Decision tree algorithm applied to MIMIC-III database for the prediction of acute kidney injury in ICU patients

Gao Wenpeng[1], Lyu Haijin[2], Zhou Lang[1], Guo Shengwen[3]

1 *Department of Biomedical Engineering, School of Material Science and Engineering, South China University of Technology, Guangzhou 510006;*

2 *SICU, The Third Affiliated Hospital, Sun Yat-sen University, Guangzhou 510630;*

3 *School of Automation Science and Engineering, South China University of Technology, Guangzhou 510640*

Corresponding author: GUO Shengwen (E-mail: shwguo@scut.edu.cn)

Abstract:

Objective Acute kidney injury (AKI) is one of the most common complications and fatal factors in intensive care unit (ICU). Accurate prediction of AKI risk and identification of key factors related to AKI can provide effective guidance for clinical decision-making and intervention for patients with AKI risk. Methods A total of 30 020 patients in ICU (including 17 222 AKI patients and 12 798 Non-AKI patients) were selected from the public database MIMIC-III in this study, and basic information, physiological and biochemical indicators, drug use, and comorbidity during their stay in ICU were collected. All patients were randomly divided into training sets and independent testing sets according to the ratio of 4:1, and logistic regression, random forest, and lightgbm were applied to construct models for AKI predication in three time points including 24 h, 48 h and 72 h, respectively. The 10-fold cross validation was used to train and validate various models to predict the occurrence of AKI, and obtain important features. Furthermore, 24 h prediction models were used to predict AKI every 24 h during the 7-day window. **Results** lightgbm achieved the best performance with AUC values of 0.90, 0.88, 0.87 for 24 h, 48 h, and 72 h prediction, respectively, and F1 values were 0.91, 0.88, and 0.86. In prediction of every 24 h, the success rates of identifying AKI patients were 89%, 83%, and 80% in one day, two days and three days in advance, respectively. It was found that the length of stay in ICU, body weight, albumin, systolic blood pressure, bicarbonate, glucose, white blood cell count, body temperature, diastolic blood pressure and blood urea nitrogen played vital roles in predicting AKI for ICU patients. Using only 24 important features, the models could still achieve prominent prediction performance. **Conclusions** Based on basic information, physiological and biochemical indicators, drug use, and comorbidity, machine learning methods can be adopted to effectively predict AKI risk for ICU patients at several time points, and determine the dominant factors relative to AKI.

Keywords: Acute kidney injury; Intensive care unit; Machine learning; Risk prediction; Important feature

DOI: 10.3969/j.issn. 1002-3208.2021.06.010.

CLC classification No.: r318 document mark code a Article No.: 1002-3208 (2021) 06-0609-09

Description format of this article gaowenpeng, lvhaijin, zhoulang, et al Application of decision tree algorithm in prediction of acute renal injury in ICU patients based on mimic-iii database [j] Beijing Biomedical Engineering, 2021, 40 (6): 609-617 GAO Wenpeng, LYU Haijin, ZHOU Lang, et al. Decision tree algorithm applied to MIMIC-III database for the prediction of acute kidney injury in ICU

patients[J]. Beijing Biomedical Engineering, 2021, 40(6) : 609-617.

0 Introduction

Acute kidney injury (AKI) is common in intensive care unit (ICU) patients, with high incidence rate and mortality [1-2]. KDIGO (kidney disease: The specific criteria for AKI in the AKI clinical practice guidelines [3] published by improving global outcomes are: Increase of serum creatinine ≥ 26.5 within 48 h μmol/L or the serum creatinine increased to more than 1.5 times of the baseline value within 7 days, or the urine volume was less than 0.5 ml/(kg · h) and the duration was not less than 6 h. Studies have shown that AKI leads to higher treatment costs, adverse clinical reactions and the development of chronic kidney disease in ICU patients [3], and is an independent influencing factor of high mortality in ICU patients [4-5].

As serum creatinine is a non-specific marker of AKI, the diagnosis of AKI has a certain lag [6], and the clinical urine volume is not easy to monitor and the operation error is large, so looking for important clinical factors affecting AKI and making early prediction is the key to timely intervention and guiding treatment of AKI risk patients in ICU. So far, there are usually two methods for early prediction of AKI: one is to find specific biomarkers, and the other is to establish risk prediction models based on statistics or machine learning methods. The clinical application of biomarkers is limited due to the high cost of the method, the small number of samples that can be included and the large impact of individual differences.

