BMC Infectious Diseases

RESEARCH

Open Access

Development of a risk prediction model for secondary infection in severe/critical COVID-19 patients

Yinmei Zhang[1†], Mingmei Lin[2†], Zhenchao Wu[3], Zhongyu Han[1], Liyan Cui[1*] and Jiajia Zheng[1*]

Abstract

Objective This study aimed to develop a predictive model for secondary infections in patients with severe or critical COVID-19 by analyzing clinical characteristics and laboratory indicators.

Method A total of 307 patients with severe or critical COVID-19 admitted to Peking University Third Hospital from December 2022 to February 2023 were retrospectively analyzed, including 156 patients with secondary infection and 151 patients without secondary infection. The Boruta algorithm identified significant variables, and eight machine learning models were evaluated based on area under the curve (AUC) performance. The optimal model selected was further assessed, with model interpretability provided using SHapley Additive exPlanations (SHAP).

Result Nine predictive factors were identified: Mechanical Ventilation, Procalcitonin (PCT), Interleukin-8 (IL-8), Interleukin-6 (IL-6), Blood Urea Nitrogen, Glucose, Creatine Kinase, Lactate Dehydrogenase, and Mean Platelet Volume (MPV). The random forest model demonstrated the best performance, with further evaluation showing an average AUC of 0.981 (CI 0.965–0.998) on the training set and 0.836 (CI 0.761–0.912) on the test set. SHAP analysis identified MPV, PCT, and IL-8 as the strongest predictors of secondary infections.

Conclusion We developed an effective predictive model for secondary infection risk in severe COVID-19 patients using readily available clinical parameters, enabling early clinical intervention. This machine learning approach demonstrates potential for improving patient management.

Clinical trial This study does not involve clinical trial interventions. Therefore, clinical trial registration was not applicable.

Keywords Severe/critical COVID-19, Secondary infections, Laboratory indicators, Machine learning, Risk prediction model

[†] Yinmei Zhang and Mingmei Lin contributed equally to this work.

*Correspondence:
Liyan Cui
cliyan@163.com
Jiajia Zheng
zhengjiajia@bjmu.edu.cn

[1] Department of Laboratory Medicine, Peking University Third Hospital, Haidian District, No. 49 North Garden Road, Beijing 100191, People's Republic of China
[2] Department of Obstetrics and Gynecology, Peking University Third Hospital, Haidian District, No. 49 North Garden Road, Beijing 100191, People's Republic of China
[3] Department of Respiratory and Critical Care Medicine, Peking University Third Hospital, Haidian District, No. 49 North Garden Road, Beijing 100191, People's Republic of China

Introduction

The novel coronavirus (SARS-CoV-2) causing COVID-19 has led to over 750 million infections and 6.8 million deaths worldwide. Following China's December 2022 policy adjustments, the number of patients infected with COVID-19 has significantly increased in many regions, with the Omicron variant being the predominant strain [1, 2]. 5-32% of patients infected with the COVID-19 have the risk of progressing to severe or critical conditions, characterized by acute respiratory distress syndrome, organ dysfunction, and the development of secondary infections [3].Secondary infections lead to a significant threat to the healthcare system and are associated with increased mortality rates [4, 5].

Severe/critical COVID-19 patients often experience secondary infections caused by bacteria or fungi [6]. Understanding the risk factors associated with secondary infections in these patients is crucial for early identification, timely intervention, and improving patient outcomes [7, 8]. Several Clinical character and laboratory test results have been confirmed to be related to the severity of COVID-19, such as peripheral blood leukocyte count, lymphocyte count, and levels of inflammatory factors [9, 10]. However, research on laboratory predictive indicators for secondary infections in COVID-19 remains quite limited. The current epidemiological patterns of COVID-19 show a certain complexity and volatility globally, necessitating accurate predictive tools to identify patients with secondary infections among critically COVID-19 patients.

This research undertook a retrospective examination of the demographic characteristics and laboratory parameters, encompassing routine blood tests, biochemical analyses, inflammatory markers, and cytokine levels, of 307 patients with severe or critical COVID-19 who were admitted to Peking University Third Hospital between December 2022 and February 2023. Our aim is to develop a machine learning-based predictive model for the assessment of severe or critically ill COVID-19 patients who develop secondary infections [11, 12]. To enhance clinical interpretability, SHapley Additive exPlanations (SHAP) were employed to quantify feature contributions in the optimal model [13]. By precisely identifying significant variables associated with the likelihood of secondary infections, this study offers a valuable early warning mechanism for forecasting secondary infections in patients with severe or critical COVID-19, thereby providing substantial support for timely clinical interventions and enhancing patient prognosis.

Methods

Study population

A retrospective analysis was performed on 310 adult patients with severe or critical COVID-19 pneumonia admitted to Peking University Third Hospital between December 2022 and February 2023. Laboratory confirmation of COVID-19 infection was defined as a positive polymerase chain reaction (PCR) result from nasopharyngeal swab samples or bronchoalveolar lavage fluid. Following screening, three patients were excluded due to incomplete clinical information or case data, resulting in a final cohort of 307 patients included in the study (Fig. 1).

The assessment of severe disease was based on the Diagnosis and Treatment plan for COVID-19 (Trial version 10) [14], the criteria for severe disease were any of the following in adults and could not be explained by other causes other than COVID-19 infection: (1) The presence of dyspnea with a respiratory rate (RR) ≥ 30 breaths per minute; (2) Under resting conditions, an oxygen saturation level ≤ 93% when breathing ambient air; (3) A ratio of arterial oxygen partial pressure (PaO2) to fraction of inspired oxygen (FiO2) ≤ 300 mmHg; (4) Progressive worsening of clinical symptoms with pulmonary imaging indicating a significant progression of lesions by > 50% within 24–48 h. The criteria for critical illness included patients who met at least one of the following conditions: (1) Requirement for mechanical ventilation due to respiratory failure; (2) Development of shock; (3) Presence of organ failure necessitating intensive care unit (ICU) monitoring and treatment.

Secondary infections were defined as new bacterial, fungal, or other pathogen-driven infections occurring 48 h to 14 days after hospital admission, excluding preexisting infections at admission. The criteria for diagnosing secondary infections were as follows: (1) Positive etiological culture with successful isolation of pathogenic microorganisms from relevant clinical specimens; (2) Clinical manifestations consistent with secondary infections, including respiratory tract infections (e.g., fever, cough, sputum production, wheezing), urinary tract infections (e.g., dysuria, urinary frequency/urgency), biliary infections (e.g., fever, jaundice, right upper quadrant pain), or skin/soft tissue infections (e.g., localized erythema, purulent discharge); (3) Elevated Laboratory Inflammatory Markers: Increased levels of inflammatory indicators such as Procalcitonin, C-Reactive Protein, and white blood cell count.

