Submitted: 11 September, 2022 Accepted: 09 November, 2022 Published: 08 September, 2023 DOI:10.22514/sv.2023.046

Open Access
Signa Vitae

ORIGINAL RESEARCH

Mortality prediction of patients with sepsis in the emergency department using machine learning models: a retrospective cohort study according to the Sepsis-3 definitions

Eun-Tae Jeon[1,2], Juhyun Song[3,*], Dae Won Park[4], Ki-Sun Lee[2], Sejoong Ahn[5], Joo Yeong Kim[5], Jong-hak Park[5], Sungwoo Moon[5], Han-jin Cho[5]

[1] Department of Neurology, Korea University Ansan Hospital, 15355 Ansan, Republic of Korea
[2] Medical Science Research Center, Korea University Ansan Hospital, 15355 Ansan, Republic of Korea
[3] Department of Emergency Medicine, Korea University Anam Hospital, 02841 Seoul, Republic of Korea
[4] Division of infectious Diseases, Department of Internal Medicine, Korea University Ansan Hospital, 15355 Ansan, Republic of Korea
[5] Department of Emergency Medicine, Korea University Ansan Hospital, 15355 Ansan, Republic of Korea

***Correspondence**
songcap97@hotmail.com
(Juhyun Song)

Abstract

Although clinical scoring systems and biomarkers have been used to predict outcomes in sepsis, their prognostic value is limited. Therefore, machine learning (ML) models have been proposed to predict the outcomes of sepsis. This study aims to propose ML algorithms that create robust models for predicting mortality in patients with sepsis diagnosed using the Sepsis-3 definitions in the emergency department. This study was performed using a prospectively collected registry of adult patients with sepsis between January 2016 and February 2020. Among the 810 patients, 607 (75%) and 203 (25%) patients were assigned to the training and test sets, respectively. The primary outcome was 30-day mortality. Using the values of the area under the receiver operating characteristic curve (AUROC), the area under the precision-recall curve (AUPRC), balanced accuracy, and Brier score, we compared the performances of different ML algorithms with that of the logistic regression models and clinical scoring systems. The ML models' performance was superior to that of the clinical scoring systems. A light gradient boosting machine achieved the highest AUROC among the ML models in predicting 30-day mortality. Most of the ML models had significantly higher AUROC and balanced accuracy than the logistic regression models. All the ML models exhibited higher AUPRC and lower Brier scores compared to the scoring systems and logistic regression model. The ML models can be used as supportive tools for predicting mortality in sepsis patients. In future studies, the performance of the proposed models will be validated using more data from different hospitals or departments.

Keywords

Emergency department; Machine learning; Mortality; Sepsis; Septic shock

1. Introduction

Sepsis is a global health problem with high mortality despite advances in antimicrobial agents and resuscitation [1–3]. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) defined sepsis as a life-threatening organ dysfunction caused by infection [4]. The early prediction of clinical outcomes is important because it can guide early intervention and help reduce mortality among patients with sepsis [5].

The prognostic value of the quick sepsis-related organ failure assessment (qSOFA) score, systemic inflammatory response syndrome (SIRS), early warning score (EWS), acute physiology and chronic health evaluation II, and SOFA score have been validated in patients with sepsis or suspected infections [6–9]. Biomarkers such as lactate, procalcitonin, C-reactive protein, and presepsin have also been used to predict the clinical outcomes in patients with sepsis [10–12]. However, these clinical severity scores and sepsis biomarkers were shown to exhibit only moderate to poor values for prognosticating sepsis.

Due to the limited prognostic values of these clinical severity scores and biomarkers, various machine learning (ML) models using medical data have been proposed as more powerful and accurate tools to predict clinical outcomes in sepsis. A recent study showed that an ML model provided the early identification of patients at risk for circulatory failure in the intensive care unit (ICU) [13]. Another study validated ML-based mortality prediction in patients with suspected infections in the emergency department [14]. The study showed that the accuracy rates of ML methods in predicting mortality were superior to those of pre-existing screening tools, such as SIRS and qSOFA [14]. ML models using the first 6 hours of clinical data in sepsis patients could accurately predict severity, mortality, and length of stay in the ICU [15]. However, to our knowledge, few ML-based studies have been conducted where patients with sepsis diagnosed using the latest Sepsis-3 definitions in the emergency department (ED) have been included to predict mortality.

In this study, we aim to create various ML algorithms to realize a more robust model for predicting short-term mortality among patients with sepsis diagnosed according to the Sepsis-3 definitions in the ED. The prognostic values of the individual models were compared with each other and the logistic regression model and clinical severity scores.

2. Materials and methods

2.1 Study design and participants

This study was performed using a prospectively collected registry of ED patients who were diagnosed with sepsis and septic shock according to the Sepsis-3 definitions in a tertiary care teaching hospital in Korea. From January 2016 to February 2020, adults (aged $\geq$18 years) who visited the ED with an actual or suspected infection and an increase in SOFA score of $\geq$2 points were enrolled by ED physicians on duty. If the patients had baseline SOFA scores, we used the standard of an increase in the SOFA score of at least two points. If the patients had no previous SOFA score, we reviewed their medical records with laboratory results and determined the association between the present infection and the SOFA score. The exclusion criteria were: (a) age $<$18 years, (b) death within 12 hours of ED presentation, (c) unknown outcome (30-day mortality), and (d) ED visit for trauma care. Finally, two infectious disease experts carefully reviewed all the patients. Clinical variables were collected from the patients and different ML algorithms were used to predict 30-day mortality. The prognostic values of the different ML algorithms were compared with that of the logistic regression models and established clinical scoring systems. This study followed the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) reporting guidelines [16].

2.2 Dataset and definitions

Sepsis is as a life-threatening organ dysfunction caused by a dysregulated host response to infection [4]. Septic shock is a subset of sepsis where profound circulatory, cellular, and metabolic abnormalities pose a greater risk of mortality compared to sepsis alone [4, 17]. The qSOFA score uses three criteria: low blood pressure (systolic blood pressure: $\leq$100 mmHg), high respiratory rate ($\geq$22 breaths/min), and altered mental status (Glasgow coma score: $<$15), assigning one point for each criterion with the final score ranging from zero to three points. A positive qSOFA score was defined as the presence of $\geq$2 qSOFA points near the onset of infection. The diagnostic criterion for sepsis includes an increase in the SOFA score by $\geq$2 points due to the present infection. The SOFA score is used to track a person's status to determine the extent of organ function or the rate of failure [18–20]. The score is based on six different scores, each for the respiratory, cardiovascular, hepatic, coagulation, renal, and neurological systems. When the physiological parameters do not match any row, zero points are given. The criteria for septic shock include vasopressor requirements to maintain a mean arterial pressure of 65 mmHg and serum lactate level $>$2 mmoL/L despite adequate fluid resuscitation. The National Early Warning Score (NEWS) (2012) and NEWS2 (2017) are tools developed by the Royal College of Physicians to improve the detection and response to clinical deterioration in adult patients and are critical to patient safety and improving patient outcomes [21]. The modified early warning score (MEWS) is a composite score used by physicians to determine the severity of illness in various clinical situations [22]. It evaluates the risk of mortality based on vital signs (systolic blood pressure, heart rate, respiration rate, and body temperature), saturation of percutaneous oxygen (SpO_2), and level of consciousness. Determining a MEWS involves assigning a number between zero and three to each of the six items. A total score equal to or more than five points of MEWS is associated with an increased mortality. In the present study, we defined follow-up (F/U) lactate levels within 12 hours as the maximum values of lactate levels measured within 12 hours of ED presentation, except initial lactate levels.

2.3 Data splitting and preprocessing

Fig. 1 shows this study's ML flowchart. Variables with more than 40% missing data were excluded [23], and 25% of the data were randomly separated with stratifications of sepsis-related death. The hold-out data were used only in the final model evaluation as a test set, and the remaining 75% of the data were used as a training set with a leave-one-out cross-validation strategy. Multivariate imputation by chained equations [24] and isolation forests [25] were used for imputation and outlier detections, respectively. Min-max scaling and one-hot encoding were applied, and **Supplementary material** provides the details of the preprocessing procedures. Included variables for model development with information of their missing rate and variable type are presented in **Supplementary Table 1**.

