

Review began 08/07/2024
Review ended 08/14/2024
Published 08/18/2024

© Copyright 2024
Mei et al. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

DOI: 10.7759/cureus.67121

Early Warning Models Using Machine Learning to Predict Sepsis-Associated Chronic Critical Illness: A Study Based on the Medical Information Mart for Intensive Care Database

Yulin Mei ¹, Meng Li ², Yuqi Li ¹, Ximei Sheng ¹, Chunyan Zhu ³, Xiaoqin Fan ³, Lei Zhang ³, Aijun Pan ³

1. Department of Critical Care Medicine, Wannan Medical College, Wuhu, CHN 2. Department of Intensive Care Unit, First Affiliated Hospital of Anhui Medical University, Hefei, CHN 3. Department of Critical Care Medicine, The First Affiliated Hospital of USTC, Division of Life Science and Medicine, University of Science and Technology of China, Hefei, CHN

Corresponding author: Aijun Pan, aijunpan868@ustc.edu.cn

Abstract

Background

Patients with chronic critical illness (CCI) experience poor prognoses and incur high medical costs. However, there is currently limited clinical awareness of sepsis-associated CCI, resulting in insufficient vigilance. Therefore, it is necessary to build a machine learning model that can predict whether sepsis patients will develop CCI.

Methods

Clinical data on 19,077 sepsis patients from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database were analyzed. Predictive factors were identified using the Student's *t*-test, Mann-Whitney *U* test, or χ^2 test. Six machine learning classification models, namely, the logistic regression, support vector machine, decision tree, random forest, extreme gradient enhancement, and artificial neural network, were established. The optimal model was selected on the basis of its performance. Calibration curves were used to evaluate the accuracy of model classification, while the external validation dataset was used to evaluate the performance of the model.

Results

Thirty-seven characteristics, such as elevated alanine aminotransferase, rapid heart rate, and high Logistic Organ Dysfunction System scores, were identified as risk factors for developing CCI. The area under the receiver operating characteristic curve (AUROC) values for all models were above 0.73 on the internal test set. Among them, the extreme gradient enhancement model exhibited superior performance (F1 score = 0.91, AUROC = 0.91, Brier score = 0.052). It also exhibited stable prediction performance on the external validation set (AUROC = 0.72).

Conclusion

A machine learning model was established to predict whether sepsis patients will develop CCI. It can provide useful predictive information for clinical decision-making.

Categories: Other, Internal Medicine, Emergency Medicine

Keywords: external validation, calibration curve, predictive model, chronic critical illness, sepsis

Introduction

Due to the rapid technological development of the intensive care unit (ICU), the survival rate of patients in intensive care has increased. However, this has led to an increasing number of patients with chronic critical illness (CCI) relying on long-term mechanical ventilation and other critical care treatments. CCI is a devastating disease with a mortality rate comparable to that of malignancy [1,2]. Patients with CCI are prevalent in modern critical care medicine, accounting for 5-11% of the total number of ICU inpatients [1,3], while patients with CCI account for 30-50% of ICU resources [2]. The in-hospital mortality rate for patients with CCI is as high as 30.9% [3], and the five-year mortality rate is as high as 82% [4]. These statistics show that CCI imposes a huge burden on society, families, and individuals.

Pathological and physiological changes during CCI lead to persistent inflammation, immunosuppression, and catabolism syndrome (PICS), which is characterized by severe immune imbalance, nutritional metabolism disorders, and sustained damage to multiple organs [5]. Both the treatment and prevention of CCI require early intervention, including enteral nutrition, immunotherapy, bedside rehabilitation, and

How to cite this article

Mei Y, Li M, Li Y, et al. (August 18, 2024) Early Warning Models Using Machine Learning to Predict Sepsis-Associated Chronic Critical Illness: A Study Based on the Medical Information Mart for Intensive Care Database. Cureus 16(8): e67121. DOI 10.7759/cureus.67121

mechanical ventilation management [6,7]. For example, nutritional support therapy not only corrects the patient's malnutrition but also plays a role in regulating metabolic disorders, inhibiting inflammatory reactions, and regulating immune function, which have a positive impact on the outcome of CCI. At present, many clinicians have a limited understanding of CCI, and their vigilance toward CCI is relatively poor. In such cases, effective intervention in the early stages of sepsis cannot be provided to prevent CCI. This leads to high mortality rates, high medical expenses, and poor prognosis in the later stages of sepsis. Therefore, early warning of CCI is crucial. Several studies have analyzed risk factors for CCI. For instance, a prospective cohort study involving 453 patients identified mechanical ventilation, Glasgow Coma Scale <15, inadequate caloric intake, and high body mass index as independent predictors of CCI. However, owing to its small sample size and use of traditional models, its predictive performance was limited (area under the receiver operating characteristic curve (AUROC) = 0.803).

The traditional linear regression prediction method has low accuracy and poor processing ability for noisy data and is prone to overfitting when the data volume is small. In addition, linear regression methods are not suitable for nonlinear data analysis. By contrast, as a subset of artificial intelligence, machine learning can search for laws in historical data and predict future data based on those laws. Machine learning methods include clustering, classification, decision trees, Bayesian inference, neural networks, deep learning, and other algorithms [8]. With the rapid development of artificial intelligence in various fields, big data analytics can now identify the risk of disease and predict its occurrence. Previous studies on sepsis-associated CCI have focused on identifying risk factors, but few studies have used machine learning algorithms to predict CCI. Models built by machine learning algorithms can predict the occurrence of CCI, which in turn can provide guidance for early intervention. As a result, the length of stay (LOS) in the ICU may be shortened and even the mortality rate may be reduced. Therefore, machine learning models for predicting sepsis-associated CCI may have a wide range of clinical applications [4].

This study used the Medical Information Mart for Intensive Care (MIMIC)-IV database (version 2.0), which includes comprehensive clinical data from a large and diverse patient population. All datasets from different sources can be obtained in the MIMIC repository (<https://mimic.physionet.org/>). In this retrospective study, six machine learning models for predicting CCI, namely, logistic regression (LR), support vector machine (SVM), decision tree (DT), random forest (RF), extreme gradient enhancement (XGBoost), and artificial neural network (ANN) models, were successfully established to provide a clinical decision-making basis in potentially high-risk populations for the prevention and treatment of CCI [9].

