistic Regression	82.55	0.8	0.33	0.62	0.43
Support Vectors	81.21	0.77	0.07	1	0.12
Linear DA	81.21	0.8	0.37	0.55	0.44
Naive Bayes	80.54	0.8	0.4	0.52	0.45
Random Forest	79.19	0.76	0.4	0.48	0.44
Nearest Neighbors	77.85	0.6	0.1	0.33	0.15
AdaBoost	77.85	0.72	0.23	0.41	0.3
Gradient Boosting	77.18	0.75	0.3	0.41	0.35
Decision Tree	76.51	0.64	0.43	0.42	0.43

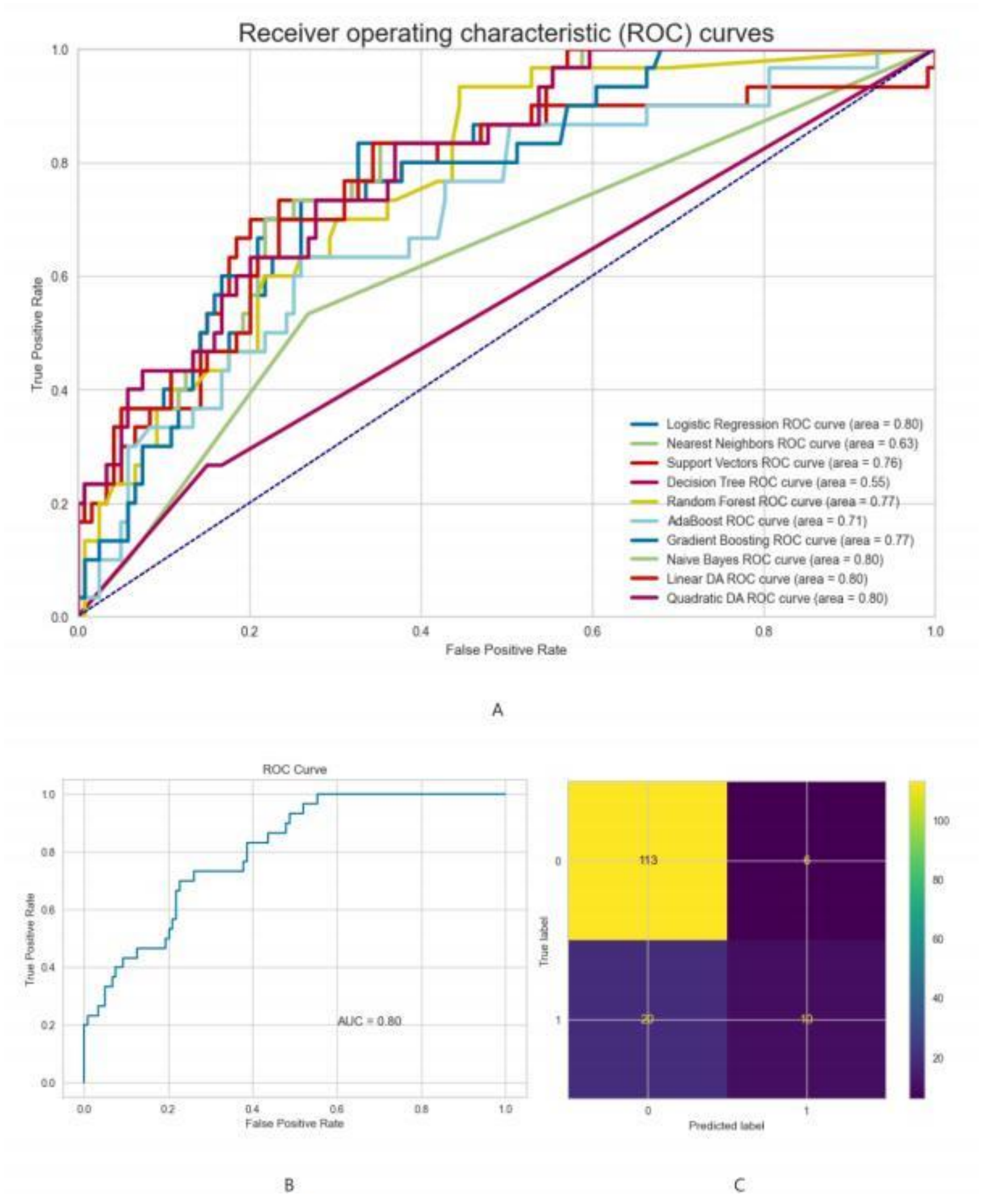


Figure 3: Model Evaluation for the Outcome Variable "Presence or Absence of Fungal Infection"
(A) ROC curves often models; (B) ROC curves of the optimal models after parameter tuning; (C) Confusion matrices of the optimal models after parameter tuning.AUC: Area Under the Curve.