2.4 Feature selection and feature importance analyses

Recursive feature elimination [26] was utilized to obtain the best feature set maximizing the model performance of the area under the receiver operating characteristic curve (AUROC) in cross-validation. Recursive feature elimination works with the following iterative procedures: (1) training a classifier, (2) computing the ranking criterion of all features, and (3) removing the feature with the lowest rank. A light gradient boosting machine (LightGBM) [28], which can handle categorical variables without data conversion, was used for the feature importance analyses. Contributions of each feature to the model prediction were measured with Shapley Additive exPlanations (SHAP) values, where positive and negative values indicated a positive and negative effect on the prediction score, respectively [27]. The relative importance of the features was evaluated and ranked using the mean absolute SHAP value. Hierarchical clustering was performed to minimize the effects of multicollinearity, causing the underestimation of the relative importance of the features [29]. **Supplementary material** describes the details.

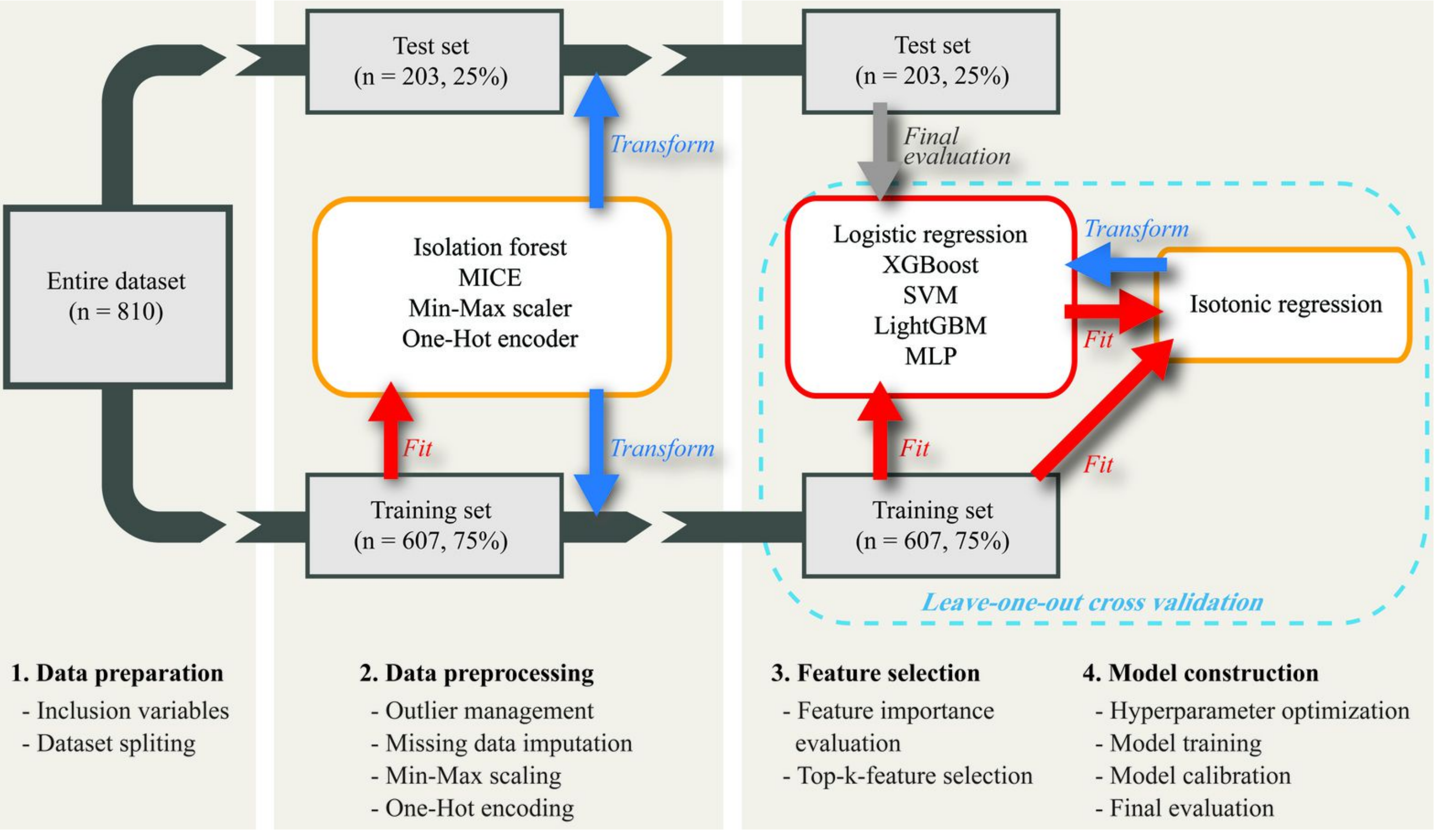

FIGURE 1. Machine learning flowchart. MICE: multiple imputation by chained equations; XGBoost: extreme gradient boosting; SVM: support vector machine; LightGBM: light gradient boosting machine; MLP: multilayer perceptron.

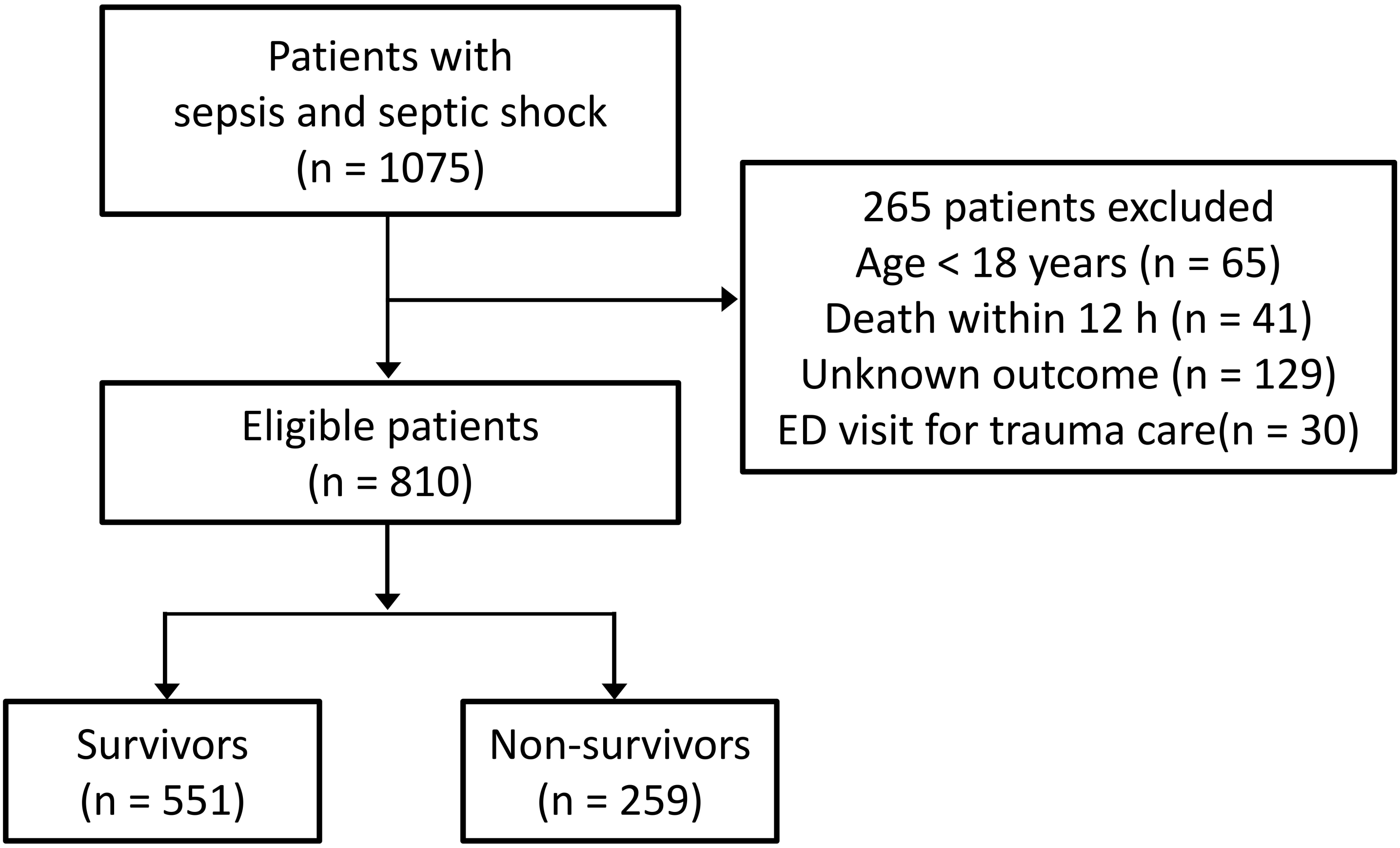

FIGURE 2. Flowchart of the study population.

