Heliyon 9 (2023) e18263

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Early identification of bloodstream infection in hemodialysis patients by machine learning

Tong Zhou [a], Zhouting Ren [a], Yimei Ma [a], Linqian He [a], Jiali Liu [b], Jincheng Tang [a], Heping Zhang [a,*]

[a] Department of Nephrology, Affiliated Hospital of North Sichuan Medical College, Nanchong, China
[b] Department of Clinical Medicine, North Sichuan Medical College, Nanchong, China

ARTICLE INFO

Keywords:
Hemodialysis
Bloodstream infection
Machine learning
Prediction

ABSTRACT

Background: Bloodstream infection (BSI) is a prevalent cause of admission in hemodialysis (HD) patients and is associated with increased morbidity and mortality. This study aimed to establish a diagnostic, predictive model for the early identification of BSI in HD patients.
Methods: HD patients who underwent blood culture testing between August 2018 and March 2022 were enrolled in this study. Machine learning algorithms, including stepwise logistic regression (SLR), Lasso logistic regression (LLR), support vector machine (SVM), decision tree, random forest (RF), and gradient boosting machine (XGboost), were used to predict the risk of developing BSI from the patient's clinical data. The accuracy (ACC) and area under the subject working curve (AUC) were used to evaluate the performance of such models. The Shapley Additive Explanation (SHAP) values were used to explain each feature's predictive value on the models' output. Finally, a simplified nomogram for predicting BSI was devised.
Results: A total of 391 HD patients were enrolled in this study, of whom 74 (18.9%) were diagnosed with BSI. The XGboost model achieved the highest AUC (0.914, 95% confidence interval [CI]: 0.861–0.964) and ACC (86.3%) for BSI prediction. The four most significant co-variables in both the significance matrix plot of the XGboost model variables and the SHAP summary plot were body temperature, dialysis access via a non-arteriovenous fistula (non-AVF), the procalcitonin levels (PCT), and neutrophil-lymphocyte ratio (NLR).
Conclusions: This study created an effective machine-learning model for predicting BSI in HD patients. The model could be used to detect BSI at an early stage and hence guide antibiotic treatment in HD patients.

1. Introduction

Bloodstream infection (BSI) is the second leading cause of death among hemodialysis (HD). The mortality rate of BSI ranges from 19% to 25% of patients [1] [–] [4]. Furthermore, BSI also prolongs the hospital stay and the treatment costs for both patients and society. The early identification of BSI in HD patients is crucial for clinicians to provide timely and appropriate treatment necessary to reduce the mortality from BSI.

* Corresponding author. Department of Nephrology, Affiliated Hospital of North Sichuan Medical College, 1 Maoyuan Road, Nanchong city, Sichuan Province, 637000, China.
E-mail address: hepingzhang790316@163.com (H. Zhang).

https://doi.org/10.1016/j.heliyon.2023.e18263
Received 7 December 2022; Received in revised form 8 July 2023; Accepted 12 July 2023
Available online 13 July 2023

A blood culture could be used to identify pathogens within the blood and to guide antibiotic treatment. Although this test is still considered the gold standard for diagnosing bloodstream infections, it takes at least 12 to 48 h to obtain the results, making it difficult to provide early antibiotic treatment to counteract the infection [5]. Therefore, there is a need to create a diagnostic model to predict BSI at an early stage to provide timely, more effective treatment. However, the currently most available models for predicting BSI [6–8] are for non-hemodialysis patients. Recent studies have shown that dialysis increases the risk factors for the development of BSI. Several factors can increase the risk of developing BSI in HD patients, such as the access used for the dialysis (odds ratio (OR): 11.2; $p < 0.001$), previous hospitalizations (OR: 6.63; $p < 0.005$), history of diabetes (OR:2.651; $p < 0.001$), hypoproteinemia (OR:1.973, $p < 0.05$) [9–11]. However, the large number of potential risk factors makes it difficult for the clinician to determine the risk for the patient. Therefore, there is a need to develop predictive models to facilitate clinical decision-making. Sasaki et al. [12] used a logistic regression model to score the risk of developing bacteremia in HD patients (BAC-HD score) in an outpatient setting. However, the small number of crucial predictors included in this model resulted in poor predictive performance following external validation of the model. Therefore, to establish a more precise diagnostic prediction model for BSI in HD patients, the model must contain more reliable predictors and include patients from a wider cohort.

Machine learning algorithms are increasingly used in healthcare to identify variables linked to clinical outcomes and improve the clinical decision-making process. When compared with conventional regression analysis, machine learning provides better modeling of complex relationships [13,14]. Therefore in this study, we aimed to develop and validate a nomogram for the early prediction of BSI in HD patients. To achieve this aim, we compared the performance of 6 machine learning algorithms; stepwise logistic regression (SLR), lasso logistic regression (LLR), support vector machine (SVM), decision tree, random forest (RF), and gradient boosting machine (XGboost) were used to identify additional clinical features that could be predictive of BSI. As opposed to pre-existing models we incorporated additional predictive features such as procalcitonin levels (PCT) and neutrophil-lymphocyte ratio (NLR) in the development of the predictive nomogram.

2. Methods

2.1. Study population and data collection

A retrospective analysis of all medical records of HD patients treated at the Department of Nephrology, Affiliated Hospital of North Sichuan Medical College, Nanchong, Sichuan Province, China, from January 2018 to March 2022, was performed. Patients with a dialysis duration of less than two weeks, no blood culture results available, below 18 years, or with acute renal failure were excluded from this study (Fig. 1). The createDataPartition method [15] in the R language caret package was used to randomly split the patients into training, and testing sets using a ratio of 7:3. The training set was used to train the model, while the testing set was used to evaluate the performance of the model (Fig. 1). A 10-fold cross-validation was performed in the training set.

2.2. Ethical considerations

The study was approved by the ethics committee of the Affiliated Hospital of North Sichuan Medical College (File Number: 2022ER476-1). The need to obtain written informed consent was waived due to the retrospective nature of the study and that the data retrieved from the medical records were anonymous.