Table 3: Model Evaluation Table for the Outcome Variable "Hospital Stay Exceeding 30 Days"

Classifier	Accuracy	ROC_AUC	Recall	Precision	F1
Logistic Regression	83.89	0.77	0.46	0.76	0.57
Random Forest	83.89	0.72	0.43	0.79	0.56
Gradient Boosting	80.54	0.73	0.43	0.62	0.51
Linear DA	79.87	0.77	0.46	0.59	0.52
Quadratic DA	79.19	0.72	0.46	0.57	0.51
Naive Bayes	77.85	0.76	0.49	0.53	0.51
AdaBoost	77.18	0.71	0.37	0.52	0.43
Support Vectors	76.51	0.58	0.03	0.5	0.05
Decision Tree	74.5	0.71	0.66	0.47	0.55
Nearest Neighbors	73.83	0.54	0.09	0.3	0.13

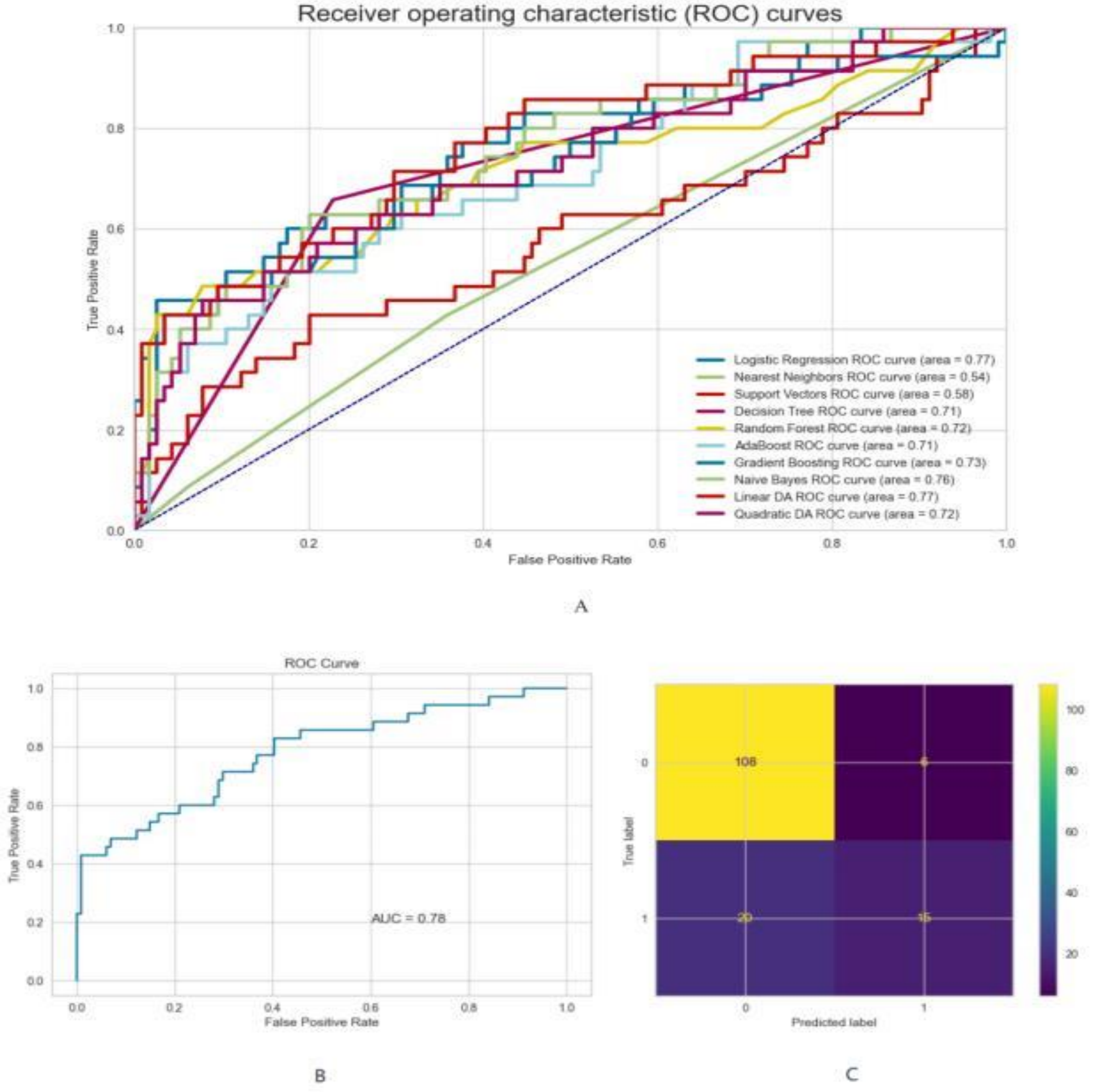


Figure 4: Model Evaluation for the Outcome Variable "Hospital Stay Exceeding 30 Days"
(A) ROC curves often models;(B) ROC curves of the optimal models after parameter tuning; (C) Confusion matrices of the optimal models after parameter tuning.

Table 4: Model Evaluation Table for the Outcome Variable "Mortality"

Classifier	Accuracy	ROC_AUC	Recall	Precision	F1
Support Vectors	95.3	0.59	0	NaN	NaN
Logistic Regression	94.63	0.46	0	0	NaN
Nearest Neighbors	94.63	0.51	0	0	NaN
Random Forest	93.96	0.68	0	0	NaN
Gradient Boosting	93.96	0.73	0.14	0.25	0.18
Linear DA	93.96	0.63	0	0	NaN
AdaBoost	93.29	0.72	0.14	0.2	0.17
Naive Bayes	93.29	0.74	0.14	0.2	0.17
Quadratic DA	93.29	0.72	0.14	0.2	0.17
Decision Tree	89.93	0.54	0.14	0.1	0.12