2.5 Model construction

A conventional statistical model-logistic regression—was built as a baseline comparator. Four popular and promising models were constructed: extreme gradient boosting (XGBoost) [30], support vector machine (SVM) [31], LightGBM, and multilayer perceptron (MLP) [32]. XGBoost (version 1.3.2, The XGBoost Contributors, New York, NY, USA) and LightGBM (version 3.2.0, Microsoft, Wilmington, Del, USA) are gradient-boosted tree-based ensemble models using a depth-wise algorithm and a leaf-wise growth algorithm, respectively. SVM is a model for classification, regression, and outlier detection using an optimal hyperplane in a multi-dimensional space. MLP is a feed-forward neural network model with a basic architecture consisted of fully connected layers. Bayesian optimization [33] was used for hyperparameter settings to maximize the AUROC in cross-validation. **Supplementary material** provides the details of the constructed models. All processes including development of SVM and MLP models were implemented in Python version 3.8.2 (Python Software Foundation, Wilmington, Del, USA) with scikit-lean (version 0.24.1) and tensorflow (version 2.0.0) libraries.

2.6 Statistical analysis

Descriptive statistics were represented as the number (percentage), mean (SD), or median (Q1 and Q3). We used the Shapiro-Wilk test for normality and Levene's test for homoscedasticity. The Chi-squared test, independent *t*-test, or Mann-Whitney U test were used for comparison analysis.

The AUROC was chosen as the primary evaluation metric with a threshold of 0.50. The AUROC is classified into excellent (0.9–1.0), good (0.8–0.9), fair (0.7–0.8), poor (0.6–0.7), and fail (0.5–0.6). The AUROC with its 95% confidence interval (CI) was calculated and compared between constructed models using Delong's method [34]. *p*-values were adjusted using the Benjamini-Hochberg method for multiple comparisons.

During cross validation for developing the models, each fold was calibrated to validation data after fitting at training data using isotonic regression. The isotonic regression fits a non-decreasing real function to one dimensional data. Prediction scores from the calibrated models allow reflection for predictive probability. Error after the post-hoc calibration was evaluated using calibration curves and the Brier score which was the mean squared error of the predicted probability [35]. The AUPRC, balanced accuracy, and F1 score were calculated as the secondary outcome metrics for model performance. The significance level was set at $p < 0.05$.

3. Results

3.1 Baseline characteristics

During the study period, 1075 consecutive patients with sepsis were screened. Fig. 2 shows a flowchart of the study population. Patients aged <18 years (n = 65), death within 12 hours of ED presentation (n = 41), unknown 30-day mortality (n = 129), and ED visits for trauma care (n = 30) were excluded. After excluding 265 patients, a final total of 810 patients were finally enrolled in the analysis.

Table 1 shows the baseline characteristics of the patients for 30-day mortality outcomes (**Supplementary Tables 2 and 3** for 7-day and 14-day mortalities, respectively). The median age of the enrolled patients was 75 years (interquartile range (IQR): 65 to 82 years), and 337 (41.6%) were women. Among the patients, 259 (32.0%) were non-survivors. There were significant differences in age, vital signs except for heart rate, comorbidities (diabetes mellitus, malignancy, chronic liver disease, and cerebrovascular disease), source of infection (respiratory and genitourinary infections), laboratory findings (platelet, bilirubin, lactate, and arterial blood gas analysis), septic shock, and clinical severity scores (NEWS, NEWS2, MEWS, and SOFA scores) between non-survivors and survivors. The missing data rates were less than 3.1% except for three variables: F/U lactate levels within 12 hours (missing rate, 35.1%), lactate clearance (missing rate, 20.2%), and procalcitonin (missing rate, 16.7%) (**Supplementary Table 1**).

3.2 Performance of the models and scoring systems

LightGBM achieved the highest AUROC values in prediction for 7-day (0.891 (0.841–0.941)), 14-day (0.893 (0.844–0.941)), and 30-day (0.871 (0.823–0.919)) mortalities among the ML models (Table 2 and Fig. 3(A)), but there were no significant differences between LightGBM and MLP in predictions for 7-day, 14-day, and 30-day mortalities. All the ML models except XGBoost in predictions for 7-day and 14-day mortalities exhibited significantly higher AUROC values compared to the logistic regression.

SOFA achieved the highest AUROC values among the scoring systems in predicting 7-day (0.680 (0.594–0.766)), 14-day (0.647 (0.565–0.730)), and 30-day (0.658 (0.579–0.736)) mortalities. MEWS had no discrimination ability in predicting 30-day mortality, of which the confidence interval of AUROC spanned 0.50. All the ML models and logistic regression had significantly higher AUROC values compared to SOFA scores.

ML models exhibited higher performance in the AUPRC (Table 2 and Fig. 3(B)) and Brier score (Fig. 4 for 30-day mortality and **Supplementary Figs. 1,2** for 7-day and 14-day mortalities) compared to the scoring systems and logistic regression. In particular, LightGBM had the highest AUPRC values in prediction for 7-day (0.702 (0.547–0.847)) and 14-day (0.786 (0.670–0.884)) mortalities, while MLP had the highest AUPRC values in predicting 30-day mortality (0.774 (0.662–0.865)). MLP in the prediction for 7-day (0.789) and 14-day (0.774) mortalities and LightGBM in predicting 30-day mortality (0.762) exhibited the highest balanced accuracy. **Supplementary Table 4** presents the cross-validation performance.

3.3 Importance of variables

Thirty-three variables were selected as the best feature set by the recursive feature elimination in predicting 30-day mortality, and all selected variables in the predictions for 7-day and 14-day mortality were included. Fig. 5 shows the overall contribution of the variables (left bar plot) and the impacts of individual values for the model prediction (the right violin

TABLE 1. Baseline characteristics of the study population.