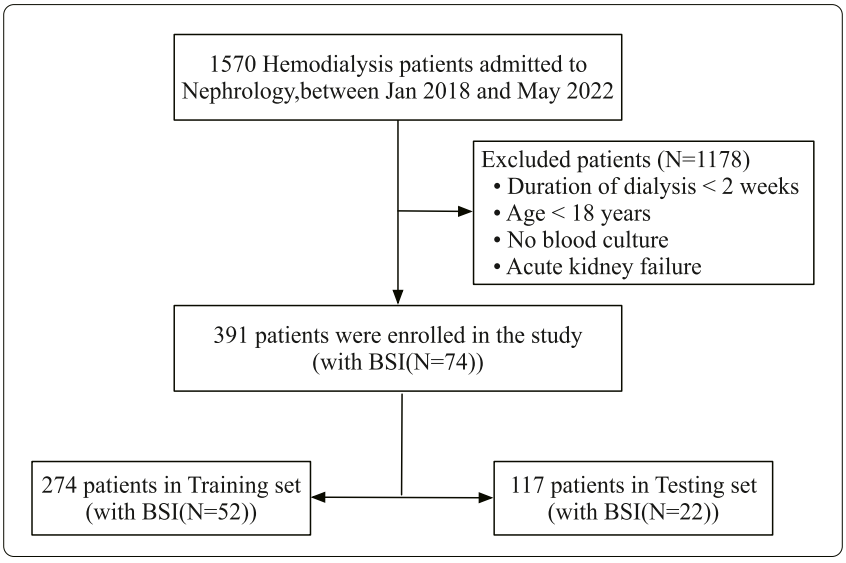


Fig. 1. Participant flow diagram.

2.3. Predictors and outcome measures

The predictors used to develop the machine learning models were selected based on previous studies [12,16] [–] [20] and expert judgment. The patient's demographic characteristics (age and gender) and clinical history (body temperature, systolic and diastolic blood pressure, heart rate, chills, vomiting, dialysis access, dialysis duration, history of bloodstream infection, history of diabetes, history of malignancy, history of antibiotic usage within one week, history of using steroid, history of using immunosuppressant) were extracted from the patient's medical records. The laboratory measurements, including leukocyte count, neutrophil count, neutrophil percentage, platelet count, neutrophil-lymphocyte ratio (NLR), lymphocyte-monocyte ratio (LMR), platelet-lymphocyte ratio, albumin, alkaline phosphatase, C-reactive protein (CRP), and procalcitonin (PCT), were also extracted. All data regarding clinical symptoms, vital signs, and laboratory measures were gathered within 24 h of obtaining the blood cultures. The body temperature, heart rate, respiratory rate, and blood pressure were also obtained on the same day as the blood cultures. In addition, we also included a non-arteriovenous fistula (non-AVF) and the dialysis access point as potential predictors in our study [9,21].

2.4. Identification of BSI

The presence and type of BSI were determined by a positive blood culture result. Two experienced laboratory physicians evaluated the blood cultures, and the presence or absence of BSI was determined by a qualified nephrologist.

2.5. Sample size and missing data

There is no universally acknowledged approach for computing sample size in the development and validation of clinical prediction models. Therefore, our same sample size calculation was based on that used in previous similar studies [12,22] and the number of desired outcome events. Since not all medical records provided the full laboratory information, the knnImputation function [23] in the R language DMwR package was used to fill in the missing CRP (17.6%) and PCT (3.3%) data.

2.6. Statistical analysis

The data analysis was conducted with the R software (version 4.2.0). The differences in the predictive variables were compared between the training and testing sets. The Wilcoxon rank sum test was used to compare the continuous variables, and the Chi-square test was used for the categorical variables. A p-value below 0.05 was deemed a statistically significant difference for all statistical tests.

2.7. Development and validation of the machine learning models

Six supervised machine learning techniques, SLR, LLR, SVM, decision tree, RF, and XGboost, were used to establish relevant diagnostic models. The SLR models have the advantage of reducing the multicollinearity problem and improving computational efficiency. However, in contrast to the models constructed using SLR, the screening of variables based on LLR is more effective in preventing over-fitting of the model and boosting model efficiency [24].

Logistic regression models select variables with a linear relationship between the independent and dependent variables and exclude independent variables with non-linear correlations. Due to this, we propose three tree-based algorithms: decision trees, RF, and XGboost. The decision trees model categorizes the data based on a simple, interpretable set of rules. However, these models are susceptible to overfitting. The RF algorithm is composed of decision trees and significantly lowers the risk of overfitting. XGboost is an ensemble learning technique based on gradient boosting that uses iterative computing of weak classifiers to produce accurate classification with high scalability and fast computational performance [25,26]. SVM is a classification analysis algorithm with distinct advantages in tackling small samples, nonlinearity, and high-dimensional pattern classification.

All the models were trained and validated on the testing dataset. The prediction performance of the model was evaluated by comparing the accuracy (ACC) and the area under the subject working curve (AUC) on the validation dataset. The R Metrics package is used to calculate the ACC, while the pRoc package was utilized to create the ROC curves and measure the AUC. Moreover, to better interpret the results of the model with the highest overall diagnostic value, we used Shapley Additive Explanation (SHAP) values [27] to assess the value for each feature within the model according to the training dataset. The SHAP summary plot shows the high and low SHAP values for the top 20 features. The extent of the SHAP values enabled us to quantify the significance of the impacts of the predictor variables on the results. The higher the SHAP value of a feature, the higher the predictive value of this feature. An importance matrix was used to illustrate all the features predictive of BSI. The SHAP dependence plot was used to explain how a single feature influences the output of the machine learning algorithm. The methodological flowchart of the study is shown in Supplementary Fig. 1.

2.8. Development of the nomogram

The best performing model was used to build a nomogram that clinicians could use to predict the occurrence of BSI. The nomogram was built using the rms package in R language. Calibration curves were plotted to test the performance of the nomogram on the training set and testing datasets. If the model has a better fit, the predictions will be more accurate when the Logistic regression calibration curve is closer to the diagonal.

3. Results

3.1. Participant characteristics

The data of 1570 HD patients were initially retrieved from the medical records. A total of 1179 patients were excluded as they did not meet the eligibility criteria (Fig. 1). Finally, 391 patients were included, and 18.9% (n = 74) of these patients had BSI. The data were divided into training (n = 274) and validation (n = 117) datasets, containing 52 and 22 cases of BSI, respectively.

The majority (58.4%) of the patients were men, and the average age of the participants was 58 years. The average duration of the dialysis treatment was 10 months. The remaining variables are described in more detail in Table 1. Arteriovenous endovascular fistula (AVF) was selected as the dialysis access method in 51.4% of the patients. *Staphylococcus aureus* accounted for 27 (36.5%) of the BSI, and *Staphylococcus epidermidis* accounted for 15 (20.3%) of the BSI cases (Table 2).

Table 1
Characteristics of the participants.