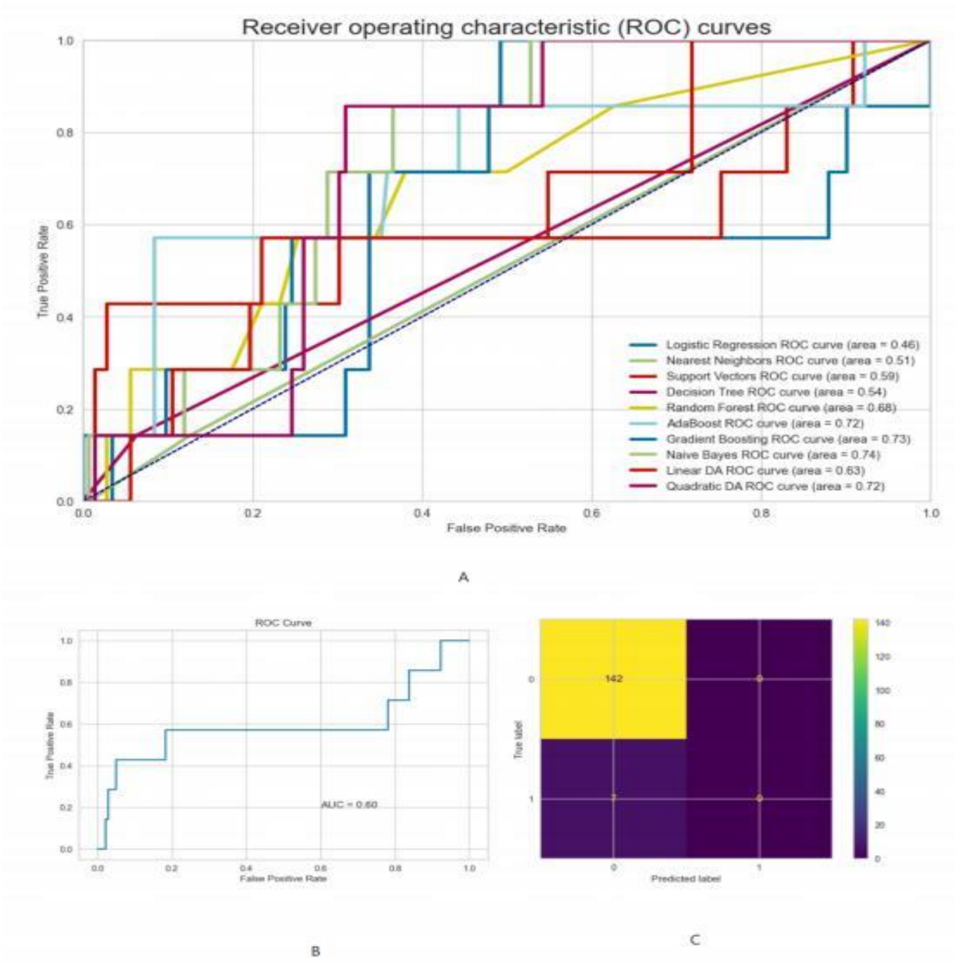


Figure 5: Model Evaluation for the Outcome Variable "Mortality"

(A) ROC curves often models; (B) ROC curves of the optimal models after parameter tuning; (C) Confusion matrices of the optimal models after parameter tuning.

2.4 Model Interpretation and Risk Factor Prediction

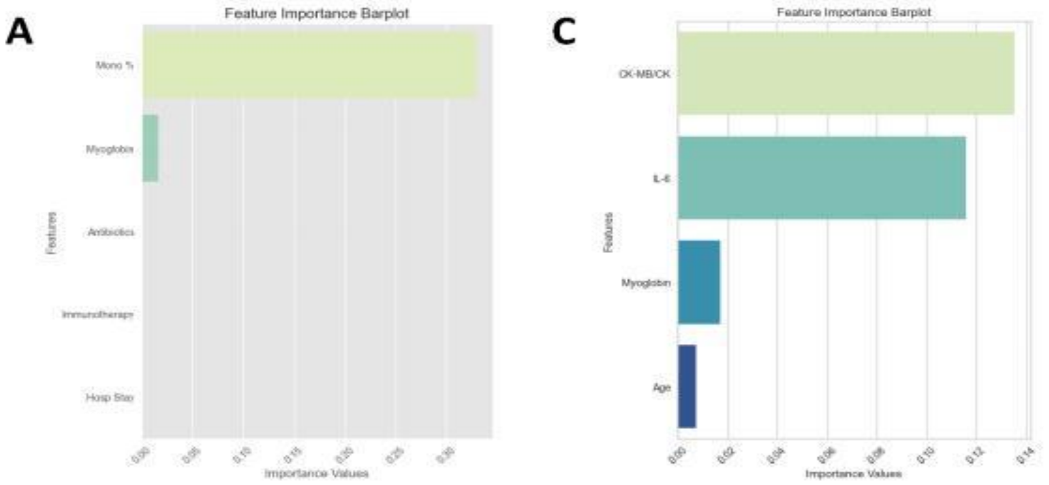
Shapley feature selection is a method for filtering features based on the contribution of each feature measured by SHAP values. By calculating SHAP values, key features that significantly influence the outcome variable can be identified, thereby enhancing the interpretability and reliability of the model[15]. The higher the SHAP value of a feature, the greater the risk of disease progression inpatients with that feature.

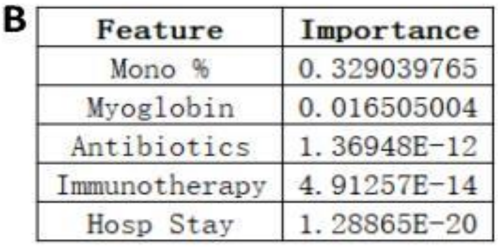
Feature weight importance measures the degree of importance of a feature within a machine learning model, allowing for an understanding of the contribution and relative significance of the feature in model predictions, and thus selecting the most critical features for model prediction[16] .