Variables	All (n = 810)	Survivors on 30-day (n = 551)	Non-Survivors on 30-day (n = 259)	*p* value
Demographics				
Female (sex)	337 (41.6%)	225 (40.8%)	112 (43.2%)	0.567
Age, years	75 (65, 82)	74 (62, 81)	78 (69, 83)	<0.001***
Vital signs and O_2 saturation				
Body temperature, °C	37.1 (36.4, 38.0)	37.2 (36.5, 38.1)	36.7 (36.0, 37.6)	<0.001***
Systolic blood pressure, mmHg	96 (80, 125)	98 (83, 126)	93 (74, 118)	<0.001***
Diastolic blood pressure, mmHg	60 (50, 73)	60 (51, 74)	57 (47, 72)	<0.001***
Heart rate, bpm	108 (90, 124)	106 (90, 124)	109 (92, 124)	0.266
Respiration rate, /min	24.5 (5.8)	23.9 (5.4)	25.7 (6.4)	<0.001***
SpO_2, %	92.2 (8.7)	93.4 (7.1)	89.8 (11.0)	<0.001***
Comorbidity				
Diabetes Mellitus	313 (39.2%)	198 (35.9%)	115 (44.4%)	0.020*
Hypertension	423 (52.9%)	288 (52.3%)	135 (52.1%)	0.976
Malignancy	149 (18.6%)	70 (12.7%)	79 (30.5%)	<0.001***
Chronic lung disease	187 (23.4%)	121 (22.0%)	66 (25.5%)	0.277
Chronic liver disease	46 (5.8%)	20 (3.6%)	26 (10.0%)	<0.001***
Chronic kidney disease	107 (13.4%)	79 (14.3%)	28 (10.8%)	0.219
Cardiovascular disease	143 (17.9 %)	100 (18.1 %)	43 (16.6 %)	0.698
Cerebrovascular disease	442 (55.3%)	327 (59.3%)	115 (44.4%)	<0.001***
Organ transplantation	11 (1.4 %)	10 (1.8 %)	1 (0.4 %)	0.193
AIDS	3 (0.4%)	2 (0.4%)	1 (0.4%)	0.573
Others	209 (26.2%)	145 (26.3%)	64 (24.7%)	0.737
Unknown	11 (1.4%)	6 (1.1%)	5 (1.9%)	0.522
Infection source				
Respiratory	534 (65.9%)	348 (63.2%)	186 (71.8%)	0.019*
Genitourinary	301 (37.2%)	222 (40.3%)	79 (30.5%)	0.009**
Gastrointestinal	91 (11.2%)	57 (10.3%)	34 (13.1%)	0.294
Bacteremia	61 (7.5%)	39 (7.1%)	22 (8.5%)	0.569
Others	49 (6.0%)	33 (6.0%)	16 (6.2%)	0.958
Laboratory findings				
White blood cell, $10^3/\mu L$	11.9 (7.7, 17.5)	11.7 (7.9, 17.2)	13.1 (6.9, 18.5)	0.255
Platelet, $10^3/\mu L$	200 (126, 284)	205 (134, 285)	187 (100, 274)	0.011*
Glucose, mg/dL	176.8 (141.2)	175.1 (133.3)	180.5 (156.6)	0.612
Bilirubin, mg/dL	1.1 (1.7)	0.9 (1.4)	1.4 (2.3)	0.002**
Creatinine, mg/dL	1.8 (1.9)	1.8 (1.9)	1.9 (1.8)	0.226

TABLE 1. Continued.

Variables	All (n = 810)	Survivors on 30-day (n = 551)	Non-Survivors on 30-day (n = 259)	*p* value
C-reactive protein, mg/L	12.4 (10.1)	12.0 (10.1)	13.4 (10.0)	0.078
Procalcitonin, ng/mL	13.9 (27.1)	13.6 (27.0)	14.4 (27.5)	0.731
Initial lactate, mg/dL	4.3 (3.7)	3.6 (3.2)	5.7 (4.3)	<0.001***
F/U lactate within 12hr, mg/dL	3.6 (3.3)	2.8 (2.5)	5.1 (4.0)	<0.001***
Lactate clearance, %	28.3 (70.8)	35.2 (52.0)	14.6 (96.4)	<0.001***
Arterial Blood Gas Analysis				
pH	7.4 (0.1)	7.4 (0.1)	7.3 (0.1)	<0.001***
$PaCO_2$, mmHg	36.6 (14.2)	36.3 (12.9)	37.2 (16.6)	0.445
PaO_2, mmHg	84.9 (53.9)	86.1 (54.6)	82.2 (52.2)	0.333
HCO_3-, mEq/L	21.2 (7.5)	22.0 (6.8)	19.7 (8.5)	<0.001***
SaO_2, %	90.9 (10.3)	92.2 (8.2)	88.0 (13.2)	<0.001***
Clinical severity				
Septic shock	365 (54.9%)	194 (35.2%)	171 (66.0%)	<0.001***
Glasgow coma scale	10 (8, 13)	10 (8, 13)	10 (7, 12)	<0.001***
SOFA score	8.0 (6.0, 11.0)	7.0 (5.0, 10.0)	10.0 (8.0, 12.0)	<0.001***
NEWS score	11 (9.0, 13.0)	10 (8.0, 12.0)	12 (10.0, 14.0)	<0.001***
NEWS2 score	11 (9.0, 13.0)	11 (8.0, 12.0)	12 (10.0, 14.0)	<0.001***
MEWS score	6.0 (5.0, 8.0)	6.0 (4.5, 7.5)	6.0 (5.0, 8.0)	<0.001***
qSOFA score of 3	204 (25.2%)	114 (20.7%)	90 (34.7%)	<0.001***
Treatment				
Time to antibiotics, min	96 (46, 162)	96 (44, 166)	95 (51, 158)	0.417
Steroid administration within 12hr	84 (10.4%)	52 (9.4%)	32 (12.4%)	0.251
Antibiotics within 3hr	630 (77.8%)	425 (77.1%)	205 (79.2%)	0.580
Source control				
Antibiotics only	766 (94.6%)	520 (94.4%)	246 (95.0%)	0.850
Emergent surgery	5 (0.6%)	5 (0.9%)	0 (0.0%)	0.291
Percutaneous drainage	23 (2.8%)	13 (2.4%)	10 (3.9%)	0.330
Endoscopic intervention	14 (1.7%)	11 (2.0%)	3 (1.2%)	0.572
Removal of infected device	2 (0.2%)	2 (0.4%)	0 (0.0%)	0.832

*The values are expressed as n (%), mean (SD), or median (Q1, Q3). *p < 0.05, **p < 0.01, and ***p < 0.001. Abbreviations: AIDS, acquired immune deficiency syndrome; F/U, follow up; MEWS, modified early warning score; NEWS, national early warning score; NEWS2, national early warning score 2; qSOFA, quick sequential organ failure assessment; SaO_2, Arterial oxygen saturation; SOFA, sequential organ failure assessment; SpO_2, Saturation of percutaneous oxygen; $PaCO_2$, partial pressure of carbon dioxide; PaO_2, partial pressure of oxygen.*

TABLE 2. Performance of the models on test set for 7-day, 14-day, and 30-day mortalities.

Model	AUROC (95% CI)			AUPRC (95% CI)		
	7-day	14-day	30-day	7-day	14-day	30-day
Scoring systems						
qSOFA	0.59 (0.52–0.66)	0.57 (0.50–0.64)	0.57 (0.50–0.64)	0.42 (0.32–0.52)	0.47 (0.37–0.57)	0.49 (0.39–0.60)
SOFA	0.68 (0.59–0.77)	0.65 (0.57–0.73)	0.66 (0.58–0.74)	0.35 (0.20–0.49)	0.39 (0.25–0.53)	0.43 (0.29–0.56)
NEWS	0.63 (0.54–0.72)	0.63 (0.55–0.71)	0.63 (0.55–0.71)	0.32 (0.18–0.45)	0.39 (0.25–0.53)	0.41 (0.27–0.55)
NEW2	0.62 (0.53–0.71)	0.62 (0.54–0.71)	0.62 (0.54–0.70)	0.32 (0.18–0.45)	0.39 (0.24–0.53)	0.41 (0.27–0.55)
MEWS	0.59 (0.50–0.68)	0.59 (0.51–0.68)	0.57 (0.49–0.65)	0.33 (0.20–0.45)	0.38 (0.25–0.51)	0.39 (0.26–0.52)
Baseline model						
LogReg	0.82 (0.74–0.89)	0.84 (0.77–0.90)	0.81 (0.74–0.88)	0.56 (0.38–0.73)	0.69 (0.55–0.82)	0.7 (0.57–0.81)
ML models						
XGBoost	0.85 (0.78–0.91)	0.84 (0.79–0.90)	0.84 (0.78–0.89)*	0.62 (0.46–0.77)	0.71 (0.59–0.82)	0.72 (0.60–0.82)
SVM	0.84 (0.78–0.90)*	0.85 (0.79–0.91)*	0.85 (0.79–0.91)*	0.61 (0.44–0.77)	0.69 (0.53–0.84)	0.76 (0.65–0.85)
LightGBM	0.89 (0.84–0.94)*	0.89 (0.84–0.94)*	0.87 (0.82–0.92)*	0.7 (0.55–0.85)	0.79 (0.67–0.88)	0.76 (0.66–0.85)
MLP	0.89 (0.83–0.94)*	0.88 (0.83–0.93)*	0.86 (0.81–0.92)*	0.69 (0.51–0.84)	0.77 (0.64–0.87)	0.77 (0.66–0.87)

**Significant difference between ML models and logistic regression ($p < 0.05$). p values were corrected by Benjamini-Hochberg method. Abbreviations: MEWS, modified early warning score; NEWS, national early warning score; NEWS2, national early warning score 2; qSOFA, quick sequential organ failure assessment; SOFA, sequential organ failure assessment; ML, machine learning; XGBoost, extreme gradient boosting; SVM, support vector machine; LightGBM, light gradient boosting machine; MLP, multilayer perceptron; AUROC, area under the receiver operating characteristic curve; CI, confidence interval.*

plot). Fig. 6 shows the partial SHAP dependence plots for the top six important variables, while **Supplementary Fig. 3** presents the plots for the other variables.

