BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Machine learning-based prognostic model for 30-day mortality prediction in Sepsis-3

Md. Sohanur Rahman[1], Khandaker Reajul Islam[2], Johayra Prithula[1], Jaya Kumar[2*], Mufti Mahmud[3], Mohammed Fasihul Alam[4], Mamun Bin Ibne Reaz[5], Abdulrahman Alqahtani[6,7] and Muhammad E. H. Chowdhury[8*]

Abstract

Background Sepsis poses a critical threat to hospitalized patients, particularly those in the Intensive Care Unit (ICU). Rapid identification of Sepsis is crucial for improving survival rates. Machine learning techniques offer advantages over traditional methods for predicting outcomes. This study aimed to develop a prognostic model using a Stacking-based Meta-Classifier to predict 30-day mortality risks in Sepsis-3 patients from the MIMIC-III database.

Methods A cohort of 4,240 Sepsis-3 patients was analyzed, with 783 experiencing 30-day mortality and 3,457 surviving. Fifteen biomarkers were selected using feature ranking methods, including Extreme Gradient Boosting (XGBoost), Random Forest, and Extra Tree, and the Logistic Regression (LR) model was used to assess their individual predictability with a fivefold cross-validation approach for the validation of the prediction. The dataset was balanced using the SMOTE-TOMEK LINK technique, and a stacking-based meta-classifier was used for 30-day mortality prediction. The SHapley Additive explanations analysis was performed to explain the model's prediction.

Results Using the LR classifier, the model achieved an area under the curve or AUC score of 0.99. A nomogram provided clinical insights into the biomarkers' significance. The stacked meta-learner, LR classifier exhibited the best performance with 95.52% accuracy, 95.79% precision, 95.52% recall, 93.65% specificity, and a 95.60% F1-score.

Conclusions In conjunction with the nomogram, the proposed stacking classifier model effectively predicted 30-day mortality in Sepsis patients. This approach holds promise for early intervention and improved outcomes in treating Sepsis cases.

Keywords Machine learning, Stacking-based meta-classifier, 30-day mortality prediction, Sepsis, Prognostic model

*Correspondence:
Jaya Kumar
jayakumar@ukm.edu.my
Muhammad E. H. Chowdhury
mchowdhury@qu.edu.qa
[1] Department of Electrical and Electronics Engineering, University of Dhaka, Dhaka 1000, Bangladesh
[2] Department of Physiology, Faculty of Medicine, University Kebangsaan Malaysia, 56000 Kuala Lumpur, Malaysia
[3] Department of Computer Science, Nottingham Trent University, Clifton Lane, Nottingham NG11 8NS, UK
[4] Department of Public Health, College of Health Sciences, QU Health, Qatar University, Doha 2713, Qatar
[5] Department of Electrical Engineering, Independent University, Bangladesh, Dhaka, Bangladesh
[6] Department of Biomedical Technology, College of Applied Medical Sciences in Al-Kharj, Prince Sattam Bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia
[7] Department of Medical Equipment Technology, College of Applied, Medical Science, Majmaah University, Majmaah City 11952, Saudi Arabia
[8] Department of Electrical Engineering, Qatar University, Doha 2713, Qatar

Introduction

Sepsis is a severe medical condition that renders the human immune system very vulnerable to any form of infection, hence posing a significant threat to overall health. As a result, the internal organs of the patient experience significant harm. Sepsis is responsible for an annual mortality rate of 5.3 million individuals globally [1]. It can be defined as a state in which an infection triggers a combination of pathological and physiological disturbances in an individual's health [2]. The diagnostic technique for such a condition has significantly changed with the introduction of the Sepsis-3 definition. This updated approach, as outlined in the Third International Consensus Definitions for Sepsis and Septic Shock, released in February 2016, emphasizes the crucial facts about how infection is associated with organ failure. [3]. As stated by relevant studies conducted on the adult population, there has been a noticeable increase in the incidence of Sepsis among adults yearly. Concurrently, the fatality rate associated with this condition remains consistently high, ranging from 30 to 50% [4, 5]. Based on a study done in April 2017, a total of 15,722 fatalities were recorded either within the hospital setting or during 30 days following discharge [6]. Hence, the timely identification and diagnosis of Sepsis are of utmost importance, as they furnish vital insights for healthcare professionals to evaluate patients' status and enhance their chances of survival with expeditious and suitable therapies. The presence of intricate factors related to the unclear definitions of Sepsis syndrome, undetermined sources of infection, and the increasing mortality rates associated with Sepsis, there is a pressing need to develop a dependable and efficient predictive model to determine patients' health outcomes. The predictive models discussed can provide substantial evidence, facilitating informed decision-making in clinical judgment and promoting the efficient allocation of public healthcare resources.

A purely clinical approach can also be followed to minimize the mortality of Sepsis patients in a hospital. A systematic review [7] involving six experimental reports included a revised Sepsis protocol to distinguish the initial indicators of Sepsis and prompt the caregivers to act accordingly. Even though it did not affect the patients' hospital stay, it successfully reduced the mortality in Sepsis patients by 22.6%, following the mentioned protocol. Another retrospective study [8] followed a 1-h bundle of Sepsis care, which demonstrated a mortality of 18.0% in the subjects with and 30.3% without the bundle of care. But it often becomes cumbersome to follow all the protocols as well as the bundle of care that needs to be taken for every patient- that can be a daunting task for the caregivers. This is where Machine Learning (ML) and deep learning algorithms come into play, which can predict the mortality of Sepsis patients in a certain time window. As a result, it may help to reduce staff workloads and lead to better allocation of limited resources.

In contemporary times, the application of Artificial Intelligence (AI) and ML algorithms has significantly advanced in the biomedical domain. These advancements have proven crucial in illness identification and have provided valuable insights in clinical settings. It also contributed significantly to Corona Virus Disease 2019 (COVID-19) outbreak in 2020. A study proposed different image enhancement techniques and novel Unet architecture to assist in detecting COVID-19 from Chest X-ray (CXR) images, which were not dependent on usual nasal swabs [9]. As a result, it alleviated significant workload from the health caregivers [9]. Another study employed a mix of image data and clinical data employing nomograms to predict mortality risk prediction for COVID-19 patients [10]. Islam et al. [11] proposed a stacking-based architecture to predict early intensive care unit (ICU) requirements for critically ill COVID-19 patients. It proposed a scoring system for the patients in question, which could be an important factor in future works relating to ICU prediction. Chowdhury et al. [12] proposed a clinical work that consisted of an ML model, nomogram, and scoring method named LNLCA to use as an early warning tool for predicting the mortality risk of COVID-19 patients. In another study [13] blood biomarkers like Age, Lymphocyte count, D-dimer, CRP, and Creatinine (ALDCC), information acquired at hospital admission as core predictors to assist in improving health care using ML models. It achieved an Area Under the Curve (AUC) score of 0.987, 0.999, and 0.992 for the development and internal and external validation cohorts, respectively.

Predicting the mortality outcome of the patients within 30 days can easily maneuver caregivers to provide extensive care and service to the patients at risk of mortality. The death rate within 30 days for a cohort of 2874 individuals diagnosed with Sepsis was found to be 29.8% in clinical research [14]. The study prospectively examined the performance of a deep learning algorithm in predicting mortality during a 30-day timeframe. The Artificial Neural Network (ANN) achieved an AUC score of 0.873 by exploiting clinical and laboratory biomarkers, including blood pressure, heart rate, length of stay in the ICU, and hospital-related factors. A separate clinical investigation has indicated that utilizing blood-based biomarkers can effectively enable the early detection of mortality in patients diagnosed with Sepsis or Septic shock within a 28-day timeframe. The research encompassed a cohort of 66 individuals diagnosed with Sepsis or Septic shock, from whom 14 blood-based biomarkers were obtained within the initial 24 h following admission to the ICU.

The mortality rate was recorded at 25.6%. The findings of the study indicate that a combination of IL-6 (Interleukin-6), NT-proBNP (N-terminal prohormone of brain natriuretic peptide), and INR (International normalized ratio) may serve as a distinct set of variables that could potentially be used as an effective predictor for early mortality, surpassing the predictive capabilities of conventional scoring systems. Based on the observed predictive performance with an AUC of 0.890, it is plausible to suggest that these biomarkers can improve the clinical prognosis of seriously ill patients diagnosed with Sepsis or septic shock [15].

Kwon et al. [16] proposed machine-learning models utilizing quick Sequential Organ Failure Assessment (qSOFA) variables outperformed the qSOFA score in predicting three-day mortality among 447,926 emergency department (ED) patients with suspected infection, with an Area Under the Receiver Operating Characteristic Curve (AUROC) of 0.86 compared to 0.78 for qSOFA scores. This suggests the potential for improved accuracy in mortality prediction using machine-learning approaches in emergency departments. In their study, Kovach et al. [17] compared four Sepsis-related prognostic scoring systems in 10,981 adult patients with suspected infection. It found that multivariate scores like SOFA and National Early Warning Score (NEWS) outperformed simpler scores like qSOFA and Systemic Inflammatory Response Syndrome (SIRS) in predicting hospital mortality, ICU transfer, and ICU length of stay, emphasizing their potential utility in inpatient hospital settings. Machine learning-based models using MIMIC-III data accurately predict in-hospital death risk for Sepsis patients, with gradient boosting machines (GBM) showing the highest performance [18]. These models offer potential assistance in ICU clinical decision-making, potentially improving patient outcomes. A nomogram prediction model was developed to assess the prognosis of Sepsis patients with lung infection using data from the MIMIC-III database [19]. The nomogram outperformed other scoring systems and can help improve in-hospital survival by guiding treatment strategies for these patients. Van Doorn et al. [20] proposed to develop ML models for predicting 31-day mortality in Sepsis patients in the emergency department, outperforming internal medicine physicians and clinical risk scores. The models achieved higher sensitivity and diagnostic accuracy, indicating their potential for improved risk stratification. Yao et al. [21] developed a predictive model for in-hospital mortality in 3,713 postoperative Sepsis patients using Extreme Gradient Boosting (XGBoost) and Logistic Regression (LR), with XGBoost outperforming. This suggests the potential for ML in early warning systems for Sepsis after major surgeries. Yang et al. [22] created a user-friendly nomogram to predict 30-day mortality in Sepsis-associated encephalopathy (SAE) patients. Using data from MIMIC III, they developed a predictive model that outperformed existing systems, showing the nomogram's potential in evaluating SAE patient prognosis and guiding future treatments. In the study of Liu et al. [23] on septic patients, Neutrophil-to-Lymphocyte Ratio (NLR) and Interleukin-6 (IL-6) were identified as independent predictors of 28-day mortality. Combining NLR and IL-6 significantly improved the accuracy of mortality prediction, demonstrating their potential clinical relevance. A model for predicting 1-year mortality in Sepsis patients using Stochastic Gradient Boosting (SGB) outperforms traditional scoring systems, achieving an AUROC of 0.8039 in validation [24]. This suggests that customized models based on assembly algorithms like SGB offer improved accuracy for long-term mortality prediction in Sepsis compared to standard severity scores.

A retrospective study [25] on 799,522 ED patients using a gradient boosting model demonstrated a high predictive ability for early mortality (up to 2 days post-ED registration) based on data available at triage in the emergency department. This model can potentially improve patient categorization and resource allocation in EDs. Faisal et al. [26] showed an LR model for in-hospital mortality prediction based on a first blood test and physiological measurements compared favorably with alternative ML methods, showing good performance and robustness across two different hospitals. A real-time Early Warning System (EWS) was validated for predicting inpatient mortality risk, achieving a high accuracy rate. It utilizes 54,246 inpatient admission patient data from Electronic Medical Records (EMR) and offers timely alerts for clinicians, enhancing patient care and outcomes [27]. Brajer et al. [28] prospectively validate an ML model for predicting in-hospital mortality at admission. It used electronic health record data from various hospital cohorts, with in-hospital mortality rates ranging from 1.6% to 3.0%. The model demonstrated good discrimination with the area under the receiver operating characteristic curves ranging from 0.84 to 0.89 and the area under the precision-recall curves from 0.13 to 0.29, suggesting its feasibility for system-wide implementation. A study [29] of 445 septic patients found that six key variables were significant predictors of both 7-day and 30-day mortality. These variables collectively showed high sensitivity (0.84 for 7-day, 0.87 for 30-day) and negative predictive value (0.96 for 7-day, 0.95 for 30-day), suggesting their potential importance in future Sepsis mortality prediction tools, but further validation in diverse cohorts is needed. In a study of 2,510 Sepsis patients [30], support vector machine (SVM) and ANN AI models achieved the best discrimination

with AUC-ROC values of 0.69, suggesting their potential for improving Sepsis classification and prognosis, which could aid clinical decision-making. Soffer et al. [31] proposed an ML model developed for in-hospital mortality prediction at admission to medical wards, achieving an impressive AUC of 0.924, with a sensitivity of 0.88 and specificity of 0.83, demonstrating its potential to enhance clinical decision-making. The model also yielded a high negative predictive value of 0.99, offering valuable insights for patient care. Perng et al. [32] employed deep-learning algorithms to predict mortality in suspected infected patients in an emergency department. The Convolutional Neural Network plus SoftMax achieved an accuracy rate of 87.01% within 72 h and 81.59% within 28 days, surpassing other methods, including SIRS and qSOFA, offering valuable support for early critical patient identification.

Liu et al. [33] reported a study on Sepsis mortality prediction using novel heart rate n-variability (HRnV) measures. They found that the final predictive model, including vital signs, HRV parameters, and HRnV parameters, achieved an AUC of 0.77, outperforming established clinical scores indicating potential for rapid and accurate risk stratification. HRnV measures offer valuable inherent information for innovative ECG analysis and risk monitoring. Cheng et al. [34] proposed different ML models using dynamic vital signs predicted in-hospital mortality in septic patients with an accuracy of 90.5%, 81.7%, and 83.5% at 6-h lead time, and 82.8%, 75.9%, and 80.5% at 48-h lead time using convolutional neural networks (CNNs), long short-term memory (LSTM), and Random Forest (RF), respectively. Performance was best when the lead time was closer to the event. A Sepsis severity prediction model and score were developed and externally validated using administrative data from five US states by Ford et al. [35]. The model showed good to excellent discrimination (C-statistics ranged from 0.709 to 0.838), providing reliable risk adjustment for administrative data in severe Sepsis cases. Using a rule-based method on 2021 Sepsis ICU patients from MIMIC-III, 77 risk prediction rules were generated by Wu et al. [36]. A prediction model based on 62 of these rules achieved an average AUC of 0. outperforming existing methods. External validation on 1468 Sepsis patients further supported the superiority of this rule-based approach, highlighting the importance of factors like the Glasgow Coma Scale, serum potassium, and serum bilirubin in predicting patient mortality. This method not only predicts in-hospital deaths accurately but also enhances our understanding of Sepsis complexity. In a study with 200 ICU patients by Selcuk et al. [37], ML outperformed SAPS II and APACHE II, achieving 85.25% accuracy in Sepsis mortality prediction. The best ML method for SOFA's prediction accuracy at 73.47%. An ensemble of eight ML methods improved APACHE II performance by approximately 2%, highlighting ML's potential for superior ICU mortality prediction with fewer variables even in small datasets. Machine learning models, including LR, RF, XGBoost, and neural networks, were applied to predict in-hospital mortality for adult Sepsis patients using data from 923,759 hospitalizations [38]. Compared to LR, all ML models demonstrated superior discriminative ability (AUC: 0.878–0.893 vs. 0.786). ML models also showed higher sensitivity, specificity, positive predictive value, and negative predictive value. The Super Learner model yielded similar results. These findings suggest that ML approaches enhance mortality prediction accuracy and could support research on Sepsis care disparities and policy initiatives. Machine learning models were developed using data from 21,680 Sepsis-3 patients by Bao et al. [39]. The top three performing models in terms of AUC in the test set were light GBM, GBM, and XGBoost. The light GBM model, with adjusted parameters, achieved an impressive AUC of 0.99 in the train set and 0.96 in the test set, demonstrating its potential for accurate prediction of Sepsis patient mortality, thereby enhancing clinical decision-making and patient outcomes.