Model interpretation employs the aforementioned two methods and ANOVAto analyze the risk factors for each outcome variable.

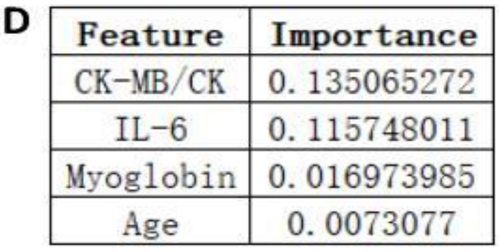
ANOVA is used to evaluate the importance of features for the outcome variables "presence of fungal infection" and "death." As shown in Figures 6A and 6B, the percentage of monocytes is the most significant risk factor for the occurrence of fungal infections in COVID- 19 patients. As depicted in Figures 6C and 6D, the ratio of CK-MB to CK and interleukin-6 are closely related to the risk of patient mortality.

Shapley feature selection and feature weight importance analysis are utilized for the predictive analysis of risk factors for the outcome variable "hospital stay exceeding 30 days." As shown in Figure 7, fungal infection is the most significant risk factor leading to a hospital stay exceeding 30 days for COVID- 19 patients, while the use of immunotherapy and antibiotics also significantly affects the length of hospital stay for patients.



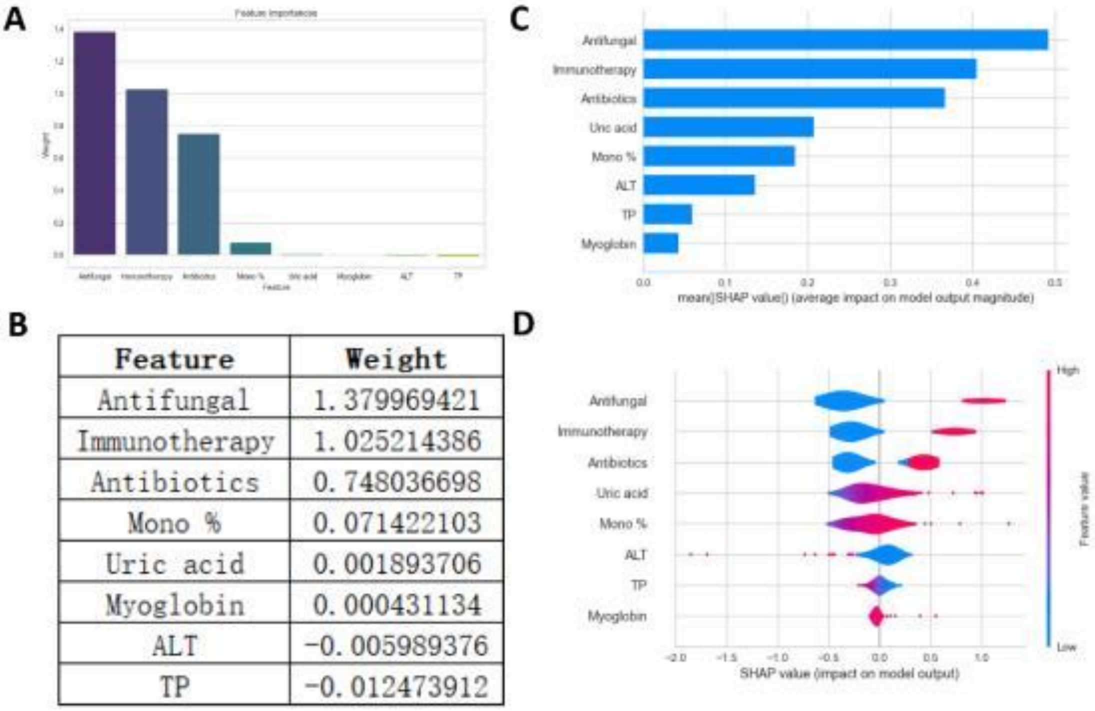

B

Feature	Importance
Mono %	0.329039765
Myoglobin	0.016505004
Antibiotics	1.36948E-12
Immunotherapy	4.91257E-14
Hosp Stay	1.28865E-20


D

Feature	Importance
CK-MB/CK	0.135065272
IL-6	0.115748011
Myoglobin	0.016973985
Age	0.0073077

Figure 6: Model Interpretation and ANOVA Feature Importance Analysis (A) (B) Feature Importance Analysis of the Outcome Variable “Presence or Absence of Fungal Infection”; (C) (D) Feature Importance Analysis of the Outcome Variable “Death or Survival”. Mono%: Monocyte Percentage; Myoglobin: Myoglobin Levels; Antibiotics: Antibiotic Administration; Immunotherapy: Immune Therapy; Hosp Stay: Hospital Stay Duration; IL-6: Interleukin-6.



Feature	Weight
Antifungal	1.379969421
Immunotherapy	1.025214386
Antibiotics	0.748036698
Mono %	0.071422103
Uric acid	0.001893706
Myoglobin	0.000431134
ALT	-0.005989376
TP	-0.012473912

Figure 7: Model Interpretation for "Hospital Stay Exceeding 30 Days"
(A) (B) Analysis of Feature Weight Importance; (C) (D) Shapley Feature Selection. Mean(|SHAP value|): SHAP Value; Antifungal: Utilization of Antifungal Medication, indicating the presence of fungal infection; Uric Acid: Uric Acid Levels; ALT: Alanine Aminotransferase; TP: Total Protein.

