Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Development and Validation of machine learning models for early prediction of severe Respiratory Syncytial Virus pneumonia in children

Nan Huang
the First People's Hospital of Zunyi (The Third Affiliated Hospital of Zunyi Medical University)
Danni Zheng
the First People's Hospital of Zunyi (The Third Affiliated Hospital of Zunyi Medical University)
Kai Zhu
the First People's Hospital of Zunyi (The Third Affiliated Hospital of Zunyi Medical University)
Min He
the First People's Hospital of Zunyi (The Third Affiliated Hospital of Zunyi Medical University)
Xiaoli Yang
the First People's Hospital of Zunyi (The Third Affiliated Hospital of Zunyi Medical University)
Jie Jiang
the First People's Hospital of Zunyi (The Third Affiliated Hospital of Zunyi Medical University)
Shengyi Wang
the First People's Hospital of Zunyi (The Third Affiliated Hospital of Zunyi Medical University)
Yunrong Li
the First People's Hospital of Zunyi (The Third Affiliated Hospital of Zunyi Medical University)
Xiang Zhang
139852425230163.com

the First People's Hospital of Zunyi (The Third Affiliated Hospital of Zunyi Medical University)

Research Article

Keywords: Respiratory Syncytial Virus, severe RSV pneumonia, machine learning, random forest, SHAP method

Posted Date: September 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10960901/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Respiratory Syncytial Virus (RSV) is a leading cause of severe respiratory infections, particularly pneumonia, in children, especially those under the age of five. The incidence of severe RSV pneumonia has risen in recent years, underscoring the need for an effective predictive tool. Therefore, the development and validation of a model to predict the severity of RSV pneumonia in pediatric patients is essential.

Methods

A retrospective analysis was conducted involving 180 children with confirmed RSV pneumonia, categorized into severe (n = 62) and non-severe (n = 118) groups. Clinical, laboratory, and imaging data were collected at admission. Features selection was performed using LASSO regression and the Boruta algorithm. Subsequently, eight machine learning models were developed, including Logistic Regression, Multilayer Perceptron, Support Vector Machine, Decision Tree, Random Forest (RF), AdaBoost, and Extreme Gradient Boosting. Model performance was evaluated using standard metrics, such as area under the curve (AUC), sensitivity, specificity, accuracy, and F1 score. Additionally, Decision Curve Analysis (DCA) was employed to assess the clinical utility of the models. SHAP analysis was performed to evaluate model interpretation.

Results

Seven features were identified for model development. Among the models, RF exhibited the best performance, with an AUC of 0.853, sensitivity of 0.895, specificity of 0.714, accuracy of 0.778, and F1-score of 0.739 in the testing cohort. DCA demonstrated that the RF model provided a significant net benefit, indicating its clinical utility for early intervention. SHAP analysis revealed that respiratory rate and shortness of breath were the most influential features in predicting severe RSV pneumonia.

Conclusion

The RF model demonstrated strong predictive performance for identifying severe RSV pneumonia in children, providing a novel approach for early detection and timely intervention. Notably, this study is the first to apply machine learning for predicting severe RSV pneumonia in pediatric patients using clinical data collected at admission.

1. Introduction

Respiratory Syncytial Virus (RSV) is a highly contagious, single-stranded negative-sense RNA virus. As a major respiratory pathogen, it is a leading cause of acute lower respiratory infections (ALRIs), particularly pneumonia, and represents a significant source of morbidity and mortality in infants and children[1,2], especially those aged between two months and five years[3]. It is estimated that there were 33 million RSV-associated ALRI cases worldwide among children under five, with 3.6 million requiring hospitalization[4]. Although most infected children present with mild symptoms initially, some may rapidly progress to severer pneumonia, and even in the most critical cases, it can be life-threatening[5]. RSV is responsible for one in every 50 deaths in children aged 0–5 years, with higher mortality rates—one in every 28 deaths—in infants aged 28 days to 6 months[4]. Pneumonia caused by RSV infection imposes high burden on developing countries, where the majority of hospitalizations and deaths occur[4,6]. Therefore, the early identification of high-risk children with severe RSV pneumonia is of great clinical significance for guiding monitoring, optimizing treatment, and improving prognosis.

Previous studies have used binary logistic regression models[7] or generalized linear regression models[8] to analyze the risk factors associated with severe RSV pneumonia. However, these models, which assume linear relationships between the variables, are limited in their ability to capture the complex, multidimensional interactions of clinical features. Consequently, their predictive performance and interpretability require further improvement. In recent years, machine learning (ML) models have been widely applied in the medical field and have demonstrated excellent performance in disease prediction[9,10]. For instance, Xie and colleagues developed a prediction model using ML algorithms to accurately identify severe Mycoplasma pneumoniae pneumonia in children[11]. Additionally, ML has also been used to develop prognostic models for severe pneumonia to predict mortality risk in the ICU[12]. Despite these advances, the use of ML models for predicting severe RSV pneumonia in children remains limited.

Hence, we conducted a retrospective study using ML algorithms to develop an early diagnostic prediction model for severe RSV pneumonia in children. This model aims to facilitate the early identification of severe cases, guide clinical intervention and reduce the length of hospital stay.

2. Methods

2.1 Subjects

We performed a retrospective analysis of RSV infection in children with pneumonia. All subjects were admitted to the Department of Pediatrics at Zunyi First People's Hospital between September 2024 and August 2025. Patients were included if they fulfilled the following criteria: 1) aged 0–60 months; 2) presence of clinical manifestations of infection, such as fever, cough, and moist rales; 3) imaging findings consistent with pneumonia[13]; 4) diagnosed with RSV infection based on the PT-PCR and/or immunofluorescence methods according to the Chinese pediatric guideline for diagnosis[14]. Exclusion criteria included:

1) those diagnosed with immunodeficiency, severe metabolic diseases, or major organ dysfunction; 2) did not sign informed consent; 3) with missing data more than 20%. Ultimately, we included 180 children with RSV infection. According to the Guideline for the management of community-acquired pneumonia in children[13], the 180 patients were divided into severe cases (62 patients) and non-severe cases (118 patients).

To ensure the protection of patients' rights and adhere to ethical standards, this study was conducted in accordance with the Declaration of Helsinki and was approved by the Medical Ethics Committee of Zunyi First People's Hospital (Approval number: Ethical Review (2025)-1-739). Written informed consent was obtained from all parents or guardians.

2.2 Data collection

Demographic and clinical data were extracted from the electronic medical records. The following clinical data were extracted: 1) clinical characteristics: age, gender, weight, delivery mode, prematurity, low birth weight, breast feeding, season of onset, past medical history, clinical symptoms and signs, and hospital duration; 2) laboratory data: hemoglobin, white blood cell count, neutrophil count, lymphocyte count, monocyte count, red blood cell count, hematocrit, platelet, hematocrit, platelet, C-reactive protein (CRP), Lymphocyte percentage, blood urea nitrogen, alanine aminotransferase (ALT), aspartate aminotransferase (AST), lactic dehydrogenase (LDH), red blood cell distribution width standard-deviation (RDW-SD), red blood cell distribution width coefficient of variation (RDW-CV); 3) pulmonary imaging findings: diffuse or patch infiltrate, increased texture, small airway disease, pulmonary consolidation, and pleural effusion; 4) complications: anemia, liver damage, febrile convulsion, myocardial damage. 5) co-infection: Streptococcus pneumoniae, Staphylococcus aureus, Haemophilus influenzae, Mycoplasma pneumoniae; 6) past medical history: Neonatal birth complications, congenital heart disease, asthma history, history of pneumonia, febrile seizure history, history of wheezing. The laboratory data and imaging results were collected within 48 hours of admission.

