Submitted: 09 October, 2022 Accepted: 25 November, 2022 Published: 08 September, 2023 DOI:10.22514/sv.2023.087

Open Access
Signa Vitae

ORIGINAL RESEARCH

XGBoost model predicts acute lung injury after acute pancreatitis

Weiwei Lu[1,2,†], Xi Chen[3,†], Wei Liu[4,†], Wenjie Cai[5], Shengliang Zhu[1], Yunkun Wang[6,*], Xiaosu Wang[1,*]

[1]Department of Gastroenterology, Yueyang Hospital of Integrated Traditional Chinese and Western Medicine, Shanghai University of Traditional Chinese Medicine, 200437 Shanghai, China
[2]Department of General Practice, Xinhua Hospital, Shanghai Jiaotong University School of Medicine, 200092 Shanghai, China
[3]Department of Emergency and Critical Care Medicine, Changzheng Hospital, Naval Medical University, 200003 Shanghai, China
[4]Department of Emergency, Xinhua Hospital, Shanghai Jiao Tong University School of Medicine, 200092 Shanghai, China
[5]School of Health Science and Engineering, University of Shanghai for Science and Technology, 200093 Shanghai, China
[6]Department of Pediatric Neurosurgery, Xinhua Hospital, Shanghai Jiaotong University School of Medicine, 200092 Shanghai, China

***Correspondence**
wangyunkun@xinhuamed.com.cn (Yunkun Wang);
xswangxs0084@163.com (Xiaosu Wang)

† These authors contributed equally.

Abstract

To develop an XGBoost model to predict the occurrence of acute lung injury (ALI) in patients with acute pancreatitis (AP). Using the case database of Xinhua Hospital affiliated to Shanghai Jiaotong University School of Medicine, 1231 cases suffering from AP were screened, and after 137 variables were identified, the clinical characteristics of the samples were statistically analyzed, and the data were randomly divided into a training set (75%) to build the XGBoost model and a test set (25%) for validation. Finally, the performance of the model was evaluated based on accuracy, specificity, sensitivity, and subject characteristics working characteristic curves. The model performance is also compared with that of three other commonly used machine learning algorithms (support vector machine (SVM), logistic regression, and random forest). The age and laboratory tests of patients with AP combined with ALI differed from those of patients without combined acute lung injury. The area under the receiver operating characteristic (ROC) curve of the test set after model evaluation was 0.9534, the specificity was 0.7333, and the sensitivity was 0.7857, with arterial partial pressure of oxygen, bile acid, aspartate transaminase, urea nitrogen, and arterial blood pH as its most important influencing factors. In this study, the XGBoost model has advantages compared with other three machine learning algorithms. The XGBoost model has potential in the application of predicting acute lung injury after acute pancreatitis.

Keywords

Acute pancreatitis; Acute lung injury; XGBoost; Predictive model

1. Introduction

Acute pancreatitis (AP) is a common digestive emergency and can be divided into mild acute pancreatitis (MAP) and severe acute pancreatitis (SAP), MAP patients have a good prognosis after treatment, while about 10%–20% of patients have SAP, whose disease progresses rapidly and variable, and the death rate can be as high as 20%–30% [1]. Among them, systemic inflammatory response syndrome (SIRS) in the early stage of the disease process leads to multi-organ failure such as acute lung injury (ALI), acute renal failure, shock, *etc.* can account for about 60% of the causes of death, and complications such as infection in the middle and late stages of the disease process account for about 40% of deaths [2]. Acute lung injury is a common complication in the course of SAP, with approximately 30% of patients developing acute respiratory distress syndrome (ARDS) [3]. Clinical treatment of acute lung injury has no specific drugs, most use mechanical ventilation, but the effect is limited, but at the same time easy to complicate ventilator-related lung injury, further aggravating the patient's condition [4]. Therefore, it is very important to predict the occurrence of ALI in AP patients in clinical work so as to guide the treatment. In recent years, the application of machine learning technology in the medical field has been intensified, and it has unique advantages in handling big data, high-dimensional data, and conducting prediction studies [5]. This study explores the XGBoost model to predict concurrent ALI in AP patients to provide some ideas for the diagnosis and treatment of the disease in clinical practice.

