scientific reports

Check for updates

OPEN

Predicting responsiveness to fixed-dose methylene blue in adult patients with septic shock using interpretable machine learning: a retrospective study

Shasha Xue[1,4], Li Li[2,3,4], Zhuolun Liu[1], Feng Lyu[1], Fan Wu[1], Panxiao Shi[1], Yongmin Zhang[1]✉, Lina Zhang[2,3]✉ & Zhaoxin Qian[2,3]

This study aimed to develop an interpretable machine learning model to predict methylene blue (MB) responsiveness in adult patients with refractory septic shock and to identify key factors influencing MB responsiveness using the SHapley Additive exPlanations (SHAP) approach. We retrospectively analyzed data from 416 adult patients with refractory septic shock who received MB treatment at Xiangya Hospital of Central South University between June 2018 and October 2022. MB responders were defined as patients who, within 6 hours after MB administration, exhibited either a reduction in average norepinephrine equivalence (NEE) of $\geq$ 10% or an increase in mean arterial pressure of $\geq$ 10 mmHg without an associated increase in NEE. The incidence of MB responders was 38.2%(n=159). Statistical and machine learning methods were used for feature selection, resulting in two datasets (ST and ML). Each dataset was randomly divided into a training set (75%) for model development and a testing set (25%) for internal validation. Prediction models were developed using logistic regression, support vector machine (SVM), random forest, light gradient boosting machine (LightGBM), and explainable boosting machine (EBM). The models were evaluated regarding discrimination, calibration, and clinical benefit. The SVM model trained on the ML dataset demonstrated the best predictive performance, with an area under the curve (AUC) of 0.74 (95% CI 0.62–0.84), 76% accuracy, 36% sensitivity, and 94% specificity. Although the model's sensitivity was low, its high specificity and the safety profile of MB underscore its clinical relevance. The model showed superior net benefit within a 24–85% threshold probability, as determined by decision curve analysis. The SHAP analysis identified the average NEE dose within 6 hours before MB initiation as the most important factor influencing MB responsiveness (P<0.01), with higher doses positively correlating with a greater likelihood of response. Lactate levels were identified as the second most important factor. The optimal model was externally validated in an independent cohort from the same institution, achieving an AUC of 0.75 and an accuracy of 74%.

Keywords Septic shock, Methylene blue, Vasopressor, Interpretable machine learning, SHapley additive exPlanations

Septic shock is the leading cause of mortality in intensive care unit (ICU) patients[1,2], and vasoplegia is one of the major pathophysiological features of the disease. The combination of vasopressors with different mechanisms to suppress the use of catecholamines has been highlighted in this field[3–5]. Methylene blue (MB) is a selective inhibitor of nitric oxide synthase and acts as a vasopressor by partially blocking the vasodilatory effect of nitric oxide. MB has the advantages of safety[6], inexpensiveness, and easy availability, especially in developing countries. Several meta-analyses recently demonstrated that MB significantly increased blood pressure in patients with vasoplegic shock of various etiologies, decreased the use of norepinephrine (NE), and

[1]School of Computer Science and Engineering, Central South University, Changsha 410083, China. [2]Department of Critical Care Medicine, Xiangya Hospital, Central South University, Changsha 410008, China. [3]National Clinical Research Center for Geriatric Disorders, Xiangya Hospital, Changsha 410008, China. [4]Shasha Xue and Li Li have contributed equally to this work. ✉email: zhangyongmin@csu.edu.cn; zln7095@csu.edu.cn

even improved patient outcomes[7–10]. However, these conclusions remain controversial due to the limitations of existing studies, including small sample sizes, heterogeneity in MB application methods, varying inclusion criteria, and inconsistent outcome measures.

Most importantly, it is unknown whether differences in application methods, such as dosage, timing, and population, affect the effectiveness and benefits of MB. In a retrospective, observational study, the rate of MB responsiveness was reported to be approximately 40% among patients with refractory vasodilatory shock[11]. With this in mind, identifying MB responders and influencing factors is a critical challenge that needs to be addressed.

Machine learning (ML) has emerged as a powerful tool in clinical medicine for predicting treatment effectiveness in advantage of its ability to process complex, non-linear relationships and interactions within large, high-dimensional datasets[12,13]. Compared to traditional statistical methods, ML can better capture variability in treatment response by learning patterns from multifaceted clinical data. This approach has the potential to reduce ineffective use, prevent unnecessary risks, and increase the effectiveness of a given therapy. To enhance the interpretability of our machine learning model, we utilized SHapley Additive exPlanations (SHAP), an advanced framework designed to elucidate the contribution of each variable to the model's predictions. By offering insights into the valuable predictors of MB responsiveness, SHAP ensures model transparency, which is critical for clinical acceptance and application. Together, these methods provide a robust and interpretable solution to address variability in MB response.

To the best of our knowledge, no articles have yet been published using interpretable models to predict the effectiveness of MB as a vasopressor in patients with septic shock. This study conducted a retrospective study based on the sepsis-specific clinical database of Xiangya Hospital of Central South University. Patients were categorized into two groups according to their responsiveness to MB drugs: MB responder group and MB non-responder group. A multiple feature selection method was used for feature optimization, a variety of machine learning methods were used to build a prediction model, and the SHAP approach was introduced to interpret and analyze the optimal model established to explore the key factors affecting the response of septic shock patients to MB. To assess the optimal model's robustness and generalizability, external validation was performed using independent patient cohorts from the same institution gathered during distinct time periods.

Methods

This study aimed to evaluate the responsiveness to methylene blue in patients with septic shock through the application of machine learning techniques. The overall study design is summarized in the flowchart presented in Fig. 1.

Study design

We conducted a retrospective study on adult septic shock patients who received MB treatment for vasopressor support at Xiangya Hospital of Central South University from June 2018 to October 2022. The data were sourced from the hospital's Sepsis-Specific Clinical Database, which includes cases from seven independent adult ICUs. Xiangya Hospital is a tertiary teaching hospital in Central and Southern China. This study was approved by the

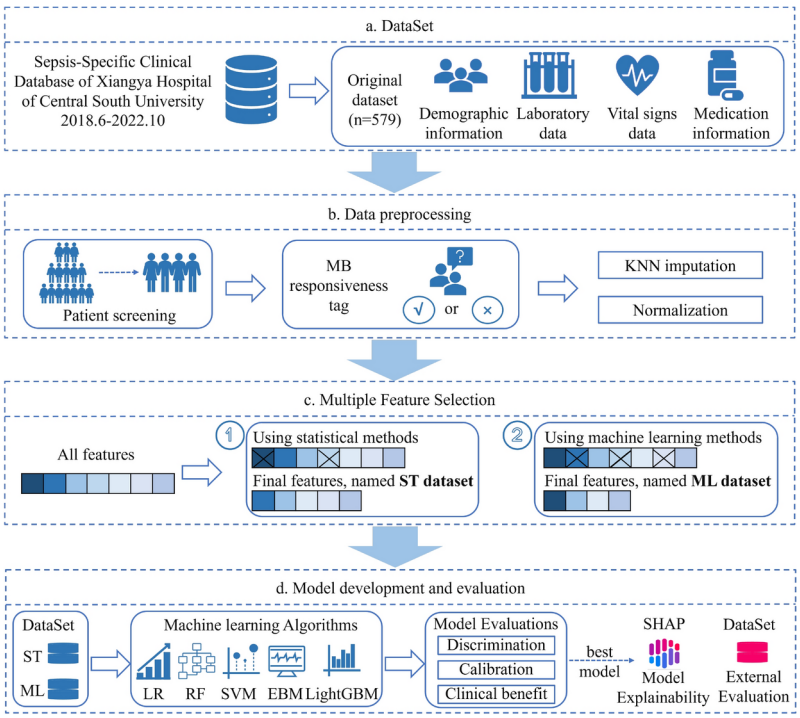

Fig. 1. The integral flowchart of the whole study.

