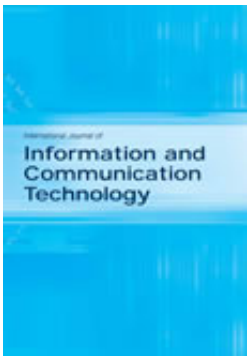
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

An explicable machine learning approach for predicting 30-day septic mortality for ICU patients

Liang Zhou, Ruiqian Wu, Shanshan Wang, Temitope Emmanuel Komolafe, Jiachen Guo, Zhiping Fan, Zhiwei Zhang

Article History:

Received:	06 September 2024
Last revised:	30 October 2024
Accepted:	31 October 2024
Published online:	28 April 2025

An explicable machine learning approach for predicting 30-day septic mortality for ICU patients

Liang Zhou
Shanghai University of Medicine and Health Sciences,
Shanghai, 201318, China
Email: wenzhou6@sjtu.edu.cn

Ruiqian Wu
University of Shanghai for Science and Technology,
Shanghai, 200093, China
Email: ruiqianwu@outlook.com

Shanshan Wang
Shanghai Jiao Tong University,
Shanghai, 200030, China
Email: nancybond@sjtu.edu.cn

Temitope Emmanuel Komolafe,
Jiachen Guo and Zhiping Fan
Shanghai University of Medicine and Health Sciences,
Shanghai, 201318, China
Email: teakomo@mail.ustc.edu.cn
Email: guojc@sumhs.edu.cn
Email: fanzp@sumhs.edu.cn

Zhiwei Zhang*
Shanghai Chest Hospital,
Shanghai, 200030, China
Email: zhangzhiwei@shchest.org
*Corresponding author

Abstract: This study developed MorSNX, a clinician-friendly model combining neural networks and XGBoost, to predict 30-day mortality risk in ICU patients with sepsis using vital signs and clinical data from the MIMIC-IV database. The top 25 predictive features were identified through backward stepwise regression, and SHAP values and decision curve analysis (DCA) enhanced interpretability. Validation with the eICU database demonstrated superior performance (AUC 0.9563), with significant clinical utility across decision thresholds, outperforming traditional models and scores, particularly at a 0.4 probability threshold. MorSNX offers a robust, interpretable tool for sepsis prognostication in critical care.

Keywords: sepsis; mortality risk; machine learning; XGboost; clinical decision support.

Reference to this paper should be made as follows: Zhou, L., Wu, R., Wang, S., Komolafe, T.E., Guo, J., Fan, Z. and Zhang, Z. (2025) 'An explicable machine learning approach for predicting 30-day septic mortality for ICU patients', *Int. J. Information and Communication Technology*, Vol. 26, No. 9, pp.1–22.

Biographical notes: Liang Zhou received his PhD degree from Donghua University in 2012. He is currently an Associate Professor in Shanghai University of Medicine and Health Sciences, Shanghai, China. His main research interests include big data analysis and intelligent medical auxiliary diagnosis.

Ruiqian Wu is a researcher in biomedical engineering, focusing on machine learning and mathematical modelling. Her work aims to develop practical algorithms and models to address challenges in clinical and biomedical applications, such as disease prediction and personalised medicine. With an emphasis on interdisciplinary collaboration, she strives to integrate computational methods into healthcare to support better patient outcomes and clinical decision-making.

Shanshan Wang holds a Master's degree in Public Administration from Shanghai Jiao Tong University. Her research interests include public health management and intelligent medicine.

Temitope Emmanuel Komolafe is an Assistant Professor at the Collaborative Research Centre, Shanghai University of Medicine and Health Sciences. His research focuses on intelligent medicine, medical image processing, and precision and diagnostic medicine. He earned his PhD in Biomedical Engineering from the University of Science and Technology of China, Hefei in 2021. He worked as a postdoctoral researcher at the School of Biomedical Engineering, Shanghai Tech University, from 2021 to 2023. He has authored numerous peer-reviewed journal articles and currently serves as a reviewer and Editor for *Data Mining and Management* (*Frontiers in Big Data*) as well as a book editor for Taylor & Francis.

Jiachen Guo received his PhD in Physical Electronics from Fudan University, Shanghai, China. His research interests are in the areas of intelligent sensing and detection technology, intelligent devices and active health.

Zhiping Fan received her PhD degree from Donghua University in 2022. She is currently a Lecturer in Shanghai University of Medicine and Health Sciences, Shanghai, China. Her research interests include intelligent auxiliary diagnosis, on-line fault detection and control, modelling and optimisation of process.

Zhiwei Zhang holds a Master's degree in Management and qualified as a Senior Engineer. Her research interests include smart hospital and intelligent healthcare systems. She has received awards such as third prize of Shanghai Science and Technology Progress Award, and second prize of Hospital Science and Technology Innovation Award from the China Hospital Association.

1 Introduction

Sepsis is characterised by life-threatening organ dysfunction resulting from a dysregulated host response to infection (Singer et al., 2016), poses a significant global health threat, claiming over 6 million lives annually. However, the true incidence and mortality rates of sepsis remain elusive due to variations in reporting criteria across different countries. This lack of standardised reporting may lead to substantial underestimation, particularly as populations age (Bauer et al., 2020; Cecconi et al., 2018). Despite the critical care provided in ICUs, sepsis remains a leading cause of mortality among ICU patients worldwide. Early identification and improved treatment of septic patients hinge on accurate mortality prediction (Schvetz et al., 2021).

In clinical sepsis diagnosis, commonly adopted scoring systems such as the sequential organ failure assessment (SOFA), acute physiology and chronic health evaluation score-III (APACHE-III), and logistic organ dysfunction score (LODS) have been established. However, these systems' accuracy can be influenced by the subjective judgment of physicians and healthcare teams. Variability in results may arise due to differences in individual experiences and preferences among physicians.

In a meta-analysis conducted in 2020 by Fleuren et al., it was found that ML models, notably eXtreme gradient boosting (XGBoost), exhibit remarkable accuracy in predicting sepsis episodes and prognosis. These ML models excel in identifying potential patients, offering advantages such as high accuracy, adeptness in handling large-scale datasets, and automatic feature selection and weight assignment, as evidenced by studies conducted by Zheng et al. (2023) and Wang et al. (2022).

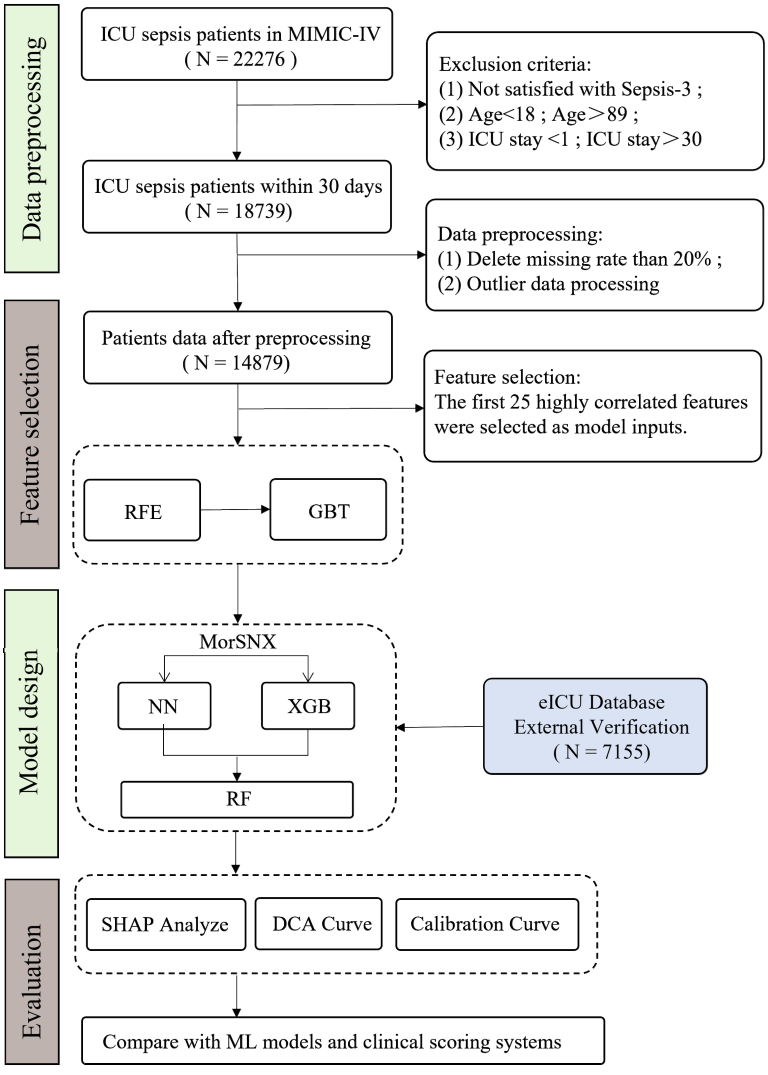
In recent years, studies have explored the use of deep learning (DL) models to analyse high-dimensional time series data for predicting mortality among ICU patients, as demonstrated by Wang et al. (2024). While many ML methods have demonstrated promising performance in clinical prediction, their 'black box' nature has hindered their applicability with real clinical data (Sabut et al., 2022). In critical decision-making scenarios like clinical prediction, the demand for models to be both accurate and interpretable is paramount (Rasheed et al., 2002). Notably, Hu et al. (2022) applied Shapley's game theory to predict sepsis mortality using the SHAP model, identifying six key features in ML models. Analysing these features aids healthcare professionals in understanding their impact on sepsis mortality, revealing patient progression mechanisms and enhancing risk factor. This innovative approach has inspired other researchers to use the SHAP method for disease prediction, leading to significant advancements in both clinical practice and scientific understanding (Wang et al., 2022).

