Li *et al. BMC Medical Informatics and Decision Making* (2025) 25:99
https://doi.org/10.1186/s12911-025-02878-z

BMC Medical Informatics and Decision Making

RESEARCH **Open Access**

Machine learning-based prediction model for patients with recurrent Staphylococcus aureus bacteremia

Yuan Li[1,2†], Shuang Song[1], Liying Zhu[1], Xiaorun Zhang[1], Yijiao Mou[1], Maoxing Lei[1], Wenjing Wang[1*] and Zhen Tao[1*]

Abstract

Background Staphylococcus aureus bacteremia (SAB) remains a significant contributor to both community-acquired and healthcare-associated bloodstream infections. SAB exhibits a high recurrence rate and mortality rate, leading to numerous clinical treatment challenges. Particularly, since the outbreak of COVID-19, there has been a gradual increase in SAB patients, with a growing proportion of (Methicillin-resistant Staphylococcus aureus) MRSA infections. Therefore, we have constructed and validated a pediction model for recurrent SAB using machine learning. This model aids physicians in promptly assessing the condition and intervening proactively.

Methods The patients data is sourced from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database version 2.2. The patients were divided into training and testing datasets using a 7:3 random sampling ratio. The process of feature selection employed two methods: Recursive Feature Elimination (RFE) and Least Absolute Shrinkage and Selection Operator (LASSO). Prediction models were built using Extreme Gradient Boosting (XGBoost), Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), and Artificial Neural Network (ANN). Model validation included Receiver Operating Characteristic (ROC) analysis, Decision Curve Analysis (DCA), and Precision-Recall Curve (PRC). We utilized SHAP (SHapley Additive exPlanations) values to demonstrate the significance of each feature and explain the XGBoost model.

Results After screening, MRSA, PTT, RBC, RDW, Neutrophils_abs, Sodium, Calcium, Vancomycin concentration, MCHC, MCV, and Prognostic Nutritional Index(PNI) were selected as features for constructing the model. Through combined evaluation using ROC、 DCA and PRC, XGBoost demonstrated the best predictive performance, achieving an AUC value of 0.76 (95% CI: 0.66–0.85) in ROC and 0.56 (95% CI: 0.37–0.75) in PRC. Building a website based on the Xgboost model. SHAP illustrated the feature importance ranking in the XGBoost model and provided examples to explain the XGBoost model.

[†]Yuan Li the first author of the paper

*Correspondence:
Wenjing Wang
njmuwwj@yeah.net
Zhen Tao
tz1010@126.com

Full list of author information is available at the end of the article

Conclusions The adoption of XGBoost for model development holds widespread acceptance in the medical domain. The prediction model for recurrent SAB, developed by our team, aids physicians in timely diagnosis and treatment of patients.

Keywords Staphylococcus aureus bacteremia, Machine learning, Prediction model, Readmission, Web app

Background

Staphylococcus aureus is one of the most common causes of both hospital-acquired and community-acquired bloodstream infections. Staphylococcus aureus bacteremia (SAB) is characterized by a unique capability to disseminate through the bloodstream and affect various organs throughout the body [1]. Approximately 20% of SAB patients die in 30 days [2]. Methicillin-resistant Staphylococcus aureus (MRSA) is a cause of staph infection that is difficult to treat because of resistance to some antibiotics, leading to higher mortality and recurrence rates. Weiner-Lastinger et al. analyzed the National Healthcare Safety Network (NHSN) database in the United States, which publishes Standardized Infection Ratios (SIRs). Before the COVID-19, the SIRS for Methicillin-Resistant Staphylococcus aureus Bacteremia (MR-SAB) showed a significant annual decline. However, in the last two quarters of 2020, the national MR-SAB SIR increased by 23% and 34%, respectively, compared to the previous year. Some states reported a staggering 99% growth during this period, underscoring the profound impact of the pandemic on infection prevention and control [3]. Following the resolution of an initial SAB infection, even with standardized antibiotic treatment, approximately 2–20% of patients will experience a recurrence, known as recurrent SAB [1]. Recurrent SAB can lead to an increased incidence of complications such as acute kidney injury, venous thrombosis, and cardiovascular and cerebrovascular diseases, thereby raising the mortality rate among hospitalized patients [4]. Strengthening care and extending the duration of antibiotic treatment for patients identified by the model as high risk for recurrent SAB can reduce their risk of readmission, thereby effectively improving patient prognosis. Predicting the readmission rate of SAB patients is essential for clinicians to optimize treatment strategies and allocate resources effectively. In recent years, a multitude of machine learning (ML) algorithms, such like bagging, boosting, and stacking ensembles, which are a collection of data analysis methods that learn from data to develop algorithms, have been widely used in health science, agronomy, finance and other fields [5–7]. These methods have demonstrated superiority over conventional statistical approaches [8]. It is increasingly common to develop ML models to predict the occurrence, recurrence, and complications of diseases. These models can alleviate the medical burden and reduce patient mortality rates, a benefit that is gradually being recognized by the public [9, 10]. However, there is no existing research utilizing ML algorithms to predict readmissions for SAB. It is challenging to identify which SAB patients require further treatment and care interventions. In this study, we compare various algorithms, establish and validate a model, and develop a website for clinical practitioners to use, aims to assist physicians in developing more effective clinical treatment plans to reduce the readmission rates of patients with SAB.

Materials and methods

Database

The MIMIC-IV (Medical Information Mart for Intensive Care IV) version 2.2 is an open-access database that offers extensive clinical data from patients admitted to the Beth Israel Deaconess Medical Center spanning the years 2008 to 2019 [11, 12]. The database encompasses a range of clinical data including demographic details, vital signs, imaging studies, laboratory test outcomes, a comprehensive data dictionary, and documentation featuring International Classification of Diseases codes(ICD-9 and ICD-10). Additionally, it contains validated hourly physiological records from bedside monitors monitored by ICU nurses. As the health information from MIMIC-IV database was de-identified, patient consent was not required for its use. We have acquired Credentialing and Certification through training provided by PhysioNet in order to use the aforementioned databases(PhysioNet ID:12168208) [13].

Study patients

Patients diagnosed with SAB were included based on the International Classification of Diseases, ninth Revision (ICD-9), or Tenth Revision (ICD-10) codes. Additionally, patients meeting any of the following three criteria will be excluded: (1): age < 18 years old. (2): not first admission to the hospital due to SAB. (3): died in hospital.

Recurrent SAB was defined as readmission with positive staphylococcus aureus blood cultures at least 14 days after the first discharge [14].