Laboratory parameters analysis

A retrospective analysis was conducted on the laboratory parameters of the patients. This analysis encompassed an evaluation of complete blood count parameters, including white blood cells(WBC), red blood cells(RBC), hemoglobin (Hb), mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), platelet count (PLT), red blood cell distribution

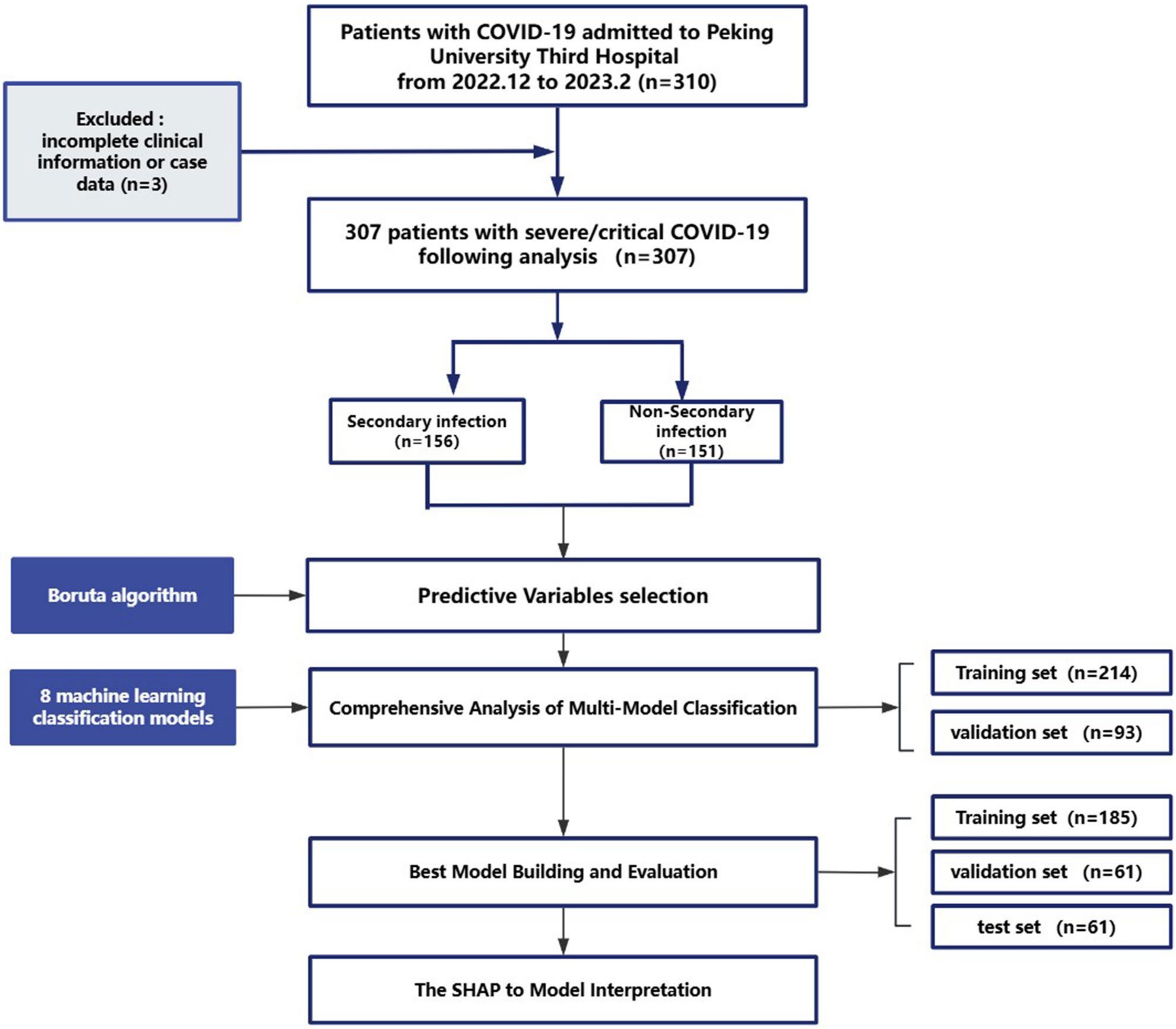
Patients with COVID-19 admitted to Peking University Third Hospital from 2022.12 to 2023.2 (n=310)
Excluded : incomplete clinical information or case data (n=3)
307 patients with severe/critical COVID-19 following analysis (n=307)
Secondary infection (n=156)
Non-Secondary infection (n=151)
Boruta algorithm
Predictive Variables selection
8 machine learning classification models
Comprehensive Analysis of Multi-Model Classification
Training set (n=214)
validation set (n=93)
Best Model Building and Evaluation
Training set (n=185)
validation set (n=61)
test set (n=61)
The SHAP to Model Interpretation

Fig. 1 The flow chart of the study

width coefficient of variation(RDW-CV), lymphocyte percentage(Lymph%), absolute lymphocyte count, neutrophil percentage(Neut%), absolute neutrophil count, eosinophil percentage (Eos%), absolute eosinophil count, basophil percentage(Baso%), absolute basophil count, monocyte percentage(Mono%), absolute monocyte count, mean platelet volume (MPV) and platelet distribution width (PDW). Additionally, serum infectious markers and biochemical analyses were performed, measuring troponin T(Tnt), ferritin, procalcitonin(PCT), C-reactive protein(CRP), lipase(LIPA), alanine aminotransferase(ALT), total protein(TP), albumin(ALB), total bilirubin(TB), alkaline phosphatase(ALP), aspartate aminotransferase(AST), creatine kinase isoenzyme MB(CK-MB), creatine kinase(CK), lactate dehydrogenase(LDH), urea, total carbon dioxide(CO2), potassium(K), sodium(Na), chloride (Cl), calcium(Ca), glucose(Glu), creatinine(Cr), Urea Nitrogen(UREA), amylase(AMY), uric acid(UA), γ-glutamyl transferase(GGT), magnesium(Mg), and cholinesterase(CHE). Furthermore, cytokine assays were conducted to assess levels of Interleukin-1β(IL-1β), interleukin-2(IL-2), interleukin-4(IL-4), interleukin-5(IL-5), interleukin-6(IL-6), interleukin-8(IL-8), interleukin-12p70(IL-12p70), interleukin-10(IL-10), interleukin-17(IL-17), interferon-alpha(IFN-α), interferon-gamma(IFN-γ), and tumor necrosis factor-alpha(TNF-α), along with the measurement of D-dimer as a clotting indicator.

Laboratory analyze instruments included Sysmex XN-10 (Japan) for complete blood counts, Beckman Automatic Biochemical Analyzer 5800 (USA) for biochemical assessments, Werfen ACL/TOP 700 (Spain) for coagulation analysis, and the BD FACS CANTOII (USA) for the detection of cytokines by flow cytometry.