Medical Ethics Committee of Xiangya Hospital of Central South University, and the requirement for written informed consent was waived because all personal information was anonymized. All research methods were strictly carried out following relevant guidelines and regulations.

Study population

Since June 2018, ICU physicians at Xiangya Hospital of Central South University have incorporated a bolus infusion of 2 mg/kg MB as adjunctive vasopressor therapy for septic shock patients. This approach is implemented when the norepinephrine equivalent (NEE) dose exceeds 0.5 mcg/kg/min for two hours following adequate fluid resuscitation. MB (2 mg/kg) was dissolved in 100 ml of sterile saline, and infused intravenously over 20–30 minutes. The physicians in charge determined the timing and population of MB administration.

The inclusion criteria were as follows: (1) confirmed diagnosis of septic shock[14,15]; (2) age $\geq$ 18 years; and (3) first use of MB as a second-line vasopressor therapy (at least 6 hours after NE initiation). The electronic health record was reviewed to determine whether patients met Sepsis-3 criteria for septic shock within the first 24 hours of ICU admission: (1) evidence of a suspected infection, defined as the combination of administration of antibiotics (oral or parenteral) and a body fluid culture specimen obtained (blood, urine, or cerebrospinal fluid); (2) presence of organ dysfunction, defined as 2 or more Sequential Organ Failure Assessment (SOFA) points; (3) requiring vasopressor therapy to maintain a mean arterial pressure (MAP) of at least 65 mmHg and a serum lactate level exceeding 2 mmol/L, despite adequate volume resuscitation. The exclusion criteria were as follows: (1) ICU stayed < 24 hours after MB administration; (2) MB for other purposes; (3) repeated use of MB; (4) the first use of MB was less than 6 hours away from the next medication of MB; and (5) inability to assess MB responsiveness due to missing data.

Predictor variables and data processing

The extracted variables included demographic characteristics, comorbidities, medications, vital signs, laboratory results, and supportive therapies such as mechanical ventilation and continuous renal replacement therapy. The hourly pumping doses of vasoactive drugs, including NE, dopamine, and terlipressin, were calculated. The dose and duration of glucocorticoids and MB were also recorded. The time interval between the initiation of NE and MB was defined as T_{NE}, the time interval between the diagnosis of shock and the initiation of MB as T_{shock}, the average dose of NEE 6 hours before MB initiation was defined as NEE_{pre}, and the average NEE 6 hours after MB initiation was defined as NEE_{post}.

The NEE dose was calculated as[16]: NE dose (mcg/kg/min) + 1/100 $\times$ dopamine dose (mcg/kg/min) + 0.06 $\times$ phenylephrine dose (mcg/kg/min) + 10 $\times$ terlipressin dose (mcg/kg/min) + 0.2 $\times$ MB dose (mg/kg/h) + 8 $\times$ metaraminol dose (mcg/kg/min) + 0.02 $\times$ hydroxocobalamin B12 dose (g) + 0.4 $\times$ midodrine dose (mcg/kg/min). The change in NEE(NEE%) was calculated as (NEE_{post} - NEE_{pre}) / NEE_{pre} $\times$ 100%. All patients were divided into two groups: MB responders and MB non-responders. Patients were defined as MB responders if they met either of the following criteria within 6 hours after MB administration: (1) a reduction in the average NEE dose by at least 10% (NEE% < -10%), with MAP $\geq$ 65 mmHg; (2) a stable NEE dose ($-10\% \leq$ NEE% < 10%), but an increase in MAP of $\geq$ 10 mmHg. Patients who did not meet either of these criteria were classified as MB non-responders. These thresholds were established by integrating findings from studies on vasopressin responsiveness[17,18], the effective half-life of methylene blue[5], and clinical significance.

A total of 24 candidate predictor variables were collected. Outliers in continuous variables were identified using the interquartile range (IQR) method and visually inspected using boxplots and histograms. For variables where outliers represented clinically plausible extreme values, these data points were retained. For outliers likely arising from errors, values were corrected or excluded based on clinical plausibility. Variables with a missing value ratio of more than 20% were excluded. The k-nearest neighbor (KNN) imputation method was used to fill in missing values of the remaining variables[19–21]. Specifically, the KNN imputation method combined with the random forest algorithm based on the scikit-learn library was employed for training. Meanwhile, the performance of the KNN imputation method was evaluated based on the root mean square error (RMSE) and the average area under the receiver operating characteristic curve (AUC) (details in Supplementary Fig. S1–S3). The interpolation dataset corresponding to the K value with the best performance was selected as the dataset for subsequent multiple-feature selection.

Feature selection and statistical analysis

Categorical variables were summarized as counts and percentages, denoted as n (%), while continuous variables were reported as means ± standard deviations (SDs) for normally distributed data or as medians with interquartile ranges (IQRs) for non-normally distributed data. All statistical analyses and modeling were conducted using PyCharm software (Python version 3.9.13). Intergroup differences for categorical variables were assessed using the chi-square test, while continuous variables were compared using either the independent sample t-test (for normally distributed data) or the Wilcoxon rank-sum test (for non-normally distributed data). To avoid overfitting of the prediction model and multicollinearity among predictors, a multiple-feature selection approach was adopted (details in Supplementary Table S1 and Supplementary Fig. S4–S6). First, the candidate predictors were selected through univariate analysis, with a threshold for inclusion set at a P value < 0.30[22,23]. The threshold was set higher than the conventional 0.20 to ensure that variables with potential predictive value, even if they showed weaker univariate associations, were not excluded prematurely. This approach was particularly important given the complex and multifactorial nature of septic shock, where interactions among variables may enhance their predictive utility in multivariable modeling. Then, Pearson correlation coefficient tests were performed to explore the relationships between variables, producing the ST dataset (derived from statistical methods). Moreover, various machine learning techniques, including random forest-based recursive feature elimination and CatBoost, have been integrated for feature engineering, resulting in the ML dataset (derived

from machine learning methods)[24,25]. Both the ST and ML datasets were prepared for subsequent modeling. We acknowledge that utilizing the entire dataset for feature selection may inadvertently introduce a risk of data leakage. To minimize this potential issue, we ensured that feature selection was solely for identifying candidate predictors, without model training or testing, thus preventing the direct use of the testing dataset in the selection process.

Machine learning model development and evaluation

The ST and ML datasets were randomly split into a training dataset (75%) for model development and a testing dataset (25%) for internal validation. The 75/25 split was chosen to provide sufficient data for model training while maintaining a sufficient sample size for internal validation to evaluate the reliability of the models. Five popular machine learning classification algorithms, including logistic regression (LR), random forest (RF), explainable boosting machine (EBM), support vector machine (SVM), and light gradient boosting machine (LightGBM), were applied in the model development. Among these, EBM is a glass-box model based on generalized additive models, which demonstrates comparable accuracy to other machine learning models while shedding light on the contribution of each feature to the final prediction[26,27]. The hyperparameters of the models were tuned via a grid search and 5-fold cross-validation on the basis of the training dataset. The 5-fold cross-validation method was selected to balance computational efficiency and performance reliability, allowing for robust evaluation without excessive computational cost. The optimal hyperparameter combinations were determined based on accuracy evaluated on the training data. To minimize the risk of overfitting, we employed several strategies. Logistic regression used L1 and L2 regularization, with the optimal penalty strength determined through grid search and 5-fold cross-validation. For tree-based models such as random forest and LightGBM, built-in regularization techniques were optimized, including tuning parameters like max depth, minimum data in leaves, and feature fraction. Importantly, all training and hyperparameter optimization processes were confined to the training dataset, ensuring that no information from the testing dataset influenced model development.

