

Comparison of Mortality Predictive Models of Sepsis Patients Based on Machine Learning

Ziyang Wang[†], Yushan Lan[†], Zidu Xu, Yaowen Gu, Jiao Li*

Institute of Medical Information/Medical Library, Chinese Academy of Medical Science & Peking Union Medical College, Beijing 100020, China

ABSTRACT

Objective To compare the performance of five machine learning models and SAPS II score in predicting the 30-day mortality amongst patients with sepsis.

Methods The sepsis patient-related data were extracted from the MIMIC-IV database. Clinical features were generated and selected by mutual information and grid search. Logistic regression, Random forest, LightGBM, XGBoost, and other machine learning models were constructed to predict the mortality probability. Five measurements including accuracy, precision, recall, F1 score, and area under curve (AUC) were acquired for model evaluation. An external validation was implemented to avoid conclusion bias.

Results LightGBM outperformed other methods, achieving the highest AUC (0.900), accuracy (0.808), and precision (0.559). All machine learning models performed better than SAPS II score (AUC=0.748). LightGBM achieved 0.883 in AUC in the external data validation.

Conclusions The machine learning models are more effective in predicting the 30-day mortality of patients with sepsis than the traditional SAPS II score.

Key words: MIMIC-IV; sepsis; machine learning; risk prediction

INTRODUCTION

Sepsis is a life-threatening organ dysfunction caused by a dysregulated host response to infection, as defined by the Third International Consensus^[1]. In developed countries, 2% of inpatients had sepsis, and sepsis may occur in 6%–30% of the patients who stay in intensive-care units (ICUs)^[2]. As a common cause of death in ICUs, sepsis leads to over 5.3 million deaths annually^[3]. In China, it has been reported that the prevalence of sepsis was 20.6% and the 90-day mortality rate was 35.3%^[4].

Mortality risk prediction in sepsis has been widely explored. Zhao L *et al.* have constructed a COX pro-

portional hazards model, using platelets as a prognostic marker of sepsis^[5]. Chen H *et al.* used multiple regression to analyze the relationship between central venous pressure and outcomes of septic patients^[6]. However, these traditional regression models assumed a linear independent relationship between variables, making it difficult to take a large number of valuable variables into account. In addition, the Simplified Acute Physiology Score (SAPS II) can be used to assess the severity of sepsis and estimate the mortality risk but with limited prediction accuracy.

Currently, machine learning methods have been widely used in mortality risk prediction^[7–10]. HOU N *et al.* has used XGBoost to predict the 30-day mortality of septic patients and found the reported model performance was significantly better than that of the traditional regression model^[11]. Wang used random forest algorithm to predict the onset of sepsis^[12]. In our current study, we constructed five machine learning models for mortality prediction in sepsis patients, including Logistic regression (LR), Bayesian network (BN), Random forest (RF), XGBoost, and LightGBM trained on

Received April 21, 2022; accepted August 10, 2022; published online September 20, 2022.

*Corresponding author E-mail: li.jiao@imicams.ac.cn

[†]The first two authors contributed equally to this article.

© The authors 2022. Published by Chinese Academy of Medical Sciences. This is an open access article attributed under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0>).

the data from MIMIC-IV database. We used accuracy, F1 score, area under the receiver operating characteristic curve (AUC), and other metrics to evaluate the performance of these models, with an attempt to figure out the best machine learning model for identifying mortality risk in septic patients. We also compared them with the traditional SAPS II. Finally, we adopted the SHapley Additive exPlanations (SHAP) algorithm to interpret the prediction results and visualize the contributions of risk factors. The workflow of our research is shown in **Fig. 1**.

MATERIALS AND METHODS

Data sources

We obtained the clinical data of sepsis patients from the MIMIC-IV database^[13], and all the co-authors were approved to extract data for this study (Certificate Number: 44835985). MIMIC-IV is a large publicly available and single-center database, comprising de-identified health-related data in 2008-2019 from patients who were admitted to the critical care units of the Beth Israel Deaconess Medical Center, and the data were collected from Metavision bedside monitors^[13]. MIMIC-IV database contains many clinical features, such as age, gender, vital signs, fluid balance, vital status, laboratory tests, and medications.

Using PostgreSQL (6.1) and Navicat Premium (15.0.27), we extracted the clinical data based on the

Sepsis-3 (SOFA score>2) patient concept table. The inclusion criteria were: a) patients who were older than 18 years and younger than 89 years; b) the length of ICU stay was over 24 hours; and c) once the patients had multiple admissions with sepsis, only the last admission was analyzed. Besides, we used other concept tables to extract 47 related features, including: a) demographic information such as age, and gender; and b) indicators of laboratory tests such as blood routine and urine output. The detailed process of patients selection for data extraction is shown in **Fig. 2**.

