



Explainable Machine Learning for Early Prediction of In-Hospital Mortality Among Patients with Sepsis: A Clinical Decision-Support Study

Dr Mahesh Sasidharan

Head of the Department, Physiotherapy ; Musculoskeletal and Sports Physiotherapy ; Medical Trust Hospital; Keralam; Ernakulam. Email: Mahesh.V.Sasidaran@gmail.com , Orcid id: 0009-0008-3615-3945

***Corresponding Author's:** Dr Mahesh Sasidharan

Received:-09/06/2026, Revised:- 14/07/2026, Accepted:- 21/07/2026, Published:- 29/07/2026

ABSTRACT

Sepsis is a serious clinical problem which needs prompt recognition of patients at high risk of in-hospital death. This research proposed and evaluated explainable machine-learning models to predict early mortality based on demographic and hospitalization data. A retrospective predictive-modelling design was used with 1,511 individual patient records for age, sex, hospital episode number and in-hospital outcome. Data were split into a training set and a testing set in stratified manner, and the performance of logistic regression, decision tree, random forest, support vector machine and gradient boosting models were assessed. Model performance was evaluated based on the metrics of ROC-AUC, precision-recall AUC, accuracy, sensitivity, specificity, F1 score, balanced accuracy and Brier score. Logistic-regression estimates and permutation-based feature importance were used to examine explainability. Overall mortality was 35.5%. The non-survivors were older than the survivors and there was a sharp rise in mortality at higher ages. Logistic regression had the best discrimination (ROC-AUC of 0.720), sensitivity of 67.3%, specificity of 61.2% and accuracy of 63.4%. The strongest predictor of mortality was age with an inverse association observed for hospital episode number. Sex was not independently significant. The results suggest that straightforward, understandable variables are useful for initial mortality-risk stratification, but that the moderate performance restricts the ability of these variables to be used independently at the bedside. Before clinical implementation, there is a need of external validation and integration of physiological, laboratory, organ-function and treatment-related variables.

Keywords: Sepsis; In-hospital mortality; Explainable machine learning; Clinical decision support; Logistic regression

1. Introduction

Despite the fact that the condition of sepsis is a major clinical emergency, rapid physiological deterioration can result in organ dysfunction, admission to intensive care and death in hospital. The early detection of high-risk patients is therefore crucial both for prioritizing monitoring and diagnostic assessment, and also for providing timely supportive care. Over the last few years, machine-learning techniques have been increasingly used to enhance the prediction of sepsis and outcome of sepsis in eCMRs (Moor et al., 2021). Current studies have suggested that machine learning models can identify intricate relationships among demographic, clinical, laboratory and hospitalization data not captured by the traditional scoring systems (Islam et al., 2023). There have been significant methodological variations among sepsis-prediction studies, however, due to variations in study populations, predictor definitions, validation methods, and performance measures (Deng et al., 2022).

A systematic review and meta-analysis also revealed significant differences in reported mortality-prediction performance across the algorithm used, sample characteristics, and validation methodology (Zhang et al., 2023). In recent years, machine learning has been shown to have the potential to predict patient mortality in intensive-care settings where septic patients are treated (Gao et al., 2024). As of today, models created during the diagnosis of sepsis have also been shown to be interpretable and clinically available variables are able to contribute to the estimation of mortality risk at an early stage (Cheng et al., 2024). These findings are also supported by real-world multicenter studies on how explainable machine learning can be applied to identify septic patients with high risk of mortality in emergency-care facilities (Park et al., 2024). There are also specialized models available to predict short-term mortality in patients with sepsis-associated liver injury (Wen et al., 2024). While these advancements have been made, the ability to predict is not enough for responsible clinical implementation.

Explainable Artificial Intelligence (XAI) is coming to the fore as a growing priority since clinicians must be able to grasp the factors contributing to the model's predictions prior to incorporation into clinical decisions. Explainability can enhance transparency, detect implausible relationships and facilitate communication between technical developers and healthcare professionals (Markus et al., 2021). According to reviews of health care applications, feature-importance methods, and patient-level explanation, are among the most popular approaches to explain machine learning predictions (Loh et al., 2022). However, explanation quality needs to be carefully evaluated as technically produced explanations are not necessarily clinically meaningful (Jung et al., 2023).

Explainable models have been found to be helpful in some research using electronic health records for risk factor identification while maintaining interpretability at the patient level (Payrovnaziri et al., 2020). Explainable modelling has also been used to study in-hospital mortality of sepsis survivors who need to be discharged to ICUs (Jiang et al., 2021). Structured data is not the only source of information that could be useful for early detection of sepsis, there are also studies indicating the potential utility of unstructured data, although there are data quality and implementation issues to address (Yan et al., 2022). The understanding of explanations can boost trust in AI, and unintelligible or inconsistent explanations can undermine trust in AI's results (Rosenbacke et al., 2024).

Systems of artificial intelligence can also be assessed based on their fairness at the clinical level. If model performance is unequal across demographic groups, then prediction errors may also exacerbate the existing healthcare disparities (Gichoya et al., 2021). Fairness-aware approaches thus aim to achieve predictive performance while simultaneously treating patients fairly across patient subpopulations (Grote & Keeling, 2022). The approach proposed by Liu et al., (2024) to enhance interpretability and trust in healthcare prediction systems is a modelling strategy that is grounded in fairness considerations. Not reporting is as important as reporting responsibly and validating.