The model's predictive performance in the testing dataset was preliminarily evaluated using several metrics, including AUC, accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 score. To calculate the 95% confidence intervals (CIs) for these metrics, 1,000 bootstrap resamples were performed. The F1 score, defined as the harmonic mean of precision and recall, was employed as an additional performance metric, particularly useful in the context of class imbalance. The F1 score was calculated as follows:

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1 score played a significant role in this study as it provided a balanced measure of the model's ability to correctly identify responders and non-responders. The calibration of the models was evaluated by plotting calibration curves and calculating the Brier score, which represents the mean squared error of the probabilistic predictions[28,29]. A lower Brier value indicates better calibration. Decision curve analysis (DCA) was conducted to evaluate the clinical utility of the decision models[30]. Statistical tests were used to compare the performance of different models. Specifically, the Friedman test was applied to determine whether significant differences existed in AUC, accuracy, and F1 scores across the machine learning models. If significant differences were observed, a Nemenyi post hoc test was conducted to identify pairwise differences and evaluate model performance. The optimal model was selected based on these analyses, and its robustness and generalizability were further validated using an independent dataset collected from the same institution but spanning a different period. External validation helps evaluate the performance of the model in the time queue and provides additional evidence for its clinical applicability. The machine learning models were built and evaluated with the following software packages: pandas, matplotlib, scikit-learn, interpret, and LightGBM.

Machine learning explainable tool

SHAP analysis is a unified approach for explaining the output of machine learning models[31], elucidating the contribution of each feature to the predicted outcome. All features are considered contributors to SHAP analysis. Each feature receives its own SHAP value, providing explanations at both the local and global levels[32]. Local interpretability focuses on how each feature influences the model's prediction for an individual, and global interpretability focuses on how each feature affects the model predictions across the entire dataset.

Results

Patient characteristics

From June 2018 to October 2022, a total of 579 patients received MB infusions as second-line vasopressors at Xiangya Hospital of Central South University. Ultimately, 416 patients were included for statistical analysis, with 159 patients (38.2%) being responders. The entire process of patient inclusion and variable selection is illustrated in Fig. 2. The majority of patients were from comprehensive ICU (91.6%, 381 cases), so the ICU types were dichotomously categorized into comprehensive ICU and specialty ICU. The latter included neurosurgical, neurologic, respiratory, emergency, cardiothoracic surgical, and cardiovascular medical ICUs. The dataset was randomly divided into training data (75%, 312 cases) and testing data (25%, 104 cases), with MB responders accounting for 40.4% (126/312) and 31.7% (33/104), respectively. After feature engineering, a total of 13 predictive variables were identified. Table 1 presented the relevant variable information of all septic shock patients receiving methylene blue treatment, which we compared into the overall, MB responders group, and

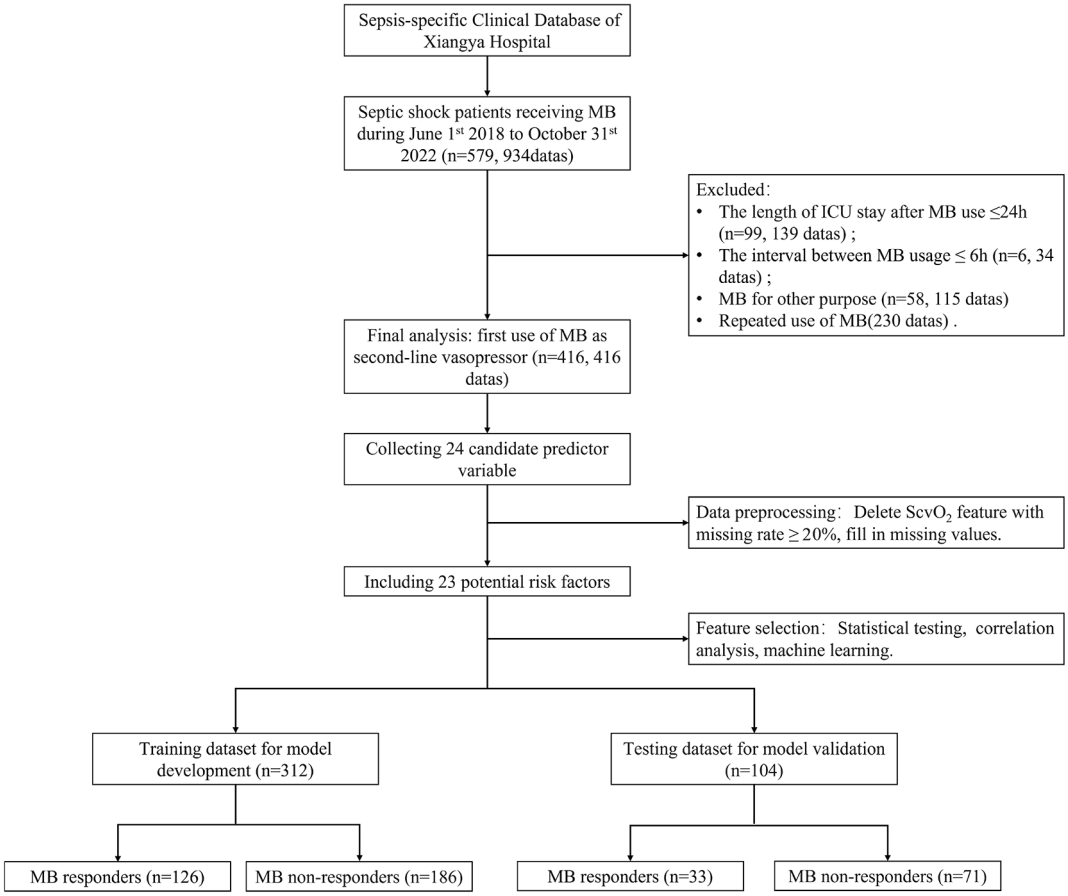

Fig. 2. Flowchart of patient inclusion and variable selection.

MB non-responders group. Compared with non-responders, MB responders had significantly shorter T_{NE} and higher NEE_{pre} doses (p<0.01).

Machine learning model performance

The hyperparameter tuning ranges and the optimal combinations of hyperparameters for each model are detailed in Supplementary Tables S2 and S3. GridSearchCV with 5-fold cross-validation was used for hyperparameter optimization, ensuring robust and reliable model performance. The correlation heatmaps between variables in the ST dataset and ML dataset are shown in Fig. 3. As illustrated in the figures, no multicollinearity issues were found in either dataset. The LR, RF, EBM, SVM, and LightGBM models were established on both training datasets. Figure 4a,b shows that the AUC values of the five models based on the ST testing dataset were 0.69(95% CI 0.57–0.80), 0.68(95% CI 0.57–0.79), 0.70(95% CI 0.58–0.80), 0.74(0.63–0.83), and 0.73(95% CI 0.62–0.83), while the AUC values of the five models based on the ML testing dataset were 0.73(95% CI 0.61–0.83), 0.71(95% CI 0.60–0.82), 0.66(95% CI 0.55–0.78), 0.74(95% CI 0.63–0.84), and 0.71(95% CI 0.60–0.82). Table 2 displays eight evaluation metrics calculated from both testing datasets to preliminarily assess and compare the performance of all the models. From the perspective of the AUC, the SVM model outperformed the other machine learning models, achieving an AUC of 0.74 in both the ST and ML datasets. In contrast, the EBM model in the ML dataset exhibited the poorest generalization ability, with an AUC of 0.66. The F1 score values of the five models based on the ST testing dataset were 0.53(95% CI 0.37–0.68), 0.53(95% CI 0.36–0.68), 0.54(95% CI 0.38–0.68), 0.55(95% CI 0.39–0.70), and 0.53(95% CI 0.36–0.68), while the F1 score values of the five models based on the ML testing dataset were 0.45(95% CI 0.27–0.61), 0.48(95% CI 0.31–0.64), 0.46(95% CI 0.30–0.60), 0.49(95% CI 0.31–0.64), and 0.57(95% CI 0.41–0.71). The F1 scores of the five models varied across the ST and ML datasets, with the EBM model achieving the highest F1 score of 0.57 in the ML dataset and the LR model showing the lowest F1 score of 0.45. Overall, these models demonstrated moderate F1 scores and relatively low sensitivity. We emphasize that this step represents a preliminary screening process. The selection of the final model involves additional evaluation using calibration curves and decision curve analysis to identify the most clinically relevant model.