Data extraction

Using the official code and raw data, we placed it within pgAdmin 4 (version 7.1). SQL (Structured Query Language) is the standard language for interacting with relational databases like PostgreSQL, which is commonly managed by pgAdmin 4. We utilized SQL to extract patient data from pgAdmin 4, including: (1)

Demographic characteristics: gender, age, weight on admission. (2) Complications and procedures: hypertension, diabetes, choronic kidney disease, hemodialysis, and valve replacement. (3) The blood routine and the biochemical indicators. (4) pathogen, vancomycin's duration hours and concentration. (5) Some inflammation and nutrition related indexs, Here are the calculation formulas:

- NLR (Neutrophil-to-Lymphocyte Ratio) = (Neutrophil Count) / (Lymphocyte Count).
- PLR (Platelet-to-Lymphocyte Ratio) = (Platelet Count) / (Lymphocyte Count).
- MLR (Monocyte-to-Lymphocyte Ratio) = (Monocyte Count) / (Lymphocyte Count).
- PNI (Prognostic Nutritional Index) = Albumin + (5 * Lymphocyte Count).
- SII (Systemic Immune-Inflammatory Index) = (Platelet Count * Neutrophil Count) / (Lymphocyte Count).
- SIRI (Systemic Immune-Inflammation Index) = (Neutrophil Count * Monocyte Count) / (Lymphocyte Count).

Data processing and statistical analyses

The missing values are imputed using chained equations for multiple imputation, implemented through the 'mice' package by R(R version 4.2.1) [15]. Using the 'compareGroups' R package, we conducted a Shapiro-Wilk test for normality. Continuous variables with a normal distribution are presented as mean (SD, standard deviation) and compared using an independent samples t-test. Non-normally distributed variables are presented as median (25%, 75%) and compared using the Kruskal-Wallis test. Categorical variables are described as percentages and compared using the chi-square test [16]. The patients were randomly divided into training and test groups in a 7:3 ratio, Variables are displayed and compared in Table 1.

Feature selection

We included a total of 45 features. Considering the imbalanced nature of the dataset, we employed the Synthetic Minority Oversampling Technique (SMOTE) to preprocess the training set by 'DMwR' package [17], aims to achieve more accurate prediction and improve model performance. We utilized Recursive Feature Elimination (RFE) to iteratively rank and eliminate the least relevant features from the complete set, selecting the most important ones based on specific criteria. Through cross-validation, we identified 19 features that exhibited the highest accuracy(Fig. 1A). Subsequently, we used 10-fold cross-validation to determine the tuning parameter (λ) in the LASSO model, plotting a partial likelihood deviance curve against log(λ). Employing both the minimum criterion and one standard error (1-SE criterion), we identified the min λ value as 0.034 based on the 10-fold cross-validation. This criterion led to the selection of 16 features(Fig. 1B). To refine the feature set, we took the intersection of these features, resulting in the identification of 11 key features: MRSA, PTT, RBC, RDW, Neutrophils_abs, Sodium, Calcium, Vancomycin_concentration, MCHC, MCV, PNI.

Constructing and validating models

The ML algorithms incorporated in this study encompass Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), eXtreme Gradient Boosting (XGBoost), and Artificial Neural Network (ANN). To prevent overfitting and enhance model accuracy, hyperparameter optimization and 10-fold cross-validation were executed using GridSearchCV. The R packages utilized include 'caret', 'randomForest', 'e1071', 'xgboost', 'nnet', 'rms', among others [18–22]. The predictive performance of the five models was evaluated using ROC curves, Decision Curve Analysis (DCA) and Precision-Recall (PR) Curve. SHAP (SHapley Additive exPlanations) is a game theoretic approach to explain the output of any ML model, It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions [23]. We applied the XGBoost model to the training dataset that included only the selected features. Subsequently, we utilized SHAP to display the importance of these selected features and to illustrate how each feature influences the predicted values for individual patients. All process was shown in Flowchart. Here are some parameters to construct models and the SHAP calculating formula:

Models parameters:

- RF: nTree = 5500;
- SVM: kernel = radial, gamma = 0.1, cost = 10;
- XGBoost: Iteration = 548, nrounds = 500, verbose = 1;
- ANN: size = 15, decay = 0.01, maxit = 2000;
- LR: default parameters.

The formula for calculating SHAP values for a feature *jj* in a model can be expressed as follows:

- ϕj(f)=∑S⊆N∖{j}|S|!(|N|−|S|−1)!|N|![f(S∪{j}) − f(S)]ϕj(f) = S⊆N∖{j}∑|N|!|S|!(|N|−|S|−1)![f(S∪{j}) − f(S)]

Where:

- ϕj(f)ϕj(*f*) is the SHAP value for feature *jj*.
- *ff* is the model prediction function.
- N*N* is the set of all features.
- S*S* is a subset of features excluding feature *jj*.
- |S||*S*| is the number of elements in set S*S*.

Table 1 Aseline characteristics of the patients

	no-recurrent	**recurrent**	***p*.value**
	***N*=711**	***N*=101**	
Age	60.9 [46.8;71.8]	64.1 [51.8;75.1]	0.137
Gender:			1.000
Female	272 (38.3%)	39 (38.6%)	
Male	439 (61.7%)	62 (61.4%)	
Weight	80.0 [67.4;95.0]	77.2 [67.0;93.8]	0.652
Hypertension:			1.000
No	484 (68.1%)	69 (68.3%)	
Yes	227 (31.9%)	32 (31.7%)	
Diabetes:			0.852
No	468 (65.8%)	68 (67.3%)	
Yes	243 (34.2%)	33 (32.7%)	
CKD:			0.443
No	509 (71.6%)	68 (67.3%)	
Yes	202 (28.4%)	33 (32.7%)	
Hemodialysis:			0.048
No	636 (89.5%)	83 (82.2%)	
Yes	75 (10.5%)	18 (17.8%)	
Valvereplace:			0.017
No	695 (97.7%)	94 (93.1%)	
Yes	16 (2.25%)	7 (6.93%)	
Hematocrit	29.2 [26.3;33.1]	27.4 [25.7;29.1]	<0.001
Hemoglobin	9.50 [8.55;10.9]	8.80 [8.30;9.60]	<0.001
MCH	29.6 [28.0;31.0]	29.8 [28.6;31.2]	0.234
MCHC	32.7 (1.43)	32.2 (1.18)	0.001
MCV	90.2 [85.8;94.2]	91.8 [88.1;97.3]	0.001
Platelet	250 [173;338]	238 [158;351]	0.365
RBC	3.30 [2.90;3.70]	3.00 [2.80;3.30]	<0.001
RDW	15.2 [14.0;16.5]	16.3 [15.3;17.6]	<0.001
WBC	9.50 [7.05;12.1]	10.8 [7.80;13.1]	0.003
Neutrophils_abs	8.00 [5.20;11.4]	10.9 [6.90;13.0]	<0.001
Lymphocytes_abs	1.10 [0.70;1.60]	1.00 [0.70;1.40]	0.474
Monocytes_abs	0.60 [0.40;0.80]	0.60 [0.40;0.90]	0.657
Albumin	29.2 [25.5;33.0]	27.1 [23.7;30.5]	0.001
Aniongap	13.4 [12.2;15.0]	13.5 [12.2;16.1]	0.410
Bicarbonate	25.4 [23.4;27.1]	25.0 [23.4;26.7]	0.530
BUN	6.30 [4.30;11.9]	9.50 [5.30;16.2]	0.001
Calcium	2.10 [2.00;2.20]	2.00 [2.00;2.10]	<0.001
Chloride	101 [98.2;104]	102 [98.7;105]	0.280
Creatinine	88.4 [63.8;155]	109 [66.9;199]	0.095
Glucose	6.50 [5.70;7.90]	6.60 [6.00;8.20]	0.071
Sodium	137 (3.51)	138 (3.32)	0.021
Potassium	4.10 [3.80;4.30]	4.00 [3.80;4.20]	0.258
INR	1.30 [1.10;1.50]	1.50 [1.20;1.80]	<0.001
PT	14.1 [12.7;16.5]	16.3 [13.6;19.7]	<0.001
PTT	31.6 [28.2;39.8]	38.9 [30.9;53.0]	<0.001
ALT	26.0 [15.6;45.2]	23.5 [15.0;40.8]	0.228
AST	31.8 [20.4;49.9]	34.2 [23.0;50.7]	0.368
Bilirubin_total	0.50 [0.30;0.90]	0.70 [0.50;1.60]	<0.001
Pathogen:			<0.001
SA	685 (96.3%)	80 (79.2%)	
MRSA	26 (3.66%)	21 (20.8%)	
Total_vancomycin_hours	91.0 [43.0;220]	178 [86.0;334]	<0.001