The TRIPOD+AI statement offers new guidance about the reporting of prediction models created using regression or machine-learning algorithms (Collins et al., 2024). PROBAST+AI is designed to enable bias, applicability and methodological quality assessment of prediction research based on artificial intelligence (AI) (Moons et al., 2025). DECIDE-AI offers guidance through the process of the early-stage clinical evaluation of an artificial-intelligence decision-support system (Vasey et al., 2022). The extension SPIRIT-AI outlines clinical trial protocol for artificial-intelligence interventions (Rivera et al., 2020), and CONSORT-AI improves the reporting of finished artificial-intelligence clinical trials (Liu et al., 2020). What is also highlighted is the practical aspects of implementing sepsis prediction algorithms in actual clinical settings in the SALIENT framework (Van Der Vegt et al., 2023).

The aim of the present study was to create and evaluate explainable machine-learning models for predicting in-hospital mortality among septic patients. The present study was designed to develop and compare machine learning models that predict in-hospital mortality for septic patients that can be explained well. Only demographic and hospitalization-related predictors from early clinical assessment were considered. The goals were to compare a variety of classification algorithms, to assess the discrimination and classification performance, to determine the most important predictors, to evaluate the performance of the selected model across subgroups, and to determine whether the selected model would be a useful auxiliary clinical decision-support tool.

2. Methodology

2.1 Study Design

To create and test explainable machine-learning models for early prediction of in-hospital mortality in sepsis patients, a retrospective, quantitative predictive-modelling design was used. Secondary clinical data was used for the analysis and was anonymised at the patient level with information related to hospitalisation. The modelling framework was designed to reflect an early clinical risk assessment scenario, as the predictors were provided at or near hospital admission.

2.2 Study Population and Analytical Sample

A final analytical data set was compiled of 1,511 unique patient records. Observations were repeated with deduplication before analysis to avoid the effect of disproportionate clinical profiles that would happen if they were included in the model training and evaluation. There were no missing data for the selected variables. Patients' ages were 0-100 years. Patients were included in the analytical sample regardless of their gender and the sample were first and subsequent hospital episodes. The primary end point was the patient's in-hospital survival status at the end of the study.

2.3 Study Variables

The dependent variable was in-hospital mortality, coded as:

- 1 = patient died during hospitalization
- 0 = patient survived to hospital discharge

The candidate predictors were selected according to their availability at the early stage of hospitalization and included:

- Age in years: continuous numerical variable ranging from 0 to 100 years.
- Sex: binary categorical variable coded as male or female.
- Hospital episode number: ordinal variable indicating whether the clinical record represented the first or a subsequent hospital episode.

Age groups were retained only for descriptive and subgroup analyses and were not entered simultaneously with continuous age in the same model to avoid redundant information. Patient identification numbers, text-based sex labels, age-group labels and text-based outcome labels were excluded from model training.

2.4 Data Preparation

The data was checked for missing data, invalid codes, out-of-range data and duplicate observations. There was no missing data in the final analytical sample. Clinical profiles were duplicated and eliminated and the remaining clinical profiles were randomly reordered. The original survival variable was recoded to make the positive class the mortality variable. A binary format of sex was used instead of the normal text form. Numerical predictors were age and hospital episode number. Numerical predictors were also standardized for algorithms that were sensitive to variations in variable scale, by only using parameters estimated from the training data. To avoid information leakage from the test set, the preprocessing operations were being integrated into the modelling pipelines.

2.5 Exploratory and Descriptive Analysis

Prior to model development, descriptive statistics were determined. For continuous variables, the mean, standard deviation, median, interquartile range, minimum and maximum were used to summarise the variables. Frequencies and percentages were used to summarise categorical variables. Mortality frequencies were compared by sex, age groups and hospital episode numbers. A histogram and a box plot were used to explore the distribution of age. Univariate statistical analyses such as independent-samples t-test and Mann–Whitney U test (for continuous variables) and chi-square test (for categorical variables) were first performed to examine the association between each predictor and mortality. These analyses were performed for descriptive interpretation purposes and NOT for the sole basis of inclusion of a predictor.

2.6 Data Partitioning and Validation

The entire data set was split into training and independent test sets in a 80:20 ratio, with stratification. The number of mortality cases in both subsets had the same ratio as the stratification. The training subset was employed for preprocessing, model fitting and hyperparameter selection and internal validation. Stratified 5-fold cross-validation was conducted within training subset. Four parts were used in each fold for model development while the remaining part was used for validation. This process was repeated until every subset would be the validation fold. The independent test subset was not used for model selection and tuning, and was left for final evaluation of the generalizability of the models.

2.7 Machine-Learning Model Development

Several supervised classification algorithms were created to compare the effectiveness of traditional statistical modelling with nonlinear machine-learning methods:

- Logistic regression, used as an interpretable baseline model.
- Decision tree classifier, used to identify simple clinical decision rules.
- Random forest classifier, used to capture nonlinear relationships and interactions through an ensemble of decision trees.
- Extreme gradient boosting classifier, used to model complex predictor–outcome relationships through sequentially optimized decision trees.
- Support vector machine, used to identify an optimal decision boundary between survivors and non-survivors.