With the establishment of open source critical care database and the popularization of electronic health records (EHR) in hospitals, the availability of clinical data of ICU patients has been continuously improved [7], thus providing sufficient data support for AKI prediction research, and relevant research has gradually increased. For example, Haines et al. [8] collected the demographic information of 830 patients in the ICU of the Royal Hospital in London and the hematological indicators within 24 hours after admission, and used logistic regression to predict AKI. The results showed that the area under curve (AUC) value of the receiver operator characteristic curve (ROC curve) predicted AKI 1~3 was 0.70, and the AUC value predicted AKI 2~3 was 0.91. Malhotra et al. [9] collected the demographic information, complications, vital signs, hematological indicators and intervention measures of 207 patients in ICU of two independent hospitals, and used multivariable regression analysis to predict AKI. The AUC value of the independent test set was 0.81. Some scholars also use the open source critical care database to predict AKI. For example, Li Qianhui [10] extracted vital signs and hematological indicators of 1690 patients (840 AKI patients) from the medical information mart for intensive care (mimic) critical care database, and used logistic regression, adaboost and multi-layer perceptron models to predict AKI early. The results showed that the multi-layer perceptron had the best performance, with an f1.5 score of 0.944.

Zhang Yuan et al. [11] extracted the demographic information, vital signs and

Author unit:

1 Department of Biomedical Engineering, School of materials science and engineering, South China University of Technology (Guangzhou 510006)

2 Surgical ICU of the Third Affiliated Hospital of Sun Yat sen University (Guangzhou 510630)

3 School of Automation Science and engineering, South China University of Technology (Guangzhou 510640)

Corresponding author: Guoshengwen. E-mail: shwguo@ scut. Edu. Cn

hematological indicators of 1166 patients (884 AKI patients) from the mimic database. They used logistic regression, random forest and lightgbm models to predict the 24 hours before AKI. They found that the lightgbm model was the best, and the AUC value was 0.92. Zimmerman et al. [12] extracted the demographic information of 23950 patients in the mimic database and the vital signs, hematological indicators and intervention measures within 24 hours after admission. They used logistic regression, random forest and multi-layer perceptron to make early prediction of AKI. The average AUC value was 0.783.

At present, the research on AKI prediction of ICU patients mainly has the following deficiencies: (1) The included sample size is insufficient, especially the sample size of AKI patients is generally small, which makes the model unreliable; (2) Inadequate utilization of clinical or database information, and possible omission of important influencing factors; (3) The prediction time is not timely, and the continuous early warning function is lacking, which leads to the clinicians do not have enough time to intervene.

In view of the above shortcomings, based on the demographic information, admission information, medication use, vital signs, hematological indicators, critical illness score, complications, intervention measures and other 8 types of clinical information of 30020 ICU patients, the study randomly divided the training set and independent test set according to 4:1, and separately applied three machine learning algorithms, namely, logistic regression, random forest and lightgbm, to establish AKI prediction models at 24 h, 48 h and 72 h, Evaluate, compare and analyze the performance of different models, and use the optimal model for continuous 24-h prediction to identify the important factors related to AKI events.

1 Research data

1.1 Data source

The data used in this study are from the mimic-iii database [13]. Mimic-iii is a public and free multi parameter intensive care database provided by MIT, which contains the hospitalization records of 46520 patients admitted to the ICU of Beth Israel Deaconess Medical Center in Boston from June 1, 2001 to October 31, 2012. It has the characteristics of large sample size and rich clinical information.

1.2 Data filtering

Inclusion criteria: age > 18 years; ICU length of stay > 24 h. Patients with the following complications were excluded: kidney stones; Ureteral calculi; renal carcinoma; Carcinoma of renal pelvis; Urinary tract obstructive disease. The first measured creatinine value belongs to the normal range (31.8~116.0 μmol/L), with the first measured creatinine value as the baseline creatinine value; For patients whose creatinine value does not fall within the normal range for the first time, 116.0 is taken on the premise of excluding chronic kidney disease μmol/L as the baseline creatinine value; Patients who have been in ICU for many times, if the time interval between two consecutive ICU stays exceeds 48 hours, will be included according to different samples.

According to the above criteria, 30020 patients were finally included, including 17222 AKI patients, accounting for 57.4%.

1.3 Variable inclusion

(1) Demographic information: age, sex, body mass, height.

(2) Admission information: admission mode and ICU type.

(3) Drug use: antibiotics, diuretics, tacrolimus, rifampicin, amphotericin, cisplatin.

(4) Vital signs: mean arterial pressure, heart rate, respiratory rate, systolic blood pressure, diastolic blood pressure, body temperature.

(5) Health score of critically ill patients: Elixhauser comorbidity score、SAPS II、SOFA score、APSIII.

(6) Complications: hypertension, diabetes, myocardial infarction, heart failure, sepsis, cancer.

(7) Hematological indicators: creatinine, hemoglobin, albumin, ph, bicarbonate, alkali residue, lactic acid, potassium, chlorine, sodium, white blood cell count, glucose, blood urea nitrogen, bilirubin.

(8) Interventions: mechanical ventilation and renal replacement therapy.