2. Materials and Methods

2.1 Data Acquisition and Ethics

The data were obtained from our hospital, using our case information database, in which information involving patient privacy has been hidden by the database system itself. A total of 1231 AP cases from 2015 to 2020 were screened. Screening criteria: all patients had AP as the primary diagnosis. Exclusion criteria: those with missing arterial blood gas examination.

2.2 Diagnostic criteria

Diagnostic criteria of AP: It meets the relevant diagnostic criteria of AP in the "Guidelines for the diagnosis and treatment of acute pancreatitis in China (2021)" compiled by the Pancreatic Surgery Group of the Chinese Medical Association's Surgery Branch in 2021 [6]: (1) persistent pain in the upper abdomen. (2) Serum amylase and/or lipase concentration at least 3 times higher than the upper limit of normal. (3) Abdominal imaging results show imaging changes consistent with acute pancreatitis. Acute pancreatitis can be diagnosed by meeting two of the above three criteria. Diagnostic criteria of ALI: The diagnostic criteria related to ALI in the Guidelines for the diagnosis and treatment of acute lung injury/acute respiratory distress syndrome (2006) compiled by the Critical Care Medicine Branch of the Chinese Medical Association in 2006 [7]: (1) acute onset; (2) oxygenation index (PaO_2/oxygen concentration (FiO_2)) $\leq$200 mmHg (1 mmHg = 0.133 kpa) (regardless of the level of positive end-expiratory pressure (PEEP)); (3) orthopantomograph showing patchy shadows in both lungs; (4) pulmonary artery pressure $\leq$18 mmHg, or no clinical evidence of increased left atrial pressure. If PaO_2/FiO_2 $\leq$300 mmHg and other criteria mentioned above were met, the diagnosis of ALI was made.

2.3 Selection and treatment of independent variables

Combining clinical experience with relevant paper studies [8, 9], 137 potential variables associated with acute pancreatitis and acute lung injury were initially screened. The 10 most significant influencing factors were finally established using the XGBoost classifier, which included age and laboratory tests (Table 1).

2.4 Modeling

An XGBoost model was built for machine learning with the open-source Python software (version 3.7, produced by Python Software Foundation). The data were randomly divided into a training set (75%) and a test set (25%), and the training set was subjected to a 10-fold cross-validation method to determine the optimal parameters (max-depth = 9, learning-rate = 0.3). The test set was used to evaluate the model performance.

The other dataset splitting strategy using 5-fold cross-validation method was applied to show whether the model performance was stable (Table 2).

2.5 Evaluation of the model

The models were evaluated after they were built to verify their suitability for application in the detection of diseases. In this study, the performance of the model was evaluated based on the calculation of accuracy, specificity, sensitivity, and receiver operating characteristic (ROC) curves.

2.6 Model Comparison

In addition, we built common machine learning algorithms such as SVM, logistic regression, and random forest to compare their evaluation results with the XGBoost model.

2.7 Statistical Methodology

Statistical analyses were performed according to the following formulas:

$$Sensitivity = \frac{TP}{TP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + TF + FP + FN} \tag{4}$$

$$F1 = \frac{2 \times Sensitivity \times Precision}{Sensitivity + Precision} \tag{5}$$

Where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

ROC curve analysis and the area under the curve (AUC) were calculated using the Scikit-learn package to compare the performance of each model.

3. Results

3.1 Basic clinical information of the patient

A total of 1231 patients with acute pancreatitis were included in the study, of which a total of 125 patients, or 10.15% of all patients, had combined lung injury.

3.2 XGBoost model results

Fig. 1 shows the test results of XGBoost model, out of 308 samples, 294 samples were correctly predicted, and its accuracy rate was 95.45%.

3.3 Model evaluation and variable weights

Table 3 shows the comparison of the performance of the four machine learning models, and it can be seen that the AUC

TABLE 1. Description of 10 indicators related to ALI.

