BMC Medical Informatics and Decision Making

RESEARCH

Open Access

# Development and application of an early prediction model for risk of bloodstream infection based on real-world study

Xiefei Hu[1], Shenshen Zhi[1], Yang Li[2], Yuming Cheng[3], Haiping Fan[4], Haorong Li[5], Zihao Meng[5], Jiaxin Xie[4], Shu Tang[5*] and Wei Li[1*]

## Abstract

**Background** Bloodstream Infection (BSI) is a severe systemic infectious disease that can lead to sepsis and Multiple Organ Dysfunction Syndrome (MODS), resulting in high mortality rates and posing a major public health burden globally. Early identification of BSI is crucial for effective intervention, reducing mortality, and improving patient outcomes. However, existing diagnostic methods are flawed by low specificity, long detection times and high demands on testing platforms. The development of artificial intelligence provides a new approach for early disease identification. This study aims to explore the optimal combination of routine laboratory data and clinical monitoring indicators, and to utilize machine learning algorithms to construct an early, rapid, and universally applicable BSI risk prediction model, to assist in the early diagnosis of BSI in clinical practice.

**Methods** Clinical data of 2582 suspected BSI patients admitted to the Chongqing University Central Hospital, from January 1, 2021 to December 31, 2023 were collected for this study. The data were divided into a modeling dataset and an external validation dataset based on chronological order, while the modeling dataset was further divided into a training set and an internal validation set. The occurrence rate of BSI, distribution of pathogens, and microbial primary reporting time were analyzed within the training set. During the feature selection stage, univariate regression and ML algorithms were applied. First, Univariate logistic regression was used to screen for predictive factors of BSI. Then, the Boruta algorithm, Lasso regression, and Recursive Feature Elimination with Cross-validation (RFE-CV) were employed to determine the optimal combination of predictors for predicting BSI. Based on the optimal combination, six machine learning algorithms were used to construct an early BSI risk prediction model. The best model was selected by models' performance, and the Shapley Additive Explanations (SHAP) method was used to explain the model. The external validation set was used to evaluate the predictive performance and generalizability of the selected model, and the research findings were ultimately applied in clinical practice.

**Results** The incidence of BSI among inpatients at the Chongqing University Central Hospital was 12.91%. Following further feature selection, a set of 5 variables was determined, including white blood cell count, standard bicarbonate,

*Correspondence:
Shu Tang
tangshu@cqupt.edu.cn
Wei Li
liwei0111@cqu.edu.cn

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

base excess of extracellular fluid, interleukin-6, and body temperature. BSI early risk prediction models were constructed using six machine learning algorithms, with the XGBoost model demonstrating the best performance, achieving an AUC value of 0.782 in the internal validation set and an AUC value of 0.776 in the external validation set. This model is made publicly available as an online webpage tool for clinical use.

**Conclusions** This study successfully identified a set of 5 features by analyzing routine laboratory data clinical monitoring indicators among hospitalized patients. Based on this set, a machine learning-based early risk prediction model for BSI was constructed. The model is capable of early and rapid differentiation between BSI and non-BSI patients. The inclusion of minimal risk prediction factors enhances its applicability in clinical settings, particularly at the primary care level. To further improve the model's real-world applicability and more convenient for clinical use, the online application of the model could greatly improve the efficiency of BSI diagnosis and reducing patients' mortality.

**Keywords** Bloodstream infection, Risk prediction, Real-world, Model construction

## Introduction

Bloodstream Infection (BSI) refers to the invasion of microorganisms into the bloodstream, leading to a systemic infection, which can cause damage to all organs of the body. It is prone to inducing sepsis and multiple organ dysfunction syndrome (MODS), and associated with a high mortality rate [1, 2]. Annually, approximately 1.2 million patients are diagnosed with BSI in Europe [3]. Every hour of delayed treatment for BSI raises the mortality rate by 8%, reaching 58% after a 6-hour delay [4]. In the ICU, factors like weakened immunity, frequent risky procedures, multiple complications, and extended hospital stays heighten the risk of BSI, making it a common issue in these settings. Untreated BSI can rapidly lead to sepsis, progressing to MODS, causing poor outcomes and life-threatening conditions [5–7]. Early detection, appropriate antibiotics, and addressing the source of BSI greatly reduce morbidity and mortality rates [8].

Blood culture is the benchmark for diagnosing BSI but has limitations such as low positivity rates, long turnaround times, contamination risks, and challenges in detecting certain pathogens with standard culturing [9]. Machine learning (ML), a vital part of artificial intelligence, has advanced analytical powers that can independently detect disease patterns in data and forecast unknown results [10, 11]. Compared to traditional diagnostic and therapeutic approaches, ML offers a deeper insight into complex relationships. In recent years, ML has shown significant promise in disease screening, diagnosis, prognosis prediction, and risk analysis [12]. Developing early prediction models for BSI using ML is crucial for enhancing early diagnosis, treatment, and personalized healthcare. However, current prediction models often require a large number of features [13–14]. While including more features can improve the predictive ability of the models, it can also lead to increased complexity, requiring more data for training, and reducing interpretability and generalizability of the model. This poses a challenge in practical clinical settings, especially in primary care facilities where extensive testing and comprehensive patient data collection may not be feasible. Therefore, researchers and clinicians need to find a balance—ensuring predictive accuracy while minimizing the number and complexity of required features—to make these predictive models effective in resource-limited environments, such as grassroots healthcare institutions.

Consequently, this study aims to analyze routine laboratory/clinical data to identify key predictive factors that play a significant role in the early diagnosis of BSI. The goal is to find the optimal combination of these factors and use machine learning algorithm to develop a broadly applicable early risk prediction model for BSI. This model aims to facilitate early and rapid prediction of BSI in a variety of clinical settings and will be validated and implemented in real-world scenarios.

## Methods

### Study population

This study was a secondary analysis of a retrospective observational study conducted from 2021 to 2023 among inpatients at the Chongqing University Central Hospital. The inclusion criteria were (1) age ≥ 18 years; (2) inpatients; (3) had at least one blood culture examination performed during hospital stay. The exclusion criteria were (1) The blood culture results indicated a probable contaminant; (2) Data missing rate ≥ 30%. Clinical or laboratory parameters related to BSI were collected for each adult patient. For patients with multiple positive BC samples, only the first episode was included. For those with multiple negative BC samples, a single episode was randomly selected.

### Outcome

The outcome assessed was BSI, defined as the growth of a clinically significant pathogen in at least one BC bottle. Potential contaminants were defined by the Center for Disease Control and Prevention (CDC)/National Health Safety Network (NHSN) guidelines for Laboratory Confirmed Bloodstream Infection (LCBI) and were not

classified as BSIs. These potential contaminants include coagulase-negative Staphylococci, Corynebacterium species, Bacillus species, Diphtheroids, Aerococcus, and Propionibacterium species [13].

### Dataset

At the target medical centers, we constructed datasets that included demographics, clinical and laboratory parameters, including microbiology, available within 3 h before and after BC sampling time [15].

The dataset included as follows: (i) blood cells; (ii) liver function; (iii) renal function; (iv) hemagglutination; (v) blood gas analysis; (vi) electrolytes; (vii) inflammatory markers; (viii) blood culture; (ix) clinical features. All examination indicators were captured based on the time of sample collection.