Characteristic	All Patients (n = 391)	Training Set (n = 274)	Testing Set (n = 117)	P Value
Sex , n (%)				0.83
Male	229 (58.6)	159 (58.0)	70 (59.8)	
Female	162 (41.4)	115 (42.0)	47 (40.2)	
Age (years)	58 (50,70)	59.5 (51,70)	57 (49–71)	0.58
Vital signs				
Body temperature (°C)	37.4 (36.9,38.6)	37.4 (36.9,38.6)	37.4 (36.9,38.6)	0.99
Systolic blood pressure (mmHg)	137 (121,156)	136 (122,155)	138 (122,158)	0.43
Diastolic blood pressure (mmHg)	78 (68–87)	78 (68,86)	76 (68,90)	0.79
Heart rate (/min)	92 (83–106)	92 (83,106)	92 (84,105)	0.73
Breathing rate (/min)	20 (20–21)	20 (20,21)	20 (20,21)	0.60
Symptoms , n (%)				0.13
Chill	46 (11.8)	33 (12.0)	13 (11.7)	0.93
Vomit	64 (16.4)	45 (16.4)	19 (16.2)	1.00
Dialysis duration(m), n (%)	10 (2,26.5)	12 (2,36)	6 (1,24)	**0.04**
Dialysis Access, n (%)				
AV fistula	201 (51.4)	149 (54.4)	52 (44.4)	
Central venous catheter	64 (16.4)	46 (16.8)	18 (115.4)	
Femoral vein catheter	125 (32.0)	78 (28.5)	47 (40.2)	
AV graft	1 (0.2)	1 (0.4)	0 (0,0)	
History of BSI, n (%)	25 (6.4)	18 (6.6)	7 (6.0)	1.00
Comorbidities , n (%)				
Diabetes mellitus	121 (30.9)	75 (27.4)	46 (39.3)	**0.03**
Malignancy	29 (7.4)	24 (8.8)	5 (4.3)	0.18
Medication , n (%)				
Steroid use	43 (11.0)	32 (11.7)	11 (9.4)	0.63
Immunosuppressant use	32 (8.2)	25 (9.1)	7 (6.0)	0.40
Antibiotics use within one week	94 (24.0)	64 (23.4)	30 (25.6)	0.72
Causes of CKD, n (%)				**0.03**
Diabetic nephropathy	115 (29.4)	71 (25.9)	44 (37.6)	
Hypertensive nephrosclerosis	9 (2.3)	9 (3.3)	0 (0.0)	
Chronic glomerulonephritis	63 (16.1)	48 (17.5)	15 (12.8)	
Others and unknown	204 (52.2)	146 (53.3)	58 (49.6)	
Laboratory findings				
Leukocyte count (10^9/L)	7.16 (5.4,10.6)	6.9 (5.2,10.3)	7.3 (5.4–10.3)	**<0.01**
Neutrophil count (10^9/L)	5.7 (4.1,8.9)	5.3 (3.9,8.4)	6.3 (4.5,9.8)	**0.01**
Neutrophil percentage (%)	79.7 (71.8,86.9)	78.7 (71.8,86.8)	81.9 (72.2,87)	0.26
Platelet count (10^9/L)	149 (106,197)	148 (101,197)	152 (113,196)	0.23
Neutrophil/Lymphocyte ratio	7.0 (4.1,13.3)	6.7 (4.1–13.2)	8.1 (4.2,14.3)	0.30
Lymphocyte/Monocyte ratio	1.8 (1.1,2.8)	1.7 (1.1,2.8)	1.9 (1.2,2.8)	0.58
Platelet/Lymphocyte ratio	187 (114,274)	187 (120,273)	187 (104,280)	0.79
Albumin (g/L)	35.6 (31.2,39.4)	35.9 (31.2–39.6)	35 (31–38.9)	0.45
Alkaline phosphatase (U/L)	89 (70,115)	89.5 (69.2,114)	89 (72–116)	0.51
C-reacted protein (mg/L)	22.1 (6.7,65.0)	22.6 (6.5,68.6)	20.1 (7.9,62.0)	0.48
Procalcitonin (ug/L)	1.2 (0.4,4.7)	1.3 (0.4,4.1)	1.0 (0.4,5.8)	0.57
Type of infection, n (%)				0.49
Pulmonary Infection	233 (59.6)	162 (59.1)	71 (60.7)	
Urinary tract infections	25 (6.4)	15 (5.5)	10 (8.5)	
BSI	74 (18.9)	52 (19.0)	22 (18.8)	
ARBSI	52 (13.3)	34 (12.4)	18 (15.4)	
Others	22 (5.6)	15 (8.0)	7 (6.0)	
No infection	37 (9.5)	30 (10.9)	7 (6.0)	

Data are presented as median (interquartile range) or number (%). *Abbreviations*: AV fistula, arteriovenous endovascular fistula; AV graft, arteriovenous endovascular graft; CKD, Chronic Kidney Disease; BSI, Bloodstream infection; ARBSI, Access-related Bloodstream infection.

3.2. Model comparisons

Fig. 2 shows the ACC and AUC obtained by all models on the testing dataset. LLR had the highest ACC (88.0%), followed by XGboost (86.3%), RF (86.3%), SLR (85.5%), SMV (84.6%), and decision tree (82.9%) on the testing dataset. However, XGboost had the highest AUC (0.914, range: 0.861–0.968), followed by LLR (0.903, range: 0.836–0.970), RF (0.900, range: 0.836–0.964), SVM(0.883, range: 0.816–0.951), SLR(0.855, range: 0.770–0.933), decision tree (0.844, range: 0.740–0.968). The 10-fold cross-validation results showed that XGboost showed the highest mean AUC (0.891) and ACC (0.886) values (Supplementary Table 1).

Since the XGboost model exhibited the highest AUC, it was chosen as the best prediction model and analysised further by SHAP values.

3.3. Optimal features analysis

Body temperature, PCT, age, non-AVF, NLR, LMR, white blood cell count, platelet count, diastolic blood pressure, and neutrophil percentage were identified as the top 10 most important variables by the XGboost algorithm (Fig. 3). The SHAP summary plot of the XGboost model and the top 20 features predictive of BSI are illustrated in Fig. 4. The variable importance matrix plots and SHAP summary plots identified that body temperature, non-AVF, PCT, and NLR were the four most important predictor factors for BSI.

Fig. 5 shows the results of the SHAP dependence plots. The values on the y-axis represent the SHAP value of the feature, and the X-axis represents the changes in feature values. SHAP values above zero for specific features indicate an increased risk of BSI.

3.4. Nomogram and calibration curve

Table 3 displays the regression coefficients and intercepts of the optimum logistic regression model developed with the first 4 significant predictor variables. Fig. 6A illustrates the BSI prediction nomogram. The calibration curves used to evaluate the model's performance on the training and validation datasets are illustrated in Fig. 6B and C, respectively.