Modeling and evaluation

The feature selection process was conducted using the Boruta algorithm. The cohort was divided into training (70%) and validation (30%). Several machine learning models were then constructed and applied, including Extreme Gradient Boosting (XGBoost), Logistic Regression, Light Gradient Boosting Machine (LightGBM), Random Forest, Complement Naive Bayes (CNB), Multi-Layer Perceptron (MLP), Support vector machine (SVM) and K-Nearest Neighbor (KNN). 10-fold cross-validation was applied to the training set. Nested cross-validation was used for verification.

The Receiver Operating Characteristic (ROC) curve was constructed, and the area under the curve (AUC) was calculated to assess the diagnostic accuracy or predictive performance of the selected model. Decision curve analysis (DCA) was performed to evaluate the clinical utility of the predictive model. Calibration curves were generated to assess predictive accuracy, while precision-recall (PR) curves were plotted to evaluate model performance.

The best-performing model was selected based on these analyses. The dataset was divided into a training set (60%), a validation set (20%), and a test set (20%). The selected optimal model was then validated and tested using these sets. Feature importance and model contributions were assessed using SHAP, which quantified each feature's contribution to the prediction. SHAP values for individual samples were computed to explain the model's predictions in detail.

Data preprocessing and descriptive statistics

For continuous variables, the mean and standard deviation (SD) were reported for normally distributed data, while the median and interquartile range (IQR) were used for non-normally distributed variables, Categorical variables were described using frequencies and percentages. Normality tests were conducted to assess the distribution of the data, and the appropriate descriptive statistics were applied based on the distribution characteristics. Continuous variables were inspected for outliers and adjusted via Winsorization at the 1st and 99th percentiles to retain clinically plausible extremes while minimizing undue influence on model training.

For comparisons of normally distributed continuous data between groups, Welch's t-test was used, while the Wilcoxon rank-sum test was applied to non-normally distributed variables. Categorical variables were compared using the Pearson χ^2 test. All statistical analyses were performed using R software (version 4.2.2) and MSTATA software (www.mstata.com). All hypothesis tests were two-tailed, with a significance level set at $P < 0.05$.

Table 1 Comparison of basic data between secondary infection group and non-secondary infection group

General and Clinical Characteristics	Secondary Infection		*p*-value
	No (*N* = 151)	Yes (*N* = 156)	
Age	83 (75, 88)	84 (74, 88)	0.551
Gender			0.367
Male (%)	106 (70.2%)	102 (65.4%)	
Hypertension	86 (57.0%)	97 (62.2%)	0.351
Diabetes mellitus	52 (34.4%)	72 (46.2%)	0.036
Cardiovascular Disease	60 (39.7%)	56 (35.9%)	0.488
Renal disease	14 (9.3%)	33 (21.2%)	0.004
Malignancies	10 (6.6%)	30 (19.2%)	0.001
Hormone therapy	132 (87.4%)	136 (87.2%)	0.950
Immunosuppressive therapy	39 (25.8%)	41 (26.3%)	0.928
Mechanical ventilate	7 (4.6%)	50 (32.1%)	< 0.001
Invasive manipulation	12 (7.9%)	52 (33.5%)	< 0.001

Results

Demographic and clinical characteristics

The present study included a cohort of 307 patients diagnosed with severe or critical COVID-19. Within this population, 156 patients developed secondary infections, whereas 151 patients did not. There were no statistically significant differences in age and gender between the two groups. The incidence of comorbidities such as diabetes, renal disease, malignancies, as well as the proportion of patients requiring mechanical ventilation and invasive interventions, was significantly higher in the secondary infection group compared to the non-secondary infection group (Table 1).

Comparison of laboratory parameters between patients with secondary infection and those without secondary infection in severe/critical patients

A comparative analysis of laboratory parameters was conducted between patients who developed secondary infections and those who without secondary infections. Table 2 displays the laboratory parameters statistically significant intergroup differences.

Predictive variables of secondary infection in patients with severe/critical COVID-19 infection

Variable selection was performed using the Boruta algorithm. Following the screening, nine features were identified as important: Mechanical Ventilation, Procalcitonin, Interleukin-8, Interleukin-6, Urea Nitrogen, Glucose, Creatine Kinase, Lactate Dehydrogenase and Mean Platelet Volume for model development (Fig. 2). The Variance Inflation Factor (VIF) and Tolerance values were employed to evaluate potential multicollinearity among the parameters. All VIF values were below 5, and Tolerance values exceeded 0.1, indicating a low degree of multicollinearity (Table 3).

Table 2 Laboratory parameters with statistical differences of patients between secondary infection and non-secondary infection groups

Laboratory parameters	Secondary infection		*p*-value
	No (*N* = 151)	Yes (*N* = 156)	
NT Pro-BNP(pg/mL)	1,044 (427, 2,534)	1,830 (587, 5,137)	0.005
Tnt(ng/mL)	0.04 (0.02, 0.09)	0.06 (0.03, 0.18)	0.012
D-dimer(μg/mL)	0.74 (0.37, 2.07)	1.02 (0.49, 2.67)	0.040
PCT(ng/mL)	0.12 (0.08, 0.25)	0.45 (0.16, 1.21)	< 0.001
CRP(mg/dL)	15 (7, 41)	27 (11, 60)	0.004
IL-2 (pg/mL)	1.73 (1.44, 2.14)	1.88 (1.52, 2.44)	0.010
IL-6 (pg/mL)	24 (8, 67)	45 (16, 134)	< 0.001
IL-8(pg/mL)	14 (9, 25)	26 (14, 58)	< 0.001
IFN-γ(pg/mL)	2.02 (1.55, 2.74)	2.16 (1.71, 3.13)	0.034
TNF-α(pg/mL)	1.84 (1.40, 2.58)	2.02 (1.61, 2.89)	0.030
TB(μmol/L)	10 (8, 14)	12 (9, 16)	0.007
CK(U/L)	44 (26, 84)	84 (45, 220)	< 0.001
LDH(U/L)	246 (199, 364)	306 (238, 430)	< 0.001
UREA(mmol/L)	9 (7, 13)	12 (7, 22)	< 0.001
Na(mmol/L)	138 (135, 141)	140 (137, 144)	< 0.001
Cl(mmol/L)	103.6 (100.4, 106.6)	104.1 (101.4, 108.7)	0.023
Ca(mmol/L)	2.11 (2.00, 2.19)	2.06 (1.96, 2.16)	0.037
Glu(mmol/L)	7.2 (5.1, 9.4)	7.9 (6.2, 10.2)	0.003
Cr(mmol/L)	71 (59, 105)	91 (62, 183)	0.003
WBC(10^9/L)	7.4 (5.6, 10.6)	8.8 (6.7, 12.0)	0.005
RDW-CV(%)	12.90 (12.30, 13.80)	13.25 (12.60, 14.40)	0.002
Lymph%(%)	10 (6, 17)	8 (4, 12)	0.008
Neut%(%)	84 (75, 90)	87 (81, 92)	0.003
Neut count(10^9/L)	6.2 (4.4, 9.5)	7.6 (5.7, 10.6)	0.002
MONO%(%)	5.20 (3.65, 6.70)	4.45 (2.60, 6.01)	0.030
MPV(fL)	10.60 (9.75, 11.20)	10.94 (10.30, 11.60)	0.001
PDW(fL)	12.30 (10.45, 13.90)	12.70 (11.60, 14.57)	< 0.001

Comprehensive analysis of multi-model classification

After selecting the relevant characteristic factors from all variables, multiple machine learning (ML) classification models were applied for comprehensive analysis. including Extreme Gradient Boosting (XGBoost), Logistic Regression, Light Gradient Boosting Machine (LightGBM), Random Forest, Complement Naive Bayes (CNB), Multi-Layer Perceptron (MLP), Support vector machine (SVM) and K-Nearest Neighbor (KNN).