Table 1 (continued)

	no-recurrent ***N*=711**	**recurrent** ***N*=101**	***p*.value**
Vancomycin_concentration	15.4 [7.00;19.6]	18.6 [15.5;20.7]	<0.001
NLR	7.20 [4.30;11.7]	9.90 [6.90;14.2]	<0.001
PLR	235 [152;355]	252 [144;363]	0.932
MLR	0.50 [0.30;0.80]	0.50 [0.40;0.90]	0.454
PNI	35.3 [30.9;39.5]	32.1 [28.9;36.5]	<0.001
SII	1861 [896;3169]	2156 [1333;4084]	0.018
SIRI	4.20 [2.10;7.90]	5.60 [3.60;10.0]	0.002

Categorical data were showed as frequency (percentage). Continuous variables with normal distributions were presented as the mean (SD, standard deviation) and compared with independent samples t tests. Non-normally distributed variables are expressed as the median(interquartile ranges). Abbreviation: CKD (Chronic Kidney Disease), MCH (Mean Corpuscular Hemoglobin), MCHC (Mean Corpuscular Hemoglobin Concentration), MCV (Mean Corpuscular Volume), RBC (Red Blood Cell Count), RDW (Red Cell Distribution Width), WBC (White Blood Cell Count), BUN (Blood Urea Nitrogen), INR (International Normalized Ratio), PT (Prothrombin Time), PTT (Partial Thromboplastin Time), ALT (Alanine Aminotransferase), AST (Aspartate Aminotransferase), SA (Staphylococcus aureus), MRSA (Methicillin-Resistant Staphylococcus aureus), NLR (Neutrophil-to-Lymphocyte Ratio), PLR (Platelet-to-Lymphocyte Ratio), MLR (Monocyte-to-Lymphocyte Ratio), PNI (Prognostic Nutritional Index), SII (Systemic Immune-Inflammatory Index), SIRI (Systemic Immune-Inflammation Index).

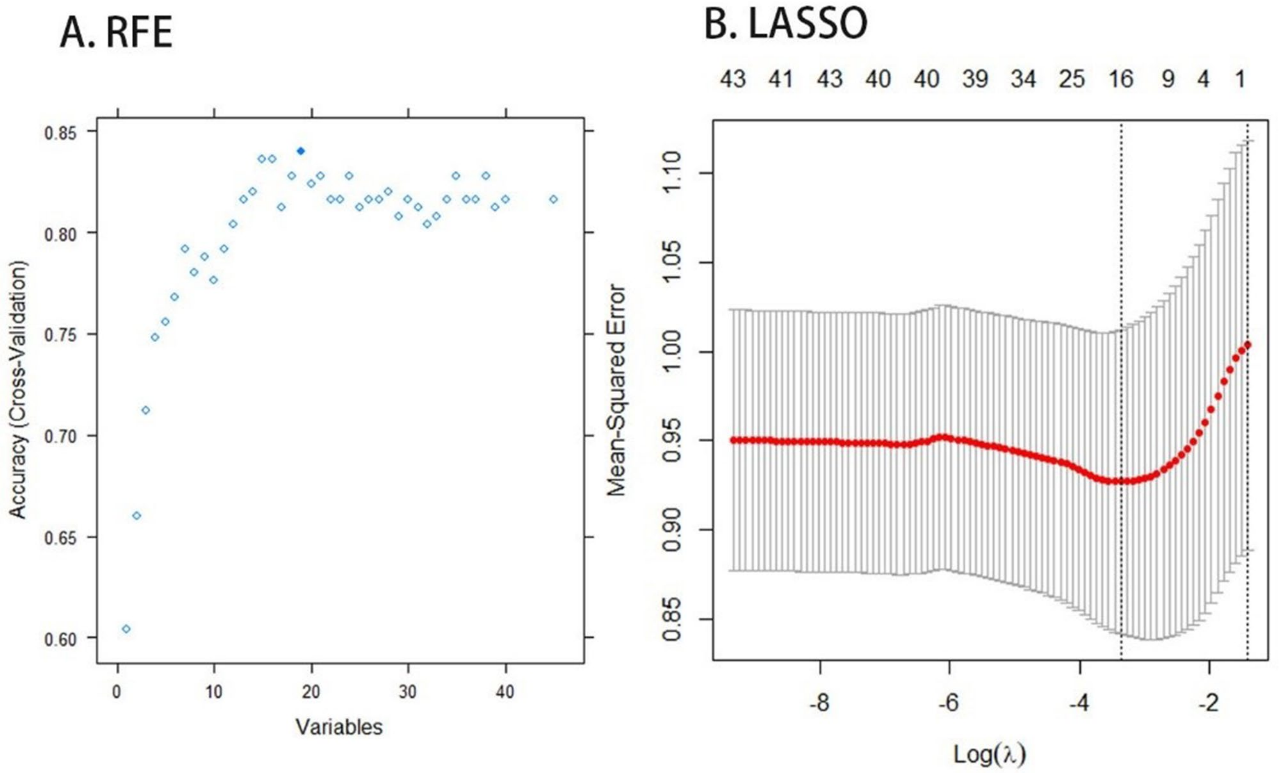


Fig. 1 **A**. Illustrates a scatter plot depicting the relationship between the number of variables and accuracy in the RFE method. **B**. Dotted vertical lines are utilized to represent the minimum criteria (lambda.min) and one standard error of the minimum criteria (lambda.1se) as optimal values

- $|\mathrm{N}||N|$ is the total number of features.