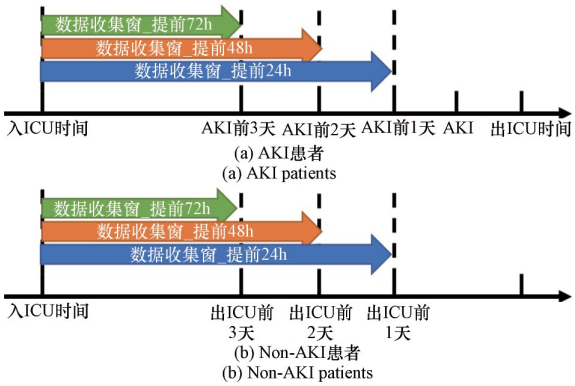
Among them, the admission modes include electric (planned admission), emergency (emergency, not life-threatening) and emergency (emergency, life-threatening); ICU types include SICU, MICU, CCU, tsicu and CSRU; "0/1" for complications and intervention measures, 0 for none, 1 for yes; In the score of critical patients, Elixhauser comorbidity score is the score for patients' complications, SAPS II is the simplified acute physiology score, SOFA score is the score for organ failure, and APS III is the score for acute physiology and chronic health status.

1.4 Data collection time window

In this study, AKI was predicted 24 hours, 48 hours and 72 hours in advance. Referring to the data collection methods of Peng et al. [14], for patients with AKI, the data collection range is from entering ICU to 24 hours, 48 hours and 72 hours before AKI; For non AKI patients, the data collection range is 24 hours, 48 hours and 72 hours before entering the ICU and leaving the ICU. The data collection window is shown in Figure 1.

1.5 Feature construction

According to the time window of data collection, after obtaining the characteristic parameters, the first test value, minimum value, maximum value, mean value, standard deviation, etc. Of vital signs and hematology indicators within the time window are calculated respectively, and the statistical characteristics are included in the characteristic queue to reflect the statistical distribution characteristics of the characteristics. Finally, the characteristic dimension is 102.



Time of entering ICU
3 days before Aki
2 days before Aki
1 day before Aki
Time out of ICU
Data collection window... H in advance

Figure 1 data collection windows for AKI patients and Non-AKI patients

2 Research methods

2.1 Logistic regression

Logistic regression [15] is a classical generalized linear analysis model.

If x and θ、H and Y respectively represent training data, model parameters, predictive output function and real label, so the classification problem is actually a Bernoulli distribution:

$$P(y=1|\ x;\theta)=h_\theta(x) \tag{1}$$

$$P^{'}=P(y=0|\ x;\theta)=1-h_\theta(x) \tag{2}$$

Equations (1) and (2) can be combined into:

$$P' = \left(h_\theta(x)\right)^y \left(1-h_\theta(x)\right)^{1-y} \quad (3)$$

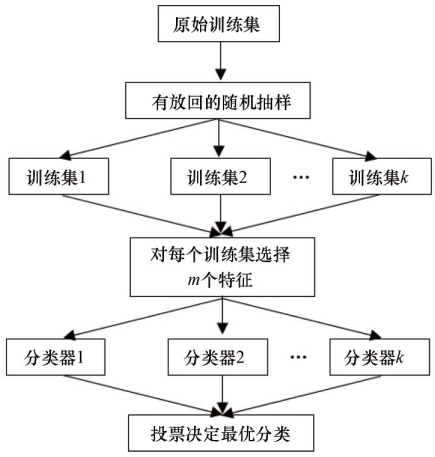
Take the maximum likelihood of equation (3) as the objective function:

$$L(\theta) = \prod_1^n \left(h_\theta\left(x^i\right)\right)^{y^i} \left(1-h_\theta\left(x^i\right)\right)^{1-y^i} \quad (4)$$

Applying gradient descent method to $L(\theta)$ Find the logarithm, and then compare the model parameters θ to find partial derivatives.

Different from linear regression, logistic regression maps the continuous value predicted by linear equation into two discrete values of 1 /0 by introducing a monotonically differentiable sigmoid function as the output function.

The calculation cost of logistic regression is not high, and it is easy to understand and implement. However, it is sensitive to the multicollinearity of independent variables in the model, easy to under fit, low classification accuracy, and difficult to deal with the problem of data imbalance.



Original training set
Random sampling with return
Training set 1
Training set 2
Training set K
Select m features for each training set
Sorter 1
Sorter 2
Classifier K
Voting for optimal classification

Figure 2 Flowchart of random forest

2.2 Random forest

The random forest [16] takes the decision tree as the basic classifier, uses the bagging idea of ensemble learning, randomly selects data subsets and features from the original data set through random sampling with return, and constructs multiple decision trees for classification. The output category is the mode of the output category of a single tree.

By combining multiple classifiers, random forest can often obtain better generalization performance than a single learner, but it is easy to over fit on noisy data. The algorithm flow chart of random forest is shown in Figure 2.

2.3 Lightgbm

Lightgbm (light gradient boosting machine) is an optimization of gradient boosting decision tree (gbdt) [17], which mainly includes two algorithms: gradient based one side sampling (Goss) and exclusive feature bundling (EFB).

