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 interpretation of an XGBoost-based model for predicting prolonged ICU stay in patients with acute exacerbation of chronic obstructive pulmonary disease

Wensheng Chang
Affiliated Hospital of Shandong Medical University
Haimin Li
Affiliated Hospital of Shandong Medical University
Chunqing Hu
Affiliated Hospital of Shandong Medical University
Quanmei Shang
Affiliated Hospital of Shandong Medical University
Xiuxiu Liu
937776268@qq.com
Affiliated Hospital of Shandong Medical University
Hang Yang
Macao Polytechnic University

Research Article

Keywords: risk factors, SHAP、XGBoost、prediction

Posted Date: August 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10277166/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: Prolonged hospital length of stay (LOS) in AECOPD patients significantly increases healthcare resource consumption and economic burden, and there is an urgent need to develop early prediction tools based on admission metrics to optimize clinical decision-making.

Objective: The aim of this study was to construct a prediction model based on indicators accessible at admission to predict the risk of prolonged LOS in patients with AECOPD, and to reveal the key influencing factors through interpretable analysis.

Methods: The medical records of 520 AECOPD patients admitted to the ICU from December 2020 to December 2025 from a tertiary general hospital in Shandong Province were retrospectively included, including the collection of demographic characteristics (age, gender, occupation, etc.), clinical indicators (comorbidities, medication history, vital signs, etc.), and test indicators (routine blood counts, coagulation function, blood gas analysis, etc.). Lasso regression was used to screen the variables and construct the XGBoost prediction model. The SHAP method was used to visualize the decision-making mechanism of the model. The SHAP values of each variable were calculated to quantify its contribution to the prediction results, determine the importance ranking, and reveal the potential associations and interactions between variables. The performance of the model was evaluated by plotting the ROC curve, calibration curve, DCA curve, and calculating the sensitivity, specificity, F1-Score and other indicators.

Results: A total of 520 AECOPD patients were enrolled in this study, of which 123 developed prolonged LOS, with an incidence rate of 23.7%.The XGBoost model performed well in both the training set and the testing set, with AUC values (0.995; 0.891), sensitivity (0.976; 0.756), specificity (0.954; 0.896). Moreover, both the calibration curves and DCA curves indicated that the model had an overall good performance. Five independent predictor variables were included in the XGBoost model. The importance of their SHAP values in the model is ranked as follows: FIB, Tni, $PaCO_2$, CPHD, and CPR.

Conclusion: The XGBoost prediction model constructed in this study has good predictive efficacy and clinical interpretability, and the combinations of key variables revealed in the model can provide a scientific basis for the development of individualized intervention plans for prolonged LOS in AECOPD patients. Multicenter prospective studies are needed to further validate the reliability and generalizability of the model and to promote its clinical application.

1. Introduction

Chronic obstructive pulmonary disease (COPD) is a chronic respiratory disease characterized by persistent airflow limitation worldwide, and its prevalence and disease burden are increasing year by year. According to the World Health Organization (WHO), COPD has been ranked as the fourth leading cause of death in the world, leading to more than 3.5 million deaths (about 5% of global deaths) in 2021 (1). The prevalence rate of COPD in China is as high as 13.7% in the population over 40 years of age and the trend of the disease prevalence is still intensifying due to the impact of factors such as population aging, air pollution, and cigarette smoking (2).Patients with COPD often hospitalized repeatedly due to acute exacerbation of chronic obstructive pulmonary disease (AECOPD), which are characterized by worsening symptoms such as dyspnea, increased sputum production, or purulent sputum, and require intensive therapeutic interventions, and the mortality rate of hospitalization for severely ill patients can be as high as 38.2% (3). Prolonged Length of Stay (LOS) is a central challenge in the clinical management of AECOPD. Although the average length of stay is 7–10 days, approximately 20%-30% of patients are hospitalized for more than 11 days, which significantly increases healthcare costs and is strongly associated with poor prognosis (4, 5). The driving mechanisms of prolonged LOS are complex and involve multidimensional factors such as infection severity, comorbidity burden, treatment responsiveness, and social support (6); therefore, early identification of high-risk patients at risk for prolonged LOS is important for optimizing resource allocation, developing individualized intervention strategies, and improving prognosis.

In recent years, with the rapid development of big data and artificial intelligence technologies, machine learning and deep learning models have been increasingly used in the medical field, providing new possibilities for AECOPD prediction. These models are able to extract useful information from large-scale medical data, learn complex disease progression patterns, and

generate personalized predictions. Indeed, it has been shown (7, 8) that machine learning models have the potential to predict the risk of hospitalization and mortality in COPD patients, providing strong support for the development of more accurate prediction models. However, despite the achievements of machine learning models in COPD prediction, the development of prediction models specifically targeting prolonged LOS in AECOPD patients is still relatively underdeveloped. Most of the existing prediction models focus on the prediction of risk of death or readmission (9, 10), while less attention has been paid to the specific outcome of prolonged LOS (11). Prolonged LOS not only reflects the severity of the disease and the difficulty of treatment, but is also associated with higher healthcare costs and poorer patient prognosis. Therefore, the development of a machine learning model that can accurately predict prolonged LOS in AECOPD patients is important for guiding clinical decisions, optimizing patient management strategies, and reducing healthcare costs. In this study, we not only aim to construct such a predictive model, but also pay special attention to the interpretability of the model. To achieve this goal, we used the SHapley Additive exPlanations (SHAP) method.

SHAP is an explanatory method based on game theory that quantifies how much each feature contributes to the model's predicted outcome (12). The SHAP values allow us to visualize which factors have the greatest impact on the prolonged LOS in AECOPD patients and how these factors specifically contribute to the model prediction process. This interpretability not only enhances our understanding of the model's decision-making process, but also helps clinicians to better accept and trust the model, thus facilitating its application in actual clinical decision-making.

Therefore, the main objective of this study was to construct and validate a predictive model for prolonged LOS in AECOPD patients based on clinical indicators at the time of admission to the ICU and combined with the SHAP interpretation, with the aim of providing a more accurate and effective tool for the clinical management and resource allocation of AECOPD patients.

2. Methods

2.1 Research design and data sources

We collected medical records of 520 AECOPD patients admitted to the ICU of a tertiary general hospital in Shandong Province from December 2020 to December 2025 for a single-center retrospective cohort study.

Inclusion criteria: ①Meet the diagnostic criteria for AECOPD as defined by the GOLD 2023 guidelines (13); ②Age ≥ 18 years old; ③Relevant clinical indicators are collected within 24 hours of admission.

Exclusion criteria: ①Admission for other reasons, AECOPD is not the main diagnosis; ②Patients with other serious diseases in combination (such as malignant tumors, severe hepatic and renal insufficiency, etc.); ③Incomplete medical records or > 20% missing data in key variables.

Outcome indicator: prolonged LOS (calculated based on the 75th percentile of the LOS distribution, taking LOS greater than the 75th percentile as prolonged LOS).

2.2 Data collection:

We collected clinical records of patients and included 41 variables, including:

Basic information: sex, age, profession, history of smoking, history of drinking, oxygen therapy, fever, respiratory rate, SpO_2

Disease history: COPD, chronic pulmonary heart disease (CPHD), coronary heart disease(CHD), heart failure, cerebrovascular disease, chronic kidney disease, hypertension, diabetes, hypoproteinemia

Therapeutic measures: Non-invasive Ventilation(NIV), glucocorticoids, bronchodilators

Laboratory indices within 24 hours of admission: WBC, RDW, ESR, PLT, HB, NEUT, LYMP, Absolute Eosinophil Count, ALB, FIB, CRP, PCT, Tni, NT-proBNP, UA, Arterial Blood Gas pH, $PaCO_2$, PaO_2, Sodium, Potassium

2.3 Variable handling and screening

Missing value processing: processing was performed using multiple interpolation or deletion of missing > 20% of the variables.