2.3 Methods for missing value

Missing values were addressed using K-Nearest Neighbors imputation (KNN, k = 3). Specifically, missing values in continuous variables were imputed based on a distance-weighted average of the nearest neighbors, while missing values in categorical variables were imputed using a weighted voting strategy based on the continuous feature space. Euclidean distances were calculated on standardized continuous variables to identify the nearest neighbors, and the imputed category was determined by majority voting among the neighbors. All continuous variables were subsequently standardized using Z-score normalization.

2.4 Statistical analysis

All statistical analysis were performed using R studio (version 4.3.1), and Python (version 3.9). For continuous variables, the Shapiro - Wilk test was used to assess normality. Normally distributed data were expressed as mean ± standard deviation, while non-normally distributed data were presented as median (interquartile range, IQR). Categorical variables were reported as the number of cases (n) and percentages. The independent sample t-test, Mann-Whitney U test, and chi-square test were conducted to compare the distribution differences in the clinical data between the severe and non-severe groups. $P < 0.05$ was considered statistically significant.

2.5 Feature selection

The cohort was randomly divided into a training set and a testing set in a 7:3 ratio using the stratified sampling. The training set was used for feature selection and model development, while the testing set was reserved for independent validation. Feature selection followed a multi-stage, progressively restrictive procedure. Variables potentially associated with severe RSV infection were initially screened using univariate analysis with a liberal threshold ($P < 0.05$), followed by the removal of highly collinear features (Pearson $|r| > 0.90$). To enhance robustness, a dual feature selection strategy was then applied. Least Absolute Shrinkage and Selection Operator (LASSO) regression, with ten-fold cross-validation, was employed to determine the optimal penalty parameter λ for sparse feature selection. Concurrently, the Boruta algorithm was used to identify relevant features by comparing their importance against shadow features. Only variables consistently selected by both methods were retained for subsequent modeling.

2.6 Model development and evaluation

Based on features identified through the intersection of LASSO regression and the Boruta algorithm, eight ML models were developed and systematically compared, including Logistic Regression (LR), Multilayer Perceptron (MLP), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Decision Tree (DT), Random Forest (RF), AdaBoost, and Extreme Gradient Boosting (XGBoost). Within the training set, hyperparameters were optimized using RandomizedSearchCV combined with ten-fold cross-validation to balance model fitting and generalization.

Model performance was evaluated in terms of discriminative ability and overall predictive performance. The primary metrics included the area under Receiver Operating Characteristic (ROC) Curve (AUC), average precision (AP), sensitivity (also referred to as recall), specificity, Brier score, and F1 score. The 95% confidence interval (CI) for AUC was estimated using a non-parametric bootstrap procedure with 2000 resamples. Additionally, Decision Curve Analysis (DCA) was performed to evaluate the clinical utility of the models.

Given the potentially severe clinical consequences of missed diagnoses in severe pneumonia, a sensitivity-oriented threshold optimization was applied during model evaluation. The classification threshold was determined based on cross-validated predictions from the training cohort and subsequently fixed and applied to the validation cohort to assess model performance before and after threshold adjustment.

2.7 Model interpretation

Model interpretability was assessed using Shapley Additive exPlanations (SHAP). Global feature importance was visualized using SHAP summary dot (beeswarm) and bar plots. To elucidate nonlinear effects, interaction patterns, and individual-level prediction mechanisms, partial dependence plots (PDPs),

feature interaction networks, and waterfall plots were employed.

3. Results

3.1 Patient characteristics

A total of 180 children with confirmed RSV pneumonia were included in the final analysis, consisting of 118 non-severe cases (65.6%) and 62 severe cases (34.4%). The baseline clinical characteristics, laboratory test results, and imaging findings of the two groups are summarized in Table 1.

Table 1
Patients' characteristics between the severe and non-severe groups at admission

Variable	Non-severe group (n = 118)	Severe group (n = 62)	*P* value
Age (Q1, Q3), months	12.00 (7.00, 33.75)	5.00 (2.00, 8.75)	< 0.001
Weight, (Q1, Q3), kg	10.00 (8.00, 13.38)	7.20 (6.00, 10.00)	< 0.001
Sex, male, n (%)	68 (57.63%)	43 (69.35%)	0.169
Delivery mode, n (%)			1.000
Vaginal delivery	37 (31.36%)	20 (32.26%)	
Cesarean delivery	81 (68.64%)	42 (67.74%)	
Prematurity, n (%)	9 (7.63%)	10 (16.13%)	0.131
Low birth weight, n (%)	5 (4.24%)	5 (8.06%)	0.469
Breast feeding, n (%)	102 (86.44%)	50 (80.65%)	0.422
Season of onset, n (%)			0.292
Spring	45 (38.14%)	26 (41.94%)	
Summer	16 (13.56%)	12 (19.35%)	
Autumn	7 (5.93%)	6 (9.68%)	
Winter	50 (42.37%)	18 (29.03%)	
Co-infection, n (%)	73 (61.86%)	33 (53.23%)	0.337
Streptococcus pneumoniae	22 (18.64%)	14 (22.58%)	0.666
Staphylococcus aureus	12 (10.17%)	10 (16.13%)	0.357
Haemophilus influenzae	22 (18.64%)	15 (24.19%)	0.496
Mycoplasma pneumoniae	34 (28.81%)	5 (8.06%)	0.002
Clinical characteristics, n (%)			
Fever	63 (53.39%)	19 (30.65%)	0.006
Cough	105 (88.98%)	62 (100.00%)	0.016
Asthma	20 (16.95%)	27 (43.55%)	< 0.001
Shortness of breath	12 (10.17%)	33 (53.23%)	< 0.001
Rhinorrhea	41 (34.75%)	24 (38.71%)	0.717
Difficulty breathing	0 (0.00%)	2 (3.23%)	0.225
Diarrhea	13 (11.02%)	5 (8.06%)	0.714
Vomiting	14 (11.86%)	15 (24.19%)	0.054
Wheezing sounds	21 (17.80%)	25 (40.32%)	0.002
Moist rales	100 (84.75%)	59 (95.16%)	0.068
Respiratory rate (Q1, Q3), breaths/min	34.00 (28.00, 36.75)	48.50 (41.25, 52.00)	< 0.001
Increased respiratory rate[a]	8 (6.78%)	31 (50.00%)	< 0.001
Hospital duration (Q1, Q3), days	6.00 (5.25, 7.00)	7.00 (6.00, 8.00)	0.004
Laboratory test results			
Hemoglobin, M (SD), g/L	122.66 (10.37)	116.63 (10.66)	< 0.001
White blood cell count (Q1, Q3), 10^9/L	10.05 (6.98, 13.30)	9.20 (7.09, 12.12)	0.303

Note: [a]The criteria of increased respiratory rate: Neonates (0–1 month): >60 breaths/min; Infants (2–11 months): >50 breaths/min; Children (1–5 years): >40 breaths/min.