Goss algorithm distinguishes the training data of different gradients, and randomly samples the smaller gradient data while retaining the larger gradient data, so as to reduce the amount of calculation and improve the operation efficiency. Define o to represent the training set of a fixed node. The training set instance is $x_1, x_2, \cdots, x_n$, the feature dimension is s, the segmentation feature is j, and the information gain is V[17]. At each gradient iteration, the negative gradient direction of the loss function of the model data variable is expressed as $g_1, g_2, \cdots, g_n$, then the information gain of the segmentation feature j at the segmentation point d is:

$$\hat{V}_{j|O}(d) = \frac{1}{n}\left(\frac{\left(\sum_{\{x_i \in O: x_{ij} \leqslant d\}} g_i\right)^2}{n^i_{l|O}(d)} + \frac{\left(\sum_{\{x_i \in O: x_{ij} > d\}} g_i\right)^2}{n^i_{R|O}(d)}\right) \quad (5)$$

In Goss algorithm, the training data are first sorted in descending order according to the gradient of the data. The data of the first a% with the largest gradient is reserved as data subset a, and then the data subset B is obtained by random sampling from the remaining data.

EFB reduces feature dimensions through feature bundling to improve computing efficiency. The number of original features is feature, and the number of combined features is bundle. The feature complexity of this method ranges from O1 (data × Feature) to O2 (data × Bundle), because the bundle is much smaller than the feature, the model can greatly accelerate the training process of gbdt without affecting the final accuracy.

In addition, lightgbm discretizes the continuous floating-point eigenvalues into k integers to construct a histogram with a width of K. After traversing the data once, the histogram accumulates the statistics required for discretization, and then when splitting the nodes, it can find the best segmentation point according to the discrete values on the histogram to reduce the consumption of memory. Lightgbm also discards the level wise decision tree growth strategy used by most gbdts, and uses the leaf wise strategy with depth constraints, which can reduce more errors and obtain better accuracy under the same splitting times.

2.4 Performance evaluation index

Accuracy, sensitivity, F1 value and AUC were used as evaluation indicators.

$$\text{precision} = \frac{TP}{TP + FP} \quad (6)$$

$$\text{sensitivity} = \frac{TP}{TP + FN} \quad (7)$$

$$F1 = 2 \times \frac{\text{precision} \times \text{sensitivity}}{\text{precision} + \text{sensitivity}} \quad (8)$$

Where: True negative (TN), called true negative rate, indicates the actual number of negative samples predicted to be negative samples; False positive (FP), called false positive rate, indicates the number of samples that are actually negative samples and predicted to be positive samples; False negative (FN), called false negative rate, indicates the number of samples that are actually positive samples and predicted to be negative samples; True positive (TP), called true positive rate, indicates the number of samples predicted to be positive samples.

3 Experimental results

3.1 Prediction results

30020 patients were randomly divided into training set and independent test set according to 4:1. The number of training set and independent test set were 24016 (including 13778 AKI patients and 10238 non AKI patients) and 6004 (including 3444 AKI patients and 2560 non AKI patients) respectively. In the model training phase, ten fold cross validation is used. After the training, the performance of the trained model is evaluated by using independent test sets.

The results of ten fold cross validation and independent test set of each model are shown in Table 1 and Table 2. When different models predict the same time point, except that the logistic regression model has the highest AKI sensitivity after 24 hours, the performance of logistic regression, random forest and lightgbm models increases in turn; As the prediction time point increases from 24 h, 48 h to 72 h, the prediction difficulty increases, and the prediction performance of the same model decreases gradually. Lightgbm has high accuracy and sensitivity, and the difference is

Table 1 Results of 10-fold cross-validation of different time points

Forecast time	Model	Accuracy (95%ci)	Recall rate (95%ci)	F1 value (95%ci)	AUC (95%CI)
24 h	Logistic regression	0.69 (0.68-0.70)	0.93 (0.92-0.94)	0.79 (0.78-0.80)	0.82 (0.81-0.83)
	Random forest	0.88 (0.87-0.89)	0.88 (0.87-0.89)	0.88 (0.87-0.89)	0.94 (0.93-0.95)
	Lightgbm	0.92 (0.91-0.93)	0.89 (0.88-0.90)	0.90 (0.89-0.91)	0.96 (0.95-0.97)
48 h	Logistic regression	0.71 (0.70-0.72)	0.82 (0.81-0.83)	0.76 (0.75-0.77)	0.80 (0.79-0.81)
	Random forest	0.88 (0.87-0.89)	0.85 (0.84-0.86)	0.86 (0.85-0.87)	0.93 (0.92-0.94)
	Lightgbm	0.90 (0.89-0.91)	0.86 (0.75-0.87)	0.88 (0.87-0.89)	0.94 (0.93-0.95)
72 h	Logistic regression	0.82 (0.81-0.83)	0.40 (0.39-0.41)	0.54 (0.53-0.55)	0.80 (0.79-0.81)
	Random forest	0.86 (0.85-0.87)	0.81 (0.79-0.83)	0.83 (0.82-0.84)	0.92 (0.90-0.94)
	Lightgbm	0.86 (0.78-0.94)	0.84 (0.79-0.89)	0.84 (0.80-0.88)	0.94 (0.93-0.95)

