Am J Transl Res 2024;16(7):2864-2876
www.ajtr.org /ISSN:1943-8141/AJTR0156724

Original Article

Predictive value of gradient boosting decision trees for postoperative atelectasis complications in patients with pulmonary destruction

Zhongming Tang, Jifu Tang, Wei Liu, Guoqiang Chen, Chenggang Feng, Aiping Zhang

Department of Thoracic Surgery, Chest Hospital of Guangxi Zhuang Autonomous Region, Liuzhou, Guangxi Zhuang Autonomous Region, China

Received March 21, 2024; Accepted May 28, 2024; Epub July 15, 2024; Published July 30, 2024

Abstract: Objective: To explore the application value of a gradient boosting decision tree (GBDT) in predicting postoperative atelectasis in patients with destroyed lungs. Methods: A total of 170 patients with damaged lungs who underwent surgical treatment in Chest Hospital of Guangxi Zhuang Autonomous Region from January 2021 to May 2023 were retrospectively selected. The patients were divided into a training set (n = 119) and a validation set (n = 51). Both GBDT algorithm model and Logistic regression model for predicting postoperative atelectasis in patients were constructed. The receiver operating characteristic (ROC) curve, calibration curve and decision curve were used to evaluate the prediction efficiency of the model. Results: The GBDT model indicated that the relative importance scores of the four influencing factors were operation time (51.037), intraoperative blood loss (38.657), presence of lung function (9.126) and sputum obstruction (1.180). Multivariate Logistic regression analysis revealed that operation duration and sputum obstruction were significant predictors of postoperative atelectasis among patients with destroyed lungs within the training set (P = 0.048, P = 0.002). The ROC curve analysis showed that the area under the curve (AUC) for GBDT and Logistic model in the training set was 0.795 and 0.763, and their AUCs in the validation set were 0.776 and 0.811. The GBDT model's predictions closely matched the ideal curve, showing a higher net benefit than the reference line. Conclusions: GBDT model is suitable for predicting the incidence of complications in small samples.

Keywords: Destroyed lung, atelectasis, gradient boosting decision tree, prediction, machine learning

Introduction

Destroyed lung refers to the extensive tissue damage caused by fungal infection, silicosis, extensive bronchiectasis or tuberculosis, alongside recurrent pulmonary bacteria. This condition leads to irreversible functional loss, presenting as long-term and recurrent clinical episodes. Due to the presence of abundant microorganisms, the lung becomes a major infection source, leading to repeated infections [1]. Normally, CD4(+) and CD8(+) T cells can identify and attack foreign pathogens and abnormal cells. However, vascular destruction inside the destroyed lung prevents these cells from reaching the lesion, impairing the protective function of the human immune system [2]. Additionally, the characteristically thick-walled lesions hinders drug penetration, which complicates conservative drug treatments [3]. Therefore, surgical treatment is an option for patients with destroyed lungs.

Postoperative pulmonary complications, including atelectasis which accounts for 8.4% of such complications, represent a significant risk during the perioperative period of thoracic surgery, with an incidence ranging from 15% to 40%. Atelectasis forms a pathophysiological basis linked to potentially preventable morbidity and mortality [4, 5]. The complications include hypoventilation or pneumonia. While significant atelectasis can manifest with hypoxemic symptoms, other symptoms are often attributable to the underlying cause or superimposed pneumonia. The small airways within the atelectatic region are subject to repeated collapse and reopening during breathing, exacerbating lung injury.

https://doi.org/10.62347/IEQE3348

Early prediction and intervention of postoperative atelectasis in patients with destroyed lungs are crucial for reducing incidence rates. However, obtaining large study samples is often impractical, especially for rare diseases. Here, the Gradient Boosting Decision Tree (GBDT) model, an ensemble decision tree approach suited for small sample sizes and effective in supervised learning classification, is explored [6]. GBDT is a machine learning algorithm that can effectively identify the influencing factors from a large amount of data, regardless of the original modelling rules. Compared with the traditional Logistic regression model, the GBDT model is more suitable for processing clinically unordered data, such as gender and disease type, due to its high interpretability [7]. However, no existing literature explores the use of GBDT for predicting postoperative atelectasis in patients with destroyed lungs. This study aims to fill that gap by analyzing the factors influencing postoperative atelectasis using the GBDT model, thereby offering insights for postoperative care and treatment strategies.

Materials and methods

Subjects

A total of 170 patients who were pathologically diagnosed with destroyed lungs and received surgical treatment in the Chest Hospital of Guangxi Zhuang Autonomous Region between January 2021 and May 2023 were retrospectively selected. Of these, 25 patients developed atelectasis while 145 did not. This study was approved by the ethics committee of the Chest Hospital of Guangxi Zhuang Autonomous Region. Inclusion criteria: (1) patients with an age of 18 years or older; (2) patients who received surgical treatment in our hospital; (3) patients with a diagnosis of destroyed lung by chest CT, displaying a large high-density shadow in the damaged lung, extensive fibrosis, and tubular or cystic bronchiectasis with necrosis and cavitation [8]; (4) patients with complete clinical data, encompassing baseline information, and pre-, intra-, and postoperative metrics. Exclusion criteria: (1) patients with a prior history of chest surgery; (2) patients with concurrent pulmonary conditions, such as asthma, chronic obstructive pulmonary disease (COPD), and lung cancer. The study design is shown in **Figure 1**.

Data collection

The data of patients were collected from the hospital's electronic medical record system, including baseline data [gender, age, height, weight, history of smoking, diabetes, hypertension, COPD (chronic obstructive pulmonary disease), bronchiectasis, type of lung damage, electrolyte abnormalities], preoperative indicators [lung function, preoperative fasting blood glucose, white blood cell count, neutrophil count, platelet count, fibrinogen, CRP (C-reactive protein), hs-CRP (high sensitive C-reactive protein)], intraoperative indicators (type of operation, operation time, intraoperative blood loss), and postoperative indicators (postoperative pain score, hypoxemia, pleural effusion, sputum obstruction). The preoperative indicators of the patients were measured the day before the operation, and the postoperative indicators were taken as the values of the first measurement of each indicator after the operation.

Model construction and validation

The model was built using R (version 4.2.1). All patients were randomly divided into a training set (n = 119, including 19 atelectasis patients and 100 non-atelectasis patients) and a validation set (n = 51, including 6 atelectasis patients and 45 non-atelectasis patients) according to a ratio of 7:3. The training set was used for model construction and the validation set was used to evaluate the prediction performance of the model. With the complication of atelectasis as the outcome variable, the GBDT model and logistic regression model were established using the training set. The predictive performance of the model was evaluated in the validation set, by calculating and drafting the receiver operator characteristic (ROC) curve, calibration curve, and decision curve of the predictive model. Delong test was used to compare the difference in AUCs (area under the curve) between the different models.