Materials And Methods

Data sources

This study utilized data from the MIMIC-IV database (version 2.0), which contains clinical data and waveform data of bedside monitoring devices. The former includes demographic information, diagnostic information, laboratory testing information, microbial culture reports, medical imaging information, and vital sign observations, while the latter includes vital sign parameters and event records (medical measures and drugs). In addition, we utilized an independent external validation set from the MIMIC-III version 1.4 "CareVue" cohort (2001-2008). Using both databases, we carefully studied related courses and obtained permission for their use (record ID 50611006). As the MIMIC databases are approved by the institutional review boards of Beth Israel Deaconess Medical Center and MIT and all patient information in the databases is anonymized, informed consent was not required for the use of the MIMIC-IV database [10,11].

Study population and grouping

Inclusion criteria for this study were based on the diagnostic criteria for CCI proposed by Gardner et al. [5]: ICU hospitalization exceeding 14 days, together with persistent organ dysfunction characterized by a cardiovascular sequential organ failure assessment (SOFA) score ≥ 1 or a score ≥ 2 for any other organ. Patients were categorized into CCI and non-CCI groups based on the diagnostic criteria for CCI. We referred to diagnostic codes of the ninth revision of the International Classification of Diseases (ICD-9), available in the MIMIC-IV and MIMIC-III databases [4,12]. Exclusion criteria were as follows: patient age, <18 years; LOS in the ICU, <24 hours; pregnancy; proportion of missing data, $\geq 25\%$; craniocerebral injury (Figure 1).

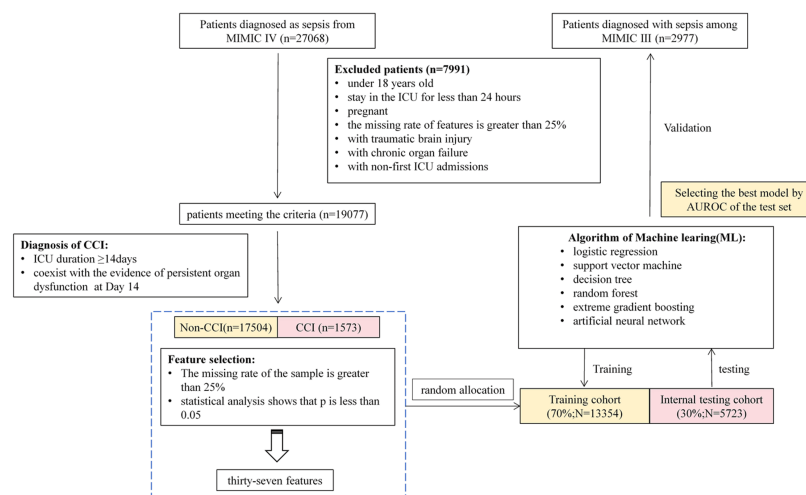


FIGURE 1: Model establishment and validation flowchart.

Clinical variables

Several variables were extracted from the database, including demographics, vital signs, underlying disease, laboratory indicators, scoring systems, medication treatments, mechanical ventilation, extracorporeal membrane oxygenation, and renal replacement therapy. All data were collected within 24 hours of ICU admission. The outcome variable was whether patients developed CCI. In cases where the variables were measured more than once, the worst first-day laboratory variables were used.

All data collection processes were performed using PostgreSQL (version 11.19, PostgreSQL Global Development Group, USA) and Stata/MP (version 17.0, StataCorp LLC, USA).

Data filling and discretization

Owing to substantial variability in missing data among different age groups, segmental imputation based on age groups was not performed (Figure 2). Features with a missing rate $\geq 40\%$ were excluded. Missing continuous data were imputed using the mean or median of their respective groups. There was no missing classification data and hence no need to fill it in.

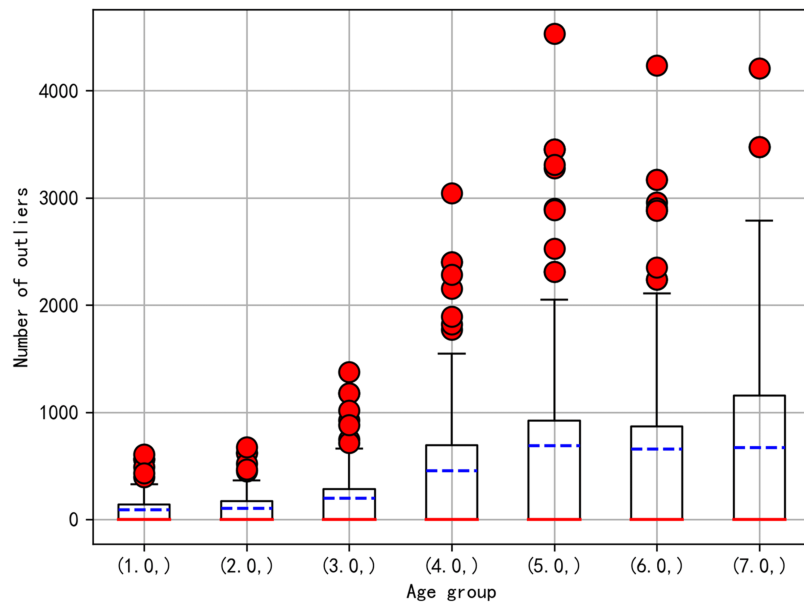


FIGURE 2: Missing-value box plot.

The x-axis represents different age groups, measured in units of 10 years (1.0 represents 18 to 30 years old; 2.0 represents 30 to 40 years old, and so on). The y-axis represents the number of missing patients.

Data discretization enhances data mining efficiency and accuracy, optimizes algorithms for filtering noise and identifying outliers, and mitigates overfitting. Therefore, we converted 41 continuous features (e.g., glucose, Na⁺ concentration, heart rate, diastolic blood pressure, and urea) into discrete data. For instance, Na⁺ concentrations were classified into hypoosmolar, iso-osmolar, and hyperosmolar based on thresholds of 135 mEq/L and 145 mEq/L. Heart rate was categorized as bradycardia, normal, and tachycardia using thresholds of 60 beats/min and 100 beats/min. This method focused particularly on features with a high prevalence of outliers. We constructed a box plot encompassing all features and manually established the range for normal values (mean $\pm$ 2.5 * standard deviation). Values beyond this range were designated as outliers, each denoted by an open circle (Figure 3).