This study introduces MorSNX (mortality prediction for sepsis using neural networks and XGBoost), a model that integrates DL and ML techniques to predict 30-day mortality among ICU sepsis patients. The choice of a 30-day time frame allows for the consideration of treatment effects, which may take time to manifest. By encompassing the entirety of patients' ICU stays, this timeframe provides a comprehensive evaluation of treatment outcomes. Data from the medical information mart for intensive care (MIMIC-IV), including records of 21,128 ICU septic patients, were used for both model training and independent testing. Additionally, a dataset comprising 31,045 patients from the electronic intensive care unit (eICU) collaborative research database was employed for external independent validation.

The model's performance was benchmarked against seven commonly used ML models and clinical scoring systems. Results demonstrate superior discriminative and

calibration capabilities, surpassing many existing models. Notably, the model exhibited high clinical utility as evidenced by decision curve analysis. This study presents a novel and accurate prognostic prediction method for septic patients. MorSNX offers clinicians an easily understandable and practical clinical tool for mortality risk assessment in the ICU setting.

Figure 1 The overall proposed framework for this study (see online version for colours)



2 Materials and methods

2.1 Data

2.1.1 Data source

Data for this study were sourced from two databases: MIMIC-IV (v2.2) and eICU (v2.0). The MIMIC database, initiated in 2003 by the laboratory of computational physiology at Massachusetts Institute of Technology (MIT), Beth Israel Deaconess Medical Center (BIDMC), Harvard Medical School (HMS), and Philips Healthcare (Goldberger et al., 2000), has been continuously updated, with MIMIC-IV (v2.2) released in June 2023 as an upgrade to MIMIC-III (Johnson et al., 2023). It encompasses comprehensive medical and surgical data for ICU patients at BIDMC from 2008 to 2019, ensuring patient privacy and confidentiality through desensitisation procedures such as de-identification, date and time offsetting, and sensitive field fuzzification.

The eICU database, used for model validation, contains over 200,000 ICU admissions from 334 ICUs across the USA from 2014 to 2015. It is a de-identified dataset rich in clinical data, widely used in healthcare studies and applications (Pollard et al., 2017). Accessing these databases requires passing the CITI PROGRAM exam and formally requesting permission via the MIMIC and eICU website (https://physionet.org/). The author has obtained research access with certification number 55393124. The study process is illustrated in Figure 1.

2.1.2 Inclusion criteria

This study used Navicat for PostgreSQL (v15.0) along with pgAdmin (v4.0) to identify sepsis patients based on the Sepsis-3 criteria (Deutschman, 2016). Based on the inclusion criteria of Sepsis-3, we included patients who met the following criteria as sepsis deaths: death occurring in the ICU due to sepsis, despite adequate fluid resuscitation, and necessitating the use of vasoactive agents to maintain a mean arterial pressure (MAP) of 65 mmHg or higher, coupled with a blood lactate level greater than 2 mmol/L. In addition, in this study, we limited the patient's age to: $18 \leq \text{Age} \leq 89$, and the patient's ICU stay time limit to: $1 < \text{ICU stay} \leq 30$ days.

This study extracted 51 characteristic variables related to sepsis and mortality, considering the pathophysiological process and clinicians' assessments of sepsis patients. These variables include 5 demographic attributes, 31 laboratory metrics, and 15 vital signs indicators. In medical monitoring, patient vital signs and laboratory readings are continuously recorded over short time spans, resulting in substantial clinical data accumulation. To address this, our method adopts an approach that captures dynamic physiological changes in patients. Specifically, when extracting laboratory indicators, we select variables based on data characteristics and estimate their maximum, minimum, and average values, providing a more comprehensive perspective on patient physiological processes.

2.1.3 Statistical analysis

This study stratified patients under follow-up into two groups: survivors and non-survivors within a 30-day period. We then compared variables across these groups. Continuous variables were presented as medians with quartiles [$M(Q_L,Q_U)$], while

categorical variables were represented as counts and percentages. We conducted a preliminary normality analysis for each variable, using the Kolmogorov-Smirnov test for normally distributed continuous variables and the Mann-Whitney U test for non-normally distributed continuous variables (Chiew et al., 2020). To compare the two groups, we used the chi-square test. Statistical significance was set at $P<0.05$. All analyses were performed using SPSS software (v26).

2.1.4 Data preprocessing

In clinical settings, dealing with incomplete and inconsistent data can be challenging. To tackle this issue, we performed essential data preprocessing on the MIMIC-IV database, including detecting missing values, handling outliers, and oversampling. Missing values were identified by calculating their proportion in the dataset, with a threshold set at 0.2, the proportion of missing value for each feature is shown in Table S1 in the supplementary materials. Any patient data or feature variables with missing rates exceeding this threshold were removed. For data with missing rates below 20%, we used imputation techniques, i.e. means for numerical data, modes for character data, and zeros for null entries.

We applied a 3σ outlier detection method based on standard deviation to identify significant deviations from the dataset's average ($\mu \pm 3\sigma$), where μ and σ represent the mean and standard deviation, respectively, as shown in the supplementary Table S2. To ensure clinically meaningful handling of outliers, we differentiated between clinically plausible outliers (e.g., extreme physiological values due to a patient's condition) and data entry errors. Clinically plausible outliers were retained to preserve the dataset's integrity, while data entry errors were removed. For cases with a small number of abnormal values, we replaced them with the mean or mode depending on the column's characteristics.

Additionally, to ensure consistency in preprocessing across different datasets, we applied the same rules and criteria for handling missing values and outliers regardless of the dataset's source. This approach guarantees that preprocessing is consistent, reproducible, and robust, which is critical for ensuring the reliability of the model across different patient populations. Table S5 in the supplementary materials shows the results we obtained after applying the same preprocessing strategies (including missing value processing and outlier detection) on the external dataset eICU. In addition, we plot the outlier probability density plots for the external dataset in Figure 5 in the supplementary to visualise the distribution of outliers after applying these strategies. These results provide an important validation basis for the application of the model on different datasets, ensuring the consistency and effectiveness of the method.

After categorising 14,879 sepsis patients for mortality markers, the dataset exhibits a severe imbalance issue, with 3,141 patient deaths far fewer than 11,738 patient survivals. This imbalance could affect the model's performance, favouring the predominant survival category. To ensure that our model is able to learn sufficient information and accurately reflect its capability to recognise the minority class, we have adopted the following strategies for splitting our dataset into training and test sets. We employed stratified sampling to maintain the proportion of deceased patients (the minority class) in the training and test sets consistent with the original dataset. This approach preserves the representativeness of class proportions and reduces bias towards the majority class. Furthermore, to balance the class distribution in the training set, we performed random

sampling among the surviving patients (the majority class) to select a number of samples equal to the number of deceased patients, thereby creating a balanced training dataset. While this method reduces the volume of the majority class samples used during training, it enhances the model's ability to recognise the minority class. A similar approach is applied to the construction of the test set to ensure the accuracy of model evaluation. To address potential inconsistencies between different datasets, we ensured that stratified sampling and preprocessing rules were applied uniformly across all datasets, thereby guaranteeing consistent handling of data imbalance and ensuring that the model's performance is robust across diverse data sources.