We collected datasets from two time periods: the dataset from January 2021 to April 2023 was randomly split into training and validation sets comprising 70% and 30% respectively. The training sets were used for modelling, while the validation sets for internal validation. The dataset from May 2023 to December 2023 was used for external validation of the best model.

### Data preprocessing

Data cleaning and preprocessing are critical steps in the data analysis process, aimed at transforming raw data into a format suitable for statistical analysis or ML modeling [16, 17]. In this study, data cleaning and preprocessing primarily involved the removal of duplicate data, analysis and treatment of outliers, imputation of missing values, data standardization, and balancing of data categories.

In this study, to ensure the uniqueness of each patient's record and to avoid the duplication of patient information, we used the hospital admission number as the unique identifier for patients. During the data collection process, we made sure that each patient's admission number was recorded only once, meaning that only one record per patient was included in the study. We conducted outlier detection and analysis on the variables contained within the data. Upon identifying outliers, we reviewed the original records of the corresponding patients and assessed these outlier values. If an outlier was deemed unreasonable, it was removed; otherwise, the record was retained. The determination of outliers was primarily based on the opinions of clinical experts, with detailed information provided in Table 1. For missing data, we employed forward and backward filling algorithms to impute the values. In our study, we utilized the min-max standardization method for data preprocessing. This method does not rely on a specific data distribution and maintains the original characteristics of the data while simplifying the model training process, making it particularly effective for handling non-normally distributed data. Due to the imbalance in the proportion of positive and negative samples in the collected dataset, we used random oversampling techniques to balance the dataset. This is a common strategy in the field of machine learning.

**Table 1** Vital-signs values assumed to be plausible

| Items | Values |
|---|---|
| Temperature (Celsius) | 32–42 |
| Heart-Rate (bpm) | 30–190 |
| Systolic arterial pressure (mmHg) | 30–250 |
| Diastolic arterial pressure (mmHg) | 15–175 |
| Mean arterial pressure (mmHg) | 15–200 |

### Feature selection and modeling

Feature selection: (i) In this study, the initial method for selecting predictive factors involved univariable logistic regression. Univariable logistic regression allowed for the assessment of whether each biomarker was independently associated with BSIs, thus enabling the preliminary selection of predictive factors for model development. (ii) The study also incorporated the Boruta algorithm [18–20], Lasso regression [21–22], and Recursive Feature Elimination with Cross-validation (RFE-CV) [23–25] to optimize the results obtained from the univariable logistic regression analysis.

Modeling: In this study, we used the Light Gradient Boosting Machine (LightGBM), eXtreme Gradient Boosting (XGBoost), Gradient Boosting Decision Tree (GBDT), Random Forest (RF), Support Vector Machine (SVM) and Gaussian Naive Bayes (GNB) algorithms to predict the risk of BSI in inpatients by analyzing clinic/laboratory data [26–31]. Throughout the model development phase, we implemented a grid search technique to refine the hyperparameters.

### Validation and explanation

We evaluated the performance of the model by applying several different indices, namely (i) AUC, (ii) accuracy, (iii) sensitivity, and (iv) specificity. The performance assessment for selecting the best model will primarily be based on the AUC value. First, we conducted an assessment on the internal validation set, which comprised 30% of the original data that was initially set aside for validation purposes only. After model selection, we used the Shapley Additive Explanations (SHAP) algorithm from model-agnostic approaches to explain the best-performing model [32–34]. Finally, the dataset from May 2023 to December 2023 was utilized for external validation of the optimal model.

### Statistical analysis

Binary variables were presented as counts and percentages, and their significance was assessed using the Chi-square test or Fisher's exact test. Continuous variables that were normally distributed were compared with a t-test and reported as means ± SEM. For variables with a non-normal distribution, the Mann–Whitney U test was applied. A P-value of less than 0.05 was deemed statistically significant. All statistical analyses were conducted in the Beckman Coulter DxAI platform (https://www.xsmartanalysis.com/beckman/login/).

## Results

### Patient characteristics

Our model construction database initially contained 5,057 inpatients suspected of having BSI. Following a series of exclusions, 43 patients were under the age of 18, 70 patients had blood culture results suspected to be contaminated, and 2,621 patients had a data missing rate exceeding 30%. Ultimately, 2,323 adult inpatients were included in this study, of which 300 patients developed BSI, accounting for 12.9% of the study population. The missing rates of the included patients were shown in Supplementary Fig. 1.The patient selection process is illustrated in Fig. 1. The baseline characteristics of the patients are presented in Supplementary Table 1.The training and internal validation datasets comprised of 1,626 and 697 patients, respectively. A total of 74 variables, including age, sex, Temperature, White Blood Cell Count (WBC), D-dimer, and other laboratory or clinical parameters related to BSI, were collected for each patient. A comparison of basic information between the two sets were shown in Supplementary Table 2.

For external validation of the model, 259 patients were included, of whom 34 developed BSI (13.13%). The baseline characteristics of the patients are presented in Supplementary Table 3.

### Variables of importance

The model's accuracy increased as more variables were incorporated. However, increasing the number of variables did not correspond with the practicality needed for clinical application. In order to identify the most significant features, we employed univariate logistic regression to preliminarily screen the variables associated with BSI within the training set. We identified 27 variables that are crucial for predicting BSI, which were shown in Supplementary Table 4.

Based on the results of the univariate logistic regression analysis, the individual indicators that were screened (WBC, EOS, EOS%, Neu%, Mon, Mon%, RDW, Hct, PLT, A/G, Alb, CHE, PA, Cr, UA, Urea, Fib, SB, AB, BEf, Lac, TCO2, Cl, Mg, IL-6, hs-CRP, and T) were used separately to predict whether patients had BSI. As shown in Fig. 2, the AUC values for Neu%, Cr, Urea, and T exceeded 0.600, while the AUC values for the remaining indicators were all below 0.600. The efficacy of single indicators for predicting BSI was poor.

We utilized the Boruta algorithm, Lasso regression, and RFE-CV to further reduce the number of variables.

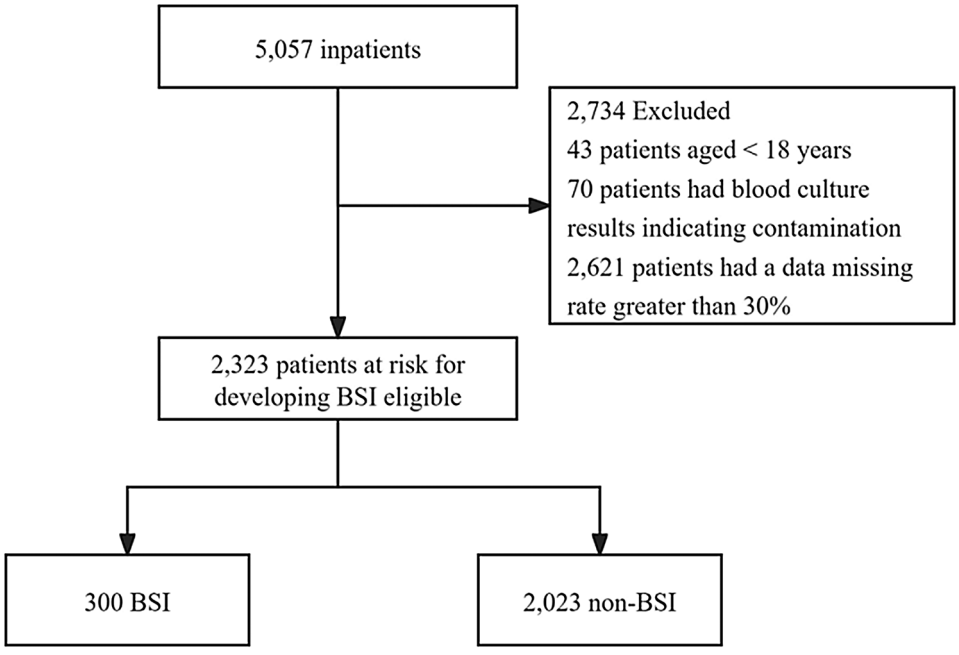


**Fig. 1** Flow chart depicting number of patients who were included in analysis after exclusion criteria. The total included encounters were divided into those with and without BSI