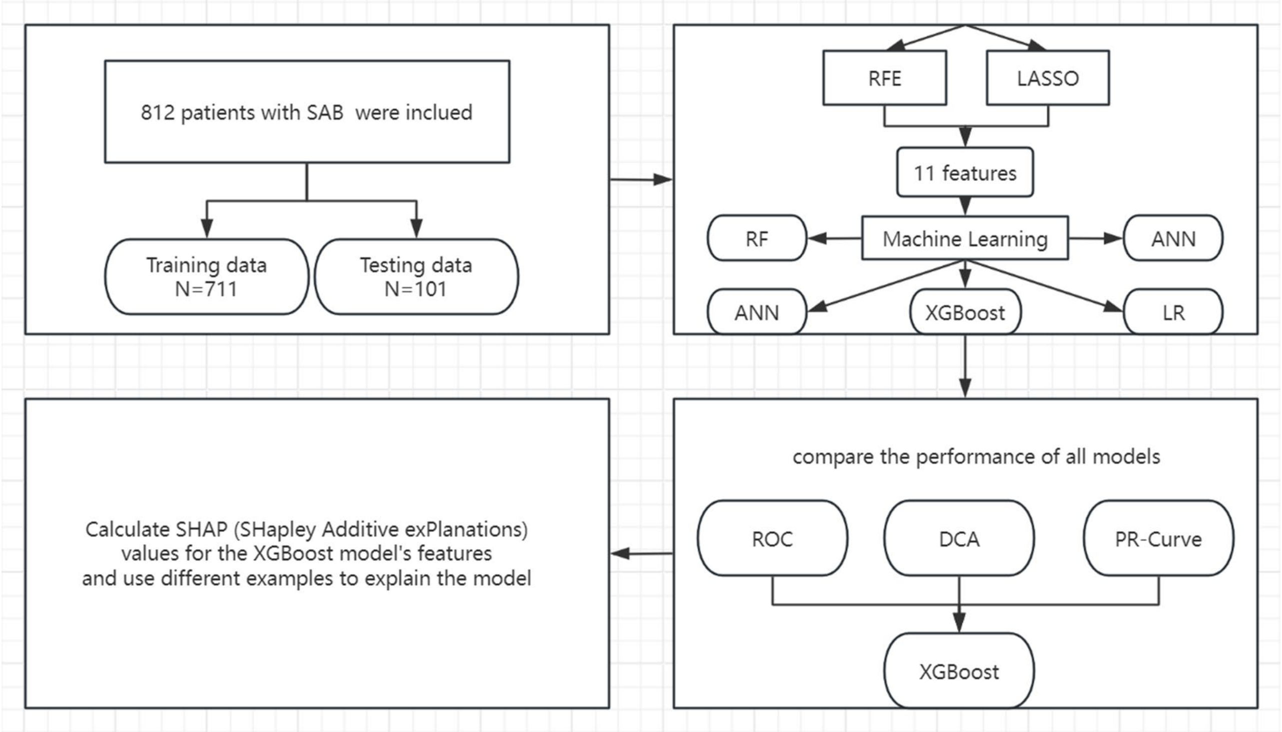
Explanation of the Components:

- The summation is over all subsets SS of features excluding the feature jj.
- The term $\mathrm{f(S{\cup}\{j\})}f(S\cup\{j\})$ represents the model prediction with feature jj included.
- The term $\mathrm{f(S)}f(S)$ represents the model prediction without feature jj.
- The fraction $\mathrm{|S|!(|N|-|S|-1)!|N|!}|N|!|S|!(|N|-|S|-1)!$ is a weighting term that accounts for the different ways of ordering the features.

Results

Characteristics

Among patients infected with Staphylococcus aureus, there are 711 patients without recurrent SAB. and 101 patients with recurrent SAB upon hospital readmission. Their baseline characteristics are shown in Table 1.


812 patients with SAB were inclued
Training data N=711
Testing data N=101
RFE
LASSO
11 features
Machine Learning
RF
ANN
ANN
XGBoost
LR
compare the performance of all models
ROC
DCA
PR-Curve
XGBoost
Calculate SHAP (SHapley Additive exPlanations) values for the XGBoost model's features and use different examples to explain the model

Flowchart 1 Flowchart of the entire process

There were no statistically significant differences in baseline characteristics and medical history between the two groups of patients. Patients undergoing valve replacement surgery and hemodialysis during hospitalization are more likely to readmission for SAB (*P*<0.05). Complete blood count (CBC) is a vital examination in clinical practice, RBC, RDW, MCHC, MCV, and neutrophils_abs exhibit notable difference between the two groups (*P*<0.001). There are statistically significant differences in albumin, bun, calcium, and sodium within the biochemistry test between the two groups (*P*<0.05). INR, PT, and PTT, which are indicators related to coagulation, exhibit marked differences between the two groups (*P*<0.001). The pathogen, duration of vancomycin administration in hours, and concentration are also crucial features for readmitted patients(*P*<0.001). Additionally, calculated indices from blood and biochemical tests, including NLR, PNI, SII, and SIR, are also crucial for identifying high-risk patients for readmission(*P*<0.05).

Model comparison

The AUROC values obtained for the five ML models (XGBoost, ANN, LR, RF, and SVM) in the testing cohort are presented in Fig. 2.**A**: XGBoost (0.76, 95% CI: 0.66–0.85), ANN (0.70, 95% CI: 0.62–0.79), LR (0.72, 95% CI: 0.61–0.82), RF (0.75, 95% CI: 0.65–0.84), and SVM (0.70, 95% CI: 0.61–0.80). The XGBoost model demonstrated superior performance compared to the others based on AUC values. Moreover, the DCA plot (Fig. 2.B) revealed that the XGBoost model exhibited a stable and relatively high standardized net benefit across various thresholds compared to the other models. In the precision-recall curve, the XGBoost model also show the best PRAUC value 0.56(95%CI: 0.37–0.75) in Fig. 2.

Shap value

The XGBoost model demonstrated outstanding predictive performance. We calculated SHAP values for the model's features. Figure 3.**A** showed the features importance of the XGBoost model based on SHAP value. The SHAP summary plot illustrates the impact of each feature on the output values of the model in Fig. 3.**B.** PTT, Neutrophils_abs, RDW were identified as the three most critical features influencing the outcomes of this model. These factors demonstrated the highest impact, emphasizing their pivotal roles in the XGBoost model. Figure 4 displayed the SHAP force plot for the patients in the testing group, illustrating how different features influence the XGBoost model in each patient.