Data pre-processing

We obtained the raw data of 19,903 patients who were diagnosed with sepsis and calculated the proportion of missing values for each feature. The features with a missing ratio greater than 5% are listed in **Table 1**. Features with a missing ratio higher than 10% were deleted, and features with a missing ratio of less than 10% were filled with the mode. Finally, 12,664 patients were included for the final analysis.

Feature selection

We used the grid search and the maximal information coefficient (MIC) to select valuable features.

$$MIC = H(X;Y) = H(X) - H(X|Y) \quad (1)$$

In the formula (1), $H(X)$ is the information coefficient, and $H(X;Y)$ is the mutual information of X and Y .

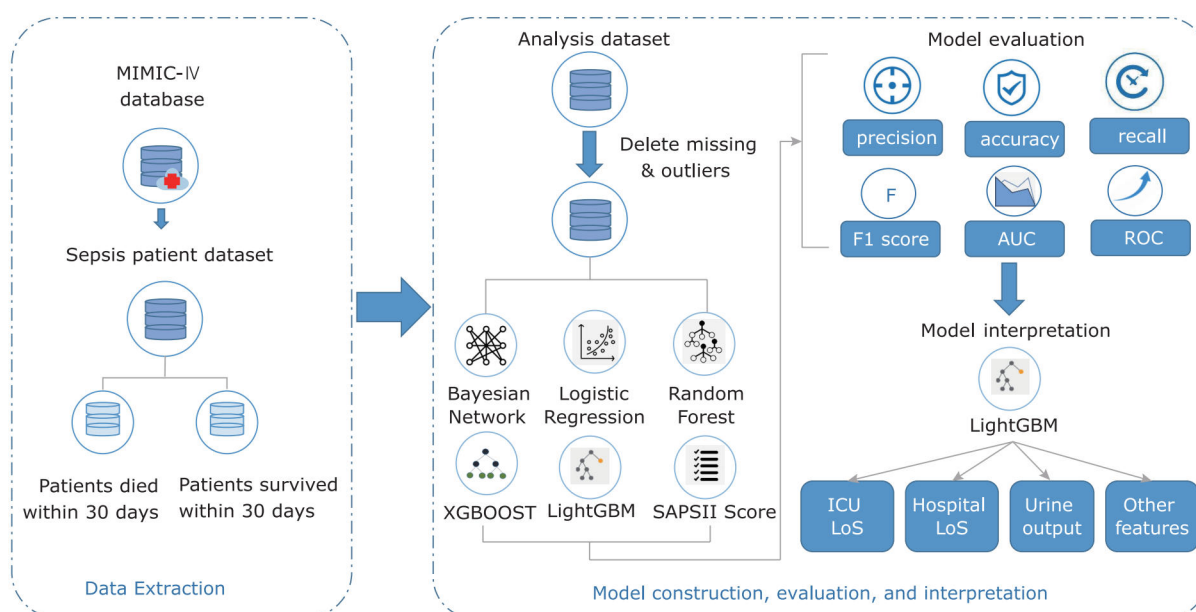


Figure 1. Workflow of research to compare mortality predictive models of sepsis patients based on machine learning
AUC: area under curve; ROC: Receiver operating characteristic; ICU: intensive care unit; LoS: length of stay

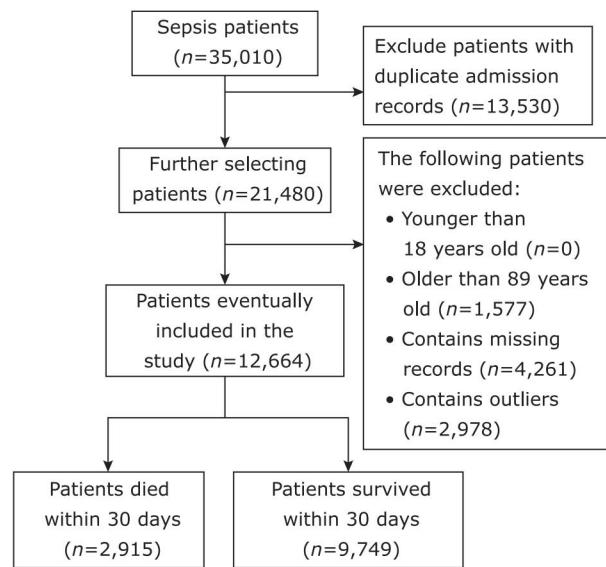


Figure 2. The patient selection processes for data extraction from the MIMIC-IV database.

Table 1. Features of raw data with missing ratio > 5% obtained from MIMIC-IV (n=19,903)

Feature	Missing records (n)	Missing ratio
C reactive protein _max	15,373	77.2%
Height_mean	7,128	35.8%
vent_status	1,803	9.1%
Inhalation_max	1,159	5.8%

It represents the reduction in uncertainty of Y when X is known. Features were incorporated into grid search in descending order of MIC values.