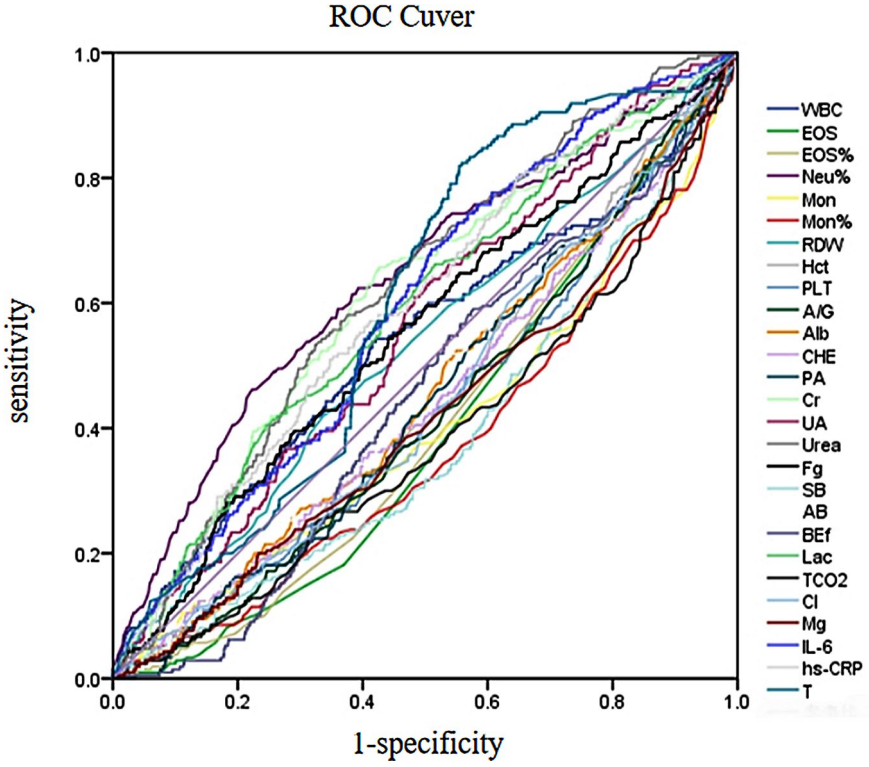


**Fig. 2** The ROC curves of predictive factors identified by univariate logistic regression analysis

As shown in Fig. 3, the Boruta algorithm indentified 19 variables such as Hct, Fib, UA, Cl, Alb, hs-CRP, WBC, TCO2, Urea, AB, Cr, Mg, Mon, IL-6, Mon%, BEf, SB, Neu%, and T. The Lasso regression analysis highlighted 15 features that help minimize the model's prediction error: WBC, EOS, PLT, PA, Lac, UA, TCO2, AB, SB, BEf, Na, Cl, hs-CRP, IL-6, and T. Meanwhile, the RFE-CV method selected the top five feature indicators based on their contribution rankings, which are WBC, SB, BEf, IL-6, and T. Ultimately, by taking the intersection of the results from these three algorithms, we identified the 5 key features that contribute the most to the model's predictive capability: WBC, SB, BEf, IL-6, and T.

#### Classification results

Based on the selected 5 key features (WBC, SB, BEf, IL-6, and T), we constructed six early prediction models for BSI risk using machine learning algorithms: the LightGBM model, the XGBoost model, the GBDT model, the RF model, the SVM model, and the GNB model. The model construction process involved hyperparameter optimization using grid search techniques.

As shown in Fig. 4, the average AUC values for the XGBoost, LightGBM, RF, GBDT, GNB, and SVM models on the internal validation set were 0.782 (95% CI: 0.715–0.849), 0.700 (95% CI: 0.627–0.773), 0.772 (95% CI: 0.704–0.841), 0.723 (95% CI: 0.650–0.797), 0.562 (95% CI: 0.483–0.642), and 0.528 (95% CI: 0.446–0.611), respectively. The XGBoost model had the highest AUC value of 0.782, while the SVM model had the lowest AUC value of 0.528. For the hyperparameters of each model, please refer to Supplementary Table 5 in the supplementary Information section.

As shown in Table 2, the RF model had the highest accuracy rate at 0.882; the GNB model had the highest sensitivity at 0.747; and the XGBoost model had the highest specificity at 0.824. Considering the AUC values and the evaluation metrics, the XGBoost model emerged as the best model.

In the external validation, the AUROC of the XGBoost model decreased to 0.776 (95% 0.685–0.864), with an accuracy of 0.685, sensitivity of 0.647, and specificity of 0.800. The calibration curve was close to the 45° line, indicating a good fit between the model's predictions and the actual values. The results are shown in Fig. 5.

#### Model interpretation and online application

To better understand the prediction results of the XGBoost model and the basis for decision-making, the SHAP algorithm was used to quantify the contribution of each feature to the model's predictive outcomes. Figure 6a displays the ranking of feature contributions in the XGBoost model, with the indicators ranked from highest to lowest contribution being SB, BEf, IL-6, T, and WBC.

For individual patients, as shown in Fig. 6b and c, the figure uses color coding to represent the impact of