Table 2 Results of independent test sets at different time points

Forecast time	Model	Accuracy rate	Recall	F1 value	AUC
24 h	Logistic regression	0.67	0.94	0.79	0.66
	Random forest	0.88	0.89	0.89	0.86
	Lightgbm	0.92	0.89	0.91	0.90
48 h	Logistic regression	0.71	0.84	0.77	0.72
	Random forest	0.89	0.85	0.87	0.86
	Lightgbm	0.91	0.86	0.88	0.88
72 h	Logistic regression	0.82	0.35	0.50	0.65
	Random forest	0.87	0.81	0.84	0.86
	Lightgbm	0.87	0.84	0.86	0.87

small. The random forest is also similar. Therefore, the F1 value and AUC value of the two are also high, but the accuracy and sensitivity of the logistic regression are quite different, that is, the logistic regression can not effectively balance the precision and recall, and its F1 value and AUC value are low. Through comprehensive comparison, lightgbm's prediction performance formula at three time points: True negative (TN), called true negative rate, indicates the actual number of negative samples predicted to be negative samples; False positive (FP), which is called false positive rate, indicates that in fact, it is the best to predict positive samples from negative samples. On the independent test set, the AUC values of AKI predicted at 24 h, 48 h and 72 h are 0.90, 0.88 and 0.87 respectively. The ROC curves predicted by different models at different time points are shown in Figure 3.

Lightgbm with the best performance is used to predict AKI patients for 24 hours continuously, that is, from the first day of admission to ICU, the risk of AKI after 24 hours is predicted, and on the second day, according to the data of the same day and before that day, the risk of AKI is predicted for 24 hours until being transferred out of ICU. Take the time of clinical diagnosis of

AKI minus the time when the model first predicts the success of AKI as the lead time (days), and count the days and proportion of successful prediction (Table 3). Table 3 shows that the success rate of lightgbm prediction for 24 consecutive hours is high, and the success rate of AKI risk patients predicted 1, 2 and 3 days in advance is 89%, 83% and 80% respectively. When 1587 AKI patients were diagnosed according to KDIGO standard, the model could know that 1272 of them had AKI risk 3 days in advance. At this time, clinicians were given 3 days to intervene.

T

able 3 Success rates of AKI prediction of 1-3 days in advance using lightgbm

Forecast days in advance	Number of AKI patients	Number of successful forecasts	Success rate
1	3 444	3 079	0. 89
2	2 695	2 240	0. 83
3	1 587	1 272	0. 80

3.2 Important features

Feature importance can reflect the contribution of each feature to the prediction ability of the model. According to the number of times the features are used in the model training process, all feature weight lists of lightgbm at the three time points of 24 h, 48 h and 72 h are obtained, and the top 35 features are selected, as shown in Figure 4.

Most of the leading features of the three time point prediction models are the same, Including length of stay, body mass, minimum body temperature, maximum/minimum leukocyte count, maximum bicarbonate, minimum glucose, maximum diastolic blood pressure, minimum/maximum blood urea nitrogen, APS III score, maximum body temperature/first measured value, first measured value of hemoglobin, first measured value of heart rate, minimum/maximum serum creatinine, maximum systolic blood pressure, first measured value of heart rate, maximum/maximum heart rate, maximum/maximum glucose/first measured value The first measured value of shrinkage pressure and the maximum value of chlorine. It is worth noting that in the three time point prediction models, the length of hospitalization and body mass are among the top 2, suggesting that these two indicators should be observed first in the early prediction of AKI; the maximum value of leukocyte count, the maximum value of bicarbonate and the minimum value of body temperature all rank in the top 10, which should be paid special attention; However, the importance of serum creatinine was not included in the top 10, indicating that the role of serum creatinine in the model was not ideal. Only creatinine as the diagnostic standard of AKI was not sensitive enough and had a certain lag.