The most important variable was septic shock, which had a positive effect on the model prediction, followed by F/U lactate levels within 12 hours followed by the initial lactate levels. When the F/U lactate level within 12 hours was less than 3.7 mmoL/L, it consistently exhibited a negative effect on the model prediction. However, it reversed at 3.7 mmoL/L, and SHAP values increased sharply with positive effects for the model prediction. Initial lactate levels exhibited a similar tendency with a cut-off of 5.3 mmoL/L. The following three variables were considered: malignancy, age, and arterial oxygen saturation (SaO_2). Malignancy had a positive effect on model prediction. Age exhibited a positive correlation with model prediction, resulting in the shape of the sigmoid curve and the cut-off value of 75.1 years. SaO_2 demonstrated a negative correlation with model prediction and a cut-off value of 87.5%.

4. Discussion

It was demonstrated that the performance of 7-day, 14-day, and 30-day mortality predictions in sepsis was greatly improved by ML. Among the various methods validated, LightGBM exhibited the best predictive value for sepsis-related mortality. To construct ML models for mortality prediction, we used clinical variables which can be promptly acquired in the usual ED clinical pathway. Although, our study did not include the variables in ICU or general ward, it demonstrated that ML models constructed using variables in ED can effectively predict outcome among patients with sepsis. In a previous study, 35 clinical variables were used with the relevant vector machine (RVM, a variant of SVM) model to predict sepsis-related mortality among 354 ICU patients. The study demonstrated that RVM had an AUROC of 0.80, which was lower than our SVM results (AUROC 0.850, 95% CI, 0.794–0.906). Our samples were ED-based patients, and the size of the samples were larger than that of the previous study. The variables

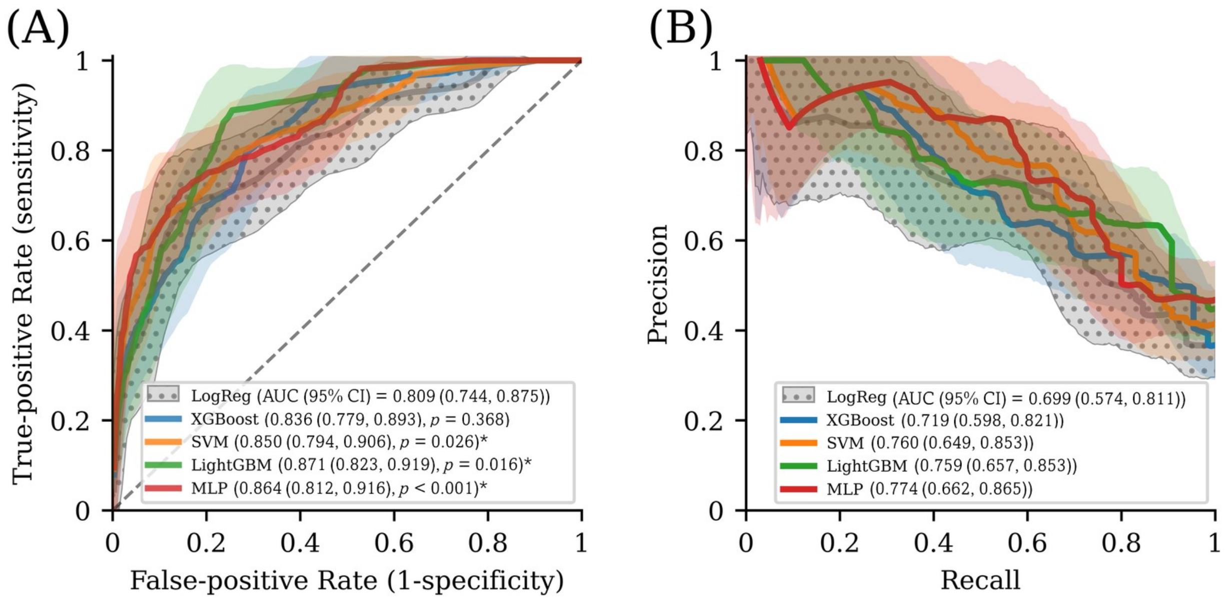

FIGURE 3. Test set performance in prediction for 30-day mortality. A, Solid lines and shades representing receiver operating characteristics curves and its 95% confidence intervals. An asterisk (*) indicates a significant difference ($p < 0.05$) in comparison with logistic regression. B, Solid lines and shades representing precision-recall curves and its 95% confidence intervals. Only the confidence intervals of the baseline model (logistic regression (LogReg)) are represented with a polka dot pattern in both plots. AUC, area under the receiver operating characteristic curve; CI, confidence interval; XGBoost, extreme gradient boosting; SVM, support vector machine; LightGBM, light gradient boosting machine; MLP, multilayer perceptron.

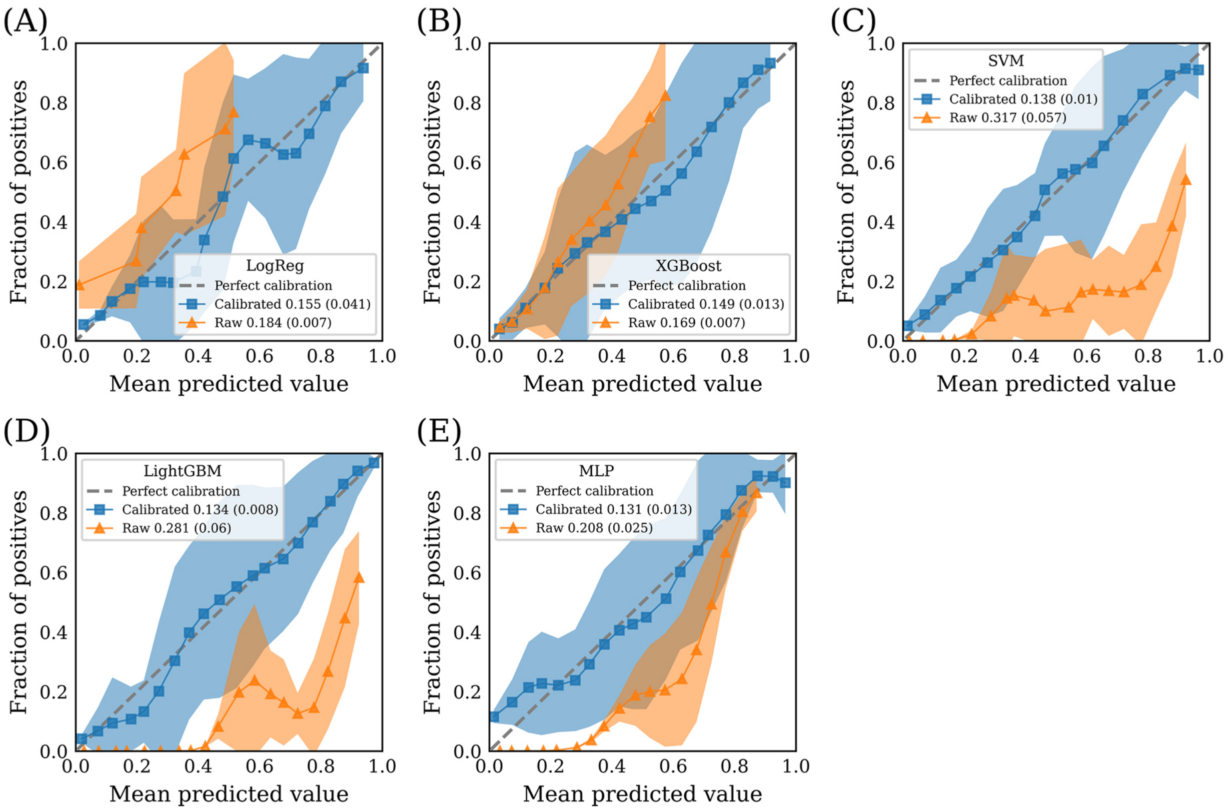

FIGURE 4. Post-hoc calibration of models in the prediction for 30-day mortality. Orange and blue lines represent the calibration curves of raw and calibrated models (isotonic regression), respectively. The legend displays the mean (SD) value of brier scores for model prediction, while solid lines and shades represent the mean and ± SD of the calibration curves. (A) Calibration curves of logistic regression; (B) Calibration curves of XGBoost; (C) Calibration curves of support vector machine; (D) Calibration curves of light gradient boosting machine; (E) Calibration curves of multilayer perceptron. XGBoost, extreme gradient boosting; SVM, support vector machine; LightGBM, light gradient boosting machine; MLP, multilayer perceptron.