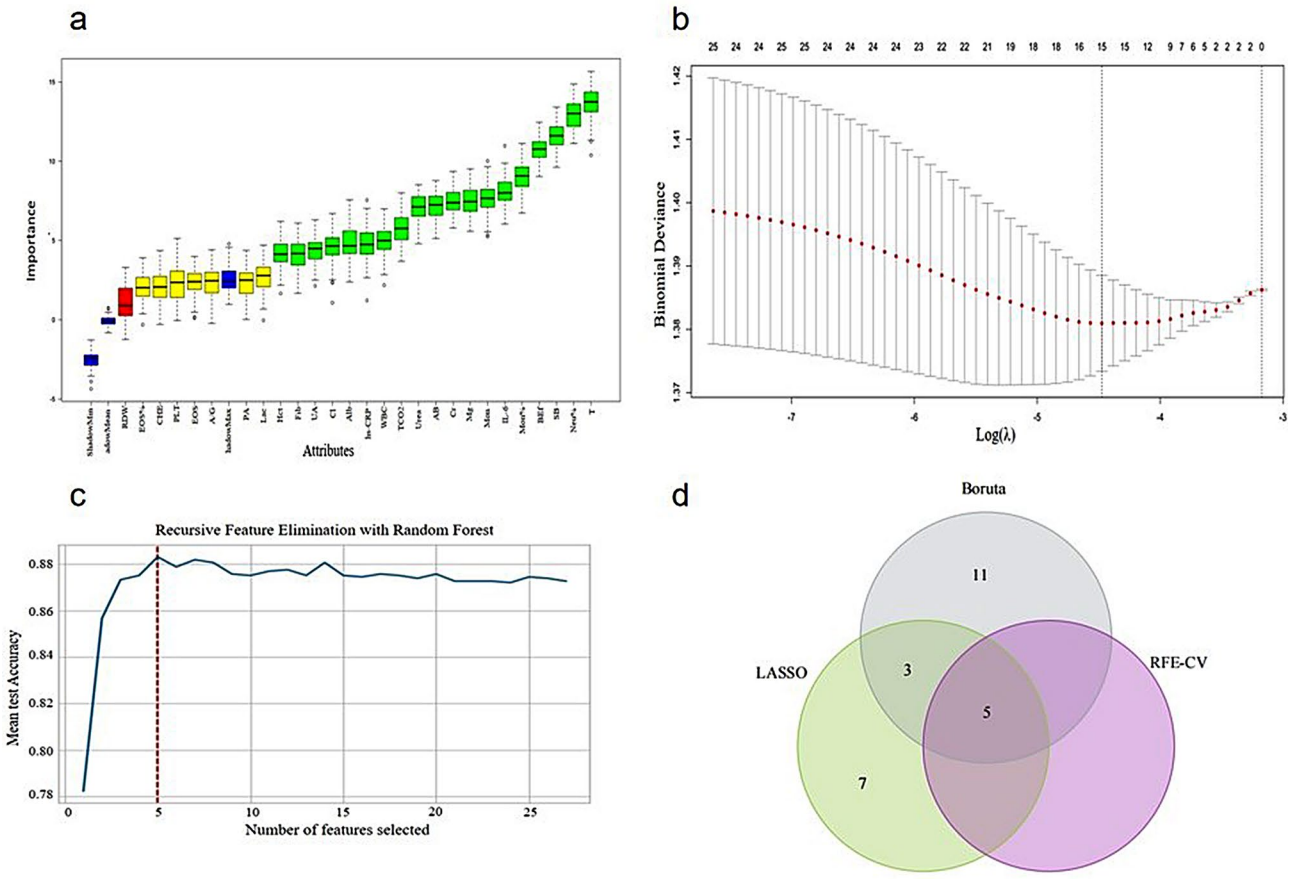


**Fig. 3** Selection of key features for BSI. (**a**) Variable Selection Plot of Boruta; (**b**) Variable Selection Plot of Lasso; (**c**) Variable Selection Plot of RFE-CV; (**d**) Venn graph displaying 5 features shared by Boruta, Lasso and RFE-CV

features on the prediction. Blue indicates features that negatively influence the prediction (leftward arrows, which correspond to a decrease in SHAP values), and red signifies features that positively affect the prediction (rightward arrows, indicating an increase in SHAP values). The base value represents the average model output for the training set, and the SHAP values for an individual patient's model output are indicated by f(x). In Fig. 6b, the f(x) value is below the base value (0.03 compared to 0.20), which suggests the model predicts a low risk of BSI for this patient. In contrast, in Fig. 6c, the f(x) value exceeds the base value (0.62 compared to 0.20), leading the model to predict a high risk of BSI for the patient.

To enhance the practicality and broad applicability of the constructed model in clinical practice, early risk prediction for patients can be conducted via an online link. The URL for the online prediction tool is: [http://www.xsmartanalysis.com/model/list/predict/model/html?mid=13885&symbol=11im71SWNC211Qj91806].

## Discussion

In this study, we developed a machine learning algorithm that utilizes clinical data to predict the risk of BSI in adult patients suspected of bacteremia. Traditional blood culture methods typically require several days to yield results, whereas our model can predict the likelihood of BSI within 3 h before and after blood culture collection. The XGBoost model outperformed other models, achieving an AUC of 0.782, with high specificity, closely aligning with the 45-degree line on the calibration curve. Therefore, this model was identified as the optimal model and will be used for subsequent external validation and clinical application. Clinicians can access the model online and input values for WBC (white blood cell count), SB (standard bicarbonate), T (body temperature), BEf (base excess), and IL-6 (interleukin-6) to obtain BSI risk predictions. It is particularly noteworthy that our model relies on only five common indicators to predict the occurrence of BSI, which significantly enhances the model's applicability and facilitates its widespread use in medical institutions at all levels.

Our study findings revealed a 12.91% incidence rate of BSI among hospitalized patients between 2021 and 2023.