The Receiver Operating Characteristic (ROC) curve was generated, and the area under the curve (AUC) was calculated to evaluate the diagnostic accuracy and predictive performance of the selected model (Fig. 3A-B). Additionally, calibration curves, Decision Curve Analysis (DCA), and the Precision-Recall (PR) curve were assessed to further validate model performance (Fig. 3C-F). Among the evaluated models, the Random Forest model demonstrated the highest AUC in both the training set (0.991 (95% CI 0.983–0.999)) and validation set (0.823 (95% CI 0.675–0.967)), making it the optimal choice for predicting secondary infections (Table 4).

The best model building and evaluation

The Random Forest machine learning method was used for classification, with secondary Infection as the target variable. The model includes 9 variables: Mechanical Ventilation, Procalcitonin, Interleukin-8, Interleukin-6, Urea Nitrogen, Glucose, Creatine Kinase, Lactate Dehydrogenase, and Mean Platelet Volume.

The Random Forest machine learning method analysis and 10-fold cross-validation were performed to build and evaluate the model on the training set. The results showed an average AUC of 0.981(0.965–0.998) on the training set and 0.836 (0.761–0.912) on the test set (Figs. 4A-B). These results suggest that the Random Forest machine learning method model is suitable for classification tasks in this dataset. Calibration analysis suggested a reasonable degree of reliability in the model's probabilistic predictions, as indicated by a Brier score of 0.169 in the test set. The calibration curve showed an acceptable alignment with the ideal curve, suggesting a good consistency between predicted probabilities and observed outcomes (Fig. 4C). The decision curve analysis (DCA) demonstrated good clinical utility, with the model's net benefit exceeding the "intervene all" and "intervene none" baselines across a relatively wide threshold probability range, showed its robustness for practical clinical decision-making (Fig. 4D).

The SHAP to model interpretation

To enhance the interpretability of the selected variables, we assessed feature importance using SHapley Additive exPlanations (SHAP), which quantified the contribution of each feature to the model's predictions. As shown in Fig. 5A, the ranking of the nine risk factors was determined based on their mean absolute SHAP values. The results indicated that Mean Platelet Volume, Procalcitonin, and Interleukin-8 were the most influential factors associated with secondary infections. Figure 5B illustrates the significance of various features and the direction of their influence within the model. Each feature importance line represents the contributions of all patients to the outcome. Red dots signify that the feature values have a positive impact on the model's prediction of secondary infection, whereas blue dots indicate a negative impact.

Figure 5C intuitively visualizes the SHAP values of individual samples and their impact on the model's prediction. The contribution of each feature is represented by color-coded bars, with the length of the bar indicating the magnitude of its influence on the final prediction. Red bars denote positive contributions, meaning

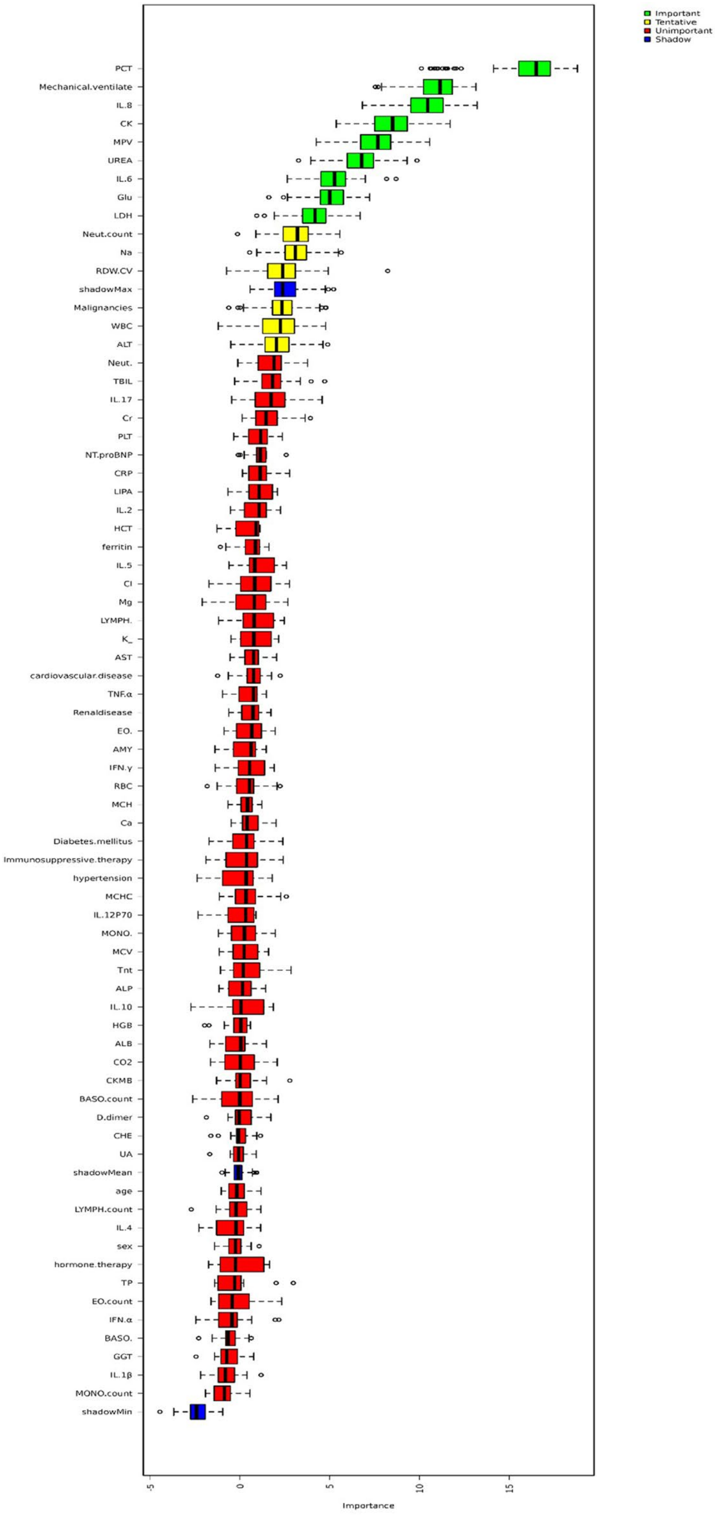

Fig. 2 Ranking of clinical variables for predicting by Boruta algorithm. The plot demonstrates boxplot of important attributes in green, tentative attributes in yellow, non-important attributes in red, and shadow attributes in blue box, respectively