To avoid overfitting, the variables were screened using the Least Absolute Contraction and Selection Operator (LASSO) (14).

2.4 Model Development and evaluation

XGBoost model: According to the ratio of 7:3, the research data was randomly divided into the training set and the testing set. The data in the training set were used to develop a prediction model using the "XGBoost" package. We adjusted the model parameters and then output the optimal parameter configurations to build the XGBoost model.

Model evaluation was conducted by drawing the Receiver Operating Characteristic Curve (ROC) and the calibration curve, performing Decision Curve Analysis (DCA), and by calculating the Area Under Curve (AUC) value of the model, accuracy, sensitivity, specificity, and F1 - score to evaluate the predictive performance of the model.

2.5 Interpretability analysis

SHAP value can deeply analyze the importance of variables and their contribution to model prediction from both global and individual levels, and visually explain the model by drawing Feature Importance Bar Plot, Beeswarm Plot, and Force Plot. Specifically, Feature Importance Bar Plot calculates the average SHAP absolute value of each variable to intuitively show the overall impact of the variable on model prediction; Beeswarm Plot distributes the SHAP value of each data point along the horizontal axis and uses color to represent the size of the variable value, where each point represents a sample, the horizontal position of the point represents the contribution of the variable to the prediction (SHAP value), and the depth of the color reflects the high or low value of the variable; Force Plot selects true positive or true negative samples to show the contribution of each variable in a single sample, thereby specifically explaining the model prediction results of the sample.

2.6 Statistical analysis

This study used SPSS 26.0 software as well as R4.3.0 software for statistical analysis and data visualization. Continuous variables were expressed using ($\bar{x} \pm s$) or M (P_{25}, P_{75}), and comparisons between groups were made using t-test or the Mann-Whitney U test; categorical variables were expressed using the number of cases and percentages, and comparisons between groups were made using chi-squared (χ^2) test. The “glmnet”, “shapviz”, “XGBoost”, “pROC”, “rms”, and “rmda” packages in R4.3.0 software were used for data organization and visualization.

3. Results

3.1 Baseline characteristics

A total of 520 AECOPD patients were included in this study, with a mean age of (73.62 ± 7.88) years, 340 males (65.38%), and a LOS of 9 (7, 12) days, with a cutoff point of 12 days based on the P_{75} of LOS, of which 123 patients (23.7%) had a prolonged LOS (LOS > 12). The basic clinical data of the two groups are shown in Table 1, and there was a statistical difference between the two groups in the 12 variables of CPHD, Tni, R, SpO_2, WBC, RDW, NEUT, ALB, FIB, CRP, $PaCO_2$, and PaO_2 ($p < 0.05$).

Table 1
The baseline and clinical characteristics in prolonged LOS patients and no prolonged LOS patients.

Variables		All patients (n = 520)	No prolonged LOS (n = 397)	Prolonged LOS (n = 123)	Statistical value	p
Age		73.62 ± 7.88	73.55 ± 7.79	73.85 ± 8.21	-0.364	0.716
Gender	Male	340(65.38)	256(64.48)	84(68.29)	0.602	0.438
	Female	180(34.62)	141(35.52)	39(31.71)		
Profession	No	101(19.42)	82(20.65)	19(15.45)	1.651	0.648
	Retiree	92(17.69)	69(17.38)	23(18.70)		
	Workers	27(5.19)	20(5.04)	7(5.69)		
	Farmers	300(57.69)	226(56.93)	74(60.16)		
COPD	No	245(47.12)	192(48.36)	53(43.09)	1.048	0.306
	Yes	275(52.88)	205(51.64)	70(56.91)		
CPHD	No	347(66.73)	302(76.07)	45(36.59)	65.946	<0.001
	Yes	173(33.27)	95(23.93)	78(63.41)		
CHD	No	418(80.38)	315(79.35)	103(83.74)	1.15	0.283
	Yes	102(19.62)	82(20.65)	20(16.26)		
Heart Failure	No	495(95.19)	379(95.47)	116(94.31)	0.275	0.600
	Yes	25(4.81)	18(4.53)	7(5.69)		
Cerebrovascular Disease	No	432(83.08)	334(84.13)	98(79.67)	1.326	0.249
	Yes	88(16.92)	63(15.87)	25(20.33)		
Chronic Kidney Disease	No	508(97.69)	388(97.73)	120(97.56)	0.012	0.912
	Yes	12(2.31)	9(2.27)	3(2.44)		
Fever	No	431(82.88)	325(81.86)	106(86.18)	1.232	0.267
	Yes	89(17.12)	72(18.14)	17(13.82)		
Hypertension	No	317(60.96)	234(58.94)	83(67.48)	2.876	0.090
	Yes	203(39.04)	163(41.06)	40(32.52)		
Diabetes	No	463(89.04)	351(88.41)	112(91.06)	0.673	0.412
	Yes	57(10.96)	46(11.59)	11(8.94)		
History of smoking	No	164(31.54)	131(33.00)	33(26.83)	1.655	0.198
	Yes	356(68.46)	266(67.00)	90(73.17)		
History of drinking	No	404(77.69)	303(76.32)	101(82.11)	1.817	0.178
	Yes	116(22.31)	94(23.68)	22(17.89)		
Oxygen therapy	No	385(74.04)	289(72.80)	96(78.05)	1.348	0.246
	Yes	135(25.96)	108(27.20)	27(21.95)		

Variables		All patients (n = 520)	No prolonged LOS (n = 397)	Prolonged LOS (n = 123)	Statistical value	p
Hypoproteinemia	No	341(65.58)	260(65.49)	81(65.85)	0.005	0.941
	Yes	179(34.42)	137(34.51)	42(34.15)		
NIV	No	432(83.08)	334(84.13)	98(79.67)	1.326	0.249
	Yes	88(16.92)	63(15.87)	25(20.33)		
Glucocorticoids	No	114(21.92)	92(23.17)	22(17.89)	1.534	0.216
	Yes	406(78.08)	305(76.83)	101(82.11)		
Bronchodilators	No	53(10.19)	44(11.08)	9(7.32)	1.455	0.228
	Yes	467(89.81)	353(88.92)	114(92.68)		
PCT(ng/mL)	≤ 0.046	142(27.31)	115(28.97)	27(21.95)	2.329	0.127
	>0.046	378(72.69)	282(71.03)	96(78.05)		
Tni(μg/L)	≤ 0.03	353(67.88)	312(78.59)	41(33.33)	88.219	<0.001
	>0.03	167(32.12)	85(21.41)	82(66.67)		
NT-proBNP(pg/ml)	≤ 125	132(25.38)	107(26.95)	25(20.33)	2.177	0.140
	>125	388(74.62)	290(73.05)	98(79.67)		
Respiratory rate(times)		21(19, 23)	21(18, 23)	22(21, 24)	-4.442	<0.001
SpO_2		94(91,96)	95(92,97)	92(87,95)	-4.871	<0.001
WBC(×10^9/L)		7.5(5.7,10.7)	7.40(5.6,10.2)	8.30(6.2,12.2)	-2.324	0.020
RDW(%)		43.2(41.1,46.1)	43.0(40.8,45.9)	44.2(42.0,47.1)	-2.842	0.004
ESR(mm/h)		35(18,35)	35(18,35)	35(18,38)	-0.189	0.850
PLT(×10^9/L)		220(183,267)	220(183,263)	225(182,289)	-1.272	0.203
HB(g/L)		132(121,143)	133(123,143)	131(119,146)	-0.5	0.617
NEUT(×10^9/L)		5.7(3.9,8.9)	5.6(3.8,8.4)	6.3(4.2,10.6)	-1.976	0.048
LYMP(×10^9/L)		1.2(0.8,1.7)	1.3(0.8,1.7)	1.1(0.7,1.6)	-1.487	0.137
Absolute Eosinophil Count(×10^9/L)		0.02(0,0.1)	0.02(0.0,0.1)	0.01(0.0,0.1)	-0.283	0.777
ALB(g/L)		36.5(33.1,39.4)	36.8(33.5,39.8)	35.4(31.7,38.0)	-3.285	<0.001
FIB(g/L)		3.8(2.9,4.9)	3.5(2.7,4.4)	4.9(4.4,5.2)	-10.355	<0.001
CRP(mg/L)		17(5.0,32.3)	14.4(4.5,32.3)	32.3(8.4,42.7)	-2.934	0.003
UA(μmol/L)		260.1(198.4,330.2)	261.1(198.7,333.3)	245.0(186.5,320.2)	-1.178	0.239
Arterial Blood Gas pH		7.43(7.40,7.46)	7.43(7.40,7.50)	7.43(7.40,7.50)	-1.131	0.258
$PaCO_2$(mmHg)		44(40,51)	42(39,49)	50(45,57)	-7.071	<0.001
PaO_2(mmHg)		70(57,79)	71(60,79)	66(49,76)	-2.731	0.006
Sodium(mmol/L)		136(133,140)	136(133,140)	137(134,139)	-0.407	0.684