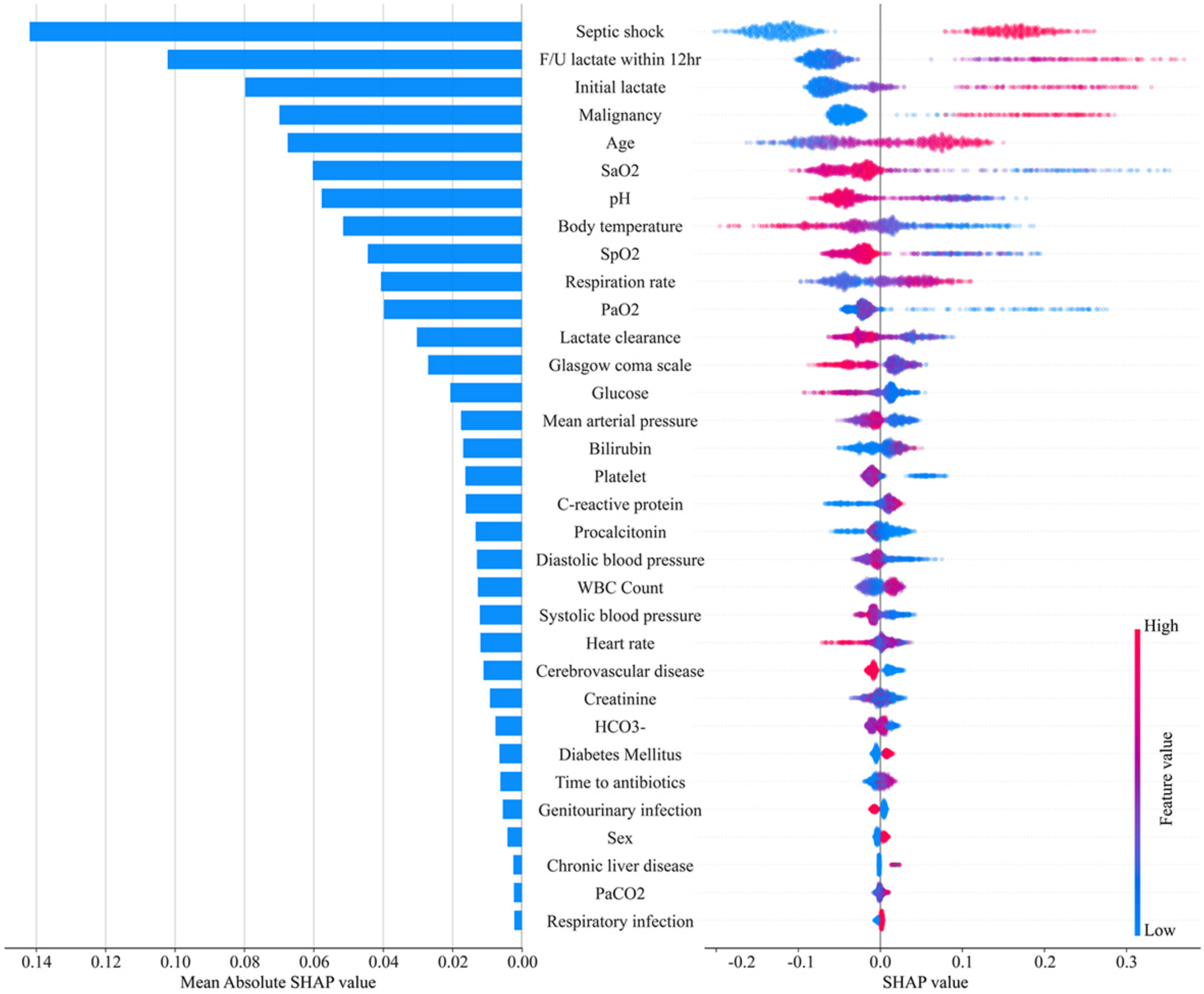

FIGURE 5. Importance of the selected variables. The left bar plot shows the overall contribution of the variables to model prediction. The right violin plot shows the impacts of individual values of the variables and directionality for model prediction. The red color represents the large value in continuous variables or the affirmative response in categorical variables. Supplementary Table 3 presents the variable types. F/U, follow up; HCO_3, bicarbonate; pH, potential of hydrogen; SaO_2, Arterial oxygen saturation; SpO_2, Saturation of percutaneous oxygen; PaO_2, partial pressure of oxygen; $PaCO_2$, partial pressure of carbon dioxide; SHAP, Shapley Additive exPlanations; WBC, white blood cell.

using the SHAP method were evaluated by minimizing the effect of multicollinearity and selecting the most promising feature set in the recursive feature elimination scheme. This variable selection process was believed to improve the results. Another recent study showed that ML algorithms exhibited a higher accuracy rate for mortality prediction compared to the conventional regression model and existing medical scores such as SIRS and qSOFA [14]. In that study, 53 clinical variables were used with SVM to predict 28-day mortality among the 42,220 ED patients with suspected infection, and the model achieved AUROC 0.90 (95% CI; 0.89–0.90). The difference between their performance and these results may be explained by the sample numbers and subject population. Only patients with sepsis were included, and the patient numbers were small. However, the homogenous population enabled in-depth feature analysis. Although homogeneity made it difficult to discriminate and lower performance was inevitable, the more relevant and informative results of the importance of the variables for sepsis were presented.

In recent years, the model's performance and interpretability have become important. The contribution to the model predictions of individual values of variables in every patient using SHAP values was analyzed. Although the relatively small sample size of this study may limit generalizability because of the data dependency of SHAP values, clear patterns and directionalities of the contributions were demonstrated. Consequently, our model can be applied as an interpretable model and can provide intuitive information for individual patients.

Interestingly, contrary to the best performance of LightGBM, XGBoost had the lowest performance among the ML models. They were identical in base architecture as a gradient-