	PaO_2	Bile acid	AST	Urea nitrogen	pH	Creatinine	GFR (MDRD)	Albumin	Na	Age	ALI
count	1231	1231	1231	1231	1231	1231	1231	1231	1231	1231	1231
mean	6.86	26.38	7.68	3.75	3.77	53.72	98.63	31.78	100.80	53.90	0.10
std	6.96	58.19	9.03	3.41	3.67	27.27	69.87	12.31	60.93	21.95	0.30
min	0.70	0.00	0.00	0.50	0.06	22.00	14.17	4.83	3.29	1.00	0.00
25%	0.70	1.10	3.00	0.50	0.06	34.00	58.96	31.00	3.29	38.00	0.00
50%	3.00	2.70	7.00	4.00	7.13	53.00	97.32	35.80	137.70	58.00	0.00
75%	12.10	12.20	10.00	5.60	7.40	69.00	127.20	39.40	140.00	69.00	0.00
max	30.41	441.50	140.00	41.10	7.58	357.00	574.79	51.80	147.00	101.00	1.00

AST: Aspartate Transaminase; GFR: Glomerular filtration rate; MDRD: Modified diet in renal disease.

TABLE 2. 5-fold cross-validation results using the XGBoost method.

XGBoost	Accuracy	F1	Precision	Sensitivity	Specificity	AUC
mean	0.957	0.748	0.896	0.648	0.992	0.953
std	0.014	0.094	0.052	0.120	0.003	0.015

AUC: area under the curve.

TABLE 3. Comparison of models' performance on ALI detection.

Algorithm	Accuracy	F1	Precision	Sensitivity	Specificity	AUC
XGBoost	0.955	0.759	0.733	0.786	0.978	0.953
SVM	0.886	0.546	0.447	0.700	0.906	0.856
Logistic Regression	0.883	0.514	0.432	0.633	0.910	0.863
Random Forest	0.945	0.712	0.724	0.700	0.975	0.945

AUC: area under the curve; SVM: support vector machine.

values of all four machine learning models are greater than 0.8, among which the AUC value of XGBoost model (0.9534) is the highest (Fig. 2), the precision is 0.7333, and the sensitivity is 0.7857, all of which are higher than the other three algorithmic models.

The variable importance ranking of the XGBoost model is shown in Fig. 3, where arterial partial pressure of oxygen, bile acid, glutamic oxalacetic transaminase, urea nitrogen, and arterial blood pH are the top five important variables heat map of the 10 most important influencing factors is shown in Fig. 4.

We also analyzed the distribution of specific cases for each of the 10 most important influencing factors, as shown in Fig. 5.

4. Discussion

In this study, an XGBoost model of acute pancreatitis combined with acute lung injury was developed, and a total of 137 indicators were included for the study, and finally the 10 most important influencing factors were derived, and the model was assessed to have a good predictive ability.

We have compared performance of the four machine learning models: XGBoost, SVM, Logistic Regression and Random Forest. XGBoost, eXterme Gradient Boosting, mainly uses a parallel Boosting tree and follows the boosting algorithm, which continuously generates a new tree based on the training and prediction results of the previous tree. It combines multi-threading, data compression, and slicing to make the algorithm as efficient as possible. SVM (support vector machine) is a binary classification model, which is based on the principle of finding a decision boundary or hyperplane in the feature space that maximizes its interval from the training set. Logistic Regression is essentially a linear classifier, so it does not handle correlation between features well. Although the results are average, it wins because the model is clear and the probability science behind it can stand up to criticism. The parameters it fits represent the impact of each feature on the result. Random forest can handle high-dimensional data and can obtain feature importance based on the Information Gain during splitting and can achieve good results in the case of unbalanced data. In our study XGBoost model has the best performance in data learning et analysis.

In recent years, many studies have applied XGBoost models to process data in medical fields such as genes, drugs, and diseases [10], demonstrates its high efficiency. In contrast, there are fewer studies on prediction models for the occurrence of ALI in AP patients. The final test set of this study showed an accuracy of 0.9545, an AUC value of 0.9534 for the model