Variables	All patients (n = 520)	No prolonged LOS (n = 397)	Prolonged LOS (n = 123)	Statistical value	p
Potassium(mmol/L)	3.9(3.6,4.2)	3.9(3.6,4.2)	3.9(3.6,4.3)	-1.127	0.260

COPD,Chronic Obstructive Pulmonary Disease;CPHD,Chronic Pulmonary Heart Disease;CHD,Coronary Heart Disease;NIV,Non-invasive Ventilation;PCT,Procalcitonin;Tni,Troponin I;BNP,Brain Natriuretic Peptide;SpO_2,Peripheral Oxygen Saturation;WBC,White Blood Cell Count;RDW,Red Cell Distribution Width;ESR,Erythrocyte Sedimentation Rate;PLT,Platelet Count;HB,Hemoglobin Count;NEUT,Neutrophil Count;LYMP,Lymphocyte Count;ALB,Albumin Count;FIB,Fibrinogen;CRP,C-reactive Protein;UA,Uric Acid;$PaCO_2$,Partial Pressure of Carbon Dioxide in Arterial Blood;PaO_2,Partial Pressure of Oxygen in Arterial Blood

3.2 Variable screening results and model development

Considering the possible correlation between different variables, Lasso regression was used to screen these 12 variables during the optimization of our model to screen risk factors that accurately predict the occurrence of prolonged LOS. As shown in Fig. 1 and Fig. 2, we utilized 10-fold cross-validation for lambda validation to calculate the minimum value of lambda (λmin) and one standard deviation of the smallest lambda value (λ1SE) in order to screen out the non-zero coefficients of the characteristic variables in Lasso regression. Using lambda 1SE values, the model fits well, screens fewer variables, and is simpler. Therefore, the five predictor variables corresponding to non - zero coefficients at lambda 1SE were selected and the five variables CPHD, Tni, FIB, CRP, and $PaCO_2$ were included in the model.

According to the ratio of 7:3, the research data was randomly divided into the training set and the testing set. We used the data in the training set to develop the prediction model using the "XGBoost" package. Then, we adjusted the model parameters and output the optimal parameter configurations. Ultimately, we set the learning rate (eta) to 0.1, the maximum depth (max_depth) to 4, and the number of iteration rounds to 100 to build the XGBoost model.

3.3 Evaluation of model performance

To assess the predictive efficacy of the XGBoost model, we plotted the ROC curves of the predictive model (Fig. 3A and 3B) and calculated the AUC values. The results showed that the XGBoost model was excellent in predicting the risk of prolonged LOS in AECOPD patients, with an AUC value as high as 0.995 in the training set, and an AUC value of 0.891 in the testing set. Further analysis of the model performance metrics showed that the sensitivity of the XGBoost model in the training set was 0.976, specificity was 0.954, Negative Predictive Value (NPV) was 0.993, and Positive Predictive Value (PPV) was 0.860, the accuracy was 0.959, and F1 score was 0.914. When the model is internally validated using the testing set, the XGBoost model still shows excellent performance. For specific data, refer to Table 2. In addition to the ROC curve evaluation, we also analyzed the model using the calibration curve. The results (Fig. 4A and 4B) show that the XGBoost model's calibration curve performs well, and the predicted probabilities are more consistent with the actual probabilities, reflecting the strong fitting ability of the model. In addition, we plotted DCA curves, and the analysis results (Fig. 5A and 5B) showed that within the threshold probability interval of 0 to 1, compared with the all-intervention strategy (All) and the no-intervention strategy (None), the XGBoost model was able to bring a more significant net benefit to the patients, leading to higher net clinical benefits. This suggests that the XGBoost model has outstanding application value in individualized diagnostic and therapeutic decision-making and can be used to provide more effective treatment for patients. It also provides a reliable reference basis for clinicians to make treatment decisions.

Table 2
The predictive capacity of XGBoost model.

Training set	AUC	Sensitivity	Specificity	NPV	PPV	Accuracy	F1 score
	0.995(0.991–0.999)	**0.976**	**0.954**	**0.993**	**0.860**	**0.959**	**0.914**
Testing set	0.891(0.836–0.946)	0.756	0.896	0.912	0.721	0.859	0.738

#AUC: area under the curve.

#NPV: negative predictive value.

#PPV: positive predictive value.

Figure 3A ROC curve of the model on the training set.

Figure 4A Calibration curve of the model on the training set.

Figure 5A Clinical decision curve of the model on the training set.

3.4 SHAP Interpretation

The magnitude of the contribution of each variable to the predictive results of the model was obtained using the SHAP algorithm, as shown in (Fig. 6), the feature importance bar plot of the model were sorted by the SHAP value, and the higher the SHAP value, the higher the predictive value of the variable to the model, in which FIB was the strongest predictive variable, followed by Tni, PaCO, CPHD, and CPR. The beeswarm plot, as shown in (Fig. 7), each point represents a sample, and the horizontal axis, the positive or negative SHAP value indicates the positive or negative effect of the corresponding variable, the yellow dots indicate an increase in the probability of prolonged LOS in patients with AECOPD, and the purple dots indicate a decrease in the probability of prolonged LOS in patients with AECOPD, and from the figure we can see that the patients suffering from CPHD are concentrated in the portion where the yellow dots are clustered on the right side of the horizontal axis, which indicates that the probability of prolonged LOS in patients with combined CPHD is greatly increased.

SHAP individual force plot mainly reveal the magnitude and direction of the contribution of each variable to the individual in the XGBoost model. We show the individual force plot for two different outcome patients, as shown in Fig. (8A, 8B), where yellow and purple colors represent the positive and negative effects on the prediction of the outcome, respectively, and E[f(x)] denotes the SHAP reference value, and f(x) is the predicted SHAP value of the individual. For the "true-positive" individuals (Fig. 8A), the SHAP value predicted by the XGBoost model is 0.756, which exceeds the reference value, indicating that the patient is at risk of prolonged LOS, and for the "true-negative" individuals (Fig. 8B), the SHAP value predicted by the XGBoost model is -3, which does not exceed the reference value, indicating that this patient is not at risk of prolonged LOS.

4. Discussion

4.1 Key findings

In this study, AECOPD patients were categorized into normal LOS and prolonged LOS groups according to LOS > 12 days, in which the incidence of prolonged LOS group was 23.7%, which was similar to the results of previous studies. To ensure the accuracy of the model and to reduce the problem of multicollinearity, we performed variable screening by Lasso regression, and identified the five variables of FIB, Tni, $PaCO_2$, CPHD, and CRP at admission as the core predictors of the occurrence of prolonged LOS in AECOPD patients. These variables covered three major dimensions: coagulation index (FIB), cardiac-related indexes (Tni, CPHD), and inflammation-related index (CRP) and respiratory function index ($PaCO_2$), and their synergistic effect further improved the predictive performance of the model.