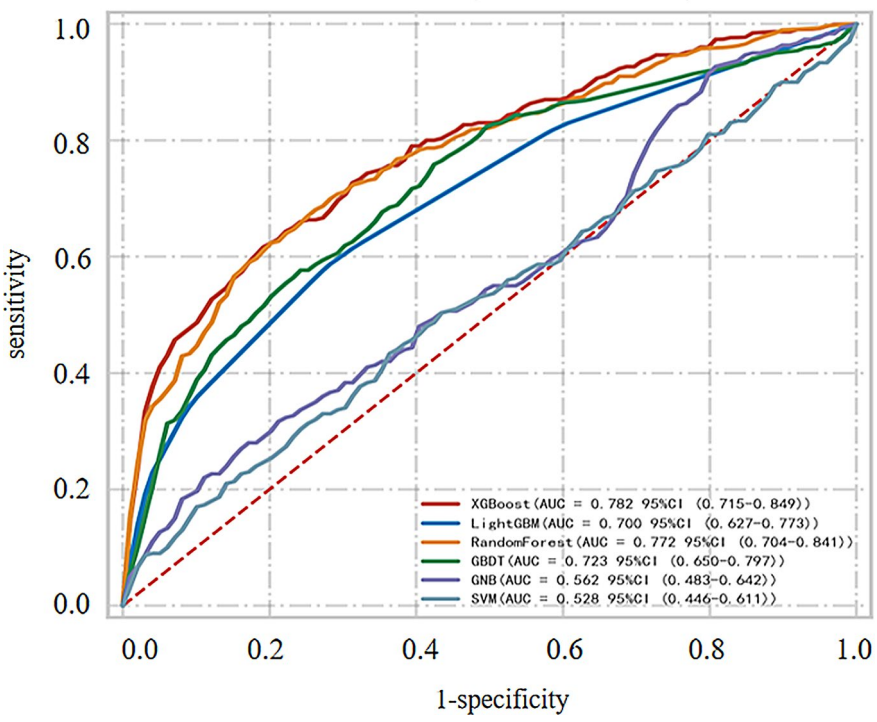


**Fig. 4** ROC curves of six models in the internal validation set

**Table 2** Evaluation metrics results of six models

| Model | AUC (95%CI) | Accuracy (95%CI) | Sensitivity (95%CI) | Specificity (95%CI) | PPV (95%CI) | NPV (95%CI) | F1-score (95%CI) |
|---|---|---|---|---|---|---|---|
| XGBoost | 0.782(0.715–0.849) | 0.763(0.724–0.802) | 0.633(0.518–0.749) | 0.824(0.720–0.927) | 0.309(0.267–0.352) | 0.936(0.932–0.940) | 0.41(0.372–0.447) |
| LightGBM | 0.700(0.627–0.773) | 0.83(0.816–0.844) | 0.527(0.435–0.619) | 0.788(0.691–0.886) | 0.346(0.291–0.401) | 0.905(0.896–0.914) | 0.408(0.378–0.438) |
| RF | 0.772(0.704–0.841) | 0.882(0.874–0.889) | 0.653(0.574–0.733) | 0.799(0.740–0.857) | 0.665(0.555–0.774) | 0.889(0.884–0.894) | 0.652(0.585–0.718) |
| GBDT | 0.723(0.650–0.797) | 0.729(0.689–0.769) | 0.643(0.541–0.745) | 0.738(0.647–0.829) | 0.257(0.225–0.290) | 0.92(0.913–0.928) | 0.365(0.327–0.402) |
| GNB | 0.562(0.483–0.642) | 0.47(0.253–0.688) | 0.747(0.520–0.973) | 0.416(0.174–0.657) | 0.155(0.142–0.168) | 0.918(0.884–0.953) | 0.254(0.223–0.286) |
| SVM | 0.528(0.446–0.611) | 0.595(0.389–0.801) | 0.51(0.250–0.770) | 0.642(0.395–0.889) | 0.169(0.136–0.203) | 0.882(0.850–0.914) | 0.24(0.190–0.289) |

This rate was higher than that reported in a 6-year retrospective study in the U.S. (12.91% vs. 5.90%) [35], likely due to the hospital's status as a national critical care center, which treats a higher volume of critically ill patients susceptible to BSI. The data collection period coincided with the COVID-19 pandemic, which may also have contributed to the increased BSI rates [36]. Finnish research has indicated higher BSI incidence and mortality in the elderly, particularly those over 80 years old [37–38]. Our study population had a median age of 68.0 years for the cohort and 72.0 years for BSI cases, which may explain the higher incidence. Given the aging population, addressing BSI in elderly patients is particularly crucial. It is important to note that pre-admission BSI cases were not excluded, which could have contributed to the higher incidence rate by including community-acquired BSI.

Early diagnosis of BSI is vital for lowering mortality and enhancing patient outcomes. As artificial intelligence evolves, ML algorithms are becoming pivotal in medicine, particularly for BSI diagnosis. Studies like Roimi's achieved an AUC of 0.930 with 50 features [13], Zhang's LSTM model reached 0.892 with over 100 features [15], and Zoabi et al. reported 0.810 with 25 features [39]. While more features can improve model performance, extensive data collection complicates practical use, especially in primary care where early BSI diagnosis is challenging. This study initially narrowed 74 predictors to 27 via univariate logistic regression, but single-factor prediction was inadequate. Further analysis led to feature selection using ML methods, including Boruta, Lasso, and RFE-CV, pinpointing 5 key indicators for early BSI risk, including SB, BEf, IL-6, Temperature, and WBC. WBC, IL-6, and Temperature are standard in infectious disease management and are key in BSI diagnosis. When a patient develops a BSI, the increase in WBC count, IL-6 levels, and body temperature is generally considered to be a result of the pathogen invasion activating the immune system, triggering an inflammatory

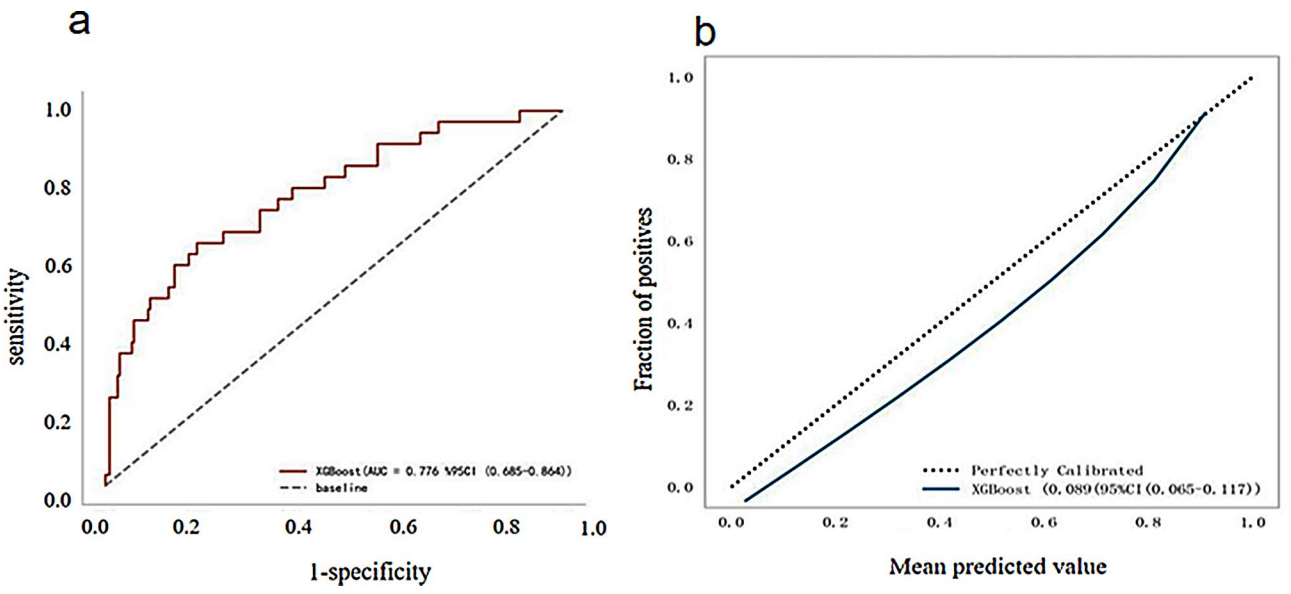


**Fig. 5** Performance evaluation of the XGBoost model. (**a**) ROC curve of external validation set in the XGBoost model; (**b**) calibration curve of XGBoost model

response. WBC are mobilized as immune cells to combat the infection, IL-6 is released as a pro-inflammatory cytokine to enhance the immune response, and the rise in body temperature serves as a defense mechanism to create an environment unfavorable for the survival of the pathogen [40–45]. Blood gas analysis, often focusing on TCO2 and pH, has seen less research on SB and BEf for early BSI detection. Some studies have indicated that during the early stages of infectious diseases, more pronounced changes occur in SB and BEF. Research suggests that the systemic inflammatory response induced by infection can impair the normal function of the circulatory system, thereby affecting tissue oxygenation. Even when the blood pH of patients has not shown significant fluctuations, the SB level begins to decline in the context of hypoxia [46]. When BSI patients experience acid-base balance disorders, BEF exhibits marked abnormalities. Song-Mao Ouyang and colleagues have observed statistically significant differences in BEF values between infected and non-infected patients ($P<0.05$) [47]. These five indicators are easier to obtain compared to the numerous features required by other studies. This means that the model can be more conveniently applied in other medical institutions.