Zhang et al. [40] developed a mortality risk score for Sepsis-3 patients in ICU. The study encompassed a total of 5,443 patients diagnosed with Sepsis-3, of which 16.7% had mortality. The Sepsis mortality risk score categorizes admitted patients into four risk groups: low risk (3.2%), moderate risk (12.4%), high risk (30.7%), and very high risk (68.1%). Based on the decision curve analysis conducted in the study, the scoring scheme demonstrated a net positive advantage. Despite several predictive prognostic models have been developed to identify patients at an elevated risk of mortality from Sepsis at an early stage, there is a notable deficiency in the availability of robust and relevant ML algorithms that can accurately identify the most significant predictive markers for patient mortality. The important aspects of successful resource allocation and treatment planning are identifying and prioritizing patients with serious risks. In light of this observation, it is conceivable to implement a system for ongoing surveillance of patients at high risk during their hospital stay, which would involve the early prediction of clinical outcomes. Similarly, reducing admissions for patients with minimal risk of complications can effectively mitigate the strain on healthcare facilities. This specific study could have been better if it had more clinical information related to deceased outcomes as well as missing values. Some biomarkers were null without having a measure of any information. Another stage of evaluation of outcomes could have greatly improved the level

of study as it only included outcomes within the 30-day timeframe but no information of any specific outcomes after 30 days of evaluation.

The main aim of this study is to construct a machine-learning framework utilizing classical machine learning algorithms that can efficiently evaluate the risk of mortality in patients diagnosed with Sepsis or septic shock within 30 days. This will help offer them the appropriate healthcare support. To accomplish the aforementioned objective, advanced ML methods were adopted to estimate predictors for Sepsis mortality within a 30-day timeframe. The study involved the identification of the predominant biomarkers from a dataset containing 77 biomarkers. These biomarkers were found to substantially impact differentiating between the survival and mortality of Sepsis patients. The aforementioned high-ranking characteristics were employed to estimate a nomogram based on multivariable LR in conjunction with the most effective classification model. This nomogram was later subjected to validation to assess its predictive capabilities. Furthermore, an open-source software prototype has been developed and hosted on a cloud server for public use and validation. This prototype can be accessed at http://34.16.212.11/sepsis_morality. This platform allows for real-time validation and provides transparency regarding the model's predictive capabilities.

In this work, we concentrated on developing a stacking-based meta-classifier that integrates classical machine learning methods, thereby optimizing performance while maintaining model interpretability. The key contributions of this study are outlined below:

- We have incorporated data balancing techniques, specifically SMOTE-TOMEK LINK, to address class imbalance, thereby ensuring robust and unbiased model performance.
- We have employed advanced feature ranking methods, including XGBoost, Random Forest, and Extra Tree, to meticulously select 15 key biomarkers, ensuring precise and reliable predictions.
- We have developed a stacking-based meta-classifier, utilizing the top-performing models to significantly enhance the prediction accuracy for 30-day mortality in Sepsis-3 patients.
- We have integrated a nomogram into our framework, providing clear clinical interpretability of the biomarkers' significance, which aids in practical decision-making for healthcare providers.

The article proceeds with a detailed examination of the methodology, which is presented in the following section. This methodology section delves into the theoretical foundations of each algorithm. Subsequently, the results obtained from the algorithms described in the methodology are presented. This is followed by a thorough and critical discussion of the results, including clinical justifications. Finally, the article concludes with a summary of the overall work.

Methodology

The methodology of this project followed a systematic approach to address the dataset's challenges and successfully arrived at a valid conclusion with the aid of essential algorithms and methods. The initial analysis stage involved basic preprocessing of the Sepsis dataset, which included filling data gaps and converting categorical data into numerical values. Feature values were standardized using Min–Max normalization to ensure consistency. Subsequently, XGBoost, RF, and Extra Tree (ET) algorithms were employed to rank and select the relevant features. Various classical ML techniques were used to train the models in this study. Additionally, a multivariate LR-based Nomogram was created to validate the research findings. Model explainability was assessed through the SHAP summary graphic, which is also addressed in this methodology. While it was the final step, it played a crucial role. An overview diagram in Fig. 1 provides a comprehensive visual representation of all the procedures conducted in this study.

Dataset description and preprocessing

The dataset utilized in this experiment was shared by Hou et al. [41]. The studied dataset had been sufficiently and appropriately cleaned in advance. The three columns, namely patient ID, ICU admission ID, and first service used during clinic admission, were excluded from the analysis due to their lack of feature value. Several biomarkers, such as the thirty-day mortality, blood culture positivity, gender, presence of metastatic cancer, diabetes, vent, quick Sepsis-related organ failure assessment (qSOFA) score, respiration score, and Glasgow Coma Scale (GCS) score, are essential for identifying Sepsis patients at high risk of mortality. These biomarkers are represented in categorical Boolean form. ML models are limited in recognizing and processing non-numerical examples, as they rely on mathematical modeling. Consequently, it becomes necessary to translate these instances into a numerical binary representation. Following the conversion process, any instances of missing data were first checked whether missing-at-random and addressed by employing Multivariate Imputation using Chained Equations (MICE) imputation technique [42, 43]. This method replaced missing values using mathematical averaging and regression techniques, specifically within a

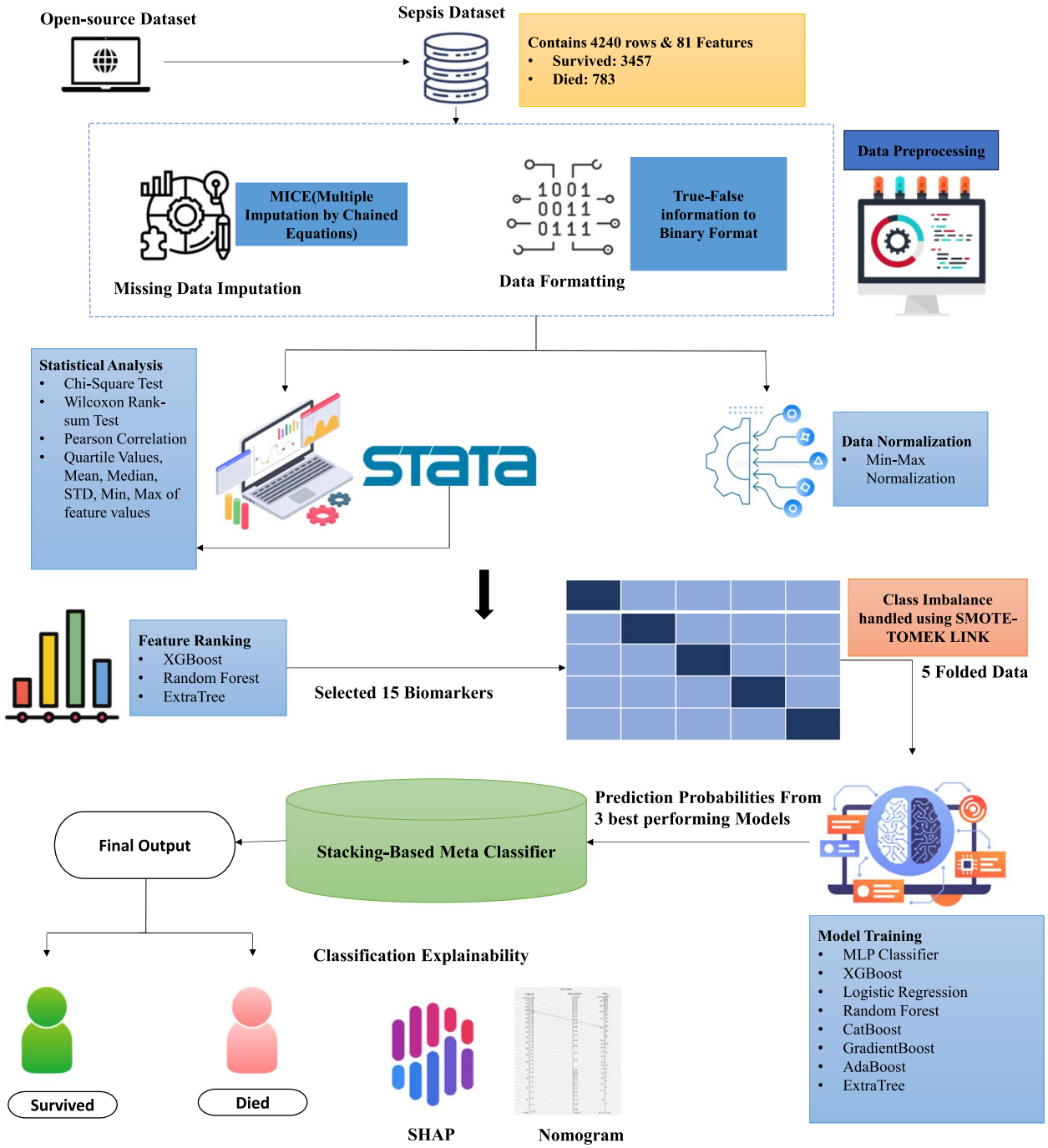

Fig. 1 Step-by-step overview of the methodology for 30-Day Mortality Prediction in Sepsis-3

single-column vector. The expiration flag of thirty days served as the ground truth label in this project, functioning as an indicator of patients' survival or mortality. The distribution of our ground truth labels exhibited significant imbalance across both classes, with 3,457 for the survivor class and 783 for the death class.

Statistical analysis

Statistics for each key biomarker used in training a robust model included the count of missing values, the mean, the degree of data deviation for each class, as well as the minimum–maximum range and interquartile values. Two statistical techniques were employed to examine

the link between the provided biomarkers and the target trait: the Wilcoxon Rank-sum method and the Chi-squared test, and the statistical significance of results was reported in terms of p-values [44]. Given the likely outliers in the biomarkers, a non-parametric (Wilcoxon Rank-sum) approach was adopted.

The mean of a specific attribute denotes the measure of central tendency and population average for the corresponding column. However, the mean can often yield biased results when there are outliers in the sample. In this scenario, the median is utilized to obtain the precise middle value of a variable that is not influenced by extreme values. The terms "min" and "max" refer to the lowest and highest values obtained by a specific predictor. The quartile values Q1 and Q3 determine the 25th and 75th percentiles of a given set of data points, respectively.

The Chi-square test focuses on binary outcomes, establishing null and alternative hypotheses. At the same time, the Wilcoxon Rank-sum method assesses categorical data, combining samples and rank sums to calculate p-values. The statistical analyses were carried out using Stata/MP 15.0 software.

Data normalization

The presence of a varied range of data in independent columns, along with uneven scaling, can potentially introduce bias in the outcomes of classification models. The predictive performance of a classification task can be estimated by examining the results obtained when the model undergoes training on a dataset that closely resembles the target dataset in terms of all variables and possesses correct labeling [45]. Hence, the normalization process is imperative to ensure a sound model training process and provide a smooth outcome.

To fit the entire dataset into the range of 0 and 1, the succeeding mathematical formula was applied to do the scaling of the data based on the Min–Max normalization process:

$$x_{scaled} = \frac{x_{f\_}x_{min}}{x_{max} - x_{min}}$$

where x_{scaled} is scaled feature value, x_f, x_{min} and x_{max} is the feature in question and the minimum and maximum value of that specific feature. Min–Max normalization keeps the underlying relationship between data points by scaling them down to a range of values between 0 and 1. Since this study involves a binary classification problem, it is important to ensure that the values in the training set fall within the range of 0 to 1. This helps to achieve consistency in predictions. In addition, Min–Max normalization is not susceptible to outlier values.

Feature ranking

Feature ranking and selection offer a significant advantage in enhancing research outcomes and optimizing time consumption through the use of the most pertinent feature subsets for model training. Tree-based algorithms like XGBOOST, Random Forest, and Extra Tree are widely used for feature ranking and selection because of their ability to effectively handle big datasets and apply dimensionality reduction. Tree-based algorithms utilize splitting to manipulate more relevant sets of information for the training process. Applying all three ranking algorithms can offer a complete viewpoint on the significance of features in the research. XGBOOST operates with weak learners, while Random Forest and Extra Tree include randomization in the decision-making process. While the dataset had 77 independent feature variables and 1 ground truth label, we analyzed 15 common features as core features based on the feature ranking effort carried out during this study. The relative significance of each feature is given by the feature ranking technique, which determines the score given the conditions in which the model underwent training.

XGBoost

XGBoost is a highly adaptable and powerful ML tool that has broad usage across numerous tasks, such as feature ranking and assessment of feature relevance. XGBoost has an inherent mechanism for computing feature significance scores, enabling the user to gain insights into the relative contributions of different features towards the performance of its predictive capacity proven by the model. XGBoost employs a system that ranks features according to their significance in facilitating precise predictions inside a gradient-boosting ensemble. Determining feature relevance in XGBoost involves an evaluation of the frequency with which a feature is utilized for data splitting across all trees in the ensemble, as well as the extent to which each feature helps decrease the error (loss) function. Features that are utilized with greater frequency and provide a more substantial contribution to mistake reduction are considered of greater importance. The hyperparameters used to perform this endeavor were: max_depth = 4, which represents the longest path in a tree-based network; learning_rate = 0.2, which employs a slow running rate to prevent overfitting and reduce prediction losses; reg_lambda = 1, which is the default value in the model; n_estimators = 150, representing a moderate number of boosting rounds the model will go through; and subsample parameter, which is crucial for utilizing training samples in each boosting round. Typically, the default value for this parameter is

1. However, in our study, we chose a subsample value of 0.9 to introduce some variability into the model through incorporating just a portion of the training data. The variable subsample_bytree operates in a same manner as subsample, except it specifically involves the columns. In our study, we assigned the variable the same value as the subsample.