Statistical methods

SPSS 23.0 was used for statistical analysis. Quantitative data conforming to a normal distribution were expressed as ($\bar{x} \pm s$), and the intergroup difference was examined by independent sample t-test. The quantitative data conforming to skewed distribution were expressed as medi-

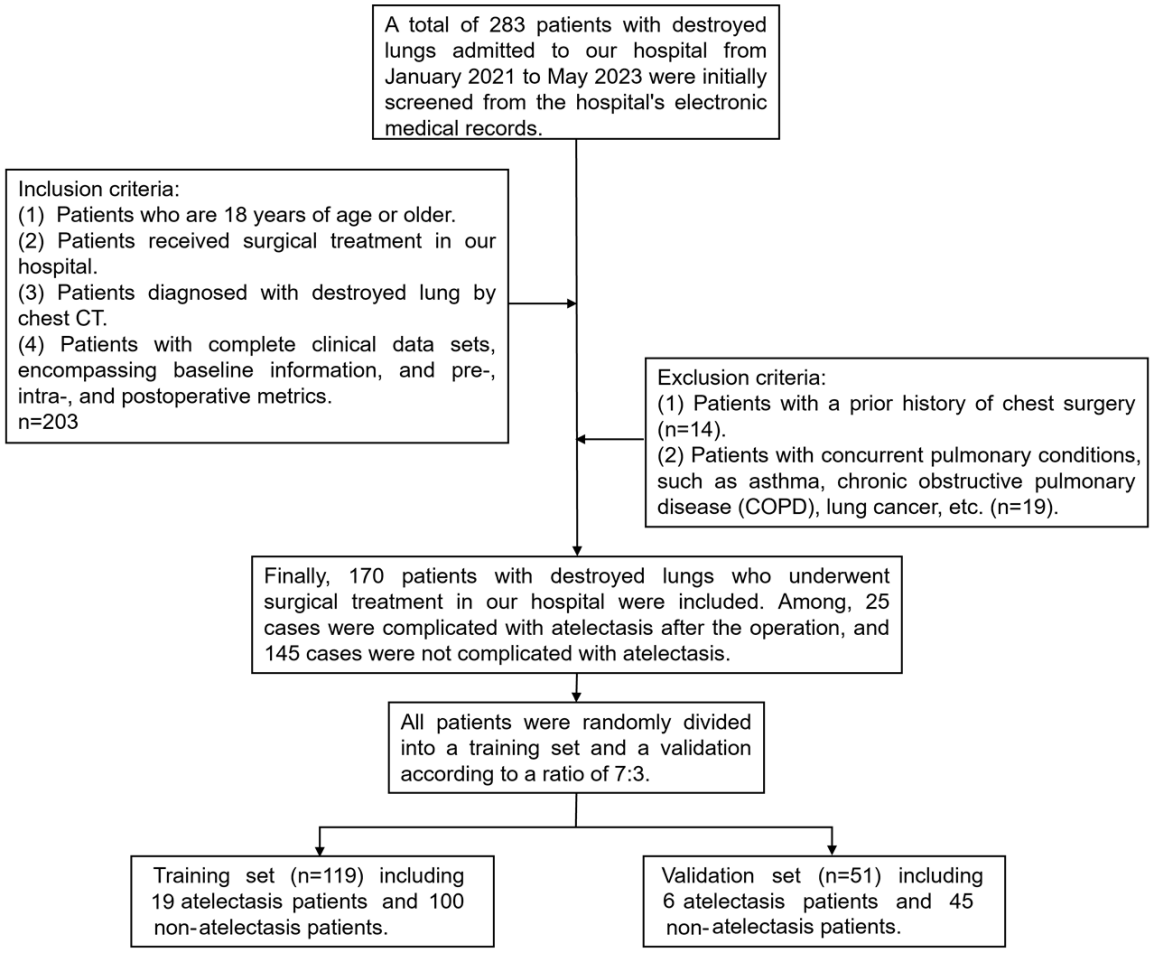

Figure 1. Study flow chart.

an (M) and quartile (P_{25}, P_{75}), and the comparison between groups was performed using the Wilcoxon rank sum test. Count data were shown as count (%) and inter-group difference was examined by x^2 tests. $P<0.05$ was considered as statistically significant.

Results

Single-factor analysis of clinical data of patients with destroyed lung

The single factor analysis of the clinical data showed that the differences in lungs function ($x^2 = 5.170$, $P = 0.023$), operation duration ($Z = -3.135$, $P = 0.002$), intraoperative bleeding volume ($t = 2.186$, $P = 0.039$), and sputum obstruction ($x^2 = 29.899$, $P<0.001$) between the postoperative atelectasis group and the non-atelectasis group were statistically significant. Other variables, such as gender, age, height, weight, history of smoking, diabetes, hypertension, presence of COPD, presence of bronchiectasis, type of lung damage, preoperative fasting blood glucose, white blood cell count, neutrophil count, platelet count, fibrinogen, CRP, hs-CRP, electrolyte abnormalities, operation type, postoperative pain score, hypoxemia, and pleural effusion, were not statistically different between the two groups (all $P>0.05$). The results in detail are shown in **Table 1**.

GBDT model and logistic model construction

All patients were randomly divided into a training set (n = 119) and a validation set (n = 51) according to a 7:3 ratio. There was no difference in patient data between the training set and the validation set (**Table 2**). Indicators with statistical significance in single-factor analysis (lung function, operation duration, intraopera-