For the grid search-based feature selection, we adopted three models (LR, BN, and RF) to search for

the best number of features and bins. Then the AUC scores of the above models with different numbers of features were calculated, therefore the optimum numbers of included features could be determined once the AUC reached the ceiling. The results of grid searches are shown in **Fig. 3A**.

Fig. 3A shows that the model performance was increased as the number of included features increased, and the model achieved the optimum performance when the number of model features increased to 19 (shown in **Fig. 3A** with red dashed line). Therefore, we selected the top 19 features with the largest MIC value to construct machine learning models, and the selected features are listed in **Table 2**. In addition, as Bayesian networks only accept discrete-type features, we adopted a percentile discretization to discretize the continuous features. In this situation, we also used grid searches to determine the best discretization strategy, and the results of grid searches are shown in **Fig. 3B**.

Fig. 3B indicates that the model performance increased as the number of bins increased, and discretizing the continuous feature into 5 bins was a sub-optimal point. The more discretization bins are, the more difficult to explain the model. Considering the performance of bins and the average AUC of BN, we finally discretized the continuous variables into 5 bins. The feature selection results are listed in **Table 2**, where the features are arranged in descending order of the maximal information coefficient (MIC). The features discretization strategies are shown in **Table S1**.

The feature we selected had 18 continuous variables and one category variable. The categorical vari-

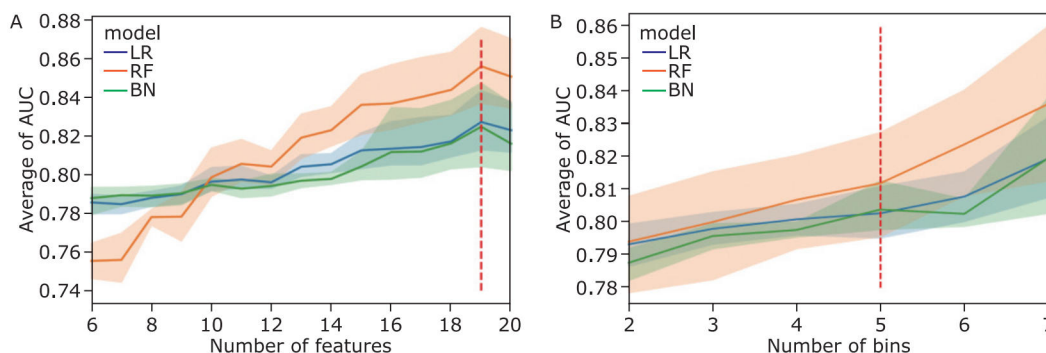


Figure 3. Performance of models at different number of features.

Three lines represent AUC scores in different models with top n features selected by MIC in (A), with different number of bins for continuous features discretization in (B). The color shadow area represents the fluctuation of AUC in different folds. The red dashed line represents the feature numbers we chose.

LR: Logistic Regression; RF: Random Forest; BN: Bayesian Network.

Table 2. MIC values of the features in feature selection.

NO	Feature	MIC
1	urine_max	0.0441
2	los_hospital	0.0418
3	aniongap_max	0.0370
4	specimen_count	0.0347
5	antibiotic_num	0.0346
6	bun_max	0.0325
7	bun_min	0.0303
8	aniongap_min	0.0295
9	bicarbonate_min	0.0273
10	inr_max	0.0271
11	creatinine_max	0.0270
12	vent_status	0.0224
13	los_ICU	0.0223
14	charlson_score_max	0.0185
15	bicarbonate_max	0.0181
16	chloride_max	0.0180
17	heart_rate_max	0.0173
18	sodium_max	0.0160
19	spo2_mean	0.0158

MIC: maximal information coefficient of the features and outcome. In this study we chose death within 30 days as the predicted outcome. aniongap: anion gap; vent_status: ventilation; los: length of stay; inr: international normalized ratio; bun: blood urea nitrogen; spo2: oxygen saturation; num: number of the events occurred in ICU; ICU: intensive care unit.

able vent_status means ventilation status. It has five values: low flow oxygen ($n=4,105$, 32%), high flow oxygen ($n=366$, 3%), non-invasive mechanical ($n=263$, 2%), invasive mechanical ($n=7,679$, 61%), and tracheotomy ($n=251$, 2%). In order to make the model more effective, we calculated two features including specimen_count and antibiotic_num, which represent the positive specimen and the number of antibiotic use in ICU.

Model construction

We constructed five machine learning models, including Logistic Regression (LR), Bayesian Network (BN), Random Forest (RF), XGBoost, and LightGBM. All experiments were performed by Python. Scikit-learn, Statsmodels, and other Python packages were used to develop the proposed and comparative models, respectively.

LR is a statistical method for classification and regression. LR can model the probability of a discrete

outcome given an input variable. As opposed to linear regression, LR does not require a linear relationship between input and output variables. So it is widely used to predict diseases and handle other classification questions^[6]. To optimize the algorithm, we use L-BFGS method to improve the effectiveness of iteration and adopted L2 regularization to avoid model overfitting.