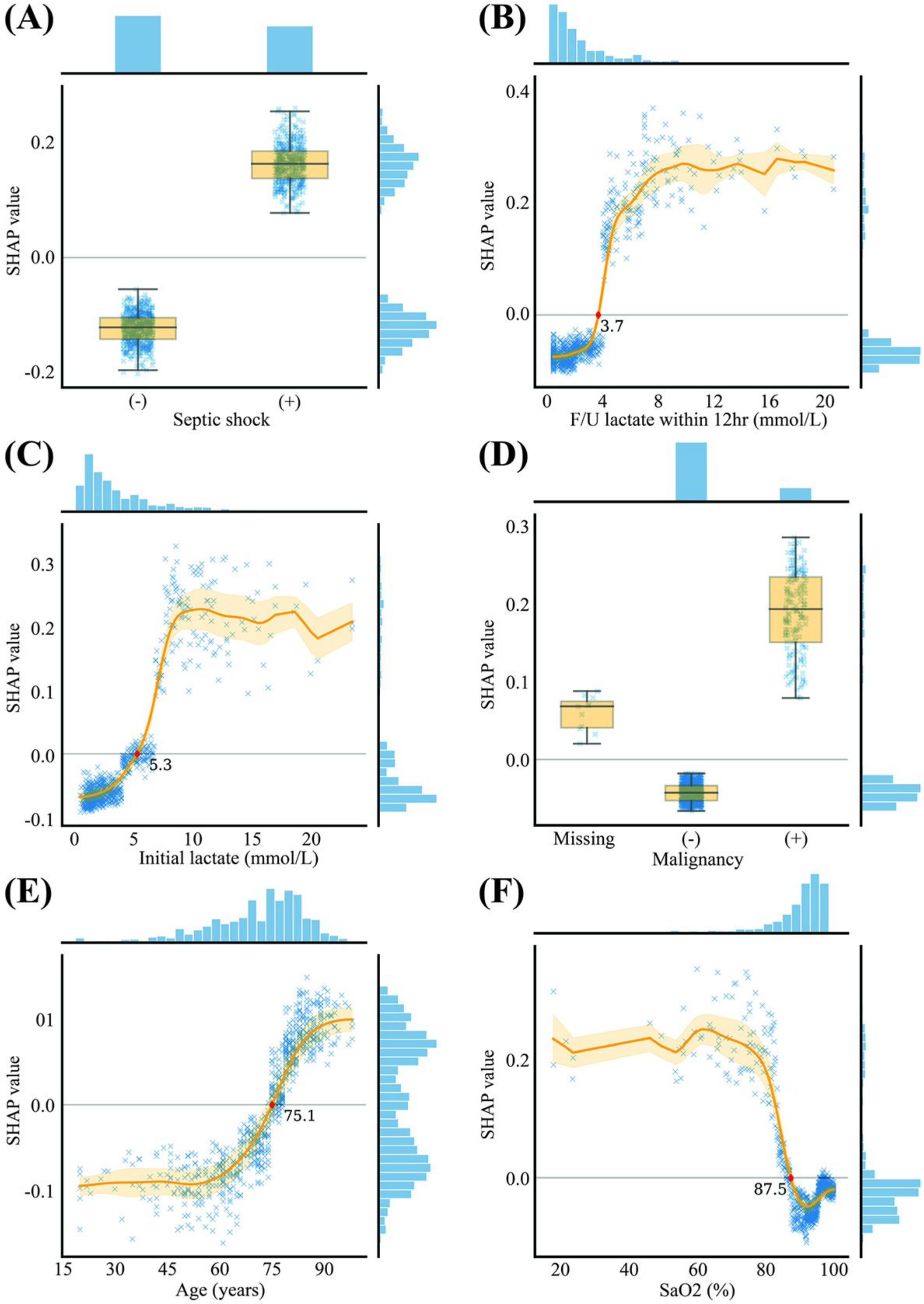

FIGURE 6. Partial SHAP dependence plots for the representative top six important variables. (A), (D) Categorical variables are plotted with a scatter plot and a box plot with whiskers of 1.5 times the interquartile ranges. (B), (C), (E), (F) Continuous variables are plotted with a scatter plot and a regression line represented with the orange line of mean and shade of SD. A red diamond represents a cut-off value of the variable. Histograms on the right and top of each plot are distributions of the SHAP and values of variables. SHAP, Shapley Additive exPlanations; F/U, follow up; SaO_2, Arterial oxygen saturation.

boosted tree, but LightGBM implements a leaf-wise growth algorithm that splits a leaf node with maximum delta loss. Unlike the depth-wise algorithm of XGBoost, the leaf-wise algorithm could make it vulnerable to overfitting. However, it is highly effective in minimizing training loss and was thought that the intensive training strategy was suitable for our data of sepsis patients in the ED. Similarly, MLP, which was a powerful feature extractor but could easily be overfit, almost yielded the highest performance. MLP showed the highest AUPRC among the models. AUPRC is more informative for model performance than AUROC for skewed data [36]. The number of non-survivors included in the present study was almost a half of that of survivors. Therefore, our results showed that MLP might be effective in the target population of the present study.

Our study showed that the most important variable evaluated by recursive feature elimination was septic shock followed by F/U lactate levels, initial lactate levels, malignancy, age, and SaO_2. Septic shock is a subset of sepsis in which profound circulatory, cellular, and metabolic abnormalities pose a greater risk of mortality than sepsis alone [4, 17]. Other important features are subsequent and initial lactate levels, which have been well investigated in previous studies. These studies demonstrated that lactate levels and lactate clearance could be used to predict mortality among patients with sepsis [12, 37], which was in accordance with this study's results. A propensity score matching analysis demonstrated that a history of early stage solid malignancy was an independent risk factor for 28-day mortality in patients with sepsis [38]. A descriptive analysis in China reported that increasing age was independently associated with increased sepsis-related mortality [39]. Similar to this study, a recent cross-sectional analysis using ML revealed that low oxygen saturation was an important variable for predicting short-term mortality in patients with sepsis [40].

In this study, while respiratory infection was included in the best feature set for predicting 14-day and 30-day mortalities, it was not included in the best feature set for predicting 7-day mortality. The early activation of both innate and adaptive immune responses is involved in the pathogenesis of sepsis. The peak mortality rates during the early period are caused by an overwhelming inflammatory response, also known as "cytokine storm," which consists of fever, refractory shock, inadequate resuscitation, and pulmonary or cardiac failure [41]. Meanwhile, mortality in the later period is caused by persistent immunosuppression with secondary infections that result in organ dysfunction [41]. Our results suggest that 7-day mortality (mortality during the early period) was mainly caused by cytokine storms and may be less influenced by organ dysfunction caused by respiratory infections versus 14-day and 30-day mortalities. Although advanced ICU treatment has recently improved short-term mortality, patients still die in later periods due to persistent immunosuppression, immune dysfunction, or chronic catabolism.

There were some limitations to the present study. First, the proposed ML models were trained with ED data from one tertiary care teaching hospital and may not be applicable to ICUs or other primary hospitals. For the purposes in other populations, specific data corresponding to the populations should be utilized because of the data-dependent nature of ML algorithms. Second, due to the single-center design and absence of external validation, our models may not be generalizable to other institutions. Further studies had better to include multi-center data including independent datasets. Third, although we used 30-day mortality as a primary outcome, other adverse outcomes such as endotracheal intubation, ICU admission, and extracorporeal membrane oxygenation could aid ED physicians in promptly identifying sepsis patients with poor outcomes and act on the disease. Further studies are recommended to investigate the effects of ML algorithms for not only mortality but also the adverse outcomes. Fourth, our study did not utilize time-series data for model construction and analysis. Because the time-varying characteristics of the selected variables are important to 30-day mortality, further studies need to consider the utilization of time-series data. Fifth, our study mainly compared the performance for the ML algorithms, logistic regression, and other clinical scoring systems, rather than suggest novel models or techniques. Hence, the technical novelty is lowered to some extent.

5. Conclusions

This study demonstrated that the prognostic values of ML models were superior to those of existing clinical scoring systems in patients with sepsis. Furthermore, among the ML models, the performance of LightGBM and MLP were better compared to the logistic regression model. In future studies, the performance of our proposed model will be validated using more data from different hospitals or departments.

AVAILABILITY OF DATA AND MATERIALS

The data presented in this study are available on reasonable request from the corresponding author.

AUTHOR CONTRIBUTIONS

JS, EJ and DWP—designed the research study. JS, JYK, SA, SM, JP, HC and EJ—performed the research. EJ, DWP and KL—provided help and advice on the analysis. EJ and KL—analyzed the data. JS and EJ—wrote the manuscript. All authors contributed to editorial changes in the manuscript. All authors read and approved the final manuscript.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

The study was approved by the Institutional Review Board (IRB) of Korea University Medical Center (2020AS0163). Informed consent was waived because of the retrospective analysis of the data.

ACKNOWLEDGMENT

We would like to thank Namyoung Baek for her contribution to our project. We would also like to thank eWorldEditing for English language editing.

FUNDING

This research was supported by the National Research Foundation of Korea (NRF) through a grant from the Korean Government (MSIT) (Grant numbers: 2020R1C1C1010362). The funding bodies had no role in the study design, conduct, review, report, or in the decision to submit the article for publication.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

SUPPLEMENTARY MATERIAL

Supplementary material associated with this article can be found, in the online version, at https://oss.signavitae.com/mre-signavitae/article/1673247857667129344/attachment/Supplementary%20material.docx.