(Q1, Q3), (first quartile, third quartile); M (SD), mean (standard deviation); CRP, C-reactive protein; ALT, alanine aminotransferase; AST, aspartate aminotransferase; LDH, lactic dehydrogenase; RDW-SD, red blood cell distribution width standard-deviation; RDW-CV, red blood cell distribution width coefficient of variation.

Variable	Non-severe group (n = 118)	Severe group (n = 62)	*P* value
Neutrophil count (Q1, Q3), 10^9/L	3.60 (2.12, 6.30)	2.25 (1.51, 4.47)	0.002
Lymphocyte count (Q1, Q3), 10^9/L	3.90 (2.42, 6.28)	5.39 (3.70, 7.40)	0.016
Monocyte count (Q1, Q3), 10^9/L	0.70 (0.53, 1.06)	0.77 (0.54, 1.10)	0.726
Red blood cell count (Q1, Q3), 10^{12}/L	4.66 (4.40, 4.90)	4.50 (4.00, 4.80)	0.017
Hematocrit, M (SD), %	37.59 (3.05)	36.07 (2.61)	0.001
Platelet (Q1, Q3), 10^9/L	322.50 (253.25, 415.50)	432.00 (336.50, 521.75)	< 0.001
CRP (Q1, Q3), mg/L	2.61 (0.62, 9.64)	1.19 (0.38, 4.97)	0.078
Lymphocyte percentage (Q1, Q3), %	45.50 (30.85, 59.52)	55.67 (44.62, 71.20)	< 0.001
Blood urea nitrogen, M (SD), mmol/L	3.38 (1.14)	2.67 (0.96)	< 0.001
ALT (Q1, Q3), U/L	19.50 (15.00, 30.75)	25.50 (17.25, 40.00)	0.021
AST (Q1, Q3), U/L	43.00 (35.00, 54.00)	45.00 (36.25, 60.25)	0.445
LDH (Q1, Q3), U/L	302.00 (271.75, 344.75)	303.50 (272.00, 339.25)	0.768
RDW-SD (Q1, Q3)	38.95 (37.20, 41.20)	40.30 (38.42, 42.06)	0.014
RDW-CV (Q1, Q3)	13.10 (12.60, 13.90)	13.60 (12.72, 14.53)	0.021
Complications, n (%)	44 (37.29%)	26 (41.94%)	0.655
Anemia, n (%)	7 (5.93%)	5 (8.06%)	0.818
Liver damage	14 (11.86%)	5 (8.06%)	0.594
Febrile convulsion	0 (0.00%)	1 (1.61%)	0.743
Myocardial damage	3 (2.54%)	1 (1.61%)	1.000
Past medical history, n (%)	22 (18.64%)	17 (27.42%)	0.243
Neonatal birth complications	3 (2.54%)	2 (3.23%)	1.000
Congenital heart disease	3 (2.54%)	2 (3.23%)	1.000
Asthma history	1 (0.85%)	0 (0.00%)	1.000
History of pneumonia	7 (5.93%)	7 (11.29%)	0.326
Febrile seizure history	1 (0.85%)	3 (4.84%)	0.232
History of wheezing	2 (1.69%)	2 (3.23%)	0.897
Pulmonary imaging results, n (%)			
Diffuse or patch infiltrate	53 (44.92%)	47 (75.81%)	< 0.001
Increased texture	23 (19.49%)	10 (16.13%)	0.725
Small airway disease	2 (1.69%)	2 (3.23%)	0.897
Pulmonary consolidation	2 (1.69%)	2 (3.23%)	0.897
Pleural effusion	1 (0.85%)	0 (0.00%)	1.000

Note: [a]The criteria of increased respiratory rate: Neonates (0–1 month): >60 breaths/min; Infants (2–11 months): >50 breaths/min; Children (1–5 years): >40 breaths/min.

(Q1, Q3), (first quartile, third quartile); M (SD), mean (standard deviation); CRP, C-reactive protein; ALT, alanine aminotransferase; AST, aspartate aminotransferase; LDH, lactic dehydrogenase; RDW-SD, red blood cell distribution width standard-deviation; RDW-CV, red blood cell distribution width coefficient of variation.

This study included 59 variables, of which 23 showed significant differences in univariate analysis ($P < 0.05$). Clinically, severe cases differed significantly from non-severe cases in terms of age, weight, fever, cough, asthma, shortness of breath, wheezing sounds, increased respiratory rate, respiratory rate, and hospital duration. Laboratory tests revealed significant differences in hemoglobin, neutrophil count, lymphocyte count, red blood cell count, hematocrit, platelet, CRP, Lymphocyte percentage, blood urea nitrogen, ALT, RDW-SD, RDW-CV. Imaging findings also indicated a significant difference, with diffuse or patch infiltrate being more prevalent in the severe group compared to non-severe group ($P < 0.05$).

3.2 Feature selection

Following univariate analysis, 23 variables potentially associated with disease severity were identified ($P < 0.05$). A highly collinear variable, hematocrit, was excluded based on Pearson correlation analysis ($|r| > 0.90$) (Fig. 1A), leaving 22 candidate variables for further selection. To enhance model robustness, a dual-feature selection approach combining LASSO regression and the Boruta algorithm was applied. LASSO regression, optimized using ten-fold cross-validation, achieved a peak cross-validation area under the curve (CV-AUC) of 0.878 (Fig. 1B). The corresponding coefficient paths demonstrated progressive shrinkage toward zero as the penalty parameter increased (Fig. 1C), resulting in the selection of 15 features with non-zero coefficients. Simultaneously, the Boruta algorithm identified 10 features with consistently higher importance scores, surpassing the maximum importance of shadow features across all iterations (Fig. 1D). The relative significance of these features was further assessed using Boruta importance boxplots (Fig. 1E). By intersecting the feature sets selected by LASSO and Boruta (Fig. 1F), seven robust predictors were ultimately retained for model development: weight, respiratory rate, increased respiratory rate, shortness of breath, RDW-CV, lymphocyte percentage, and blood urea nitrogen.

3.3 Development and assessment of the machine learning model

In this study, we developed and systematically compared eight machine learning models: LR, MLP, SVM, LDA, DT, RF, AdaBoost, and XGBoost. Seven robust features identified through LASSO regression and the Boruta algorithm were selected for model development. Hyperparameters for all models were optimized using RandomizedSearchCV combined with ten-fold cross-validation to ensure an optimal balance between model fitting and generalization.

The predictive performance of eight machine learning algorithms was comprehensively evaluated in both the training and testing cohorts. The ROC curves, shown in Fig. 2A and 2B, were utilized to assess prediction model performance. Table 2 presents the performance metrics—AUC, sensitivity, specificity, accuracy, AP, F1-score, and Brier score—for each model in both the training and testing cohorts.