Subsequently, these variables were used to develop the XGBoost predictive model, and our results showed that the XGBoost model had excellent predictive performance in terms of the risk of prolonged LOS in patients with AECOPD, with the highest

AUC values (0.995; 0.891) in both the training and testing sets. In addition, other evaluation metrics of the model, such as accuracy, sensitivity, and specificity, further supported this conclusion. The calibration curve plot of the model showed a small deviation between the predicted, actual observed and ideal curves of the model, indicating a high degree of calibration and good stability of the model. The DCA curve of the model showed that the model had a high net benefit in the threshold probability interval of 0–1, which also further verified the better clinical application value of the XGBoost model, suggesting that the model can be used as an important support tool for clinical decision-making.

Although the XGBoost model exhibits good predictive performance, its "black-box" nature poses a challenge for clinicians to intuitively comprehend the model's decision-making logic. To address this issue, this study integrated SHAP analysis with the XGBoost model to offer an intuitive interpretation of the model. Specifically, it not only quantifies the contribution of each variable to the model's predictions on a global scale by calculating the SHAP values for each variable but also generates detailed local interpretations for individual samples. This dual-interpretation mechanism, encompassing both global and local perspectives, equips clinicians with a visualization tool to intuitively understand the model's decisions, thereby significantly enhancing the model's interpretability and clinical applicability.

In the XGBoost model, we incorporated five key variables: FIB, Tni, $PaCO_2$, CPHD, and CPR, which belong to the four major dimensions of coagulation, cardiac function, inflammatory response, and respiratory function. Among these variables, FIB has the strongest predictive value. As an important component of the coagulation system, elevated levels of fibrinogen (FIB) not only reflect the hypercoagulable state in the body but are also closely associated with the inflammatory response (15). A study that included 240 COPD patients (comprising 120 AECOPD and 120 stable COPD patients) showed (16) that FIB levels were significantly higher in the AECOPD group than in the stable COPD group ($P < 0.05$), and that FEV1 as a percentage of the predicted value showed a significant negative correlation with FIB ($r=-0.656$, $P < 0.001$), suggesting that FIB levels increased with the degree of airflow limitation. Another study found (17) that FIB was a significant risk factor for poor prognosis in AECOPD, and that patients with AECOPD with FIB ≥ 4 g/L experienced a poor prognosis approximately 1.2 times higher than the normal group (OR = 1.196, 95% CI: 1.075–1.317). The median FIB of 4.9 (4.4, 5.2) g/L in the patients who developed prolonged LOS in the AECOPD group in this study also reflected a strong association between the level of FIB and prolonged LOS. This may be due to the fact that a high level of FIB may predict a higher risk of thrombosis (18), which affects blood perfusion and gas exchange in the lungs, thus prolonging LOS. Therefore, FIB may serve as an important biomarker for predicting prolonged LOS in patients with AECOPD.

Troponin (Tni), as a sensitive and specific marker of myocardial injury (19), is valuable in the assessment of COPD patients. In a multicenter cohort study that included 2085 patients with COPD (20), after comprehensive consideration of multiple respiratory and cardiovascular markers as well as associated risk factors, it was found that high-sensitivity troponin I (hs - Tni) was a key predictor of all-cause mortality, providing additional prognostic information for risk assessment in COPD patients. Similarly, the results of a study conducted by Kallis C et al. (21) in AECOPD patients showed that elevated troponin levels in AECOPD patients are expected to be a useful tool for predicting the occurrence of major adverse cardiovascular events (MACE) in COPD patients. The data from the study showed that, compared to patients with normal Tni levels, patients with Tni above the upper limit of normal had a significantly increased risk of MACE (adjusted hazard ratio 2.20, 95% CI: 1.75–2.77) and a higher risk of COPD readmission (adjusted hazard ratio 1.37, 95% CI: 1.02–1.83). When COPD patients are in an acute exacerbation of the disease, it is often accompanied by a more severe hypoxic state as well as an inflammatory response, which further exacerbates the risk of myocardial injury, which in turn leads to elevated Tni levels. Elevated Tni levels not only imply myocardial injury but also affect the pumping function of the heart, trigger pulmonary circulation disorders, and aggravate the degree of hypoxia in the body (22). The combination of these factors makes patients need to undergo a longer hospitalization and rehabilitation process. This shows that Tni is of non-negligible importance for the assessment of the condition of AECOPD patients and the prediction of LOS.

Several studies have explored the comorbidity profile of COPD patients in depth. A study pointed out (23) that among the many comorbidities suffered by COPD patients, CPHD is the most common one, accounting for about 40.99%. Meanwhile, most comorbidities were significantly correlated with prolonged LOS (on average 10 to 13 days longer) and increased daily hospitalization costs (ranging from 10 to 930 RMB) in patients. Of particular note, the presence of CPHD was also strongly

associated with an elevated one-year readmission rate in patients (adjusted odds ratio aOR: 1.41, 95% CI: 1.37–1.46). In addition, You L et al. similarly showed (24) that comorbid CPHD was an independent risk factor affecting the cost of medication and LOS in patients with AECOPD, implying that patients with AECOPD with comorbid CPHD had a heavier disease burden and a relatively poorer prognosis compared to other patients. In this study, it was analyzed and clarified that comorbid CPHD in AECOPD patients was indeed an independent risk factor for their prolonged LOS ($p < 0.05$). An in-depth analysis of the underlying mechanisms suggests that the presence of CPHD significantly elevates the risk of cardiovascular events, such as arrhythmia and heart failure, during hospitalization, which are undoubtedly key factors contributing to prolonged LOS, further aggravating the patients' treatment burden and physical impairment (25).

CRP is an acute-phase reactive protein, and its elevated level reflects the presence of inflammation or infection in the body (26). Chen L et al. showed (27) that smokers had a higher risk of developing COPD and had a corresponding increase in serum levels of CRP, and that there was an association between the two. Wang R et al. also showed (28) that in middle-aged and elderly COPD patients, CRP levels were significantly higher than those in the control group (OR: 1.20; 95% CI: 1.11–1.30) and were one of the independent factors influencing poor prognosis. High levels of CRP not only reflect increased inflammation in the lungs but may also indicate the presence of a systemic inflammatory response, which may lead to impaired multi-organ function and thus prolonged LOS (29). Therefore, CRP is an important indicator for assessing the inflammatory state and predicting prolonged LOS in patients with AECOPD.

Partial pressure of arterial blood carbon dioxide ($PaCO_2$) is one of the most important variables reflecting the gas exchange function of the lungs. A study by Zuoyou L et al. conducted a multiple regression analysis of 225 patients with chronic lung diseases (30), and the results showed that $PaCO_2$ was an independent variable associated with the elevation of the mean pulmonary arterial pressure. This suggests that clinicians managing chronic lung diseases should closely monitor and control the patients' $PaCO_2$ levels and take appropriate therapeutic measures to maintain patients' lung health and quality of life. Another study (31) showed that compared with patients with COPD or obstructive sleep apnea (OSA) alone, patients with COPD combined with OSA had significantly higher $PaCO_2$ levels and lower PaO_2 levels, with a statistically significant difference ($p < 0.001$). Moreover, these patients also had significantly lower sleep duration and efficiency and a poorer prognosis. Therefore, $PaCO_2$ level is considered one of the independent factors influencing the prognosis of patients with COPD combined with OSA.In patients with AECOPD, elevated $PaCO_2$ levels may be caused by an imbalance in the ventilation/perfusion ratio or alveolar hypoventilation due to inflammation in the lungs, which reflects a severe impairment of the patient's lung function (32). This impairment may lead to serious complications such as disturbances in acid-base balance and impaired consciousness, which in turn require more prolonged monitoring and treatment.