Random forest

In our study, the RF algorithm was used not only as a classical tool but also for feature ranking tool by leveraging its ability to assign relevance scores to features. These relevance scores are determined by means of each feature's contribution to reducing Gini impurity, when data splitting is processed across the decision trees in the forest. Features that consistently lead to significant reductions in impurity are considered more important and which in turn provide more relevance for the model's predictions. Random forest's usage as a ranking tool is advantageous because it employs the feature ranking into the overall classification framework. This two-pronged approach ensures model's accuracy and interpretability by focusing on the most impactful features. In our study, we chose Random Forest for feature ranking because it offers a reliable, interpretable, and computationally efficient way to identify the most relevant features. This method allowed us to focus our analysis on the key variables that contribute most significantly to the prediction task, thereby enhancing the overall accuracy and robustness of our model. Additionally, the feature ranking provided by Random Forest helped in reducing the complexity of our model by eliminating less important features, which could otherwise contribute to overfitting or increase computational costs without adding significant predictive value. Random Forest has been widely used as an effective feature ranking tool in various studies, demonstrating its robustness and accuracy in identifying key predictive features across many literatures [46, 47]. In our study, we chose Random Forest for feature ranking due to its reliability and computational efficiency in identifying the most relevant features. This allowed us to focus our analysis on the key variables that contribute most significantly to the prediction task, thereby enhancing the overall accuracy and robustness of our model. Additionally, the feature ranking provided by Random Forest helped in reducing the complexity of our model by eliminating less important features, which could otherwise contribute to overfitting or increase computational costs without adding significant predictive value.

Extra tree

The ET algorithm shares similarities with the RF algorithm while ranking features based on importance. The algorithm constructs a collection of decision trees inside an ensemble framework and determines the relevance of attributes in terms of their ability to facilitate accurate predictions. The ranking of features in ET is established based on their respective contributions in minimizing impurity or error throughout the data partitioning process in the decision trees that comprise the ensemble. Extra Tree went through the default hyper-parameter values without any specific value change.

Model training and performance metrics

After employing the three previously discussed strategies to extract meaningful biomarkers, a total of eight conventional machine-learning models were trained. The dataset was partitioned into an 80:20 ratio, with 80% allocated for training and 20% for testing. This process was repeated five times using Stratified k-Fold cross-validation from the scikit-learn (sklearn) package. This cross-validation technique employs a stratified sampling scheme, maintaining the original class ratio in the divided sets. Thus, each fold's training and testing sets preserve the same ratio of positive and negative class samples as the original dataset. This method ensures a more generalized training stage for each fold. Since 20% of the data was used for testing in each fold and 5 folds were utilized, the entire dataset was used for testing at some point during the training phase. The ground truth labels were greatly imbalanced. Hence, the training dataset was refined by employing the SMOTE-TOMEKLINK oversampling technique to raise the number of data points associated with the minority class. This outcome was achieved using the application of Tomek linkages for data cleaning purposes. This cleaning stage followed in Tomek linkages effectively find and get rid of the samples which have close resemblance between the majority and the minority class. The integration of incrementing and decrementing data points algorithm was achieved using the Synthetic Minority Over-sampling Technique (SMOTE) in conjunction with Tomek connections, which nullified class imbalance in the training set [48]. The rationale behind employing such algorithm is to remove any data points that might pose ambiguity between different classes and as well as successfully handle the imbalance problem. Hyperparameter used in this algorithm was only sampling_strategy = 'majority' which would down-sample the instances of the majority class equal to the minority class.

The training phase was employed on a total of 9 classical ML models. The Multi-Layer Perceptron (MLP) classifier is an artificial neural network (ANN) commonly employed in supervised ML applications, particularly to solve classification problems related to the current study. The neural network in question is classified as a feed-forward network, recognized by the characteristics of

a single directional stream of data from the input layer, via intermediate hidden layers, and finally to the output classification layer, without forming any cyclic connections [49]. Another training model was XGBoost, which is an extremely robust ML algorithm that finds extensive application in the domain of classification tasks. The proposed methodology is an ensemble learning technique that blends in the predictions generated by numerous decision trees to enhance the accuracy of predictions [50]. Our study also included LR classifier, while rooted in traditional statistics, has become a cornerstone in the field of ML algorithm for binary classification tasks, wherein the objective is to forecast one of two potential outcomes, which in our case is the prediction of alive or death condition of the patients [51, 52]. In this classifier, classification is done by representing the probability of a given input belongs to a specific class with the help of logistic function. The model calculates the weighted sum of the input features, which is then subsequently converted into probability. The class determination process is done by comparing the probability with a threshold of 0.5. Probability score more than 0.5 indicates the positive class and less than this threshold dictates negative class.

The RF algorithm is one of the most prolifically used algorithms for classification tasks. This specific algorithm is a type of ensemble learning method that merges many decision trees. In this approach, each specific tree is trained on a randomly selected portion of the available data pointers and follows a randomized selected subset of the available features. Using randomness in the model improves its robustness and capacity for generalization [53]. The algorithm referred to as "Extra Tree," which is an expression for Extremely Randomized Trees, is an ML technique that bears a strong connection to the RF algorithm. Similar to RF, ET is an ensemble learning technique employed for classification and regression applications. ET is an ML algorithm that constructs a collection of decision trees. Yet, it defines itself by adding a further degree of randomization across the tree-building approach, boosting its immunity to overfitting [54]. Furthermore, AdaBoost was also used, which is an abbreviated form of Adaptive Boosting. In classification, the AdaBoost algorithm is utilized to aggregate the outcomes found by a collection of weak classifiers, often decision trees with restricted depth, to construct a powerful classification tool. The AdaBoost algorithm is structured to improve the accuracy of weak classifiers by assigning greater importance to data points that have been wrongly labeled in preceding training cycles [55]. Another tool used in this study was Gradient Boosting (GB) classifier, more precisely Gradient Boosting Trees, which is a highly efficient ensemble learning method that develops an additive model in a step-by-step way by iteratively training weak learners, often decision trees with restricted depth. Each learner is trained to rectify the errors caused by the preceding learners. This phenomenon leads to the emergence of a robust learner that amalgamates the individual strengths exhibited by numerous learners with relatively lower performance [56]. Catboost is recognized for its user-friendly interface, efficient functionality, and strong capability to manage categorical variables. The term "CatBoost" is derived from the phrase "Categorical Boosting," which signifies its proficiency in handling categorical data [57]. Lastly, we also made use of Decision Tree (DT) classifier; a type of supervised learning algorithm commonly used for classification tasks. Decision trees operate by recursively splitting the dataset based on specific features that provide the highest information gain, creating a tree-like structure where each internal node represents a decision, and each leaf node represents a class label. Specifically, we used a decision tree classifier with the criterion='entropy' parameter, which aligns closely with the principles of the C4.5 algorithm, a well-known method that uses entropy to guide the splitting process at each node. This approach allowed us to replicate the core aspects of C4.5, ensuring that our model effectively handles classification tasks by focusing on reducing impurity and maximizing information gain at each step [58]. Our selection of classification models, including the entropy-based decision tree classifier, was guided by a thorough review of the literature, where such models have consistently demonstrated strong performance in medical data classification and prediction tasks. By adhering to established methodologies from previous studies, we ensured that our approach was both grounded in proven techniques and capable of meaningful comparisons with existing research, thereby enhancing the robustness and relevance of our findings [46, 59].

The evaluation of the results obtained during the training phase was not limited solely to accuracy. The objective of the findings was to provide a rationale for the outcome indicated by the models. Considering this, various evaluation metrics were employed to substantiate the validity of the outcomes, as Accuracy alone was not a reliable option [60–63]. Several metrics used in calculating the performance of the model can be seen in the following equations:

$$Accuracy_{class\_x} = \frac{TP_{class_x} + TN_{class\_x}}{TP_{class_x} + TN_{class\_x} + FP_{class\_x} + FN_{class\_x}} \quad (1)$$

$$Precsion_{class\_x} = \frac{TP_{class\_x}}{TP_{class\_x} + FP_{class\_x}} \quad (2)$$

$$Recall_{class\_x} = \frac{TP_{class\_x}}{TP_{class\_x} + FN_{class\_x}} \quad (3)$$

$$Specificity_{class_x} = \frac{TN_{class_x}}{TN_{class_x} + FP_{class_x}} \tag{4}$$

$$F1 - Score_{class_x} = 2\frac{Precsion_{class_x} \times Recall_{class_x}}{Precsion_{class_x} + Recall_{class_x}} \tag{5}$$

where *class_x* = *Suvival and Death.*

In this context, TP, TN, FP, and FN represent True Positive, True Negative, False Positive, and False Negative, respectively. These evaluation metrics provide a thorough and rigorous assessment of a model's performance from various perspectives. Accuracy measures the overall correctness of the model's predictions compared to the actual labels, but this alone does not suffice to evaluate performance quality, particularly in cases of class imbalance. Precision and recall evaluate the accuracy of positive class predictions and the model's ability to capture all positive instances, respectively. Specificity focuses on the accuracy of negative class predictions. The F1-score balances precision and recall. It is crucial to consider these indicators collectively to achieve a comprehensive understanding of the model's performance, as considering all metrics provides a generalized outcome based on the weighted performance of all classes as well as the performance of individual classes.

The use of such evaluation metrics is firmly established in many biomedical researches. Some notable works may include studies such as Yong et al. [64] who used accuracy to validate their performance and Zhang et al. [65] who made of AUC and F1 score to explain their clinical findings. Our study showcased the utilization of ROC, AUC, and confusion matrix as tools to gain insights into the classification performance. The AUC metric serves as a measure of the quality of Binary Classification. A higher AUC score indicates a better outcome in terms of classification. On the other hand, the confusion matrix offers insights into the distribution of performance for each class.

Validation using logistic regression-based nomogram in the mortality prediction

A nomogram based on LR was generated using Stata/MP software version 15.0, utilizing the Nomolog plugin developed by Alexander Zlotnik [66]. The Logit function in Stata is utilized to estimate LR, which serves as a binary classifier. The Binary LR model provides a probabilistic prediction for two distinct groups. In the context of this study, the label '1' denotes the death class, whereas the label '0' represents the survived class. LR is especially important as it is coherent in retrieving probability outcomes for survived and death classes [67]. The probability of an event occurring (P_x) can be expressed as a fraction, where the numerator represents the likelihood of the event happening, and the denominator represents the likelihood of it not happening. The aforementioned relationship is seen in Eq. 6. The characterization of odds differs from that of probabilistic outcomes, as odds can range from 0 to infinity, but probabilistic outcomes are limited from 0 to 1.

When the LR is in the workflow, the logarithm of probabilities can be expressed as a linear result, incorporating several feature vectors. These feature vectors may include binary variables, such as the gender of patients, as well as continuous variables, like age and blood lactate count. The Log-odds, also known as the linear outcome (LO) in Eq. 7, represents the probabilistic value associated with a specific occurrence. Equations (6–9) are utilized to establish the foundational framework for determining the probability (P_x) of adverse outcomes through the application of LR.

$$Odds = \frac{P_x}{1 - P_x} \tag{6}$$

$$LO = \ln(Odds) = \ln\left(\frac{P_x}{1 - P_x}\right) = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n \tag{7}$$

$$\frac{P_x}{1 - P_x} = e^{b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n} = e^{LO} \tag{8}$$

$$P_x = \frac{e^{LO}}{1 + e^{LO}} = \frac{1}{1 + e^{-LO}} \tag{9}$$

The Nomogram architecture was applied to incorporate the top-selected features identified by applying three ranking methodologies. A calibration belt curve was produced with internal validation to assess the calibration performance of LR. The present study additionally employed Decision Curve Analysis (DCA) to distinguish the limiting value associated with each of the features. All the aforementioned areas of study were conducted by applying the statistical software, Stata.

Model explainability

Given its relevance and applicability across diverse practical contexts, examining the underlying reasoning behind a model's specific forecasted result is often imperative. However, the interpretability of sophisticated models, such as ensemble or deep learning models, remains a daunting task even for professionals, despite their capability of achieving high accuracy on large and modern datasets. This situation poses a quandary in which the pursuit of precision clashes with the capacity to shed light on the rationale behind the model's outcomes. To tackle this matter, numerous algorithms have been recently established to assist users in interpreting the results of complex models. However, there is still a lack of clarity regarding the connections between these strategies and

the specific situations in which one method should be prioritized over another.

The SHAP framework, known as SHapley Additive exPlanations, is a broad framework and collection of approaches employed to facilitate the interpretability of models, specifically within the domain of ML and predictive modeling. The principal objective of the SHapley Additive exPlanations (SHAP) framework is to help accommodate users in comprehending how the input features of a model contribute to its predictions. The fundamental goal of the SHAP framework is to provide insights for individual predictions extracted by an ML model. Rather than offering a broad overview of the model's behavior, this specific interpretable architecture exploits the game theory to its purpose by integrating Shapley values in the work. The Shapley values assign a significance value to each feature inside a predicted outcome, quantifying the extent to which each feature has affected the overall prediction. The aforementioned values are derived from the principle of equity in allocating credit for a prediction across several attributes [68, 69].

In our study, SHAP values were calculated by considering every possible combination of features and their average marginal contribution across those features. We used the SHAP framework to gain insights into the model's predictions for various attributes. To provide a comprehensive perspective, SHAP values were analyzed at both the global and local levels. Global SHAP analysis examined the influence of each parameter across the entire dataset, uncovering significant trends during model training. This helped us understand how different features generally impacted on the model's predictions. On the other hand, local SHAP analysis focused on individual predictions, calculating feature contributions for specific prediction points. This enhanced our ability to interpret and explain the model's decision-making process on a case-by-case basis.

Results

A comparative analysis of the efficacy of classical ML methods and a meta-classifier based on stacking was reported in this section. The validation process was further investigated using a Nomogram based on LR, and the impact on clinical outcomes was evaluated by analyzing the top features.

Statistical analysis

The dataset has 77 biomarkers, including 15 characteristics, patient gender, and outcome variables. These variables were organized in a table using several feature selection algorithms. The biomarkers encompass various factors such as the patients' age, minimum and average blood lactate count, SOFA score (a crucial predictor of ICU Mortality), average rate of respiration by the patients, length of stay for both ICU and Hospitals, minimum and average blood urea nitrogen, maximum temperature recorded in Celsius, percentage of oxygen saturation, urinary output, and finally, two scoring methods. The Elisxhauser comorbidity index and the logistic organ dysfunction score were included in the variables list, which were utilized as the independent features for training the model. For these specific biomarkers, Chi-square test was applied in discrete or categorical valued features such as Gender, and Wilcoxon Rank-sum test was applied in continuous valued features which were the rest of the features the models were trained on. To imply the statistical significance level, 0.05 or less than that value of p indicates the events are not due to the randomness of the distribution rather it has strong statistical association.

The investigation involved 41.89% female participants and 58.11% male. In the survival and expiry classes, where males dominated in percentage (58.52% and 56.32% respectively for the Survived and Death classes, each), females showed 41.48% and 43.68% in that order. According to the Chi-square test results, the patients' gender has a weak association with the target feature. The majority of the passed-away patients were between the ages of 68.72 ± 15.06, while the survivors were between 56.89 ± 16.83, meaning that older patients suffered the majority of the fatalities.

Table 1 contains the patients' major characteristics, including age and gender. As all of the listed biomarkers were extracted using the appropriate feature selection algorithm, all of them demonstrated a high statistically significant association with the target feature with the target variable as they show a p-value less than 0.05 except for gender. Gender was not extracted from the selection algorithm.