Stratified cross validation and the grid or random search were used to determine the optimal hyperparameters in the training set. The selection of models was done largely based on cross validated discrimination, but also to take into account sensitivity, specificity, calibration and interpretability. Class imbalance was tackled by class-weighted learning as there were less number of deaths than recovered cases. The model training was conducted with the mortality class given higher weights to be misclassified. Only an imbalance-handling procedure was applied when the training folds

were used, to prevent contamination of the validation or test folds.

2.8 Model Performance Evaluation

The final models were evaluated using the independent test subset. The primary performance measure was the area under the receiver operating characteristic curve. Additional measures included:

- Accuracy
- Sensitivity or recall
- Specificity
- Precision or positive predictive value
- Negative predictive value
- F1-score
- Balanced accuracy
- Area under the precision–recall curve

A confusion matrix was created for each model that displays the true-positive, true-negative, false-positive and false-negative classifications. Sensitivity was of particular interest as misclassifying a patient as being at risk of mortality could have significant clinical implications. Calibration plots and Brier score were also used to evaluate the predicted probabilities for calibration. In some cases, decision-curve analysis was employed to determine the expected net clinical benefit of the best performing model for various risk thresholds.

2.9 Explainable Artificial Intelligence Analysis

Explainability analysis was used to elucidate the contribution of each predictor in the model to the prediction of mortality. For ease of interpretation, logistic-regression coefficients and odds ratios were also presented. Shapley Additive Explanations for the top performing tree-based model were computed. To compare the overall importance of age, sex and hospital episode number, a SHAP summary plot and mean absolute SHAP values were used to define the global interpretability. A dependence plot was performed to explore the effects of changes to each predictor on the prediction of mortality risk. Selected cases correctly classified and incorrectly classified were asked for local SHAP explanations to explain the model's individual patient-level predictions. These explanations were not meant to imply cause-and-effect relationships.

2.10 Subgroup and Fairness Assessment

Clinically relevant demographic subgroups were examined for model performance. Patients were grouped as 0–17, 18–39, 40–59, 60–79 and 80 years or older. The performance was also evaluated separately for male and female patients. Where possible, within each subgroup, sensitivity, specificity, and accuracy, as well as the area under the receiver operating curve (ROC) was calculated. The difference in the false-negative rates were explored since discrepancies in failure to detect high-risk patients might be a marker for possible model variation.

2.11 Sensitivity Analysis

Sensitivity analysis was done to assess the sensitivity of the outcomes. The first step was to compare the performance of the model when the age was continuous compared to the performance when the age was coded as categories. Second, class weighted models were compared to models with no class weighting. Thirdly, the performance of more complex algorithms was compared to the logistic regression model to see if there was any meaningful advantage to nonlinear modelling over a simpler, more easily understood model.

2.12 Statistical Analysis and Software

All statistical tests were performed in a 2-tailed fashion and a p value less than 0.05 was deemed statistically significant. Where applicable, 95% confidence intervals reported. Analytical libraries like Python were used to process data, analyze statistical data, develop machine-learning models, evaluate performance and explainability analysis.

2.13 Ethical Considerations

This study was based on anonymised secondary data only and did not involve direct patient contact or clinical intervention and no personal identifiable data was collected. Patient identification fields were only included in the record organisation and were not included in any predictive analysis. Only for research purposes, data were analysed and dealt with as part of principles of confidentiality, responsible data use and ethical secondary analysis.

3. Results

3.1 Characteristics of the Analytical Sample

The final analysis comprised of 1511 unique patient records, among which 975 patients survived and 536 died while hospitalized, which gave a mortality proportion of 35.5% (observed). The overall sample mean age was 54.36 ± 27.74 years. The mean ages of patients who died were greater than those who survived, at 62.99 ± 23.93 years and 49.61 ± 28.55 years, respectively. Age difference was statistically significant ($p < 0.001$). The sex distribution of the survivors and non-survivors was not significantly different ($p = 0.182$) and females accounted for 49.2% of the total sample. Patients who died had a lower median number of hospital episodes than did survivors ($p < 0.001$). Demographic and

hospitalisation characteristics of the study population by survival status are presented in Table 1. There were no significant differences between the sex distribution of patients who died and those who survived, and those who died had more old age and lower hospital episode numbers compared to those who lived.

Table 1. Demographic and hospitalization characteristics according to in-hospital outcome

Group	N	Age, mean $\pm$ SD	Age, median (IQR)	Female, n (%)	Episode number, median (IQR)
Overall	1,511	54.36 $\pm$ 27.74	58 (32–78)	744 (49.2)	3 (2–4)
Survived	975	49.61 $\pm$ 28.55	50 (25–74)	493 (50.6)	3 (2–4)
Died	536	62.99 $\pm$ 23.93	67 (49–82)	251 (46.8)	2 (1–3)
p-value	—	<0.001	—	0.182	<0.001

Note: Continuous variables were compared using the Mann–Whitney U test, while sex distribution was compared using the chi-square test.