Table 2
Performance metrics for machine learning algorithms in training and testing cohorts.

Model type	Training cohort								Testing cohort			
	AUC (95%CI)	Sensitivity (Recall)	Specificity	Precision	Accuracy	AP	F1-score	Brier Score	AUC (95%CI)	Sensitivity (Recall)	Specificity	Precision
RF	0.958 (0.923–0.983)	0.953	0.807	0.719	0.857	0.917	0.820	0.086	0.853 (0.723–0.955)	0.895	0.714	0.630
XGBoost	0.948 (0.908–0.978)	0.860	0.879	0.787	0.873	0.888	0.822	0.089	0.838 (0.711–0.943)	0.842	0.828	0.727
SVM	0.930 (0.883–0.968)	0.767	0.916	0.825	0.865	0.852	0.795	0.107	0.827 (0.699–0.933)	0.737	0.828	0.700
DT	0.929 (0.885–0.966)	0.976	0.783	0.700	0.849	0.799	0.815	0.102	0.829 (0.716–0.927)	0.947	0.686	0.621
LogReg	0.927 (0.881–0.967)	0.837	0.843	0.735	0.841	0.845	0.783	0.108	0.806 (0.679–0.917)	0.789	0.714	0.600
LDA	0.925 (0.878–0.966)	0.767	0.916	0.825	0.865	0.848	0.795	0.113	0.833 (0.709–0.940)	0.737	0.828	0.700
AdaBoost	0.921 (0.876–0.960)	0.976	0.783	0.700	0.849	0.775	0.816	0.109	0.830 (0.706–0.923)	0.947	0.686	0.621
MLP	0.911 (0.855–0.959)	0.721	0.904	0.795	0.841	0.837	0.756	0.133	0.818 (0.675–0.935)	0.737	0.800	0.667

Note: RF, Random Forest; XGBoost, Extreme Gradient Boosting; SVM, Support Vector Machine; DT, Decision Tree; LogReg, Logistic Regression; LDA, Linear D Perceptron; AUC, the area under Receiver Operating Characteristic Curve ; AP, average precision.

Among the eight models evaluated, the RF model demonstrated exceptional performance, achieving the highest AUC value of 0.958 in the training cohort and 0.853 in testing cohort. It also exhibited high sensitivity, with values of 0.953 in the training cohort and 0.895 in the testing cohort, indicating strong ability to identify positive cases while maintaining good generalizability. Additionally, RF showed a low Brier score (0.086 in the training cohort and 0.148 in the testing cohort), reflecting its accurate probabilistic predictions and further underscoring its robustness and reliability.

DCA, depicted in Fig. 2C, provided additional support for the clinical utility of the RF model. The analysis revealed that RF consistently offered a higher net benefit across a range of threshold probabilities. This suggests that the RF model was effective in identifying high-risk patients without leading to a significant increase in unnecessary interventions, thus enabling more efficient allocation of clinical resources.

To prioritize clinical safety and minimize the risk of missing severe cases, threshold optimization was performed using a fixed-sensitivity strategy based on out-of-fold predictions from the training cohort (Fig. 2D). An optimal threshold of 0.370 was selected, which increased sensitivity to 94.7%, accurately identifying 18 out of 19 severe cases, while maintaining an acceptable specificity of 68.6% (Fig. 2E).

3.4 Interpretation of machine learning models

To enhance the clinical interpretability of ML predictive model, SHAP analysis was applied to elucidate how the RF model utilizes various clinical features in its decision-making process. The SHAP summary plot (Fig. 3A) displayed the global feature importance, with respiratory rate and shortness of breath emerging as the most influential features in predicting severe RSV pneumonia. Both features were associated with higher SHAP values, indicating that an increased respiratory rate and the presence of shortness of breath significantly elevated the model's predicted severity. The color gradient in the plot reflects the feature values, with higher values shown in red and lower values in blue. Notably, PDPs further identified non-linear risk thresholds of clinical relevance that could inform early intervention strategies. For instance, a respiratory rate exceeding approximately 37 breaths per minute and an RDW-CV greater than 13.9% were associated with a sharp increase in predicted risk, suggesting critical thresholds beyond which disease severity escalates rapidly. These quantitative cut-offs provide actionable reference points for bedside risk stratification and clinical decision-making.

Beyond individual predictors, the SHAP interaction network (Fig. 3B) revealed significant synergistic effects between variables. The most prominent interactions were observed between respiratory rate and increased respiratory rate, as well as between respiratory rate and weight, suggesting that the prognostic impact of respiratory parameters is dynamically influenced by physiological context, particularly the patient's development and body condition, rather than acting in isolation.

Finally, individual-level model interpretability was illustrated using SHAP waterfall plots (Fig. 3C). In a representative patient with a predicted probability of severe RSV pneumonia of 0.614, the model’s output was influenced by a combination of reinforcing and mitigating clinical factors. An elevated respiratory rate contributed the largest positive effect on the risk prediction (SHAP value + 0.16), followed by the presence of shortness of breath (SHAP value + 0.12). In contrast, several variables exerted protective effects, including higher body weight (SHAP value − 0.04), lower blood urea nitrogen (SHAP value − 0.03), reduced red cell distribution width (RDW-CV; SHAP value − 0.03), and a lower lymphocyte percentage (SHAP value − 0.02). These opposing contributions were aggregated additively from the baseline prediction (E[f(X)] = 0.496), resulting in the final individualized risk estimate. This example highlights the model’s ability to transparently integrate heterogeneous clinical signals and produce personalized predictions based on additive reasoning, thereby providing a clearer understanding of the decision-making process.

4. Discussion

RSV pneumonia is a leading cause of severe respiratory illness in pediatric populations, particularly among children under five years of age[3]. Globally, RSV infection remains one of the most prevalent causes of hospitalization and mortality in children[15–17], placing a significant burden on healthcare systems worldwide. Early detection of severe RSV pneumonia is therefore essential for initiating appropriate interventions and improving patient outcomes. The clinical manifestations of RSV pneumonia are highly variable, ranging from mild symptoms that resolve spontaneously to severe cases that progress rapidly, often requiring intensive care, and in some instances, leading to respiratory failure or even death[18–21]. One of the primary challenges in managing RSV pneumonia is the lack of reliable early screening tools, which complicates the differentiation between mild and severe cases at the initial point of diagnosis. Given the critical need for early identification of high-risk patients, ML methods present a promising solution to enhance diagnostic accuracy. By analyzing complex, multidimensional clinical datasets, ML models have the potential to uncover intricate patterns and interactions among risk factors that may not be immediately apparent through traditional clinical evaluation. This capability is particularly vital in resource-constrained settings, where early identification of high-risk patients can empower clinicians to optimize interventions and therapeutic strategies, ultimately improving clinical outcomes. As such, the integration of ML-based models for the prediction of severe RSV pneumonia could significantly transform early risk stratification and management strategies in pediatric care.