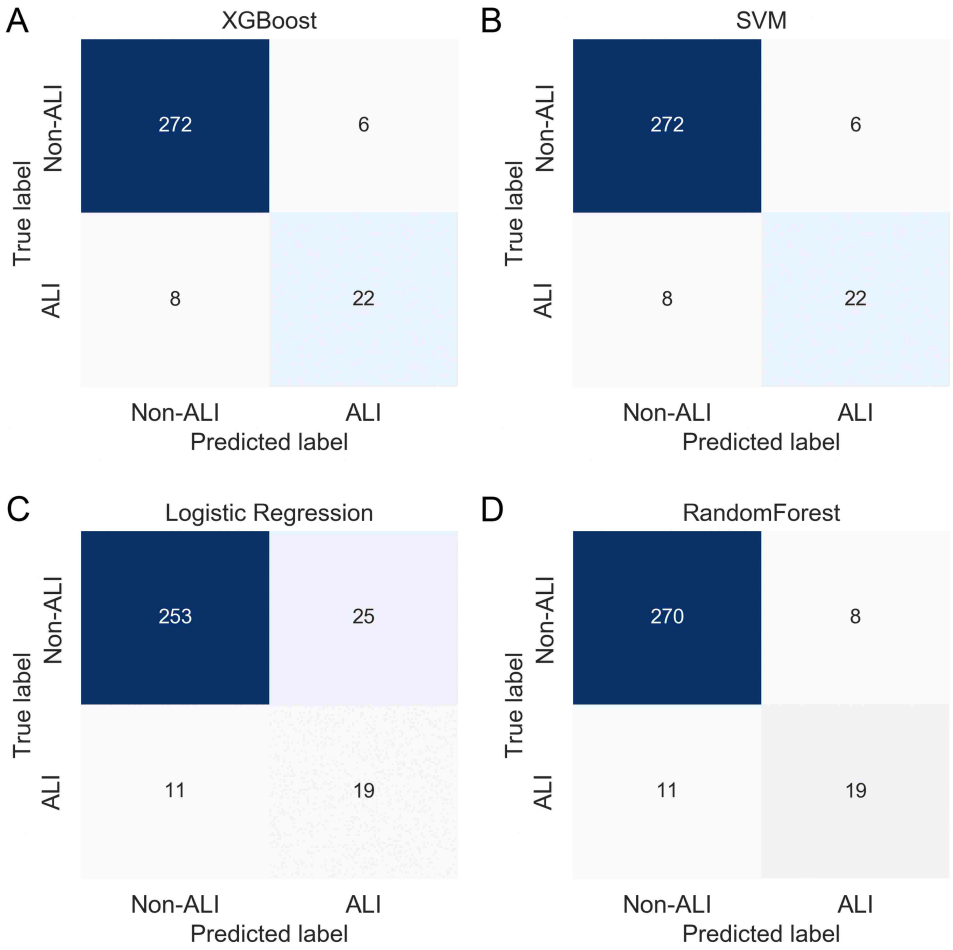

FIGURE 1. Confusion matrix for classification results. ALI: acute lung injury; SVM: support vector machine.