Our model also has its limitations. As evidenced by the research results, the AUC values across all models, including the top-performing XGBoost model, remain suboptimal. We attribute this outcome to several key factors. First, the real-world patient dataset utilized in this study contains substantial missing values. While vital sign records exhibit relatively complete documentation, other laboratory test indicators suffer from severe data scarcity. The forward and backward filling methods employed for data imputation may have compromised data authenticity. We conducted additional experiments predicting BSI occurrence using only vital signs with fewer missing entries, yet the predictive performance remained unsatisfactory. Second, our BSI prediction framework incorporates only five clinical indicators—an intentional design choice to enhance clinical applicability. Although this feature scarcity inherently limits model predictive capability, we prioritized practical utility over theoretical performance. Excessive feature requirements would render the model operationally burdensome in real-world healthcare settings. This trade-off between predictive accuracy and clinical feasibility represents a deliberate compromise to ensure hospital adoption potential. Future iterations could explore balancing feature parsimony with enhanced predictive power through advanced feature engineering or multimodal data integration. Additionally, factors such as the empirical use of antibiotics by patients and the presence of multiple underlying diseases may impact the model's performance. The calibration curve results of the XGBoost model suggest a tendency to overestimate the risk of mortality. However, considering the severity of BSI as an infectious disease, clinicians may prefer overestimating the risk of death as a risk management strategy over missing a diagnosis [48–50]. In cases of BSI, early identification and intervention are vital for enhancing patient survival rates. Thus, despite the model's potential to overestimate risk, it still offers clinicians a more cautious foundation for treatment decisions, preventing the delay of necessary treatment due to underestimating the risk [51]. Addressing the issues highlighted by the calibration plot, future research will focus on developing a comprehensive decision support system

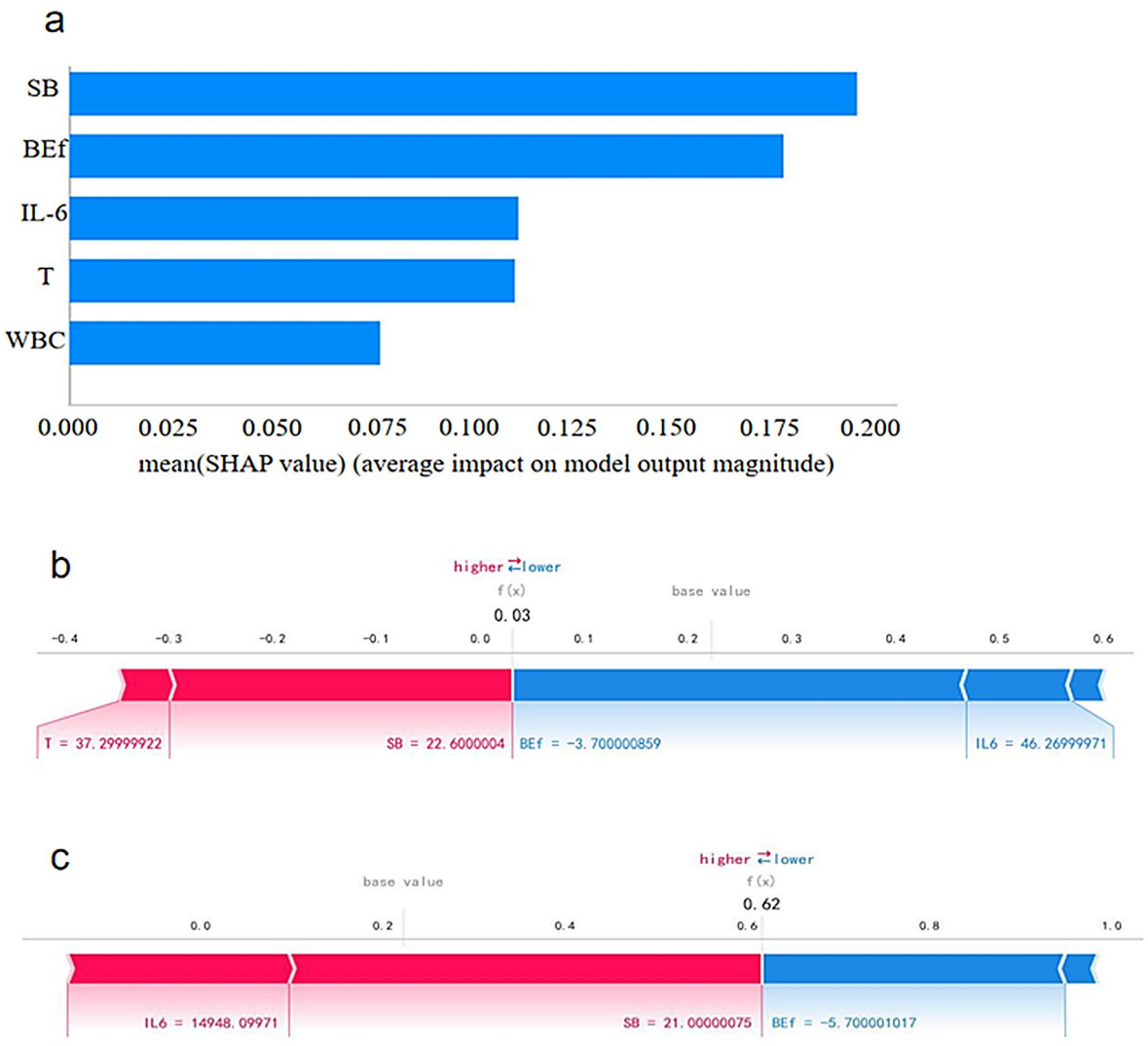


**Fig. 6** Model Interpretation of XGBoost. (**a**) Importance ranking of features; (**b**) Example of Low-risk Patient; (**c**) Example of hight-risk patient

that merges clinical experience with model predictions. Depending on the evaluation results, we will consider whether model adjustments are necessary to enhance calibration, while preserving its high sensitivity and ability to effectively identify high-risk patients.

This study aims to develop an early BSI risk prediction model utilizing standardized, cost-effective, and readily accessible laboratory indicators, with the goal of providing clinicians with an accurate yet simple diagnostic tool. Leveraging machine learning techniques, we have established a prediction framework intentionally designed for universal adaptability, ensuring seamless implementation across diverse healthcare facilities. Future research directions should prioritize multicenter validation studies to strengthen the model's generalizability and robustness. Subsequent investigations should further examine the model's predictive performance through clinical datasets spanning varied demographic populations, temporal contexts, and environmental settings, thereby comprehensively evaluating its clinical utility in decision-making processes. From clinical practice perspectives, our findings empower healthcare providers to enhance early BSI risk identification capabilities, enabling timely therapeutic interventions that optimize patient prognosis. For patient care, this translates to personalized treatment protocols and accelerated recovery trajectories. Regarding healthcare policy, the evidence-based insights derived from this research inform strategic optimization of medical resource allocation, fostering sustainable improvements in healthcare system efficiency.

In conclusion, this work not only advances methodological approaches to BSI risk assessment but also delivers practical solutions and conceptual frameworks that bridge clinical practice with health policy formulation. The developed model serves as both a predictive

instrument and a catalyst for transforming infection management paradigms.