Feature ranking

A variety of feature rating algorithms were applied to correspond with the project. Numerous elements that were shared by all of those strategies were made possible by these ranking systems. To perform this, the XGBoost, RF, and ET methods were used. By assessing each feature's impact, these models reduce loss and boost prediction accuracy. Finally, we identified the characteristics in question that had been statistically explained in the statistical analysis stage by taking into account a total of 25 top features from individual models and applying the set intersection approach, which resulted in 15 relevant biomarkers left for model training. Figure 2 displays the bar plots for the previously described ranking models. The bars' length denotes the significance of that particular feature shown over the mortality prediction in 30 days.

Table 1 Statistical overview of the predictor biomarkers of Sepsis patients

	Item	Survived	Death	Total	Technique	Test Statistic	*P*-value
1	Gender				Chi-square test	$X^2 = 1.27$	0.261
	• Male (%)	2023(58.52%)	441(56.32%)	2464(58.11%)			
	• Female (%)	1434(41.48%)	342(43.68%)	1776(41.89%)			
2	Age				Rank-sum test	Z = -10.564	< 0.001
	• N(missing)	3457(0)	783(0)	4240(0)			
	• Mean ± SD	56.89 ± 16.83	68.72 ± 15.06	63.15 ± 16.73			
	• Median	53.61	71.64	65.01			
	• Q1, Q3	51.65, 75.26	58.78, 81.32	55.80, 77.04			
	• Min, Max	17.24, 88.97	16.78, 88.98	16.78, 88.98			
3	Blood Lactate Minimum				Rank-sum test	Z = -12.60	< 0.001
	• N(missing)	3457(0)	783(0)	4240(0)			
	• Mean ± SD	1.55 ± 0.82	2.43 ± 2.11	1.71 ± 1.22			
	• Median	1.4	1.7	1.4			
	• Q1, Q3	1, 1.9	1.2, 2.8	1, 2			
	• Min, Max	0.3, 12.1	0.4, 20.3	0.3, 20.3			
4	Sequential Organ Failure Assessment				Rank-sum test	Z = -17.48	< 0.001
	• N(missing)	3457(0)	783(0)	4240(0)			
	• Mean ± SD	5.24 ± 2.88	8.03 ± 4.3	5.76 ± 3.37			
	• Median	5	7	5			
	• Q1, Q3	3, 7	5, 11	3, 7			
	• Min, Max	2, 21	2, 21	2, 21			
5	Blood Lactate Mean				Rank-sum test	Z = -12.74	< 0.001
	• N(missing)	3457(0)	783(0)	4240(0)			
	• Mean ± SD	2.13 ± 1.3	3.47 ± 2.90	2.38 ± 1.77			
	• Median	1.8	2.4	1.9			
	• Q1, Q3	1.3, 2.55	1.6, 4.25	1.35, 2.75			
	• Min, Max	0.3, 16.8	0.4, 20.85	0.3, 20.85			
6	Average Respiration Rate				Rank-sum test	Z = -12.68	< 0.001
	• N(missing)	3456(1)	783(0)	4239(1)			
	• Mean ± SD	19.48 ± 4.07	21.74 ± 4.71	19.9 ± 4.3			
	• Median	18.84	21.3	19.24			
	• Q1, Q3	16.53, 21.8	18.1, 24.7	16.8, 22.41			
	• Min, Max	9.54, 40.37	10.17, 40.58	9.54, 40.58			
7	ICU Length of Stay				Rank-sum test	Z = -2.59	< 0.05
	• N(missing)	3457(0)	783(0)	4240(0)			
	• Mean ± SD	5.17 ± 6.8	5.15 ± 5.07	5.2 ± 6.5			
	• Median	2.79	3.39	2.88			
	• Q1, Q3	1.63, 5.82	1.69, 6.85	1.64, 6.01			
	• Min, Max	0.17, 101.74	0.31, 30.89	0.17, 101.74			
8	Minimum Blood Urea Nitrogen				Rank-sum test	Z = -14.7	< 0.001
	• N(missing)	3456(1)	782(1)	4238(2)			
	• Mean ± SD	23.67 ± 19.61	35.24 ± 26.19	25.8 ± 21.45			
	• Median	17	27	19			
	• Q1, Q3	12, 28	17, 45	12,32			
	• Min, Max	1, 182	3, 181	1, 182			

Table 1 (continued)

	Item	Survived	Death	Total	Technique	Test Statistic	*P*-value
9	Hospital Length of Stay				Rank-sum test	Z = 15.77	< 0.001
	• N(missing)	3457(0)	783(0)	4240(0)			
	• Mean ± SD	11.63 ± 10.71	7.015 ± 6.37	10.78 ± 10.21			
	• Median	8.54	5.16	7.94			
	• Q1, Q3	5.34, 14.15	2.1, 9.81	4.80, 13.42			
	• Min, Max	0.25, 206.43	-0.43, 30.44	-0.43, 206.42			
10	Max Temperature (Celsius)				Rank-sum test	Z = 7.086	< 0.001
	• N(missing)	3375(82)	764(19)	4139(101)			
	• Mean ± SD	37.66 ± 0.84	37.4 ± 1.14	37.61 ± 0.91			
	• Median	37.55	37.33	37.55			
	• Q1, Q3	37.05, 38.17	36.72, 38	37, 38.17			
	• Min, Max	32.94, 42	32.2, 40.94	32.2, 42			
11	International Standard Ratio Max				Rank-sum test	Z = -12.64	< 0.001
	• N(missing)	3236(221)	760(23)	3996(244)			
	• Mean ± SD	1.61 ± 1.34	2.12 ± 1.77	1.71 ± 1.45			
	• Median	1.3	1.5	1.3			
	• Q1, Q3	1.2, 1.6	1.2, 2.3	1.2, 1.7			
	• Min, Max	0.6, 24	0.9, 21.5	0.6, 24			
12	Minimum Oxygen Saturation (%)				Rank-sum test	Z = 10.19	< 0.001
	• N(missing)	3457(0)	783(0)	4240(0)			
	• Mean ± SD	92.01 ± 5.84	87.57 ± 12.57	91.19 ± 7.74			
	• Median	93	91	93			
	• Q1, Q3	91, 95	86, 94	90, 95			
	• Min, Max	11, 100	1,100	1, 100			
13	Mean Blood Urea Nitrogen				Rank-sum test	Z = -14.72	< 0.001
	• N(missing)	3457(0)	783(0)	4240(0)			
	• Mean ± SD	26.60 ± 21.66	38.55 ± 27.46	28.80 ± 23.31			
	• Median	19.5	30.5	21			
	• Q1, Q3	13.5, 32	19, 49.5	14, 35.5			
	• Min, Max	0, 216.5	0, 194.5	0, 216.5			
14	Urine Output				Rank-sum test	Z = -18.1	< 0.001
	• N(missing)	3457(0)	783(0)	4240(0)			
	• Mean ± SD	2037.71 ± 1576.87	1276.79 ± 1358.36	1897.19 ± 1566.8			
	• Median	1740	970	1600.5			
	• Q1, Q3	1100, 2658	406, 1652	940, 2530			
	• Min, Max	0, 50,515	0, 12,210	0, 50,515			
15	Elisxhauser Comorbidity Index				Rank-sum test	Z = -13.76	< 0.001
	• N(missing)	3457(0)	783(0)	4240(0)			
	• Mean ± SD	3.08 ± 6.89	6.86 ± 7.00	3.78 ± 7.06			
	• Median	2	7	3			
	• Q1, Q3	-1, 8	2, 12	0, 9			
	• Min, Max	-23, 28	-19, 30	-23, 30			
16	Logistic Organ Dysfunction Score				Rank-sum test	Z = -21.37	< 0.001
	• N(missing)	3457(0)	783(0)	4240(0)			
	• Mean ± SD	4.7 ± 2.56	7.54 ± 3.5	5.23 ± 2.97			
	• Median	4	7	5			
	• Q1, Q3	3, 6	5, 10	3, 7			
	• Min, Max	0, 16	0, 20	0, 20			
17	Outcome (%)	3457(81.53%)	783(18.47%)	4240			

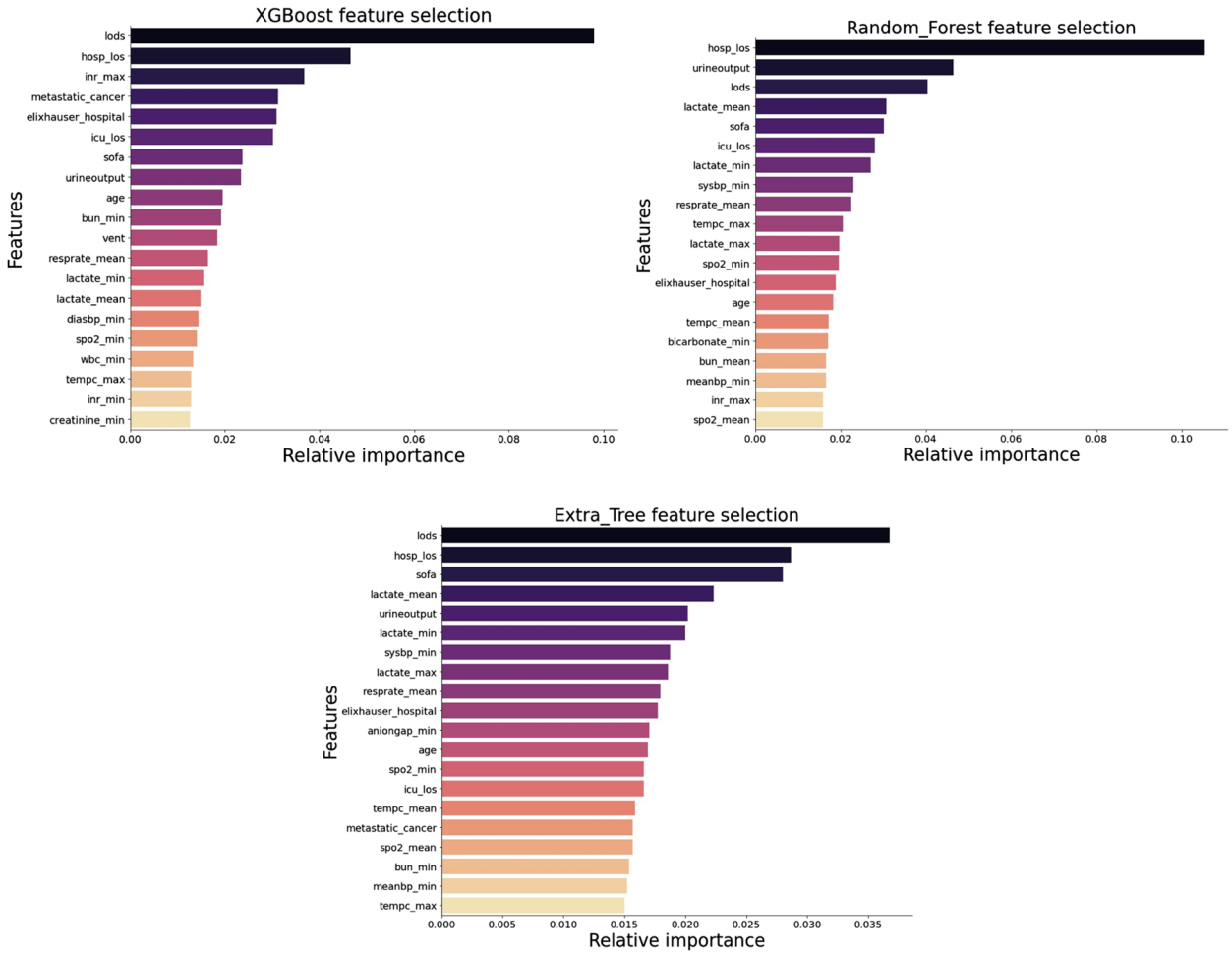

Fig. 2 Bar plots presenting Feature Ranking using XGBoost, RF, and ET techniques

After all common feature variables were selected from the three-ranking system, the characteristics were fed into an LR formulation to track how performance changed as more features were added. The model training phase came next.

Performance comparison of classical machine learning models and stacking-based meta-classifier

To evaluate the influence of the selected biomarkers and their role in model training, an examination was carried out on the results pertaining to the highest-ranking feature (top-1), the second highest-ranking feature (top-2), and up to the fifteenth highest-ranking feature (Top-15). The aforementioned findings were subsequently subjected to analysis employing the LR method. Table 2 provides empirical evidence to support the proposition presented before.

According to the information provided in the table, it is possible to deduce that there was a discernible improvement in the overall performance of the model for every increment in the increased feature value. When 15 characteristics were utilized, the LR model produced the most accurate results, surpassing the performance of earlier feature combinations. To train the model, there was consequently no apparent requirement for additional steps to isolate those properties from one another.

The evaluation process involved subjecting the folder data to training using eight classical ML models. The model selection for the meta-classifier was determined by considering a balanced performance across the overall performance criteria previously described in the preceding section. From the table, we observe the following model performances: The MLP Classifier achieves an accuracy of 80.5%, precision of 84.19%, recall of 80.04%, specificity of 72.86%, and an F1-score of 81.44%. While its overall performance is decent, its specificity lags behind, indicating that the model detects true positives well but may produce a higher number of false positives. The XGB model attains an accuracy of 82.85%, precision

Table 2 Individual Feature performance in Logistic Regression

Feature Incrementation	Accuracy	Precision	Recall	Specificity	F1-Score
Top-1 Features	68.61	75.20	68.61	51.20	71.14
Top-2 Features	71.46	78.53	71.46	60.84	73.93
Top-3 Features	71.23	78.40	71.22	60.59	73.73
Top-4 Features	71.67	73.44	71.68	64.04	74.26
Top-5 Features	70.99	73.24	70.99	63.99	73.70
Top-6 Features	71.27	79.31	71.28	63.95	73.93
Top-7 Features	76.01	83.54	76.01	75.01	78.24
Top-8 Features	75.68	83.52	75.69	75.23	77.98
Top-9 Features	75.92	83.67	75.92	75.58	78.18
Top-10 Features	76.04	83.46	76.03	74.72	78.24
Top-11 Features	76.08	83.45	76.08	74.63	78.28
Top-12 Features	75.14	83.40	75.14	75.30	77.52
Top-13 Features	75.73	83.58	75.73	75.44	78.02
Top-14 Features	75.71	83.22	75.70	74.15	77.95
Top-15 Features	76.63	83.81	76.62	75.45	78.75

of 83.95%, recall of 82.86%, specificity of 66.28%, and an F1-score of 83.33%, outperforming the MLP Classifier in accuracy, recall, and F1-score, though its specificity is lower, suggesting challenges in accurately identifying true negatives. Logistic Regression shows an accuracy of 76.63%, precision of 83.81%, recall of 76.62%, specificity of 75.45%, and an F1-score of 78.75%, with balanced performance across metrics but lower accuracy and recall compared to other models. The Extra Tree model excels with an accuracy of 94.72%, precision of 95.34%, recall of 94.72%, specificity of 94.75%, and an F1-score of 94.88%, demonstrating outstanding performance across all parameters. The AdaBoost model records an accuracy of 79.98%, precision of 82.97%, recall of 79.98%, specificity of 67.70%, and an F1-score of 81.09%, showing consistent performance though with lower specificity. The Gradient Boost model achieves an accuracy of 86.91%, precision of 86.32%, recall of 86.91%, specificity of 64.43%, and an F1-score of 86.54%, performing the worst in specificity among the models. The CatBoost model shows excellent overall performance with an accuracy of 87.95%, precision of 87.53%, recall of 87.95%, specificity of 68.12%, and an F1-score of 87.69%, though it faces similar specificity issues as most models. Finally, the Random Forest model achieves an accuracy of 82.83%, precision of 83.92%, recall of 82.83%, specificity of 66.18%, and an F1-score of 83.29%, demonstrating favorable performance with notably high precision and recall rates.