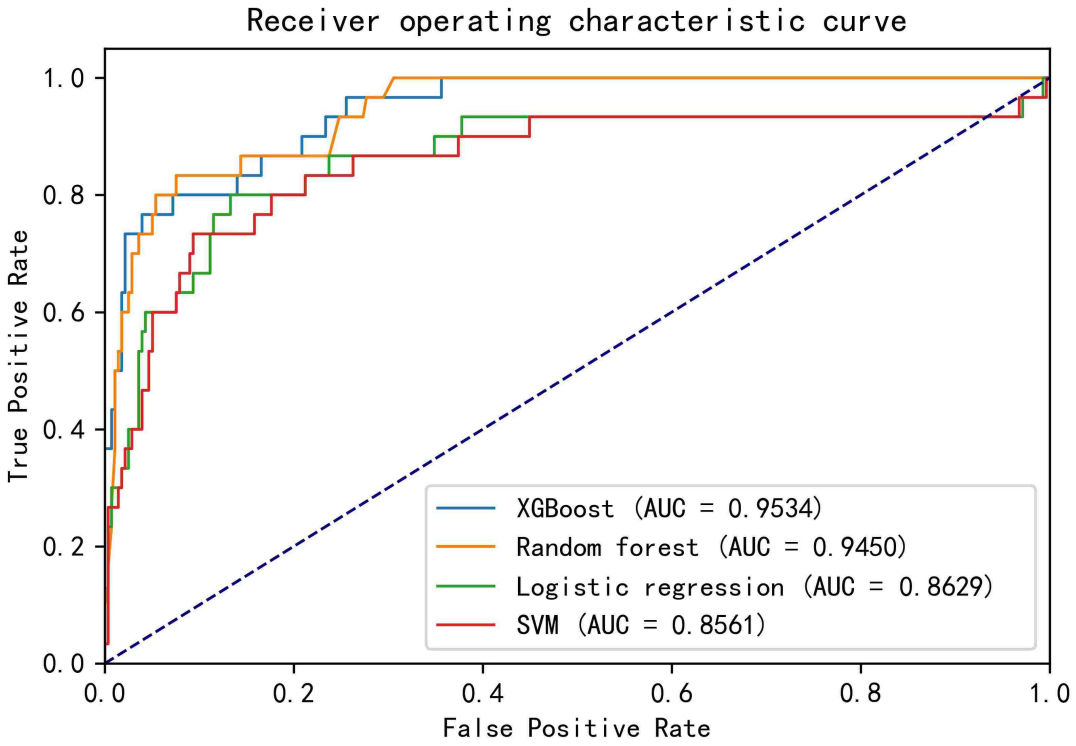

FIGURE 2. Receiver operating characteristic curve. AUC: area under the curve; SVM: support vector machine.

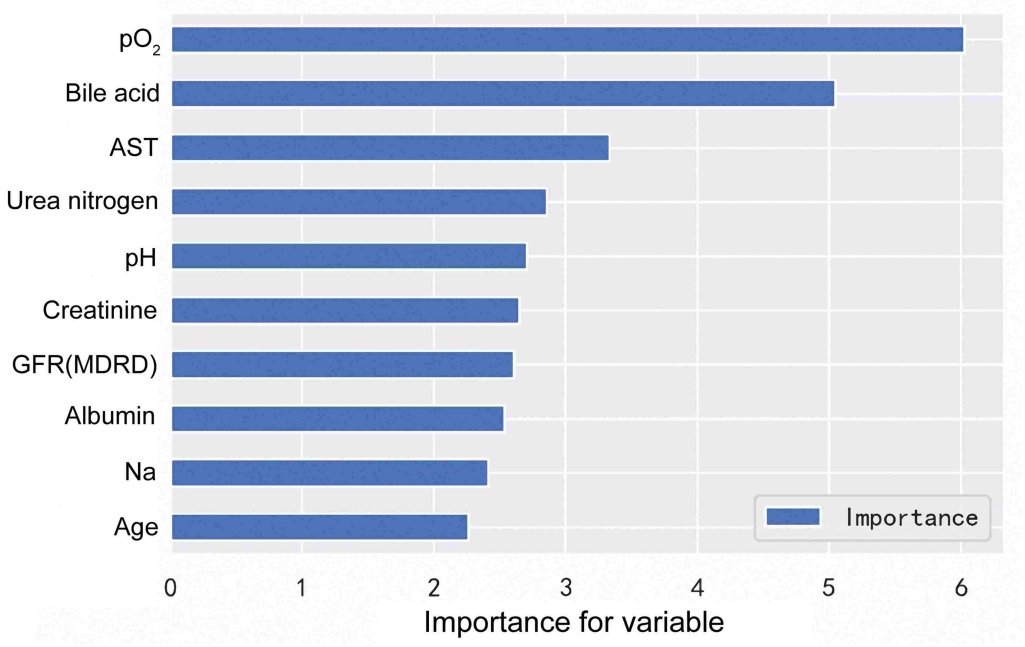

FIGURE 3. The 10 most important influencing factors. AST: Aspartate Transaminase; GFR: Glomerular filtration rate; MDRD: Modified diet in renal disease.