To mitigate the randomness introduced by specific data splits and to enhance the robustness of our model evaluation, we implemented 5-fold cross-validation. In each fold, the entire dataset is re-sampled using stratified sampling to generate new training and test sets. This ensures that each test set is independent and has not appeared in the training set, thus preventing data leakage.

By adopting this methodology, we ensure fair representation of each class during both training and testing, and we are also able to comprehensively evaluate the performance of the model through the results of the 5-fold cross-validation, including its generalisation ability and recognition of the minority class. This approach aids in developing a predictive model that is both fair and effective.

2.2 Model development

2.2.1 Feature selection

Our proposed MorSNX model combines the recursive feature elimination (RFE) algorithm with a gradient boosting tree (GBT) methodology. By adopting 3-fold cross-validation, this study aims to enhance model generalisation and identify the optimal feature subset, as previously outlined. The training process iteratively eliminates the least important feature in each iteration, and the important feature is evaluated. Model performance is assessed via cross-validation for robustness and reliability. The best predictive model performance subset is chosen from the estimated results. The model employs GBT to iteratively train decision trees for improved performance. In each iteration, GBT adjusts tree weights based on the previous round's error, registering performance scores and retained feature subsets. After cross-validation, the model selects the best feature subset. These top 25 features are ranked for relevance and used as inputs for the prediction model from the original 43 features.

To evaluate the efficacy of our feature selection method, this search compared it against mutual information and correlation coefficient methods, utilising the XGBoost model. We employ accuracy as the primary evaluation metric, with detailed results provided in Table S3 and Figure S1 in the supplementary materials.

2.2.2 Model design for MorSNX

The MorSNX model, a novel approach tailored for predicting septic mortality in intensive care units, is introduced in Figure 2. This model leverages a stacking-based architecture, combining the strengths of neural networks and XGBoost as base learners, with random forest (RF) serving as the meta-learner. This design effectively addresses diverse data processing needs, maximising the advantages of each model component.

In terms of base models' proficiency, XGBoost is strategically adapted for its adept handling of tabular and structured datasets. Meanwhile, neural networks are leveraged for their proficiency in managing large-scale structured datasets, particularly adept at capturing high-dimensional data and complex nonlinear relationships. Conversely, RF is incorporated as the meta-model due to its robust stability and intrinsic resistance to overfitting, thus serving as a reliable adjudicator for the final classification output.

The amalgamation of heterogeneous model predictions endows MorSNX with a comprehensive multi-perspective view of the data landscape, allowing it to capture and synthesise underlying data features and patterns more holistically (Divina et al., 2018; Sujan et al., 2022). This strategic blending mitigates the risk of individual model underperformance by distributing predictive responsibility among the models, thus leading to a more consistent and reliable prediction outcome. In essence, MorSNX represents a classification paradigm that effectively integrates the diverse model dynamics to serve a unified purpose: accurately classifying sepsis mortality risk. This enhances the model's versatility and predictive performance, making it a valuable tool for ICU clinicians.

In this study, we applied Keras-based neural network model with two hidden layers, including a dropout layer with a 0.2 deactivation probability to prevent overfitting. The output function of each hidden layer can be expressed as:

$$a^i = \frac{e^{w^{[i]}x+b^{[i]}}}{e^{w^{[i]}x+b^{[i]}}+1} \tag{1}$$

where a^i is the output of the hidden layer, $w^{[i]}$ is the weight matrix of hidden layer i, x is the 25 feature vectors of the input, and $b^{[i]}$ is the bias term of the hidden layer.

XGBoost handles both classification and regression problems, enhancing flexibility and robustness (Chen and Guestrin, 2016). The objective function of our model can be written as:

$$\mathcal{L}(\varnothing) = -\left[\sum_{i=1}^{N}\left(y_i \log\left(s_i\right)+\left(1-y_i\right)\log\left(1-s_i\right)\right)\right]+\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2 \tag{2}$$

This formula consists of a loss function and a regularisation term, i denotes a certain sample, and y_i, s_i denotes the true and predicted values of the i^{th} sample, respectively, T represents the number of leaf nodes, w represents the weight, and γ and λ indicate penalty terms for various complexity terms.

The optimal value of the weight of each leaf node is:

$$w_j^* = -\frac{\left(\frac{y_j}{p_j^2}+\frac{1-y_j}{\left(1-p_j\right)^2}\right)}{\left(\frac{1-y_j}{1-p_j}-\frac{y_j}{p_j}\right)+\lambda} \tag{3}$$

The optimal value of the objective function is:

$$obj^* = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\frac{y_j}{p_j^2}+\frac{1-y_j}{\left(1-p_j\right)^2}\right)^2}{\left(\frac{1-y_j}{1-p_j}-\frac{y_j}{p_j}\right)+\lambda}+\gamma T \tag{4}$$

By estimating this objective function, XGBoost minimises total error, enhancing predictive performance on new data (detailed algorithm implementation details can be found in the supplementary material).

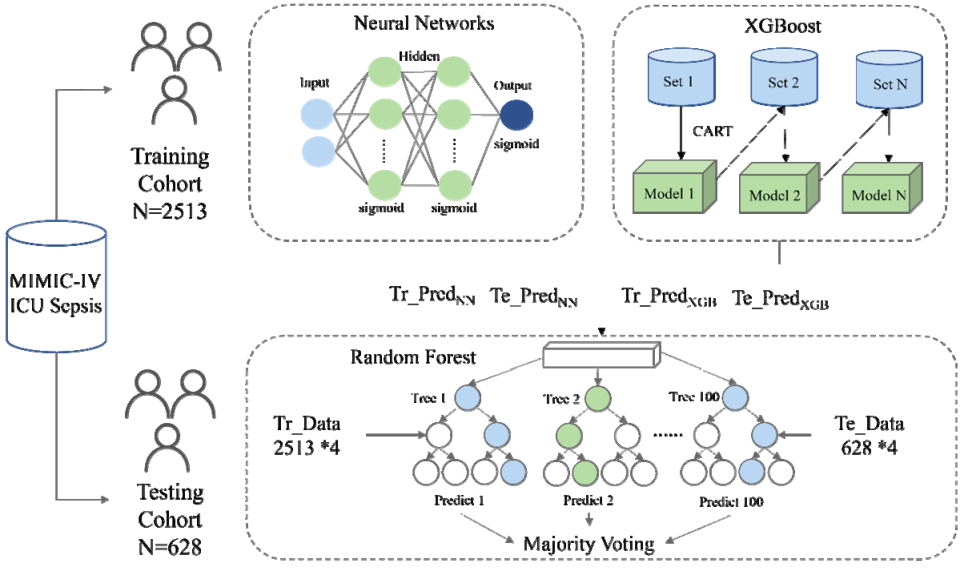
It is essential to fine-tune hyperparameters for efficient resource utilisation before adding each base model. Bayesian optimisation, an adaptive approach, selects promising parameter settings based on previous experiments, continually improving parameter combinations without predefined grids. This approach greatly enhances parameter optimisation efficiency (Liu et al., 2024). Combining the characteristics of the base model, the Bayesian hyperparameter optimisation process can be represented by the following equation:

$$\Phi_{EI}(x)=\int_{-\infty}^{loss^*(x)}\left(loss^*(x)-loss(x)\right)N\left(loss(x);\mu(x),K\left(x,x'\right)d\ loss(x)\right. \tag{5}$$

The equation (5) above represents the expected improvement in hyperparameter optimisation, $\Phi_{EI}(x)$ denotes the expected improvement under a given hyperparameter configuration, $loss^*(x)$ is the value of the optimal objective function to which the model is fitted, and N is a Gaussian distribution, which represents the probability distribution under a given mean and covariance function.