Table 1. Single factor analysis of clinical data

Variable	Atelectasis (n = 25)	Non-atelectasis (n = 145)	$\chi^2/t/Z$	*P*
Gender			3.353	0.067
Male	22 (88.00)	98 (67.59)		
Female	3 (12.00)	47 (32.41)		
Age (years)	49.00 ± 13.63	46.67 ± 14.07	0.649	0.522
Height (cm)	164.76 ± 6.88	163.51 ± 8.68	0.913	0.370
Body weight (kg)	54.80 ± 9.19	54.65 ± 10.78	-0.225	0.824
Smoking history			0.121	0.728
Yes	3 (12.00)	11 (7.59)		
No	22 (88.00)	134 (92.41)		
Pulmonary function			5.170	0.023
With	13 (52.00)	42 (28.97)		
Without	12 (48.00)	103 (71.03)		
Diabetes			0.050	0.823
Yes	1 (4.00)	11 (7.59)		
No	24 (96.00)	134 (92.41)		
Hypertension			0.018	0.892
Yes	1 (4.00)	5 (3.45)		
No	24 (96.00)	140 (96.55)		
COPD			1.289	0.256
Yes	4 (16.00)	10 (6.90)		
No	21 (84.00)	135 (93.10)		
Bronchiectasis			0.038	0.845
Yes	2 (8.00)	10 (6.90)		
No	23 (92.00)	135 (93.10)		
Destroyed area			7.449	0.114
Upper left lobe	11 (44.00)	56 (38.62)		
Lower left lobe	2 (8.00)	36 (24.83)		
Upper right lobe	10 (40.00)	64 (44.14)		
Right middle lobe	0 (0.00)	13 (8.97)		
Lower right lobe	2 (8.00)	7 (4.83)		
Fasting blood glucose (mmol/L)	4.95 (4.39, 6.72)	5.10 (4.62, 5.88)	-0.229	0.819
Electrolyte abnormalities			0.095	0.758
Yes	8 (32.00)	42 (28.97)		
No	17 (68.00)	103 (71.03)		
White blood cells (10^9/L)	6.73 ± 1.85	7.09 ± 2.37	-1.087	0.288
Neutrophils (10^9/L)	4.36 ± 1.58	4.61 ± 2.17	-0.658	0.517
Platelets (10^9/L)	266.00 (218.5, 310.00)	269.00 (217.00, 327.00)	-1.063	0.288
Fibrinogen (g/L)	3.37 (2.42, 4.28)	3.15 (2.45, 4.17)	-1.171	0.241
CRP (mg/L)	12.80 (4.40, 12.80)	11.60 (2.25, 18.34)	-0.057	0.954
hs-CRP (mg/L)	5.20 (2.25, 8.18)	6.80 (2.15, 6.87)	-0.657	0.511
Surgery type			0.003	0.959
Minimally invasive	14 (56.00)	82 (56.55)		
Open	11 (44.00)	63 (43.45)		
Surgical method			1.954	0.162
Pneumonectomy	1 (4.00)	25 (17.24)		
Partial lobectomy	24 (96.00)	120 (82.76)		
Operation duration(h)	5.47 (4.26, 6.59)	4.03 (3.11, 4.83)	-3.135	0.002

Intraoperative bleeding volume (ml)	1620.00 ± 1084.36	1177.38 ± 1121.44	2.186	0.039
Postoperative pain score	6.36 ± 1.50	6.40 ± 1.29	-0.104	0.918
Pleural effusion			0.898	0.343
Yes	3 (12.00)	7 (4.83)		
No	22 (88.00)	138 (95.17)		
Sputum obstruction			29.899	<0.001
Yes	7 (28.00)	7 (4.83)		
No	18 (72.00)	138 (95.17)		
Postoperative hypoxemia			0.360	0.548
Yes	4 (16.00)	14 (9.66)		
No	21 (84.00)	131 (90.34)		

COPD: chronic obstructive pulmonary disease; CRP: C-reactive protein; hs-CRP: high sensitive C-reactive protein.

Table 2. Comparison of clinical data between the training set and validation set

Variable	Training set (n = 119)	Validation set (n = 51)	$\chi^2/t/Z$	*P*
Gender			0.540	0.463
Male	86 (72.27)	34 (66.67)		
Female	33 (27.73)	17 (33.33)		
Age (years)	48.97 ± 13.30	44.74 ± 14.63	1.844	0.067
Height (cm)	163.71 ± 8.65	163.65 ± 8.00	0.288	0.775
Body weight (kg)	54.95 ± 10.54	54.02 ± 10.61	0.210	0.834
Smoking history			0.566	0.452
Yes	11 (9.24)	3 (5.88)		
No	108 (90.76)	48 (94.12)		
Pulmonary function			1.568	0.211
With	42 (35.29)	13 (25.49)		
Without	77 (64.71)	38 (74.51)		
Diabetes			0.067	0.796
Yes	8 (6.72)	4 (7.84)		
No	111 (93.28)	47 (92.16)		
Hypertension			0.586	0.444
Yes	5 (4.20)	1 (1.96)		
No	114 (95.80)	50 (98.04)		
COPD			2.039	0.153
Yes	12 (10.08)	2 (3.92)		
No	107 (89.92)	49 (96.08)		
Bronchiectasis			0.001	0.980
Yes	9 (7.56)	3 (5.88)		
No	110 (92.44)	48 (94.12)		
Destroyed area			1.260	0.868
Upper left lobe	102 (85.71)	16 (31.37)		
Lower left lobe	52 (43.70)	12 (23.53)		
Upper right lobe	104 (87.39)	22 (43.14)		
Right middle lobe	18 (15.13)	4 (7.84)		
Lower right lobe	12 (10.08)	3 (5.88)		
Fasting blood glucose (mmol/L)	5.10 (4.60, 6.08)	4.89 (4.46, 5.74)	-0.131	0.896
Electrolyte abnormalities			1.214	0.270
Yes	32 (26.89)	18 (35.29)		
No	87 (73.11)	33 (64.71)		

White blood cells (10^9/L)	7.11 ± 2.41	6.87 ± 2.03	1.198	0.237
Neutrophils (10^9/L)	4.68 ± 2.20	4.33 ± 1.81	1.138	0.260
Platelets (10^9/L)	266.00 (223.00, 236.00)	293.00 (232.00, 360.00)	-0.056	0.955
Fibrinogen (g/L)	3.20 (2.50, 4.19)	3.14 (2.27, 4.17)	-0.994	0.320
CRP (mg/L)	12.20 (3.10, 18.34)	11.90 (2.10, 18.34)	-0.393	0.694
hs-CRP (mg/L)	6.87 (3.30, 6.87)	5.40 (1.10, 6.87)	-0.630	0.529
Surgery type			0.164	0.685
Minimally invasive	66 (55.46)	30 (58.82)		
Open	53 (44.54)	21 (41.18)		
Surgical method			0.138	0.710
Pneumonectomy	19 (15.97)	7 (13.73)		
Partial lobectomy	100 (84.03)	44 (86.27)		
Operation duration (h)	4.08 (3.17, 5.50)	3.83 (3.17, 4.92)	-0.492	0.623
Intraoperative bleeding volume (ml)	1294.62 ± 1195.66	1120.78 ± 935.13	-0.340	0.735
VAS score	6.50 ± 1.31	6.14 ± 1.31	0.505	0.615
Pleural effusion			0.482	0.488
Yes	6 (5.04)	4 (7.84)		
No	113 (94.96)	47 (92.16)		
Sputum obstruction			0.566	0.452
Yes	11 (9.24)	3 (5.88)		
No	108 (90.76)	48 (94.12)		
Postoperative hypoxemia			0.107	0.744
Yes	12 (10.08)	6 (11.76)		
No	107 (89.92)	45 (88.24)		

COPD: chronic obstructive pulmonary disease; CRP: C-reactive protein; hs-CRP: high sensitive C-reactive protein.