BN was developed from Bayesian methods. It can explain causal relationships in complex structures by dealing with missing data and uncertain knowledge^[14]. The construction of BN includes two steps: structure learning and parameter learning. For structure learning, a previous study had indicated that the score-based method is more robust than the constraint-based method^[15]. Therefore, we used the hill-climbing search algorithm as the adopted score-based structure learning algorithm. For parameter learning, the maximum likelihood estimation method was adopted to learn the parameters of the BN.

RF is a bagging-based ensemble method that contains multiple decision trees. Each decision tree consists of multiple nodes, and each node represents a classification rule based on a single variable^[16]. In our current study, we applied 100 estimators in the prediction model. In addition, the quality of split-measured criteria was "Gini".

XGBoost, a boosting method developed from gradient boosting decision tree (GBDT), can efficiently and flexibly handle missing data and achieve higher performance than GBDT^[17]. We used gbtrees as our booster to construct this model. Besides, the number of iterations was 100 and the learning rate was 0.5. LightGBM is a gradient boosting method developed by Microsoft and combines many advantages of XGBoost.

LightGBM can handle large and structured data and has ultra-high training speed^[18]. We configured the number of iterations as 100 and used all observations in the training data for each iteration to construct LightGBM. To achieve better performance of our model, we set the learning rate as 0.05.

The new Simplified Acute Physiology Score (SAPS II) was used to assess the severity of sepsis and estimate the mortality rate. The mortality rate calculated by SAPS II can be formulated as^[19]:

$$SAPS\ II\ mortality = \frac{e^x}{1 + e^x} \quad (2)$$

$$x = -7.7631 + 0.0737 \times Score + 0.9971 \times [\ln(Score + 1)] \quad (3)$$

where Score is the SAPS II score, which is the worst

score in the past 24 hours measured on the SAPS Scale. Previous study has demonstrated that the SAPS II score is an acceptable predictor of mortality in sepsis^[20].

Model evaluation and comparison

Five metrics of these models were used, including accuracy, precision, recall, F1 score, and AUC score. To verify the stability of those models, we used 10-fold cross-validation to estimate the above metrics. Moreover, we adopted the SAPS II as a baseline method and calculated the mortality risks and the above metrics for performance comparisons. Furthermore, we used *t* test to compare the performance of these five machine learning models. External validation was implemented to avoid the bias of the conclusion.

Model interpretability

Machine learning models have been proven to achieve better performance than traditional methods in predicting the mortality of patients with sepsis. However, these models are often described as "black-box models", and it is often difficult to interpret the model predictions with relevant clinical knowledge. Therefore, in this study, we used the SHAP, a method widely used in machine learning interpretation^[21,22], to analyze the impacts of different features on the model results and provide a

reasonable interpretation of the model. Formula (4) was adopted when explaining SHAP of the tree model:

$$g(x') = \phi_0 + \sum_{j=1}^M \phi_j \quad (4)$$

where g is the interpretation function, ϕ_0 is the mean of the predicted values for all samples (called base value), and ϕ_j is Shapley value of the j^{th} feature. The Shapley value is the average marginal contribution of the features in all possible feature combination^[23]. M is the number of features. Our code is available in https://github.com/zityang-W/MIMIC_IV_Sepsis.

RESULTS

Descriptive statistical analysis

A total of 12,664 ICU patients with sepsis were enrolled in this study, including 5,249 males (41.5%) and 7,415 females (58.5%). Up to 2,915 patients (23.0%) died within 30 days, while 9,749 (77.0%) survived. The average age of patients who died within 30 days and those who survived were 68.32 years and 65.89 years, respectively. Among patients with sepsis, there were more women than men, and the average age of death was higher than the average age of survival. Descriptive statistical analysis was performed for each feature, and results of continuous and categorical fea-

Table 3. Comparisons of features between survived and dead patients ($n=12,664$)