The algorithmic flow of the MorSNX model is depicted in Figure 2. It operates on a dataset consisting of 14,879 patient records with 25 features as input. This dataset is partitioned into a test set (628 × 25) and a training set (2,513 × 25) for the meta-model, with a split ratio of 0.2 to 0.8. Using 5-fold cross-validation, the 2513 training data is split into five subsets, with 502 records (1-fold) for base model testing and 2011 records (4-fold) for base model training. Initially, the first-layer neural network base model is trained to obtain the NN_1 model. The trained model predicts the base model's test data, resulting in 502 × 2 training data recorded as Tr_Pred_{NN1}. This process repeats for the 4-fold validation, resulting in four sets of 502 × 2 training records denoted as Tr_Pred_{NNi}, where 'i' connotes the fold number. The first-layer neural network training is completed, yielding a total of 2513 × 2 training records. Similarly, the data from the 5-fold split is used to train the second-layer XGBoost-based model, resulting in one set of 502 × 2 and four sets of 2011 × 2 training records. Predictions are made for each dataset and labelled as Tr_PredXGBi, where 'i' signifies the fold number. Tr_Pred_{NN} and Tr_Pred_{XGB} data are merged into Tr_Data, serving as input to the metamodel. The 628 test sets from the meta-model undergo 5-fold cross-validation in the neural network base model. Predicted values are averaged across folds, resulting in 628 × 2 test sets labeled as Te_Pred_{NN} after the first-layer base model training. The second-layer base model test set is called Te_Pred_{XGB}. Te_Pred_{NN} and Te_Pred_{XGB} combine to form Te_Data. Tr_Data is used to train the RF metamodel, which is then tested on Te_Data.

Figure 2 The comprehensive algorithmic flow of the MorSNX model (see online version for colours)



To mitigate overfitting, an early-stopping strategy is implemented, which monitors the validation loss. Training is stopped upon reaching the minimum value to prevent overfitting to the training data.

2.2.3 Model comparisons

We conducted experimental comparisons on various ML models, including RF, XGBoost, k-nearest neighbours (KNN), light gradient boosting machine (LightGBM), and decision trees (DT). We optimised these models using the recursive feature elimination method and compared different feature selection strategies. To validate our model, we also compared clinical scoring systems for predicting the risk of death in ICU sepsis patients, including SOFA, impact of acute physiology score III (APSIII) score, LODS score, OASIS score, simplified acute physiology score II (SAPSII) score, and systemic inflammatory response syndrome (SIRS) score.

2.2.4 Model interpretability

The ability to provide explanations for features is more likely to gain acceptance and be applied by clinicians compared to high-dimensional AI models. These models excel at complex data processing and prediction but lack transparency in their internal processes, limiting their ability to offer detailed justifications for their decisions, especially in medical diagnosis.

SHAP, based on game theory and first proposed by UCLA professor Lloyd Shapley, aims to equitably distribute the benefits of cooperation by considering each player's contribution to the project. In the field of ML, SHAP performs post-interpretation of models, interpreting the predicted value as the sum of the attributed values of input features to measure their contribution to the model (Baptista et al., 2022).

The core formula is:

$$g\left(Z_j\right) = \varphi_0 + \sum_{j=1}^{M} \varphi_j Z'_j \tag{6}$$

where g represents the explanatory model, M is the number of input features, in this study, the value of M is set at 25, Z_j is the parameter that determine whether the feature exists or not for a certain sample, φ_0 is a constant that represents the predicted mean value of all the training samples, and φ_j is the attributed value of each feature known as the shapely value of the feature.

3 Results

3.1 Data analysis

This study analysed data from 18,739 sepsis patients in MIMIC-IV database, all with stays under 30 days. Each patient had 51 demographic and clinical variables. To ensure data integrity, records with a missing rate above 20% for patient and feature information were excluded, resulting in 14,879 patient records with 47 features. Statistical analysis (see Table 1) confirmed the removal of statistically non-significant features.

Table 1 Demographic and clinical characteristics for the ICU admissions included into MIMIC-IV dataset

Features	*Survive*	*Death*	*P*
Number	11,738	3,141	
Baseline variables and in-hospital factors			
Age_mean	66.60 (55.78, 76.29)	69.62 (59.44, 79.04)	<0.001
Sex(%)			
Female	7,178 (61%)	1,815 (58%)	0.612
Male	4,546 (39%)	1,326 (42%)	
Vital signs			
los_icu(Days)	3.12 (1.83, 6.32)	5.12 (2.69, 9.65)	<0.001
los_hospital(Days)	8.96 (5.68, 15.61)	7.83 (3.66, 14.49)	<0.001
Hematocrit_max(times/min)	34.30 (30.40, 38.60)	33.10 (28.70, 38.40)	<0.001
Hematocrit_min(times/min)	29.10 (25.00, 33.60)	28.10 (23.90, 33.20)	0.060
HeartRate_max(times/min)	103.00 (91.00, 118.00)	113.00 (97.00, 129.00)	<0.001
Resprate_max(times/min)	27.00 (24.00, 32.00)	30.00 (26.00, 35.00)	<0.001
Temperature_max(℃)	37.44 (37.06, 38.00)	37.33 (36.89, 38.00)	<0.001
Sofa_score	3.00 (2.00, 4.00)	4.00 (3.00, 6.00)	<0.001
Sapsii	37.00 (29.00, 45.00)	50.00 (40.00, 62.00)	<0.001
Sapsii_prob_max	0.20 (0.10, 0.35)	0.46 (0.25, 0.72)	<0.001
Charlson_score_max	5.00 (3.00, 7.00)	7.00 (5.00, 9.00)	<0.001
Sbp_max(mmHg)	145.00 (133.00, 160.00)	144.00 (129.00, 160.00)	<0.001
Sbp_min(mmHg)	88.00 (80.00, 97.00)	83.00 (73.00, 93.00)	<0.001
Sbp_mean(mmHg)	113.68 (106.00, 123.76)	109.00 (100.96, 120.55)	<0.001

Note: The abbreviations of some features can be found in the attachment

Table 1 Demographic and clinical characteristics for the ICU admissions included into MIMIC-IV dataset (continued)

Features	*Survive*	*Death*	*P*
Vital signs			
Spo2_max(%)	99.00 (97.00, 100.00)	98.00 (96.00, 99.00)	<0.001
Spo2_min(%)	93.00 (90.00, 95.00)	91.00 (87.00, 94.00)	<0.001
Spo2_mean(%)	97.48 (96.00, 98.67)	96.88 (95.08, 98.51)	<0.001
Dbp_max(mmHg)	83.00 (73.00, 96.00)	85.00 (73.00, 99.00)	0.324
Dbp_min(mmHg)	45.00 (39.00, 51.00)	42.00 (35.00, 49.00)	<0.001
Dbp_mean(mmHg)	60.40 (54.92, 66.84)	59.39 (52.94, 66.62)	<0.001
Urine_max(mL)	1,725.00 (1,114.00, 2,512.25)	980.00 (425.00, 1,727.00)	<0.001
Laboratory parameters			
Aniongap_max(mEq/L)	15.00 (13.00, 18.00)	18.00 (15.00, 22.00)	<0.001
Aniongap_min(mEq/L)	12.00 (10.00, 14.00)	14.00 (12.00, 17.00)	<0.001
Bicarbonate_max(mmol/L)	24.00 (22.00, 27.00)	23.00 (20.00, 26.00)	<0.001
Bicarbonate_min(mmol/L)	22.00 (19.00, 24.00)	19.00 (16.00, 23.00)	<0.001
Inr_max	1.30 (1.20, 1.60)	1.60 (1.20, 2.30)	<0.001
Sodium_max(mmol/L)	140.00 (138.00, 142.00)	140.00 (137.00, 144.00)	<0.001
Sodium_min(mmol/L)	137.00 (135.00, 140.00)	137.00 (133.00, 140.00)	0.022
Chloride_max(mmol/L)	107.00 (103.00, 111.00)	106.00 (100.50, 111.00)	<0.001
Chloride_min(mmol/L)	103.00 (99.00, 106.00)	101.00 (96.00, 105.00)	<0.001
Bun_max(mmol/L l)	20.00 (14.00, 33.00)	34.00 (21.00, 56.00)	<0.001
Bun_min(mmol/L)	17.00 (12.00, 27.00)	27.00 (17.00, 47.00)	<0.001
Wbc_max(10^9/L)	14.10 (10.30, 18.80)	15.00 (10.30, 21.15)	0.389
Wbc_min(10^9/L)	9.80 (7.00, 13.10)	10.60 (6.80, 15.15)	0.055
Hemoglobin_max(10^9/L)	11.30 (9.90, 12.80)	10.70 (9.20, 12.40)	<0.001
Hemoglobin_min(10^9/L)	9.70 (8.30, 11.20)	9.10 (7.70, 10.80)	<0.001
Creatinine_max(mg/dL)	1.00 (0.80, 1.60)	1.60 (1.00, 2.60)	<0.001
Creatinine_min(mg/dL)	0.90 (0.70, 1.30)	1.20 (0.80, 2.00)	<0.001
Sus_anti_period	0.29 (0.05, 0.77)	0.33 (0.14, 0.97)	0.022
Cul_anti_period	0.29 (0.05, 0.77)	0.33 (0.14, 0.97)	0.022
Antibiotic_num	4.00 (2.00, 7.00)	7.00 (4.00, 11.00)	<0.001
Specimen_count	4.00 (2.00, 7.00)	7.00 (4.00, 11.00)	<0.001
Positive_culture	0.00 (0.00, 0.00)	0.00 (0.00, 1.00)	<0.001
C-reactive protein(max)	51.7 (9.1, 132.525)	63.15 (17,151.02)	0.002