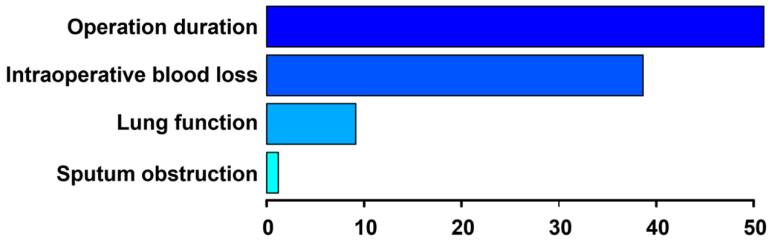

Figure 2. Relative importance of included features within the GBDT model. GBDT: Gradient Boosted Decision Tree.

tive blood loss, and sputum obstruction) were included in the GBDT model. The prediction model was constructed based on the training set data, and the relative importance score of the four indicators was obtained by GBDT. The importance scores in descending order were 51.037 for operation duration, 38.657 for intraoperative blood loss, 9.126 for presence of lung function, and 1.180 for sputum obstruction. The results are shown in **Figure 2**.

The operation duration (measured value), intraoperative blood loss (measured value), lung function (assignment: yes = 1, no = 0), and sputum obstruction (assignment: yes = 1, no = 0) were used as independent variables, while the presence of postoperative atelectasis was used as dependent variable (assignment: occurrence = 1, no occurrence = 0). Multivariate logistic stepwise regression analysis was performed. The results showed that after removing the insignificant factors, the operation duration and sputum obstruction were the influencing factors of postoperative atelectasis in patients with destroyed lungs in the training set ($P = 0.048$, $P = 0.002$). Based on this, a logistic regression model was constructed, and the formula is Y = -3.551 + 0.315 × operation duration + 2.281 × sputum obstruction, see **Table 3**.

GBDT model and logistic model validation

The ROC curve showed that the AUC of the GBDT model in the training set was 0.795 [95% CI (0.685, 0.906)], the specificity was 0.800,

Table 3. Influencing factors of postoperative atelectasis in patients with destroyed lungs

Variable	*β*	*SE*	*Wald* χ^2	*P*	*OR*	*95% CI*
Operation duration	0.315	0.159	3.923	0.048	1.371	1.003-1.873
Sputum obstruction	2.281	0.720	10.042	0.002	9.783	2.387-40.090
Constant	-3.551	0.860	17.064	<0.001	0.029	-

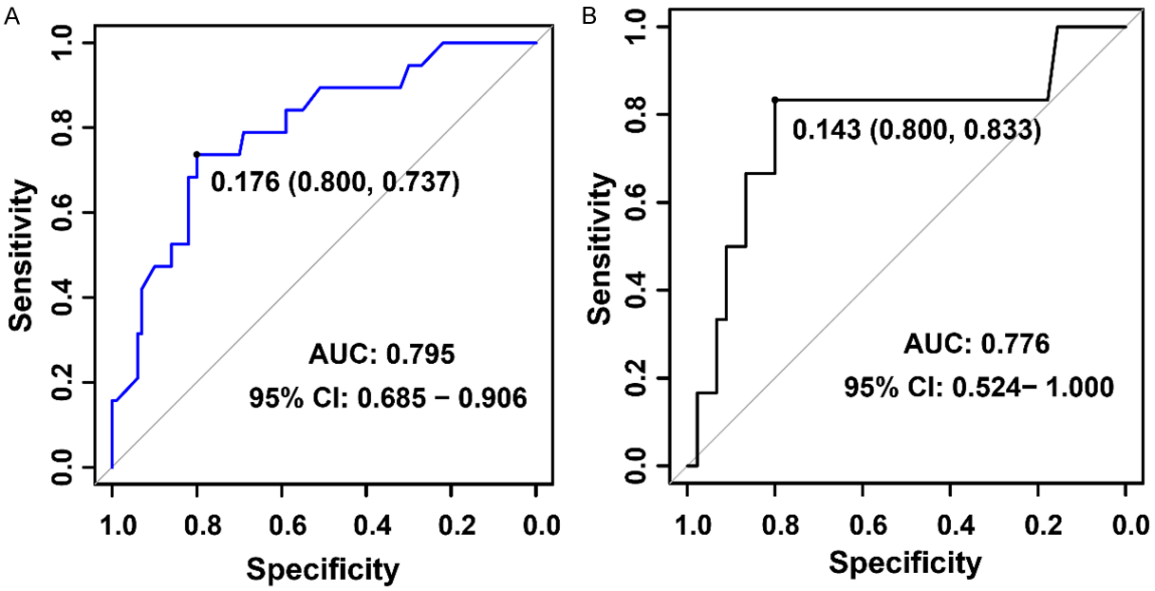

Figure 3. The ROC curves of GBDT model in the (A) training set and (B) validation set in predicting postoperative atelectasis in patients with destroyed lung. ROC: receiver operator characteristic curve; GBDT: Gradient Boosted Decision Tree.

and the sensitivity was 0.737; while the AUC in the validation set was 0.776 [95% CI (0.524, 1.000)], the specificity was 0.800, and the sensitivity was 0.833, as shown in **Figure 3**. The AUC of the logistic regression model in the training set was 0.763 [95% CI (0.620, 0.906)], the specificity was 0.820, and the sensitivity was 0.684; while the AUC in the validation set was 0.811 [95% CI (0.672, 0.950)], the specificity was 0.578, and the sensitivity was 1.000, as shown in **Figure 4**. The results of the Delong test showed that there was no significant difference in AUC between the training set and validation set of the GBDT model and the Logistic regression model (Z = 0.348, P = 0.728; Z= -0.415, P = 0.678). The calibration curve showed that the actual curves predicted by both GBDT model and logistic regression model were close to the ideal curve, as shown in **Figures 5**, **6**. The decision curve of the GBDT model showed that the net benefits for patients were higher than the two extreme curves (**Figure 7**), suggesting that the model had good clinical effectiveness. The decision curve of the logistic regression model showed that the net benefits for patients in the training set were higher than the two extreme curves (**Figure 8**).