Table 3 Collinearity statistics

	VIF	Tolerance
Mechanical.ventilate	1.039862	0.9616660
PCT	1.101437	0.9079049
IL.6	1.235436	0.8094306
IL.8	1.400664	0.7139470
CK	1.205617	0.8294508
LDH	1.239101	0.8070370
Glu	1.119566	0.8932029
MPV	1.021273	0.9791698
UREA	1.377652	0.7258728

the feature increases the predicted value, while blue bars represent negative contributions, indicating a decrease in the predicted value. We present an example in which a patient's laboratory and clinical feature results are as follows. Using the random forest method, the predicted probability of developing a secondary infection is 0.74.

Discussion

Since December 2022, China has experienced a significant surge in COVID-19 cases, predominantly driven by the Omicron JN.1 sublineage—a recombinant variant derived from XBB and JN.1 lineages [15, 16]. Although infection rates stabilized after March 2023, localized outbreaks persist, with SARS-CoV-2 maintaining low-level circulation as of 2024. Secondary infections in severe COVID-19 patients substantially increase mortality risks and healthcare burdens, underscoring the need for early prediction mechanisms to guide clinical interventions.

The occurrence of secondary infections can result in changes in various laboratory parameters [17]. In this

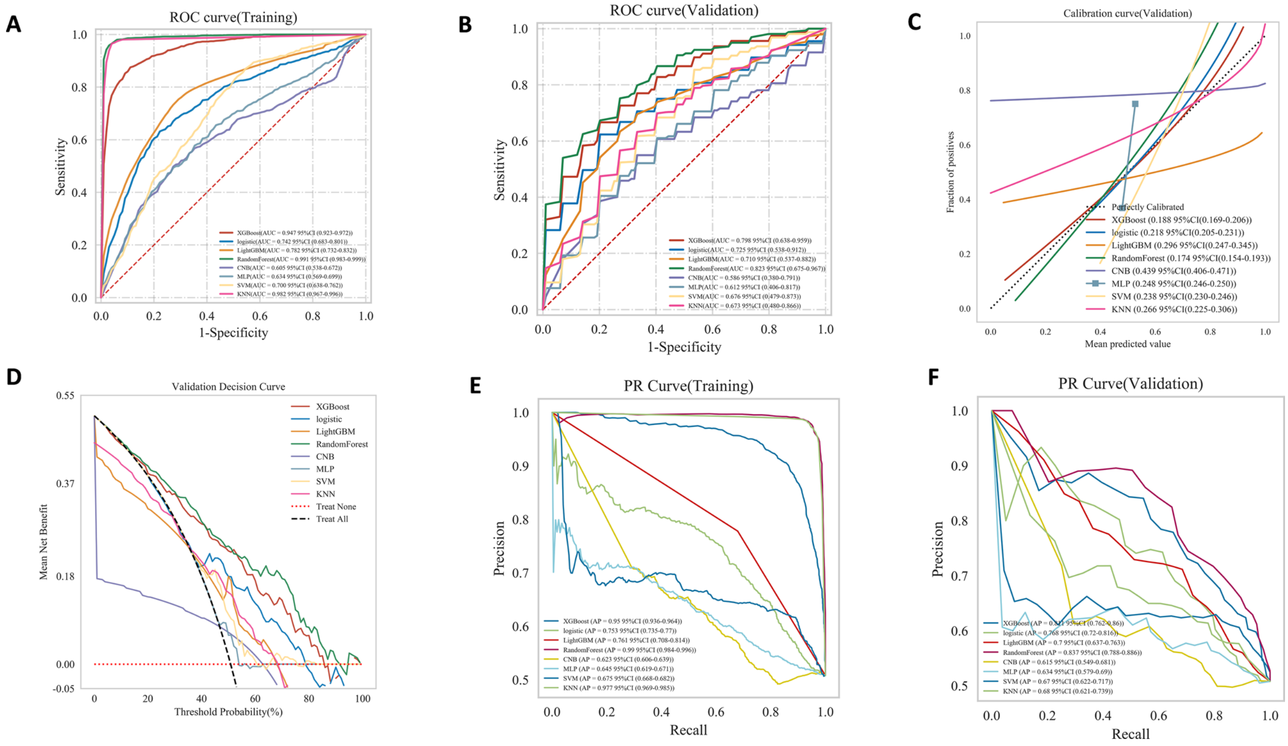

Fig. 3 Multiple machine learning model comprehensive analysis. (**A**) The ROC and AUC of training sets. (**B**) The ROC and AUC of Validation sets. (**C**) The calibration curve of the validation set, the x-axis is the average prediction probability, the case coordinate represents the actual probability of the event. The dashed diagonal line serves as the reference, and the other solid lines represent the different model fitting lines. (**D**) The DCA of the validation set. The black dashed line indicates that all patients are assumed to have a secondary infection, and the red dotted line indicates that no patients are assumed to have a secondary infection. (**E**) The PR curve and AP of Training sets. (**F**) The PR curve and AP of validation sets. The y-axis is precision and the x-axis is recall

Table 4 The AUC value of the machine learning Medels

	XGBoost	logistic	LightGBM	Random Forest	CNB	MLP	SVM	KNN
AUC of Training set	0.947 (0.923–0.972)	0.742 (0.683–0.801)	0.782 (0.732–0.832)	0.991 (0.983-0.999)	0.605 (0.538–0.672)	0.634 (0.569–0.699)	0.700 (0.638–0.762)	0.982 (0.967–0.996)
AUC of Validation set	0.798 (0.638–0.959)	0.725 (0.538–0.912)	0.710 (0.537–0.882)	0.823 (0.675-0.967)	0.586 (0.380–0.791)	0.612 (0.406–0.817)	0.676 (0.479–0.873)	0.673 (0.480–0.866)