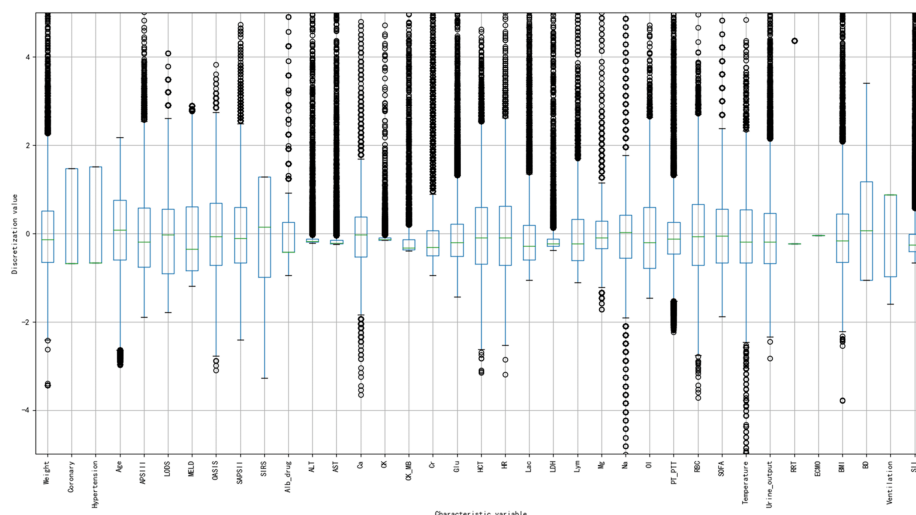


FIGURE 3: Discrete-value box plot.

BMI: body mass index; BD: basis diseases; ECMO: extracorporeal membrane oxygenation; RRT: renal replacement therapy; MV: mechanical ventilation; SOFA: Sequential Organ Failure Assessment scores; SIRS: Systemic Inflammatory Response Syndrome scores; APSIII: simplified acute physiology score III; SAPSII: simplified acute physiology score II; MELD: Model for End-Stage Liver Disease scores; OASIS: Oxford acute severity of illness scores; LODS: Logistic Organ Dysfunction System scores; OI: oxygenation index; urine_output: daily urine output; ALT: glutathione aminotransferase; LDH: lactate dehydrogenase; Glu: blood glucose; CK_MB: creatine kinase isoenzyme; PTT: activated partial thromboplastin time; AST: alanine aminotransferase; CK: creatine kinase; RBC: red blood cell; HCT: hematocytometry; SII: systemic inflammatory response index; Alb_drug: albumin dosage within 24 hours

Construction and evaluation of predictive models

Feature Selection

Student's t-test, the Mann-Whitney U test, and the χ^2 test were used to screen for statistically significant factors between groups.

Data Partitioning

The patients were randomly divided into training and testing sets in a 7:3 ratio. The training set comprised 13,353 individuals, and the testing set comprised 3,473 individuals. Considering the markedly smaller number of positive samples compared with the number of negative samples, positive samples in the training set were oversampled to match the number of negative samples before building the model.

Comprehensive Analysis of Classification Models

Using the sklearn library (version 1.3.2) in Python, multiple classification models were developed, including LR, SVM, DT, RF, XGBoost, and ANN. The training set underwent fivefold cross-validation, followed by evaluation using the testing set to assess the predictive performance of various models. Calibration curves were generated using Python (sklearn, version 1.3.2) to further assess the models' predictive accuracy. Subsequently, a comprehensive evaluation of predictive models was conducted to validate their usefulness in decision support. The optimal model was ultimately selected.

External Validation

Further evaluation of model performance was conducted using an external validation set from MIMIC-III.

Statistical analysis

Continuous variables that followed a normal distribution are represented as mean $\pm$ standard deviation and were compared using Student's t-test. Non-normally distributed data are represented as median (interquartile range) and were compared using the Mann-Whitney U test. To balance the large number of negative samples, random sampling was used to select an equivalent number of negative samples for statistical analysis alongside the positive samples. Each feature underwent 1,000 iterations of testing, and the resulting average value was reported as the p-value. A threshold of $p < 0.05$ was considered statistically significant. Statistical analyses were performed using IBM SPSS Statistics for Windows, version 27.0.1

(released 2020, IBM Corp., Armonk, NY) (and Python (version 3.10.9, Python Software Foundation, USA).

For a detailed introduction to the used code, see https://github.com/15990036592/1.git.

Results
Baseline data comparison

Table 1 presents the fundamental characteristics of the study population. After excluding patients with substantial missing data, the study included 19,077 sepsis patients from the MIMIC-IV dataset, of which 1,573 (8.2%) were classified as CCI patients. After statistical testing, 37 statistically significant characteristic variables were selected, including age, aspartate aminotransferase (AST), and systemic immune inflammation index.

Characteristic	CCI (n = 1,573)	非CCI (n = 17,504)	P-value
Age (years), mean (SD)	63 (15)	67 (15)	<0.001
Gender (male)(%)	895 (56.9%)	10,389 (59.4%)	0.061
Weight (kg), median (IQR)	83 (68.99)	80 (68.94)	<0.001
BMI (kg/m^2), median (IQR)	29.3 (24.8, 32.9)	29.3 (25.8, 29.8)	0.005
Hypertension (%)	373 (23.7%)	5412 (30.9%)	<0.001
Coronary (%)	323 (20.5%)	5729 (32.7%)	<0.001
BD (times), mean median (IQR)	1 (0,1)	1 (0, 2)	<0.001
ECMO (%)	14 (0.9%)	24 (0.13%)	<0.001
RRT (%)	140 (8.9%)	810 (4.6%)	<0.001
MV (%)	1,315 (77.5%)	9948 (48.7%)	<0.001
SOFA, mean (SD)	7 (4)	6 (3)	<0.001
SIRS, median (IQR)	3 (3,4)	3 (2, 3)	<0.001
APSIII, mean (SD)	80 (25)	52 (22)	<0.001
SAPSII, mean (SD)	44 (14)	39 (13)	<0.001
MELD, median (IQR)	15 (9,22)	13 (9, 20)	0.005
OASIS, mean (SD)	43 (8)	35 (9)	<0.001
LODS, mean (SD)	9 (3)	6 (3)	<0.001
Heart rate (times/min), mean (SD)	110 (21)	105 (19)	<0.001
Temperature (celsius), median (IQR)	37.6 (37.2, 38.2)	37.4 (37.1, 37.9)	<0.001
PH, median (IQR)	7.41 (7.38, 7.46)	7.42 (7.39, 7.45)	0.548
OI (mmHg), median (IQR)	128 (84, 184)	168 (127, 200)	0.001
Urine_output (ml), median (IQR)	1387 (810, 2115)	1625 (1005, 2335)	<0.001
Magnesium (mg/dL), median (IQR)	2.2 (2.0, 2.5)	2.2 (2.0, 2.4)	0.034
Creatinine (mg/dL), median (IQR)	1.2 (0.8, 1.8)	1.1 (0.8, 1.6)	0.002
ALT (U/L), median (IQR)	37 (27, 53)	34 (28, 40)	0.003
LDH (U/L), median (IQR)	458 (382)	348 (255)	<0.001
Glu (mg/dL), median (IQR)	168 (134, 214)	158 (128, 195)	<0.001
Sodium (mEq/L), median (IQR)	141 (138, 143)	140 (137, 142)	0.015
Ck_MB (U/L), median (IQR)	6.2 (16.2)	3.9 (11.1)	0.002
Lactate, median (IQR)	2.1 (1.2, 3.8)	1.8 (0.8, 2.8)	<0.001