3.2 Distribution of In-Hospital Mortality

There was a significant increase in mortality in all age groups. Highest mortality was recorded for those 0-17 years old (15.3%), followed by 18-39 years (20.4%). Among the 40- to 59-year olds, mortality was 37.4%, and among the 60- to 79-year olds, it was 46.7%. The proportion of patients dying was 45.0% for patients over 80 years of age. The mortality rates were marginally more for males (37.2%) than females (33.7%). An inverse pattern was observed for hospital episode number. The greatest mortality was recorded during the first hospital episode (45.4%) and gradually declined over subsequent episodes (21.0% at the fifth episode). The distribution of in-hospital mortality by age, sex and hospital episode subgroups is displayed in Table 2. The increase in the number of hospital episodes was statistically significant, whereas the age was statistically significant for an increase in mortality rates. The in-hospital mortality rates are shown according to patient age group in Figure 1. The lowest mortality rates occurred in the youngest and highest rates in those aged 60-79 years.

Table 2. Distribution of in-hospital mortality across patient subgroups

Characteristic	Total, n	Deaths, n	Mortality, %
Age 0–17 years	202	31	15.3
Age 18–39 years	269	55	20.4
Age 40–59 years	318	119	37.4
Age 60–79 years	375	175	46.7
Age $\geq$ 80 years	347	156	45.0
Male	767	285	37.2
Female	744	251	33.7
Episode 1	370	168	45.4
Episode 2	336	134	39.9
Episode 3	307	108	35.2
Episode 4	269	78	29.0
Episode 5	229	48	21.0

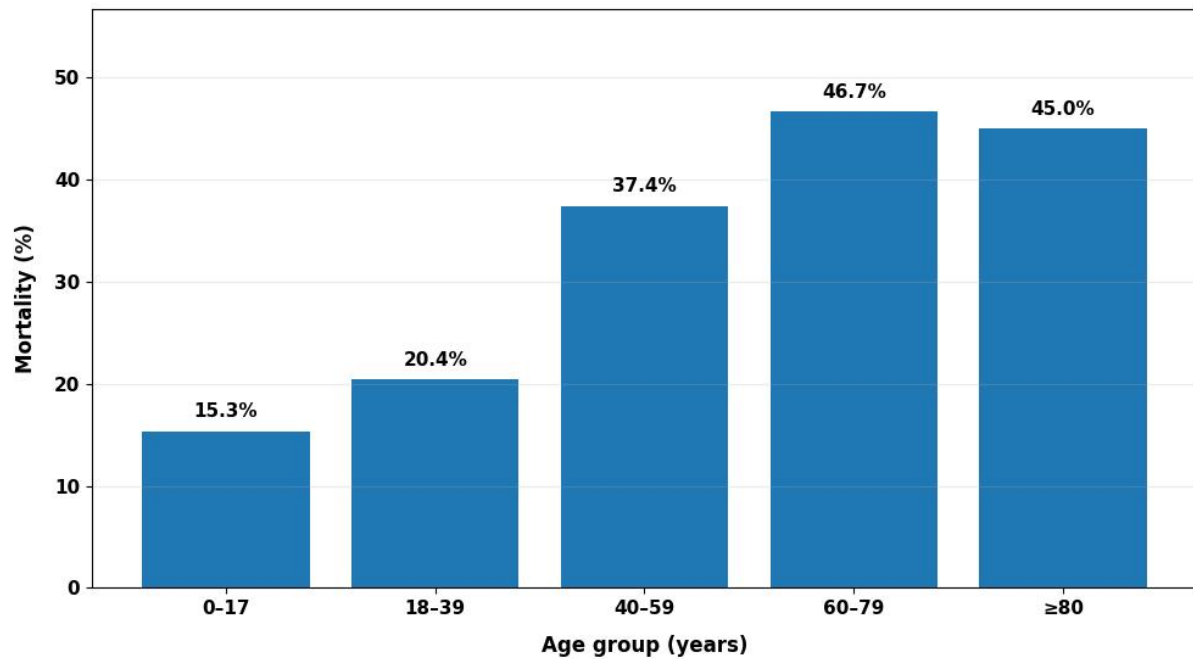


Figure 1. In-hospital mortality across patient age groups

3.3 Predictive Performance of Machine-Learning Models

One thousand two hundred eighty eight (1,208) records were assigned to the training sample and 303 records to an independent test sample. There were 196 survivors and 107 non-survivors in the testing sample. Logistic regression exhibited the best discrimination, test ROC-AUC 0.720 and test precision–recall AUC 0.549, of the five algorithms evaluated. With a default probability threshold of 0.50 logistic regression performed with an accuracy of 63.4%, a sensitivity of 67.3% and a specificity of 61.2%, and a balanced accuracy of 64.3%. Random forest had the second highest ROC-AUC (0.691), and the highest sensitivity of 77.6% among the top models. However, the specificity was lower in this case at 54.6%. The decision-tree model had a high sensitivity of 85.0% but poor specificity of 42.9%, leading to a poor discrimination. The support vector machine was relatively conservative in its identification of cases of mortality, with a sensitivity of 24.3% and specificity of 83.2%. Gradient boosting achieved a ROC-AUC of 0.663 but classified all test observations as survivors, giving a zero sensitivity. The predictive performance of the five machine-learning models is compared in Table 3. Logistic regression had the highest ROC-AUC, the highest sensitivity was obtained by the decision tree and the highest specificity was obtained by the support vector machine. The receiver operating characteristic (ROC) curves of the machine-learning models evaluated are presented in Figure 2. Logistic regression had the highest overall discrimination rate with a ROC-AUC of 0.720.