Model application

To make our findings accessible to clinical practitioners, researchers, patients, and their families, we have created a prognostic prediction system. This system can be accessed on the following website: https://bifeinitong.shinyapps.io/sa_recurrent/.

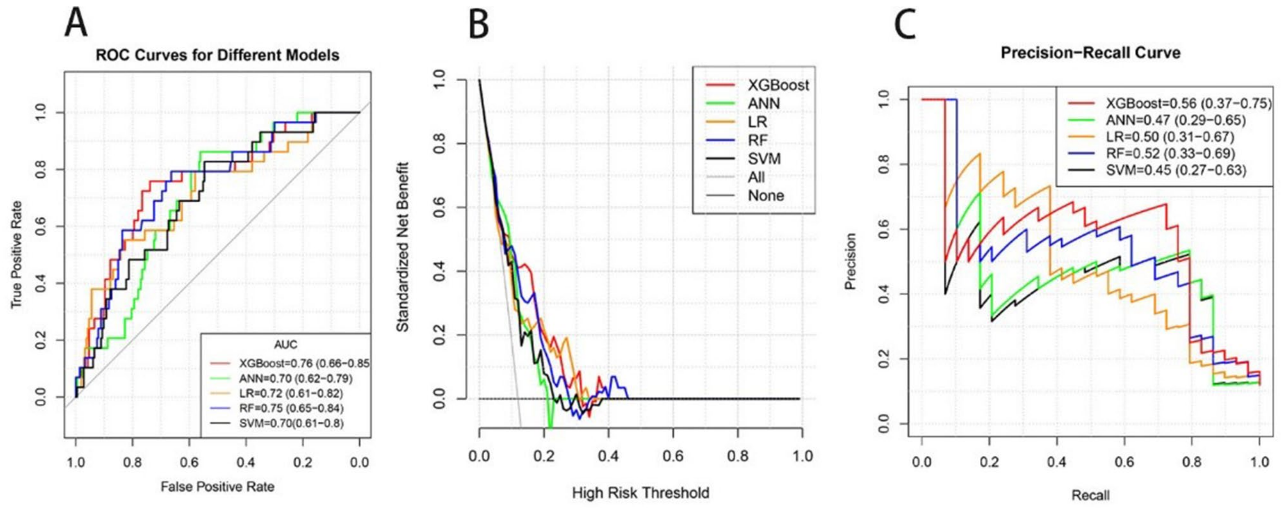


Fig. 2 **A**: The ROC curves comparison of the five models in testing cohort. **B**: The DCA curve comparison of the five models in testing cohort. **C**: The Precision-Recall curve comparison of the five models in testing cohort. Red line=XGBoost model, green line=ANN model, darkorange line=LR model, blue line=RF model, black line=SVM model. The models' auc and 95% confidence interval show in the legend

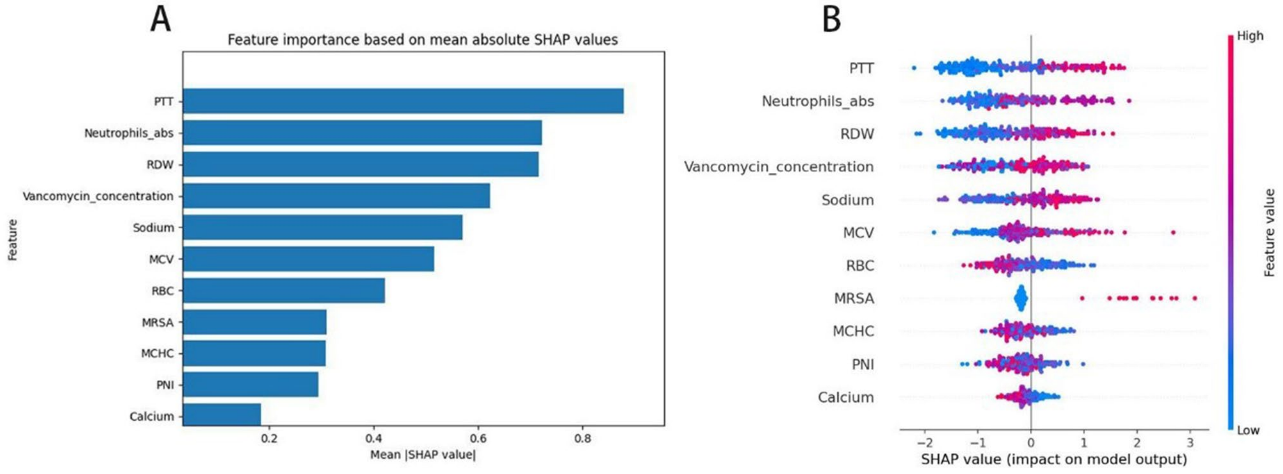


Fig. 3 **A**: feature importance based on shap values. **B**: dot is created for each feature attribution value for the model of each patient, and thus one patient is allocated one dot on the line for each feature. Dots are colored according to the values of features for the respective patient and accumulate vertically to depict density. Red represents a high feature value (in this case death), whereas blue represents a low feature value. The further away a point is from the baseline SHAP value of zero, the stronger it effects the output

Discussion

In this study, we collected data from the MIMIC-IV database to analyze and compare patients with SAB and recurrent SAB. Then, we employed RFE and LASSO methods to select 11 essential features. Subsequently, we constructed predictive models using five different ML approaches, and their predictive capabilities were compared through AUC and DCA. Furthermore, we analyzed the SHAP values of the model to demonstrate the importance of each feature in influencing the output results. Using different examples, we demonstrate how these features influence the performance of the XGBoost model. In addition, we have developed a website that enables the classification of patients with SAB. This platform facilitates the identification of patients with a high predicted incidence of recurrent SAB, allowing for more optimized clinical treatment. As a result, it aims to reduce the likelihood of hospital readmission for SAB.