Discussion

Destroyed lung is characterized by chronic fibrous lesion in lung tissues caused by pulmonary tuberculosis or severe pulmonary infection. During the disease progression, the lung tissue undergoes extensive damage, leading to irreversible lung function loss [9]. In the surgical treatment of destroyed lung, atelectasis is a common postoperative intrapulmonary complication. Due to the collapse of the lung caused by surgical trauma, inflammatory reactions and other factors, lung tissue cannot expand normally. Without timely intervention, it may progress to respiratory failure, posing a serious risk to patient survival [10]. Therefore, predicting postoperative atelectasis is of great significance for early prevention and timely treatment

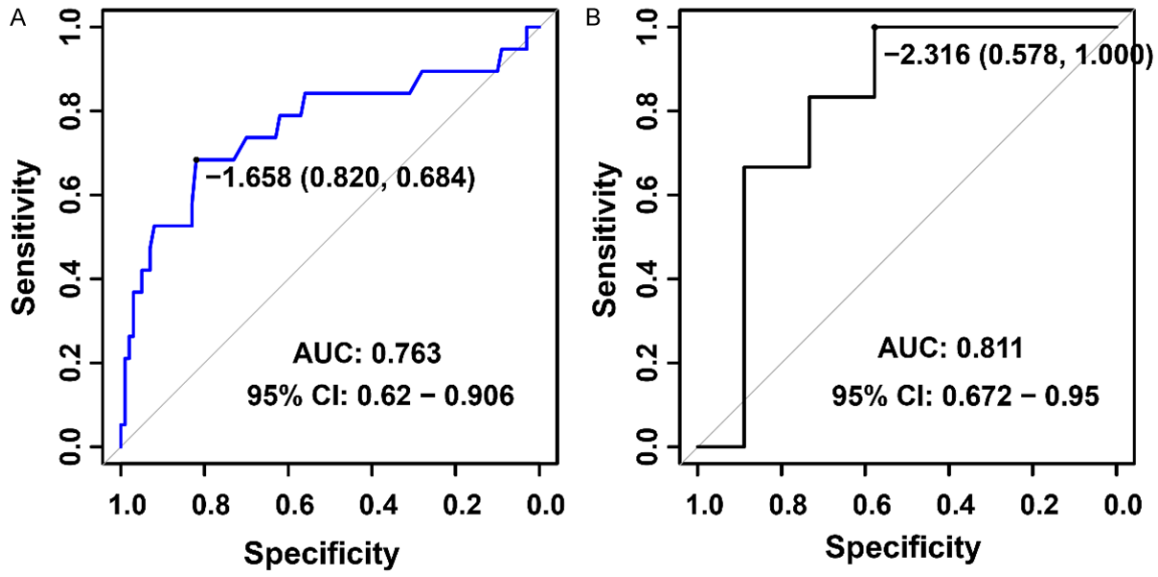

Figure 4. The ROC curves of Logistic model in the (A) training set and (B) validation set in predicting postoperative atelectasis in patients with destroyed lung. ROC: receiver operator characteristic curve.

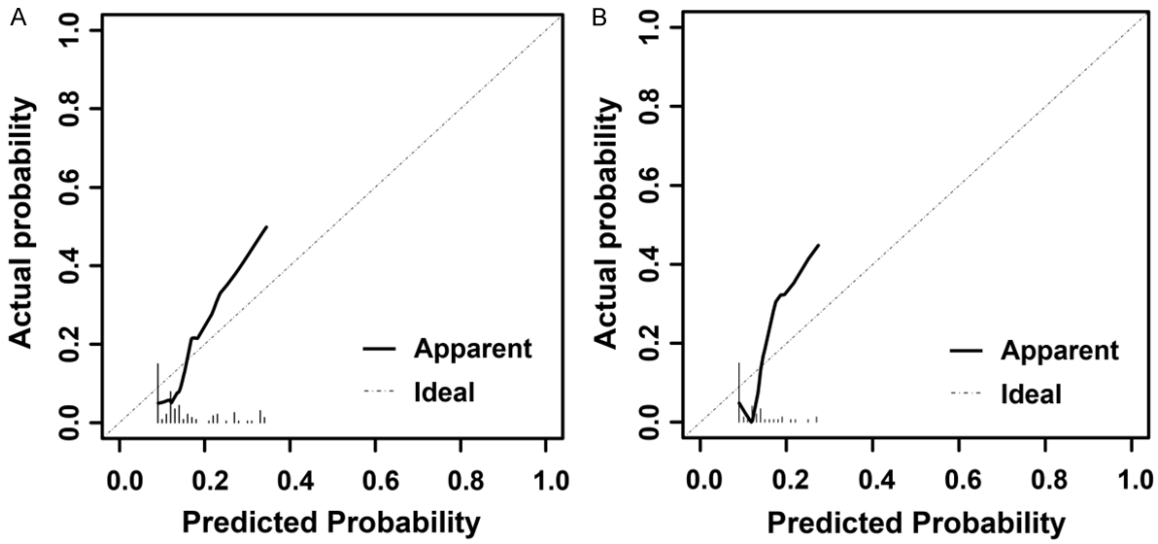

Figure 5. The calibration curves of GBDT model in the (A) training set and (B) validation set in predicting postoperative atelectasis in patients with destroyed lung. GBDT: Gradient Boosted Decision Tree.

in patients with destroyed lungs. In recent years, the Gradient Boosting Decision Tree (GBDT) model has received extensive attention and demonstrated ideal performance in practical applications [11-13].