In this study, we utilized a robust feature selection process, incorporating LASSO regression and the Boruta algorithm, to identify seven key features associated with severe RSV pneumonia in pediatric patients under five years of age. Subsequently, we developed and evaluated eight distinct ML models to predict the risk of severe RSV pneumonia. Among the models tested, RF demonstrated superior performance, achieving an AUC of 0.853, sensitivity of 0.895, accuracy of 0.778, specificity of 0.714, precision of 0.630, F1 score of 0.739, and a Brier score of 0.148. These results highlight RF’s ability to accurately identify high-risk cases while balancing false positives. DCA further confirmed that RF provided the highest net benefit across a range of threshold probabilities, reinforcing its clinical utility for early intervention. To the best of our knowledge, this study is the first to apply ML models to predict severe RSV pneumonia in pediatric patients using clinical data collected at admission. The RF model’s strong performance suggests its potential for enhancing early diagnosis and guiding clinical decision-making in future practice.

Recently, numerous studies have explored the application of ML algorithms for disease prediction and diagnosis, including the identification of severe pneumonia. For instance, in the study by Yang et al.[22], seven key predictive factors—such as age, gender, and blood-based inflammatory biomarkers—were identified to develop six different ML models. Among these, the XGBoost model emerged as the most effective, achieving an AUC of 0.901 and an accuracy of 0.803[22], thereby demonstrating superior performance in this context. Similarly, Bing et al. [23] constructed five ML models to predict severe Mycoplasma pneumoniae pneumonia in children. Remarkably, the diagnostic model based on the LightGBM algorithm successfully distinguished severe Mycoplasma pneumoniae pneumonia from other respiratory conditions, attaining an outstanding AUC = 0.975[23]. In summary, ML models have exhibited considerable

promise in the accurate prediction of diseases and the assessment of their severity. The exceptional performance of these optimal models, marked by high accuracy, underscores their potential for clinical application, where they could play a transformative role in early diagnosis and risk assessment.

Despite significant advances in the prediction of other types of pneumonia, research focusing on RSV pneumonia remains limited. Our study addressed this gap by providing the first evidence of the role of ML models in predicting the progression to severe RSV pneumonia using easily accessible clinical data collected at admission. We demonstrate that key clinical indicators—respiratory rate, shortness of breath, increased respiratory rate, weight, RDW-CV, blood urea nitrogen, and lymphocyte percentage—can serve as reliable predictors for severe disease progression. SHAP analysis provided transparent insights into how individual variables influenced the model's predictions. Respiratory rate emerged as the dominant predictor, which is consistent with extensive clinical evidence showing that an elevated respiratory rate reflects underlying pulmonary dysfunction and impending respiratory compromise[24–26]. An increased respiratory rate is a core component of clinical severity scoring systems for pediatric pneumonia[27], where it correlates with greater work of breathing, impaired gas exchange, and an increased need for advanced respiratory support. These metrics have been validated as significant predictors of clinical deterioration in children with acute respiratory illness[28–31]. Similarly, shortness of breath, a direct manifestation of respiratory distress, was also identified as a key predictor of severe RSV pneumonia. This symptom signals the body's compensatory response to impaired gas exchange, indicating escalating respiratory compromise. Shortness of breath correlates with worsening hypoxemia and an heightened likelihood of requiring respiratory support, making it a critical marker of disease severity in pediatric pneumonia[32,33]. Weight was found to be associated with the progression to critical illness, consistent with findings by Gong et al., which showed that weight is a strong predictor of severe pneumonia[34]. Importantly, body weight serves as an indirect marker of nutritional status and immune function in pediatric populations. Low weight, often indicative of malnutrition, weakens immune defenses, impairs respiratory function, and increases susceptibility to severe RSV pneumonia[35,36]. Additionally, RDW-CV, an index of red blood cell size variability, was identified as another significant predictor. RDW has been associated with increased systemic inflammation and adverse clinical outcomes in various diseases, including severe infections such as RSV pneumonia[37,38]. Similarly, blood urea nitrogen played a role in risk stratification, potentially acting as an early indicator of renal hypoperfusion[39], muscle catabolism[40], and heightened sympathetic activation during sepsis[36]. Lymphocyte percentage was also recognized as an important predictor. This suggests that immune system dysfunction plays a critical role in the progression of RSV pneumonia to more severe stages. Previous studies have shown that a reduction in lymphocytes is associated with increased disease severity and a higher likelihood of respiratory failure in viral infections, such as influenza virus pneumonia and community-acquired pneumonia, where lymphopenia has been linked to intensive care admission and mortality risk[41,42]. In summary, our study underscores the importance of these key clinical features. By incorporating them into a predictive model, we developed an RF model with high accuracy for identifying patients at risk of severe disease progression. This model can assist clinicians in early identification, enabling timely intervention and improving patient outcomes.

Although the results of this study are promising, several limitations should be acknowledged. First, the dataset used for model development was collected from a single center, which may limit the generalizability of the findings to other populations or healthcare systems. Multi-center studies involving a larger and more diverse cohort of patients are needed to validate the model's robustness and external applicability. Second, the retrospective design of the study means that the model was trained using historical data; therefore, its prospective validation in real-world clinical settings is essential.

5. Conclusion

We have successfully developed a ML model to predict the progression of severe RSV pneumonia based on clinical data collected at admission. Seven features were selected using a combination of LASSO regression and the Boruta algorithm. Among the eight ML models evaluated in this study, the RF model exhibited the best overall performance. To further elucidate the model's predictions, the SHAP method was employed, providing transparent insights into the contribution of each feature. In summary, our study presents a novel approach for predicting the progression to severe RSV pneumonia, utilizing easily accessible clinical indicators. By leveraging these predictive features, healthcare providers can identify high-risk patients at an earlier stage, enabling timely interventions, and ultimately improving patient outcomes. Future research should focus on validating this model across diverse populations and refining it with additional biomarkers to further enhance its clinical applicability and predictive accuracy.

Abbreviations

RSV: Respiratory Syncytial Virus; ALRIs: acute lower respiratory infections; ML: machine learning; CRP: C-reactive protein; Lymphocyte percentage, ALT: alanine aminotransferase; AST: aspartate aminotransferase; LDH: lactic dehydrogenase; RDW-SD: red blood cell distribution width standard-deviation; RDW-CV: red blood cell distribution width coefficient of variation; LASSO: Least Absolute Shrinkage and Selection Operator; LR: Logistic Regression; MLP: Multilayer Perceptron; SVM: Support Vector Machine; LDA: Linear Discriminant Analysis; DT: Decision Tree; RF: Random Forest; XGBoost: Extreme Gradient Boosting; AUC: area under the curve; DCA: Decision Curve Analysis; SHAP: Shapley Additive exPlanations.

Declarations

Ethics approval and consent to participate

This study was conducted in accordance with the principles of the Declaration of Helsinki. It was approved by the Ethics Committee of the First People's Hospital of Zunyi City (Ethics Review Approval Number: (2025)-1-739).

Funding

This work was supported by the Zunyi Science and Technology Plan Project (Grant No. Zun Shi Ke He HZ (2019) 170).

Author Contribution

Nan Huang and Danni Zheng wrote the main manuscript. Nan Huang , Kai Zhu, Min He, Xiaoli Yang and Jie Jiang were conducted for statistical analysis, prepared figures and tables. Shengyi Wang and Yunrong Li were responsible for data collection. Xiang Zhang provided fund project support and reviewed the manuscript.