3 DISCUSSION

Machine learning has demonstrated significant potential in predicting the risk factors for diseases, and by establishing models for early disease prediction, it can improve patient prognosis and reduce the incidence of adverse outcomes. Current research indicates that machine learning models are capable of predicting the risk of specific mortality or the need for ventilator use in COVID- 19 patients[17]. Moreover, a study on ICU patients found that machine learning models can accurately predict the prognosis of COVID- 19 patients in the ICU[18] .

This study, based on machine learning models, has achieved the prediction of risk factors for COVID- 19 patients and identified risk factors for three important outcome variables: monocyte percentage, CK-MB/CK ratio, IL-6, fungal infection, immunotherapy, and antibiotic use. A substantial amount of research has confirmed that the aforementioned clinical characteristics are significant factors affecting the prognosis of COVID- 19 patients.

Among these, research has shown that macrophages derived from monocytes can influence the severity of COVID- 19 by regulating gene expression[19]. The findings of this study reveal that the most significant risk factor for the occurrence of fungal infections in COVID- 19 patients is the monocyte percentage.

A retrospective study on the elevation of cardiac biomarkers in severe COVID- 19 patients found that CK can predict the prognosis of COVID- 19 patients[20]. In addition, studies have shown that the level of IL-6 upon hospital admission can predict the risk of disease progression in severe COVID- 19 patients[21]. These findings are consistent with CK-MB/CK and IL-6 as risk factors affecting the mortality risk of COVID- 19 patients in this study.

Research on the impact of fungal infections, immunotherapy, and antibiotic use on the prognosis of COVID- 19 patients is diverse. Some studies have found that severe COVID- 19 patients often experience fungal infections, and secondary fungal infections can lead to high mortality rates in COVID- 19 patients[22,23]. Other studies have indicated that early administration and adequate dosage of passive antibody therapy before hospitalization are key to effectively preventing clinical progression in COVID- 19 patients[24]. A retrospective study found that the prophylactic use of antibiotics may increase the incidence of multidrug-resistant bacterial colonization[25] .

Ten machine learning algorithms were used to establish a predictive model, and the model's predictive results can be confirmed by existing research. Furthermore, by comprehensively analyzing laboratory test data, imaging materials, and patient medical history data, a large number of clinical characteristics were extracted and analyzed. This helps to bridge the gap between reliable and practical predictive models, thereby enhancing the accuracy of the predictions[4] .

However, there are some limitations to this study. Firstly, the data collected for the study were retrospective data from a single center, which carries the risk of selection bias. A multicenter prospective study could mitigate this deficiency, thereby making the study results more reliable and generalizable. Secondly, although characteristics such as high-sensitivity C-reactive protein (hs-CRP) and the number of vaccine doses may affect the prognosis of patients with COVID- 19, due to limitations in data collection, they were not included as predictive variables [26,27]. Future research could further explore the impact of these characteristics on the progression of the disease inpatients with COVID- 19. In addition, the adverse outcomes for patients with COVID- 19 are quite broad, and this study only investigated the risk factors for three main outcome variables. Researchers could establish predictive models for other outcome variables to more accurately predict the risk factors for COVID- 19 and thus improve patient prognosis.

4 SUMMARY

The percentage of monocytes, CK-MB/CK ratio, IL-6 levels, fungal infections, immunotherapy, and antibiotic use are all risk factors that impact the prognosis of patients with COVID- 19. Among these, the percentage of monocytes is closely related to the risk of secondary fungal infections in patients with COVID- 19; the CK-MB/CK ratio and the level of interleukin-6 (IL-6) can affect the mortality rate of patients; fungal infections, immunotherapy, and the use of antibiotics affect the length of hospital stay for patients to varying degrees.

REFERENCES

[1] World Health Organization. WHO coronavirus (COVID-19) dashboard overview[EB/OL].[EB/OL]//datadot. [2024-04-11]. http://data.who.int/dashboards/covid19/cases.

[2] ZHOU C, JIANG Y, SUN L, 等. Secondary pulmonary infection and co-infection in elderly COVID-19 patients during the pandemics in a tertiary general hospital in Beijing, China.[J/OL].

Frontiers in microbiology, 2023, 14. http://www.ncbi.nlm.nih.gov/pubmed/37901822. DOI:10.3389/fmicb.2023.1280026.

[3] SILVA RCda, DELIMASC, DOS SANTOS REIS WP M, 等. Comparison of DNA extraction methods for COVID-19 host genetics studies.[J/OL]. PloS one, 2023, 18(10). http://www.ncbi.nlm.nih.gov/pubmed/37903126. DOI:10.1371/journal.pone.0287551.

[4] SU Y, LI Y, CHEN W, 等. Automated machine learning-based model for predicting benign anastomotic strictures in patients with rectal cancer who have received anterior resection.[J/OL]. European journal of surgical oncology : the journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology, 2023, 49(12). http://www.ncbi.nlm.nih.gov/pubmed/37857102. DOI:10.1016/j.ejso.2023.107113.

[5] ZHONG L, SHI L, ZHOU L, 等. Development of a nomogram-based model combining intra- and peritumoral ultrasound radiomics with clinical features for differentiating benign from malignant in Breast Imaging Reporting and Data System category 3-5 nodules.[J/OL]. Quantitative imaging in medicine and surgery, 2023, 13(10). http://www.ncbi.nlm.nih.gov/pubmed/37869276. DOI:10.21037/qims-23-283.