The application of machine learning models based on artificial intelligence in the medical field is a hot spot in current research (33, 34). It can not only help doctors to carry out accurate and rapid identification of the risk of disease occurrence and effectively improve the efficiency of clinical work, but also enable doctors to take targeted treatment measures based on the results of the model identification, thus improving the prognosis of patients and enhancing their quality of life.

Patients with AECOPD who fail to effectively control risk factors may experience prolonged LOS, and the current inadequacy of assessment for prolonged LOS has become a critical issue that needs to be addressed. Therefore, it is essential to conduct an adequate assessment to identify relevant risk factors. This study focused on identifying risk variables associated with prolonged LOS in patients with AECOPD using a machine learning-based model. The coagulation-based variable FIB, cardiac-related variables TnI and CPHD, inflammation-related variable CRP, and respiratory function variable $PaCO_2$ together constitute an important combination of variables for predicting prolonged LOS in AECOPD patients. These metrics reflect disease status from different perspectives and are interrelated, collectively impacting patient clinical outcomes. By comprehensively evaluating these variables, we were able to more accurately predict the risk of prolonged LOS in AECOPD patients. In addition, with the help of the SHAP method, we simplified the XGBoost model into an easy-to-understand graph. This visualization enhanced the comprehensibility of the model, making it more valuable to clinicians and providing strong support for clinical decision-making. As a result, doctors could optimize the treatment plan and reduce the number of unnecessary LOS for patients, ultimately improving their quality of life.

4.2 Limitations

Although the XGBoost model in this study demonstrated good prediction performance, we must still acknowledge its limitations. Firstly, the patient data utilized in this study were derived from a single-center, retrospective, and limited sample. This implies that the sample's representativeness may be inadequate to fully encompass the characteristics of patients from diverse regions and medical settings. Consequently, this could introduce a certain degree of bias into the study results and, to some extent, impact the model's generalization ability. Secondly, regarding the collection of laboratory variables, this study only incorporated the test data from the day the patients were admitted to the ICU due to resource constraints. It did not take into account the dynamic variables related to the therapeutic response. Since the progression of a disease is a dynamic process, the variables associated with a patient's treatment course will change over time. By constructing the model solely based on the data from the ICU admission day, we overlook the dynamic evolution of these variables during the treatment process. This results in the model's inability to comprehensively capture the features of the patient's condition and its changing trends, thereby reducing the comprehensiveness and accuracy of the model's predictions to a certain degree.

4.3 Innovativeness and Future Directions

This study exhibits innovation in several aspects. In terms of model development and feature selection, the XGBoost model is integrated with Lasso regression. This combination effectively circumvents the issue of multicollinearity and enhances the model's stability to a certain degree. Regarding model interpretation, the SHAP method is introduced. It visually illustrates the specific contribution of each variable to the prediction of prolonged LOS. Moreover, the analysis results are transformed into a practical clinical decision support tool. Based on this tool, doctors can accurately assess the risk and formulate a personalized treatment plan in accordance with the patient's condition.

In the future, this study will advance in the following three aspects. First, we will expand the sample size and establish multicenter cooperation. The current sample may have limitations in terms of geographic and population characteristics. By expanding the sample, we can cover more different patient groups, and through multicenter cooperation, we can integrate the data from various healthcare institutions. Through multicenter external validation, we can then test the generalization and stability of the model in different environments and patient groups, so as to enhance the reliability of the model. Second, the model can be used in clinical practice to collect dynamic data, such as changes in physiological variables during the treatment process, so as to improve the model and enhance the accuracy of the model's prediction of clinical events. Third, we will develop an embedded clinical decision support system (CDSS) to integrate the optimized model deeply into the hospital information system to achieve real-time data interaction. This way, clinicians can obtain decision-making suggestions based on model analysis, taking into account the individual patient and real-time condition, providing a more robust and personalized decision-making framework and improving the quality of medical care.

5. Conclusion

In this study, we used an XGBoost model based on admission variables to develop a model that can effectively predict prolonged LOS in AECOPD. This model includes five variables: FIB, Tni, PaCO2, CPHD, and CPR, and it performs well in terms of discrimination, calibration, and clinical application value. We used the SHAP method to visualize the model, presenting the complex and abstract model operation mechanism as an intuitive graphic. This clearly demonstrates the direction and degree of the influence of each variable on the patient's prolonged LOS, helping clinicians quickly understand the decision-making basis of the model's prediction and conduct personalized risk assessment according to the patient's specific situation. In the future, with further optimization and validation of the model, its clinical application scenarios will continue to expand, providing technical support for accurate diagnosis and treatment.

Abbreviations

AECOPD
acute exacerbation of chronic obstructive pulmonary disease

ALB
albumin
AUC
area under the receiver operating characteristic curve
CDSS
clinical decision support system
CHD
coronary heart disease
CI
confidence interval
COPD
chronic obstructive pulmonary disease
CPHD
chronic pulmonary heart disease
CRP
C-reactive protein
DCA
decision curve analysis
ESR
erythrocyte sedimentation rate
FEV1
forced expiratory volume in one second
FIB
fibrinogen
GOLD
Global Initiative for Chronic Obstructive Lung Disease
HB
hemoglobin
ICU
intensive care unit
LASSO
least absolute shrinkage and selection operator
LOS
length of stay
LYMP
lymphocyte count
MACE
major adverse cardiovascular events
NEUT
neutrophil count
NIV
non-invasive ventilation
NPV
negative predictive value
NT-proBNP
N-terminal pro-B-type natriuretic peptide
OR
odds ratio

OSA
obstructive sleep apnea
PaCO2
partial pressure of arterial carbon dioxide
PaO2
partial pressure of arterial oxygen
PCT
procalcitonin
PLT
platelet count
PPV
positive predictive value
RDW
red cell distribution width
ROC
receiver operating characteristic
SHAP
SHapley Additive exPlanations
SpO2
peripheral oxygen saturation
TnI
troponin I
UA
uric acid
WBC
white blood cell count
WHO
World Health Organization
XGBoost
eXtreme Gradient Boosting.

Declarations

Ethics approval and consent to participate

This retrospective study was approved by the Ethics Committee of the Affiliated Hospital of Shandong Medical University (Approval No. KYLL-156). All methods were performed in accordance with relevant guidelines and regulations, including the Declaration of Helsinki. The requirement for informed consent to participate was waived by the Ethics Committee of the Affiliated Hospital of Binzhou Medical University because of the retrospective nature of the study.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing Interests

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' contributions

YH contributed to the conception and design of the study and planned the overall framework and workflow. CWS, HCQ and LXX were responsible for data collection. LHM and SQM drafted the manuscript. All authors made substantial contributions to the conduct of the study and preparation of the manuscript, critically reviewed the manuscript, and read and approved the final version.

Acknowledgements

Not applicable.