The calibration performance of the five models was evaluated on both testing datasets to assess the consistency between the predicted and actual probabilities. As shown in Fig. 4c,d, the SVM model exhibited the best agreement with the actual probabilities across all the models in both datasets. This observation was further supported by the Brier score results, which quantify the accuracy of probabilistic predictions. The Brier scores

	Overall	MB responders	MB non-responders	*P* value
Demographic informations				
Count	416	159	257	
$Gender\ female$(%)	273(65.63)	105(66.04)	168(65.37)	0.89
$Age^{*}(years)$ mean(SD)	58.29(15.81)	58.98(15.71)	57.87(15.88)	0.55
$Comprehensive\ ICU^{\#}$ (%)	381(91.59)	152(95.60)	229(89.11)	0.02
$Weight^{*}(kg)$ median(Q_1, Q_3)	59.11(50,63.25)	60(50,65)	60(50,63)	0.48
$Smoking\ history^{\#}$(%)	121(29.09)	40(25.16)	81(31.52)	0.17
$Drinking\ history$(%)	82(19.71)	31(19.50)	51(19.84)	0.93
Vital signs				
$Oxygenation\ index^{*\#}(mmHg)$ median(Q_1, Q_3)	186.98(110,281.89)	195(114.58,281.89)	172.5(109.76,281.4)	0.28
$Lactate^{*}(mmol/L)$ median(Q_1, Q_3)	3.1(1.7,6.2)	3.1(1.89,6.1)	3.1(1.6,6.2)	0.41
Intervention				
CRRT(%)	116(27.88)	46(28.93)	70(27.24)	0.71
$Terlipressin^{\#}$(%)	98(23.56)	42(26.42)	56(21.79)	0.28
Corticosteriod(%)	167(40.14)	61(38.36)	106(41.25)	0.56
$Mechanical\ ventilation^{*\#}$(%)	329(79.09)	131(82.3)	198(77.04)	0.19
Comorbidity				
$Immunosuppressive\ status$(%)	17(4.09)	5(3.14)	12(4.67)	0.45
Hypertension(%)	137(32.93)	52(32.70)	85(33.07)	0.94
Diabetes(%)	67(16.11)	26(16.35)	41(15.95)	0.91
$Coronary\ heart\ disease$(%)	43(10.34)	16(10.06)	27(10.51)	0.89
$Atrial\ fibrillation^{\#}$(%)	13(3.13)	3(1.89)	10(3.89)	0.25
$Pulmonary\ disease$(%)	7(1.68)	2(1.26)	5(1.95)	0.6
$Kidney\ disease$(%)	50(12.02)	19(11.95)	31(12.06)	0.97
$Liver\ disease^{\#}$(%)	32(7.69)	15(9.43)	17(6.61)	0.29
Severity scores				
${T_{shock}}^{*\#}(hours)$ median(Q_1, Q_3)	23.61(7.58,103.52)	20.25(5.9,61.22)	24.95(9.1,133.33)	0.04
${T_{NE}}^{*\#}(hours)$ median(Q_1, Q_3)	16.68(4.68,52.62)	14.13(2.66,39.25)	19.85(6.87,62.63)	<0.01
${NEE_{pre}}^{*\#}(mcg/kg.min)$ median(Q_1, Q_3)	0.65(0.29,1.61)	1.31(0.51,3.03)	0.52(0.22,1.02)	<0.01
SOFA median(Q_1, Q_3)	13(10,17)	13(11,16)	14(10,17)	0.72

Table 1. Major variables for septic shock patients receiving methylene blue. *: variables included in the ML dataset; #: variables included in the ST dataset; CRRT: continuous renal replacement treatment; SOFA: sequential organ failure assessment.

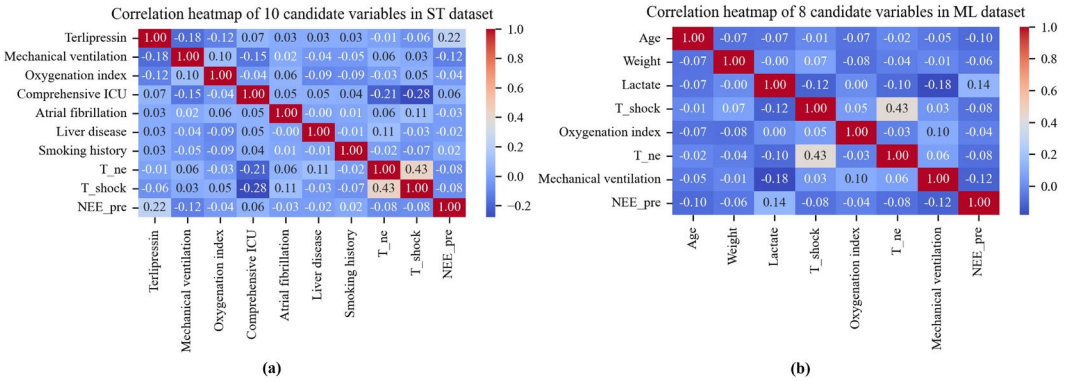

Fig. 3. Correlation matrix among variables selected via statistical (**a**) and machine learning (**b**) methods. (**a**) Correlation heatmap of 10 variables in the ST dataset. (**b**) Correlation heatmap of 8 variables in the ML dataset.