PTT (s), median (IQR)	35.8 (28.9, 47.9)	33.8 (28.9, 43.4)	0.010
Calcium (mg/dL), median (IQR)	8.4 (7.9, 8.8)	8.4 (8.0, 8.7)	0.008
AST (U/L), median (IQR)	61 (44, 94)	51 (43, 63)	<0.001
CK (U/L), median (IQR)	730 (1618)	464 (1374)	0.001
RBC (10^12/L), mean (SD)	3.3 (0.7)	3.3 (0.7)	0.003
HCT (%), median (IQR)	33.8 (29.9, 38.0)	33.0 (29.5, 36.7)	<0.001
Lymphocytes (%), median (IQR)	10.6 (4.5)	12.0 (4.9)	0.016
SII (10^9), median (IQR)	1502 (984, 2382)	1284 (822, 2067)	0.014
Alb_drug (g), median (IQR)	7.9 (16.9)	5.9 (13.0)	0.027

TABLE 1: Baseline characteristics in the positive cohort and negative cohort.

CCI: chronic critical illness; IQR: interquartile range; SD: standard deviation; BMI: body mass index; BD: basis diseases; ECMO: extracorporeal membrane oxygenation; RRT: renal replacement therapy; MV: mechanical ventilation; SOFA: Sequential Organ Failure Assessment scores; SIRS: Systemic Inflammatory Response Syndrome scores; APSIII: Simplified Acute Physiology Score III; SAPSII: Simplified Acute Physiology Score II; MELD: Model for End-Stage Liver Disease scores; OASIS: Oxford acute severity of illness scores; LODS: Logistic Organ Dysfunction System scores; OI: oxygenation index; urine_output: daily urine output; ALT: glutathione aminotransferase; LDH: lactate dehydrogenase; Glu: blood glucose; CK_MB: creatine kinase isoenzyme; PTT: activated partial thromboplastin time; AST: alanine aminotransferase; CK: creatine kinase; RBC: red blood cell; HCT: hematocytometry; SII: systemic inflammatory response index; Alb_drug: albumin dosage within 24 hours

Comprehensive model analysis

We developed six machine learning models, namely, LR, SVM, DT, RF, XGBoost, and ANN. The area under the receiver operating characteristic curve (AUROC) for each of the six models was used to compare model performance on both training and testing sets. A larger AUROC indicates stronger discriminative ability (Figure 4A, 4C) [13,14]. Classification reports are summarized in Table 2. In the test set, except for the ANN, the AUROCs of all models were above 0.8, and the XGBoost model performed the best (AUROC = 0.91, accuracy = 0.93, specificity = 0.93, sensitivity = 0.93). While the AUROC evaluates prediction accuracy, precision-recall curves are more clinically applicable [15]. In the training and testing sets, XGBoost still showed the best performance, having the highest average precision (AP) values of 0.90 and 0.58, respectively (Figure 4B, 4D). Altogether, XGBoost was proven to be the optimal model.

Model	Training set					Test set				
	Accuracy	Specificity	Sensitivity	F1-score	AUC	Accuracy	Specificity	Sensitivity	F1-score	AUC
LR	0.90	0.95	0.88	0.89	0.82	0.89	0.95	0.87	0.88	0.81
SVC	0.95	0.99	0.92	0.93	0.95	0.89	0.95	0.86	0.87	0.82
DT	0.94	0.96	0.95	0.94	0.96	0.91	0.94	0.92	0.91	0.83
RF	0.91	0.93	0.92	0.90	0.88	0.90	0.93	0.92	0.90	0.86
XGBoost	0.94	0.94	0.94	0.92	0.99	0.93	0.93	0.93	0.91	0.91
ANN	0.89	0.94	0.88	0.89	0.73	0.90	0.93	0.92	0.90	0.72

TABLE 2: Predictive performance of models in the training and test sets.

LR: logistic regression, SVM: support vector machine, DT: decision tree, RF: random forest, XGBoost: extreme gradient boosting, ANN: artificial neural network, AUROC: area under the receiver operating characteristic curve

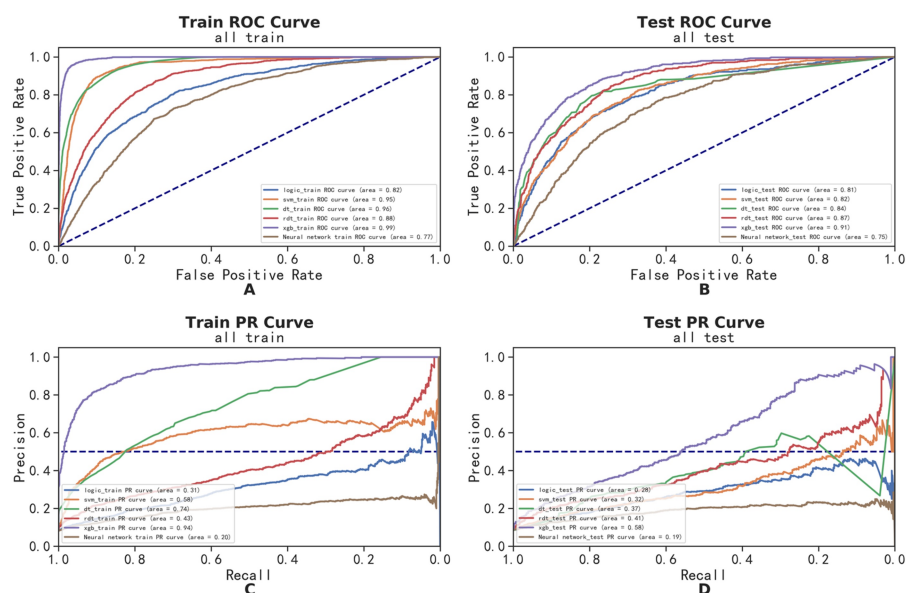


FIGURE 4: Machine learning models' comprehensive analysis.

(A) Train receiver operating characteristic (ROC) curve. (B) Train precision-recall (PR) curve. (C) Test ROC curve. (D) Test PR curve. The ratio of the training set to the test set is 7:3.