Note: The abbreviations of some features can be found in the attachment

To evaluate the model's generalisability and applicability, validation was conducted using the eICU database, adhering to identical cohort inclusion and exclusion criteria. Due to inherent differences between the MIMIC-IV and eICU databases in terms of feature availability and data collection methods, not all features selected for training in the MIMIC-IV dataset were present in the eICU dataset. Additionally, certain features were removed due to missing value thresholds. To address this, we aligned the feature

sets by selecting common features available in both datasets and applied appropriate preprocessing techniques. This ensures that the model is evaluated fairly on the validation dataset, while maintaining clinical relevance. Detailed descriptions of the extracted variables and the characteristics of real patients can be found in Table S4 in the supplementary material. The excluded features are also labelled in Table 4.

3.2 Data preprocessing and feature selection

Outlier detection involved using the 3σ method for each feature, analysing the causes of outliers. We kept outliers related to patient diseases and addressed those arising from irregular record-keeping by replacing character data with mode values and numerical data with mean values. After data preprocessing, the study included 14,879 patient records, as depicted in Figure 1 and Table 2 in the supplementary materials.

This study compares a RFE algorithm, with the mutual information and correlation coefficient methods. The top 25 features from each feature selection method are trained using the XGBoost model, and ROC values are used for evaluation, the trained results are as shown in Figure S2 in the supplementary materials. Notably, the RFE method provides the highest accuracy, making it the chosen method for this study.

3.3 Model evaluation

After preprocessing and feature selection, we built ML models (Logistic Regression, RF, KNN, DT, LightGBM) in Python. The data was split into train and test sets (0.2 ratio) and standardised for comparable scales. Model performance is shown in Figure 3(a) and Figure 3(c). Hyperparameters were optimised using Bayesian methods for complex models and grid search for simpler ones. These models were evaluated with 5-fold cross-validation, and metrics (AUROC, precision, recall, accuracy, F1 score) were computed as in Table 2.

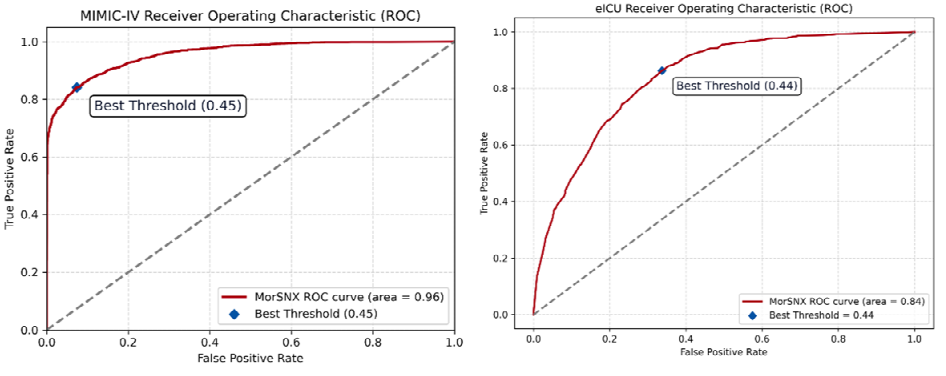
Table 2 Comparison performance of our proposed method with different models in MIMIC-IV

Model	*ROC*	*Accuracy*	*Precision*	*Recall*	*F1 score*
MorSNX	0.96	0.89	0.91	0.82	0.87
XGBoost	0.93	0.89	0.90	0.84	0.87
Neural network	0.89	0.81	0.81	0.76	0.78
KNN	0.90	0.82	0.75	0.88	0.81
LightGBM	0.89	0.81	0.88	0.67	0.76
Decision tree	0.86	0.79	0.75	0.78	0.76
Random forest	0.87	0.77	0.73	0.77	0.75
Logistic	0.82	0.75	0.73	0.67	0.70

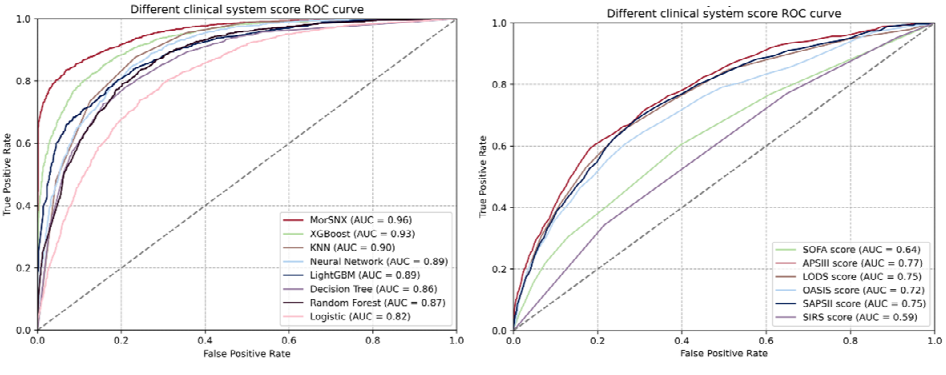
In the eICU external validation cohort, the model's parameter settings remained consistent with the original experiment. The results, as depicted in Figure 3(b), demonstrated credibility and generalisation. The ROC value was 0.8397 (95% CI: 0.8224–0.8519), with accuracy, precision, recall rate, and F1 score at 0.7560, 0.6982, 0.7965, and 0.7441, respectively. This underscores the model's versatility and effectiveness across various environments and patient groups. Due to the lack of some

indicators related to sepsis mortality risk in the eICU database, the features in the external database are inconsistent with the features used in the training set and test set, which will affect the performance of the model.

Figure 3 (a) ROC curves of MorSNX model by the 5-fold cross-validation in the MIMIC-IV testing cohort – the ROC area is 0.9563, the optimal threshold for this model is 0.44 (in this case, the true positive rate is the highest and the false positive rate is the lowest) (b) ROC curves for the eICU database validated as an external database - the area is 0.8397 (c) ROC curves of other commonly used machine learning models by the 5-fold cross-validation in the MIMIC-IV testing cohort (d) ROC curves of the six most widely used clinical scoring systems (see online version for colours)



(a) (b)



(c) (d)

In this experiment, we compared five traditional clinical scoring systems using predictive results from the MIMIC-IV database. We evaluated each system using the same scoring indexes and plotted ROC curves for each, as shown in Figure 3(d). The results indicate that our proposed model outperforms others in all aspects, with an ROC value of 0.9563 (95% CI: 0.9518–0.9626), surpassing most models in sepsis death risk prediction and outperforming the widely used clinical scoring system. Our model excels in predicting sepsis death risk in ICU.

3.4 Interpretability analysis

This study's model accurately predicts patient mortality risk and highlights the importance of specific variables for patient prognosis. Figure 4 illustrates the top 25 most important variables, ranked by importance, based on the results of the SHAP analysis. Key factors affecting patient prognosis, ranked by importance, include ventilation status, anion gap, maximum urine output, coagulation function, ICU length of stay, antibiotic administration status, sodium levels, heart rate, body temperature, blood urea nitrogen, peripheral capillary oxygen saturation and creatinine. These variables significantly influence the mortality risk in septic patients. Notably, variables such as ventilation status, anion gap, maximum urine output, coagulation function, and antibiotic administration status demonstrated strong positive associations with mortality risk, which align with well-established prognostic factors in sepsis. In this study, a SHAP scatterplot was also plotted to show the SHAP value of each feature for a single sample to help understand how the predictions of the model are contributed by different features, as shown in Figure 4.