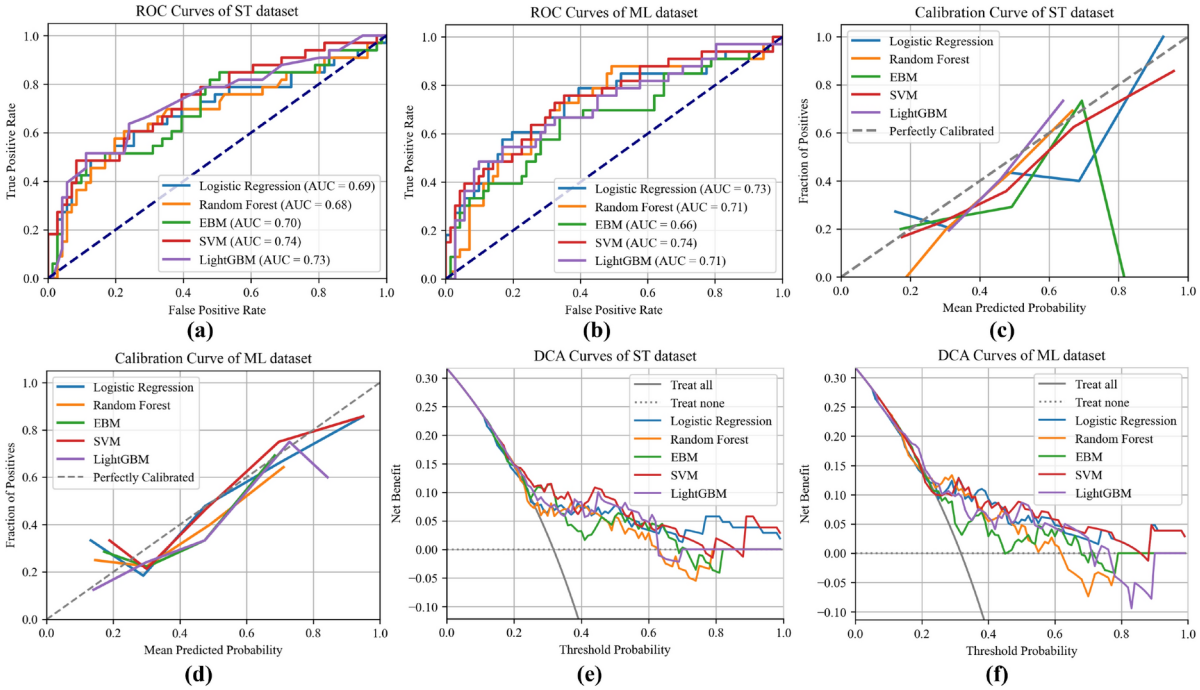

Fig. 4. Performance of each prediction model. (**a**–**b**) Receiver operating characteristic (ROC) curves for the five machine learning models based on the ST dataset (**a**) and the ML dataset (**b**), showing model performance in the test sets. The SVM model demonstrated the best overall performance. (**c**–**d**) Calibration curves for the five machine learning models based on the ST dataset (**c**) and the ML dataset (**d**), illustrating the agreement between predicted and actual risks in the test sets. (**e**–**f**) Decision curve analysis (DCA) for the five machine learning models based on the ST dataset (**e**) and the ML dataset (**f**), showing clinical net benefit across different threshold probabilities in the test sets.

Model	AUC	Accuracy (95%CI)	Specificity (95%CI)	Sensitivity (95%CI)	F1 score	PPV	NPV	BS (95%CI)
ST dataset derived from statistical methods								
LR	0.69	0.76(0.67–0.84)	0.92(0.85–0.97)	0.42(0.37–0.68)	0.53	0.70	0.77	0.19(0.16–0.22)
RF	0.68	0.74(0.65–0.82)	0.87(0.80–0.95)	0.45(0.29–0.62)	0.53	0.63	0.78	0.20(0.17–0.23)
EBM	0.70	0.74(0.65–0.83)	0.86(0.77–0.93)	0.48(0.32–0.65)	0.54	0.61	0.78	0.20(0.17–0.23)
SVM	0.74	0.76(0.68–0.85)	0.92(0.85–0.97)	0.42(0.29–0.62)	0.53	0.70	0.77	0.18(0.15–0.21)
LightGBM	0.73	0.76(0.67–0.84)	0.92(0.85–0.97)	0.42(0.27–0.59)	0.53	0.70	0.77	0.19(0.16–0.22)
ML dataset derived from machine learning methods								
LR	0.73	0.74(0.65–0.83)	0.93(0.87–0.99)	0.33(0.19–0.50)	0.45	0.69	0.75	0.19(0.15–0.22)
RF	0.71	0.73(0.64–0.81)	0.89(0.82–0.96)	0.39(0.24–0.56)	0.48	0.62	0.76	0.20(0.16–0.23)
EBM	0.66	0.70(0.62–0.79)	0.85(0.76–0.93)	0.39(0.25–0.57)	0.46	0.54	0.75	0.20(0.17–0.23)
SVM	0.74	0.76(0.67–0.84)	0.94(0.89–0.99)	0.36(0.21–0.52)	0.49	0.75	0.76	0.18(0.15–0.21)
LightGBM	0.71	0.77(0.69–0.85)	0.90(0.83–0.96)	0.48(0.32–0.66)	0.57	0.70	0.79	0.19(0.16–0.23)

Table 2. Performance of each prediction model.

for the LR, RF, EBM, SVM, and LightGBM models were 0.19, 0.20, 0.20, 0.18, and 0.19 in the ST dataset, and 0.19, 0.20, 0.20, 0.18, and 0.19 in the ML dataset, respectively. The SVM model consistently achieved the lowest Brier score in both datasets, indicating its superior calibration performance. Additionally, DCA was performed to assess the net benefit of these machine learning models in the testing datasets. The DCA curve measures the net benefit at different threshold probabilities. As illustrated in Fig. 4e,f, the solid gray line represents the assumption that all patients received MB medication, while the dashed gray line represents the assumption that no patients received MB. In both testing datasets, each machine learning model outperformed the strategy of treating all patients or no patients within a probability threshold of 26%-61%. In the ML dataset, the net benefit of the SVM model surpassed that of the other four models across a wider range of probabilities (24%-85%,Fig. 4f).

To compare the performance of five models across two datasets, we conducted statistical analyses focusing on AUC, F1 score, and Brier score. The Friedman test revealed statistically significant differences in at least

two models for each metric (AUC, F1 score, and Brier score) in both the ST and ML datasets, with $p < 0.01$. To further investigate these differences, Nemenyi post hoc tests were performed to evaluate pairwise comparisons between the models. The results, presented in Supplementary Fig. S7, demonstrate that the SVM model trained on the ML dataset outperformed other models across most metrics, with statistically significant differences in performance. Based on these findings, the SVM model built on the ML dataset was selected as the optimal model with AUC of 0.74, 76% accuracy (95% CI: 67–84 %), 94% specificity (95% CI: 89–99 %) and 36% sensitivity (95% CI: 21–52 %).

The external validation dataset was constructed by collecting data from septic shock patients who received MB treatment at the same institution between November 2022 and November 2023. A total of 62 patients were included, of whom 24 were classified as MB responders. Comprehensive information on all the modeling variables for this cohort is provided in Supplementary Table S4. The external validation of the optimal model demonstrated an AUC of 0.75 (95% CI: 0.61–0.88), an accuracy of 74%, a recall of 42%, a specificity of 95%, and an F1 score of 0.56. The receiver operating characteristic (ROC) curve with the 95% CI is depicted in Supplementary Fig. S8. These results underscore the optimal model's robustness and clinical applicability.

Explanation of the SVM model in the ML dataset with the SHAP method

The SHAP algorithm was employed to evaluate the contribution of each variable to the prediction outcome based on the SVM model in the ML dataset. Figure 5a showed the predictive power of the variables for the outcome, ranked in descending order of importance. NEE_{pre} had the strongest predictive value across all prediction ranges, followed by the level of lactate and body weight. Figure 5b listed the SHAP values of all the predictor variables on the horizontal location, where positive and negative values corresponded to positive and negative effects on the prediction outcome. The magnitude of the variable's value was coded with color; red for higher values and blue for lower values. Patients with higher NEE_{pre} values were more likely to respond to MB, whereas those with lower NEE_{pre} values were less likely. Patients with higher levels of lactate were less likely to be MB responders, and heavy patients. We cannot exclude the possibility of dose-related effectiveness of MB, since we administered a fixed dose of 2 mg/kg.

SHAP individual force plots

Figure 6 illustrated individual force plots for an MB responder and an MB non-responder respectively. The bold number is the predicted probability (f(x), model output value), and the baseline represents the predicted value when no variable is input to the model. Red poles indicate a high likelihood of MB responsiveness, while blue poles indicate a low likelihood. The length of the poles corresponds to the extent of the influence on the prediction, a longer length indicates greater impact.