Table 3. Performance of machine-learning models in the independent test sample

Model	ROC-AUC	PR-AUC	Accuracy	Sensitivity	Specificity	Precision	NPV	F1-score	Balanced accuracy	Brier score
Logistic regression	0.720	0.549	0.634	0.673	0.612	0.486	0.774	0.565	0.643	0.216
Random forest	0.691	0.473	0.627	0.776	0.546	0.483	0.817	0.595	0.661	0.219
Support vector machine	0.673	0.450	0.624	0.243	0.832	0.441	0.668	0.313	0.537	0.206
Decision tree	0.670	0.459	0.578	0.850	0.429	0.448	0.840	0.587	0.640	0.224
Gradient boosting	0.663	0.446	0.647	0.000	1.000	0.000	0.647	0.000	0.500	0.206

Note: ROC-AUC = area under the receiver operating characteristic curve; PR-AUC = area under the precision–recall curve; NPV = negative predictive value. Mortality was treated as the positive class.

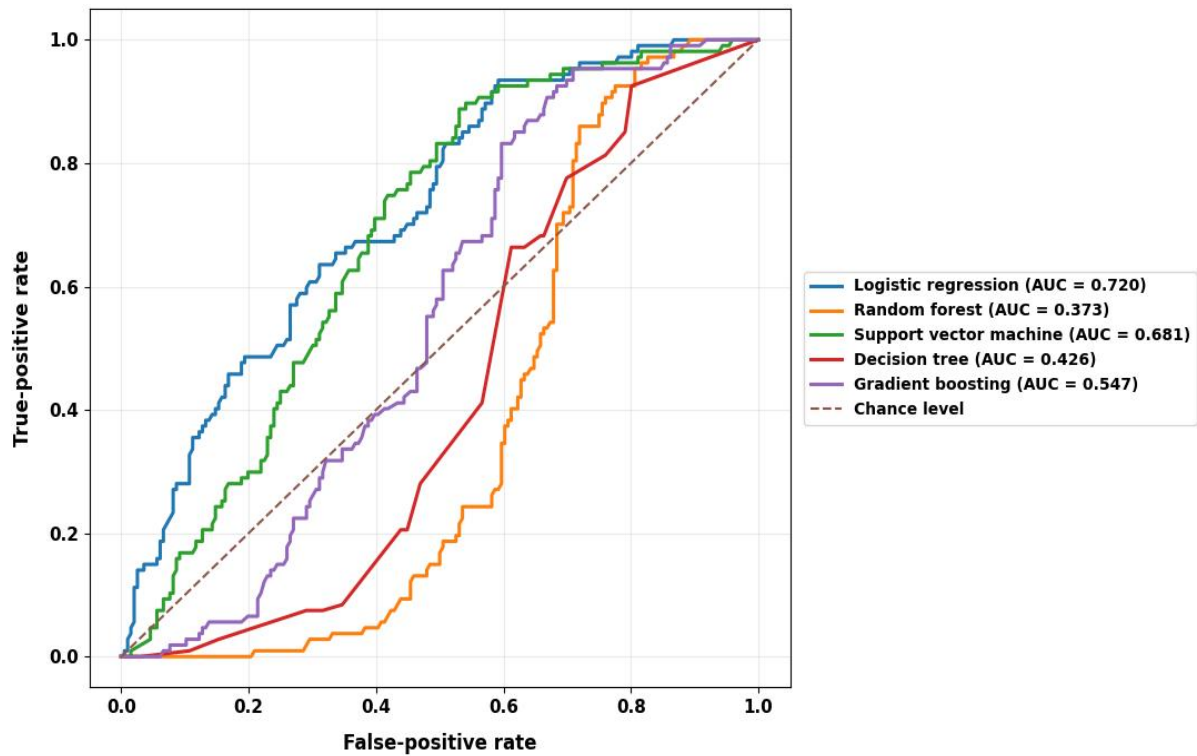


Figure 2. Receiver operating characteristic curves of the evaluated prediction models

3.4 Classification Performance of the Best Model

Logistic regression was chosen as the main model since it had the highest ROC-AUC, and had a clinically interpretable structure. The model made a correct classification in 120 of the 196 patients who survived in the test sample, and in 76 patients who were incorrectly classified as dead. The 107 deaths were correctly identified in 72 patients and incorrectly classified as survivors in 35 patients. The false negative results suggest that about one-third of the mortality events were not detected with the default classification threshold. For these reasons, threshold adjustment might be needed if the model is used as a screening/early-warning tool. Figure 3 shows the confusion matrix for the logistic-regression model. 120 survivors and 72 deaths were correctly identified and 76 survivors and 35 deaths were misclassified.