[6] LIAO LD, HUBBARD AE, GUTIERREZ J P, 等. Who is most at risk of dying if infected with SARS-CoV-2? A mortality risk factor analysis using machine learning of patients with COVID-19 over time: a large population-based cohort study in Mexico.[J/OL]. BMJ open, 2023, 13(9). http://www.ncbi.nlm.nih.gov/pubmed/37739469. DOI:10.1136/bmjopen-2023-072436.

[7] 朱英. 关于印发新型冠状病毒感染诊疗方案（试行第十版） 的通知_国务院部门文件_中国政府网 [EB/OL]. [2024-04-11]. https://www.gov.cn/zhengce/zhengceku/2023-01/06/content_5735343.htm.

[8] ELVAS L B, NUNES M, FERREIRA J C, 等. AI-Driven Decision Support for Early Detection of Cardiac Events: Unveiling Patterns and Predicting Myocardial Ischemia.[J/OL]. Journal of personalized medicine, 2023, 13(9). http://www.ncbi.nlm.nih.gov/pubmed/37763188. DOI:10.3390/jpm13091421.

[9] LI S, ZHU P, CAI G, 等. Application of machine learning models in predicting insomnia severity: an integrative approach with constitution of traditional Chinese medicine.[J/OL]. Frontiers in medicine, 2023, 10. http://www.ncbi.nlm.nih.gov/pubmed/37928471. DOI:10.3389/fmed.2023.1292761.

[10] WEILER R, DIACHENKO M, JUAREZ-MARTINEZ E L, 等. Robin's Viewer: Using deep-learning predictions to assist EEG annotation.[J/OL]. Frontiers in neuroinformatics, 2022, 16. http://www.ncbi.nlm.nih.gov/pubmed/36844437. DOI:10.3389/fninf.2022.1025847.

[11] WU Q, YU J, ZHANG M, 等. Serum lipidomic profiling for liver cancer screening using surface-assisted laser desorption ionization MS and machine learning.[J/OL]. Talanta, 2023, 268. http://www.ncbi.nlm.nih.gov/pubmed/37931569. DOI:10.1016/j.talanta.2023.125371.

[12] HANGARAGI S, NIZAMPATNAM N, KALIYAPERUMAL D, 等. An evolutionary model for sleep quality analytics using fuzzy system.[J/OL]. Proceedings of the Institution of Mechanical Engineers. PartH, Journal of engineering in medicine, 2023. http://www.ncbi.nlm.nih.gov/pubmed/37667998. DOI:10.1177/09544119231195177.

[13] JIANG J, LIU X, CHENG Z, 等. Interpretable machine learning models for early prediction of acute kidney injury after cardiac surgery.[J/OL]. BMC nephrology, 2023, 24(1). http://www.ncbi.nlm.nih.gov/pubmed/37936067. DOI:10.1186/s12882-023-03324-w.

[14] FENG B, ZHANG Z, WEI Q, 等. A prediction model for neonatal necrotizing enterocolitis in preterm and very low birth weight infants.[J/OL]. Frontiers in pediatrics, 2023, 11. http://www.ncbi.nlm.nih.gov/pubmed/37920794. DOI:10.3389/fped.2023.1242978.

[15] LADBURY C, ZARINSHENASR, SEMWAL H, 等. Utilization of model-agnostic explainable artificial intelligence frameworks in oncology: a narrative review.[J/OL]. Translational

cancer research, 2022, 11(10). http://www.ncbi.nlm.nih.gov/pubmed/36388027. DOI:10.21037/tcr-22-1626.

[16] NASIRI S, VAEZIHIR A, AHMADISHALI J. Designing soil contamination monitoring network in petroleum refineries by XGBoost weighting and geostatistical facility allocation methods.[J/OL]. Environmental science and pollution research international, 2023. http://www.ncbi.nlm.nih.gov/pubmed/37910363. DOI:10.1007/s11356-023-30452-5.

[17] GIUSTEF O, HEL, LAISP, 等. Early and fair COVID-19 outcome risk assessment using robust feature selection.[J/OL]. Scientific reports, 2023, 13(1). http://www.ncbi.nlm.nih.gov/pubmed/37923795. DOI:10.1038/s41598-023-36175-4.

[18] CHIMBUNDE E, SIGWADHI L N, TAMUZI J L, 等. Machine learning algorithms for predicting determinants of COVID-19 mortality in South Africa.[J/OL]. Frontiers in artificial intelligence, 2023, 6. http://www.ncbi.nlm.nih.gov/pubmed/37899965. DOI:10.3389/frai.2023.1171256.

[19] SIMÓN-FUENTES M, RÍOSI, HERREROC, 等. MAFB shapes human monocyte-derived macrophage response to SARS-CoV-2 and controls severe COVID-19 biomarker expression.[J/OL]. JCI insight, 2023. http://www.ncbi.nlm.nih.gov/pubmed/37917179. DOI:10.1172/jci.insight.172862.

[20] KUMAR N, AHMAD S, MAHTO M, 等. Prognostic value of elevated cardiac and inflammatory biomarkers in patients with severe COVID-19: a single-center, retrospective study.[J/OL]. Emergency and critical care medicine, 2022, 2(3). http://www.ncbi.nlm.nih.gov/pubmed/37521815. DOI:10.1097/EC9.0000000000000057.