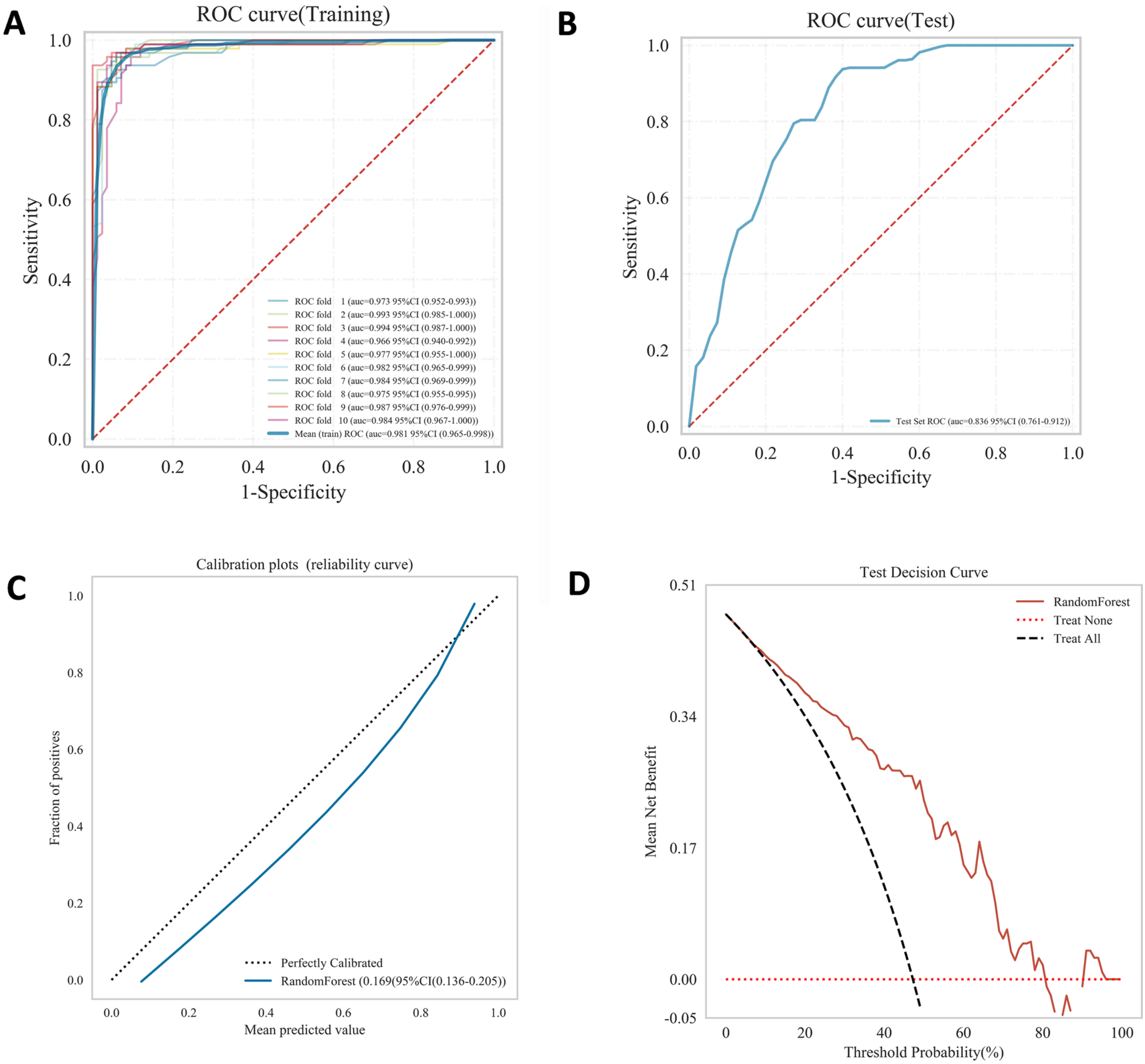

Fig. 4 The evaluation of random forest models. (**A**) The AUC of Training sets. The different colored solid lines represent 10 different results. (**B**) The AUC of Testing set. (**C**) Calibration plot of testing set. (**D**) Decision curve analysis of testing set

study, significant statistical differences were observed between the secondary infection cohort and the non-secondary infection cohort across several infection-related markers, including PCT, CRP, interleukins (IL-2, IL-6, IL-8), IFN-γ, as well as indicators of renal function like serum creatinine, urea, and calcium levels. Furthermore, noteworthy variations were also observed in hematological parameters, including white blood cell (WBC) count, lymphocyte percentage, neutrophil percentage, and absolute neutrophil count.

Feature selection for model construction was performed using the Boruta algorithm. This algorithm evaluates the importance of each variable using a random forest and compares it against randomly generated "shadow" variables, ensuring that the selected variables are genuinely informative which could handle complex, high-dimensional datasets effectively, minimizing the risk of overfitting [18]. Boruta screening revealed a strong correlation between secondary infections and several variables, including mechanical ventilation, procalcitonin, interleukin-8, interleukin-6, urea nitrogen, glucose, creatine kinase, lactate dehydrogenase, and mean platelet volume.

Prior studies have highlighted that severe COVID-19 patients requiring mechanical ventilation face an elevated risk of nosocomial secondary infections due to airway barrier disruption, bacterial colonization, mucus accumulation, and ventilator-associated lung injury,

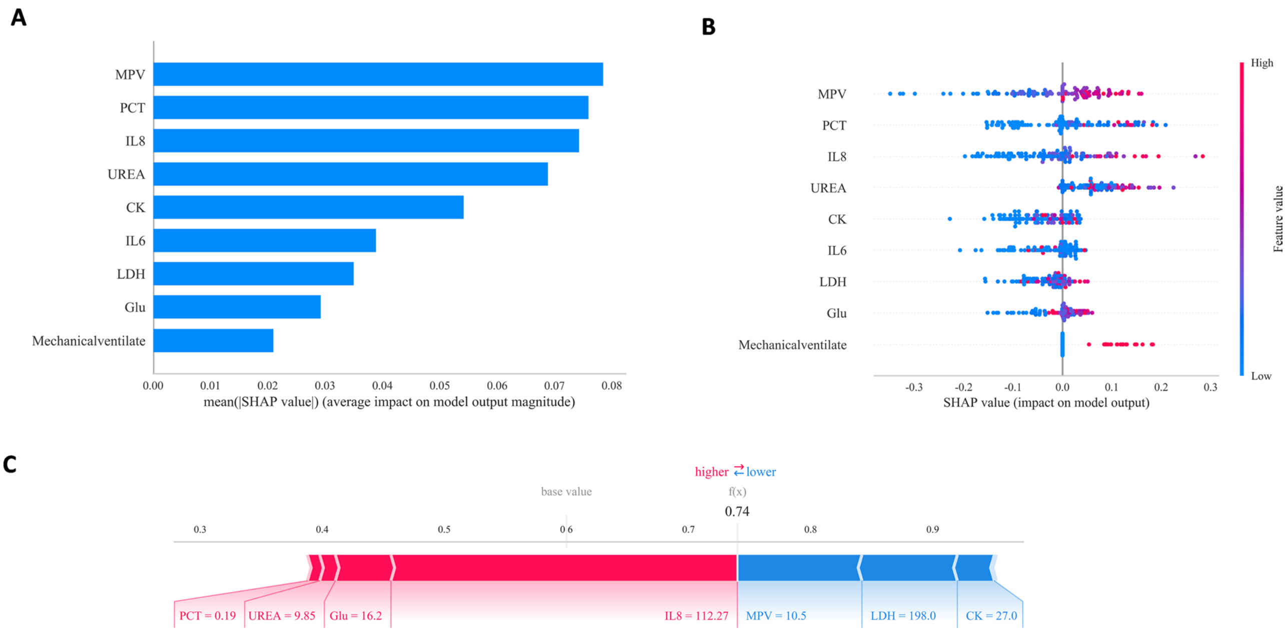

Fig. 5 SHAP interprets the model. (**A**) Feature importance ranking as indicated by SHAP. The matrix diagram describes the importance of each covariate in the development of the final prediction model. (**B**) Summary Plot illustrates each feature importance line represents the contributions of all patients to the outcome. (**C**) The force plot illustrates an example of a personalized prediction for a patient