It is generally believed that Staphylococcus aureus(SA) is a commensal bacterium, with over 30% of individuals having a permanent colonization of this bacterium. However, it is also a common cause of bacterial infections in humans [24]. Skin infection is the most common way for Staphylococcus aureus to invade the human body. When it occurs in individuals with weakened immune systems, such as those who are bedridden for extended periods or patients in the intensive care unit (ICU), it can spread to the circulatory system through the skin-blood

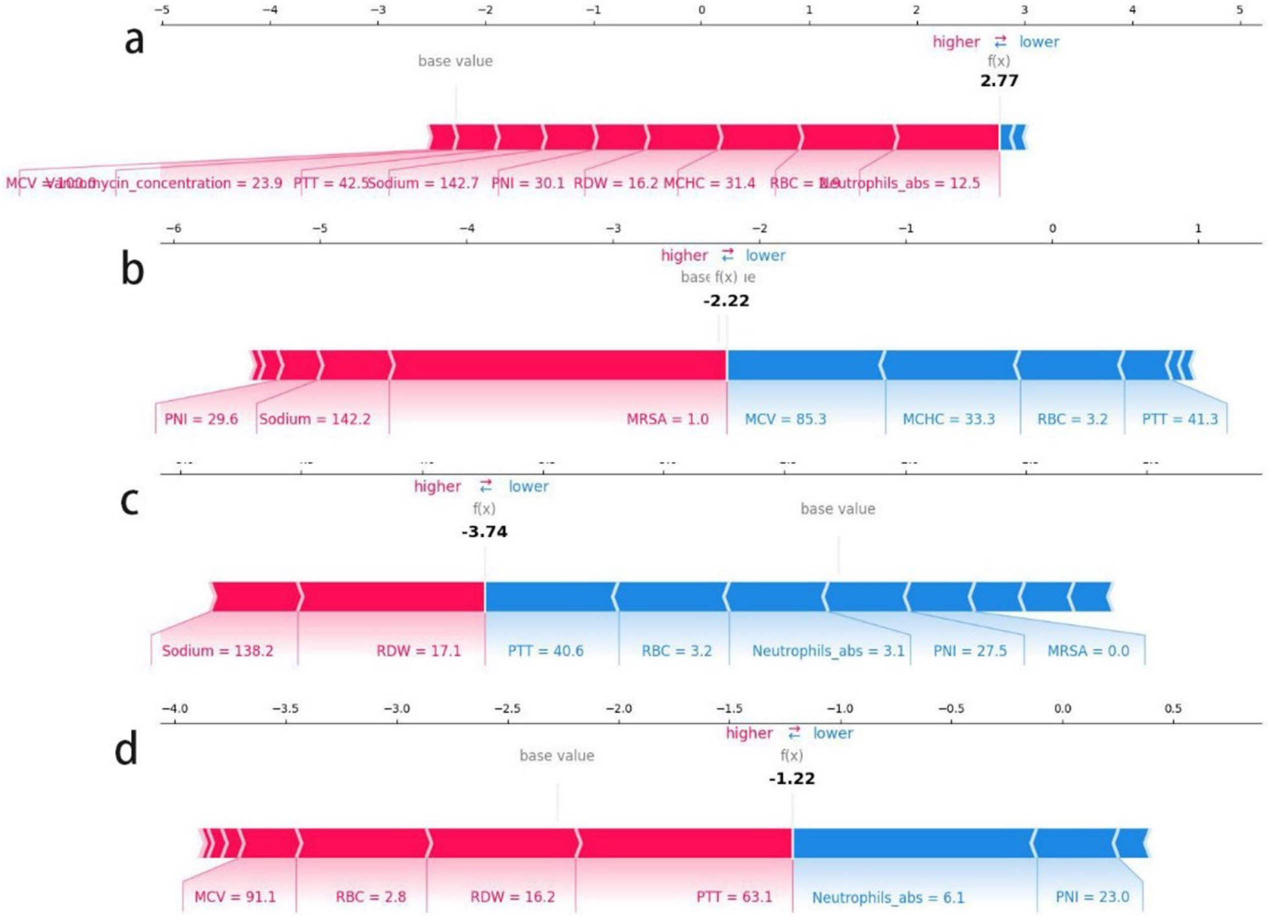


Fig. 4 Depicts the SHAP force plots generated by randomly selecting patients. In the figure, red arrows represent a positive impact on the predicted value, while blue arrows represent a negative impact on the predicted value. The larger the area of the arrow, the greater the influence of that feature. The base value in the figure represents the model's default prediction when no feature input is provided, typically the average output value across all training data. In this case, the base value is -2.28. The prediction value f(x) in the figure represents the final predicted output value of the model for the given sample

circulation route, leading to SAB [25]. Additionally, catheter-related bloodstream infections caused by SA are also a significant pathway for SAB. MRSA is a type of antibiotic-resistant bacterium, According to the antibiotic resistance threat report released by the CDC, there were 323,700 cases of MRSA infections in 2017, resulting in 10,600 patient deaths. This poses a significant burden on healthcare, emphasizing the substantial impact of MRSA [26]. Despite the decreasing trend in MRSA infections between 2012 and 2017, there appears to be a resurgence of MRSA, especially among hospitalized patients, since the outbreak of COVID-19. The proportion of MRSA infections among hospitalized patients has shown a noticeable increase [27]. Due to the limited effectiveness of many antibiotics against MRSA infections, it is often associated with high hospital mortality rates and recurrence rates [28]. To investigate the true recurrence rate of SAB, Choi et al. analyzed a single-center cohort of 756 cases. Out of the total, 69 patients (9%) experienced new episodes of SAB at least 14 days after the last positive blood culture for SAB, and the methicillin resistance shows the difference between the two groups($P = 0.04$) [14]. Our research findings are consistent with these results, providing further reason to believe that MRSA is one of the major causes of recurrent SAB.

Our study found significant differences in INR, PT, and PTT between recurrent SAB patients and those with SAB ($P < 0.001$). SHAP analysis revealed that PTT is the most influential feature in the XGBoost model. INR, PT, PTT are coagulation-related indicators used to assess a patient's coagulation function. SA produces coagulases, including Staphylocoagulase and von Willebrand factor-binding protein (vWf-binding protein). It directly engages host platelets. Additionally, SA utilizes the bacterial iron-regulated surface determinant cluster B (IsdB) to modulate its binding to glycoprotein (GP) IIb/IIIa on the platelet surface. It can also indirectly bind to GP IIb/IIIa through microbial surface component recognizing adhesive matrix molecule (MSCRAMM) on the bacterial surface, triggering a host coagulation-inflammatory response. This mechanism allows the pathogen to bypass host regulatory systems and hijack the coagulation cascade [29, 30]. Neutrophil extracellular traps (NETs) represent a critical immune mechanism for the host to counteract invading pathogens. The formation of NETs is closely linked to platelet function and the coagulation cascade. SA can evade the surveillance and clearance by NETs through interference with the coagulation system. Furthermore, studies have reported that SA can induce the formation of NETs, reside within them, and subsequently establish biofilms, thus evading the attacks from antibiotics and other immune pathways [31, 32]. It is plausible to believe that in patients with unstable coagulation functions, SA can temporarily evade the initial course of antibiotic treatment through this mechanism. When antibiotics are discontinued, the residual SA can gradually mature and disseminate, leading to a recurrent episode of SAB.