[21] ZOBEL CM, WENZEL W, KRÜGER JP, 等. Serum interleukin-6, procalcitonin, and C-reactive protein at hospital admission can identify patients at low risk for severe COVID-19 progression.[J/OL]. Frontiers in microbiology, 2023, 14. http://www.ncbi.nlm.nih.gov/pubmed/37937220. DOI:10.3389/fmicb.2023.1256210.

[22] RAMLI S R, ABDUL HADI F S, NORAMDANNA, 等. Secondary and Co-Infections in Hospitalized COVID-19 Patients: A Multicenter Cross-Sectional Study in Malaysia.[J/OL]. Antibiotics (Basel, Switzerland), 2023, 12(10). http://www.ncbi.nlm.nih.gov/pubmed/37887248. DOI:10.3390/antibiotics12101547.

[23] KUMAR D, AHMAD F, KUMAR A, 等. Risk Factors, Clinical Manifestations, and Outcomes of COVID- 19-Associated Mucormycosis and Other Opportunistic Fungal Infections.[J/OL]. Cureus, 2023, 15(9). http://www.ncbi.nlm.nih.gov/pubmed/37915866. DOI:10.7759/cureus.46289.

[24] STADLER E, CHAI K L, SCHLUB T E, 等. Determinants of passive antibody efficacy in SARS-CoV-2 infection: a systematic review and meta-analysis.[J/OL]. The Lancet. Microbe, 2023, 4(11). http://www.ncbi.nlm.nih.gov/pubmed/37924835. DOI:10.1016/S2666-5247(23)00194-5.

[25] MEMBRILLO DENOVALES F J, RAMÍREZ-OLIVENCIA G, MATA FORTE MT, 等. The Impact of Antibiotic Prophylaxis on a Retrospective Cohort of Hospitalized Patients with COVID-19 Treated with a Combination of Steroids and Tocilizumab.[J/OL]. Antibiotics (Basel, Switzerland), 2023, 12(10). http://www.ncbi.nlm.nih.gov/pubmed/37887216. DOI:10.3390/antibiotics12101515.

[26] UEDAY, YOKOGAWA N, MURATA K, 等. C-reactive protein and the neutrophil-to-lymphocyte ratio on admission predicting bacteraemia with COVID-19.[J/OL]. Annals of medicine, 2023, 55(2). http://www.ncbi.nlm.nih.gov/pubmed/37939245. DOI:10.1080/07853890.2023.2278618.

[27] CHEN Y, HU C, WANG Z, 等. Immunity Induced by Inactivated SARS-CoV-2 Vaccine: Breadth, Durability, Potency, and Specificity in a Healthcare Worker Cohort.[J/OL]. Pathogens (Basel, Switzerland), 2023, 12(10). http://www.ncbi.nlm.nih.gov/pubmed/37887770. DOI:10.3390/pathogens12101254.

4.1 ***Appendix Table：***

Feature	Test N=149	Train N=447	p.overall
Gender:			0.346
Female	47 (31.5%)	162 (36.2%)	
Male	102 (68.5%)	285 (63.8%)	
Age	73.6 (13.1)	74.3 (13.8)	0.558
BMI	22.9 (4.31)	22.8 (3.73)	0.807
Immunotherapy:			0.319
No	112 (75.2%)	315 (70.5%)	
Yes	37 (24.8%)	132 (29.5%)	
Antifungal:			0.864
No	115 (77.2%)	350 (78.3%)	
Yes	34 (22.8%)	97 (21.7%)	
Antibiotics:			0.962
No	87 (58.4%)	264 (59.1%)	
Yes	62 (41.6%)	183 (40.9%)	
Antiviral:			0.730
No	115 (77.2%)	353 (79.0%)	
Yes	34 (22.8%)	94 (21.0%)	
Hormones:			0.881
No	52 (34.9%)	151 (33.8%)	
Yes	97 (65.1%)	296 (66.2%)	
MV:			0.431
No	146 (98.0%)	430 (96.2%)	
Yes	3 (2.01%)	17 (3.80%)	
Hosp Stay	22.6 (14.8)	24.0 (16.7)	0.326
Outcome:			0.920
No	141 (94.6%)	420 (94.0%)	
Yes	8 (5.37%)	27 (6.04%)	
WBC	6.85 (3.21)	7.20 (4.42)	0.309
N	5.36 (2.87)	5.46 (3.57)	0.722
N %	76.6 (11.0)	74.4 (13.5)	0.052
L	0.91 (0.51)	0.97 (0.58)	0.240
L %	14.9 (8.72)	16.5 (10.0)	0.073
Mono	0.47 (0.25)	0.51 (0.61)	0.220
Mono %	7.21 (3.15)	7.59 (5.41)	0.296
RBC	3.78 (0.84)	3.78 (0.88)	0.945
Hb	113 (24.0)	114 (24.8)	0.630
Plt	190 (90.3)	189 (90.2)	0.895
PCT	0.94 (3.12)	0.84 (3.46)	0.757
IL-6	58.2 (150)	45.8 (72.3)	0.329
BNP	302 (573)	276 (456)	0.614
CK	117 (271)	104 (149)	0.575
CK-MB	12.5 (7.50)	12.5 (10.6)	0.960
CK-MB/CK	26.9 (22.5)	23.6 (19.4)	0.118
Myoglobin	161 (267)	154 (238)	0.776
Troponin T	93.1 (592)	54.3 (108)	0.427
D-dimer	2.49 (4.27)	2.52 (4.17)	0.945
FDP	8.47 (6.76)	8.94 (7.93)	0.480
TB	13.3 (11.0)	12.1 (6.79)	0.226