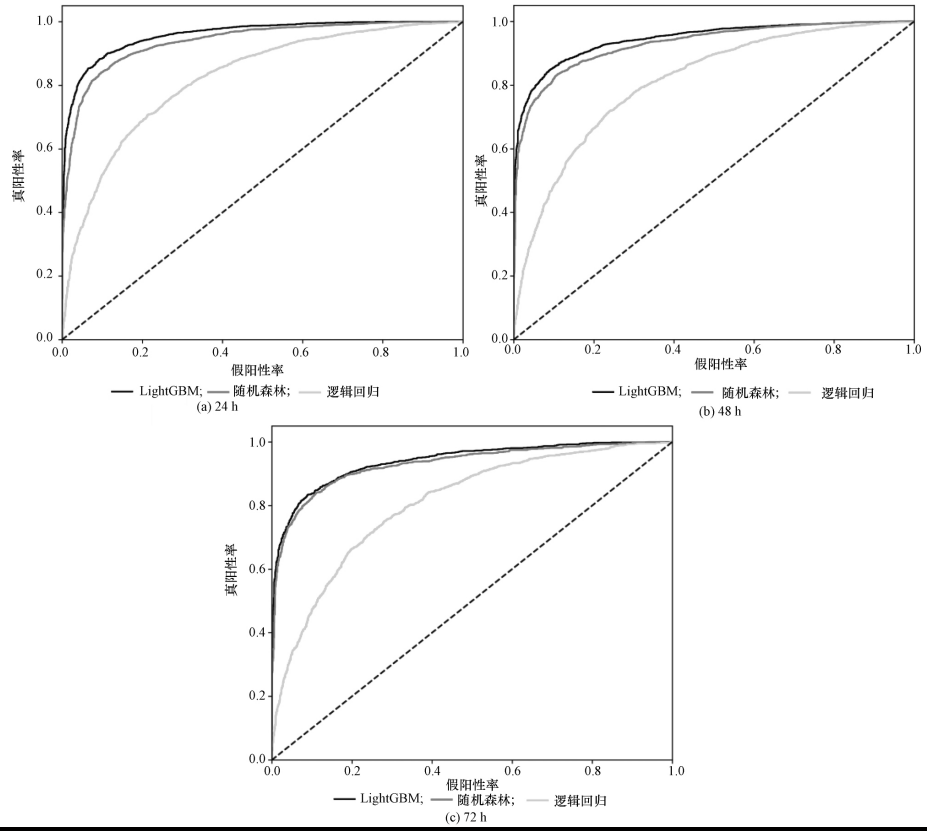
In order to further verify the role of important features and reduce the dimension of features, the lightgbm prediction model at different time points is trained and tested when only 24 important features are used. The results of the independent test set are shown in Table 4.

Table 4 Prediction results of lightgbm using only 24 important features

Forecast time /h	Accuracy rate	Recall	Value	AUC
24	0. 89	0. 87	0. 89	0. 88
48	0. 89	0. 85	0. 87	0. 86
72	0. 86	0. 81	0. 85	0. 85

According to the results in Table 4, when only 24 important features are used, the AUC values of lightgbm prediction model at 24h, 48h and 72h are 0.88, 0.86 and 0.85 respectively. Compared with the use of all features, the AUC value is reduced by no more than 2 percentage points. It shows that among all the 102

dimensional features, 24 important features contribute most to the prediction performance of lightgbm model, and only the important features can be used to continuously and effectively predict AKI.



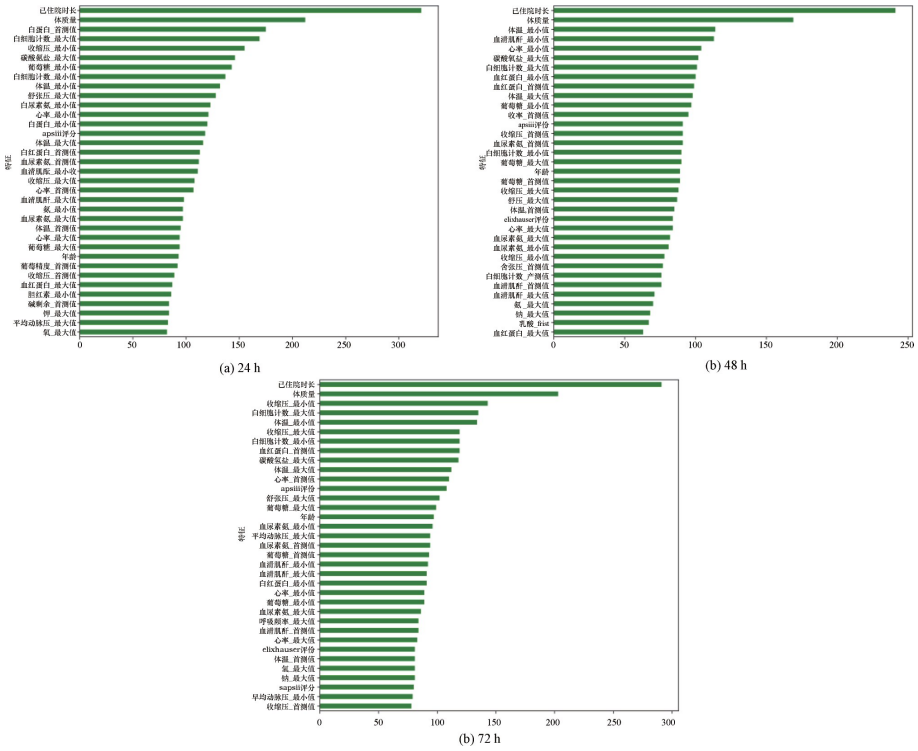
True positive rate
False positive rate
Random forest
logistic regression

Figure 3 Comparison of the ROC obtained by three prediction models

4 Discussion

The experimental results show that among the three machine learning models, lightgbm model has the best performance, followed by random forest, and the logistic regression is poor. Lightgbm with the best performance can be used to continuously predict AKI risk patients, and it still has good performance when only important features are used.

Compared with previous AKI prediction studies [8-12], the main advantages of this study are: (1) Include a large sample size. The total sample size and the number of AKI patients are larger than the relevant studies mentioned above. The sample balance is better and the results are more reliable. (2) Eight categories of clinical information are used to construct features, which reduces the possibility of missing key factors, reduces the dimension of features, and uses important features to predict, which further verifies the role of the important features obtained in this paper. (3) Three time point prediction models of 24 h, 48 h and 72 h have been established, and the continuous prediction of AKI risk patients can not only continuously monitor the patient's condition, but also reserve more time for clinicians to intervene and treat.