**Abbreviations**

| | |
|---|---|
| BSI | Bloodstream Infection |
| MODS | Multiple Organ Dysfunction Syndrome |
| RFE-CV | Recursive Feature Elimination with Cross-validation |
| SHAP | Shapley Additive Explanations |
| ML | Machine learning |
| CDC | Disease Control and Prevention |
| NHSN | National Health Safety Network |
| LCBI | Laboratory Confirmed Bloodstream Infection |
| LightGBM | Light Gradient Boosting |
| XGBoost | Extreme Gradient Boosting |
| GBDT | Gradient Boosting Decision Tree |
| RF | Random Forest |
| SVM | Support Vector Machine |
| GNB | Gaussian Naive Bayes |
| MCHC | Mean Corpuscular Hemoglobin Concentration |
| RDW | Red Cell Distribution Width |
| MCH | Mean Corpuscular Hemoglobin |
| PLT | Platelet Count |
| WBC | White Blood Cell Count |
| Neu | Neutrophil Count |
| Neu% | Neutrophils Percentage |
| Eos | Direct Eosinophil Count |
| Eos% | Eosinophils Percentage |
| Mon | Monocyte Count |
| Mon% | Monocytes Percentage |
| Baso | Direct Basophil Count |
| Baso% | Basophils Percentage |
| Lym | Lymphocyte Count |
| Lym% | Lymphocyte Percentage |
| RBC | Red Blood Cell Count |
| Hb | Hemoglobin |
| Hct | Hematocrit |
| Plateletcrit | Plateletcrit |
| AFU | α-fucosidase |
| ALP | Alkaline Phosphatase |
| ALT | Alanine Aminotransferase |
| Alb | Albumin |
| CHE | Cholinesterase |
| GCA | Glycocholic Acid |
| LDH | Lactate Dehydrogenase |
| PA | Prealbumin |
| TBA | Total Bile Acids |
| TP | Total Protein |
| TBiL | Total Bilirubin |
| A/G | Alb/Glob Ratio |
| 5-n | 5-nucleotidase |
| GGT | Gamma-glutamyl Transferase |
| Cr | Creatinine |
| Cys-C | Cystatin C |
| UA | Uric Acid |
| α1-MG | α1-microglobulin |
| β2-MG | β2-microglobulin |
| PT | Prothrombin Time |
| APTT | Activated Partial Thromboplastin Time |
| Fib | Fibrinogen |
| TT | Thrombin Time |
| INR | International Normalized Ratio |
| D-D | D-dimer |
| PT% | Prothrombin Activity |
| AB | Actual Bicarbonate |
| AG | Anion Gap |
| BEf | Base Excess of Extracellular Fluid |
| BOP | Blood Osmotic Pressure |
| FiO2 | Fraction of Inspired Oxygen |
| PCO2 | Partial Pressure of Carbon Dioxide |
| pH | Pondus Hydrogeni |
| SB | Standard Bicarbonate |
| TCO2 | Total Carbon Dioxide Partial Pressure |
| Lac | Lactic Acid |
| PO2 | Oxygen Partial Pressure |
| Na | Sodium |
| K | Potassium |
| Cl | Chlorine |
| Mg | Magnesium |
| P | Phosphorus |
| PCT | Procalcitonin |
| IL-6 | Interleukin-6 |
| hs-CRP | High-sensitive C-reactive Protein |

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-025-03020-9.

Supplementary Material 1

**Acknowledgements**
Not applicable.

**Author contributions**
Xiefei Hu: Conceptualization, Data Curation, Methodology, Software, Writing-Original draft preparation, Writing- Reviewing and Editing. Shenshen Zhi: Conceptualization, case data collection and article design. Yuming Chen and Yang Li: Conceptualization, Methodology. Haiping Fan, Haorong Li, Zihao Meng and Jiaxin Xie: Data Curation, Methodology, Software. Shu Tang and Wei Li: Overall planning. All authors contributed to the article and approved the submitted version.

**Funding**
This work was supported by the Science and Technology Research Project of the Chongqing Municipal Education Commission (grant number: KJZD-M202300101), the Emergency Medicine Chongqing Key Laboratory Talent Development Innovation Joint Fund Project (grant number: 2024RCCX06) and the Wu Jieping Medical Foundation (grant number: 320.6750.2024-23-1 1). The statements made herein are solely the responsibility of the authors.

**Data availability**
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

## Declarations

**Ethics approval and consent to participate**
The studies involving humans were approved by the Ethics Committee of Chongqing Emergency Medical Center and Chongqing University Central Hospital (Approval Ethics Review No.RS202410). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study. The research protocol was approved by the institutional review board and adhered to the ethical guidelines of the Helsinki Declaration.

**Consent for the publication**
Not applicable.

**Competing interests**
The authors declare no competing interests.

**Author details**
[1]Department of Clinical Laboratory, Chongqing Emergency Medical Center, School of Medicine, Chongqing University Central Hospital, Chongqing University, Chongqing, China
[2]Peking University Chongqing Big Data Research Institute, Chongqing, China
[3]Beckman Coulter Commercial Enterprise (China) Co., Ltd, Shanghai, China
[4]School of Medicine, ChongQing University, Chongqing, China