We employed the partial dependence plot (PDP) method to conduct an interpretability analysis of the model results, as illustrated in the Figure 3. We plotted the three key feature variables with the most significant impact on septic mortality:

a ventilation status

b anion gap

c urine output.

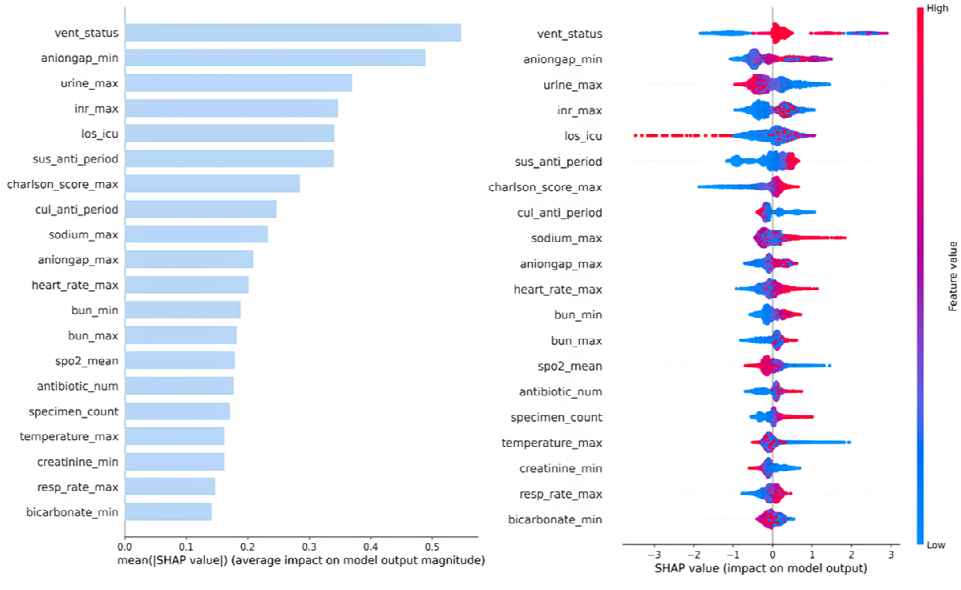
In Figure 3(a), we observed a trend in the impact of ventilation status on septic mortality as the patient's duration in the ICU increased. The risk of septic mortality is highest when the ventilation status values are 1, 2 and 3, corresponding to high-flow oxygen therapy, invasive ventilation, and tracheostomy ventilation methods, respectively. It can be seen from Figure 3(b) that when the anion gap increases to 15, the patient's risk of death reaches the highest level. In addition, urine output is inversely related to the risk of death in sepsis. When the urine output is reduced to 1,000 ml and below, the patient's death risk is also most significant, as illustrated in Figure 3(c).

This study utilises DCA to evaluate the clinical utility of different models in Figure 5. DCA evaluates the model's performance at various thresholds, aiding in the selection of the optimal treatment strategy. In Figure 5(a), the MorSNX model demonstrates positive net benefit at different thresholds, indicating its ability to distinguish between positive and negative cases. We also assessed the clinical effectiveness of this model within eICU as depicted in Figure 5(b). Furthermore, a comparison of our model with commonly used ML models is shown in Figure 5(c), where the XGBoost model exhibits similar clinical utility, suggesting its potential as a competitive alternative in this study. Figure 5(d) illustrates the clinical utility of different clinical scoring systems for predicting ICU sepsis mortality, notably, around the threshold of 0.5, various clinical scoring systems lose significance, whereas our model demonstrates optimal net benefit at most thresholds.

The calibration curve was also used to assess the binary classification model. Ideally, the calibration curve should resemble a 45-degree diagonal line, indicating consistent predicted probabilities with actual values. Figure 4(a) shows that the model in this study adheres to the diagonal line, with predictions closely matching real values, while other ML models deviate from the diagonal line as seen in Figure 4(b). Some clinical scoring

systems provide accurate probability predictions up to a confidence level of 0.4, similar to the actual event frequency. However, beyond 0.4, the model's probability predictions begin to deviate, potentially overestimating or underestimating probabilities. Overall, the proposed model's calibration curves outperform clinical scoring systems.

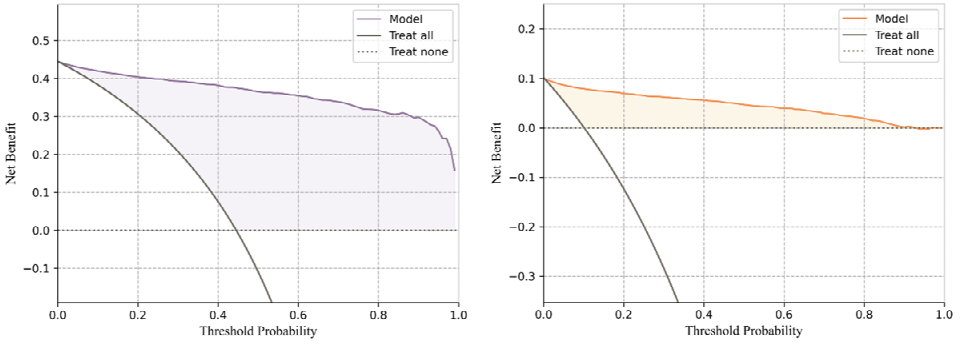
Figure 4 The contribution of each feature to MorSNX model computed by SHAP, (a) SHAP feature weights (b) SHAP scatterplot (see online version for colours)



(a) (b)

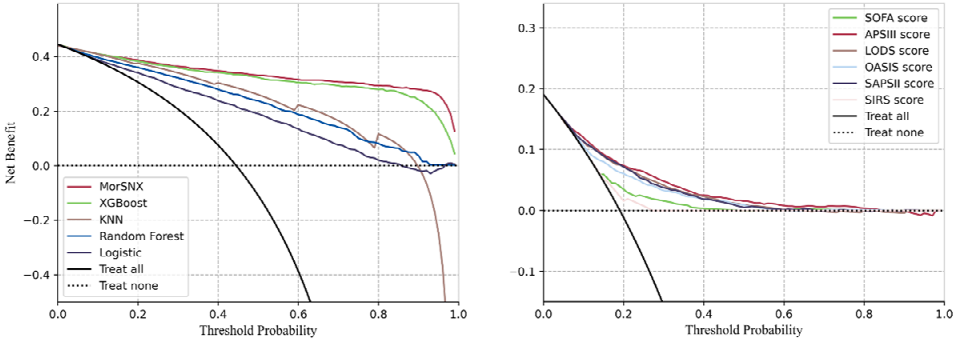
Note: X-axis indicates feature's impact on MorSNX model Y-axis represents the variables, the higher the ranking the more important the variable

Figure 5 (a) DCA curves of our MorSNX model in MIMIC-IV database (b) DCA curves of our MorSNX model in the independent test database(eICU) (c) DCA curves of traditional ML models (d) DCA curves of traditional clinical scoring systems (see online version for colours)



(a) (b)

Figure 5 (a) DCA curves of our MorSNX model in MIMIC-IV database (b) DCA curves of our MorSNX model in the independent test database(eICU) (c) DCA curves of traditional ML models (d) DCA curves of traditional clinical scoring systems (continued) (see online version for colours)



(c) (d)

4 Discussion

The model we proposed outperforms other ML and traditional critical care scoring models in various ways. Firstly, it demonstrates higher accuracy and discriminative performance in predicting patient mortality risk, outperforming previous studies. Secondly, in this study, the model used the GBT algorithm along with RFE for feature selection, automatically selecting the top 25 features most correlated with the target label from high-dimensional features. This approach ensures that the selected features align with the model's characteristics and commonly used clinical variables.

This model assessed the importance of various predicted outcomes, including ventilation status, anion gap, maximum urine output, coagulation function, ICU length of stay, antibiotic administration status, sodium levels, heart rate, body temperature, blood urea nitrogen, peripheral capillary oxygen saturation and creatinine. These factors were evaluated for clinical applicability through SHAP value analysis and DCA curve test, providing insights into model performance and its comparison with established clinical scoring systems. Despite the model's complexity, it demonstrated efficient processing times, taking 39.9 seconds for training and 38.6 seconds for testing, totalling 68.5 seconds. This suggests that the model is both accurate and efficient for real-world clinical settings.