4. Discussion

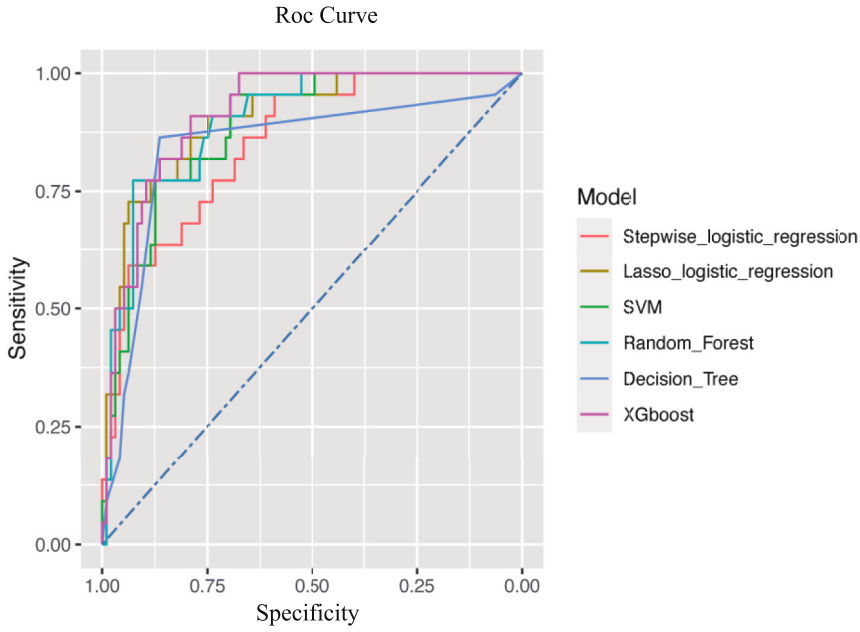
Early intervention is essential to reduce morbidity and mortality from BSI in HD patients. Previous studies have attempted to develop prediction models for patients with BSI. Shapiro created a risk prediction model to forecast the occurrence of BSI in general emergency medical patients [6]. Although several later cohorts validated the model [28,29], the AUC of the model after external validation was relatively poor in HD patients compared to the general population (AUC: 0.73 vs. 0.81) [22,28]. Sasaki et al. established a score for predicting BSI in maintenance hemodialysis patients using multicenter emergency department records [12]. However, the study's limitations were restricted predictor factors and uncertain data quality. As a result, it would be essential to build a more reliable computational method for developing a better diagnostic model to predict the incidence of BSI in HD patients.

To our knowledge, this is the first study making use of machine learning to identify clinical variables predictive of BSI in HD patients. A total of 28 clinical variables were used to train 6 machine learning models, including SLR, LLR, SVM, decision tree, RF, and XGboost. The predictive ability of these models was compared in a cohort dataset of 117 patients. The XGboost model had the highest AUC (0.914) and ACC (86.3%). The XGboost model makes use of a bootstrapping method and has powerful predictive efficacy even in limited datasets [30]. The XGboost model variable importance matrix plot (Fig. 3) and SHAP summary plot (Fig. 4) showed that body temperature, non-AVF, PCT, and NLR have a significant predictive role for BSI. These 4 predictive variables were used to construct the final nomogram.

Similar to previous studies [1,9,21], non-AVF was found to be a strong predictor for the development of BSI in HD patients. When compared with our cohort, a lower proportion of patients in the study of Sasaki et al. (26.4% vs. 48.6%) [12] utilized the non-AVF method as the access point for dialysis, as recommended by the guidelines the preferred use of AVF [31]. Given its low usage rate, the author advised against using non-AVF as a predictor for BSI [12], which appears to be at odds with the results of this study. However, A possible explanation for the increased use of non-AVF dialysis access in our study could be attributed to the fact that the patients in our cohort had been on dialysis for a shorter period (median dialysis time 10 months versus. 61 months), indicating that a substantial proportion of patients were in the early stages of dialysis, when non-AVF uses appear to be increasing as a result of delayed

Table 2
BSI pathogens.

Staphylococcus aureus	27	Candida albicans	1
Staphylococcus epidermidis	15	Propionibacterium acnes	1
Escherichia coli	5	Rhodopseudomonas pilosula	1
Klebsiella pneumoniae	5	*Candida tropicalis*	1
Enterococcus faecalis	4	Staphylococcus carnosus	1
Serratia marcescens	2	Stenotrophomonas maltophilia	1
Enterobacter cloacae	2	Staphylococcus cephalosporus	1
Streptococcus agalactiae	2	Aeromonas punctata subsp.caviae	1
Staphylococcus mimicus	1	Corynebacteriumstriatum	1
Human Staphylococcus	*1*	*Staphylococcus wolframii*	*1*



Model	Accuracy in the Test Set	AUC in the Test Set
Stepwise logistic regression	0.855	0.852(0.770-0.933)
Lasso logistic regression	0.880	0.903(0.836-0.970)
Support vector machine (SVM)	0.846	0.883(0.816-0.951)
Random Forest (RF)	0.863	0.900(0.836-0.964)
Decision Tree	0.829	0.844(0.740-0.948)
eXtreme Gradient Boosting (XGboost)	0.863	0.914(0.861-0.968)

Fig. 2. ACC and AUC values achieved by the machine learning models on the testing dataset. The XGboost model achieved the best AUC, and the LLR achieved the best ACC.

diagnosis of uremia, the failure of AVF maturation and AVF dysfunction [31–33]. Moreover, according to the dialysis outcomes and practice patterns study (DOPPS), AVF was chosen as the first vascular access in only 44% of northern China and it appears to be equally prevalent in the rest of China [34,35]. Therefore, we suggest using non-AVF in the model to evaluate the presence of BSI.

The SHAP analysis identified PCT as a strong predictor for BSI (Fig. 4). This finding was confirmed by the SHAP independent plot (Fig. 5B), which also showed that as the PCT concentration increases, the risk of BSI increases. Therefore, PCT was identified as a critical contributor to our model. The most influential biomarkers for identifying infection were the inflammatory markers PCT and CRP. Plasma PCT levels are elevated in patients with sepsis and following severe bacterial or fungal infections. However, compared with CRP, the PCT levels tend to be lower in patients with less severe inflammatory responses. Studies have shown that PCT is a better indicator for the early prediction of BSI than CRP [36] [–] [38]. Interestingly, owing to the presence of different types of infections in our model, CRP had a lower diagnostic value when compared with PCT. It is also important to note that the kidneys may not effectively clear PCT in patients with severe renal failure. As a result, PCT on its own may be a poor indicator of [39,40] the optimal cut-off value for PCT to diagnose BSI in HD patients may not be remarkably different from that of the general population. Therefore, more research is required to establish the value of PCT as a predictive biomarker for BSI.

NLR is also one of the top five variables in the XGboost model and SHAP summary plots (Figs. 3 and 4). NLR is a simple, widely available test that could be used to predict BSI. However, NLR on its own is insufficient to predict BSI and must be coupled with other proven biological markers to increase its diagnostic accuracy for bloodstream infections [18].