The calibration curve for the testing set plots the mean predicted probability on the horizontal axis and the actual probability of events on the vertical axis. A dashed diagonal line serves as a reference, while smooth solid lines represent the fits of different models. A closer fit to the reference line and a smaller value in parentheses indicate that the model's predicted values were more accurate. The calibration curve highlights that both XGBoost and RF models achieved higher accuracy. The Brier score measures the mean squared error between predicted probabilities and actual outcomes. The Brier scores for the XGBoost and RF models were 0.052 and 0.148, respectively (Figure 5).

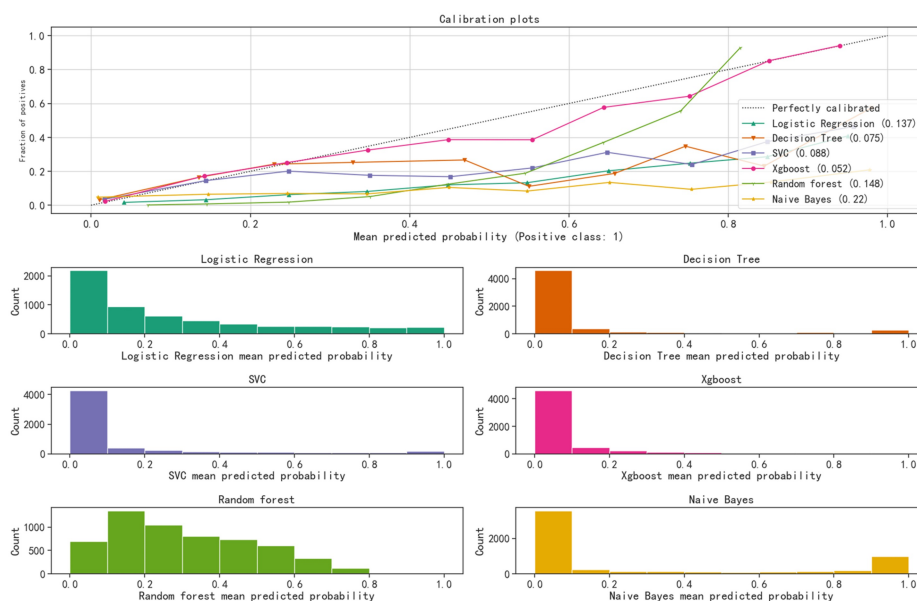


FIGURE 5: Calibration curve of the test set.

The x-axis represents the average predicted probability, and the y-axis represents the actual probability of the event. The dashed diagonal line serves as the reference line. The other smooth solid lines represent different model fitting lines. The closer a fitting line is to the reference line, the more accurate the model's prediction, as indicated by the smaller values in brackets.

On the testing set, the RF clearly outperformed other models, achieving AUROC and AP values of 0.88 and 0.40, respectively. We conducted a comparative analysis between RF and XGBoost, plotting feature importance rankings (Figure 6A, 6B), training history curves (Figure 7A, 7B), and tree diagrams. Key features in the RF model included AST, Acute Physiology and Chronic Health Evaluation III (APACHE III), and Logistic Organ Dysfunction System, while those in the XGBoost were AST, Model for End-Stage Liver Disease, and coronary heart disease. In the training history curves, the x-axis quantifies the number of training samples, and the y-axis quantifies the accuracy scores (Figure 6). As the sample size increased, accuracy decreased for both models on training sets, contrasting with minor fluctuations in validation set accuracy. A small training set yields low training error but high cross-validation error, indicating potential overfitting. Larger training sets enhance model generalization, stabilizing errors in both sets. These curves elucidate the relationship between the model performance and training set size, which facilitates model selection, optimization, and determination of whether more data are needed to improve performance. The XGBoost model exhibited a more pronounced disparity between training and validation accuracies than the RF model, suggesting a proclivity for overfitting. Conversely, the RF model demonstrated a superior balance between training and validation accuracies, indicating its robust generalization capability.

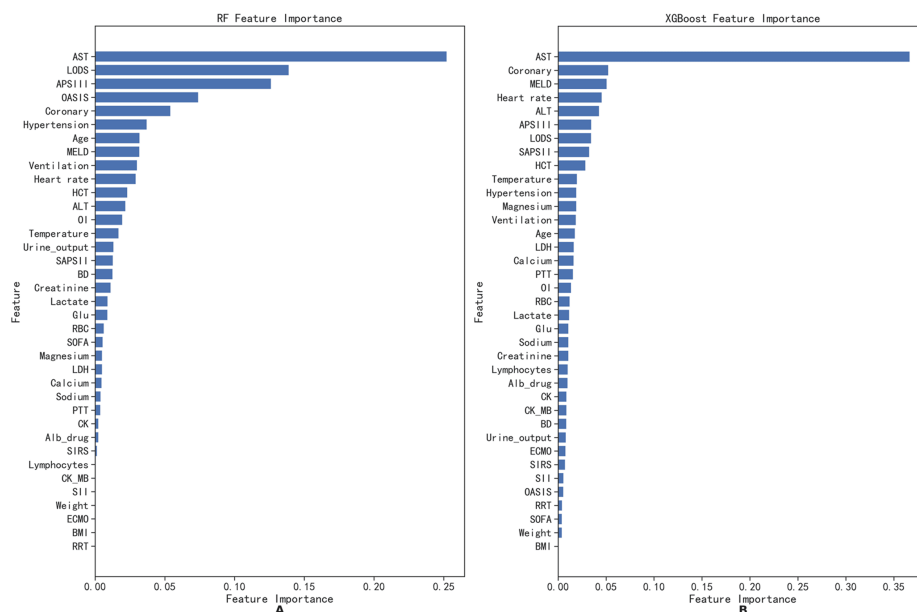


FIGURE 6: Feature importance ranking of extreme gradient enhancement (A) and random forest (B).