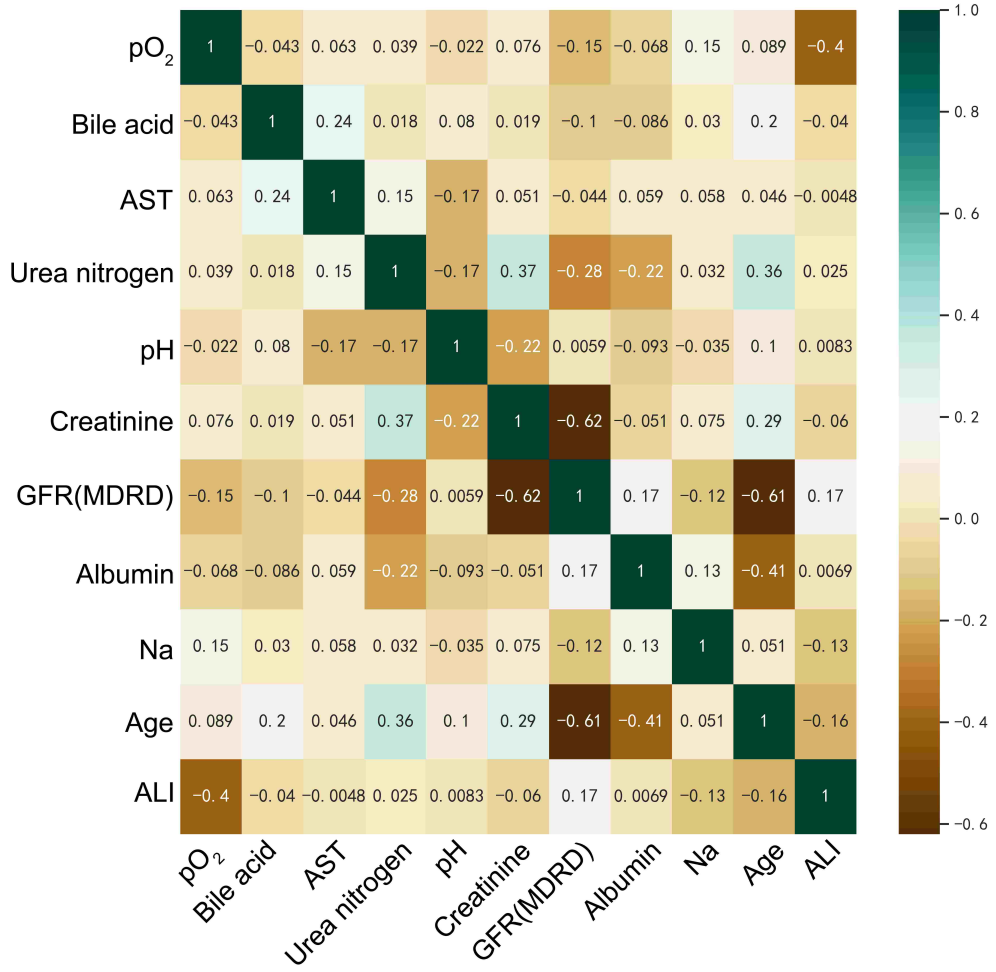

FIGURE 4. Heat map of the 10 most important influencing factors. ALI: acute lung injury; AST: Aspartate Transaminase; GFR: Glomerular filtration rate; MDRD: Modified diet in renal disease.

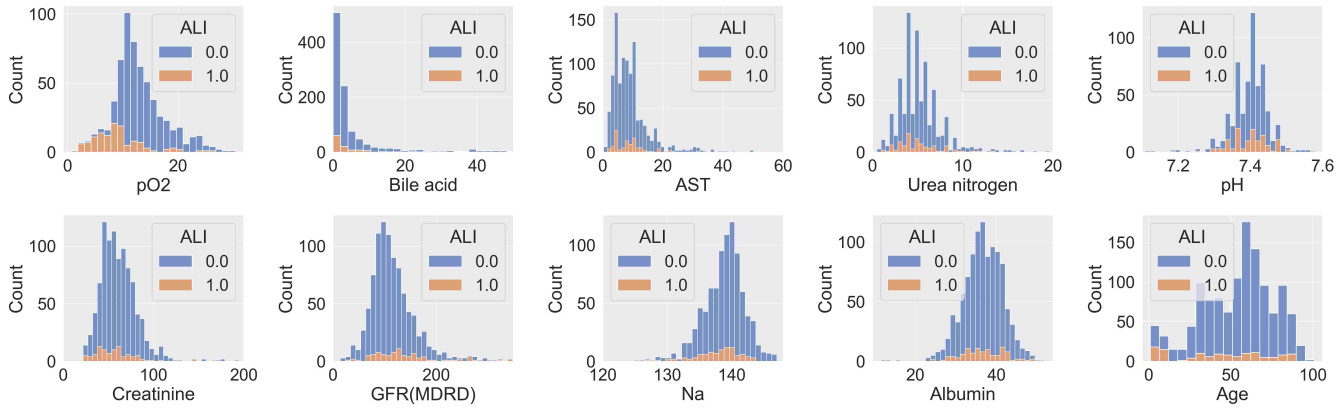

FIGURE 5. Distribution of specific cases of the ten indicators. ALI: acute lung injury; AST: Aspartate Transaminase; GFR: Glomerular filtration rate; MDRD: Modified diet in renal disease.

predictive ability, a specificity of 0.7333, and a sensitivity of 0.7857, indicating that the established XGBoost model has better predictive ability. Also, this study compared three other commonly used machine learning models and considered that the XGBoost model had better disease detection ability.