In our study, patients with higher RDW exhibited a greater likelihood of developing recurrent SAB. The RDW level reflects the heterogeneity in the size of red blood cells and indicates the body's response to oxidative stress and inflammation. It is often associated with bacterial infections. Kim et al. found that baseline RDW values and an increase in RDW can serve as a promising independent prognostic marker in patients with severe sepsis or septic shock [33]. Elevated RDW levels may indicate weakened immune function in patients, necessitating the production of a substantial amount of inflammatory oxidative factors to clear SA. It may also reflect the heightened virulence of the pathogen.

Vancomycin concentration reflects the Minimum Inhibitory Concentration (MIC) value of SA against vancomycin. High vancomycin MIC (>1.5 μg/mL) was the only independent risk factor for development of complicated bacteremia caused by methicillin-susceptible S. aureus [34]. Notably, a high vancomycin MIC is not necessarily related to therapeutic outcomes. Holmes et al. assessed 532 patients with SAB from 8 hospitals. All MRSA bacteremia patients received vancomycin, while MSSA bacteremia patients were treated with either flucloxacillin or vancomycin. An increase in vancomycin MIC was associated with an increased mortality rate in patients treated with vancomycin. However, even in MSSA bacteremia patients receiving flucloxacillin treatment, those with isolates having a vancomycin MIC of 1.5 μg/mL had a higher mortality rate compared to isolates with lower MIC (26.8% vs. 12.2%; $P<0.001$) [35]. However, a meta-analysis incorporating 13 studies involving 2089 patients found no significant differences between high and low vancomycin MIC groups in overall mortality, in-hospital mortality, late mortality, persistent bacteremia, severe sepsis or septic shock. It is essential to conduct randomized controlled trials to assess the utility of vancomycin MIC values in predicting mortality and recurrent rate.

Additionally, through comparative analysis using AUC, DCA, and PR-Curve, it was found that the XGBoost model outperformed SVM, LR, RF, and ANN in predicting the recurrence of SAB in patients. XGBoost is an efficient and scalable ML algorithm used for solving supervised learning problems such as classification, regression, and ranking. It is based on boosting algorithms, utilizing decision trees as fundamental learning models, and employs regularization techniques to control model complexity. XGBoost exhibits high flexibility and scalability, capable of handling large-scale datasets to construct precise prediction models [36]. Through the comparison of AUROC, Rahmani et al. found that XGBoost demonstrated an impressive AUROC of 0.762 for predicting the risk of Central Line-Associated Bloodstream Infection (CLABSI) 48 h after central line placement [37]. The application of XGBoost as a clinical prediction model in the medical field has been increasingly recognized and validated.

The recurrence of SAB is a complex and dangerous issue, significantly increasing patient mortality and leading to a series of severe complications. Recurrence often indicates that the initial treatment failed to completely eradicate the pathogens. Inadequate antibiotic duration or dosage is a primary cause. Additionally, undetected or incompletely treated infection foci, such as abscesses or infected medical devices, can continuously release bacteria into the bloodstream, resulting in recurrent infections [38]. Recurrent SAB can lead to severe complications such as multiple embolisms, acute kidney injury (AKI), and cardiovascular diseases, posing life-threatening risks and long-term health issues [39]. These conditions severely affect patient quality of life. Furthermore, recurrent infections typically require prolonged and more potent antibiotic treatments, which come with increased side effects and drug toxicity. From a societal and healthcare system perspective, recurrent SAB increases hospitalization duration and medical costs, exacerbating the burden on healthcare resources. This situation also imposes significant financial and psychological stress on patients and their families. Therefore, developing a predictive model for SAB recurrence is crucial. This model can assist in formulating personalized treatment plans and enable early preventive interventions, thus reducing recurrence rates and associated complications. Firstly, the predictive model can help clinicians identify high-risk patients. After initial treatment, these high-risk patients can undergo closer monitoring and follow-up, allowing for timely detection and management of infections at early stages. Early intervention helps reduce severe complications, improving patient survival rates and quality of life. Secondly, the predictive model allow for the optimization of antibiotic usage strategies. Existing treatment regimens may be ineffective for some high-risk patients, but the predictive model can provide data support to help clinicians select more appropriate antibiotics and treatment durations. This approach not only enhances treatment efficacy but also minimizes unnecessary antibiotic use, thereby reducing the risk of antibiotic resistance. Finally, by reducing recurrence rates, shortening hospital stays, and decreasing readmission rates, the predictive model effectively alleviate the burden on healthcare resources.

In summary, our developed predictive model for recurrent SAB holds significant clinical management implications and provides robust support for public health and healthcare resource optimization. Through multidisciplinary collaboration and data integration, this innovative tool offers new solutions for addressing SAB recurrence. However, our study has certain limitations. Firstly, further collection of effective inflammatory markers could enhance model performance. Secondly, additional comparisons using more deep learning models are warranted. Lastly, the model has not been validated with external databases, necessitating further clarification of its applicability to external data.

Conclusions

In conclusion, This study illustrates the application of ML utilizing clinical information and laboratory parameters in predicting SAB recurrence. We developed an XGBoost algorithm model that accurately predicts the likelihood of SAB recurrence. By integrating ML with the SHAP explainability method, our model transitions from a “black box” to an interpretable tool, making it more applicable for clinical scenarios in predicting SAB recurrence. Moreover, the inclusion of ROC, DCA, and PR curves underscores the clinical utility of the XGBoost model. Our goal is to assist clinicians in the early identification of high-risk individuals susceptible to recurrent SAB through our website, thereby facilitating the creation of personalized preventive strategies to improve SAB patient outcomes.

Acknowledgements

We thank all participants in the Nanjing First Hospital and Nanjing Medical University.

Author contributions

YL designed the study and drafted the manuscript; SS and Ly Z extracted the data; Xr Z and Yj M conducted data quality management and statistical analysis; Mx L developed the website; Wj W and ZT critically revised the manuscript. All authors contributed to the article and approved the submitted version. All authors take responsibility for all aspects of the reliability and freedom from bias of the data presented and their discussed interpretation.

Funding

Nanjing Medical Science and Technique Development Foundation, NMUB20220080.

The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Data availability

The datasets are available in the PhysioNet (https://physionet.org/content/mimiciv/2.2/).

Declarations

Ethics approval and consent to participate

MIMIC-IV database is publicly available anonymized database, approval for the ethical committee are not necessary.