Features	All ($n=12,664$)	Survival ($n=9,749$)	Death ($n=2,915$)	t/χ^2	<i>P</i> value
aniongap_max (mmol/L, mean±SD)	16.99 ± 5.37	19.43 ± 6.36	16.26 ± 4.81	24.85	<0.001
aniongap_min (mmol/L, mean±SD)	13.12 ± 3.70	14.72 ± 4.47	12.64 ± 3.29	23.32	<0.001
antibiotic_num (mean±SD)	6.31 ± 5.72	8.38 ± 6.00	5.69 ± 5.48	21.67	<0.001
bicarbonate_max (mmol/L, mean±SD)	24.34 ± 4.71	23.39 ± 5.44	24.62 ± 4.43	-11.14	<0.001
bicarbonate_min (mmol/L, mean±SD)	21.15 ± 5.21	19.46 ± 6.04	21.65 ± 4.82	-17.93	<0.001
bun_max (mg/L, mean±SD)	32.40 ± 25.09	42.77 ± 29.93	29.30 ± 22.54	22.46	<0.001
bun_min (mg/L, mean±SD)	26.98 ± 21.93	35.84 ± 26.48	24.33 ± 19.61	21.75	<0.001
charlson_score_max (mean±SD)	6.59 ± 3.03	7.63 ± 2.94	6.28 ± 2.98	21.79	<0.001
creatinine_max (mmol/L, mean±SD)	1.64 ± 1.49	2.03 ± 1.56	1.53 ± 1.45	15.45	<0.001
chloride_max (mmol/L, mean±SD)	106.21 ± 6.73	106.4 ± 6.29	105.58 ± 7.98	14.26	<0.001
inr_max (mean±SD)	1.73 ± 1.27	2.11 ± 1.62	1.61 ± 1.11	15.47	<0.001
los_hospital (days, mean±SD)	13.03 ± 13.76	11.73 ± 15.19	13.42 ± 13.28	-5.42	<0.001
los_ICU (days, mean±SD)	6.16 ± 6.80	7.14 ± 5.79	5.87 ± 7.05	9.85	<0.001
heart_rate_max (bpm, mean±SD)	107.82 ± 21.48	106.05 ± 20.48	113.73 ± 23.60	-14.54	<0.001
spo2_mean (% , mean±SD)	96.94 ± 2.25	97.09 ± 1.95	96.44 ± 2.99	14.49	<0.001
sodium_max (mmol/L, mean±SD)	140.14 ± 5.36	140.32 ± 6.65	140.08 ± 4.91	1.81	0.0696
specimen_count (mean±SD)	6.11 ± 5.51	8.14 ± 5.81	5.51 ± 5.26	21.91	<0.001
urine_max (ml, mean±SD)	1,751.39 ± 1,253.79	1,284.30 ± 1,234.80	1,891.05 ± 1,225.37	-23.32	<0.001
vent_status [n(%)]*	12,664 (100)	9,749 (77)	2,915 (23)	503.11	<0.001

*There was significant difference between survival and death groups within 30 days.

tures statistical analysis between survived and death patients are shown in **Table 3** and **4**, respectively.

Each feature selected by MIC was significantly different in groups except sodium_max. However, we had adjusted the constraints of significance to incorporate more potential features.

Model comparison

In this study, five machine learning models were constructed and their performances were validated under 10-fold cross-validation. To compare the performance of the five machine learning models with SAPS II score predictor clearly, we aggregated the ROC curves in 10-fold cross-validation of each method and plotted them in **Fig. 4**, where the solid lines indicate the average ROC curves of different methods and the gray areas indicate the fluctuation ranges of cross-validations. The results demonstrated that LightGBM outperformed other methods. Moreover, the performances of LR and BN were lower than the other three machine learning models but still higher than the SAPS II score.

We calculated the means and standard deviations of accuracy, precision, recall, F1 score, and AUC scores to evaluate the performance of these methods by 10-fold cross-validation (**Table 4**). To illustrate the differences between LightGBM and other machine learning models, we apply the *t* test to compare the metrics

among different models. All the machine learning models performed better than SAPS II score; LightGBM achieved the best performance in accuracy, precision, and AUC score. Compared to the traditional SAPS II score, the relative improvement of accuracy, precision, recall, F1 score and AUC was 11.4%, 30.0%, 38.1%, 33.1%, 20.3%, respectively. Among the machine learning models, BN performed the worst.

LightGBM significantly outperformed BN and LR in each metrics and significantly outperformed other 4 machine learning models in AUC. Therefore, LightGBM was the best-performing predictive model in our study.

To further evaluate the preference of machine learning model in predicting the 30-day mortality rate in sepsis patients, LightGBM was used as an example to validate our proposed machine learning-based methods on external data. The data supplied by Nianzong Hou^[11] was chosen as the validation dataset to avoid bias, which extracted from MIMIC-III. There are 4,560 patients in the validation dataset, with 889 death and 3,671 survivals. There were some differences between our data and validation data. The three missing features (antibiotic_num, specimen_count, and charlson_score_max, with predicted accuracy 0.942, 0.938, and 0.895, respectively) were filled by the random forest method through other identical features. In addition, the target result was deleted during the filling of the missing features. Finally, the LightGBM showed good discriminatory power, with AUC being 0.883, and accuracy being 0.852, precision being 0.612, recall being 0.709, and F1 score being 0.657.