REFERENCES

[1] Fleischmann C, Scherag A, Adhikari NK, Hartog CS, Tsaganos T, Schlattmann P, *et al*. Assessment of global incidence and mortality of hospital-treated sepsis. Current estimates and limitations. American Journal of Respiratory and Critical Care Medicine. 2016; 193: 259–272.

[2] Martin-Loeches I, Levy MM, Artigas A. Management of severe sepsis: advances, challenges, and current status. Drug Design, Development and Therapy. 2015; 9: 2079–2088.

[3] Rhodes A, Evans LE, Alhazzani W, Levy MM, Antonelli M, Ferrer R, *et al*. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock: 2016. Intensive Care Medicine. 2017; 43: 304–377.

[4] Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, *et al*. The third international consensus definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016; 315: 801–810.

[5] Morgan RW, Fitzgerald JC, Weiss SL, Nadkarni VM, Sutton RM, Berg RA. Sepsis-associated in-hospital cardiac arrest: Epidemiology, pathophysiology, and potential therapies. Journal of Critical Care. 2017; 40: 128–135.

[6] Churpek MM, Snyder A, Han X, Sokol S, Pettit N, Howell MD, *et al*. Quick sepsis-related organ failure assessment, systemic inflammatory response syndrome, and early warning scores for detecting clinical deterioration in infected patients outside the intensive care unit. American Journal of Respiratory and Critical Care Medicine. 2017; 195: 906–911.

[7] Gole AR, Srivastava SL, Neeraj. Prognostic accuracy of SOFA score, SIRS criteria, NEWS and MEWS scores for in-hospital mortality among adults admitted to ICU with suspected sepsis. Journal of the Association of Physicians of India. 2020; 68: 87.

[8] Khwannimit B, Bhurayanontachai R, Vattanavanit V. Comparison of the accuracy of three early warning scores with SOFA score for predicting mortality in adult sepsis and septic shock patients admitted to intensive care unit. Heart & Lung. 2019; 48: 240–244.

[9] Raith EP, Udy AA, Bailey M, McGloughlin S, MacIsaac C, Bellomo R, *et al*. Prognostic accuracy of the SOFA score, SIRS criteria, and qSOFA score for in-hospital mortality among adults with suspected infection admitted to the intensive care unit. JAMA. 2017; 317: 290–300.

[10] Jekarl DW, Lee S, Kim M, Kim Y, Woo SH, Lee WJ. Procalcitonin as a prognostic marker for sepsis based on SEPSIS-3. Journal of Clinical Laboratory Analysis. 2019; 33: e22996.

[11] Pierrakos C, Velissaris D, Bisdorff M, Marshall JC, Vincent J. Biomarkers of sepsis: time for a reappraisal. Critical Care. 2020; 24: 287.

[12] Ryoo SM, Lee J, Lee Y, Lee JH, Lim KS, Huh JW, *et al*. Lactate level versus lactate clearance for predicting mortality in patients with septic shock defined by Sepsis-3. Critical Care Medicine. 2018; 46: e489–e495.

[13] Hyland SL, Faltys M, Hüser M, Lyu X, Gumbsch T, Esteban C, *et al*. Early prediction of circulatory failure in the intensive care unit using machine learning. Nature Medicine. 2020; 26: 364–373.

[14] Perng JW, Kao IH, Kung CT, Hung SC, Lai YH, Su CM. Mortality prediction of septic patients in the emergency department based on machine learning. Journal of Clinical Medicine. 2019; 8: 1906.

[15] Su L, Xu Z, Chang F, Ma Y, Liu S, Jiang H, *et al*. Early prediction of mortality, severity, and length of stay in the intensive care unit of sepsis patients based on Sepsis 3.0 by machine learning models. Frontiers in Medicine. 2021; 8: 664966.

[16] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. The TRIPOD Group. Circulation. 2015; 131: 211–219.

[17] Shankar-Hari M, Phillips GS, Levy ML, Seymour CW, Liu VX, Deutschman CS, *et al*. Developing a new definition and assessing new clinical criteria for septic shock: for the third international consensus definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016; 315: 775–787.

[18] Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, *et al*. The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. On behalf of the working group on sepsis-related problems of the european society of intensive care medicine. Intensive Care Medicine. 1996; 22: 707–710.

[19] Moreno R, Vincent JL, Matos R, Mendonça A, Cantraine F, Thijs L, *et al*. The use of maximum SOFA score to quantify organ dysfunction/failure in intensive care. Results of a prospective, multicentre study. Working group on sepsis related problems of the ESICM. Intensive Care Medicine. 1999; 25: 686–696.

[20] de Mendonça A, Vincent JL, Suter PM, Moreno R, Dearden NM, Antonelli M, *et al*. Acute renal failure in the ICU: risk factors and outcome evaluated by the SOFA score. Intensive Care Medicine. 2000; 26: 915–921.

[21] Jones M. NEWSDIG: the national early warning score development and implementation group. Clinical Medicine. 2012; 12: 501–503.

[22] Subbe CP, Kruger M, Rutherford P, Gemmel L. Validation of a modified early warning score in medical admissions. QJM: Monthly Journal of the Association of Physicians. 2001; 94: 521–526.

[23] Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials—a practical guide with flowcharts. BMC Medical Research Methodology. 2017; 17: 162.

[24] Van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. Journal of Statistical Software. 2011; 45: 1–67.

[25] Liu FT, Ting KM, Zhou Z. Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data. 2012; 6: 1–39.

[26] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning. 2002; 46: 389–422.

[27] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, *et al*. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence. 2020; 2: 56–67.

[28] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, *et al*. Lightgbm: a highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems. 2017; 30: 3146–3154.

[29] Gregorutti B, Michel B, Saint-Pierre P. Correlation and variable importance in random forests. Statistics and Computing. 2017; 27: 659–678.

[30] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016; 785–794.

[31] Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST). 2011; 2: 1–27.

[32] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521: 436–444.

[33] Snoek J, Larochelle H, Adams RP. Practical bayesian optimization of machine learning algorithms. ArXiv. 2012; 12: 2951–2959.

[34] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988; 44: 837–845.

[35] Rufibach K. Use of Brier score to assess binary predictions. Journal of Clinical Epidemiology. 2010; 63: 938–939.

[36] Davis J, Goadrich M. The relationship between precision-recall and ROC curves. Proceedings of the 23rd International Conference on Machine Learning. 2006; 233–240.

[37] Lee SG, Song J, Park DW, Moon S, Cho H, Kim JY, *et al*. Prognostic value of lactate levels and lactate clearance in sepsis and septic shock with initial hyperlactatemia: a retrospective cohort study according to the Sepsis-3 definitions. Medicine. 2021; 100: e24835.

[38] Dimopoulos G, Rovina N, Patrani M, Antoniadou E, Konstantonis D, Vryza K, *et al*. Past history of stage I/II solid tumor malignancy impacts considerably on sepsis mortality: a propensity score matching analysis from the hellenic sepsis study group. BMC Infectious Diseases. 2019; 19: 831.

[39] Weng L, Zeng X, Yin P, Wang L, Wang C, Jiang W, *et al*. Sepsis-related mortality in China: a descriptive analysis. Intensive Care Medicine. 2018; 44: 1071–1080.

[40] Karlsson A, Stassen W, Loutfi A, Wallgren U, Larsson E, Kurland L. Predicting mortality among septic patients presenting to the emergency department-a cross sectional analysis using machine learning. BMC Emergency Medicine. 2021; 21: 84.

[41] Cao C, Yu M, Chai Y. Pathological alteration and therapeutic implications of sepsis-induced immune cell apoptosis. Cell Death & Disease. 2019; 10: 782.

How to cite this article: Eun-Tae Jeon, Juhyun Song, Dae Won Park, Ki-Sun Lee, Sejoong Ahn, Joo Yeong Kim, *et al*. Mortality prediction of patients with sepsis in the emergency department using machine learning models: a retrospective cohort study according to the Sepsis-3 definitions. Signa Vitae. 2023; 19(5): 112-124. doi: 10.22514/sv.2023.046.