Acknowledgement

The authors would like to thank the Science and Technology Bureau of Zunyi for their financial support (grant number: Zun Shi Ke He HZ (2019) 170).

Data Availability

The datasets used and/or analyzed during the current study are presented in the manuscript. Further inquiries can be available from the corresponding author on reasonable request.

References

1. Shang Z, Tan S, Ma D. Respiratory syncytial virus: from pathogenesis to potential therapeutic strategies. Int J Biol Sci. 2021;17(14):4073–91. 10.7150/ijbs.64762.
2. Pratt MTG, Abdalla T, Richmond PC, et al. Prevalence of respiratory viruses in community-acquired pneumonia in children: a systematic review and meta-analysis. Lancet Child Adolesc Health Aug. 2022;6(8):555–70. 10.1016/s2352-4642(22)00092-x.
3. Baum SG. Respiratory Syncytial Virus: Another Serious Respiratory Infection. Clin Infect Dis Jul. 2021;1(1):e164–5. 10.1093/cid/ciaa723.
4. Li Y, Wang X, Blau DM, et al. Global, regional, and national disease burden estimates of acute lower respiratory infections due to respiratory syncytial virus in children younger than 5 years in 2019: a systematic analysis. Lancet May. 2022;28(10340):2047–64. 10.1016/s0140-6736(22)00478-0.
5. Zhang XL, Zhang X, Hua W, et al. Expert consensus on the diagnosis, treatment, and prevention of respiratory syncytial virus infections in children. World J Pediatr Jan. 2024;20(1):11–25. 10.1007/s12519-023-00777-9.
6. Laudanno SL, Sánchez Yanotti CI, Polack FP. RSV Lower Respiratory Tract Illness in Infants of Low- and Middle-income Countries. Acta Med Acad Aug. 2020;49(2):191–7. 10.5644/ama2006-124.297.
7. Nguyen VH, Dubot-Pérès A, Russell FM, et al. Acute respiratory infections in hospitalized children in Vientiane, Lao PDR - the importance of Respiratory Syncytial Virus. Sci Rep Aug. 2017;24(1):9318. 10.1038/s41598-017-09006-6.
8. Mendoza-Cano O, Trujillo X, Huerta M et al. Risk factors for severe disease in pediatric respiratory syncytial virus infections. *Public Health*. 2024/10/01/ 2024;235:167–172. https://doi.org/10.1016/j.puhe.2024.06.023
9. Gong CX, Yue HL, Li QH, et al. Building a diagnostic prediction model for severe Mycoplasma pneumoniae pneumonia in children using machine learning. FRONTIERS PUBLIC HEALTH SEP. 2025;18:131585042. 10.3389/fpubh.2025.1585042.
10. Ding XY, Bai J, Liu R. Risk factors for airway clearance dysfunction in children with severe pneumonia: a retrospective study of LASSO model. FRONTIERS PEDIATRICS OCT. 2025;29:131638103. 10.3389/fped.2025.1638103.
11. Xie S, Wu M, Shang Y, et al. Development and validation of an early diagnosis model for severe mycoplasma pneumonia in children based on interpretable machine learning. RESPIRATORY RESEARCH MAY. 2025;13(1):182. 10.1186/s12931-025-03262-1.
12. Jeon ET, Lee HJ, Park TY, et al. Machine learning-based prediction of in-ICU mortality in pneumonia patients. Sci Rep Jul. 2023;17(1):11527. 10.1038/s41598-023-38765-8.
13. [Guidelines for the management of community-acquired pneumonia in children. (2024 revision)]. *Zhonghua Er Ke Za Zhi*. Oct 2 2024;62(10):920–930. 10.3760/cma.j.cn112140-20240728-00523
14. Guideline for the. diagnosis, treatment, and prevention of respiratory syncytial virus infection in children in China. Pediatr Investig Sep. 2025;9(3):205–20. 10.1002/ped4.70010.
15. Althouse BM, Flasche S, Minh LN, et al. Seasonality of respiratory viruses causing hospitalizations for acute respiratory infections in children in Nha Trang, Vietnam. Int J Infect Dis Oct. 2018;75:18–25. 10.1016/j.ijid.2018.08.001.
16. Rodríguez-Martínez CE, Rodríguez DA, Nino G. Respiratory syncytial virus, adenoviruses, and mixed acute lower respiratory infections in children in a developing country. J Med Virol May. 2015;87(5):774–81. 10.1002/jmv.24139.
17. Delacourt C. Sequelae of viral lower respiratory tract infections in children. *Rev Prat*. Nov 15. 2007;57(17):1919-22. Séquelles des infections virales respiratoires basses de l'enfant.
18. Nguyen STT, Tran TA, Vo GV. Severe Pneumonia Caused by Respiratory Syncytial Virus and Adenovirus in Children from 2 to 24 Months at Children's Hospital 1 in Ho Chi Minh City, Vietnam. Viruses Mar. 2024;7(3). 10.3390/v16030410.
19. Reddel HK, Bacharier LB, Bateman ED, et al. Global Initiative for Asthma Strategy 2021: executive summary and rationale for key changes. Eur Respir J Jan. 2022;59(1). 10.1183/13993003.02730-2021.