Consistent with previous studies, *S. aureus* (36.5%) and *S. epidermidis* (20.3%) were identified as the main pathogens involved in the development of BSI (Table 2) [12,21]. A possible explanation for this tendency could be the use of percutaneous tunnel catheters in 70.3% of the patients in this study. These findings may suggest that antibiotics targeting gram-positive bacteria could be selected as the primary empirical treatment option for patients with suspected BSI. Furthermore, the frequency and duration of percutaneous tunnel catheters should be reduced to prevent [41].

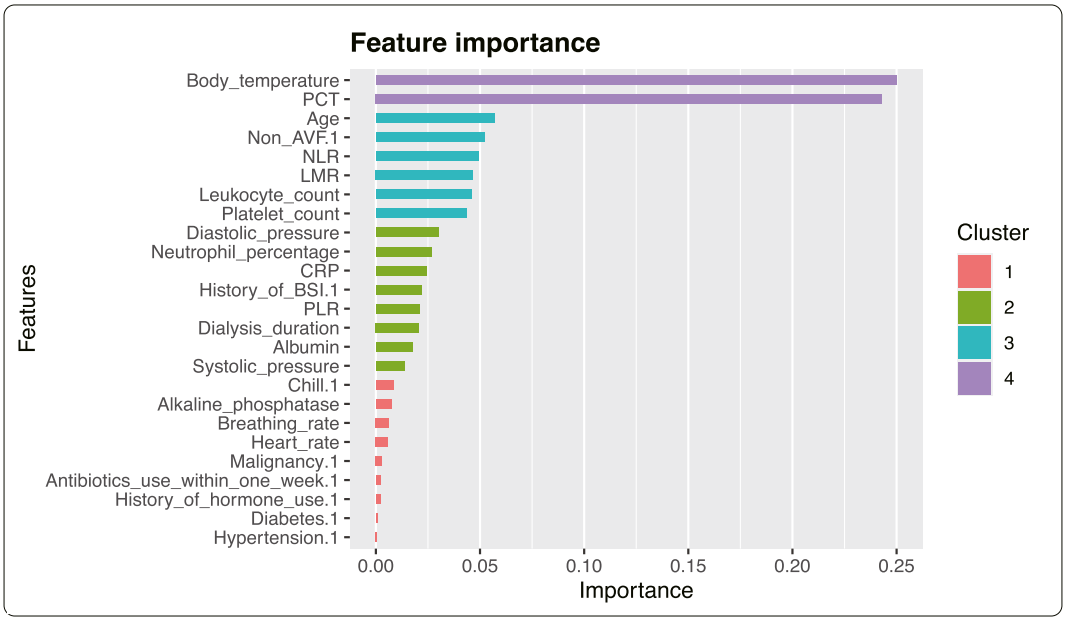


Fig. 3. Importance matrix plot of the XGboost model. This importance matrix plot depicts the importance of each covariate in building the final prediction model. Abbreviations: PCT, Procalcitonin; CRP, C-reacted protein; NLR, Neutrophil/Lymphocyte ratio; non-AVF, Non-Arteriovenous Fistula; PLR, Platelet/Lymphocyte ratio; LMR, Lymphocyte/Monocyte ratio.

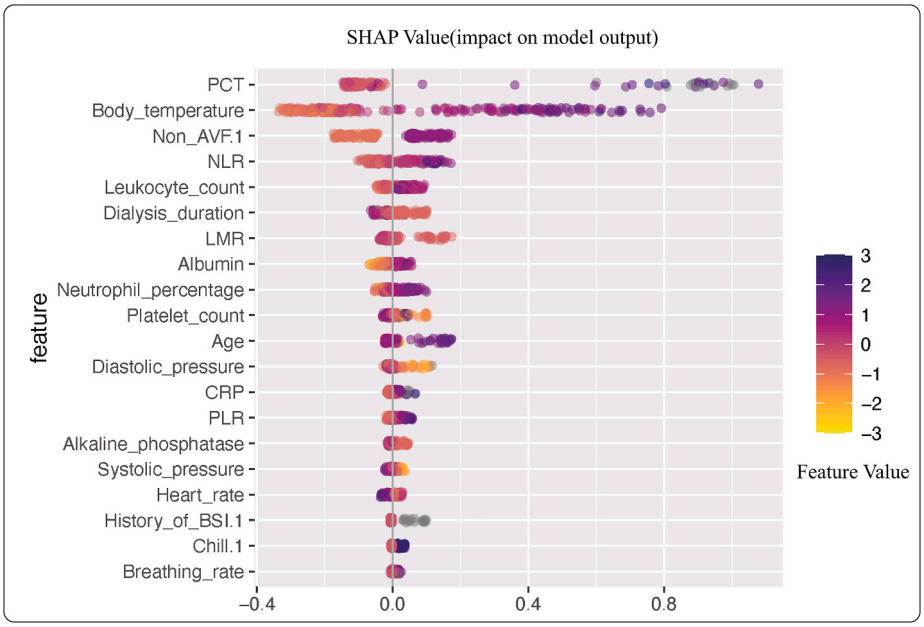


Fig. 4. SHAP summary plot of all XGboost model features. The higher the SHAP value of a feature, the higher the risk of BSI. Each patient complex contains one point for each feature attribution value, so each patient feature is represented by a point on the line. The color of the points is determined by the feature values of each patient and is piled vertically to represent the density. The yellow color on the heat map indicates a lower feature value, whereas the blue color indicates a higher feature value. Abbreviations: PCT, Procalcitonin; non-AVF, Non-Arteriovenous Fistula; CRP, C-reacted protein; NLR, Neutrophil-Lymphocyte ratio; PLR, Platelet-Lymphocyte ratio; LMR, Lymphocyte-Monocyte ratio. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

This study has certain limitations that have to be acknowledged. The sample used in our study was small, and all the data were obtained from a single center. As a result, the accuracy and generalisability of our proposed model might be limited. Therefore, external validation is indeed required to prevent model overfitting. Since the data for this study were obtained retrospectively, some