The horizontal axis represents the weight size of each feature, and the vertical axis represents the ranking of feature importance. BMI: body mass index; BD: basis diseases; ECMO: extracorporeal membrane oxygenation; RRT: renal replacement therapy; MV: mechanical ventilation; SOFA: Sequential Organ Failure Assessment scores; SIRS: Systemic Inflammatory Response Syndrome scores; APSIII: Simplified Acute Physiology Score III; SAPSII: Simplified Acute Physiology Score II; MELD: Model for End-Stage Liver Disease scores; OASIS: Oxford acute severity of illness scores; LODS: Logistic Organ Dysfunction System scores; OI: oxygenation index; urine_output: daily urine output; ALT: glutathione aminotransferase; LDH: lactate dehydrogenase; Glu: blood glucose; CK_MB: creatine kinase isoenzyme; PTT: activated partial thromboplastin time; AST: alanine aminotransferase; CK: creatine kinase; RBC: red blood cell; HCT: hematocytometry; SII: systemic inflammatory response index; Alb_drug: albumin dosage within 24 hours

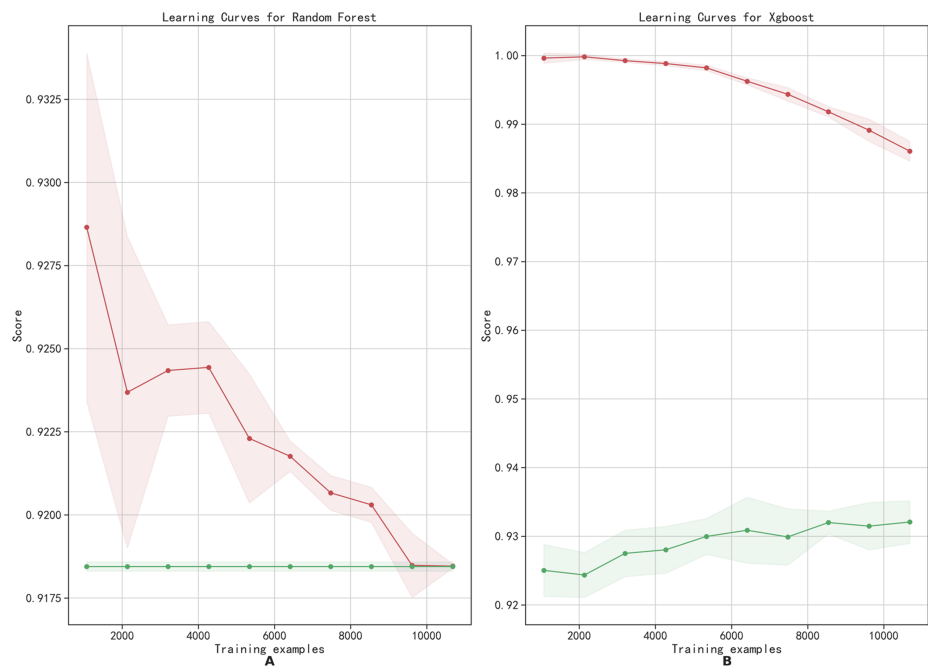


FIGURE 7: Training history curve of extreme gradient enhancement (A) and random forest (B).

The horizontal axis represents the number of training samples, and the vertical axis represents the accuracy score. The red curve represents the training set, and the green curve represents the test set.

External validation and model evaluation

To further validate the generalizability of our model and mitigate its inherent potential for overfitting, we used data on 2,977 sepsis patients obtained from MIMIC-III version 1.4. Our best-trained model was validated using this independent dataset, demonstrating that the XGBoost model maintained strong predictive performance (AUROC = 0.72) using external data (Figure 8).

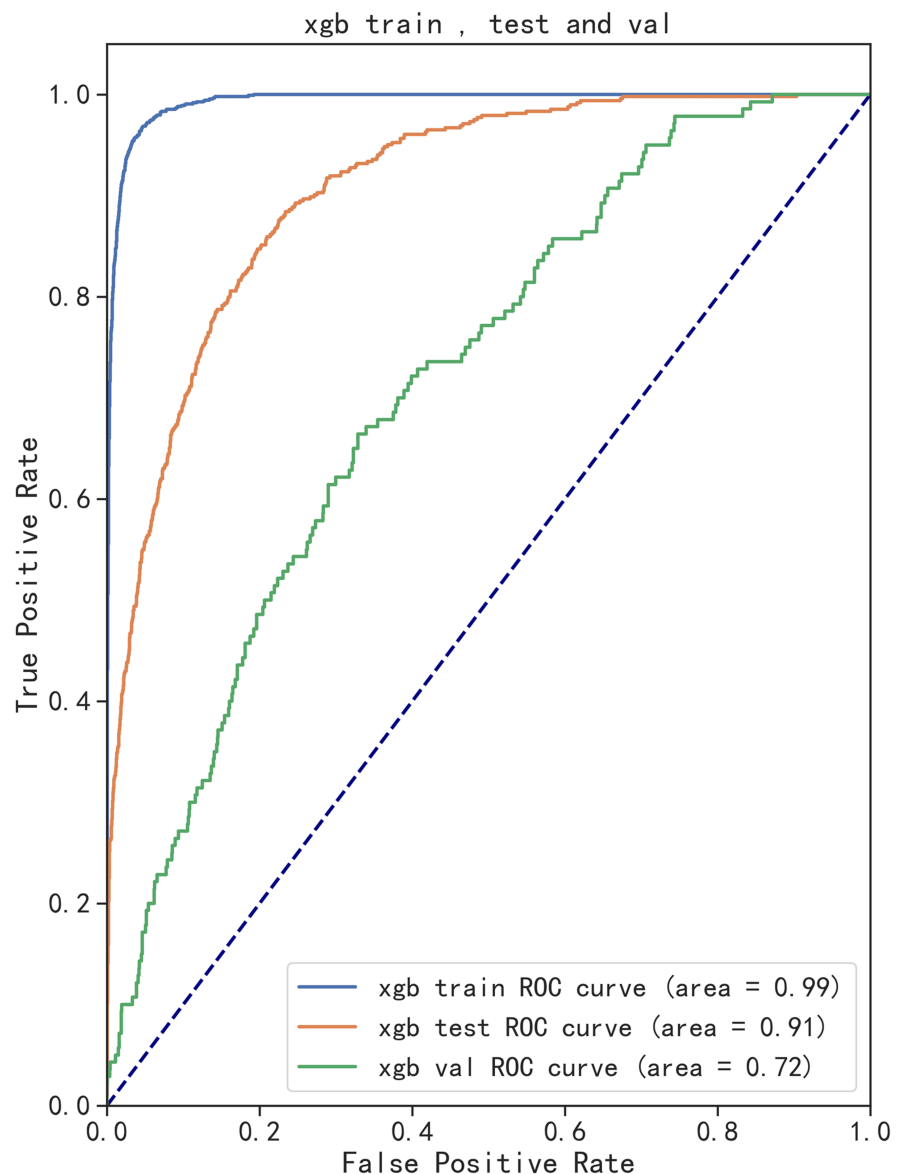


FIGURE 8: Model performance of extreme gradient enhancement on training, testing, and validation sets.

Blue represents the training set; yellow represents the testing set, and green represents the validation set.