In this study, the GBDT model was used to further verify the significance of four factors identified in the univariate difference analysis as influencing postoperative atelectasis in patients with destroyed lungs. The factors include operation duration, intraoperative bleeding volume, lung function, and sputum obstruction, with significance scores of 51.037, 38.657, 9.126, and 1.180 indicated by GBDT. Lung tissue damage can trigger an inflammatory response, leading to a series of physiological changes [14]. In addition to respiratory function, the lung is also involved in platelet production and hematopoiesis. A reduction in platelet count can result in increased intraoperative bleeding and prolonged operation, increasing

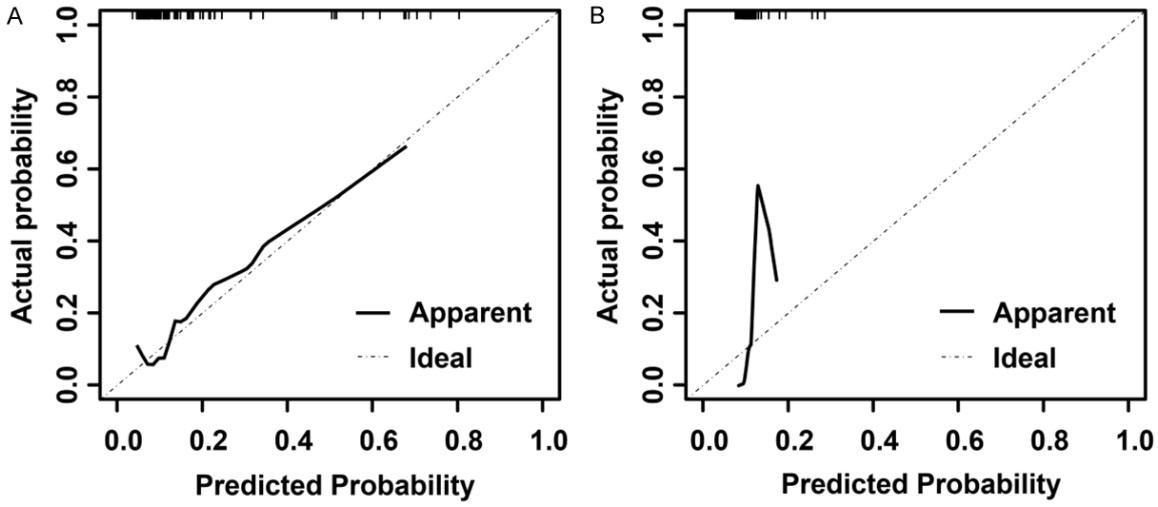

Figure 6. The calibration curves of Logistic model in the (A) training set and (B) validation set in predicting postoperative atelectasis in patients with destroyed lung.

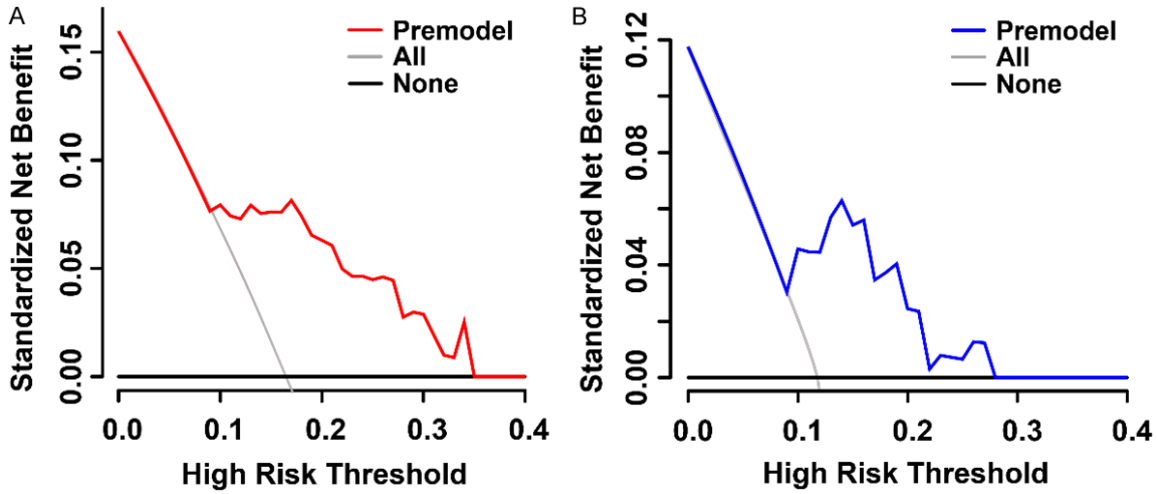

Figure 7. The decision curves of GBDT model in the (A) training set and (B) validation set in predicting postoperative atelectasis in patients with destroyed lung. GBDT: Gradient Boosted Decision Tree.

the risk and difficulty of the operation [15, 16]. Patients with destroyed lungs have serious adhesion at the lesion and require broader surgical dissection, which complicates the surgery and lengthens the operation time. Additionally, the blood vessels at the lesion may be damaged, and repeated bleeding or excessive blood loss may occur during the operation, consuming a large amount of coagulation factors. Meanwhile, prolonged surgery duration can also lead to long-term exposure of tissues, aggravating the degree of traumatic stress in patients, thus increasing the impact of surgery on body function and the risk of postoperative atelectasis [17, 18]. Therefore, both the surgery duration and the intraoperative blood loss significantly influence the likelihood of postoperative atelectasis complications. Hence, a clinically experienced surgery team is essential to prevent extended surgery times. Postoperative pulmonary infection is a common complication, stemming from a variety of reasons including bed rest, decreased immunity, epiglottic dysfunction and regurgitating aspiration [19]. These factors also contribute sputum obstructive atelectasis. In this study, the majority of patients were in critical condition with severely compromised lung function, making it challeng-

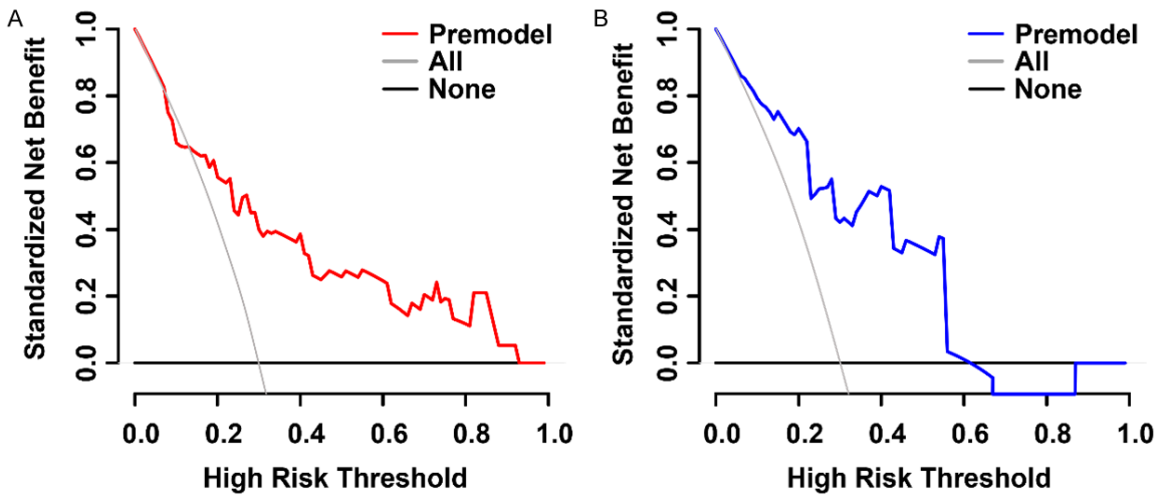

Figure 8. The decision curves of Logistic model in the (A) training set and (B) validation set in predicting postoperative atelectasis in patients with destroyed lung.