References

1. World Health Organization. Chronic obstructive pulmonary disease (COPD). [Online] Available: https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd).
2. Wang C, Xu J, Yang L, Xu Y, Zhang X, Bai C, et al. Prevalence and risk factors of chronic obstructive pulmonary disease in China (the China Pulmonary Health [CPH] study): a national cross-sectional study. Lancet. 2018;391(10131):1706–17.
3. Soler-Cataluna JJ, Martinez-Garcia MA, Roman Sanchez P, Salcedo E, Navarro M, Ochando R. Severe acute exacerbations and mortality in patients with chronic obstructive pulmonary disease. Thorax. 2005;60(11):925–31.
4. Wang H, Yang T, Yu X, Chen Z, Ran Y, Wang J, et al. Risk Factors for Length of Hospital Stay in Acute Exacerbation Chronic Obstructive Pulmonary Disease: A Multicenter Cross-Sectional Study. Int J Gen Med. 2022;15:3447–58.
5. Memon MA, Faryal S, Brohi N, Kumar B. Role of the DECAF Score in Predicting In-hospital Mortality in Acute Exacerbation of Chronic Obstructive Pulmonary Disease. Cureus. 2019;11(6):e4826.
6. Singanayagam A, Schembri S, Chalmers JD. Predictors of mortality in hospitalized adults with acute exacerbation of chronic obstructive pulmonary disease. Ann Am Thorac Soc. 2013;10(2):81–9.
7. Adibi A, Sin DD, Safari A, Johnson KM, Aaron SD, FitzGerald JM, et al. The Acute COPD Exacerbation Prediction Tool (ACCEPT): a modelling study. Lancet Respir Med. 2020;8(10):1013–21.
8. Smith LA, Oakden-Rayner L, Bird A, Zeng M, To MS, Mukherjee S, et al. Machine learning and deep learning predictive models for long-term prognosis in patients with chronic obstructive pulmonary disease: a systematic review and meta-analysis. Lancet Digit Health. 2023;5(12):e872–81.
9. Chow R, So OW, Im JHB, Chapman KR, Orchanian-Cheff A, Gershon AS, et al. Predictors of Readmission, for Patients with Chronic Obstructive Pulmonary Disease (COPD) - A Systematic Review. Int J Chron Obstruct Pulmon Dis. 2023;18:2581–617.
10. Guerra B, Gaveikaite V, Bianchi C, Puhan MA. Prediction models for exacerbations in patients with COPD. Eur Respiratory Rev. 2017;26(143).
11. Yang L, Li M, Shu J, Yang Y, Huang Q. A Risk Prediction Model for Prolonged Length of Stay in Patients with Acute Exacerbations of Chronic Obstructive Pulmonary Disease: A Retrospective Study of 225 Patients in a Single Center in Kunming, China. Med Sci Monit. 2022;28:e934392.
12. Nohara Y, Matsumoto K, Soejima H, Nakashima N. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. Comput Methods Programs Biomed. 2022;214:106584.
13. Global Initiative for Chronic Obstructive Lung Disease (GOLD). Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease (2023 Report). [Online] Available: https://goldcopd.org/

14. Xu J, Chen T, Fang X, Xia L, Pan X. Prediction model of pressure injury occurrence in diabetic patients during ICU hospitalization–XGBoost machine learning model can be interpreted based on SHAP. Intensive Crit Care Nurs. 2024;83:103715.
15. Guven B, Can M. Fibrinogen: Structure, abnormalities and laboratory assays. Adv Clin Chem. 2024;120:117–43.
16. Chen L, Xu W, Chen J, Zhang H, Huang X, Ma L, et al. Evaluating the clinical role of fibrinogen, D-dimer, mean platelet volume in patients with acute exacerbation of COPD. Heart Lung. 2023;57:54–8.
17. Hu X, Xu J, Li P, Zheng H. Correlation of Serum Clara Cell Secretory Protein 16, Plasma Fibrinogen and Serum Amyloid A with the Severity of Acute Exacerbated COPD and Their Combination in Prognosis Assessment. Int J Chron Obstruct Pulmon Dis. 2023;18:1949–57.
18. Mlačo A, Mlačo N, Begić E, Mekić M, Džubur A. D-Dimer and Fibrinogen Values according to the Localization of Deep Venous Thrombosis. Int J Angiol. 2023;32(04):243–7.
19. Campu A, Muresan I, Craciun AM, Cainap S, Astilean S, Focsan M. Cardiac Troponin Biosensor Designs: Current Developments and Remaining Challenges. Int J Mol Sci. 2022;23(14).
20. Waschki B, Alter P, Zeller T, Magnussen C, Neumann JT, Twerenbold R et al. High-sensitivity troponin I and all-cause mortality in patients with stable COPD: an analysis of the COSYCONET study. Eur Respir J. 2020;55(2).
21. Kallis C, Kaura A, Samuel NA, Mulla A, Glampson B, O'Gallagher K, et al. The Relationship Between Cardiac Troponin in People Hospitalised for Exacerbation of COPD and Major Adverse Cardiac Events (MACE) and COPD Readmissions. Int J Chron Obstruct Pulmon Dis. 2023;18:2405–16.
22. Lazar DR, Lazar FL, Homorodean C, Cainap C, Focsan M, Cainap S, et al. High-Sensitivity Troponin: A Review on Characteristics, Assessment, and Clinical Implications. Dis Markers. 2022;2022:9713326.
23. Duan Y, Fu H, Chen C, Zhao Y, Jiang S, Wang C. Comorbidity Patterns in Chronic Obstructive Pulmonary Disease and Their Associations with Service Utilization. COPD. 2024;21(1):2414793.
24. You L, Niu H, Huang K, Dong F, Yang T, Wang C. Clinical Features and Outcomes of Acute Exacerbation in Chronic Obstructive Pulmonary Disease Patients with Pulmonary Heart Disease: A Multicenter Observational Study. Int J Chron Obstruct Pulmon Dis. 2021;16:2901–10.
25. Kahnert K, Jorres RA, Behr J, Welte T. The Diagnosis and Treatment of COPD and Its Comorbidities. Dtsch Arztebl Int. 2023;120(25):434–44.
26. Levinson T, Wasserman A, Shenhar-Tsarfaty S, Halutz O, Shapira I, Zeltser D, et al. Comparative analysis of CRP as a biomarker of the inflammatory response intensity among common viral infections affecting the lungs: COVID-19 versus influenza A, influenza B and respiratory syncytial virus. Clin Exp Med. 2023;23(8):5307–13.
27. Chen L, Xiong H, Wen Q, Lv J, Sun D, Pei P, et al. The Role of Active and Passive Smoking in Chronic Obstructive Pulmonary Disease and Systemic Inflammation: A 12-year Prospective Study in China. J Epidemiol Glob Health. 2024;14(3):1332–40.
28. Wang R, Zhang W, Li Y, Jiang Y, Feng H, Du Y, et al. Evaluation of Risk Factors for Chronic Obstructive Pulmonary Disease in the Middle-Aged and Elderly Rural Population of Northeast China Using Logistic Regression and Principal Component Analysis. Risk Manag Healthc Policy. 2022;15:1717–26.
29. Khatib S, Sabobeh T, Jaber F, Abdalla K, Singh S, Salzman G. Use of Laboratory Tests and Their Prognostic Value in Patients with Stable Chronic Obstructive Pulmonary Disease. Mo Med. 2022;119(6):545–52.
30. Zuoyou L, Shiota S, Morio Y, Sugiyama A, Sekiya M, Iwakami SI, et al. Borderline pulmonary hypertension associated with chronic hypercapnia in chronic pulmonary disease. Respir Physiol Neurobiol. 2019;262:20–5.
31. Reshma SJ, George S. Pulmonary hypertension in newly diagnosed obstructive sleep apnea-chronic obstructive pulmonary disease overlap syndrome patients attending a tertiary care centre-a cross-sectional analysis. Ir J Med Sci. 2024;193(4):1917–21.
32. Boersma R, Bakker JT, de Vries M, Raveling T, Slebos DJ, Wijkstra PJ, et al. Defining a phenotype of severe COPD patients who develop chronic hypercapnia. Respir Med. 2024;234:107850.

33. Theodosiou AA, Read RC. Artificial intelligence, machine learning and deep learning: Potential resources for the infection clinician. J Infect. 2023;87(4):287–94.
34. Cote MP, Lubowitz JH, Brand JC, Rossi MJ. Artificial Intelligence, Machine Learning, and Medicine: A Little Background Goes a Long Way Toward Understanding. Arthroscopy. 2021;37(6):1699–702.