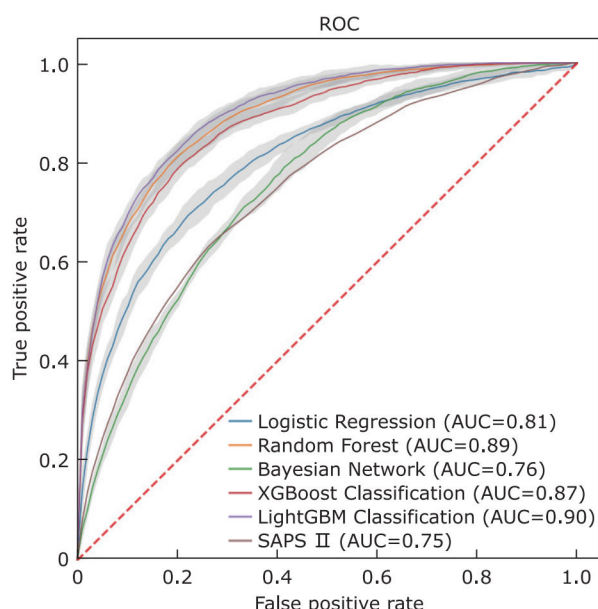


Figure 4. The ROC of the five the machine learning models and SAPS II score prediction.

AUC: area under the curve; ROC: receiver operating characteristic.

Model explanation

In this study, we took LightGBM as the state-of-the-art method and plotted the SHAP values (**Fig. 5**). The features are arranged in descending order of importance (**Fig. 5**), which shows that length of hospital stay, length of ICU stay, and maximum urine output are more important for the prediction capability of models. Longer length of stay was associated with lower 30-day mortality rate. Maximum urine output seemed to have similar impact. Higher Charlson score was associated with higher 30-day mortality rate. Excessive antibiotic use was also reported to increase 30-day mortality rate.

DISCUSSION

Sepsis is a life-threatening organ dysfunction

Table 4. Performance comparison of five machine learning models and SAPS II score predictor

Model	Accuracy	Precision	Recall	F1 Score	AUC
SAPS II Score	0.725	0.430	0.604	0.502	0.748
Logistic Regression	0.751(0.02) [#]	0.486(0.03) [#]	0.719(0.06) [#]	0.571(0.02) [#]	0.807(0.01) [#]
Bayesian Network	0.654(0.03) [#]	0.378(0.02) [#]	0.773(0.06) [*]	0.507(0.01) [#]	0.756(0.01) [#]
Random Forest	0.806(0.02) [#]	0.558(0.04)	0.807(0.05)	0.657(0.01)	0.891(0.01) [*]
XGBoost	0.795(0.02)	0.541(0.04)	0.804(0.05)	0.645(0.02) [*]	0.875(0.01) [#]
LightGBM	0.808(0.02)	0.559(0.04)	0.834(0.04)	0.668(0.02)	0.900(0.01)

Compared to the performance of LightGBM, ^{*} $P < 0.05$, [#] $P < 0.001$ (by *t*-test).

with high mortality rates in various countries but lacks specific treatment modality^[24]. Currently, sepsis management relies mainly on the early identification^[21]. Predicting the mortality of patients with sepsis can help improve the quality of patient management and prognosis and reduce the mortality rate. In this study, we constructed five machine learning models based on the clinical data of sepsis patients from the MIMIC-IV database. These models achieved desirable performances in predicting the 30-day mortality rate in sepsis patients in ICU, which can support clinical decision-making and optimize the public health resource allocation. Among the five machine learning models, the LightGBM model performed best with an average AUC score up to 0.9. This also coincides

with the findings of HOU N *et al.* who predicted the 30-day mortality rate in sepsis patients in MIMIC-III database^[11]. In addition, all the constructed machine learning models performed better than the traditional SAPS II.

Meanwhile, the SHAP values of LightGBM also show that maximum urine output plays an important role in predicting long-term outcomes in sepsis patients^[22], suggesting large urine output and high blood urea nitrogen (BUN) are risk factors for mortality in sepsis patients because these patients may have varying degrees of renal insufficiency. Currently, evidence has indicated that persistent oliguria is associated with death in patients with sepsis^[25], which may be explained that patients who have oliguria are more likely to suffer from other severe diseases (e.g. severe renal impairment), which increase the risk of death. Moreover, prolonged hospital and ICU stay may also increase the mortality, probably due to a positive correlation between length of stay and disease severity. We also find features of vent_status and antibiotics_num, which are often related to a severe condition and a poor prognosis, were associated with the 30-days death. Patients with a high Charlson score usually have more complex and severe comorbidities, which are also associated with higher risk of death. However, the positive correlation cannot make the conclusion, for the bias need to be excluded to reach a conclusion with more confidence. However, the data extracted from the MIMIC-IV database were retrospective, which might cause selection bias. To address this issue, external validation was implemented to test the stability of the machine learning model.

In our future studies, clinical guidelines will be further utilized in feature selection to improve the interpretability of the models. In addition, we will apply this method in real-world studies for the early identification of patients with sepsis.