ing to accurately assess their pulmonary capacity using standard instruments. The prognosis for these patients is generally unfavorable. Hence, the assessment of pulmonary function is also a crucial prognostic indicator for predicting the likelihood of postoperative atelectasis in patients. Sputum obstruction occurs when an excess or unusually viscous sputum produced within the respiratory tract cannot be expectorated effectively, leading to partial or complete blockage of the airways. In individuals with compromised lung function, such as those with destroyed lung, the airway's innate clearance mechanisms may be diminished, causing a buildup of sputum. Additionally, lung structural changes can increase the likelihood of sputum obstruction in the respiratory tract, which can lead to atelectasis of lung segments, lobes or even complete atelectasis of one side. It can also lead to airway obstruction, ineffective excretion of infected necrotic material, difficulty in controlling pulmonary infection, significantly decreased lung ventilatory ability, and dyspnea in patients. With the progression of the disease, respiratory failure may even occur in severe cases, requiring rescue treatment with tracheal intubation and invasive ventilatory support, which seriously endangers the life of patients [20].

In contrast, in the Logistic model after stepwise regression, only two factors remained: operation time and sputum obstruction. This analysis reveals that operation duration is the most important factor for postoperative atelectasis in patients with destroyed lungs. Stepwise regression is a technique used to create a model that performs as well as the full variable model but with fewer factors. Essentially, it screens out the most useful independent variables by employing software tools and eliminates the factors that have less influence on the dependent variables, maintaining an effective fit similar to that achieved using all variables in the model [21]. Although sputum obstruction is included in the Logistic model by stepwise regression, it is directly related to atelectasis and can even represent atelectasis in a sense. Therefore, sputum obstruction has a large range of 95% CI in the Logistic model and its importance score is lowest in the GBDT model. Clinically, it is recommended to closely observe the symptoms of patients and discharge phlegm timely to avoid the occurrence of atelectasis.

The AUC reflects the accuracy and ability of the GBDT model to distinguish between target classes [22]. In this study, the AUCs of GBDT and logistic model for postoperative atelectasis prediction in patients with destroyed lungs was 0.795 and 0.763 in the training set, 0.776 and 0.811 in the validation set. The Delong test results show that the performances of the GBDT and logistic model were similar. However, the 95% CI for the Logistic regression model was broader, indicating model instability. This may be attributed to the limited sample size, suggesting that a larger sample size is necessary for the robustness of logistic regression

model. However, obtaining a sufficiently large sample size in every study is not always feasible, particularly when dealing with diseases of low incidence, such as rare conditions. A model capable of handling small sample sizes becomes critical for identifying disease risk factors. As an integrated model of the decision tree, GBDT has a good effect on the classification of supervised learning and is suitable for the case of small samples [6]. The model's predicted curve for postoperative atelectasis in patients with destroyed lungs was close to the ideal curve in the training set; however, the prediction effect in the validation set was slightly inferior to that in the training set. This discrepancy could be attributed to biased data extracted from the validation set, exacerbated by the small total number of samples and the low incidence rate of the condition. The net benefits of patients in the training set and the validation set were higher than the reference line, demonstrating the excellent predictive efficiency of the GBDT model for postoperative atelectasis in patients with destroyed lungs. Although the predictive efficacy of the two models is similar when evaluated using ROC curve, calibration curve and decision curve, the GBDT model excels in visually ranking the importance of variables. This feature greatly aids medical staff in more effectively identifying risk factors for postoperative atelectasis in patients with lung damage.

The GBDT algorithm model differs from the linear approach of the Logistic regression model. Its robust classification capabilities make it particularly effective in handling small sample sizes, addressing some of the limitations inherent in logistic regression models [23]. Concurrently, numerous studies have highlighted that the GBDT model can improve the efficiency of disease prediction. For instance, Chen et al. [13] demonstrated that the gradient-boosting model yielded promising prognostic predictions for patients with gastric cancer, a finding of significant clinical relevance in the big data era. Similarly, Askari et al. [24] reported that the gradient-boosting random forest technique outperformed other methods in forecasting the hospital stay duration for COVID-19 patients. The GBDT model processes data using multiple independent classifications and regression trees, which are integrated into a robust classifier to improve prediction accuracy and stability. This holistic approach not only expedites data processing but also mitigates the biases associated with missing data analysis [25]. Crucially, the modular architecture of the GBDT model facilitates both disassembly for in-depth analysis and interpretation, making it particularly practical for clinical professionals. Therefore, the established GBDT model had good reliability and clinical practicability. Doctors should pay special attention to the four indicators identified by the model before and after the operation to minimize the risk of postoperative atelectasis, which is of positive significance for improving the prognosis and survival rate of patients.

In summary, the GBDT model and the Logistic model have similar predictive efficacy in postoperative atelectasis in patients with destroyed lungs; however, GBDT model exhibits better stability when dealing with small samples. In clinical application, the appropriate model can be selected according to the actual situation. However, there are still certain limitations in this study. Being a retrospective single-center study, it may be susceptible to selection bias, which could affect the accuracy of predictive factor selection during model construction. In the future, multi-center prospective trials will be conducted to verify the accuracy of the conclusions, and external data will also be used for model verification to further improve the model for clinical reference.

Acknowledgements

We acknowledged the patients involved in this work. This work was supported by a Self-funded Research Project of Health Department of Guangxi Zhuang Autonomous Region: Z2009185.

Disclosure of conflict of interest

None.