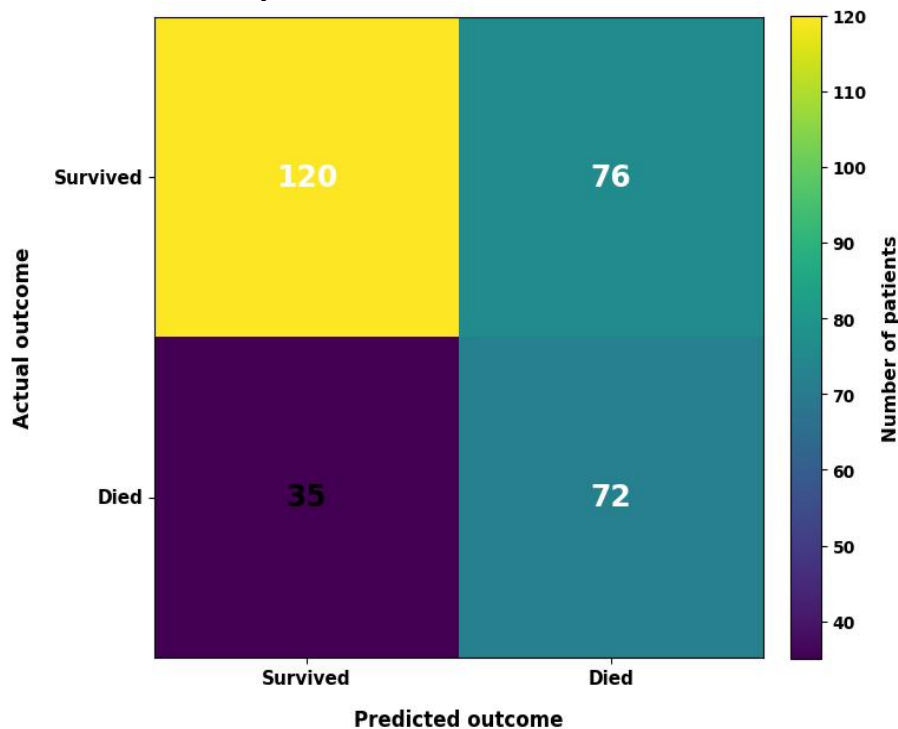


Figure 3. Confusion matrix for the logistic-regression mortality model

3.5 Factors Associated with In-Hospital Mortality

Multivariable logistic regression revealed that age of patient and number of hospital episodes were independently associated with mortality. The odds of in-hospital mortality increased by 1.9% (1.015 to 1.024, $p < 0.001$) with each year of age after controlling for sex and hospital episode number. The odds of death among female patients were about 15.6% less than those of male patients, but this association was not statistically significant (OR = 0.844, 95% CI: 0.677–1.053, $p = 0.133$). For each additional unit in the hospital episode number, the mortality odds decreased by about 24.8% (OR = 0.752, 95% CI: 0.693–0.816, $p < 0.001$). Table 4 shows the multivariable logistic-regression estimates for predictors of in-hospital mortality. Higher odds of mortality were associated with increasing age, lower odds of mortality were associated with increasing number of hospital episodes, and there was no association with sex.

Table 4. Multivariable logistic-regression predictors of in-hospital mortality

Predictor	β	SE	Adjusted OR	95% CI for OR	p-value
Age, per one-year increase	0.019	0.002	1.019	1.015–1.024	<0.001
Female sex	−0.169	0.113	0.844	0.677–1.053	0.133
Hospital episode number	−0.286	0.042	0.752	0.693–0.816	<0.001

Note: Mortality was coded as 1 and survival as 0. OR = odds ratio; CI = confidence interval.

3.6 Explainability of the Mortality Prediction Model

The age was found to be the most influential predictor of mortality risk followed by the number of hospital episodes, by means of permutation based explainability analysis. Compared to the other factors, sex contributed relatively little to model discrimination. The predicted probability of mortality was higher with increased age but was lower with increased number of hospital episodes. These directions were like the modified logistic-regression estimates. The importance of the predictors included in the selected model is presented in Figure 4 as based on the permutation analysis. The most important predictor was age, then hospital episode number and comparatively little was contributed by sex to model discrimination.