Length of hospitalization
Body mass
Albumin_ First measured value
Leukocyte count_ Maximum
Systolic blood pressure_ minimum value
Ammonia carbonate_ Maximum
Glucose_ minimum value
Leukocyte count_ minimum value
Body temperature_ minimum value
Diastolic blood pressure_ Maximum
White urea ammonia_ minimum value
Heart rate_ minimum value
Albumin_ minimum value
Apsiii score
Body temperature_ Maximum
Leucoglobin_ First measured value
Blood urea ammonia_ First measured value
Serum myometrium_ minimum value
Systolic blood pressure_ Maximum
Heart rate_ First measured value
Serum creatinine_ Maximum
Ammonia_ minimum value
Blood urea ammonia_ Maximum
Body temperature_ First measured value

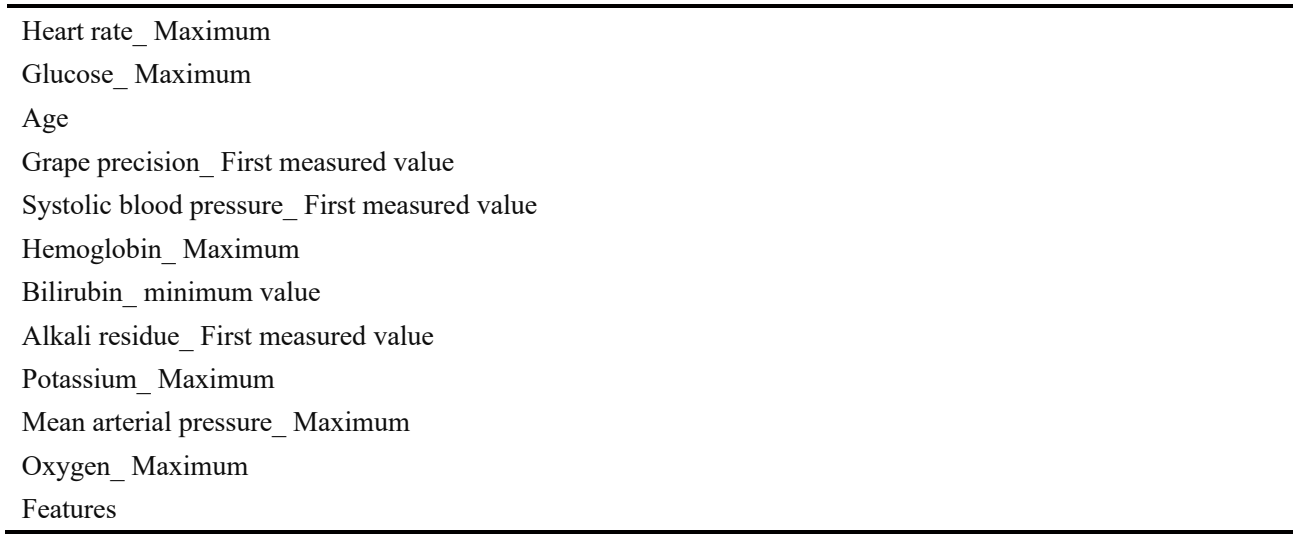


Figure 4 Top 35 important features obtained by lightgbm at three time points

In the lightgbm 24-h prediction model, the top 10 features include length of stay in hospital, body mass, first measured value of albumin, maximum white blood cell count, minimum systolic blood pressure, maximum bicarbonate, minimum glucose, minimum white blood cell count, maximum body temperature and minimum blood urea nitrogen. Most of the top 30 features in the three time point prediction models are the same.

In ICU ward, the patients have been hospitalized for a long time, which indicates that the patients' health condition is severe and the disease is more complex, resulting in an increase in the potential risk of AKI. Body mass is the most common and easily available index. In the experimental data, the incidence of AKI increased by 3.74% for every 5 kg increase in body mass.

Hypothermia can lead to decreased renal blood flow, impaired renal tubular function, and acidosis and alkalosis [18-19]. In ICU, more than 40% of patients with hypothermia will have aki[18].

Leukocytes play an important role in inflammatory response, host defense and repair, and are one of the key immunological factors in the process of most organ injury. There is a U-shaped relationship between leukocyte count and AKI risk. The high AKI risk caused by the decrease of leukocyte count may be attributed to the decrease of lymphocytes and monocytes, and the high AKI risk caused by the increase of leukocyte count may be attributed to the increase of neutrophils [20].

Bicarbonate in serum can help to increase the oxygen delivery to the kidney, and neutralize the acidosis in the kidney. Low bicarbonate levels will increase the risk of renal ischemic injury, especially in critical cases [21].

This study still has the following deficiencies: (1) The samples are from single center database, and the robustness of the model needs to be further verified by multi center data; (2) Only AKI and non AKI were predicted, and AKI was not predicted according to the kidgo diagnostic standard (level I-III).