Discussion

Our retrospective cohort study revealed that responsiveness to the first use of 2 mg/kg MB was 38.2% in refractory septic shock patients. The SVM prediction model demonstrated moderate accuracy, with AUC of 0.74, , showing consistent accuracy across both the test dataset (AUC 0.74) and the external validation dataset (AUC 0.75). The interpretable SHAP method demonstrated that NEE_{pre} was the most important factor influencing MB responsiveness, followed by the level of lactate.

The response rate of MB in our study aligns with that reported in a retrospective study conducted at Massachusetts General Hospital, in which 39.5% was reported in vasodilatory shock patients[11]. This finding indicates that the mechanism of refractory septic shock is likely highly heterogeneous, considering that MB exerts a vasocontrictive effect only on patients with elevated nitric oxide[33]. Patients without nitric oxide overexpression do not respond to MB. The type of second-line vasopressor has not reached a consensus, although early addition of non-catecholamine drugs is recommended by the latest sepsis guidelines[34]. Today, rather than suggesting a fixed type of drug, individualizing a medication regimen according to specific factors is preferable[35]. For

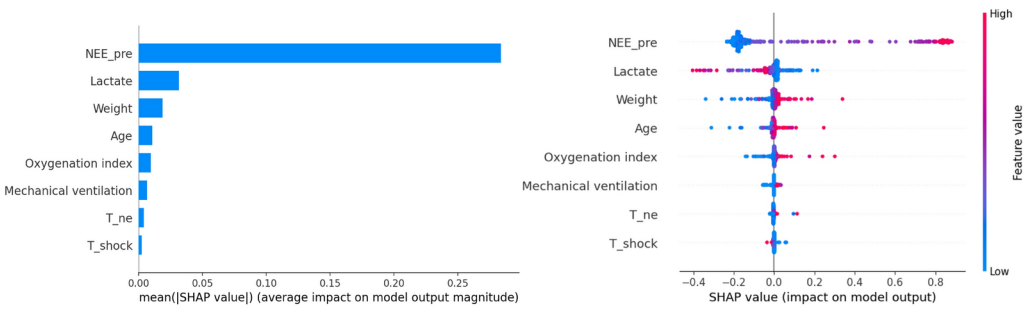

Fig. 5. Importance ranking of variables in the ML dataset and their corresponding SHAP values, highlighting the key factors that influence model predictions. (**a**) Importance ranking of the top 8 variables in the ML dataset, with the X-axis representing the importance score derived from the mean absolute SHAP value and the Y-axis listing the variables. (**b**) SHAP value plot showing the contribution of each variable to the model predictions in the ML dataset, with positive and negative impacts shown on the X-axis.

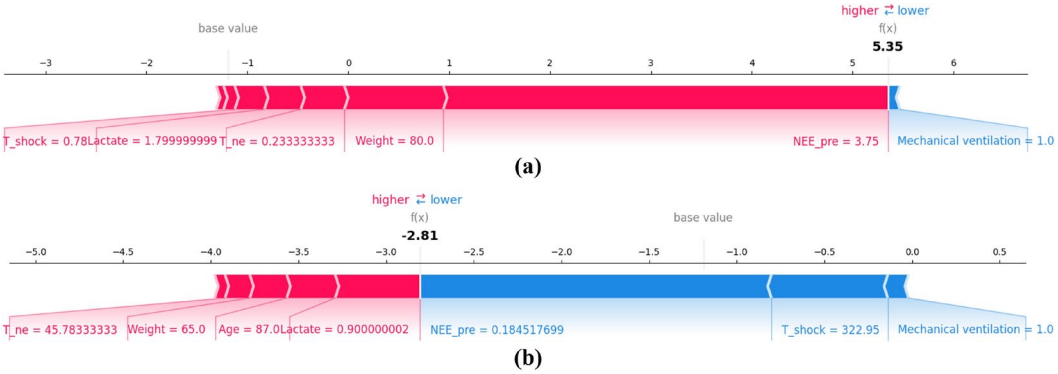

Fig. 6. Shapley Additive exPlanations (SHAP) force plot for two selected patients. (**a**) MB responder. (**b**) MB non-responder.

example, in vasodilatory shock patients, angiotensin II improved the clinical outcomes of those with high renin concentrations[36]. Notably, most types of vasopressins (except terlipressin) and angiotensin II are not available in mainland China. Unlike terlipressin, which is available in some tertiary hospitals, MB is easy to access even in primary hospitals. Our results have practical value in assisting in choosing the proper population and timing of MB initiation.

A general idea about the indication for initiating second-line vasopressors is the threshold dose of norepinephrine[34–36]. Vasopressin is the most investigated drug. Current guidelines recommend initiating vasopressin when the dose of norepinephrine reaches 0.25–0.5 mcg/kg/min[34]. Ammar suggested starting vasopressin at a norepinephrine dose of 0.1–0.2 mcg/kg/min (10–15 mcg/min)[37] and avoiding vasopressin at rates greater than 0.3 mcg/kg/min (>25 mcg/min). The tendency to add vasopression early in septic shock does not seem applicable to MB administration, since a higher NEE_{pre} correlated with an increased likelihood of MB responsiveness. Considering that the norepinephrine dose corresponds to the extent of vasoplegia, we supposed that NO participated in the mechanism of more severe vasodilatory shock. This was well proven by the fact that MB is usually considered a rescue vasopressor therapy and works effectively in cases and serial reports[38].

To the best of our knowledge, no study has used interpretable models to predict MB responsiveness in patients with septic shock. In the field of critical care research, SVM models have been widely utilized to predict the occurrence of various diseases and are considered valuable clinical decision support tools, assisting clinicians in making more accurate diagnoses and treatment decisions[39,40]. The strengths of our study include the use of SVM models to predict MB responsiveness in septic shock patients for the first time and the comparison with other models, coupled with SHAP analysis for model interpretation. While deep learning approaches, such as recurrent neural networks (RNNs) and transformers, have demonstrated potential in capturing complex temporal dependencies in physiological data[41,42], they pose significant challenges. These include high computational demands and limited transparency in their decision-making processes, which can hinder clinical adoption. In contrast, our study prioritizes clinical applicability by employing interpretable models like SVM coupled with SHAP analysis. This approach not only provides robust predictions but also ensures that clinicians can understand and trust the underlying reasoning, thereby facilitating more informed and reliable clinical decisions.

We developed and validated 5 machine learning models using two datasets derived from statistical and feature engineering methods. The best-performing SVM model surpassed a comparable machine learning study that aimed to predict vasopressin responsiveness in septic shock patients, which reported modest discrimination with AUCs of 0.59–0.61 in the training set and 0.64–0.68 in the external validation set[43]. Although the low sensitivity of our model may appear to be a limitation, it holds clinical relevance and can be addressed with further refinements. Clinically, low sensitivity may result in some MB responders being missed by the model. However, given MB's safety profile and accessibility, physicians can still consider its use as a rescue therapy for critically ill patients without contraindications, even if the model predicts non-responsiveness. Conversely, the model's high specificity ensures that patients predicted to be responders are highly likely to benefit from MB treatment, supporting its utility in guiding therapeutic decisions.