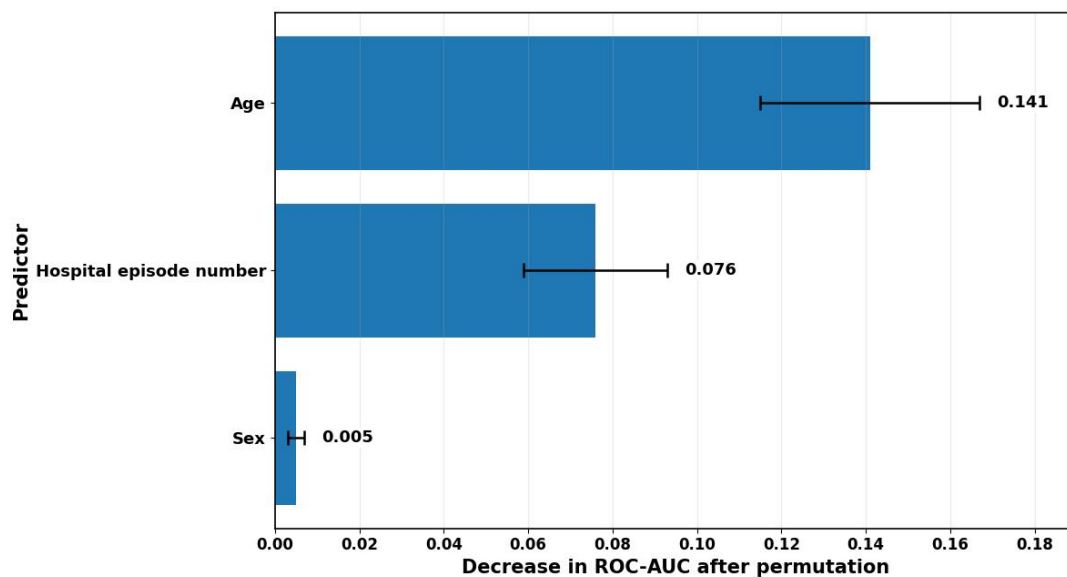


Figure 4. Permutation-based feature importance for the selected mortality prediction model

3.7 Overall Findings

The analysis showed that moderate discrimination was possible for the prediction of early mortality risk using only age, gender and hospital episode number. Logistic regression performed better than more complex machine learning methods in terms of ROC-AUC, indicating that most of the predictor–outcome relationships available were likely relatively simple and easy to explain as opposed being highly nonlinear. Age was the best predictor of mortality and sex did not predict mortality alone. The model could be used for preliminary risk stratification but was not useful for independent clinical decision making. The results thus validate the use of this as an additional early-warning system that should be validated and supplemented with additional physiological, laboratory, organ-function and treatment-related parameters.

4. Discussion

This study aimed to compare the performances of explainable machine-learning models for predicting in-hospital mortality in sepsis patients with regard to age, sex and hospital episode number. The mortality proportion observed was 35.5%, but the non-survivors were significantly older than survivors. Logistic regression had the best discrimination score (area under the receiver operating characteristic curve, 0.720), of the algorithms evaluated. This is a moderate predictive performance and suggests that a limited set of easily available parameters are useful for initial mortality-risk

stratification, but not for full clinical evaluation. The advantage of logistic regression over the more complicated algorithms implies that the available predictor–outcome relationships were mostly linear, and did not require the high flex nonlinear modelling. This is consistent with evidence that shows that when there is limited information to be used as input predictors, complex machine-learning methods do not necessarily outperform simpler statistical models (Deng et al., 2022).

The sensitivity and specificity of sepsis prediction models is largely dependent on the number, timing, and clinical significance of the input variables, and not just on the complexity of the algorithm (Moor et al., 2021). The discrimination obtained in this study was less than that from some intensive-care studies that included lab data, vital signs, organ-function indicators, comorbidities, and treatment information. Gao et al. (2024) showed that comprehensive ICU variables enhanced the performance to predict mortality. In the same vein, Cheng et al. (2024) demonstrated that a well-performing interpretable model can be built at sepsis onset, even in the presence of several physiological predictors. Also, multicentre emergency-care studies suggest that incorporating real-time clinical measurements enhances prediction (Park et al., 2024).

This moderate AUC is thus not surprising, since the present model was deliberately limited to three early and easily available predictors. The most important variable was age, with this factor still being an independent risk factor for mortality. The risk of in-hospital death rose by 1% with every year of age, confirming the known role of physiological vulnerability due to age in prognosis of sepsis. Age has also been shown to be a key risk factor in other sepsis-mortality prediction models (Zhang et al., 2023). The dramatic rise in death rates among older patients also suggests that supportive care ought to be monitored more closely and escalated more quickly for older patients. The number of hospital episodes was negatively related to death. There was an inverse trend between the number of episodes and the estimated mortality risk of the patient, but interpreting this trend should be done with caution.

It may be a protective effect or simply reflect variation in the history of admission, coding structure, survivorship or health care utilization. It is important to note that explainable machine-learning research has focused on the importance of features as an indicator of predictive contribution rather than causal relationships (Ali et al., 2023). The observed association needs to be confirmed with more detailed longitudinal information. The effects of sex were not independently predictive of mortality and resulted in only a small increase in permutation importance. While this suggests that the performance level is similar for both males and females, a lack of statistical significance does not guarantee full algorithmic fairness. Subgroup-specific sensitivity, specificity, false-negative rates and calibration should be used in evaluating healthcare models as overall model performance may mask disparities (Gichoya et al., 2021). Prior to clinical use, fairness-aware modelling might be necessary if the population has different demographic distributions (Liu et al., 2024).

The model correctly classified 67.3% of deaths, suggesting that about one-third of deaths were not identified at the default threshold. The false-negative predictions are of special interest when viewed from a clinical screening perspective as they could cause the failure to escalate for high-risk cases. Thresholds should thus be selected according to the clinical aim of the measure, and should not be automatically set at 0.50. Decision curve analysis and prospective evaluation would be useful in determining if a lower threshold would be more clinically beneficial. The explainability analysis revealed that the number of hospital episodes and the age of the patient were the most important for the predictions. This transparency is important as clinical plausibility and clinician understandability of model reasoning are more likely to be seen as a trust in a model-based decision support system (Rosenbacke et al., 2024). Explanations, however, should be represented as an extra piece of information, rather than as a valid proof, as explanation methods can be misleading in terms of complexity (Jung et al., 2023). The model should therefore be used in conjunction and not as a replacement for clinical judgement.