Figures

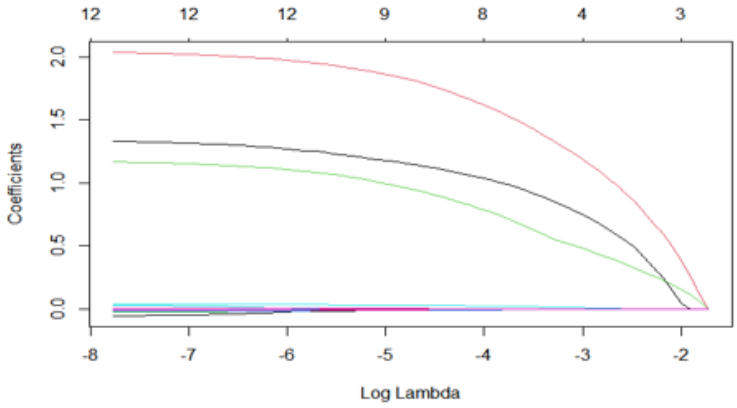


Figure 1

Fig.1 LASSO coefficient path graph.

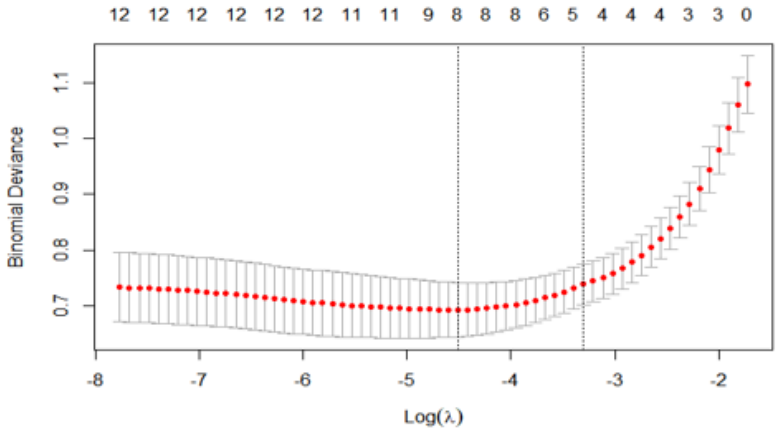


Figure 2

Fig.2 LASSO regularization path graph.

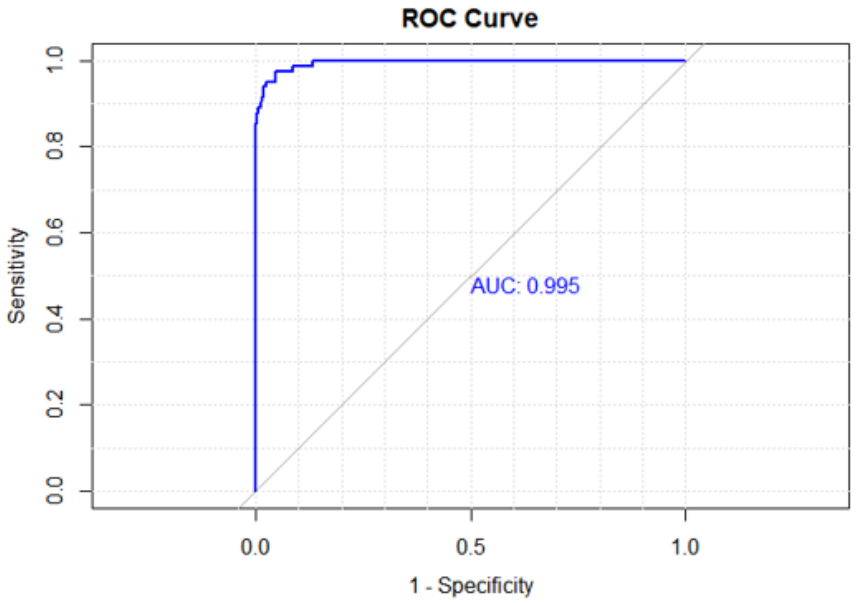


Figure 3

Fig.3A ROC curve of the model on the training set.

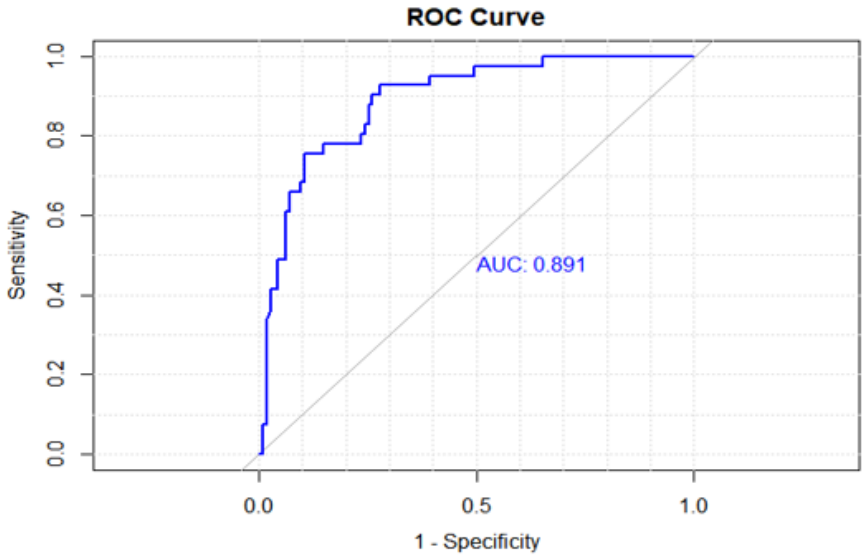


Figure 4

Fig.3B ROC curve of model in testing set.

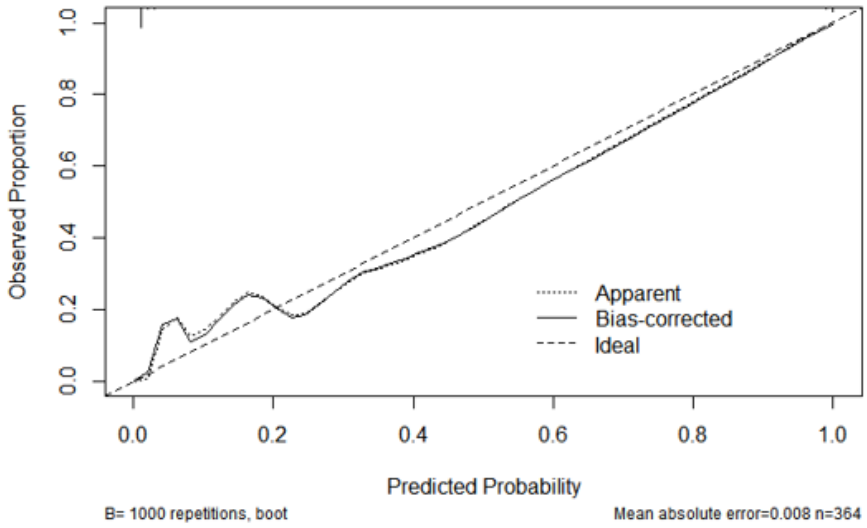


Figure 5

Fig.4A Calibration curve of the model on the training set.