Address correspondence to: Dr. Aiping Zhang, Department of Thoracic Surgery, Chest Hospital of Guangxi Zhuang Autonomous Region, No. 8 Yangjiaoshan Road, Yufeng District, Liuzhou 545005, Guangxi Zhuang Autonomous Region, China. Tel: +86-0772-3141280; E-mail: Zhangaiping5@sina.com

References

[1] Kim HC, Kim TH, Kim YJ, Rhee CK and Oh YM. Effect of tiotropium inhaler use on mortality in

patients with tuberculous destroyed lung: based on linkage between hospital and nationwide health insurance claims data in South Korea. Respir Res 2019; 20: 85.
[2] Tseng YL, Chang JM, Liu YS, Cheng L, Chen YY, Wu MH, Lu CL and Yen YT. The role of video-assisted thoracoscopic therapeutic resection for medically failed pulmonary tuberculosis. Medicine (Baltimore) 2016; 95: e3511.
[3] Vashakidze SA, Gogishvili SG, Nikolaishvili KG, Avaliani ZR, Chandrakumaran A, Gogishvili GS, Magee M, Blumberg HM and Kempker RR. Adjunctive surgery versus medical treatment among patients with cavitary multidrug-resistant tuberculosis. Eur J Cardiothorac Surg 2021; 60: 1279-1285.
[4] Mathis MR, Duggal NM, Likosky DS, Haft JW, Douville NJ, Vaughn MT, Maile MD, Blank RS, Colquhoun DA, Strobel RJ, Janda AM, Zhang M, Kheterpal S and Engoren MC. Intraoperative mechanical ventilation and postoperative pulmonary complications after cardiac surgery. Anesthesiology 2019; 131: 1046-1062.
[5] Song Q, Guo X, Zhang L, Yang L and Lu X. New approaches in the classification and prognosis of sign clusters on pulmonary CT images in patients with multidrug-resistant tuberculosis. Front Microbiol 2021; 12: 714617.
[6] Li Q, Zhao K, Bustamante CD, Ma X and Wong WH. Xrare: a machine learning method jointly modeling phenotypes and genetic evidence for rare disease diagnosis. Genet Med 2019; 21: 2126-2134..
[7] Eaton JE, Vesterhus M, McCauley BM, Atkinson EJ, Schlicht EM, Juran BD, Gossard AA, LaRusso NF, Gores GJ, Karlsen TH and Lazaridis KN. Primary sclerosing cholangitis risk estimate tool (PREsTo) predicts outcomes of the disease: a derivation and validation study using machine learning. Hepatology 2020; 71: 214-224.
[8] Jiang YH, Shen L, Dai YX, Sheng J and Liu XY. Clinical comparison of pulmonary lobectomy in patients with massive hemoptysis of pulmonary tuberculosis after bronchial artery embolization. Chinese Journal of Clinical Thoracic and Cardiovascular Surgery 2019; 26: 1190-1193.
[9] Varona Porres D, Persiva O, Pallisa E and Andreu J. Radiological findings of unilateral tuberculous lung destruction. Insights Imaging 2017; 8: 271-277.
[10] Hosoda C, Ishiguro T, Shimizu Y, Kanegane H and Takayanagi N. Mycobacteriumgenavense infection presenting as an endobronchial polyp and upper lobe atelectasis. Am J Respir Crit Care Med 2020; 202: e144-e145.
[11] Seto H, Oyama A, Kitora S, Toki H, Yamamoto R, Kotoku J, Haga A, Shinzawa M, Yamakawa M, Fukui S and Moriyama T. Gradient boosting decision tree becomes more reliable than logistic regression in predicting probability for diabetes with big data. Sci Rep 2022; 12: 15889.
[12] Li K, Shi Q, Liu S, Xie Y and Liu J. Predicting in-hospital mortality in ICU patients with sepsis using gradient boosting decision tree. Medicine (Baltimore) 2021; 100: e25813.
[13] Chen Q, Zhang J, Bao B, Zhang F and Zhou J. Large-scale gastric cancer susceptibility gene identification based on gradient boosting decision tree. Front Mol Biosci 2022; 8: 815243.
[14] Yang HH, Duan JX, Liu SK, Xiong JB, Guan XX, Zhong WJ, Sun CC, Zhang CY, Luo XQ, Zhang YF, Chen P, Hammock BD, Hwang SH, Jiang JX, Zhou Y and Guan CX. A COX-2/sEH dual inhibitor PTUPB alleviates lipopolysaccharide-induced acute lung injury in mice by inhibiting NLRP3 inflammasome activation. Theranostics 2020; 10: 4749-4761.
[15] Lefrançais E, Ortiz-Muñoz G, Caudrillier A, Mallavia B, Liu F, Sayah DM, Thornton EE, Headley MB, David T, Coughlin SR, Krummel MF, Leavitt AD, Passegué E and Looney MR. The lung is a site of platelet biogenesis and a reservoir for haematopoietic progenitors. Nature 2017; 544: 105-109.
[16] Larsen JB, Hvas AM and Hojbjerg JA. Platelet function testing: update and future directions. Semin Thromb Hemost 2023; 49: 600-608.
[17] Taylor M, Grant SW, West D, Shackcloth M, Woolley S, Naidu B and Shah R. Ninety-day mortality: redefining the perioperative period after lung resection. Clin Lung Cancer 2021; 22: e642-e645.
[18] Bigatello L and Östberg E. Pursuing the importance of postoperative atelectasis. Anesthesiology 2021; 135: 943-944.
[19] Sancho-Chust JN, Molina V, Vañes S, Pulido AM, Maestre L and Chiner E. Utility of flexible bronchoscopy for airway foreign bodies removal in adults. J Clin Med 2020; 9: 1409.
[20] Pan M, Fang G, Zheng F, Lin F, Zeng W, Qiu Y, Deng J, Chen X and Zhang J. Clinical characteristics of tracheobronchial Talaromyces marneffei infection in non-HIV-infected patients in South China. Ann Med 2023; 55: 2276310.
[21] Cui Y, Chen Z, Pan B, Chen T, Ding H, Li Q, Wan L, Luo G, Sun L, Ding C, Yang J, Tong X and Zhao J. Neddylation pattern indicates tumor microenvironment characterization and predicts prognosis in lung adenocarcinoma. Front Cell Dev Biol 2022; 10: 979262.
[22] Ngiam KY and Khor IW. Big data and machine learning algorithms for health-care delivery. Lancet Oncol 2019; 20: e262-e273.

[23] Li S, Lin Y, Zhu T, Fan M, Xu S, Qiu W, Chen C, Li L, Wang Y, Yan J, Wong J, Naing L and Xu S. Development and external evaluation of predictions models for mortality of COVID-19 patients using machine learning method. Neural Comput Appl 2023; 35: 13037-13046.

[24] Askari G, Rouhani MH and Sattari M. Prediction of length of hospital stay of COVID-19 patients using gradient boosting decision tree. Int J Biomater 2022; 2022: 6474883.

[25] Zea-Vera R, Ryan CT, Havelka J, Corr SJ, Nguyen TC, Chatterjee S, Wall MJ Jr, Coselli JS, Rosengart TK and Ghanta RK. Machine learning to predict outcomes and cost by phase of care after coronary artery bypass grafting. Ann Thorac Surg 2022; 114: 711-719.