The inherent limitations of retrospective data and the heterogeneous nature of septic shock pose significant challenges in developing a highly accurate predictive model. At the time of MB initiation, patients are often in critical condition, potentially beyond the responsiveness to any vasopressor. Moreover, the blood pressure threshold of 10 mmHg, used in this study to define MB responsiveness, exceeds the typical blood pressure improvement range of 3–8 mmHg observed after MB administration, as demonstrated by recent meta-analyses of randomized controlled trials[8,9]. Despite these limitations, this study leverages the largest retrospective cohort to date and provides valuable insights to inform future prospective randomized controlled trials. Specifically, it highlights the need to refine initiation criteria for MB as a second-line vasopressor in septic shock management. However, we acknowledge the need to further enhance the model's sensitivity to capture a larger proportion of responders. Future work could involve exploring alternative models, such as ensemble methods or deep learning

architectures, which may better handle the heterogeneity of septic shock mechanisms. Furthermore, multicenter validation using external datasets will be conducted to enhance the generalizability and robustness of the model.

The present study had several limitations. First, we obtained data from a single center and lacked external validation, however, the sample size was the largest yet applied to MB in septic shock[9]. Second, MB was used at a fixed dose in the form of a bolus infusion. In fact, the dose and duration of MB vary across studies. A randomized controlled study applied a 3 mg/kg bolus for 3 days[44] or continuous infusion of maintenance medication[45,46]. We did not include repeated use of MB to avoid interference caused by time intervals and duplicate medications. Therefore, we cannot exclude the positive correlation between the dosage or duration and the efficacy of MB[47]. Third, meta-analyses of observational and RCT studies on methylene blue (MB) as a vasopressor have confirmed its ability to increase blood pressure by approximately 5–6 mmHg and reduce catecholamine use in patients with distributive shock[7–9]. However, the absolute reduction in vasopressor requirements remains undefined due to variations in calculation methods across studies. The responder threshold we selected exceeds this typical range, and the 6-hour observation window before and after administration may dilute the proportion of identified MB responders. We believe the higher threshold and extended time window improve clinical relevance. Future research using deep learning on time-series data could further enhance the reliability and applicability of these findings.

Conclusion

In conclusion, this study indicates that the SVM machine learning model demonstrates good discrimination in predicting MB non-responders among septic shock patients. A higher NEE dose before MB initiation, which was the most important influencing factor, represented a greater likelihood of responding to MB. The interpretable SHAP method is helpful for helping physicians understand pathophysiological mechanisms and make clinical decisions.

Data Availability

The datasets used and/or analyzed during this study are not publicly available due to the confidentiality policy of the National Health Commission of China but are available from the corresponding author Lina Zhang upon reasonable request.

Received: 27 August 2024; Accepted: 10 February 2025
Published online: 01 March 2025

References

1. De Backer, D. et al. A plea for personalization of the hemodynamic management of septic shock. *Crit. Care* **26**, 372. https://doi.org/10.1186/s13054-022-04255-y (2022).
2. Meyhoff, T. S. et al. Restriction of intravenous fluid in icu patients with septic shock. *N. Engl. J. Med.* **386**, 2459–2470. https://doi.org/10.1056/NEJMoa2202707 (2022).
3. Zhong, J. et al. Association between maximum norepinephrine dosage and mortality risk in neonates with septic shock. *Sci. Rep.* **14**, 1–6. https://doi.org/10.1038/s41598-024-65744-4 (2024).
4. Teja, B., Bosch, N. A. & Walkey, A. J. How we escalate vasopressor and corticosteroid therapy in patients with septic shock. *Chest* **163**, 567–574. https://doi.org/10.1016/j.chest.2022.09.019 (2023).
5. Sacha, G. L. & Bauer, S. R. Optimizing vasopressin use and initiation timing in septic shock: A narrative review. *Chest* https://doi.org/10.1016/j.chest.2023.07.009 (2023).
6. Xue, H., Thaivalappil, A. & Cao, K. The potentials of methylene blue as an anti-aging drug. *Cells* **10**, 3379. https://doi.org/10.3390/cells1012337 (2021).
7. Zhao, C.-C. et al. Efficacy and safety of methylene blue in patients with vasodilatory shock: A systematic review and meta-analysis. *Front. Med.* **9**, 950596. https://doi.org/10.3389/fmed.2022.950596 (2022).
8. Pruna, A. et al. Methylene blue reduces mortality in critically ill and perioperative patients: A meta-analysis of randomized trials. *J. Cardiothorac. Vasc. Anesth.* https://doi.org/10.1053/j.jvca.2023.09.037 (2023).
9. Huang, X., Yan, W., Chen, Z. & Qian, Y. Effect of methylene blue on outcomes in patients with distributive shock: A meta-analysis of randomised controlled trials. *BMJ Open* **14**, e080065. https://doi.org/10.1136/bmjopen-2023-080065 (2024).
10. Porizka, M. et al. Methylene blue administration in patients with refractory distributive shock-a retrospective study. *Sci. Rep.* **10**, 1828. https://doi.org/10.1038/s41598-020-58828-4 (2020).
11. Naoum, E. E., Dalia, A. A., Roberts, R. J., Devine, L. T. & Ortoleva, J. Methylene blue for vasodilatory shock in the intensive care unit: A retrospective, observational study. *BMC Anesthesiol.* **22**, 199. https://doi.org/10.1186/s12871-022-01739-w (2022).
12. Kim, Y. J. et al. Machine learning-based model to predict delirium in patients with advanced cancer treated with palliative care: A multicenter, patient-based registry cohort. *Sci. Rep.* **14**, 11503. https://doi.org/10.1038/s41598-024-61627-w (2024).
13. Ghosh, S. K. & Khandoker, A. H. Investigation on explainable machine learning models to predict chronic kidney diseases. *Sci. Rep.* **14**, 3687. https://doi.org/10.1038/s41598-024-54375-4 (2024).
14. Rhodes, A. et al. Surviving sepsis campaign: International guidelines for management of sepsis and septic shock: 2016. *Intens. Care Med.* **43**, 304–377. https://doi.org/10.1007/s00134-017-4683-6 (2017).
15. Ahn, Y. H. et al. Association between the timing of icu admission and mortality in patients with hospital-onset sepsis: A nationwide prospective cohort study. *J. Intens. Care* **11**, 16. https://doi.org/10.1186/s40560-023-00663-6 (2023).
16. Kotani, Y., Di Gioia, A., Landoni, G., Belletti, A. & Khanna, A. K. An updated "norepinephrine equivalent" score in intensive care as a marker of shock severity. *Crit. Care* **27**, 29. https://doi.org/10.1186/s13054-023-04322-y (2023).
17. Sacha, G. L. et al. Predictors of response to fixed-dose vasopressin in adult patients with septic shock. *Ann. Intens. Care* **8**, 1–10. https://doi.org/10.1186/s13613-018-0379-5 (2018).
18. Permpikul, C. et al. Early use of norepinephrine in septic shock resuscitation (censer). a randomized trial. *Am. J. Resp. Crit. Care Med.* **199**, 1097–1105. https://doi.org/10.1164/rccm.201806-1034OC (2019).
19. Huang, M.-W., Tsai, C.-F., Tsui, S.-C. & Lin, W.-C. Combining data discretization and missing value imputation for incomplete medical datasets. *PLoS ONE* **18**, e0295032. https://doi.org/10.1371/journal.pone.0295032 (2023).
20. Nugroho, H., Utama, N. P. & Surendro, K. Comparison method for handling missing data in clinical studies. In *Proceedings of the 2020 9th International Conference on Software and Computer Applications*, 46–50, https://doi.org/10.1145/3384544.3384594 (2020).