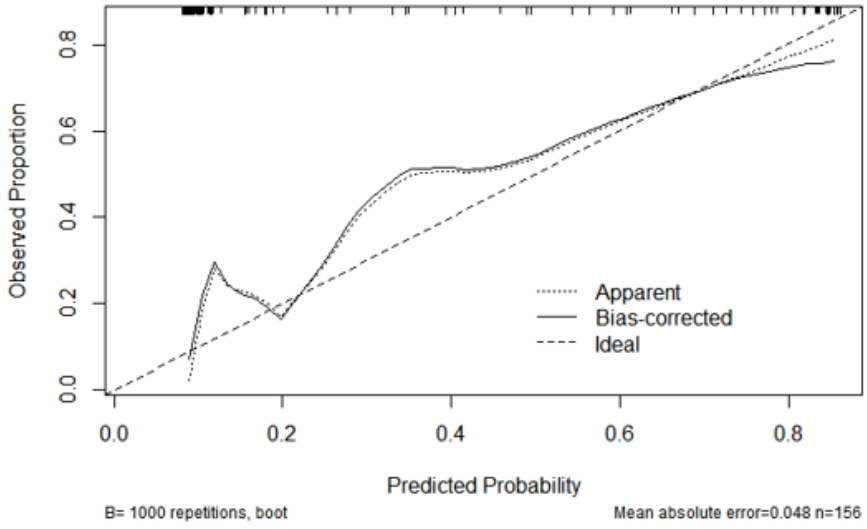


Figure 6

Fig.4B Calibration curve of the model on the testing set.

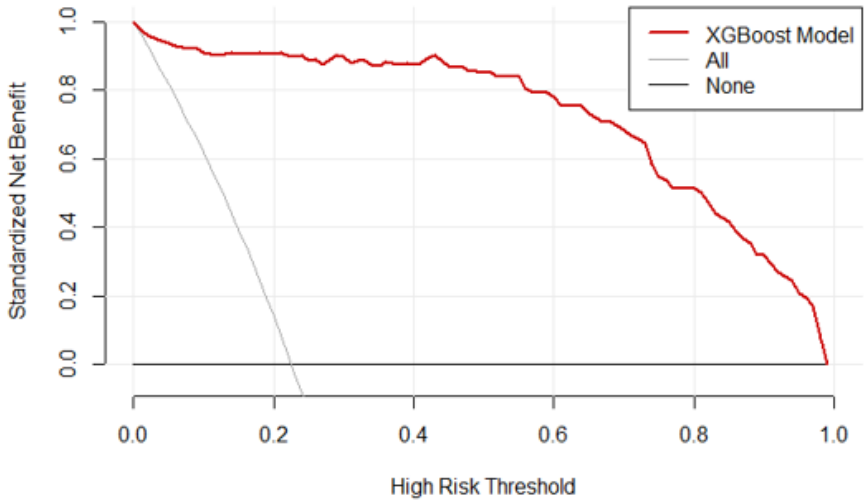


Figure 7

Fig.5A Clinical decision curve of the model on the training set.

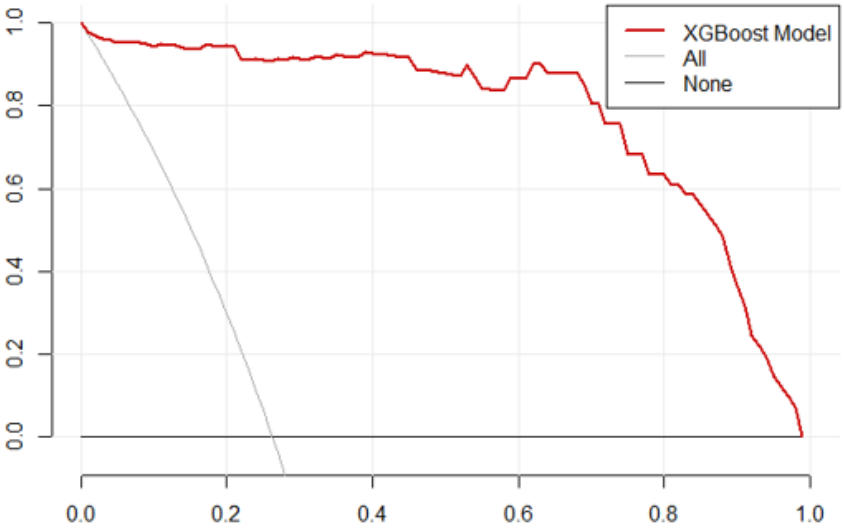


Figure 8

Fig.5B Clinical decision curve of the model on the testing set.

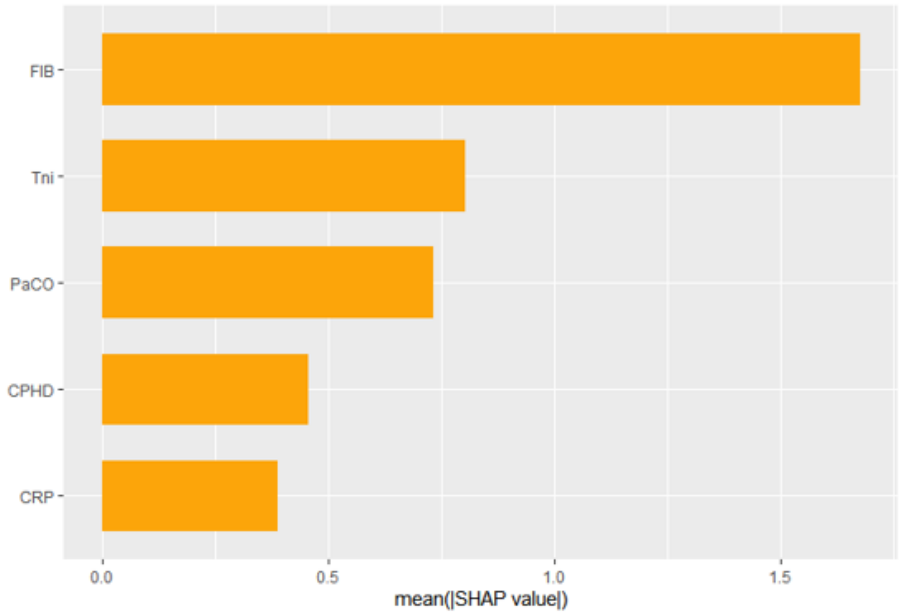


Figure 9

Fig. 6 The feature importance plot of SHAP variables.

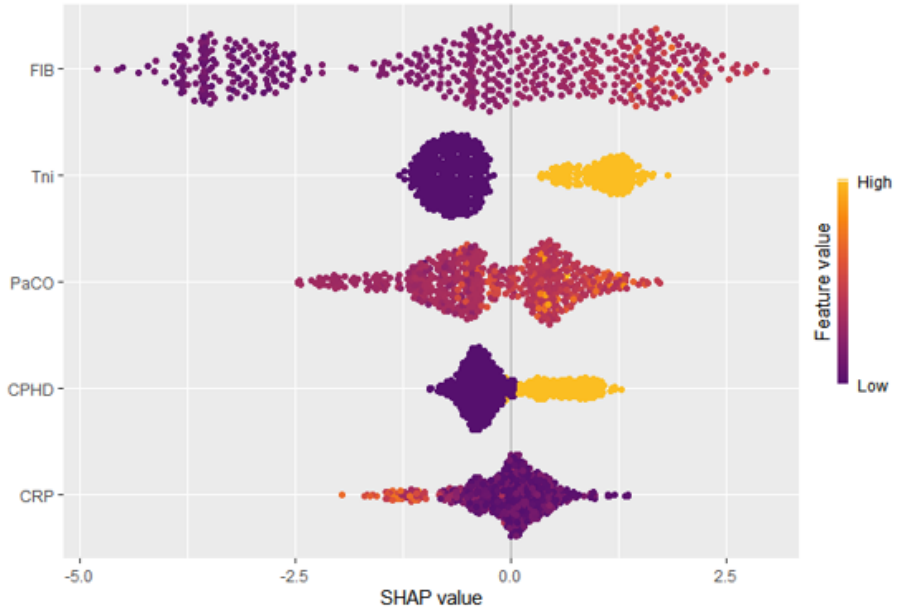


Figure 10

Fig. 7 The beeswarm plot of SHAP variables.