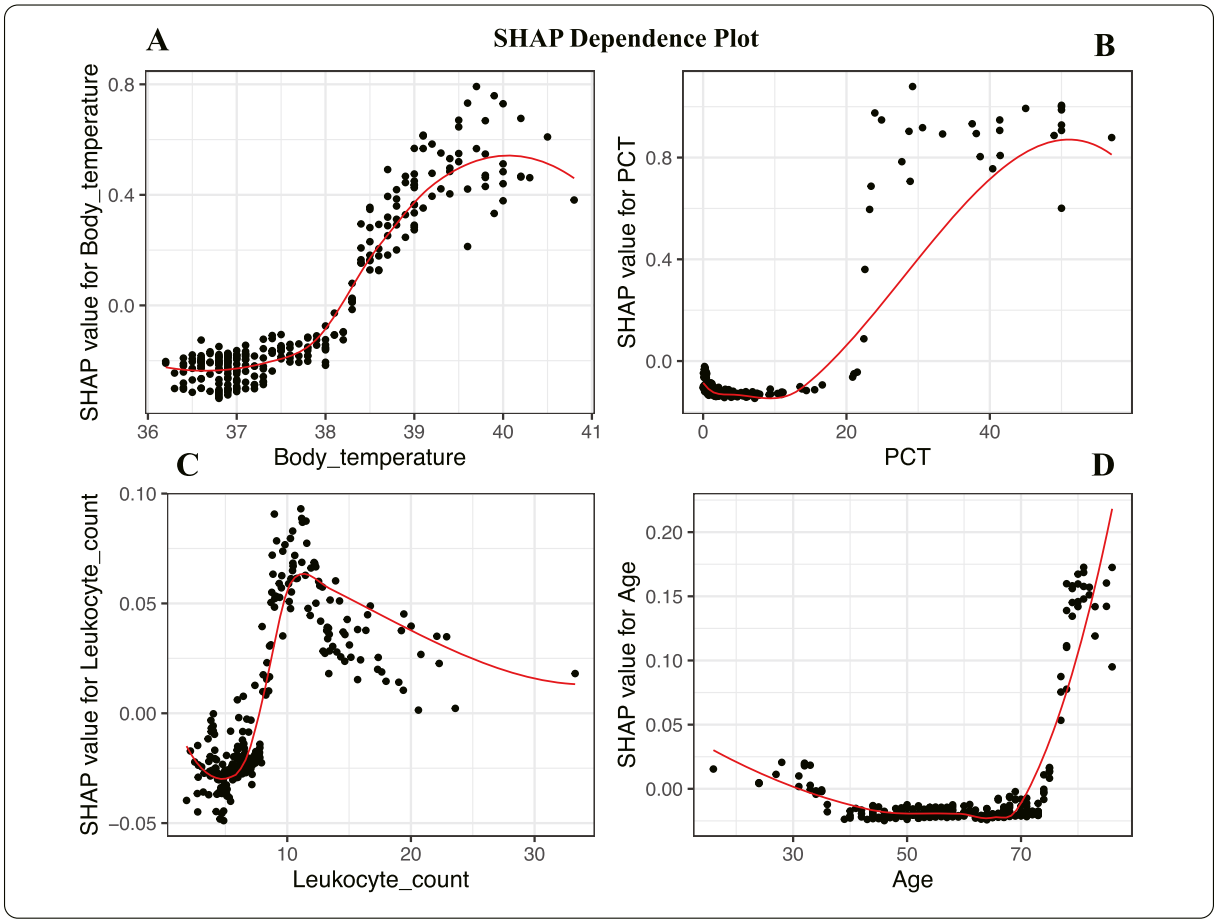


Fig. 5. SHAP dependence plot of the XGboost model. The SHAP dependence plot explains how a single feature influences the XGboost prediction model's output. Feature SHAP values above 0 indicate an elevated risk of BSI. Abbreviations: PCT, Procalcitonin.

Table 3
Logistic model based on the first 4 important variables of the XGboost model.

	β Coefficient	Odds Ratio (95%CI)	P Value
Intercept	−36.597		**<0.010**
NLR	0.003	1.003 (0.972–1.034)	0.862
Body temperature	0.875	2.400 (1.703–3.473)	**<0.010**
Non-AVF	1.944	6.988 (3.102–17.425)	**<0.010**
PCT	0.063	1.065 (1.037–1.109)	**<0.010**

Abbreviations: NLR, Neutrophil-Lymphocyte ratio; non-AVF, Non-Arteriovenous Fistula; PCT, Procalcitonin.

data were missing, potentially limiting the accuracy of our model. Therefore, more prospective studies are required to further validate the performance of this model. Although the XGboost-based nomogram could be successfully used for the early prediction of BSI in HD patients, it can not be used to predict the type of infection. Therefore, until antibiotics susceptibility tests are available, clinicians will still need to make use of their clinical judgment to select the optimal antibiotic treatment.

5. Conclusions

In this study, we compared the efficacy of 6 different machine learning models in predicting BSI in HD patients. The models were trained using 28 different variables, and body temperature, non-AVF, PCT, and NLR were identified as the main predictors of BSI. The XGboost algorithm provided the best performance on the training dataset. Internal validation showed that the model could be used to develop a feasible clinical nomogram to predict BSI in HD patients. Clinicians could use the model as a guide for the early administration of antibiotic treatment in HD patients. However, larger multicenter studies are required to validate the performance of this model.