This study ranked the importance of variables related to acute pancreatitis and acute lung injury, in which the first place was arterial partial pressure of oxygen (PaO_2), PaO_2 reflects the oxygen content in the body, PaO_2 is an early and sensitive indicator to identify acute lung injury and a key indicator to determine its severity [11], therefore, the monitoring of PaO_2 in patients with acute pancreatitis is beneficial to the prevention and early detection of acute lung injury in clinical practice.

Bile acid (BA) is a substance synthesized by the liver and secreted *via* the biliary tract. Aspartate Transaminase (AST) is an important indicator of whether liver function is impaired, and these serum chemical indicators can sensitively reflect pathophysiological changes in the liver or biliary tract. In China, a major cause of acute pancreatitis is biliary pancreatitis, and Polat *et al.* [12] concluded that these indicators have some diagnostic value in biliary pancreatitis. Tran *et al.* [13] found that bile acids play an important role in the pathogenesis of gallstone-induced AP by inducing intracellular Ca^{2+} overload. Chen *et al.* [14] discovered that bile acids may induce activation of alveolar epithelial cells and lung fibroblasts, suggesting that bile acids also play a role in acute lung injury, suggesting that bile acids may be able to play a predictive role in the early stages of AP combined with ALI.

Urea nitrogen is the main end product of human protein metabolism, which increases with the decrease of glomerular filtration rate, and creatinine is an important indicator of renal function impairment [15]. A large amount of fluid leakage in AP patients leads to insufficient effective circulating blood volume, reduced renal perfusion and combined with increased renal vascular resistance, resulting in acute kidney injury due to renal ischemia and hypoxia, manifested by an increase in urea nitrogen and creatinine and a decrease in creatinine clearance [16]. And insufficient effective circulating blood volume can aggravate pulmonary inflammation and promote the development of acute lung injury [17]. Therefore, it is very necessary to monitor renal function and electrolytes in patients with acute pancreatitis in clinical practice.

This study also found that hypoalbuminemia was a risk factor for AP complicating ALI, and the possible mechanism was that hypoalbuminemia led to a decrease in plasma colloid osmotic pressure, which aggravated the development of pulmonary edema and caused lung injury [18].

This study summarizes the importance of various influencing factors from the clinical pattern of acute pancreatitis complicated by acute lung injury, and the model assessment has certain accuracy, specificity and sensitivity, which can play a reference role in clinical diagnosis and treatment and intervene in the development of the disease to a certain extent. However, this study still has some limitations, as it is a single-center study, the sample size included is small, and some variables were discarded due to a high number of missing values, which poses a limitation to the study. There is a need to further increase the sample size and improve the collection of variable data to obtain a more accurate model. And the results of the study need to be validated with multicenter and big data. At the same time, the model contains more variables, which may lack some practicality in clinical practice, and the model needs to be further improved and the variables optimized.

5. Conclusions

In conclusion, the XGBoost model has a good predictive ability for concomitant acute lung injury in patients with acute pancreatitis, and can provide some guidance for clinical practice.

AVAILABILITY OF DATA AND MATERIALS

The data and materials presented in this study are available on reasonable request from the corresponding author.

AUTHOR CONTRIBUTIONS

WWL and XC—designed the study; WL and WJC—acquired the data; SLZ—drafted the manuscript; YKW and XSW—revised the manuscript. The article was reviewed and approved by all authors.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

The study was approved by the Ethics Committee of Xinhua Hospital, Shanghai Jiao Tong University School of Medicine (Approval No. XHEC-D-2021-055), all patients enrolled in this study has signed the broad consent.

ACKNOWLEDGMENT

Not applicable.

FUNDING

Project supported by the Shanghai Specialized Research Fund for Integrated Chinese and Western Medicine in General Hospitals (ZHYY-ZXYJHZX-201914).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. The Lancet. 2018; 392: 1789–1858.