Discussion

In our study, six machine learning models were used. By analyzing the ROC and PR curves, we found that the XGBoost model performed better than other models on the test set (AUROC = 0.91 and AP = 0.58). Generally, an AUROC of 0.7 is considered a reasonable threshold indicating good predictive performance of a model, and an AUROC > 0.8 indicates excellent model performance [16]. However, it should be noted that the AUROC cannot solely explain the performance of the model. It is also necessary to combine other indicators, such as specificity, sensitivity, and F1 score, for comprehensive evaluation. Except for LR and SVM models, all other models have specificity > 0.93, sensitivity > 0.92, and F1 score > 0.90. The XGBoost model also performed excellently in these three aspects (specificity = 0.93, sensitivity = 0.93, F1 score = 0.91), indicating that this model can reduce misdiagnosis and missed diagnosis, thereby improving the credibility of the medical system [17].

Approximately 8.2% of sepsis patients registered in the internal dataset from MIMIC-IV progressed to CCI, indicating that CCI is a rare event. The XGBoost model uses custom loss functions, regularization, and weighted voting mechanisms to handle imbalanced data. The RF model processes imbalanced data and

evaluates feature importance through ensemble learning, providing stable and accurate predictions for low-probability events [18,19]. Therefore, the XGBoost and RF models excel in predicting rare events, and our results also confirm this point.

There are 27,068 sepsis patients registered in the MIMIC-IV database. However, some patients are not suitable for inclusion in our study. First, pregnant women were excluded because they may differ significantly from general sepsis patients in terms of drug resistance, contraindications, and treatment plans. Second, because traumatic brain injury can also lead to CCI, patients with such injury were also excluded. Third, considering the negative impact of excessive missing data on model efficiency, such patients were excluded. Lastly, considering the inherent impact of chronic organ failure on SOFA scores, patients with chronic organ failure were excluded. Before model establishment, 7,991 patients were excluded, which may have helped us to avoid the bias caused by the above factors, improve model performance, reduce overfitting, and ultimately create a more reliable model.

Feature selection is crucial in machine learning. Selecting the most relevant feature from a large number of potential features is of great significance for enhancing interpretability and avoiding overfitting [20]. From 66 features, we selected 37 clinical feature variables with significant differences using statistical tests. Therefore, lasso regression was not used to eliminate more features, but the importance score for each feature was evaluated using each algorithm, which will weigh all features before building the model. This is first because lasso regression performs better on high-dimensional data, which contain many more features than the sample size, but its advantage is not marked for low-dimensional data, which contain relatively few features. Second, lasso regression may erroneously remove some useful features, leading to bias in the results. Finally, embedded methods such as DT, RF, and XGBoost automatically perform feature selection during model training. These methods not only consider the importance of features but also simultaneously perform feature selection and model training [20,21]. In our study, the performance of the model established using lasso regression to screen variables was worse than that without lasso regression (data not shown).

AST emerged as the top factor in feature importance rankings using the XGBoost and RF models. AST is an enzyme widely present in the liver, heart, muscles, and other tissues, and elevated AST levels typically indicate liver cell damage or inflammation. The liver releases many inflammatory cytokines during injury, such as tumor necrosis factor- α and interleukin-1 β , which can cause overexpression of surface molecules in antigen-presenting cells, thereby inducing a wide range of inflammatory responses through systemic circulation [22]. AST also reflects metabolic stress linked to insulin resistance and lipid metabolism disorders. Its main function is to catalyze the amino transfer reaction between aspartic acid and α -ketoglutarate, producing oxaloacetic acid and glutamic acid. Its excessive release into the bloodstream can cause abnormalities in glucose metabolism, lipid metabolism, and protein metabolism and even lead to internal metabolic disorders in the myocardium and skeletal muscles [23]. Therefore, high AST is not only a marker of liver injury but also plays an important role in PICS.

External validation is crucial for assessing model performance and generalizability. It ensures models handle unseen data in real-world applications and also develops reliable machine learning systems effectively. We used MIMIC-III as the external validation set because the time ranges covered by the MIMIC-IV and MIMIC-III databases are different, which helps to evaluate a model's generalization ability across different time periods, data structures, and patient populations. Therefore, the MIMIC-III database is very suitable for external validation of models developed using the MIMIC-IV database. The results showed that the XGBoost model also demonstrated good performance using external validation sets (AUROC = 0.72), highlighting its effectiveness in generalization beyond training data.

Several limitations of our study should be mentioned. Data from a single public database limits the generalizability of the results. Variables, such as CRP, albumin, and lymphocytes, which may be closely related to the development of CCI, were excluded because these features had too many missing values [24]. Finally, all data were collected within 24 hours after ICU admission, and we did not record data by hour or day for real-time analysis of patient physiological data. Future research may combine specific features to establish time-series models, aiming to establish dynamic warning models for predicting CCI.

Conclusions

In summary, we developed a machine learning model that demonstrated optimal performance on the internal test set and maintained robust performance on the external validation set. The model exhibits strong interpretability and predictive capability, making it suitable for predicting whether sepsis patients will develop CCI. This provides valuable predictive information for clinical decision-making.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Acquisition, analysis, or interpretation of data: Aijun Pan, Yulin Mei, Chunyan Zhu, Xiaoqin Fan, Lei Zhang

Critical review of the manuscript for important intellectual content: Aijun Pan, Yulin Mei, Chunyan Zhu, Xiaoqin Fan, Lei Zhang

Supervision: Aijun Pan

Concept and design: Yulin Mei, Meng Li, Yuqi Li, Ximei Sheng

Drafting of the manuscript: Yulin Mei, Meng Li, Yuqi Li, Ximei Sheng

Disclosures

Human subjects: Consent was obtained or waived by all participants in this study. Beth Israel Deaconess Medical Center (BIDMC) and Massachusetts Institute of Technology (MIT) issued approval 2001-P-001699/14 and 0403000206. Medical Information Mart for Intensive Care (MIMIC) was approved for research by the institutional review boards of BIDMC (2001-P-001699/14) and the Massachusetts Institute of Technology (0403000206) without a requirement for individual patient informed consent because it is deidentified and publicly available. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