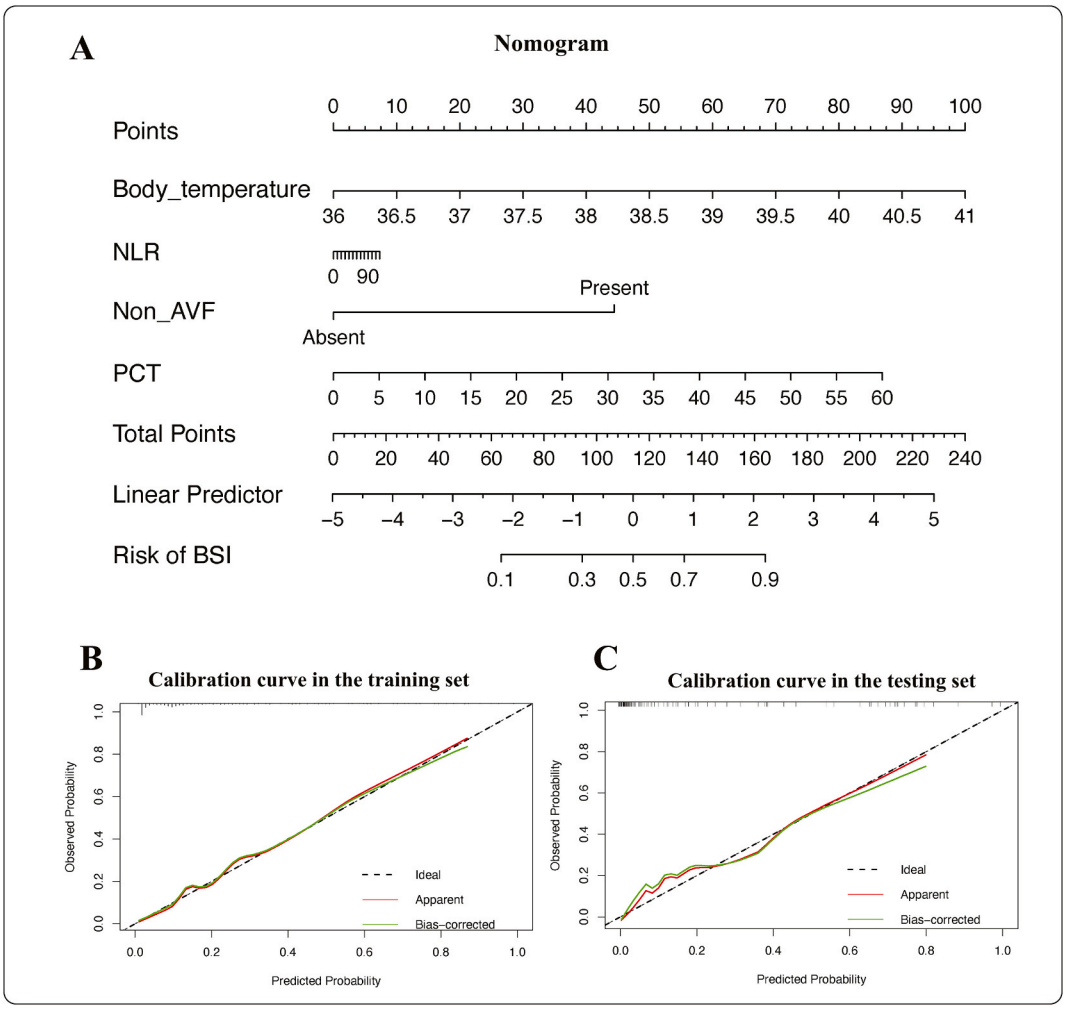


Fig. 6. (A): Nomogram used to predict BSI in HD patients based on the XGboost model. Each independent variable on the nomogram is assigned a point by drawing a line from the independent variable scale to the points scale (e.g., a body temperature of 39 °C is assigned 60 points). The total score is calculated by summing up the points assigned for each independent variable on the nomogram. The final BSI score is calculated by drawing a line from the total points scale to the risk of bias scale. (B): Calibration curve evaluating the ability of the nomogram to predict BSI in the training dataset (n = 274). (C): Calibration curve evaluating the ability of the nomogram to predict BSI in the testing dataset (n = 117). The Ideal line indicates that the model's prediction probability and the actual probability are identical, the Apparent line depicts the model's training performance, and the Bais-corrected line depicts the model's performance after 1000 iterations of sample sampling. Abbreviations: non-AVF, Non-Arteriovenous Fistula; PCT, Procalcitonin; BSI, Bloodstream infection.

Funding

This work was financially supported by the Funding Project of the Bureau of Science and Technology and Intellectual Property of Nanchong City (No.19SXHZ0164).

Author contribution statement

Heping Zhang: conceived and designed the experiments; contributed reagents, materials, analysis tools or data.

Tong Zhou: conceived and designed the experiments; performed the experiments; analyzed and interpreted the data; wrote the paper.

Zhouting Ren: conceived and designed the experiments; wrote the paper.

YiMei Ma; Linqian He; Jiali Liu; Jincheng Tan: performed the experiments.

Data availability statement

Data included in article/supp. material/referenced in article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e18263.