20. El-Nawawy A, Antonios MA, Meheissen MA, Fahim MM. Respiratory viruses associated with severe mechanically ventilated pneumonia in children. J Med Virol Feb. 2022;94(2):461–8. 10.1002/jmv.27284.
21. Ma R, Zhang C, Zhang Y, et al. The impact of respiratory syncytial virus on asthma development and exacerbation. Ann Allergy Asthma Immunol Sep. 2025;135(3):268–75. 10.1016/j.anai.2025.05.011.
22. Yang Z, Chen S, Tang X, et al. Development and validation of machine learning-based prediction model for severe pneumonia: A multicenter cohort study. Heliyon Sep. 2024;15(17):e37367. 10.1016/j.heliyon.2024.e37367.
23. He B, Li X, Dong R, et al. Development of machine learning-based differential diagnosis model and risk prediction model of organ damage for severe Mycoplasma pneumoniae pneumonia in children. Sci Rep Mar. 2025;19(1):9431. 10.1038/s41598-025-92089-3.
24. Lim WS, van der Eerden MM, Laing R, et al. Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. Thorax May. 2003;58(5):377–82. 10.1136/thorax.58.5.377.
25. Fine MJ, Auble TE, Yealy DM, et al. A prediction rule to identify low-risk patients with community-acquired pneumonia. N Engl J Med. Jan 1997;23(4):243–50. 10.1056/nejm199701233360402.
26. Öner D, Drysdale SB, McPherson C, et al. Biomarkers for Disease Severity in Children Infected With Respiratory Syncytial Virus: A Systematic Literature Review. J Infect Dis Oct. 2020;7(Suppl 7):S648–57. 10.1093/infdis/jiaa208.
27. Megied AAE, Abdel Fattah Abdel Motey M, Aziz M, Ebrahim MM. MM. Diagnostic and predictive value of Respiratory Index of Severity in Children (RISC) scoring system in community-acquired pneumonia: a prospective cross sectional study. *Egyptian Pediatric Association Gazette*. 2023/05/22 2023;71(1):24. 10.1186/s43054-023-00170-1
28. Florin TA, Tancredi DJ, Ambroggio L, et al. Predicting paediatric pneumonia severity in the emergency department: a multinational prospective cohort study of the Pediatric Emergency Research Network. Lancet Child Adolesc Health Jun. 2025;9(6):383–92. 10.1016/s2352-4642(25)00094-x.
29. Shah SN, Bachur RG, Simel DL, Neuman MI. Does This Child Have Pneumonia? The Rational Clinical Examination Systematic Review. *Jama*. Aug 1. 2017;318(5):462–471. 10.1001/jama.2017.9039
30. Florin TA, Ambroggio L, Lorenz D, et al. Development and Internal Validation of a Prediction Model to Risk Stratify Children With Suspected Community-Acquired Pneumonia. Clin Infect Dis Nov. 2021;2(9):e2713–21. 10.1093/cid/ciaa1690.
31. Daw W, Kaur R, Delaney M, Elphick H. Respiratory rate is an early predictor of clinical deterioration in children. Pediatr Pulmonol Aug. 2020;55(8):2041–9. 10.1002/ppul.24853.
32. Gong W, Gao K, Shan Z, et al. Research progress of biomarkers in evaluating the severity and prognostic value of severe pneumonia in children. Front Pediatr. 2024;12:1417644. 10.3389/fped.2024.1417644.
33. Rees CA, Haggie S, Florin TA. Narrative review of clinical prediction models for paediatric community acquired pneumonia. Paediatr Respir Rev Jun. 2025;54:19–27. 10.1016/j.prrv.2025.01.002.
34. Gong W, Gao K, Ni J, et al. Construction and verification of a risk factor prediction model for neonatal severe pneumonia. Front Med (Lausanne). 2025;12:1536705. 10.3389/fmed.2025.1536705.
35. Jakhar SK, Pandey M, Shah D, et al. Etiology and Risk Factors Determining Poor Outcome of Severe Pneumonia in Under-Five Children. Indian J Pediatr Jan. 2018;85(1):20–4. 10.1007/s12098-017-2514-y.
36. Nissen MD. Congenital and neonatal pneumonia. Paediatr Respir Rev Sep. 2007;8(3):195–203. 10.1016/j.prrv.2007.07.001.
37. Wang C, Deng R, Gou L, et al. Preliminary study to identify severe from moderate cases of COVID-19 using combined hematology parameters. Ann Transl Med. May 2020;8(9):593. 10.21037/atm-20-3391.
38. Ma X, Zhao W, Sun Q, et al. Development and validation of a machine learning model for critical progression risk in pediatric severe community-acquired pneumonia. Sci Rep Dec. 2025;2(1):43011. 10.1038/s41598-025-23209-2.
39. Han T, Cheng T, Liao Y, et al. Analysis of the Value of the Blood Urea Nitrogen to Albumin Ratio as a Predictor of Mortality in Patients with Sepsis. J Inflamm Res. 2022;15:1227–35. 10.2147/jir.S356893.
40. Sharma K, Mogensen KM, Robinson MK. Pathophysiology of Critical Illness and Role of Nutrition. Nutr Clin Pract Feb. 2019;34(1):12–22. 10.1002/ncp.10232.
41. Ruiz LA, Serrano L, Pérez S, et al. Impact of severe lymphopenia on the early prediction of clinical outcome in hospitalized patients with pneumococcal community-acquired pneumonia. Infection Oct. 2023;51(5):1319–27. 10.1007/s15010-023-01984-2.
42. Bellelli V, d'Ettorre G, Celani L, Borrazzo C, Ceccarelli G, Venditti M. Clinical significance of lymphocytopenia in patients hospitalized with pneumonia caused by influenza virus. Crit Care Oct. 2019;29(1):330. 10.1186/s13054-019-2608-1.

Figures

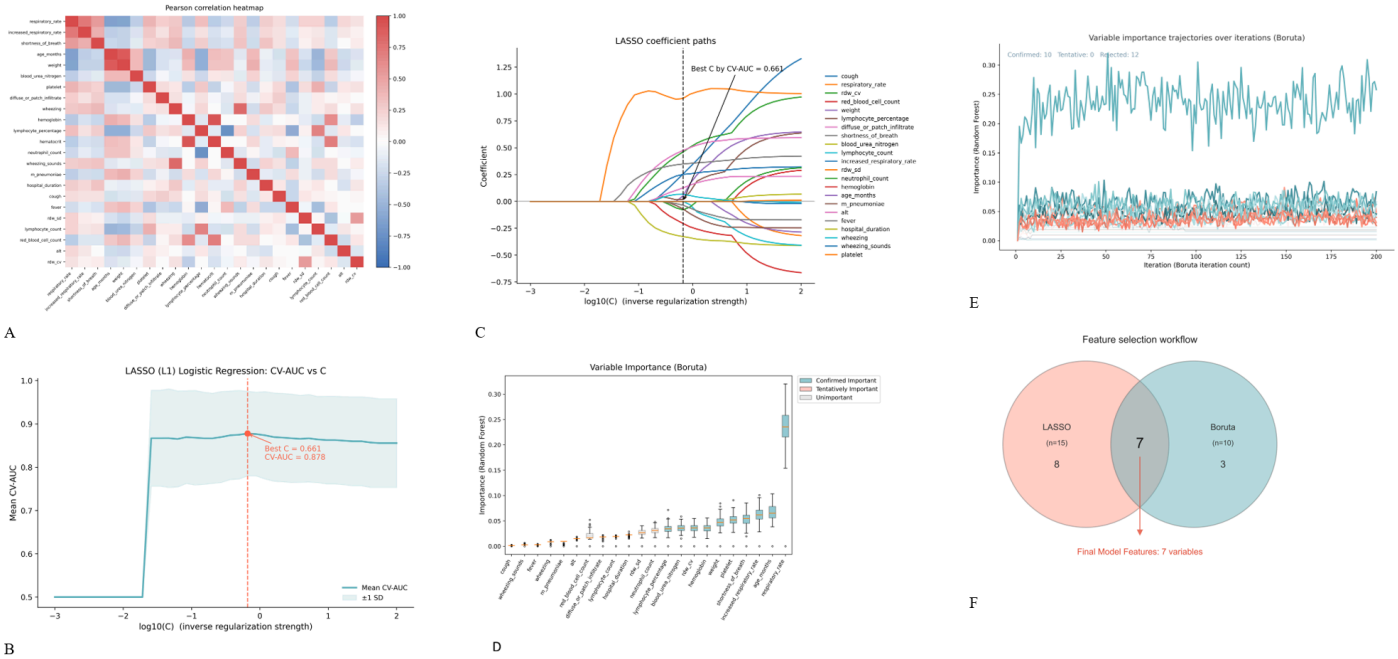


Figure 1

A Pearson correlation analysis, with one highly collinear feature (|r| > 0.90)

Fig 1B LASSO regression with ten-fold cross-validation, where the optimal λ (0.661) is determined through cross-validation and corresponds to the highest mean CV-AUC (0.878).