[5]Chongqing University of Posts and Telecommunications, Chongqing, China

Received: 19 January 2025 / Accepted: 5 May 2025
Published online: 14 May 2025

**References**

1. Lamy B, Sundqvist M, Idelevich EA. Bloodstream infections - Standard and progress in pathogen diagnostics. Clin Microbiol Infect. 2020;26(2):142–50.
2. Shanghai Society for Microbiology, Clinical Microbiology Professional Committee, Shanghai Medical Association, Critical Care Medicine Specialty Branch, Shanghai Medical Association. Critical care medicine specialty branch. Expert consensus on clinical laboratory testing pathways for bloodstream infections. Chin J Infect Dis. 2022;40(08):457–75.
3. Vincent JL, Sakr Y, Singer M, et al. Prevalence and outcomes of infection among patients in intensive care units in 2017. JAMA. 2020;323(15):1478–87.
4. Lin K, Zhang HC, Zhao YH, et al. The direct application of plasma droplet digital PCR in the ultra-early pathogen detection and warning during sepsis: case reports. J Infect Public Health. 2022;15(4):450–4.
5. Rudd KE, Johnson SC, Agesa KM, et al. Global, regional, and National sepsis incidence and mortality, 1990–2017: analysis for the global burden of disease Study. Lancet (London England). 2020;395(10219):200–11.
6. 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):e209–18.
7. Overbeek R, Leitl CJ, Stoll SE et al. The value of Next-Generation sequencing in diagnosis and therapy of critically ill patients with suspected bloodstream infections: A retrospective cohort Study. J Clin Med. 2024;13(2).
8. Liu D, Huang SY, Sun JH, et al. Sepsis-induced immunosuppression: mechanisms, diagnosis and current treatment options. Mil Med Res. 2022;9(1):56.
9. Warren BG, Yarrington ME, Polage CR, et al. Evaluation of hospital blood culture utilization rates to identify opportunities for diagnostic stewardship. Infect Control Hosp Epidemiol. 2023;44(2):200–5.
10. Swanson K, Wu E, Zhang A, et al. From patterns to patients: advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. Cell. 2023;186(8):1772–91.
11. Peiffer-Smadja N, Rawson TM, Ahmad R, et al. Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. Clin Microbiol Infect. 2020;26(5):584–95.
12. Shibue K. Artificial intelligence and machine learning in clinical Medicine. N Engl J Med. 2023;388(25):2398.
13. Roimi M, Neuberger A, Shrot A, et al. Early diagnosis of bloodstream infections in the intensive care unit using machine-learning algorithms. Intensive Care Med. 2020;46(3):454–62.
14. Li X, Xu X, Xie F, et al. A Time-Phased machine learning model for Real-Time prediction of Sepsis in critical Care. Crit Care Med. 2020;48(10):e884–8.
15. Font MD, Thyagarajan B, Khanna AK. Sepsis and septic Shock - Basics of diagnosis, pathophysiology and clinical decision making. Med Clin N Am. 2020;104(4):573–85.
16. Chicco D, Oneto L, Tavazzi E. Eleven quick tips for data cleaning and feature engineering. PLoS Comput Biol. 2022;18(12):e1010718.
17. Maletzky A, Böck C, Tschoellitsch T, et al. Lifting hospital electronic health record data treasures: challenges and Opportunities. JMIR Med Inform. 2022;30(10):e38557.
18. Yue S, Li S, Huang X, et al. Machine learning for the prediction of acute kidney injury in patients with sepsis. J Translational Med. 2022;20(1):215.
19. Zhou H, Xin Y, Li S. A diabetes prediction model based on Boruta feature selection and ensemble learning. BMC Bioinformatics. 2023;24(1):224.
20. Kong C, Zhu Y, Xie X, et al. Six potential biomarkers in septic shock: a deep bioinformatics and prospective observational study. Front Immunol. 2023;14:1184700.
21. Zhou T, Ren Z, Ma Y, et al. Early identification of bloodstream infection in Hemodialysis patients by machine learning. Heliyon. 2023;9(7):e18263.
22. Tang G, Qi L, Sun Z, et al. Evaluation and analysis of incidence and risk factors of lower extremity venous thrombosis after urologic surgeries: A prospective two-center cohort study using LASSO-logistic regression. Int J Surg (London England). 2021;89:105948.
23. Han Y, Huang L, Zhou F. A dynamic recursive feature elimination framework (dRFE) to further refine a set of OMIC biomarkers. Bioinf (Oxford England). 2021;37(15):2183–9.
24. Al Abir F, Shovan SM, Hasan MAM, et al. Biomarker identification by reversing the learning mechanism of an autoencoder and recursive feature elimination. Mol Omics. 2022;18(7):652–61.
25. Zhang Z, Wang S, Zhu Z, et al. Identification of potential feature genes in non-alcoholic fatty liver disease using bioinformatics analysis and machine learning strategies. Comput Biol Med. 2023;157:106724.
26. Amin MN, Salami BA, Zahid M et al. Investigating the bond strength of FRP laminates with concrete using LIGHT GBM and SHAPASH Analysis. Polymers. 2022;14(21).
27. Moore A, Bell M, XGBoost A, Novel Explainable AI. Technique, in the prediction of myocardial infarction: A UK biobank cohort Study. Clin Med Insights Cardiol. 2022;16:11795468221133611.
28. Seto H, Oyama A, Kitora S, et al. Gradient boosting decision tree becomes more reliable than logistic regression in predicting probability for diabetes with big data. Sci Rep. 2022;12(1):15889.
29. Wallace ML, Mentch L, Wheeler BJ, et al. Use and misuse of random forest variable importance metrics in medicine: demonstrations through incident stroke prediction. BMC Med Res Methodol. 2023;23(1):144.
30. Hao PY, Chiang JH, Chen YD. Possibilistic classification by support vector networks. Neural Netw. 2022;149:40–56.
31. Liu T, Zhang X, Chen R, et al. Development, comparison, and validation of four intelligent, practical machine learning models for patients with prostate-specific antigen in the Gray zone. Front Oncol. 2023;13:1157384.
32. Li X, Zhao Y, Zhang D, et al. Development of an interpretable machine learning model associated with heavy metals' exposure to identify coronary heart disease among US adults via SHAP: findings of the US NHANES from 2003 to 2018. Chemosphere. 2023;311(Pt 1):137039.
33. Hu C, Li L, Huang W, et al. Interpretable machine learning for early prediction of prognosis in sepsis: A discovery and validation Study. Infect Dis Therapy. 2022;11(3):1117–32.
34. Ejiyi CJ, Qin Z, Ukwuoma CC et al. Comparative performance analysis of Boruta, SHAP, and Borutashap for disease diagnosis: A study with multiple machine learning algorithms. Network (Bristol England). 2024:1–38.
35. Rhee C, Dantes R, Epstein L, et al. Incidence and trends of Sepsis in US hospitals using clinical vs claims data, 2009–2014. JAMA. 2017;318(13):1241–9.
36. Valik JK, Hedberg P, Holmberg F, et al. Impact of the COVID-19 pandemic on the incidence and mortality of hospital-onset bloodstream infection: a cohort study. BMJ Qual. Saf. 2022;31(5):379–82.
37. Kontula KSK, Skogberg K, Ollgren J et al. Population-Based Study of Bloodstream Infection Incidence and Mortality Rates, Finland, 2004–2018. Emerg. Infect. Dis. 2021;27(10):2560-9.
38. Liu T, Wang J, Yuan Y, et al. Early warning of bloodstream infection in elderly patients with Circulating microparticles. Ann Intensive Care. 2021;11(1):110.
39. Zoabi Y, Kehat O, Lahav D, et al. Predicting bloodstream infection outcome using machine learning. Sci Rep. 2021;11(1):20101.
40. Tang YH, Jeng MJ, Wang HH, et al. Risk factors and predictive markers for early and late-onset neonatal bacteremic sepsis in preterm and term infants. J Chin Med Association: JCMA. 2022;85(4):507–13.
41. Yang X, Zeng J, Yu X, et al. PCT, IL-6, and IL-10 facilitate early diagnosis and pathogen classifications in bloodstream infection. Ann Clin Microbiol Antimicrob. 2023;22(1):103.
42. Zhu Q, Li H, Zheng S, et al. IL-6 and IL-10 are associated with Gram-Negative and Gram-Positive Bacteria infection in Lymphoma. Front Immunol. 2022;13:856039.
43. Niu D, Huang Q, Yang F et al. Serum biomarkers to differentiate Gram-negative, Gram-positive and fungal infection in febrile patients. J Med Microbiol. 2021;70(7).
44. O'Grady NP, Alexander E, Alhazzani W, et al. Society of critical care medicine and the infectious diseases society of America guidelines for evaluating new fever in adult patients in the ICU. Crit Care Med. 2023;51(11):1570–86.
45. Doman M, Thy M, Dessajan J, et al. Temperature control in sepsis. Front Med. 2023;10:1292468.
46. Xie Y, Yang Y, Han Y, et al. Association between arterial blood gas variation and intraocular pressure in healthy subjects exposed to acute Short-Term hypobaric Hypoxia. Transl Vis Sci Technol. 2019;8(6):22.
47. Ouyang SM, Zhu HQ, Xie YN, et al. Temporal changes in laboratory markers of survivors and non-survivors of adult inpatients with COVID-19. BMC Infect Dis. 2020;20(1):952.
48. Qi Z, Dong L, Lin J, et al. Development and validation a nomogram prediction model for early diagnosis of bloodstream infections in the intensive care unit. Front Cell Infect Microbiol. 2024;14:1348896.

49. Cohen R, Tannous E, Natan OB, et al. An emergency department intervention to improve earlier detection of community-onset bloodstream infection among hospitalized patients. Am J Infect Control. 2024;52(6):664–9.
50. Han S, Li R, Wang H, et al. Early diagnosis of bloodstream infections using serum metabolomic analysis. Metabolites. 2024;14(12):685.
51. Choi MH, Kim D, Park Y, et al. Development and validation of artificial intelligence models to predict urinary tract infections and secondary bloodstream infections in adult patients. J Infect Public Health. 2024;17(1):10–7.

## Publisher's note