- Nelson JE, Cox CE, Hope AA, Carson SS: Chronic critical illness. *Am J Respir Crit Care Med*. 2010, 182:446-54. [10.1164/rccm.201002-0210CI](https://doi.org/10.1164/rccm.201002-0210CI)
- Ohbe H, Matsui H, Fushimi K, Yasunaga H: Epidemiology of chronic critical illness in Japan: a nationwide inpatient database study. *Crit Care Med*. 2021, 49:70-8. [10.1097/CCM.0000000000004723](https://doi.org/10.1097/CCM.0000000000004723)
- Kahn JM, Le T, Angus DC, et al.: The epidemiology of chronic critical illness in the United States*. *Crit Care Med*. 2015, 43:282-7. [10.1097/CCM.0000000000000710](https://doi.org/10.1097/CCM.0000000000000710)
- Gardner AK, Ghita GL, Wang Z, et al.: The development of chronic critical illness determines physical function, quality of life, and long-term survival among early survivors of sepsis in surgical ICUs. *Crit Care Med*. 2019, 47:566-73. [10.1097/CCM.0000000000003655](https://doi.org/10.1097/CCM.0000000000003655)
- Polcz VE, Barrios EL, Larson SD, Efron PA, Rincon JC: Charting the course for improved outcomes in chronic critical illness: therapeutic strategies for persistent inflammation, immunosuppression, and catabolism syndrome (PICS). *Br J Anaesth*. 2024, 133:260-3. [10.1016/j.bja.2024.05.005](https://doi.org/10.1016/j.bja.2024.05.005)
- Sakusic A, Gajic O: Chronic critical illness: unintended consequence of intensive care medicine. *Lancet Respir Med*. 2016, 4:531-2. [10.1016/S2213-2600\(16\)30066-2](https://doi.org/10.1016/S2213-2600(16)30066-2)
- Mancini B, Liu J, Samuelsen A, Howrylak JA, Schultz L, Bonavia AS: Comparing long-term prognosis in chronic critically ill patients: a case series study of medical versus surgical sepsis. *Medicina (Kaunas)*. 2023, 59:10.3390/medicina59091617
- Sun B, Lei M, Wang L, et al.: Prediction of sepsis among patients with major trauma using artificial intelligence: a multicenter validated cohort study. *Int J Surg*. 2024, [10.1097/JS9.0000000000001866](https://doi.org/10.1097/JS9.0000000000001866)
- Fenlon C, O'Grady L, Doherty ML, Dunnion J: A discussion of calibration techniques for evaluating binary and categorical predictive models. *Prev Vet Med*. 2018, 149:107-14. [10.1016/j.prevetmed.2017.11.018](https://doi.org/10.1016/j.prevetmed.2017.11.018)
- Liu W, Tao G, Zhang Y, et al.: A simple weaning model based on interpretable machine learning algorithm for patients with sepsis: A research of MIMIC-IV and eICU databases. *Front Med (Lausanne)*. 2021, 8:814566. [10.3389/fmed.2021.814566](https://doi.org/10.3389/fmed.2021.814566)
- Johnson AE, Stone DJ, Celi LA, Pollard TJ: The MIMIC Code Repository: enabling reproducibility in critical care research. *J Am Med Inform Assoc*. 2018, 25:32-9. [10.1093/jamia/ocx084](https://doi.org/10.1093/jamia/ocx084)
- Liu P, Li S, Zheng T, et al.: Subphenotyping heterogeneous patients with chronic critical illness to guide individualised fluid balance treatment using machine learning: a retrospective cohort study. *EClinicalMedicine*. 2023, 59:101970. [10.1016/j.eclinm.2023.101970](https://doi.org/10.1016/j.eclinm.2023.101970)
- Obuchowski NA, Bullen JA: Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine. *Phys Med Biol*. 2018, 63:07TR01. [10.1088/1361-6560/aab4b1](https://doi.org/10.1088/1361-6560/aab4b1)
- Vickers AJ, Elkin EB: Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006, 26:565-74. [10.1177/0272989X06295361](https://doi.org/10.1177/0272989X06295361)
- Li W, Guo Q: Plotting receiver operating characteristic and precision-recall curves from presence and background data. *Ecol Evol*. 2021, 11:10192-206. [10.1002/ece3.7826](https://doi.org/10.1002/ece3.7826)
- Linden A: Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis. *J Eval Clin Pract*. 2006, 12:132-9. [10.1111/j.1365-2753.2005.00598.x](https://doi.org/10.1111/j.1365-2753.2005.00598.x)
- Yi F, Yang H, Chen D, et al.: XGBoost-SHAP-based interpretable diagnostic framework for alzheimer's disease. *BMC Med Inform Decis Mak*. 2023, 23:137. [10.1186/s12911-023-02238-9](https://doi.org/10.1186/s12911-023-02238-9)
- Khan MM, Alkhathami M: Anomaly detection in IoT-based healthcare: machine learning for enhanced security. *Sci Rep*. 2024, 14:5872. [10.1038/s41598-024-56126-x](https://doi.org/10.1038/s41598-024-56126-x)
- Colangelo G, Ribo M, Montiel E, et al.: PRERISK: a personalized, artificial intelligence-based and

- statistically-based stroke recurrence predictor for recurrent stroke. *Stroke*. 2024, 55:1200-9. [10.1161/STROKEAHA.123.043691](https://doi.org/10.1161/STROKEAHA.123.043691)
20. Wang L, Wang Y, Chang Q: Feature selection methods for big data bioinformatics: a survey from the search perspective. *Methods*. 2016, 111:21-31. [10.1016/j.ymeth.2016.08.014](https://doi.org/10.1016/j.ymeth.2016.08.014)
 21. Wang M, Zhuang B, Yu S, Li G: Ensemble learning enhances the precision of preliminary detection of primary hepatocellular carcinoma based on serological and demographic indices. *Front Oncol*. 2024, 14:1397505. [10.3389/fonc.2024.1397505](https://doi.org/10.3389/fonc.2024.1397505)
 22. Staller DW, Panigrahi SS, Jayasinghe YP, et al.: A novel phosphodiesterase inhibitor for the treatment of chronic liver injury and metabolic diseases. *Hepatology*. 2024, [10.1097/HEP.0000000000000999](https://doi.org/10.1097/HEP.0000000000000999)
 23. Dufour DR, Lott JA, Nolte FS, Gretch DR, Koff RS, Seeff LB: Diagnosis and monitoring of hepatic injury. II. Recommendations for use of laboratory tests in screening, diagnosis, and monitoring. *Clin Chem*. 2000, 46:2050-68. [10.1093/clinchem/46.12.2050](https://doi.org/10.1093/clinchem/46.12.2050)
 24. Drewry AM, Samra N, Skrupky LP, Fuller BM, Compton SM, Hotchkiss RS: Persistent lymphopenia after diagnosis of sepsis predicts mortality. *Shock*. 2014, 42:383-91. [10.1097/SHK.0000000000000234](https://doi.org/10.1097/SHK.0000000000000234)