We analysed the effect of top-ranked features on sepsis. According to our predictions, ventilation status had the greatest influence on the risk of mortality in septic patients. This may be due to the clinical context where septic patients often require oxygen therapy upon ICU admission. However, excessive oxygen therapy can weaken immune defenses, increase the risk of hyperoxic acute lung injury, promote vessel constriction, and reduce coronary blood supply, thereby increasing the risk of death. Invasive ventilation, which connects patients to a ventilator, may also induce immune responses, bacterial colonisation, and inflammatory imbalances, all of which can increase infection rates (Lundberg et al., 2020; Ren et al., 2020). In addition to ventilation status, the anion gap emerged as another important factor influencing sepsis prognosis. Metabolic acidosis,

characterised by an elevated anion gap, is common in septic patients and is associated with disruptions in cardiovascular function, renal physiology, and inflammatory mediator pathways, including nitric oxide (NO) (Costa et al., 2024). While NO normally induces arterial dilation, in sepsis, dysregulated vasodilation exacerbates hypotension. Moreover, inflammatory mediator imbalances can worsen sepsis outcomes. The model also predicted a correlation between reduced urine output and increased mortality risk during ICU admission. Septic patients often experience reduced blood volume due to vasodilation and capillary leakage, impairing circulation and urine output. Inflammatory mediators such as TNF-alpha can further cause capillary leakage, exacerbating plasma volume reduction and impacting kidney function. As a result, monitoring urine output is a key indicator of circulatory and renal function, which is consistent with clinical findings regarding septic mortality (Leedahl et al., 2014).

We observed that the model's accuracy on the external eICU dataset was 0.84, slightly lower than on the MIMIC-IV dataset. This difference likely stems from variations in data collection methods, frequencies, and feature definitions between hospitals and healthcare systems. Certain key features present in MIMIC-IV, such as the Charlson score and antibiotic-culture interval, were either missing or ambiguously recorded in eICU, and thus excluded from external validation. Additionally, differences in the recording methods and distributions of some laboratory indicators further impacted the model's performance. These discrepancies, common in medical datasets, typically result in lower performance on external datasets. To improve generalisability, we applied random cross-validation during training and introduced L1 and L2 regularisation to prevent overfitting. While regularisation stabilised the model's performance on MIMIC-IV, improvements on the eICU dataset were minimal, with only slight gains in accuracy and AUC. This suggests that despite reducing feature dependency, dataset differences still limit the model's generalisation to external data.

In recent studies focusing on sepsis prediction using ML algorithms, Hu et al. (2022) conducted a comparative analysis of various ML models to predict in-hospital mortality among critically ill septic patients. Their findings highlighted the superiority of the XGBoost model, boasting an impressive AUC value of 0.884. However, it's noteworthy that their model lacked validation on a multicenter external dataset, restricting its generalisation capability to broader patient populations and healthcare settings.

Likewise, Wang et al. (2022) explored prognosis modelling for sepsis, employing LightGBM and achieving commendable interpretability. Their study identified key prognostic factors such as ICU stay duration, urine output, ventilation status, and antibiotic usage, mirroring some of our own observations. Despite this, their model exhibited lower F1 scores and precision values, indicative of substantial predictive errors, and the absence of independent validation against external datasets raises concerns regarding its robustness and clinical reliability.

Zheng et al. (2023) introduced the ShockSurv model, utilising XGBoost to predict 28-day mortality specifically in ICU septic shock patients. While innovative, their model's focus on septic shock patients limits its applicability to broader sepsis cases. Moreover, their subjective feature selection approach and lack of feature screening may compromise its generalisation ability across different clinical scenarios.

Many prior studies have leaned towards biomarker-based approaches for sepsis mortality prediction. However, reliance on biological variables not readily available in clinical practice poses implementation challenges. In contrast, our model prioritises clinically accessible data variables, enhancing feasibility for real-world application.

Addressing limitations observed in previous studies, our model boasts enhanced robustness, validated through rigorous external validation. This validation reinforces its stability and reliability. Table 3 provides a summary and comparison of previous studies in the field of sepsis mortality prediction.

Table 3 Performance comparison of existing techniques in sepsis prediction

Method	*Performance*	*Limitation*
A new model ShockSurv was built based on the XGBoost method to predict mortality in patients with septic shock (Zheng et al., 2023)	AUROC: 86.15% Precision: 71.26% Recall: 40.37% Accuracy: 86.15% F1-Score: 84.55%	a Specialising in septic shock patients. b Subjective feature selection methods.
Use XGBoost to predict in-hospital mortality in patients with severe sepsis (Hu et al., 2022)	AUROC: 88.40% Accuracy: 89.50%	a Incomplete and less reliable model evaluation indicators. b Lack of external validation.
To explore the sepsis prognostic model, the LightGBM approach was used to build the model (Wang et al., 2022)	AUROC: 90.00% Precision: 55.90% Recall: 83.40% Accuracy: 80.80% F1-Score: 66.80%	a The model has low F1 scores and precision values. b Lack of external validation.
Training and comparing eight different machine learning algorithms for predicting the probability of septic shock in patients six hours after hospital admission, RF model has the best performance (Debdipto et al., 2021)	AUROC: 94.83% Sensitivity: 83.92% Specificity: 88.14%	a Lack of key clinical characterisation variables. b Difficulty in processing missing data for laboratory variables. c Incomplete and less reliable model evaluation indicators.
Comparing the performance of machine learning models and traditional CARES models in predicting postoperative mortality, the GB model has the best performance (Chiew et al., 2020)	AUROC: 96.00% Specificity: 98.00% Precision: 20.00% Recall: 50.00% AUPOC: 23.00% F1-Score: 28.00%	a Lack of external validation. b Poor model performance, especially recall and F1 score values. c Filling in missing values with median is not representative.

Looking ahead, the potential clinical applications of MorSNX are far-reaching. This model has the potential to transform sepsis management by improving the accuracy and efficiency of predicting patient outcomes. Its seamless integration into clinical workflows can help healthcare professionals identify high-risk patients early, facilitating timely intervention and personalised treatment strategies. Additionally, MorSNX's ability to provide interpretable results adds a layer of transparency to its predictions, fostering trust among clinicians and supporting collaborative decision-making. The model outlined in this study boasts several notable advantages, including lightweight, fast response speed, and high accuracy. Considering these strengths, there is a compelling case for its integration into clinical application systems to improve the quality and efficiency of medical services and meet people's needs for active health in the modern era.

5 Conclusions

In this study, we introduce MorSNX, a robust ML model designed to predict 30-day mortality risk in ICU patients with sepsis. By integrating ML techniques with clinical data from the MIMIC-IV dataset, MorSNX outperforms existing models and clinical scoring systems in performance across various metrics such as AUROC, recall, specificity, accuracy, and F1 score.

MorSNX did not only provides quantitative metrics but also provides qualitative insights, which are crucial for clinical decision-making. The model elucidates the relative importance of the top 15 variables influencing mortality in sepsis, offering clinicians a nuanced understanding of the factors affecting patient outcomes. These insights can inform tailored clinical interventions and improve care strategies. The design of MorSNX facilitates seamless integration into hospital information systems, enhancing its practicality for real-world clinical use. By equipping medical staff with a powerful predictive tool, MorSNX has the potential to revolutionise the management of sepsis, ultimately elevating patient care and outcomes.

6 Future work

While MorSNX demonstrates robust performance in predicting 30-day mortality risk for ICU patients with sepsis, certain limitations remain that warrant further exploration. One of the key challenges encountered in this study was the discrepancy in performance between the MIMIC-IV training dataset and the external validation on the eICU dataset. Despite the introduction of L1 and L2 regularisation and the use of cross-validation, improvements in metrics such as accuracy and AUC on the eICU dataset were relatively modest. This suggests that data heterogeneity, including differences in feature availability, data distribution, and recording methods between datasets, continues to impact the model's generalisation capabilities. To address these limitations, future work will focus on the following areas:

1 *Improving generalisation:* We will explore transfer learning and multi-task learning to help the model better adapt to different ICU datasets, enhancing its ability to generalise to diverse patient populations.

2 *Expanding external validation:* Incorporating additional external ICU datasets will allow for a more thorough evaluation of the model's robustness and applicability across different clinical settings.

3 *Optimising feature selection:* We will refine feature engineering and data preprocessing strategies to address inconsistencies in feature availability and data quality, improving model performance in heterogeneous environments.

4 *Real-world integration:* Future efforts will focus on deploying MorSNX in real-time clinical settings, assessing its practical impact on decision-making and patient outcomes in various ICU environments.