Consent for publication

Not appicable.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used chatgpt3.5 in order to verify code and translate. After using this tool, the authors reviewed and edited the content as needed and takes full responsibility for the content of the publication.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Infectious Disease, Nanjing First Hospital, Nanjing Medical University, Nan jing 210006, China
[2]Nanjing Medical University, Nanjing, China

Received: 10 April 2024 / Accepted: 17 January 2025
Published online: 24 February 2025

References

1. Tong SYC, Davis JS, Eichenberger E, Holland TL, Fowler VG. Staphylococcus aureus infections: epidemiology, pathophysiology, clinical manifestations, and management. Clin Microbiol Rev. 2015;28:603–61.
2. Nambiar K, Seifert H, Rieg S, et al. Survival following Staphylococcus aureus bloodstream infection: a prospective multinational cohort study assessing the impact of place of care. J Infect. 2018;77:516–25.
3. Weiner-Lastinger LM, Pattabiraman V, Konnor RY, Patel PR, Wong E, Xu SY, Smith B, Edwards JR, Dudeck MA. The impact of coronavirus disease 2019 (COVID-19) on healthcare-associated infections in 2020: a summary of data reported to the National Healthcare Safety Network. Infect Control Hosp Epidemiol. 2022;43:12–25.
4. Bergenman O, Nilson B, Rasmussen M. Risk of infective endocarditis and complicated infection in Staphylococcus aureus bacteremia– a retrospective cohort study on the role of bacteriuria. Eur J Clin Microbiol Infect Dis. 2024;43:1419–26.
5. Buyrukoğlu G. Survival analysis in breast cancer: evaluating ensemble learning techniques for prediction. PeerJ Comput Sci. 2024;10:e2147.
6. Buyrukoğlu S. New hybrid data mining model for prediction of *Salmonella* presence in agricultural waters based on ensemble feature selection and machine learning algorithms. J Food Saf. 2021;41:e12903.
7. Buyrukoglu S. (2021) Promising Cryptocurrency Analysis using Deep Learning. In: 2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). IEEE, Ankara, Turkey, pp 372–376.
8. Handelman GS, Kok HK, Chandra RV, Razavi AH, Lee MJ, Asadi H. eDoctor: machine learning and the future of medicine. J Intern Med. 2018;284:603–19.
9. Gong J, Zhang Y, Zhong X, Zhang Y, Chen Y, Wang H. Liver function test indices-based prediction model for post-stroke depression: a multicenter, retrospective study. BMC Med Inf Decis Mak. 2023;23:127.
10. Zuo D, Yang L, Jin Y, Qi H, Liu Y, Ren L. Machine learning-based models for the prediction of breast cancer recurrence risk. BMC Med Inf Decis Mak. 2023;23:276.

11. Johnson A, Bulgarelli L, Pollard. (2023) MIMIC-IV (version 2.2). PhysioNet. https://doi.org/10.13026/6mm1-ek67
12. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023;10:1.
13. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation. 2000;101:E215–220.
14. Choi S-H, Dagher M, Ruffin F, et al. Risk factors for recurrent *Staphylococcus aureus* Bacteremia. Clin Infect Dis. 2021;72:1891–9.
15. Buuren SV, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in *R*. J Stat Soft. 2011. https://doi.org/10.18637/jss.v045.i03.
16. Subirana I, Sanz H, Vila J. Building Bivariate tables: the compareGroups Package for R. J Stat Soft. 2014. https://doi.org/10.18637/jss.v057.i12.
17. Torgo L. Data mining with R: learning with case studies. Boca Raton, Fla: CRC Press, Taylor & Francis; 2011.
18. Jr FEH (2023) rms: Regression Modeling Strategies.
19. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. (2023) e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.
20. Venables WN, Ripley BD. Modern Applied statistics with S, Fourth. New York: Springer; 2002.
21. Chen T, He T, Benesty M et al. (2023) xgboost: Extreme Gradient Boosting.
22. Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002;2:18–22.
23. Lundberg S, Lee S-I. (2017) A Unified Approach to Interpreting Model Predictions.
24. Wertheim HFL, Melles DC, Vos MC, van Leeuwen W, van Belkum A, Verbrugh HA, Nouwen JL. The role of nasal carriage in Staphylococcus aureus infections. Lancet Infect Dis. 2005;5:751–62.
25. Kwiecinski JM, Horswill AR. Staphylococcus aureus bloodstream infections: pathogenesis and regulatory mechanisms. Curr Opin Microbiol. 2020;53:51–60.
26. CDC's. Antibiotic Resistance Threats in the United States.
27. (2022) COVID-19: U.S. Impact on Antimicrobial Resistance, Special Report 2022. https://doi.org/10.15620/cdc:117915
28. Lakhundi S, Zhang K. Methicillin-Resistant Staphylococcus aureus: molecular characterization, evolution, and Epidemiology. Clin Microbiol Rev. 2018;31:e00020–18.
29. Leeten K, Jacques N, Lancellotti P, Oury C. Aspirin or Ticagrelor in Staphylococcus aureus Infective endocarditis: where do we stand? Front Cell Dev Biol. 2021;9:716302.
30. Liesenborghs L, Verhamme P, Vanassche T. Staphylococcus aureus, master manipulator of the human hemostatic system. J Thromb Haemost. 2018;16:441–54.
31. Meyers S, Crescente M, Verhamme P, Martinod K. *Staphylococcus aureus* and Neutrophil Extraceltraps Tthes: The Master Manipumeets itsts Its Match in Immunothrombosis. ATVB. 2022;42:261–76.
32. Speziale P, Pietrocola G. Staphylococcus aureus induces neutrophil extracellular traps (NETs) and neutralizes their bactericidal potential. Comput Struct Biotechnol J. 2021;19:3451–7.
33. Kim CH, Park JT, Kim EJ, et al. An increase in red blood cell distribution width from baseline predicts mortality in patients with severe sepsis or septic shock. Crit Care. 2013;17:R282.
34. Aguado JM, San-Juan R, Lalueza A, Sanz F, Rodríguez-Otero J, Gómez-Gonzalez C, Chaves F. High Vancomycin MIC and complicated methicillin-susceptible Staphylococcus aureus bacteremia. Emerg Infect Dis. 2011;17:1099–102.
35. Holmes NE, Turnidge JD, Munckhof WJ, et al. Antibiotic choice may not explain poorer outcomes in patients with Staphylococcus aureus bacteremia and high Vancomycin minimum inhibitory concentrations. J Infect Dis. 2011;204:340–7.
36. Chen T, Guestrin C. (2016) XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp 785–794.
37. Rahmani K, Garikipati A, Barnes G, Hoffman J, Calvert J, Mao Q, Das R. Early prediction of central line associated bloodstream infection using machine learning. Am J Infect Control. 2022;50:440–5.
38. Rodríguez AG, Llorach PD, Grillo S, et al. Risk factors for mortality and complications in peripheral venous catheter-associated Staphylococcus aureus bacteraemia: a large multicentre cohort study. J Hosp Infect. 2024;S0195–6701(24):00247–0.
39. López-Cortés LE, Gálvez-Acebal J, Rodríguez-Baño J. Therapy of Staphylococcus aureus bacteremia: Evidences and challenges. Enferm Infecc Microbiol Clin (Engl Ed). 2020;38:489–97.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.