compounded by exposure to multidrug-resistant organisms in clinical settings [19, 20]. Prior research has extensively studied several biomarkers linked to secondary infections in COVID-19 patients, emphasizing their clinical significance. Procalcitonin (PCT), a widely recognized marker for bacterial infections, is significantly elevated in patients with severe secondary infections. Elevated PCT levels often indicate bacterial co-infections and correlated with worse outcomes in critically ill COVID-19 patients [21]. Similarly, pro-inflammatory cytokines such as Interleukin-6 (IL-6) and Interleukin-8 (IL-8) are markedly upregulated in severe disease, driving cytokine storms that exacerbate tissue damage and immune dysregulation [22, 23].

Mean Platelet Volume (MPV), an indicator of platelet activation and inflammation, is elevated in severe COVID-19 cases and linked to secondary infections. The systemic inflammatory response in severe infections drives the production of larger, reactive platelets, contributing to endothelial damage and immune dysregulation, fostering susceptibility to secondary infections [24]. Additionally, increased levels of Lactate Dehydrogenase (LDH) and Creatine Kinase (CK) reflect extensive tissue damage and systemic inflammation, serving as markers of disease severity in severe COVID-19 cases [25, 26]. These biomarkers collectively underscore the interplay between hyperinflammation, tissue injury, and immune dysfunction in predisposing mechanically ventilated patients to secondary infections.

In addition to the 9 key parameters identified in this study, prior research has demonstrated that other biomarkers—including C-reactive protein (CRP), white blood cell (WBC) count, and neutrophil percentage—along with clinical characteristics such as diabetes mellitus, renal failure, and malignancies, have been linked to secondary infections in severe COVID-19 cases [27, 28]. Although these measures showed associations with secondary infections in univariate logistic regression analyses, they were excluded from the final predictive model. This exclusion could be attributed to the Boruta feature selection algorithm, which prioritizes variables based on their predictive relevance, coupled with potential redundancy arising from high correlations between these excluded parameters and those retained in the model.

We ultimately employed the random forest algorithm for model development. This method operates by constructing multiple decision trees and aggregating their predictions, which effectively mitigates overfitting while enhancing predictive accuracy. A key advantage of random forest lies in its capacity to generate feature importance scores, enabling the identification of the most influential predictors in the model. Furthermore, the algorithm is highly suitable for analyzing complex, high-dimensional datasets, as it accommodates missing values, handles mixed data types, and requires minimal data preprocessing [29, 30]. Notably, since our dataset comprised complete clinical and laboratory data from hospitalized severe patients, the need for data transformation and standardization was minimized, thereby preserving the inherent biological relevance of the measurements. Compared to many other machine learning algorithms, the random forest model is less prone to overfitting and thus well-suited for clinical data analysis.

Although current epidemiological data indicate a relatively stable incidence of COVID-19, secondary infections remain a persistent threat in critical care settings, particularly among immunocompromised or mechanically ventilated patients. Notably, the biomarkers central to our model—procalcitonin (PCT), interleukin-6 (IL-6), and mechanical ventilation—are well-established predictors of nosocomial infections across diverse respiratory illnesses. This broader applicability ensures our model's relevance beyond COVID-19-specific contexts, offering a promising solution for managing secondary infections in future viral pandemics or endemic respiratory diseases.

This study has several limitations. First, both the training and internal validation datasets were derived from a single center. Consequently, the results necessitate future validation using broader and more diverse populations to enhance generalizability. To address these concerns, future studies will aim to implement multi-center external validation, thereby ensuring that our findings are robust and generalizable across different clinical settings. Second, the relatively small sample sizes for the validation and test sets might affect the stability of the performance estimates, resulting in wider confidence intervals. Moreover, the model's predictive accuracy may be influenced by unmeasured confounding factors, underscoring the need for external validation in diverse cohorts to ensure robust clinical applicability. Finally, while the current predictors demonstrate utility, the exploration of novel biomarkers or additional variables holds significant potential to further refine prediction accuracy and warrants dedicated investigation.

Acknowledgements
Not applicable.

Author contributions
Yinmei Zhang, Liyan Cui and Jiajia Zheng designed the study. Zhenchao Wu and Zhongyu Han performed dataset acquisition and quality control. Yinmei Zhang and Mingmei Lin conducted data analysis and interpretation. Yinmei Zhang drafted the manuscript. Yinmei Zhang, Mingmei Lin, Liyan Cui and Jiajia Zheng critically reviewed and revised the manuscript. All authors approved the final version.

Funding
This work was supported by National Key Research and Development Program of China (YFC2303203-02), project for Beijing Haidian District Health Development Research and Cultivation Plan (HP2024-30-501002), Innovation & Transfer Fund of Peking University Third Hospital(BYSYZHKC2022114) and Joint Laboratory of Medicine and Chemistry at Peking University Thrid Hospital(BYSYJC2023005).

Data availability
Data is provided within the manuscript. The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
The study was performed in accordance with the Helsinki Declaration and its later amendments or comparable ethical standards, and was approved by the Peking University Third Hospital Medical Science Research Ethics Committee (ethical approval no. 2023-007-02). For this retrospective study utilizing anonymized clinical data, Peking University Third Hospital Medical Science Research Ethics Committee granted a waiver for informed consent (ethical approval No. 2025-337-01).