The classical result demonstrated that the MLP classifier, LR classifier, and ET Classifier exhibited the highest performance. The ET classifier exhibited superior performance in the preliminary training evaluation compared to the other models. The aforementioned three models were subsequently utilized as the foundational components for the stacking-based meta-classifier. In our study, we employed a stacking-based meta-classifier using the MLP, LR, and ET classifiers as foundational components. These models were chosen for their strong individual performance during preliminary evaluations. Their strong performance in all the metrics in individual tests made them suitable candidates for the stacking model. The stacking approach involved training these base models independently and then feeding their prediction probabilities into a meta-classifier—in our study which is all the classical models. Out of all the models, stacked LR came out as the best performing model. This meta-classifier combined the outputs of the base models to generate the final predictions, leveraging the strengths of each to enhance overall accuracy and robustness. Zhang et al.'s [70] work illustrated how stacking can significantly enhance prediction performance compared to individual base models. He presented a three-tier stacking model that estimated the probability of hospital readmission and achieved an AUC of 0.720, significantly surpassing the performance of the individual models used. This outcome served as an inspiration for our study, reinforcing the potential of stacking to elevate the predictive power of machine learning models in complex tasks. While our stacking approach utilized a simpler structure with a single meta-layer, the principle remains consistent: by combining the predictions of strong base models, the stacking technique can achieve better performance than any individual model alone. This understanding guided our decision to use stacking, with Zhang et al.'s work serving as a successful example of how this method can be applied to improve model outcomes.

The prediction probabilities for each sample from the three models were subsequently retrieved and utilized for training the Stacking algorithm. The incorporation of prediction probability weights resulted in considerable improvements across all remaining models, as evidenced by their ability to easily surpass the 90% benchmark in the Meta-Classifier outcome. Furthermore, the classical results on multiple models exhibited a significant deficiency in specificity score, which was notably enhanced in our study. In this investigation, Stacked LR had the maximum performance in several evaluation metrics. Specifically, the model achieved accuracy, precision, recall, specificity, and F1-score values of 95.52%, 95.79%, 95.52%, 93.65%, and 95.60%, respectively. Table 3 presents the comprehensive findings pertaining to the classical training and stacking training phase. The models focused on stacking architecture and achieved the highest performance have been distinguished by bolding them.

Figure 3 displays a side-by-side comparison of the ROC curves representing the performance of the classical ML model and the stacking-based meta-Classifier. During the traditional training phase, the majority of models were limited to achieving an AUC score of 0.85. The AUC score is a widely used metric in binary classification tasks. It serves as a key indicator for evaluating the performance of classification models. Among the several models assessed in the context of classical training, the ET model demonstrated exceptional performance, attaining an AUC score of 0.99. Upon the implementation of the stacking technique, it was seen that the AUROC scores for all the models exhibited a notable improvement. Notably, the LR and ET models achieved the highest scores among the evaluated models.

Table 3 Results displaying classical ML outcomes and stacking classification outcomes

	Classical ML Model Result					Stacking-Based Meta Classifier Result				
Classifiers	**Accuracy**	**Precision**	**Recall**	**Specificity**	**F1-Score**	**Accuracy**	**Precision**	**Recall**	**Specificity**	**F1-Score**
MLP	**80.5**	**84.19**	**80.04**	**72.86**	**81.44**	94.88	95.41	94.88	94.39	95.02
XGBoost	82.85	83.95	82.86	66.28	83.33	94.29	94.96	94.29	93.97	94.48
LR	**76.63**	**83.81**	**76.62**	**75.45**	**78.75**	**95.52**	**95.79**	**95.52**	**93.65**	**95.60**
ET	**94.72**	**95.34**	**94.72**	**94.75**	**94.88**	94.62	95.27	94.62	94.73	94.79
AdaBoost	79.98	82.97	79.98	67.70	81.09	93.51	94.45	93.51	93.79	93.76
GB	86.91	86.32	86.91	64.43	86.54	93.25	94.30	93.26	93.81	93.53
CatBoost	87.95	87.53	87.95	68.12	**87.69**	93.35	94.40	93.35	94.15	93.63
RF	82.83	83.92	82.83	66.18	83.29	93.87	94.69	93.87	93.97	94.09
DT	78.07	80.99	78.06	62.24	79.22	94.81	94.89	94.81	89.34	94.84

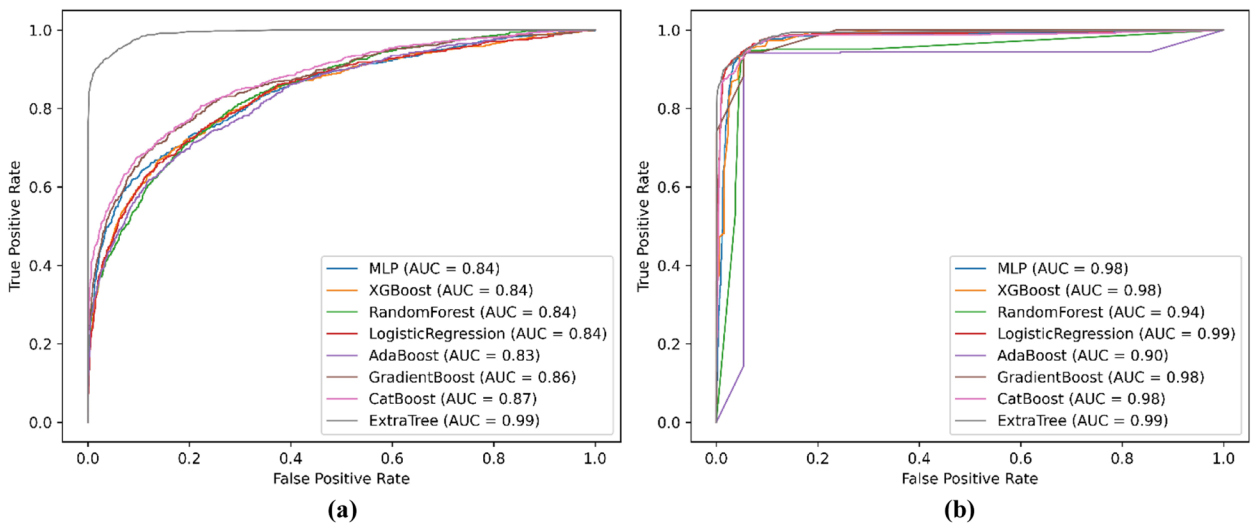

Fig. 3 Comparison of receiver operating characteristics (ROC) curves for different classical machine learning models (**a**) and stacking machine learning models (**b**)

In Fig. 4, the confusion matrix indicates a False Positive Rate (FPR) of 2.89%, meaning that out of all predictions made by the model, 2.89% incorrectly identified patients as being at high risk of death when they actually survived. In the medical field, where the stakes are high, a low FPR is essential to avoid unnecessary interventions, stress, and potential harm to patients who are not truly at risk. The model's low FPR, alongside a True Negative Rate of 97.11% and a True Positive Rate of 91.44%, suggests that it is highly effective at distinguishing between patients who are likely to survive and those at risk of death. The matrix also reflects an 8.56% False Negative Rate, where patients who actually died were incorrectly predicted to survive. In healthcare, minimizing the FNR is just as important as minimizing the FPR, as false negatives can lead to missed opportunities for critical care and treatment. This reliability is particularly important in healthcare settings, where false positives can lead to unwarranted treatments and patient anxiety. The effectiveness of the model in achieving this low FPR can be attributed to the use of the SMOTE-Tomek Link method for balancing the sample sizes in the training data. This technique ensured that the model was well-trained to recognize both categories accurately, thus minimizing the chances of false alarms and enhancing the overall trustworthiness of the predictions.

Model explainability

The classification outcome has been explained using the SHAP summary plot and SHAP values in the diagram. Figure 4 presents the contribution provided by each biomarker in the outcome in the form of a SHAP summary plot.

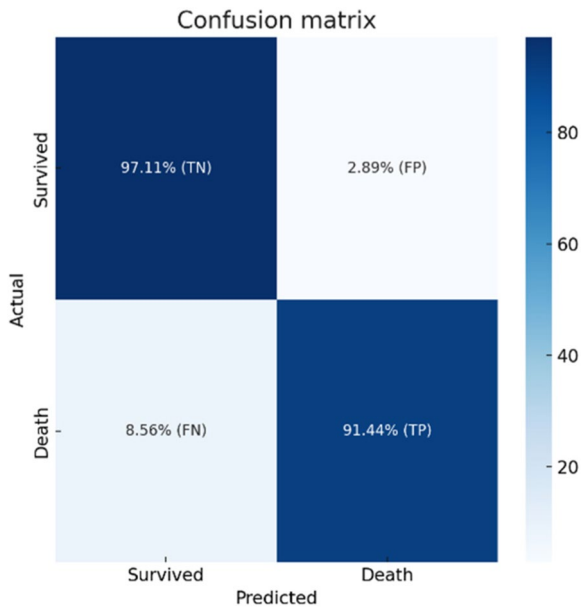

Fig. 4 Confusion matrix showing the classification outcomes for predicting patient survival and death

The SHAP values have been analyzed to interpret the contribution of features to the model's predictions. The SHAP value has a significant positive relationship with the contribution that was made to the positive class, whereas the SHAP value has a negative correlation with the contribution that was made to the negative class. For the Logistic Regression and Extra Trees models, the LODS score significantly contributed to predicting whether a patient died within a 30-day timeframe. Figure 5 shows that as the LODS value increases, the risk of patient mortality also increases. Conversely, for the MLPClassifier and the meta-classifier, the length of hospital and ICU stay, along with the average and minimum blood urea nitrogen (BUN) levels, had higher feature importance than the LODS score. The SHAP summary plot indicates that a longer hospital stay contributes to the prediction of the negative class, suggesting that patients who stayed longer were more likely to survive beyond 30 days. However, an extended ICU stay contributes to the positive class, indicating a higher likelihood of mortality. Higher mean BUN levels, indicative of better renal function, were associated with a reduced risk of mortality, suggesting that effective waste removal from the blood lowers the risk of adverse outcomes. Conversely, lower minimum BUN levels may imply weaker renal function and a precarious patient condition. A notable observation is the positive association between the length of hospital stay and patient survival, as identified by the SHAP values for both the MLPClassifier and the meta-classifier. This finding aligns with our study's results, suggesting that patients hospitalized for more than 30 days are likely to survive. Conversely, the length of ICU stay emerged as the most significant predictor of mortality, with extended ICU stays indicating a higher risk. The analysis also reveals a trade-off in feature importance, as the stacking model aligns closely with the MLPClassifier's SHAP values. This suggests that stacking multiple models can elevate the importance of features that may rank lower in individual base classifiers but perform well in the final model. This deeper understanding of feature importance and the decision-making process enhances the credibility of our stacking model.

The stacking approach used in our study combines multiple base classifiers to enhance overall model performance. While this method demonstrates superior predictive accuracy, it also introduces additional complexity. Each base classifier contributes to the final prediction, necessitating a thorough understanding of how individual models interact within the ensemble. The SHAP analysis revealed that features such as LODS, length of hospital stay, length of ICU stay, and BUN levels play varying roles across different classifiers.

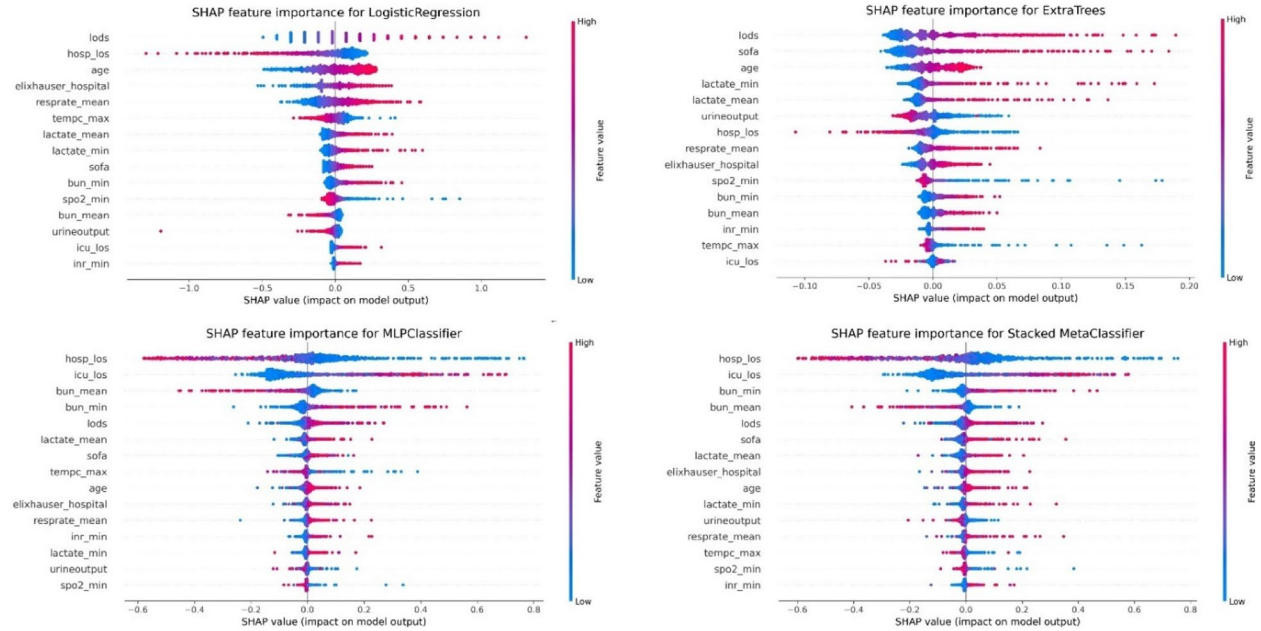

Fig. 5 SHAP analysis plot detailing the impact of each biomarker in classification outcome for the base models and stacked model

For instance, LODS significantly impacts the Logistic Regression and Extra Trees models, whereas the length of stay and BUN levels are more influential in the MLPClassifier and the meta-classifier. This complexity introduces a trade-off: while the stacking model benefits from the combined strengths of individual classifiers, it also inherits their limitations. Features that are less important in some base models may gain prominence in the final model, as observed with the MLPClassifier's SHAP values influencing the stacking model's decisions. Additionally, the ensemble method increases computational demands and may complicate the interpretability of the model. However, by leveraging the strengths of various classifiers, the stacking approach achieves a balanced and robust predictive performance. In summary, the stacking approach, despite its complexity, effectively integrates diverse feature contributions, enhancing prediction accuracy and reliability. This detailed analysis of model complexity and trade-offs provides a comprehensive understanding of the stacking model's operation and underscores its potential benefits and limitations.