21. Thomas, T. & Rajabi, E. Addressing missing data in a healthcare dataset using an improved knn algorithm. In *International Conference on Computational Science*, 223–230, https://doi.org/10.1007/978-3-030-77977-1_17 (Springer, 2021).
22. Kang, S.-J. et al. Predictors for functionally significant in-stent restenosis: An integrated analysis using coronary angiography, ivus, and myocardial perfusion imaging. *JACC Cardiovasc. Imaging* **6**, 1183–1190 (2013).
23. Chowdhury, M. Z. I. & Turin, T. C. Variable selection strategies and its importance in clinical prediction modelling. *Fam. Med. Commun. Health* https://doi.org/10.1136/fmch-2019-000262 (2020).
24. Badr, A. A. & Abdul-Hassan, A. K. (2021) Catboost machine learning based feature selection for age and gender recognition in short speech utterances. *Int. J. Intell. Eng. Syst.* https://doi.org/10.22266/ijies2021.0630.14
25. Staartjes, V. E. et al. Foundations of feature selection in clinical prediction modeling. In *Machine Learning in Clinical Neuroscience: Foundations and Applications*, 51–57, https://doi.org/10.1007/978-3-030-85292-4_7 (Springer, 2022).
26. Sarica, A., Quattrone, A. & Quattrone, A. Explainable boosting machine for predicting alzheimer's disease from mri hippocampal subfields. In *International Conference on Brain Informatics*, 341–350, https://doi.org/10.1007/978-3-030-86993-9_31 (Springer, 2021).
27. Magunia, H. et al. Machine learning identifies icu outcome predictors in a multicenter covid-19 cohort. *Crit. Care* **25**, 1–14. https://doi.org/10.1186/s13054-021-03720-4 (2021).
28. Nistal-Nuño, B. Developing machine learning models for prediction of mortality in the medical intensive care unit. *Comput. Methods Progr. Biomed.* **216**, 106663. https://doi.org/10.1016/j.cmpb.2022.106663 (2022).
29. Rozeboom, P. D. et al. Development and validation of a multivariable prediction model for postoperative intensive care unit stay in a broad surgical population. *JAMA Surg.* **157**, 344–352. https://doi.org/10.1001/jamasurg.2021.7580 (2022).
30. Hozo, I. & Djulbegovic, B. Generalised decision curve analysis for explicit comparison of treatment effects. *J. Eval. Clin. Pract.* **29**, 1271–1278. https://doi.org/10.1111/jep.13915 (2023).
31. Zoabi, Y., Deri-Rozov, S. & Shomron, N. Machine learning-based prediction of covid-19 diagnosis based on symptoms. *Npj Digit. Med.* **4**, 1–5. https://doi.org/10.1038/s41746-020-00372-6 (2021).
32. Knapič, S., Malhi, A., Saluja, R. & Främling, K. Explainable artificial intelligence for human decision support system in the medical domain. *Mach. Learn. Knowl. Extract.* **3**, 740–770. https://doi.org/10.3390/make3030037 (2021).
33. Saha, B. K. & Burns, S. L. The story of nitric oxide, sepsis and methylene blue: A comprehensive pathophysiologic review. *Am. J. Med. Sci.* **360**, 329–337. https://doi.org/10.1016/j.amjms.2020.06.007 (2020).
34. Evans, L. et al. Surviving sepsis campaign: International guidelines for management of sepsis and septic shock 2021. *Crit. Care Med.* **49**, e1063–e1143. https://doi.org/10.1007/s00134-021-06506-y (2021).
35. Hamzaoui, O., Goury, A. & Teboul, J.-L. The eight unanswered and answered questions about the use of vasopressors in septic shock. *J. Clin. Med.* **12**, 4589. https://doi.org/10.3390/jcm12144589 (2023).
36. Bellomo, R. et al. Renin and survival in patients given angiotensin ii for catecholamine-resistant vasodilatory shock. a clinical trial. *Am. J. Respir. Crit. Care Med.* **202**, 1253–1261. https://doi.org/10.1164/rccm.201911-2172OC (2020).
37. Ammar, M. A. et al. Timing of vasoactive agents and corticosteroid initiation in septic shock. *Ann. Intens. Care* **12**, 47. https://doi.org/10.1186/s13613-022-01021-9 (2022).
38. Tchen, S. & Sullivan, J. B. Clinical utility of midodrine and methylene blue as catecholamine-sparing agents in intensive care unit patients with shock. *J. Crit. Care* **57**, 148–156. https://doi.org/10.1016/j.jcrc.2020.02.011 (2020).
39. Rayan, Z., Alfonse, M. & Salem, A.-B.M. Predicting sepsis in the intensive care unit (icu) through vital signs using support vector machine (svm). *Open Bioinf. J.* https://doi.org/10.2174/1875036202114010010 8 (2021).
40. Alshanbari, H. M. et al. Prediction and classification of covid-19 admissions to intensive care units (icu) using weighted radial kernel svm coupled with recursive feature elimination (rfe). *Life* **12**, 1100. https://doi.org/10.3390/life12071100 (2022).
41. Ge, W. et al. Using deep learning with attention mechanism for identification of novel temporal data patterns for prediction of icu mortality. *Inform. Med. Unlocked* **29**, 100875 (2022).
42. Lin, Y.-W., Zhou, Y., Faghri, F., Shaw, M. J. & Campbell, R. H. Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. *PLoS ONE* **14**, e0218942. https://doi.org/10.1371/journal.pone.0218942 (2019).
43. Scheibner, A. et al. Machine learning to predict vasopressin responsiveness in patients with septic shock. *Pharmacother.: J. Human Pharmacol. Drug Ther.* **42**, 460–471. https://doi.org/10.1002/phar.2683 (2022).
44. Ibarra-Estrada, M. et al. Early adjunctive methylene blue in patients with septic shock: A randomized controlled trial. *Crit. Care* **27**, 110. https://doi.org/10.1186/s13054-023-04397-7 (2023).
45. Kirov, M. Y. et al. Infusion of methylene blue in human septic shock: A pilot, randomized, controlled study. *Crit. Care Med.* **29**, 1860–1867. https://doi.org/10.1097/00003246-200110000-00002 (2001).
46. Memis, D., Karamanlioglu, B., Yuksel, M., Gemlik, I. & Pamukcu, Z. The influence of methylene blue infusion on cytokine levels during severe sepsis. *Anaesth. Intens. Care* **30**, 755–762. https://doi.org/10.1177/0310057X0203000606 (2002).
47. Sari-Yavuz, S. et al. Methylene blue dosing strategies in critically ill adults with shock-a retrospective cohort study. *Front. Med.* **9**, 1014276. https://doi.org/10.3389/fmed.2022.1014276 (2022).

Author contributions

ZXQ, LNZ, LL, and YMZ conceived the study, and SSX performed, designed and built the machine learning models. LL, SSX, ZLL and PXS extracted data from the Xiangya database and performed the analysis. SSX, and LL interpreted the results and drafted the manuscript. ZXQ, LNZ, YMZ, FL, and FW revised the manuscript for important intellectual content. All the authors read and approved the final manuscript. The corresponding author had full access to all data in the study and had final responsibility for the decision to submit for publication.

Funding

This work was supported by the Central South University Research Programme of Advanced Interdisciplinary Studies (Grant No. 2023QYJC022) and the Project Program of the National Clinical Research Center for Geriatric Disorders (Xiangya Hospital, Grant No. 2022LNJJ05). The funders of the study had no role in the study design, data collection, analysis and interpretation, the writing of the manuscript, or the decision to submit the manuscript for publication.

Declarations

Competing interests

The authors declare that they have no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-89934-w.

Correspondence and requests for materials should be addressed to Y.Z. or L.Z.

Reprints and permissions information is available at www.nature.com/reprints.

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