References

[1] S.D. Lars, N. Mette, J. Bente, J.-F. Søren, Ø.L. Jørgen, S.H. Carl, S.O. Schmeltz, Risk and prognosis of bloodstream infections among patients on chronic hemodialysis: a population-based cohort study, PLoS One 10 (2015), e0124547.
[2] J. C.A, N. F.R, G.D. T, C. Shu-Cheng, United States Renal Data System public health surveillance of chronic kidney disease and end-stage renal disease, Kidney Int. Suppl. 5 (2015) 2–7.
[3] R. P.P, J. K.A, J. A.M, Epidemiology, surveillance, and prevention of bloodstream infections in hemodialysis patients, Am. J. Kidney Dis. 56 (2010) 566–577.
[4] N.-K.K. Emil, L.G. Hellmund, H.J. Goya, Risk of death after first-time blood stream infection in incident dialysis patients with specific consideration on vascular access and comorbidity, BMC Infect. Dis. 18 (2018) 1–12.
[5] B. R.F, S. R.C, D.M. S, N. K.F, H. Ashit, P. M.A, R. J.V, Prospective comparison of eubacterial PCR and measurement of procalcitonin levels with blood culture for diagnosing septicemia in intensive care unit patients, J. Clin. Microbiol. 47 (2009) 2964–2969.
[6] S.N. I, E. W.R, B. W.S, M. Richard, W. B.D, Who needs a blood culture? A prospectively derived and validated prediction rule, J. Emerg. Med. 35 (2008) 255–264.
[7] H. Erica, P. Payal, L. W.L, C. Anna, A. F.S, C. Vineet, A model to predict central-line–associated bloodstream infection among patients with peripherally inserted central catheters: the MPC score, Infect. Control Hosp. Epidemiol. 38 (2017) 1155–1166.
[8] R. Keyvan, G. Anurag, B. Gina, H. Jana, C. Jacob, M. Qingqing, D. Ritankar, Early prediction of central line associated bloodstream infection using machine learning, Am. J. Infect. Control 50 (2022) 440–445.
[9] T. Pc1, S. Cm, G. Cc, M. St, M. Ra, Vascular access in haemodialysis patients: a modifiable risk factor for bacteraemia and death, QJM: Int. J. Med. 100 (2007) 415–422.
[10] M. Fysaraki, G. Samonis, A. Valachis, E. Daphnis, D.E. Karageorgopoulos, M.E. Falagas, K. Stylianou, D.P. Kofteridis, Incidence, clinical, microbiological features and outcome of bloodstream infections in patients undergoing hemodialysis, Int. J. Med. Sci. 10 (2013) 1632–1638, https://doi.org/10.7150/ijms.6710.
[11] D. Fram, M.F.P. Okuno, M. Taminato, V. Ponzio, S.R. Manfredi, C. Grothe, A. Belasco, R. Sesso, D. Barbosa, Risk factors for bloodstream infection in patients at a Brazilian hemodialysis center: a case–control study, BMC Infect. Dis. 15 (2015), https://doi.org/10.1186/s12879-015-0907-y.
[12] S. Sho, H. Takeshi, K. Hiroo, N. Atsushi, U. Daisuke, I. Takahiro, F. Masahide, N. Hiroki, F. Shingo, S. Yugo, F. Shunichi, Development and validation of a clinical prediction rule for bacteremia among maintenance hemodialysis patients in outpatient settings, PLoS One 12 (2017), e0169975, https://doi.org/10.1371/journal.pone.0169975.
[13] J. Fei, J. Yong, Z. Hui, D. Yi, L. Hao, M. Sufeng, W. Yilong, D. Qiang, S. Haipeng, W. Yongjun, Artificial intelligence in healthcare: past, present and future, Stroke Vas. Neur. 2 (2017).
[14] R. Alvin, D. Jeffrey, K. Isaac, Machine learning in medicine, N. Engl. J. Med. 380 (2019) 1347–1358.
[15] J. H.R, A. George, Forecasting: Principles and Practice, 2018.
[16] F. Shu-Ling, S. M.N, L. John, G. R.D, Diagnosing sepsis–the role of laboratory medicine, Clin. Chim. Acta 460 (2016) 203–210.
[17] B. Matteo, R. Alessandro, R. Elda, D. Elisabetta, M. Maria, D. Federica, S. Assunta, C. Francesco, Role of procalcitonin in bacteremic patients and its potential use in predicting infection etiology, Expert Rev. Anti-Inf. Ther. 17 (2019) 99–105.
[18] J. Jiawei, L. Rui, Y. Xin, Y. Rui, X. Hua, M. Zhi, W. Yongqiang, The neutrophil-lymphocyte count ratio as a diagnostic marker for bacteraemia: a systematic review and meta-analysis, Am. J. Emerg. Med. 37 (2019) 1482–1489.
[19] D. R.C, P. Arun, J. G.H, S. B.N, S. Philipp, J.C.P.C. de, M. L.A.J, I. M.G, B.J. Kenneth, The utility of peripheral blood leucocyte ratios as biomarkers in infectious diseases: a systematic review and meta-analysis, J. Infect. 78 (2019) 339–348, https://doi.org/10.1016/j.jinf.2019.02.006.
[20] J. W.C, Hypoalbuminemia as surrogate and culprit of infections, Int. J. Mol. Sci. 22 (2021) 4496.
[21] K. Lalathaksha, Y. Jerry, Current concepts in hemodialysis vascular access infections, Adv. Chron. Kidney Dis. 26 (2019) 16–22.
[22] S. Sho, R. Yoshihiko, M. Minoru, Y. Shungo, T. Kentaro, H. Takeshi, F. Kiichiro, F. Shunichi, Added value of clinical prediction rules for bacteremia in hemodialysis patients: an external validation study, PLoS One 16 (2021), e0247624.
[23] T. Luis, Data Mining with R: Learning with Case Studies, 2011.
[24] M. Lukas, V.D.G. Sara, B. Peter, The group lasso for logistic regression, J. Roy. Stat. Soc. B 70 (2008) 53–71.
[25] Y. Bin, Q. Wenying, C. Cheng, M. Anjun, J. Jing, Z. Hongyan, M. Qin, SubMito-XGBoost: predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting, Bioinformatics 36 (2020) 1074–1081.
[26] L. Yingchang, L. Mingyang, L. Chao, L. Zhenzhen, Forest aboveground biomass estimation using Landsat 8 and Sentinel-1A data with machine learning algorithms, Sci. Rep. 10 (2020) 1–12.
[27] M. L.S, L. Su-In, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. (2017) 30.
[28] K. J.M, M. Julie, H.A.M. Sw, E.-E. Svend, S. Simon, K. Hans, S. H. C, I. S.N, Prediction of bacteremia in the emergency department: an external validation of a clinical decision rule, Eur. J. Emerg. Med. 23 (2016) 44–49.
[29] H.L. Eliot, D. Nicholas, V. Richard, D. D.B, G. F.L, An external validation study of a clinical prediction rule for medical patients with suspected bacteraemia, Emerg. Med. J. 33 (2016) 124–129.
[30] C. Tianqi, G. Carlos, Xgboost: a scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, n.d.: pp. 785–794.
[31] E. L.C, S. H.T, L. Timmy, S. Surendra, Y.A. S, A. Kenneth, A. Michael, A. Arif, C. A.B, H. G.M, KDOQI clinical practice guideline for vascular access: 2019 update, Am. J. Kidney Dis. 75 (2020). S1–S164.
[32] H.I. Feldman, S. Kobrin, A. Wasserstein, Hemodialysis vascular access morbidity, J. Am. Soc. Nephrol. 7 (1996) 523–535, https://doi.org/10.1681/asn.v74523.
[33] A. S.C, C. A.R, W. W.A, J. D.J, J. B.E, M. J.B, T. M.J, A. N.K, Outcomes of arteriovenous fistula creation after the fistula first initiative, Clin. J. Am. Soc. Nephrol. 6 (2011) 1996–2002.

[34] E.W. Young, D.A. Goodkin, D.L. Mapes, F.K. Port, M.L. Keen, K. Chen, B.L. Maroni, R.A. Wolfe, P.J. Held, The dialysis outcomes and practice patterns study (DOPPS): an international hemodialysis study, Kidney Int. 57 (2000) S74, https://doi.org/10.1046/j.1523-1755.2000.07413.x. –S81.
[35] R.B. Fissell, D.S. Fuller, H. Morgenstern, B.W. Gillespie, D.C. Mendelssohn, H.C. Rayner, B.M. Robinson, D. Schatell, H. Kawanishi, R.L. Pisoni, Hemodialysis patient preference for type of vascular access: variation and predictors across countries in the DOPPS, J. Vasc. Access 14 (2013) 264–272, https://doi.org/10.5301/jva.5000140.
[36] A. Marcel, B.C.G.D.R. Josette, C. H, J. G, High serum procalcitonin concentrations in patients with sepsis and infection, Lancet 341 (1993) 515–518.
[37] B. Mitra, A. Faranak, B.M. Ali, T.M. Satarzadeh, S.A. Reza, K. Hamid, Comparison of WBC, ESR, CRP and PCT serum levels in septic and non-septic burn cases, Burns 34 (2008) 770–774.
[38] R. Franz, S. Michael, E. Katherina, T. Irene, B. Marlene, H. Helmuth, M. Dieter, B. Michael, B. Heinz, Utility of sepsis biomarkers and the infection probability score to discriminate sepsis and systemic inflammatory response syndrome in standard care patients, PLoS One 8 (2013), e82946.
[39] D. Raluca, S. Dimitrie, H. Simona, M. Irina, C. Adrian, Procalcitonin: diagnostic value in systemic infections in chronic kidney disease or renal transplant patients, Int. Urol. Nephrol. 46 (2014) 461–468.
[40] G.P. Falcão, F.L. Menezes, P.I. Duque, Procalcitonin as biomarker of infection: implications for evaluation and treatment, Am. J. Therapeut. 24 (2017) e243–e249.
[41] C.C. for Disease, Prevention, Reducing bloodstream infections in an outpatient hemodialysis center–New Jersey, 2008-2011, MMWR. Morbid. Mort. Weekly Report 61 (2012) 169–173.