Evaluation of nomogram in estimating mortality outcome

A nomogram was developed using a multivariate LR model to estimate mortality prediction during 30 days for patients diagnosed with Sepsis. Initially, 15 features used for model training were also utilized in the logistic regression analysis. Hence it marks the relationship among the trained features with binary output label. Eight biomarkers were utilized in the model training phase based on the values of $P>|z|$, all of which exhibited statistical significance and had favorable feature values as all of them showing values less than 0.05. The relationship between mortality prediction within 30 days and other parameters was evaluated using multivariate LR. Table 4 provides an in-depth analysis of the regression coefficient, standard error, z-value, statistical significance, and a 95% confidence interval for each variable. The z-value serves as a key indication for determining the significance of variables in a LR model. The table presents the lowest z-value observed for the minimum and mean blood urea nitrogen count. In contrast, the Length of stay in hospital and ICU yielded a significantly larger z-value. Hence, it can be inferred that the predictive value of minimum blood urea nitrogen count is negligible, with the length of stay in the hospital and intensive care unit being the primary determinants of the regression outcome.

During the process of internal validation, it is possible to observe, in accordance with the calibration plot that is displayed in Fig. 6, that the alignment of the calibration belt corresponds closely to the diagonal line. This is indicative of a fairly substantial degree of accuracy in the outcomes that have been anticipated. Figure 7 illustrates the net advantage offered by each biomarker in the decision-making process for the outcome.

The nomogram presented in Fig. 8 incorporates the biomarkers that were previously addressed in the table. Among the eight biomarker scales depicted in the

Table 4 The LR analysis to construct the nomogram for mortality prediction in 30 days

Outcome	Coef	Std. Err	z	*P*>|z|	[95% conf. Interval]	
ICU Length of Stay	**0.23**	**0.019**	**12.92**	**0.000**	**0.19226**	**0.26272**
Hospital Length of Stay	-0.26	0.017	-15.68	0.00	-0.29713	-0.23111
Urine Output	-0.0003	0.00004	-6.43	0.00	-0.00035	-0.00019
Blood Lactate Minimum	0.31	0.04	6.97	0.00	0.22283	0.39722
Mean Blood Urea Nitrogen	-0.03	0.012	-2.60	0.01	-0.05572	-0.0078
Minimum Blood Urea Nitrogen	0.04	0.013	3.19	0.001	0.01607	0.067285
Minimum Oxygen Saturation	-0.0295	0.006	-4.77	0.00	-0.417	-0.017413
Logistic Organ Dysfunction	0.22	0.019	11.44	0.00	0.18036	0.2549322
_cons	0.79	0.61	1.29	0.20	-0.40824	1.982226

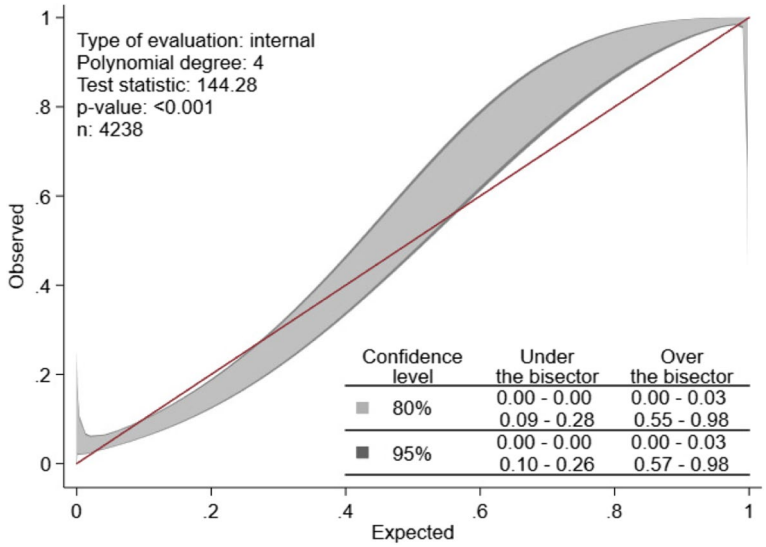

Fig. 6 Calibration curve illustrating classification for survival and death class

picture, it is evident that the duration of stay in the hospital exhibits the highest degree of prominence about the likelihood of an unfavorable outcome, specifically the mortality of a patient. This biomarker indicates a shorter duration of hospitalization is associated with a more accurate mortality prediction during 30 days. The probability score indicated on the lowest scale of the nomogram provides an estimation of the likelihood of mortality associated with a particular outcome. The predicted outcome appears to be supported by the relatively small likelihood of exceeding 0.5. Moreover, the aforementioned information aligns with the data we examined, which indicates that there were 783 deaths and 3,457 survival instances. Therefore, it is possible to deduce, based on the projected values, that the model has the potential to reliably forecast the mortality outcome for patients diagnosed with Sepsis for 30 days.

Discussion

Sepsis, a condition characterized by substantial deaths and enormous economic implications, has evolved beyond its original definition as a mere hazardous infection. According to a specific study, the death rate among ICU and hospital-treated patients was found to be 41.9% and 26.7% respectively [71]. A further investigation unveiled that the percentage of deaths within the ICU was recorded at 20.8%, while the corresponding percentage for the hospital as a whole was found to be 24.9%. Among the recorded fatalities, a significant proportion of 58.3% can be attributed to septic shock that occurred after the patients departed from the ICU [72]. Hou et al. [41] have proposed a model that utilizes XGBoost as a potential substitute for conventional LR analysis and the customary SAPS II scoring system. The objective of this model is to provide an initial estimation of mortality

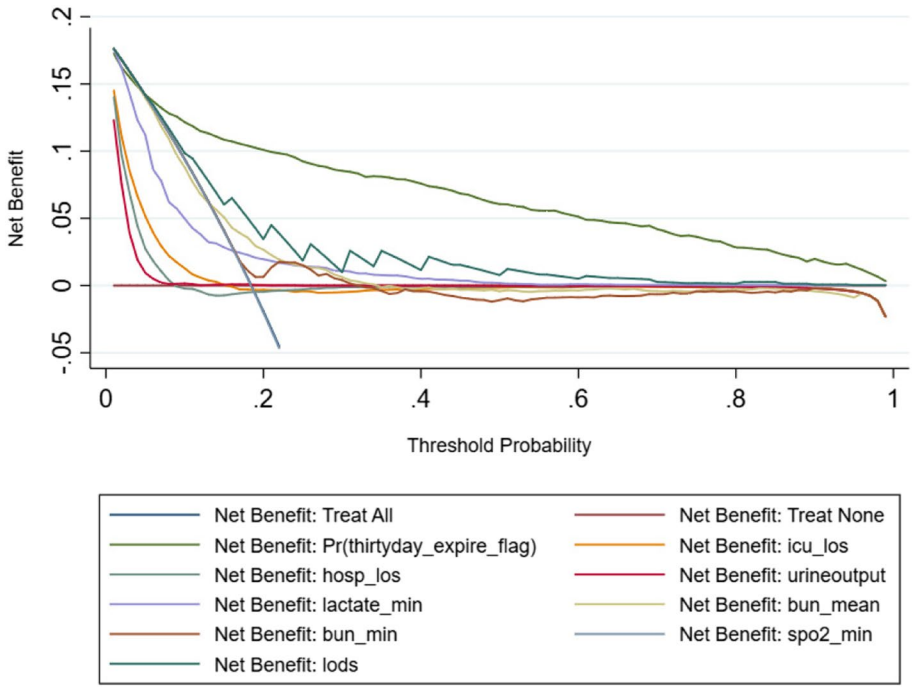

Fig. 7 Decision curves analysis showing comparison among different biomarkers to predict the death probability of patients with Sepsis

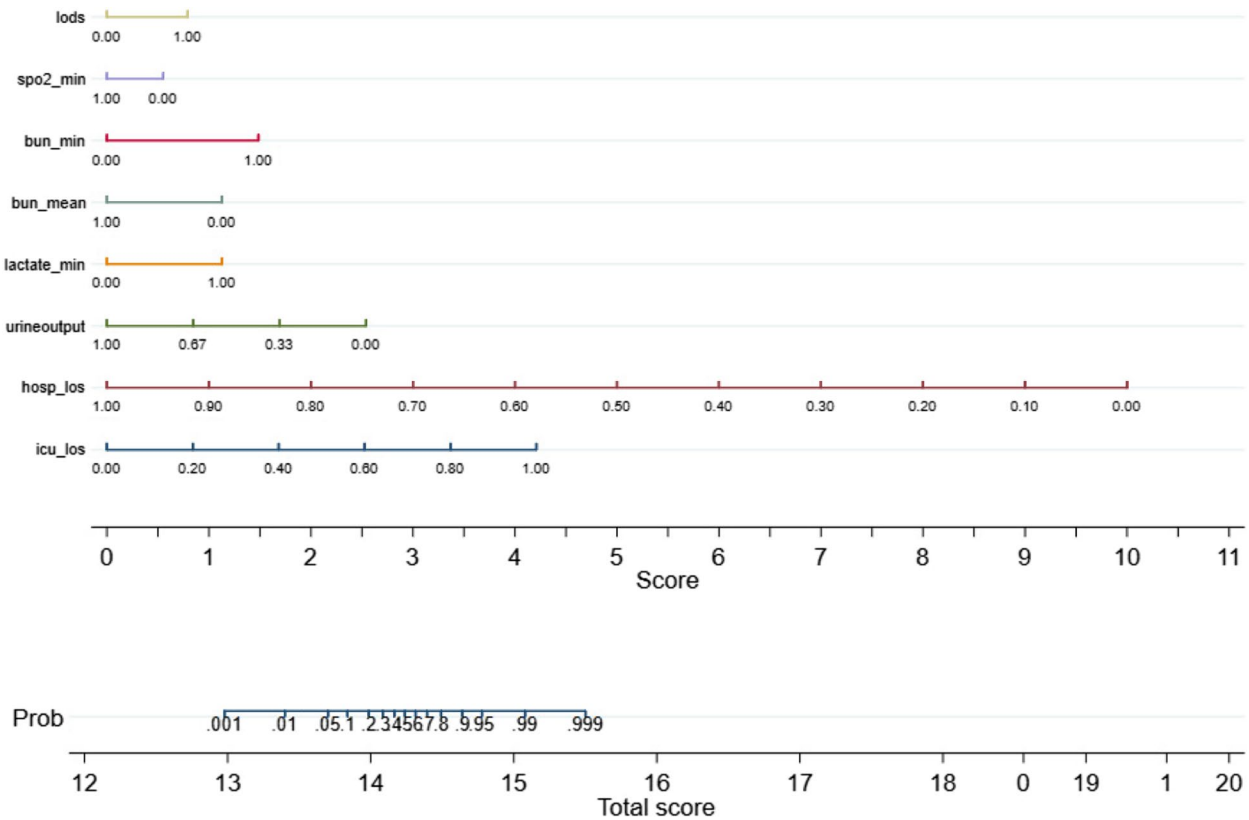

Fig. 8 A nomogram utilizing multivariate LR model is employed to estimate the probable outcome for deceased persons developed to estimate mortality using a set of eight biomarkers

rates for individual patients. The methodology employed in this research mostly involved assessing the effectiveness of the model using AUC scores and net benefit, as indicated by the examination of decision curves. The AUC score alone is insufficient to comprehensively assess class performance, particularly in class imbalance. Based on the findings of the AUCs, the values obtained were 0.819 (95% confidence interval [CI] 0.800–0.838), 0.797 (95% CI 0.781–0.813), and 0.857 (95% CI 0.839–0.876). Su et al. [14] employed ANN architecture to predict the Sepsis mortality rate in 30 days. In that study, the ANN model achieved an AUC of 0.873, while LR yielded 0.720. Additionally, the Acute Physiology and Chronic Health Evaluation (APACHE) II score demonstrated a predictive AUC of 0.629, although the SOFA score achieved a slightly lower accuracy of 0.619. The validation set yielded AUC values of 0.811, 0.752, 0.607, and 0.628 for the ANN, LR, APACHE II, and SOFA scores, respectively.

The study implemented fundamental preprocessing techniques as outlined in the methodology section, including data type formatting and data normalization. Three feature ranking algorithms were employed to identify fifteen features that have the potential to predict the probability of mortality, utilizing data obtained during the patient's admission to the hospital. These factors were: age, minimum and average blood lactate, qSOFA score, average respiration rate, length of stay in ICU and hospital, minimum and average blood urea nitrogen, maximum temperature (in Celsius), oxygen saturation (%), urine output, the Elisxhauser comorbidity index, and the Logistic Organ Dysfunction Score (LODS). So, core predictors were a mixture of Sepsis severity scores, clinical information, and pathological data. Two stages of prediction models were compared, with classical ML and Stacking ML. Deceased patients in this study predominantly consisted of older patients with high blood lactate counts and higher SOFA values. Also, they seemed to have a higher distribution of respiration rate compared to an alive patient, suggesting the deceased patient had an erratic breathing cycle among them. Their stay in the hospital and ICU also demonstrated a distinct pattern. Deceased patients had a comparatively low length of stay in the ICU as well as in the Hospital. Both of the duration of stay was strictly lied in 30 days timeframe. In addition to the length of stays, it was observed that Sepsis severity is likely to be strongly linked with high urine output with equally high urea nitrogen present in the blood. Moreover, severity could be observed with low oxygen saturation in the bloodstream of the dead patients. Patients who passed away had high LODS scores; hinting Sepsis severity is associated with the severity of organ dysfunction in critical care. In our study, we chose to use fivefold cross-validation for model testing and validation. While tenfold cross-validation is a common approach, we conducted preliminary experiments comparing both fivefold and tenfold cross-validation and found that the difference in performance metrics, such as accuracy and precision, was minimal. This indicated that increasing the number of folds did not significantly improve the model's performance estimates. Moreover, a key factor in our decision was the consideration of computational efficiency. Performing tenfold cross-validation requires training the model ten times, which doubles the training time compared to fivefold cross-validation. Given the minimal difference in results, the additional computational cost of tenfold cross-validation was not justified in our context. By opting for fivefold cross-validation, we were able to balance the need for reliable model validation with the practical considerations of computational resources. This approach allowed us to efficiently manage training time without compromising the robustness of our model's evaluation. Table 5 represents training time comparison between fivefold and tenfold:

A total of 9 machine-learning models were trained in both phases. In the first phase of training, only ET was able to produce robust results. It is due to the warm start parameter used in the ET. The "warm start" strategy is employed in an array of algorithms, such as ET (or Extremely Randomized Trees), to maximize efficiency and obtain convergence during the training of multiple models or the execution of hyperparameter tuning. It entails setting up a novel model or training procedure by applying the knowledge or variables learned from an earlier training session or iteration [73, 74]. So, exploiting stacking architecture after the classical training part was affected in the same manner and significantly improved each and every model Stacking is a powerful ensemble learning technique that combines the predictions of multiple base models to improve overall performance. In our study, we employed Stacked Logistic Regression (Stacked LR) as the meta-classifier, using output probabilities from