TP	61.6 (7.21)	62.3 (7.09)	0.318
Alb	32.8 (4.66)	33.0 (5.02)	0.659
Glob	28.8 (5.27)	29.3 (5.55)	0.316
ALT	31.6 (27.8)	34.1 (48.6)	0.441
AST	35.3 (31.4)	38.9 (66.7)	0.384
AST/ALT	1.45 (1.06)	1.39 (0.84)	0.558
LDH	249 (111)	298 (975)	0.304
Urea	9.74 (7.73)	9.42 (8.03)	0.670
Creatinine	166 (236)	157 (284)	0.712
Uric acid	321 (120)	321 (168)	0.994
Fever:			0.479
No	105 (70.5%)	299 (66.9%)	
Yes	44 (29.5%)	148 (33.1%)	
Cough:			0.576
No	44 (29.5%)	145 (32.4%)	
Yes	105 (70.5%)	302 (67.6%)	
Expectoration:			1.000
No	67 (45.0%)	200 (44.7%)	
Yes	82 (55.0%)	247 (55.3%)	
Dyspnea:			0.922
No	94 (63.1%)	278 (62.2%)	
Yes	55 (36.9%)	169 (37.8%)	
Chest pain:			1.000
No	143 (96.0%)	429 (96.0%)	
Yes	6 (4.03%)	18 (4.03%)	
Sore throat:			0.739
No	147 (98.7%)	438 (98.0%)	
Yes	2 (1.34%)	9 (2.01%)	
Hemoptysis:			0.643
No	147 (98.7%)	443 (99.1%)	
Yes	2 (1.34%)	4 (0.89%)	
Chest distress:			0.633
No	141 (94.6%)	416 (93.1%)	
Yes	8 (5.37%)	31 (6.94%)	
Palpitation:			1.000
No	143 (96.0%)	430 (96.2%)	
Yes	6 (4.03%)	17 (3.80%)	
Fatigue:			0.478
No	127 (85.2%)	393 (87.9%)	
Yes	22 (14.8%)	54 (12.1%)	
Neurological symptom:			0.823
No	141 (94.6%)	427 (95.5%)	
Yes	8 (5.37%)	20 (4.47%)	
Digestive symptom:			1.000
No	133 (89.3%)	401 (89.7%)	
Yes	16 (10.7%)	46 (10.3%)	
Smoking:			0.275
No	122 (81.9%)	345 (77.2%)	
Yes	27 (18.1%)	102 (22.8%)	
Current smoking status:			0.576
Current	24 (16.1%)	77 (17.2%)	

Not	124 (83.2%)	369 (82.6%)	
Unknown	1 (0.67%)	1 (0.22%)	
Smoking age	6.71 (14.6)	7.98 (15.0)	0.362
Cig/day	3.10 (7.84)	3.44 (7.21)	0.639
Vaccine doses (unknown for Unknown):			0.246
1 Doses	8 (5.37%)	27 (6.04%)	
2 Doses	15 (10.1%)	41 (9.17%)	
3 Doses	32 (21.5%)	138 (30.9%)	
Not	67 (45.0%)	176 (39.4%)	
Unknown	27 (18.1%)	65 (14.5%)	
CHD:			0.268
No	136 (91.3%)	391 (87.5%)	
Yes	13 (8.72%)	56 (12.5%)	
HTN:			0.185
No	86 (57.7%)	228 (51.0%)	
Yes	63 (42.3%)	219 (49.0%)	
PD:			0.803
No	137 (91.9%)	406 (90.8%)	
Yes	12 (8.05%)	41 (9.17%)	
DM:			0.911
No	115 (77.2%)	341 (76.3%)	
Yes	34 (22.8%)	106 (23.7%)	
KD:			0.202
No	125 (83.9%)	395 (88.4%)	
Yes	24 (16.1%)	52 (11.6%)	
BD:			0.661
No	139 (93.3%)	410 (91.7%)	
Yes	10 (6.71%)	37 (8.28%)	
LD:			0.590
No	143 (96.0%)	434 (97.1%)	
Yes	6 (4.03%)	13 (2.91%)	
Tumor:			1.000
No	130 (87.2%)	390 (87.2%)	
Yes	19 (12.8%)	57 (12.8%)	
BD.Yes:			0.909
No	143 (96.0%)	426 (95.3%)	
Yes	6 (4.03%)	21 (4.70%)	
ImmunoCompromised:			0.138
No	148 (99.3%)	432 (96.6%)	
Yes	1 (0.67%)	15 (3.36%)	
Other:			0.291
No	118 (79.2%)	373 (83.4%)	
Yes	31 (20.8%)	74 (16.6%)	
Respiratory rate (bpm)	20.0 (1.71)	20.1 (1.73)	0.648
O2 sat (SO2)	95.7 (4.65)	95.8 (3.52)	0.798
CT classification:			0.548
Level 1	62 (41.6%)	197 (44.1%)	
Level 2	62 (41.6%)	191 (42.7%)	
Level 3	25 (16.8%)	59 (13.2%)	

Corresponding Author:
Guoping Li
Affiliated Hospital of Southwest Jiaotong University/Third People's Hospital of Chengdu City (610014)
E-mail: lzlgp@163.com