5 Conclusion

Based on the mimic-iii database, this study extracted 8 types of clinical information of 30020 patients, including demographic information, admission information, vital signs, critical illness score, complications, hematological indicators, medication and intervention measures. Three machine learning

algorithms, namely, logistic regression, random forest and lightgbm, were used to establish AKI prediction models at three time points of 24 h, 48 h and 72 h, and to compare the prediction performance of different models and obtain important features. The results show that lightgbm has the best prediction performance, and has a recognition rate of up to 80% when continuously predicting AKI risk patients. It still has high performance when only using important features for prediction. The research results can provide continuous and effective prediction for the risk of AKI in ICU patients, clarify the important influencing factors, and provide important guidance for medical staff to carry out timely and reasonable intervention.

References

[1] Lameire NH, Bagga A, Cruz D, et al. Acute kidney injury: an increasing global concern [J]. Lancet, 2013, 382 (9887) : 170-179.

[2] Bagshaw SM, George C, Gibney RTN, et al. A multi-center evaluation of early acute kidney injury in critically ill trauma patients[J]. Renal Failure, 2008, 30(6) : 581-589.

[3] Mizuno T, Sato W, Ishikawa K, et al. KDIGO (kidney disease: improving global outcomes) criteria could be a useful outcome predictor of cisplatin-induced acute kidney injury[J]. Oncology, 2012, 82(6) : 354-359.

[4] Eriksson M, Brattstr□m O, M□rtensson J, et al. Acute kidney injury following severe trauma: risk factors and long-term outcome[J]. Journal of Trauma and Acute Care Surgery, 2015, 79 (3) : 407-412.

[5] Ostermann M, Joannidis M. Acute kidney injury 2016: diagnosis and diagnostic workup[J]. Critical Care, 2016, 20: 299.

[6] Panagidis D, Nanas S, Kokkoris S. Biomarkers of acute kidney injury in a mixed ICU population. A narrative review[J]. Health & Research Journal, 2019, 5(4) : 150.

[7] Rojas JC, Carey KA, Edelson DP, et al. Predicting intensive care unit readmission with machine learning using electronic health record data[J]. Annals of the American Thoracic Society, 2018, 15(7) : 846-853.

[8] Haines RW, Lin SP, Hewson R, et al. Acute kidney injury in trauma patients admitted to critical care: development and validation of a diagnostic prediction model [J]. Scientific Reports, 2018, 8: 3665.

[9] Malhotra R, Kashani KB, Macedo E, et al. A risk prediction score for acute kidney injury in the intensive care unit[J]. Nephrology Dialysis Transplantation, 2017, 32(5) : 814-822.

[10] Li QH. Prediction of acute kidney injury and clinical application optimization based on machine learning[D]. Beijing: Beijing Jiaotong University, 2019.

[11] Zhang Y, Feng C, Li KY, et al. Lightgbm model for predicting acute kidney injury risk in ICU patients[J]. Academic Journal of Chinese PLA Medical School, 2019, 40(4) : 316-320.

[12] Zimmerman LP, Reyfman PA, Smith A, et al. Early prediction of acute kidney injury following ICU admission using a multivariate panel of physiological measurements [J]. BMC Medical Informatics and Decision Making, 2019, 19(Suppl 1) : 16.

[13] Johnson A, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database[J]. Scientific Data, 2016 (3) : 160035.

[14] Peng C, Waitman LR, Yong H, et al. Predicting inpatient acute kidney injury over different time horizons: How early

and accurate? [J]. AMIA... Annual Symposium Proceedings/AMIA Symposium, 2018, 2017(2017) : 565-574.

[15] van Houwelingen S. Ridge estimators in logistic regression[J]. Journal of the Royal Statistical Society Series C, 1992, 41 (1) : 191-201.

[16] Criminisi A, Shotton J, Konukoglu E. Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning[J]. Foundations and Trends in Computer Graphics and Vision, 2011, 7 (2-3) : 81-227.

[17] Ke G, Meng Q, Finley T, et al. Lightgbm: a highly efficient gradient boosting decision tree [C]//NIPS'17: the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: NIPS, 2017, 3149-3157.

[18] Mallet ML. Pathophysiology of accidental hypothermia[J]. QJM, 2002, 95(12) : 775-785.

[19] Mégarbane B, Axler O, Chary I, et al. Hypothermia with indoor occurrence is associated with a worse outcome[J]. Intensive Care Medicine, 2000, 26(12) : 1843-1849.

[20] Han SS, Ahn SY, Ryu J, et al. U-shape relationship of white blood cells with acute kidney injury and mortality in critically ill patients[J]. The Tohoku Journal of Experimental Medicine, 2014, 232(3) : 177-185.

[21] Gujadhur A, Tiruvoipati R, Cole E, et al. Serum bicarbonate may independently predict acute kidney injury in critically ill patients: an observational study [J]. World Journal of Critical Care Medicine, 2015, 4(1) : 71-76.

()