Table 5 Training time comparison for classical models in fivefold and tenfold data

Classifier	5-Fold (seconds)	10-Fold (seconds)
MLP Classifier	38.87	77.46
XGB Classifier	1.645	2.429
Random Forest Classifier	6.913	14.042
Logistic Regression Classifier	0.09	0.179
Ada Boost Classifier	5.651	12.676
Gradient Boosting Classifier	32.701	68.762
Cat Boost Classifier	40.361	69.848
Extra Trees Classifier	0.464	0.497
Decision Tree	0.44	0.94

base models, including the Extra Trees (ET) classifier. While stacking often enhances performance by leveraging the strengths of different models, it is important to note that it does not always surpass the best-performing individual model. In our case, the ET classifier alone demonstrated exceptional performance, achieving high accuracy, precision, recall, specificity, and F1-score metrics, all exceeding 94%. When the Stacked LR approach was applied, we observed only marginal improvements in some metrics, and in some cases, like specificity, the performance slightly decreased compared to the standalone ET model. This outcome can be explained by the fact that the ET model was already highly effective, capturing most of the predictive power available in the data. The additional complexity introduced by stacking may not always result in significant improvements when a single model like ET is already near optimal performance. However, the Stacked LR model remains valuable because it combines the strengths of multiple models, enhancing generalization and reducing bias, which is particularly important in complex, high-stakes environments like healthcare. Even slight improvements in performance across multiple metrics can be critical in such contexts, where accuracy and reliability are paramount. Therefore, the emphasis on the Stacked LR model is justified as it strategically enhances prediction accuracy and minimizes risks, making it an essential tool for making well-informed decisions in complex scenarios. We also observed, performance of stacked ET falls behind slightly compared to the standalone ET. It is important to recognize that the performance of stacking is not guaranteed to always surpass that of the best individual model. This is because stacking relies on the quality of the base models and the ability of the meta-classifier to effectively combine their outputs. When the base models already perform very well, as is the case with the ET classifier in our study, the incremental improvement from stacking can be minimal. Additionally, because stacking introduces a layer of complexity—relying on the probabilistic outputs rather than the original data—there is a stochastic element involved, meaning that the performance of the stacking model can vary slightly depending on the specific training and validation splits**.** The AUC score achieved by this study was 0.99 in LR and ET after the Meta-Classifier was constructed, which steadily defeated the benchmarks achieved by the previously mentioned studies in this section. It resulted in an exceptionally good recognition of deceased patients and surviving patients. Provided nomogram architecture also provided a probabilistic score using those 15 biomarkers. This methodology can potentially enhance the effective allocation of healthcare resources without surpassing their capability.

A comparative result among similar literature can be observed in the Table 6. Hou et al. [41] proposed methodology achieved the highest AUC of 0.857 using the XGBoost algorithm. For the work proposed by Su et al. [14], an AUC of 0.873 was achieved through ANN. Park et al. [38] employed various predicting algorithms, earning them the highest AUC of 0.893 from Deep Neural Network (DNN). Yang et al. [22] through their LR analysis, were able to attain an AUC of 0.763 for training and 0.753 for validation. Some of the latest works like Yong et al. [64] proposed DGFSD model which was architecture upon DNN and GCN model was able to achieve an accuracy of 82.78%. Furthermore, Zhang et al. [65] achieved 0.94 AUC and 0.937 F1 score using XGBoost model where Age, AST, invasive ventilation treatment, and serum urea nitrogen performed as best contributing features. Additionally, Palmowski et al. [75] achieved an AUC of 0.84 from SVM and AUC of 0.82 from ANN. Our proposed method for the study smoothly outperformed this literature with a hefty value of 0.99 AUC from the ET model and LR from stacking.

Table 6 Comparison with similar works from the literature

Authors	Approach	Evaluation
Hou et al. [41]	Logistic regression model, SAPS-II scores prediction model and XGBoost algorithm	Logistic regression (AUC: 0.819), SAPS-II (AUC: 0.797) and XGBoost (AUC: 0.857)
Su et al. [14]	ANN-based architecture	AUC of 0.873
Park et al. [38]	Logistic Regression, Random Forest, XGBoost, Deep Neural Network (DNN), Super Learner Model	Logistic Regression (AUC: 0.878), Random Forest (AUC: 0.878), XGBoost (AUC: 0.888), DNN (AUC: 0.893), Super Learner Model (AUC: 0.883)
Yang et al. [22]	Logistic Regression Analysis	Training & Validation AUROC of 0.763 and 0.753
Palmowski et al. [75]	SVM with polynomial kernel, ANN	SVM (AUC: 0.84), ANN (AUC: 0.82)
Zhang et al. [65]	XGBoost	AUC: 0.94, F1 score: 0.937
Yong et al. [64]	DGFSD model based on DNN and Graph Convolutional Network (GCN)	Accuracy of 82.78%
Proposed Method	8 Classical Machine Learning Model and Stacking-based Meta Classifier	**Extra Tree & Stacked Logistic Regression—AUC 0.99**

The study conducted a comprehensive analysis of the empirical findings and concluded that the suggested stacking model successfully utilizes a structure capable of enhancing the outcomes generated by numerous models to understand the organization and patterns within the data. Through contemplation, various models possess the capacity to acquire information from one another, resulting in a collective improvement in their predicted accuracy, as demonstrated in this study. This implies that the stacking process exhibits promise in enhancing the predictive capacities of models, particularly in situations with intricate data patterns. This study substantiated the aforementioned claim by doing a detailed analysis of all the models, employing the weights derived by the ET algorithm. The previous assertion is confirmed by research conducted by Chiu et al., whereby they employed six classifiers (Random Forest, Support Vector Classifier, k-Nearest Neighbors, Light Gradient Boosting Machine, Bagging, and Adaboost) to construct a stacking model. The present investigation attained an impressive level of accuracy of 95.25% and an AUC score of 0.8255. [76].

Conclusion

In summary, our study demonstrates that the proposed nomogram, which utilizes multiple biomarkers as predictors based on a feature ranking scheme, achieves a high degree of accuracy and reliability in forecasting long-term outcomes for individuals with Sepsis. The model, built on a stacking-based architecture, exhibits impressive precision, accurately predicting patient outcomes well in advance of major clinical events using scoring-based biomarkers, hospital and ICU stay metrics, and pathology biomarkers. The model achieved an AUC of 0.99 and demonstrated strong performance across several metrics, with accuracy, precision, recall, specificity, and F1-score values of 95.52%, 95.79%, 95.52%, 93.65%, and 95.60%, respectively. These results suggest that the model could significantly impact clinical decision-making, particularly in resource-limited settings, and contribute to reducing mortality rates among Sepsis patients. Moreover, the stacking-based framework followed in this study offers a meticulous approach to clinical administration, potentially empowering healthcare practitioners to adopt a streamlined and effective method for patient stratification. This could help alleviate the burden on healthcare resources and improve patient outcomes through a more refined and orchestrated response. While this study highlights the potential of using Sepsis-3 clinical data to predict mortality, it is important to note that the machine learning model's performance is highly data-dependent. The model's effectiveness may vary when applied to datasets from different regions, and its robustness could be enhanced by incorporating more diverse and comprehensive datasets from across the globe. Future research should focus on exploring different models and identifying the optimal feature sets, particularly by utilizing data collected from multiple clinical centers and countries to develop a more generalized model. Additionally, future studies should consider integrating deep learning algorithms and a severity index scoring scheme to further enhance the model's predictive power and applicability. However, this study has certain limitations. The model currently only includes mortality data up to 30 days, which limits its ability to predict long-term outcomes beyond this period. Furthermore, the dataset lacks information regarding whether it was sourced from a single clinical center, which poses challenges to generalizing the findings across different healthcare settings. Addressing these limitations in future research will be crucial to further validate and extend the model's applicability in diverse clinical environments.

Abbreviations

ALDCC	Age, Lymphocyte count, D-dimer, CRP, and Creatinine
ANN	Artificial Neural Network
AI	Artificial Intelligence
AUC	Area Under the Curve
CXR	Chest X-ray
CNNs	Convolutional neural networks
COVID-19	Corona Virus Disease 2019
DNN	Deep Neural Network
EMR	Electronic Medical Records
ED	Emergency department
XGBoost	Extreme Gradient Boosting
GBM	Gradient boosting machines
IL-6	Interleukin-6
ICU	Intensive Care Unit
INR	International normalized ratio
LODS	Logistic Organ Dysfunction Score
LR	Logistic Regression
ML	Machine Learning
NEWS	National Early Warning Score
NLR	Neutrophil-to-Lymphocyte Ratio
NT-proBNP	N-terminal prohormone of brain natriuretic peptide
qSOFA	Quick Sequential Organ Failure Assessment
SMOTE	Synthetic Minority Over-sampling Technique
SIRS	Systemic Inflammatory Response Syndrome
SHAP	SHapley Additive Explanations
SGB	Stochastic Gradient Boosting
SAE	Sepsis-associated encephalopathy

Acknowledgements

Not applicable.

Disclosure of potential conflicts of interest

The authors declare that they have no conflict of interest.

Authors' contributions

Md. Sohanur Rahman: Contribution: Conceptualization, Data Curation, Methodology, Software, Writing- Original draft preparation, Writing- Reviewing and Editing. Khandaker Reajul Islam: Contribution: Data Curation, Methodology, Writing- Original draft preparation, Writing- Reviewing and Editing. Johayra Prithula: Contribution: Data Curation, Methodology, Writing- Original draft preparation. Jaya Kumar: Contribution: Validation, Supervision, Funding Acquisition, Writing- Reviewing and Editing. Mufti Mahmud: Contribution: Validation, Supervision, Funding Acquisition, Writing- Reviewing and Editing. Mohammed Fasihul Alam: Contribution: Methodology, Software, Validation, Writing- Original draft preparation Mamun Bin Ibne Reaz: Contribution:

Conceptualization, Methodology, Software, Writing- Original draft preparation, Writing- Reviewing and Editing. Abdulrahman Alqahtani: Contribution: Conceptualization, Methodology, Software, Writing- Original draft preparation, Writing- Reviewing and Editing. Muhammad E. H. Chowdhury: Contribution: Conceptualization, Methodology, Supervision, Funding Acquisition, Writing- Original draft preparation, Writing- Reviewing and Editing.

Funding
This work was made possible by High Impact grant# QUHI-CENG-23/24–216 from Qatar University and is also supported via funding from Prince Sattam Bin Abdulaziz University project number (PSAU/2024/R/1445). The statements made herein are solely the responsibility of the authors. The open-access publication cost is covered by the Faculty of Medicine UKM.

Availability of data and materials
The dataset used in this study is publicly available in supplementary information section of a study conducted by Hou et al. [41]. Please find the link to access the dataset: https://ieee-dataport.org/documents/refined-mimic-iii-30-day-mortality-prediction-sepsis-3-patients

Declarations

Ethics approval and consent to participate
This study utilizes clinical data shared by Hou et al. as a supplementary resource. Hence, the authors were not involved in the Data collection process.

Consent for the publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 5 June 2024 Accepted: 27 August 2024
Published: 9 September 2024

References

1. C. Fleischmann *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. [Online]. Available: https://www.atsjournals.org/doi/full/https://doi.org/10.1164/rccm.201504-0781OC.
2. Liu V, Escobar GJ, Greene JD, Soule J, Whippy A, Angus DC, Iwashyna TJ. "Hospital deaths in patients with sepsis from 2 independent cohorts," (in English). JAMA. 2014;312(1):90–2. https://doi.org/10.1001/jama.2014.5804.
3. Singer M, et al. "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)," (in English). JAMA. 2016;315(8):801. https://doi.org/10.1001/jama.2016.0287.
4. Sinha RK, et al. "Thrombosis and Hemostasis: PAR1 biased signaling is required for activated protein C in vivo benefits in sepsis and stroke," (in English). Blood. 2018;131(11):1163. https://doi.org/10.1182/blood-2017-10-810895.
5. Kell DB, Pretorius E. "Editorial Compilation V: To What Extent Are the Terminal Stages of Sepsis, Septic Shock, Systemic Inflammatory Response Syndrome, and Multiple Organ Dysfunction Syndrome Actually Driven by a Prion/Amyloid Form of Fibrin?," (in English). Semin Thromb Hemost. 2018;44(3):224. https://doi.org/10.1055/s-0037-1604108.
6. B. S. Bloch-Budzier, "Hospital sepsis deaths 'jump by a third'," (in English), *BBC News,* 2018. Available: https://www.bbc.com/news/health-45045438.
7. M. Taj, M. Brenner, Z. Sulaiman, and V. Pandian, "Sepsis protocols to reduce mortality in resource-restricted settings: A systematic review," *Intensive and Critical Care Nursing,* vol. 72, p. 103255, 2022/10/01/ 2022, https://doi.org/10.1016/j.iccn.2022.103255.
8. Umemura Y, et al. Hour-1 bundle adherence was associated with reduction of in-hospital mortality among patients with sepsis in Japan. PLoS ONE. 2022;17(2): e0263936. https://doi.org/10.1371/journal.pone.0263936.
9. T. Rahman *et al.*, "Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images," *Computers in Biology and Medicine,* vol. 132, p. 104319, 2021/05/01/ 2021, https://doi.org/10.1016/j.compbiomed.2021.104319.
10. T. Rahman *et al.*, "BIO-CXRNET: a robust multimodal stacking machine learning technique for mortality risk prediction of COVID-19 patients using chest X-ray images and clinical data," (in eng), *Neural computing & applications,* pp. 1–23, May 4 2023, https://doi.org/10.1007/s00521-023-08606-w.
11. K. R. Islam *et al.*, "Prognostic Model of ICU Admission Risk in Patients with COVID-19 Infection Using Machine Learning," (in eng), *Diagnostics (Basel, Switzerland),* vol. 12, no. 9, Sep 3 2022, https://doi.org/10.3390/diagnostics12092144.
12. M. E. H. Chowdhury *et al.*, "An Early Warning Tool for Predicting Mortality Risk of COVID-19 Patients Using Machine Learning," (in eng), *Cognitive computation,* pp. 1–16, Apr 21 2021, https://doi.org/10.1007/s12559-020-09812-7.
13. T. Rahman *et al.*, "Mortality Prediction Utilizing Blood Biomarkers to Predict the Severity of COVID-19 Using Machine Learning Technique," (in eng), *Diagnostics (Basel, Switzerland),* vol. 11, no. 9, Aug 31 2021, https://doi.org/10.3390/diagnostics11091582.
14. Su Y, Guo C, Zhou S, Li C, Ding N. Early predicting 30-day mortality in sepsis in MIMIC-III by an artificial neural networks model. Eur J Med Res. 2022;27(1):1–10. https://doi.org/10.1186/s40001-022-00925-3.
15. Liu J, et al. Mortality prediction using a novel combination of biomarkers in the first day of sepsis in intensive care units. Sci Rep. 2021;11(1275):1–9. https://doi.org/10.1038/s41598-020-79843-5.
16. Y. S. Kwon and M. S. Baek, "Development and Validation of a Quick Sepsis-Related Organ Failure Assessment-Based Machine-Learning Model for Mortality Prediction in Patients with Suspected Infection in the Emergency Department," (in eng), *J Clin Med,* vol. 9, no. 3, Mar 23 2020, https://doi.org/10.3390/jcm9030875.
17. Kovach CP, Fletcher GS, Rudd KE, Grant RM, Carlbom DJ. "Comparative prognostic accuracy of sepsis scores for hospital mortality in adults with suspected infection in non-ICU and ICU at an academic public hospital," (in eng). PLoS ONE. 2019;14(9): e0222563. https://doi.org/10.1371/journal.pone.0222563.
18. G. Kong, K. Lin, and Y. Hu, "Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU," (in eng), *BMC Med Inform Decis Mak,* vol. 20, no. 1, p. 251, Oct 2 2020, https://doi.org/10.1186/s12911-020-01271-2.
19. Y. Ren *et al.*, "Risk factor analysis and nomogram for predicting in-hospital mortality in ICU patients with sepsis and lung infection," *BMC Pulmonary Medicine,* vol. 22, no. 1, p. 17, 2022/01/07 2022, https://doi.org/10.1186/s12890-021-01809-8.
20. van Doorn W, Stassen PM, Borggreve HF, Schalkwijk MJ, Stoffers J, Bekers O, Meex SJR. "A comparison of machine learning models versus clinical evaluation for mortality prediction in patients with sepsis," (in eng). PLoS ONE. 2021;16(1): e0245157. https://doi.org/10.1371/journal.pone.0245157.
21. R. Q. Yao *et al.*, "A Machine Learning-Based Prediction of Hospital Mortality in Patients With Postoperative Sepsis," (in eng), *Front Med (Lausanne),* vol. 7, p. 445, 2020, https://doi.org/10.3389/fmed.2020.00445.
22. Yang Y, et al. "Development of a nomogram to predict 30-day mortality of patients with sepsis-associated encephalopathy: a retrospective cohort study," (in eng). J Intensive Care. 2020;8:45. https://doi.org/10.1186/s40560-020-00459-y.
23. Liu S, Wang X, She F, Zhang W, Liu H, Zhao X. "Effects of Neutrophil-to-Lymphocyte Ratio Combined With Interleukin-6 in Predicting 28-Day Mortality in Patients With Sepsis," (in eng). Front Immunol. 2021;12: 639735. https://doi.org/10.3389/fimmu.2021.639735.
24. García-Gallo JE, Fonseca-Ruiz NJ, Celi LA, Duitama-Muñoz JF. "A machine learning-based model for 1-year mortality prediction in patients admitted to an Intensive Care Unit with a diagnosis of sepsis," (in engspa). Med Intensiva. Apr2020;44(3):160–70. https://doi.org/10.1016/j.medin.2018.07.016.
25. Klug M, et al. "A Gradient Boosting Machine Learning Model for Predicting Early Mortality in the Emergency Department Triage: Devising a Nine-Point Triage Score," (in eng). J Gen Intern Med. Jan2020;35(1):220–7. https://doi.org/10.1007/s11606-019-05512-7.
26. Faisal M, Scally A, Howes R, Beatson K, Richardson D, Mohammed MA. "A comparison of logistic regression models with alternative machine learning methods to predict the risk of in-hospital mortality in emergency