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 7 March 2025 / Accepted: 13 May 2025
Published online: 21 May 2025

References

1. COVID-19 Forecasting Team. Past SARS-CoV-2 infection protection against re-infection: a systematic review and meta-analysis. Lancet. 2023;401:833–42. https://doi.org/10.1016/S0140-6736(22)02465-5.
2. Li B, Zhou L, Chen Z, et al. Investigation of nasal mucosal IgA responses in the population following COVID-19 Pandemic– China, September 2022–August 2023. China CDC Wkly. 2024;6(15):312–7. https://doi.org/10.46234/ccdcw2024.060.
3. Li H, Liu L, Zhang D, et al. SARS-CoV-2 and viral sepsis: observations and hypotheses. Lancet. 2020;395(10235):1517–20. https://doi.org/10.1016/S0140-6736(20)30920-X.
4. Fominskiy EV, Scandroglio AM, Monti G, et al. Prevalence, characteristics, risk factors, and outcomes of invasively ventilated COVID-19 patients with acute kidney injury and renal replacement therapy. Blood Purif. 2021;50(1):102–9. https://doi.org/10.1159/000513305.
5. Rawson TM, Moore LSP, Zhu N, et al. Bacterial and fungal coinfection in individuals with coronavirus: A rapid review to support COVID-19 antimicrobial prescribing. Clin Infect Dis. 2020;71(9):2459–68. https://doi.org/10.1093/cid/ciaa530.
6. Murgia F, Fiamma M, Serra S, et al. The impact of secondary infections in COVID-19 critically ill patients. J Infect. 2022;84(6):e116–7. https://doi.org/10.1016/j.jinf.2021.11.011.
7. Krumbein H, Kümmel LS, Fragkou PC, et al. Respiratory viral co-infections in patients with COVID-19 and associated outcomes: A systematic review and meta-analysis. Rev Med Virol. 2023;33(1):e2365. https://doi.org/10.1002/rmv.2365.
8. Zhou F, Yu T, Du R, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet. 2020;395(10229):1054–62. https://doi.org/10.1016/S0140-6736(20)30566-3.
9. Galli F, Bindo F, Motos A, et al. Procalcitonin and C-reactive protein to rule out early bacterial coinfection in COVID-19 critically ill patients. Intensive Care Med. 2023;49(8):934–45. https://doi.org/10.1007/s00134-023-07033-8.
10. Saqib MA, Siddiqui S, Qasim M, et al. Effect of COVID-19 lockdown on patients with chronic diseases. Diabetes Metab Syndr Clin Res Rev. 2020;14(6):1621–3. https://doi.org/10.1016/j.dsx.2020.08.028.
11. Ehrmann DE, Joshi S, Goodfellow SD, et al. Making machine learning matter to clinicians: model actionability in medical decision-making. NPJ Digit Med. 2023;6(1):7. https://doi.org/10.1038/s41746-023-00753-7.
12. Tai AMY, Albuquerque A, Carmona NE, et al. Machine learning and big data: implications for disease modeling and therapeutic discovery in psychiatry. Artif Intell Med. 2019;99:101704. https://doi.org/10.1016/j.artmed.2019.101704.
13. Cao S, Hu Y. Creating machine learning models that interpretably link systemic inflammatory index, sex steroid hormones, and dietary antioxidants to identify gout using the SHAP method. Front Immunol. 2024;15:1367340. https://doi.org/10.3389/fimmu.2024.1367340.
14. National Health Commission of the People's Republic of China. Diagnosis and treatment plan for COVID-19 (trial version 10). Chin J Clin Infect Dis. 2023;16(1):1–9.

15. He X, Yu J, Jiang J, et al. Heterogeneous hybrid immunity against Omicron variant JN.1 at 11 months following breakthrough infection. Signal Transduct Target Ther. 2024;9(1):180. https://doi.org/10.1038/s41392-024-01774-8.
16. Planas D, Staropoli I, Michel V, et al. Distinct evolution of SARS-CoV-2 Omicron XBB and BA.2.86/JN.1 lineages combining increased fitness and antibody evasion. Nat Commun. 2024;15(1):2254. https://doi.org/10.1038/s41467-024-46494-3.
17. Nevzorov I, Tulamo R, Albäck A, Lassila R. COVID-19 and SIC (!). J Vasc Surg. 2020;72(3):1148–50. https://doi.org/10.1016/j.jvs.2020.04.483.
18. Maeda-Gutiérrez V, Galván-Tejada CE, Galván-Tejada JI, et al. Evaluating feature selection methods for accurate diagnosis of diabetic kidney disease. Biomedicines. 2024;12(12):2858. https://doi.org/10.3390/biomedicines12122858.
19. Luyt CE, Girardis M, Paixão P. Herpes simplex virus and cytomegalovirus lung reactivations in severe COVID-19 patients: to treat or not to treat? That is (still) the question. Intensive Care Med. 2024;50:1317–9. https://doi.org/10.1007/s00134-024-07396-6.
20. Papazian L, Klompas M, Luyt CE. Ventilator-associated pneumonia in adults: a narrative review. Intensive Care Med. 2020;46(5):888–906. https://doi.org/10.1007/s00134-020-05980-0.
21. Chen Y, Geng Y, Xu X, et al. The features comparison between patients in the ICU and general wards and between patients with different outcomes: a 2020 COVID-19 study. Ann Palliat Med. 2021;10(1):672–80. https://doi.org/10.21037/apm-21-25.
22. Bohn MK, Lippi G, Horvath A, et al. Molecular, serological, and biochemical diagnosis and monitoring of COVID-19: IFCC taskforce evaluation of the latest evidence. Clin Chem Lab Med. 2020;58(7):1037–52. https://doi.org/10.1515/cclm-2020-0722.
23. Schmidt W, Pawlak-Buś K, Jóźwiak B, Leszczyński P. Identification of clinical response predictors of Tocilizumab treatment in patients with severe COVID-19 based on Single-Center experience. J Clin Med. 2023;12(6):2429. https://doi.org/10.3390/jcm12062429.
24. Yasseen BA, Elkhodiry AA, El-Messiery RM, et al. Platelets' morphology, metabolic profile, exocytosis, and heterotypic aggregation with leukocytes in relation to severity and mortality of COVID-19 patients. Front Immunol. 2022;13:1022401. https://doi.org/10.3389/fimmu.2022.1022401.
25. Orsucci D, Trezzi M, Anichini R, et al. Increased creatine kinase May predict A worse COVID-19 outcome. J Clin Med. 2021;10(8):1734. https://doi.org/10.3390/jcm10081734.
26. Kojima K, Yoon H, Okishio K, Tsuyuguchi K. Increased lactate dehydrogenase reflects the progression of COVID-19 pneumonia on chest computed tomography and predicts subsequent severe disease. Sci Rep. 2023;13(1):1012. https://doi.org/10.1038/s41598-023-28201-2.
27. De Bruyn A, Verellen S, Bruckers L, et al. Secondary infection in COVID-19 critically ill patients: a retrospective single-center evaluation. BMC Infect Dis. 2022;22(1):207. https://doi.org/10.1186/s12879-022-07206-8.
28. Chen T, Wu D, Chen H, et al. Clinical characteristics of 113 deceased patients with coronavirus disease 2019: retrospective study. BMJ. 2020;368:m1091. https://doi.org/10.1136/bmj.m1091.
29. Hu J, Szymczak S. A review on longitudinal data analysis with random forest. Brief Bioinform. 2023;24(2):bbad002. https://doi.org/10.1093/bib/bbad002.
30. Speiser JL, Miller ME, Tooze J, Ip E. A comparison of random forest variable selection methods for classification prediction modeling. Expert Syst Appl. 2019;134:93–101. https://doi.org/10.1016/j.eswa.2019.05.028.

Publisher's note