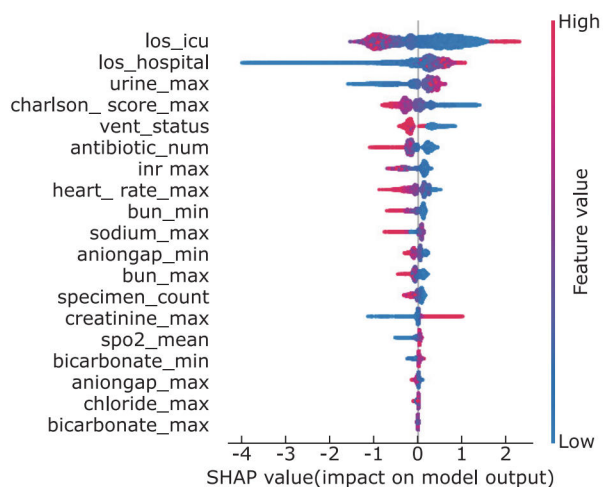


Figure 5. Feature explanation of lightGBM model using the SHapley Additive exPlanations (SHAP) value. Each point in the figure is the SHAP value of a sample, and the color of the point represents the value of the feature. Red represents high values while blue represents low values. SHAP value greater than 0 indicates that the feature is a risk factor for death, while less than 0 indicates a protective factor.

Supplementary material

Available at <http://cmsj.cams.cn/EN/10.24920/004102>.

Table S1. Results of feature selection by MIC and features discretization code

Acknowledgments

The authors thank the contributors of MIMIC-IV database and the reviewers for their suggestions to improve the manuscript.

Conflict of interests

None disclosed.

Authors' contributions

Li J proposed the research direction and revised and finalized the manuscript. Xu ZD and Gu YW contributed to the research conception and made critical revisions on the manuscript. Wang ZY and Lan YS analyzed and interpreted data and drafted the manuscript. All authors confirmed and agreed to the final version of the article.

Funding

The work was supported by the Chinese Academy of Medical Sciences "Research on Key Technologies of Medical Knowledge Management and Intelligent Knowledge Service" (2021-I2M-1-056) and "Research on Key Issues of Medical Artificial Intelligence Technology and Human-Computer Interaction" (2018-I2M-AI-016), and the National Key R&D Program of China "Construction of Precision Medicine Ontology and Semantic Network" (2016YFC0901901) and "Research and Development of Chinese Multi-Group Multi-omics Reference Database System" (2017YFC0907503).

REFERENCES

1. Singer M, Deutschman CS, Seymour CW, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* 2016; 315(8):801-10. doi:10.1001/jama.2016.0287.
2. Martin GS. Sepsis, severe sepsis and septic shock: changes in incidence, pathogens and outcomes. *Expert Rev Anti Infect Ther* 2012; 10(6):701-6. doi:10.1586/eri.12.50.
3. Song J, Park DW, Moon S, et al. Diagnostic and prognostic value of interleukin-6, pentraxin 3, and procalcitonin levels among sepsis and septic shock patients: a prospective controlled study according to the Sepsis-3 definitions. *BMC Infect Dis* 2019; 19:68. doi:10.1186/s12879-019-4618-7.
4. Xie J, Wang H, Kang Y, et al. The epidemiology of sepsis in Chinese ICUs: A national cross-sectional survey. *Crit Care Med* 2020; 48(3):e209-e218. doi:10.1097/CCM.0000000000004155.
5. Zhao L, Zhao L, Wang YY, et al. Platelets as a prognostic marker for sepsis: a cohort study from the MIMIC-III database. *Medicine* 2020; 99(45):e23151. doi:10.1097/MD.00000000000023151.
6. Chen H, Zhu Z, Zhao C, et al. Central venous pressure measurement is associated with improved outcomes in septic patients: an analysis of the MIMIC-III database. *Crit Care* 2020; 24(1):433. doi:10.1186/s13054-020-03109-9.
7. Zhu C, Xu Z, Gu Y, et al. Prediction of post-stroke urinary tract infection risk in immobile patients using machine learning: an observational cohort study. *J Hosp Infect* 2022; 122:96-107. doi:10.1016/j.jhin.2022.01.002.
8. Wang Y, Sun F, Hong G, et al. Thyroid hormone levels as a predictor marker predict the prognosis of patients with sepsis. *Am J Emerg Med* 2021; 45:42-7. doi:10.1016/j.ajem.2021.02.014.
9. Hou N, Li M, He L, et al. Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost. *J Transl Med* 2020; 18: 462. doi:10.1186/s12967-020-02620-5.
10. Feng M, McSparron JI, Kien DT, et al. Transthoracic echocardiography and mortality in sepsis: analysis of the MIMIC-III database. *Intens Care Med* 2018; 44(6): 884-92. doi:10.1007/s00134-018-5208-7.
11. Hou N, Li M, He L, et al. Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost. *J Transl Med* 2020; 18:462. doi:10.1186/s12967-020-02620-5.
12. Wang D, Li J, Sun Y, et al. A machine learning model for accurate prediction of sepsis in ICU patients. *Front Public Health* 2021; 9: 754348. doi:10.3389/fpubh.2021.754348.
13. Johnson A, Bulgarelli L, Pollard T, et al. MIMIC-IV (version 0.4). *PhysioNet* 2020. <https://doi.org/10.13026/a3wn-hq05>.
14. Uusitalo L. Advantages and challenges of Bayesian networks in environmental modelling. *Ecol Modell* 2007; 203(3): 312-18. doi:10.1016/j.ecolmodel.2006.11.033.
15. Mihaljević B, Bielza C, Larrañaga P. Bayesian networks for interpretable machine learning and optimization. *Neurocomputing* 2021; 456: 648-65. doi:10.1016/j.neucom.2021.01.138.
16. Hanco M, Grendár M, Snopko P, et al. Random forest-based prediction of outcome and mortality in patients with traumatic brain injury undergoing primary decompressive craniectomy. *World Neurosurg* 2021; 148: e450-e458. doi:10.1016/j.wneu.2021.01.002.
17. Davagdorj K, Pham VH, Theera-Umporn N, et al. XGBoost-based framework for smoking-induced noncommunicable disease prediction. *Int J Environment Res Public Health* 2020;17(18): e6513. doi:10.3390/ijerph17186513.
18. Zhang C, Lei X, Liu L. Predicting metabolite-disease associations based on lightgbm model. *Front Genet* 2021;12: 660275. doi:10.3389/fgene.2021.660275.
19. Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute

- Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA* 1993; 270(24): 2957-63. doi:10.1001/jama.270.24.2957.
20. Moreno-Torres V, Royuela A, Múñez E, et al. Better prognostic ability of NEWS2, SOFA and SAPS-II in septic patients. *Medicina Clinica* 2021; 159(5):224-9. doi:10.1016/j.medcli.2021.10.021.
21. Cohen J, Vincent JL, Adhikari NKJ, et al. Sepsis: a roadmap for future research. *Lancet Infect Dis* 2015;15(5): 581-614. doi:10.1016/S1473-3099(15)70112-X.
22. Zhang ZQ. Effect of changes in urine volume on prognosis of patients with sepsis and acute kidney injury after continuous renal replacement therapy. *Chin Med Pharm* 2021; 11(12): 178-82.
23. Lundberg S, Lee SI. A Unified Approach to Interpreting Model Predictions, carXiv: 1705.07874. Available from: <http://doi.org/1048550/arXiv.1705.07874>.
24. Dugar S, Choudhary C, Duggal A. Sepsis and septic shock: Guideline-based management. *Cleveland Clin J Med* 2020; 87(1): 53-64. doi:10.3949/ccjm.87a.18143.
25. Vincent JL, Ferguson A, Pickkers P, et al. The clinical relevance of oliguria in the critically ill patient: analysis of a large observational database. *Criti Care* 2020; 24: 171. doi:10.1186/s13054-020-02858-x.

(Edited by Liangjun Gu)

专栏：科学数据共享与重用
论著

基于机器学习的脓毒症死亡率预测模型对比研究

王梓阳[#]，兰雨姗[#]，徐子犊，顾耀文，李姣^{*}

中国医学科学院 北京协和医学院医学信息研究所 / 图书馆，北京 100020，中国

摘要

目的 比较五个机器学习模型和 SAPS II 评分在预测脓毒症患者 30 天内死亡率方面的表现。

方法 从 MIMIC-IV 数据库中提取败血症患者相关数据，生成临床特征，并通过互信息法和网格搜索进行特征筛选。构建逻辑回归、随机森林、LightGBM、XGBoost 等机器学习模型，预测脓毒症患者 30 天内死亡率。此外，还获得了包括准确率、精确度、召回率、F1 得分和受试者工作特性曲线下面积（area under the curve, AUC）在内的五个模型评估指标。最后，在外部数据集中验证了模型的效果。

结果 LightGBM 的表现优于其他方法，取得了最高的 AUC (0.900)、准确率 (0.808) 和精确度 (0.559)。所有机器学习模型的表现都优于 SAPS II 评分 (AUC=0.748)。在外部数据集的验证中 LightGBM 的 AUC 达到 0.883。

结论 机器学习模型在预测败血症患者的死亡率方面被认为是比传统的 SAPS II 评分更有效的方法。

关键词：MIMIC-IV；脓毒血症；机器学习；风险预测

基金：中国医学科学院“医学知识管理与智能化知识服务关键技术研究” (2021-I2M-1-056)；中国医学科学院“医学人工智能技术与人机交互关键问题研究” (2018-I2M-AI-016)；中国国家重点研发计划“精准医学本体和语义网络构建” (2016YFC0901901)；中国国家重点研发计划“中国人群多组学参比数据库系统研发” (2017YFC0907503)

*** 通讯作者**：李姣 li.jiao@imicams.ac.cn