Fig 1C Lasso coefficient paths for feature selection at the optimal value of λ, showing the 15 most relevant features selected by the model with non-zero coefficients.

Fig 1D Boruta feature importance assessment, with 10 features in blue indicating confirmed important features.

Fig 1E Variable importance trajectories over iterations (Boruta) in the Boruta feature selection algorithm.

Fig 1F Venn diagram highlights the overlap of 7 features between the LASSO and Boruta algorithms.

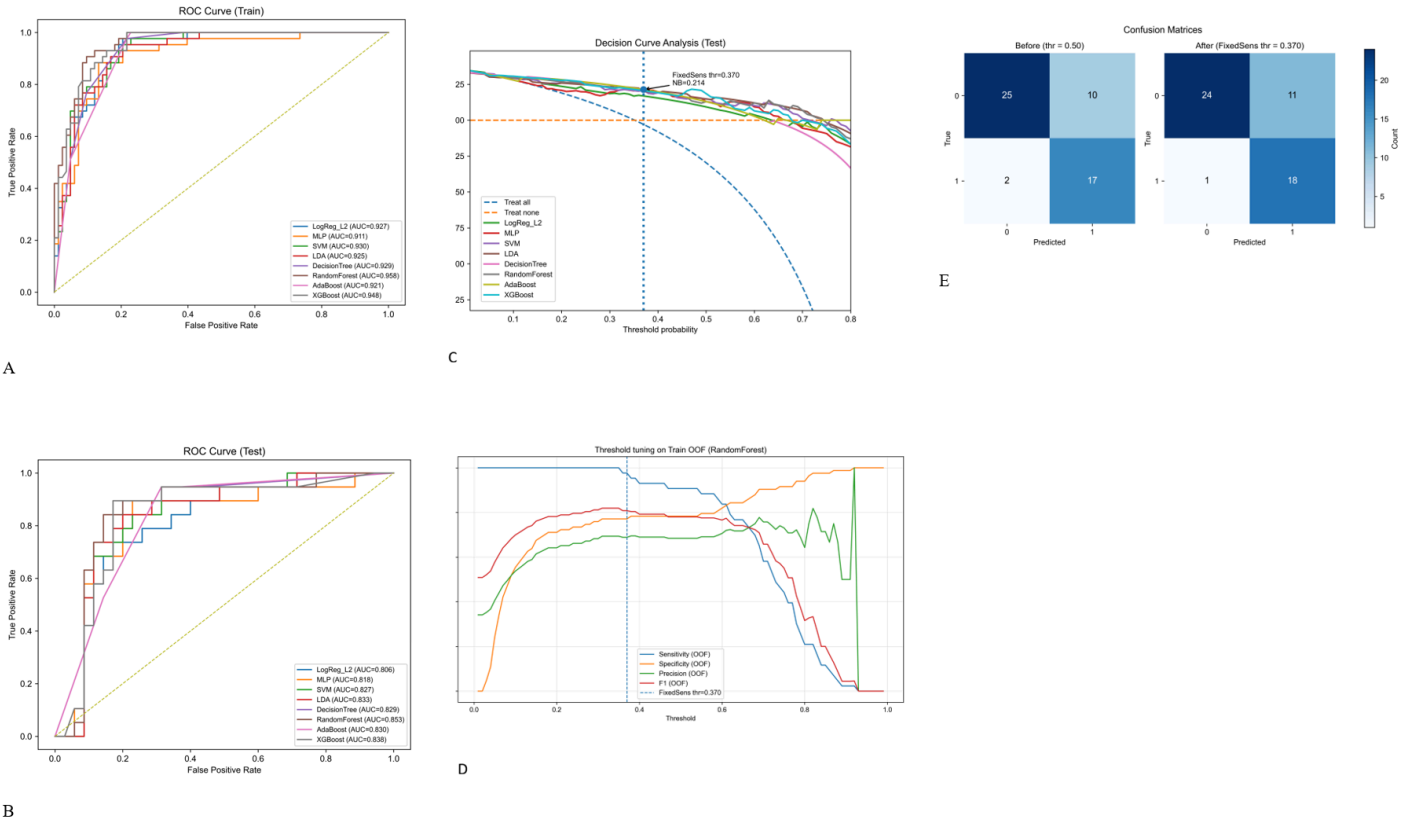


Figure 2

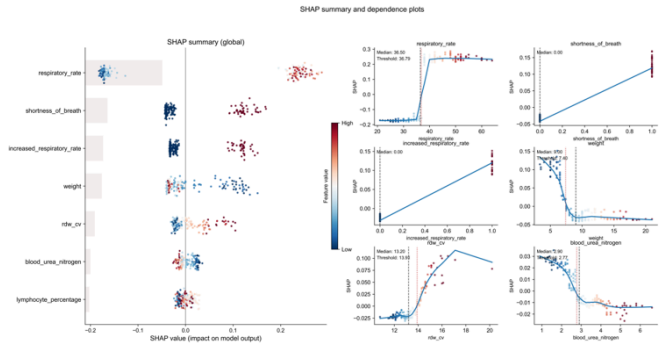
A The ROC curve for each model in the training cohort.

Fig 2B The ROC curve for each model in the testing cohort.

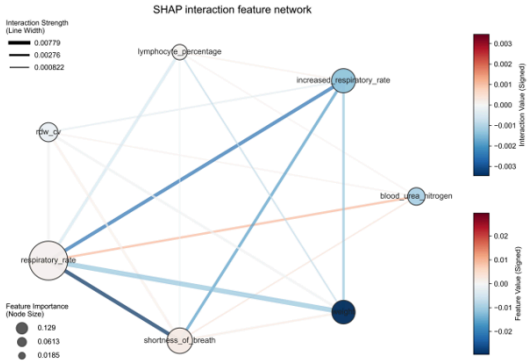
Fig 2C The decision curve analysis for each model.

Fig 2D The threshold for random forest model.

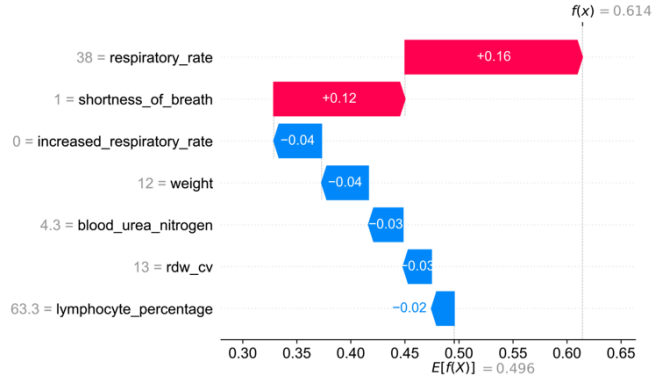
Fig 2E Confusion matrix in testing cohort of random forest model with the default and optimal threshold.



A



B



C

Figure 3

A SHAP summary plot and partial dependence plots for global feature importance.

Fig 3B SHAP Interaction Network.

Fig 3C SHAP waterfall plot for individual model interpretation.