medical admissions via external validation," (in eng). Health Informatics J. Mar2020;26(1):34–44. https://doi.org/10.1177/1460458218813600.
27. C. Ye *et al.*, "A Real-Time Early Warning System for Monitoring Inpatient Mortality Risk: Prospective Study Using Electronic Medical Record Data," (in eng), *Journal of medical Internet research,* vol. 21, no. 7, p. e13719, Jul 5 2019, https://doi.org/10.2196/13719.
28. N. Brajer *et al.*, "Prospective and External Evaluation of a Machine Learning Model to Predict In-Hospital Mortality of Adults at Time of Admission," (in eng), *JAMA network open,* vol. 3, no. 2, p. e1920733, Feb 5 2020, https://doi.org/10.1001/jamanetworkopen.2019.20733.
29. A. Karlsson, W. Stassen, A. Loutfi, U. Wallgren, E. Larsson, and L. Kurland, "Predicting mortality among septic patients presenting to the emergency department–a cross sectional analysis using machine learning," *BMC Emergency Medicine,* vol. 21, no. 1, p. 84, 2021/07/12 2021, https://doi.org/10.1186/s12873-021-00475-7.
30. Rodríguez A, Mendoza D, Ascuntar J, Jaimes F. "Supervised classification techniques for prediction of mortality in adult patients with sepsis," (in eng). Am J Emerg Med. Jul2021;45:392–7. https://doi.org/10.1016/j.ajem.2020.09.013.
31. Soffer S, Klang E, Barash Y, Grossman E, Zimlichman E. "Predicting In-Hospital Mortality at Admission to the Medical Ward: A Big-Data Machine Learning Model," (in eng). Am J Med. Feb2021;134(2):227-234.e4. https://doi.org/10.1016/j.amjmed.2020.07.014.
32. J. W. Perng, I. H. Kao, C. T. Kung, S. C. Hung, Y. H. Lai, and C. M. Su, "Mortality Prediction of Septic Patients in the Emergency Department Based on Machine Learning," (in eng), *J Clin Med,* vol. 8, no. 11, Nov 7 2019, https://doi.org/10.3390/jcm8111906.
33. Liu N, et al. "Heart rate n-variability (HRnV) measures for prediction of mortality in sepsis patients presenting at the emergency department," (in eng). PLoS ONE. 2021;16(8): e0249868. https://doi.org/10.1371/journal.pone.0249868.
34. C. Y. Cheng *et al.*, "Machine learning models for predicting in-hospital mortality in patient with sepsis: Analysis of vital sign dynamics," (in eng), *Front Med (Lausanne),* vol. 9, p. 964667, 2022, https://doi.org/10.3389/fmed.2022.964667.
35. Ford DW, Goodwin AJ, Simpson AN, Johnson E, Nadig N, Simpson KN. "A Severe Sepsis Mortality Prediction Model and Score for Use With Administrative Data," (in eng). Crit Care Med. Feb2016;44(2):319–27. https://doi.org/10.1097/ccm.0000000000001392.
36. Y. Wu, S. Huang, and X. Chang, "Understanding the complexity of sepsis mortality prediction via rule discovery and analysis: a pilot study," (in eng), *BMC Med Inform Decis Mak,* vol. 21, no. 1, p. 334, Nov 28 2021, https://doi.org/10.1186/s12911-021-01690-9.
37. M. Selcuk, O. Koc, and A. S. Kestel, "The prediction power of machine learning on estimating the sepsis mortality in the intensive care unit," *Informatics in Medicine Unlocked,* vol. 28, p. 100861, 2022/01/01/ 2022, https://doi.org/10.1016/j.imu.2022.100861.
38. J. Y. Park *et al.*, "Predicting Sepsis Mortality in a Population-Based National Database: Machine Learning Approach," (in eng), *Journal of medical Internet research,* vol. 24, no. 4, p. e29982, Apr 13 2022, https://doi.org/10.2196/29982.
39. Bao C, Deng F, Zhao S. "Machine-learning models for prediction of sepsis patients mortality," (in eng). Med Intensiva. Jun2023;47(6):315–25. https://doi.org/10.1016/j.medine.2022.06.024.
40. Zhang K, Zhang S, Cui W, Hong Y, Zhang G, Zhang Z. "Development and Validation of a Sepsis Mortality Risk Score for Sepsis-3 Patients in Intensive Care Unit," (in English). Front Med. 2021;7: 609769. https://doi.org/10.3389/fmed.2020.609769.
41. N. Hou *et al.*, "Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost," (in English), *Journal of Translational Medicine,* vol. 18, 2020, https://doi.org/10.1186/s12967-020-02620-5.
42. S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate Imputation by Chained Equations in R," *Journal of Statistical Software,* vol. 45, no. 3, pp. 1 - 67, 12/12 2011, https://doi.org/10.18637/jss.v045.i03.
43. A. Jazayeri, O. S. Liang, and C. C. Yang, "Imputation of Missing Data in Electronic Health Records Based on Patients' Similarities," *Journal of Healthcare Informatics Research,* vol. 4, no. 3, pp. 295–307, 2020/09/01 2020, https://doi.org/10.1007/s41666-020-00073-5.
44. Dahiru T. "P - value, a true test of statistical significance? A cautionary note," (in eng). Annals of Ibadan postgraduate medicine. 2008;6(1):21–6. https://doi.org/10.4314/aipm.v6i1.64038.
45. D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Applied Soft Computing,* vol. 97, p. 105524, 2020/12/01/ 2020, https://doi.org/10.1016/j.asoc.2019.105524.
46. T. Rahman *et al.*, "BIO-CXRNET: a robust multimodal stacking machine learning technique for mortality risk prediction of COVID-19 patients using chest X-ray images and clinical data," *Neural Computing and Applications,* vol. 35, no. 24, pp. 17461–17483, 2023/08/01 2023, https://doi.org/10.1007/s00521-023-08606-w.
47. J. Prithula, M. E. H. Chowdhury, M. S. Khan, K. Al-Ansari, S. M. Zughaier, K. R. Islam, and A. Alqahtani, "Improved pediatric ICU mortality prediction for respiratory diseases: machine learning and data subdivision insights," *Respiratory Research,* vol. 25, no. 1, p. 216, 2024/05/23 2024, https://doi.org/10.1186/s12931-024-02753-x.
48. "SMOTETomek — Version 0.11.0," ed, 2023.
49. H. Taud and J. Mas, "Multilayer perceptron (MLP)," *Geomatic approaches for modeling land change scenarios,* pp. 451–455, 2018.
50. T. Chen *et al.*, "Xgboost: extreme gradient boosting," *R package version 0.4–2,* vol. 1, no. 4, pp. 1–4, 2015.
51. R. E. Wright, "Logistic regression," 1995.
52. R. Bhuvana, S. Maheshwari, and S. Sasikala, "Predict the Heart Disease Using a Logistic Regression Classifier Algorithm," in *2023 12th International Conference on System Modeling & Advancement in Research Trends (SMART)*, 22–23 Dec. 2023 2023, pp. 649–652, https://doi.org/10.1109/SMART59791.2023.10428486.
53. Liu Y, Wang Y, Zhang J. "New Machine Learning Algorithm: Random Forest," in *Information Computing and Applications*. Berlin, Germany: Springer; 2012. p. 246–52.
54. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn. 2006;63:3–42.
55. Schapire RE. "Explaining adaboost," in *Empirical Inference: Festschrift in Honor of Vladimir N*. Vapnik: Springer; 2013. p. 37–52.
56. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot. 2013;7:21.
57. A. V. Dorogush, V. Ershov, and A. Gulin, "CatBoost: gradient boosting with categorical features support," *arXiv preprint* arXiv:1810.11363, 2018.
58. Quinlan JR. Induction of decision trees. Mach Learn. 1986;1:81–106.
59. K. R. Islam *et al.*, "Prognostic Model of ICU Admission Risk in Patients with COVID-19 Infection Using Machine Learning," *Diagnostics,* vol. 12, no. 9, p. 2144, 2022. [Online]. Available: https://www.mdpi.com/2075-4418/12/9/2144.
60. Yusuf M, et al. "Reporting quality of studies using machine learning models for medical diagnosis: a systematic review," (in English). BMJ Open. 2020;10(3): e034568. https://doi.org/10.1136/bmjopen-2019-034568.
61. M. Ana Carolina Alba, "Discrimination and Calibration of Clinical Prediction Models: Users' Guides to the Medical Literature," (in English), *JAMA,* vol. 318, no. 14, pp. 1377–1384, 2017, https://doi.org/10.1001/jama.2017.12126.
62. P. Yun Liu, "How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature," (in English), *JAMA,* vol. 322, no. 18, pp. 1806–1816, 2019, https://doi.org/10.1001/jama.2019.16489.
63. Valverde-Albacete FJ, Peláez-Moreno C. "100% Classification Accuracy Considered Harmful: The Normalized Information Transfer Factor Explains the Accuracy Paradox," (in English). PLoS ONE. 2014;9(1): e84217. https://doi.org/10.1371/journal.pone.0084217.
64. L. Yong and L. Zhenzhou, "Deep learning-based prediction of in-hospital mortality for sepsis," *Scientific Reports,* vol. 14, no. 1, p. 372, 2024/01/03 2024, https://doi.org/10.1038/s41598-023-49890-9.
65. G. Zhang *et al.*, "Predicting sepsis in-hospital mortality with machine learning: a multi-center study using clinical and inflammatory biomarkers," *European Journal of Medical Research,* vol. 29, no. 1, p. 156, 2024/03/06 2024, https://doi.org/10.1186/s40001-024-01756-0.
66. A. Zlotnik and V. Abraira, "A general-purpose nomogram generator for predictive logistic regression models," *Stata Journal,* vol. 15, no. 2, pp. 537–546, 2015. [Online]. Available: https://econpapers.repec.org/article/tsjstataj/v_3a15_3ay_3a2015_3ai_3a2_3ap_3a537-546.htm.
67. Anderson RP, Jin R, Grunkemeier GL. "Understanding logistic regression analysis in clinical reports: an introduction," (in English). Ann Thorac Surg. 2003;75(3):753–7. https://doi.org/10.1016/s0003-4975(02)04683-0.

68. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 4768–4777.
69. "An introduction to explainable AI with Shapley values — SHAP latest documentation," ed, 2023.
70. Zhang Z, Qiu H, Li W, Chen Y. A stacking-based model for predicting 30-day all-cause hospital readmissions of patients with acute myocardial infarction. BMC Med Inform Decis Mak. 2020;20(1):1–13. https://doi.org/10.1186/s12911-020-01358-w.
71. Fleischmann-Struzek C, et al. "Incidence and mortality of hospital- and ICU-treated sepsis: results from an updated and expanded systematic review and meta-analysis," (in English). Intensive Care Med. 2020;46(8):1552–62. https://doi.org/10.1007/s00134-020-06151-x.
72. Giacomini MG, Lopes MVCA, Gandolfi JV, Lobo SMA. "Septic shock: a major cause of hospital death after intensive care unit," (in English). Revista Brasileira de Terapia Intensiva. 2015;27(1):51. https://doi.org/10.5935/0103-507X.20150009.
73. "sklearn.ensemble.ExtraTreesClassifier," ed, 2023.
74. "Glossary of Common Terms and API Elements," ed, 2023.
75. Palmowski L, et al. Assessing SOFA score trajectories in sepsis using machine learning: A pragmatic approach to improve the accuracy of mortality prediction. PLoS ONE. 2024;19(3): e0300739. https://doi.org/10.1371/journal.pone.0300739.
76. Chiu C-C, Wu C-M, Chien T-N, Kao L-J, Li C, Jiang H-L. "Applying an Improved Stacking Ensemble Model to Predict the Mortality of ICU Patients with Heart Failure," (in English). J Clin Med. 2022;11(21):6460. https://doi.org/10.3390/jcm11216460.

Publisher's Note