[2] van Dijk SM, Hallensleben NDL, van Santvoort HC, Fockens P, van Goor H, Bruno MJ, *et al*. Acute pancreatitis: recent advances through randomised trials. Gut. 2017; 66: 2024–2032.

[3] Steer ML. Relationship between pancreatitis and lung diseases. Respiration Physiology. 2001; 128: 13–16.

[4] Zhou M. Acute lung injury and ARDS in acute pancreatitis: mechanisms and potential intervention. World Journal of Gastroenterology. 2010; 16: 2094.

[5] Handelman GS, Kok HK, Chandra RV, Razavi AH, Lee MJ, Asadi H. EDoctor: machine learning and the future of medicine. Journal of Internal Medicine. 2018; 284: 603–619.

[6] Association PSGotCM. Guidelines for the diagnosis and management of acute pancreatitis in China (2021). Chinese Journal of Practical Surgery. 2021; 41: 739–746.

[7] Chinese Medical Association ICMB. Guidelines for the diagnosis and treatment of acute lung injury/acute respiratory distress syndrome (2006). Chinese Journal of Internal Medicine. 2007; 02: 19–28.

[8] Ge P, Luo Y, Okoye CS, Chen H, Liu J, Zhang G, *et al*. Intestinal barrier damage, systemic inflammatory response syndrome, and acute lung injury: a troublesome trio for acute pancreatitis. Biomed Pharmacother. 2020; 132: 110770.

[9] Shah J, and Rana SS. Acute respiratory distress syndrome in acute pancreatitis. Indian Journal of Gastroenterology. 2020; 39: 123–132.

[10] Le NQK, Do DT, Chiu FY, Yapp EKY, Yeh HY, Chen CY. XGBoost improves classification of MGMT promoter methylation status in IDH1 wildtype glioblastoma. Journal of Personalized Medicine. 2020; 10: 128.

[11] Liu Y, Mu S, Li X, Liang Y, Wang L, Ma X. Unfractionated heparin alleviates sepsis-induced acute lung injury by protecting tight junctions. Journal of Surgical Research. 2019; 238: 175–185.

[12] Güngör B, Cağlayan K, Polat C, Seren D, Erzurumlu K, Malazgirt Z. The predictivity of serum biochemical markers in acute biliary pancreatitis. ISRN Gastroenterol. 2011; 2011: 279607.

[13] Tran QT, Tran VH, Sendler M, Doller J, Wiese M, Bolsmann R, *et al*. Role of bile acids and bile salts in acute pancreatitis: from the experimental to clinical studies. Pancreas. 2021; 50: 3–11.

[14] Chen B, Cai H, Xue S, You W, Liu B, Jiang H. Bile acids induce activation of alveolar epithelial cells and lung fibroblasts through farnesoid X receptor-dependent and independent pathways. Respirology. 2016; 21: 1075–1080.

[15] Pando E, Alberti P, Mata R, Gomez MJ, Vidal L, Cirera A, *et al*. Early changes in blood urea nitrogen (BUN) can predict mortality in acute pancreatitis: comparative study between BISAP score, APACHE-II, and other laboratory markers—a prospective observational study. Canadian Journal of Gastroenterology and Hepatology. 2021; 2021: 6643595.

[16] Nassar TI, and Qunibi WY. AKI associated with acute pancreatitis. Clinical Journal of the American Society of Nephrology. 2019; 14: 1106–1115.

[17] Koutroumpakis E, Wu BU, Bakker OJ, Dudekula A, Singh VK, Besselink MG, *et al*. Admission hematocrit and rise in blood urea nitrogen at 24 h outperform other laboratory markers in predicting persistent organ failure and pancreatic necrosis in acute pancreatitis: a post hoc analysis of three large prospective databases. The American Journal of Gastroenterology. 2015; 110: 1707–1716.

[18] Zhang W, Zhang M, Kuang Z, Huang Z, Gao L, Zhu J. The risk factors for acute respiratory distress syndrome in patients with severe acute pancreatitis: a retrospective analysis. Medicine. 2021; 100: e23982.

How to cite this article: Weiwei Lu, Xi Chen, Wei Liu, Wenjie Cai, Shengliang Zhu, Yunkun Wang, *et al*. XGBoost model predicts acute lung injury after acute pancreatitis. Signa Vitae. 2023; 19(5): 206-212. doi: 10.22514/sv.2023.087.