By addressing these areas, we aim to further enhance MorSNX's scalability and effectiveness in improving sepsis management.

Appendices/Supplementary materials are available on request by emailing the corresponding author or can be obtained under https://github.com/wuruiqian /Supplementary-material.

Acknowledgements

The work is supported by the National Natural Science Foundation of China [Grant Number: 82072228, 62376152]; the Fault Diagnosis Research in Uncertain Environment under the Background of Industry 4.0 [Grant Number: KJ2021A0866]; the Foundation of the Program of Shanghai Academic/Technology Research Leader under the Science and Technology Innovation Action Plan [Grant Number: 22XD1401300].

References

Baptista, M.L., Goebel, K. and Henriques, E.M.P. (2022) 'Relation between prognostics predictor evaluation metrics and local interpretability SHAP values', *Artificial Intelligence*, Vol. 306, p.103667, DOI: 10.1016/j.artint.2022.103667.

Bauer, M., Gerlach, H., Vogelmann, T., Preissing, F., Stiefel, J. and Adam, D. (2020) 'Mortality in sepsis and septic shock in Europe, North America and Australia between 2009 and 2019 – results from a systematic review and meta-analysis', *Critical Care*, Vol. 24, No. 1, p.239, DOI: 10.1186/s13054-020-02950-2.

Cecconi, M., Evans, L., Levy, M. and Rhodes, A. (2018) 'Sepsis and septic shock', *The Lancet*, Vol. 392, No. 10141, pp.75–87, DOI: 10.1016/S0140-6736(18)30696-2.

Chen, T. and Guestrin, C. (2016) 'XGBoost: a scalable tree boosting system', *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, San Francisco California, USA, pp.785–794, DOI: 10.1145/2939672.2939785.

Chiew, C.J., Liu, N., Wong, T.H., Sim, Y.E. and Abdullah, H.R. (2020) 'Utilizing machine learning methods for preoperative prediction of postsurgical mortality and intensive care unit admission', *Annals of Surgery*, Vol. 272, No. 6, pp.1133–1139, DOI: 10.1097 /SLA.0000000000003297.

Costa, N., da Costa Sigrist, G., Schalch, A.S., Belotti, L., Dolhnikoff, M. and da Silva, L.F.F. (2024) 'Lung tissue expression of epithelial injury markers is associated with acute lung injury severity but does not discriminate sepsis from ARDS', *Respiratory Research*, Vol. 25, No. 1, DOI: 10.1186/s12931-024-02761-x.

Debdipto, M., Venkatesh, A., Wolk, D.M. et al. (2021) 'Early detection of septic shock onset using interpretable machine learners.' *Journal of Clinical Medicine*, Vol. 10, No. 2, pp.301–301, DOI: 10.3390/jcm10020301.

Deutschman, C.S. (2016) 'Imprecise MEDICINE: THE LIMITATIONS OF Sepsis-3', *Critical Care Medicine*, Vol. 44, No. 5, pp.857–858, DOI: 10.1097/CCM.0000000000001834.

Divina, F., Gilson, A., Goméz-Vela, F., García Torres, M. and Torres, J. (2018) 'Stacking ensemble learning for short-term electricity consumption forecasting', *Energies*, Vol. 11, No. 4, p.949, DOI: 10.3390/en11040949.

Fleuren, L.M., Klausch, T.L.T., Zwager, C.L., Schoonmade, L.J., Guo, T., Roggeveen, L.F., Swart, E.L., Girbes, A.R.J., Thoral, P., Ercole, A., Hoogendoorn, M. and Elbers, P.W.G. (2020) 'Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy', *Intensive Care Medicine*, Vol. 46, No. 3, pp.383–400, DOI: 10.1007/s00134-019-05872-y.

Goldberger, A.L. et al. (2000) 'PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals', *Circulation*, Vol. 101, No. 23, DOI: 10.1161/01.CIR.101.23.e215.

Hu, C., Li, L., Huang, W., Wu, T., Xu, Q., Liu, J. and Hu, B. (2022) 'Interpretable machine learning for early prediction of prognosis in sepsis: a discovery and validation study', *Infectious Diseases and Therapy*, Vol. 11, No. 3, pp.1117–1132, DOI: 10.1007/s40121-022-00628-6.

Johnson, A.E.W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T.J., Hao, S., Moody, B., Gow, B., Lehman, L.H., Celi, L.A. and Mark, R.G. (2023) 'MIMIC-IV, a freely accessible electronic health record dataset', *Scientific Data*, Vol. 10, No. 1, p.1, DOI: 10.1038/s41597-022-01899-x.

Leedahl, D.D., Frazee, E.N., Schramm, G.E., Dierkhising, R.A., Bergstralh, E.J., Chawla, L.S. and Kashani, K.B. (2014) 'Derivation of urine output thresholds that identify a very high risk of AKI in patients with septic shock', *Clinical Journal of the American Society of Nephrology*, Vol. 9, No. 7, pp.1168–1174, DOI:10.2215/CJN.09360913.

Liu, D.H., Wu, N., Yan, Q. and Li, Y. (2024) 'Model-driven IEP-GNN framework for MIMO detection with Bayesian optimization', *IEEE Wireless Communications Letters*, Vol. 13, No. 2, pp.387–391, DOI: 10.1109/lwc.2023.3329876.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. and Lee, S-I. (2020) 'From local explanations to global understanding with explainable AI for trees', *Nature Machine Intelligence*, Vol. 2, No. 1, pp.56–67, DOI: 10.1038/s42256-019-0138-9.

Pollard, T.J., Johnson, A.E.W., Raffa, J. and Badawi, O. (2017) *The eICU Collaborative Research Database*, DOI: doi.org/10.13026/C2WM1R.

Rasheed, K., Qayyum, A., Ghaly, M., Al-Fuqaha, A., Razi, A. and Qadir, J. (2022) 'Explainable, trustworthy, and ethical machine learning for healthcare: a survey', *Computers in Biology and Medicine*, Vol. 149, p.106043, DOI: 10.1016/j.compbiomed.2022.106043.

Ren, C., Yao, R., Zhang, H., Feng, Y. and Yao, Y. (2020) 'Sepsis-associated encephalopathy: a vicious cycle of immunosuppression', *Journal of Neuroinflammation*, Vol. 17, No. 1, p.14, DOI: 10.1186/s12974-020-1701-3.

Sabut, S., Patra, P. and Ray, A. (2022) 'Deep learning approach for classifying ischemic stroke using DWI sequences of brain MRIs', *International Journal of Intelligent Systems Technologies and Applications*, Vol. 20, No. 6, p.524, DOI:10.1504/IJISTA.2022.128526.

Schvetz, M., Fuchs, L., Novack, V. and Moskovitch, R. (2021) 'Outcomes prediction in longitudinal data: study designs evaluation, use case in ICU acquired sepsis', *Journal of Biomedical Informatics*, Vol. 117, p.103734, DOI: 10.1016/j.jbi.2021.103734.

Singer, M., Deutschman, C.S., Seymour, C.W. et al. (2016) 'The third international consensus definitions for sepsis and septic shock (Sepsis-3)', *JAMA*, Vol. 315, No. 8, p.801, DOI: 10.1001/jama.2016.0287.

Sujan, R.A., Akashdeep, S., Harshvardhan, R. and Sowmya, K.S. (2022) 'Stacking deep learning and machine learning models for short-term energy consumption forecasting', *Advanced Engineering Informatics*, Vol. 52, p.101542, DOI: 10.1016/j.aei.2022.101542.

Wang, Z., Lan, Y., Xu, Z., Gu, Y. and Li, J. (2022) 'Comparison of mortality predictive models of sepsis patients based on machine learning', *Chinese Medical Sciences Journal*, Vol. 37, No. 3, p.201, DOI: 10.24920/004102.

Wang, Z., Wang, C., Peng, L. et al. (2024) 'Radiomic and deep learning analysis of dermoscopic images for skin lesion pattern decoding', *Sci Rep.*, Vol. 14, p.19781, DOI: 10.1038/s41598-024-70231-x.

Zheng, F., Wang, L., Pang, Y., Chen, Z., Lu, Y., Yang, Y. and Wu, J. (2023) 'ShockSurv: a machine learning model to accurately predict 28-day mortality for septic shock patients in the intensive care unit', *Biomedical Signal Processing and Control*, Vol. 86, p.105146, DOI: 10.1016/j.bspc.2023.105146.