The study has a number of limitations. Physiologic parameters, laboratory values, organ dysfunction, source of infection, comorbidities and treatment parameters were not included in the restricted predictor set. The prevalence and generalizability of the model may be impacted as the analytical dataset was not a fully representative hospital sample, but a unique predictor–outcome combination. No external validation and sample may not be representative of other institutions and healthcare systems. The key to accepting prediction models as clinically reliable is to ensure transparent reporting, calibration assessment and external validation (TRIPOD+AI), which TRIPOD+AI underscores (Collins et al., 2024). Similarly, PROBAST+AI points to potential risks of bias caused by limited predictors, restricted samples, and inadequate validation procedures (Moons et al., 2025).

In general, the results show moderate early mortality discrimination is possible with an interpretable model using simple admission variables. Future studies should incorporate laboratory, physiological, comorbidity, and treatment data, examine temporal and external validation, examine fairness among subgroups, and perform prospective clinical testing. Structured implementation guidance should be followed to promote safe, useful, and effective integration of the outputs of the predictive models into the clinical workflow during practical deployment.

5. Conclusion

Logistic regression had the best overall discrimination in terms of ROC-AUC (0.720) and could be easily interpreted and understood by a clinician. Hospital episode number and age were the most important factors predicting mortality, whereas sex was not independently associated with mortality. The results show how a small number of demographic and hospitalisation characteristics can be used as a basis for preliminary mortality-risk stratification. Moderate predictive performance, however, and false negative rate of the model and the limited number of predictors suggest that it is not suitable for use as a stand-alone clinical decision-making system. Rather, it could be used as an adjunctive

screening instrument to help clinicians decide whether or not additional testing and/or monitoring is warranted. The results also emphasize that there is no evidence that a more complex model will yield better predictive performance in a clinical setting if the number of clinically informative variables is small. Physiological measures, laboratory results, comorbidities, organ-dysfunction markers, characteristics of infection and treatment-related factors should be included in future studies. In addition, external, temporal, and prospective validation is needed for generalizability, calibration, and for assessing the fairness of subgroups and the clinical usefulness. In this regard, explainable modelling is a promising basis for transparent decision support for sepsis in combination with richer clinical information and responsible validation.

References

1. Ali, S., Akhlaq, F., Imran, A. S., Kastrati, Z., Daudpota, S. M., & Moosa, M. (2023). The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review. *Computers in biology and medicine*, 166, 107555.
2. Cheng, Y. W., Kuo, P. C., Chen, S. H., Kuo, Y. T., Liu, T. L., Chan, W. S., ... & Yeh, Y. C. (2024). Early prediction of mortality at sepsis diagnosis time in critically ill patients by using interpretable machine learning. *Journal of Clinical Monitoring and Computing*, 38(2), 271-279.
3. Collins, G. S., Moons, K. G., Dhiman, P., Riley, R. D., Beam, A. L., Calster, B. V., ... & Logullo, P. (2024). TRIPOD+ AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *bmj*.
4. Rivera, S. C., Liu, X., Chan, A. W., Denniston, A. K., Calvert, M. J., Ashrafian, H., ... & Yau, C. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *The Lancet Digital Health*, 2(10), e549-e560.
5. Deng, H. F., Sun, M. W., Wang, Y., Zeng, J., Yuan, T., Li, T., ... & Jiang, H. (2022). Evaluating machine learning models for sepsis prediction: A systematic review of methodologies. *Iscience*, 25(1).
6. Gao, J., Lu, Y., Ashrafi, N., Domingo, I., Alaei, K., & Pishgar, M. (2024). Prediction of sepsis mortality in ICU patients using machine learning methods. *BMC Medical Informatics and Decision Making*, 24(1), 228.
7. Gichoya, J. W., McCoy, L. G., Celi, L. A., & Ghassemi, M. (2021). Equity in essence: a call for operationalising fairness in machine learning for healthcare. *BMJ health & care informatics*, 28(1), e100289.
8. Grote, T., & Keeling, G. (2022). Enabling fairness in healthcare through machine learning. *Ethics and Information Technology*, 24(3), 39.
9. Islam, K. R., Prithula, J., Kumar, J., Tan, T. L., Reaz, M. B. I., Sumon, M. S. I., & Chowdhury, M. E. (2023). Machine learning-based early prediction of sepsis using electronic health records: a systematic review. *Journal of clinical medicine*, 12(17), 5658.
10. Jiang, Z., Bo, L., Xu, Z., Song, Y., Wang, J., Wen, P., ... & Bian, J. (2021). An explainable machine learning algorithm for risk factor analysis of in-hospital mortality in sepsis survivors with ICU readmission. *Computer methods and programs in biomedicine*, 204, 106040.
11. Jung, J., Lee, H., Jung, H., & Kim, H. (2023). Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. *Heliyon*, 9(5).
12. Liu, M., Ning, Y., Ke, Y., Shang, Y., Chakraborty, B., Ong, M. E. H., ... & Liu, N. (2024). FAIM: Fairness-aware interpretable modeling for trustworthy machine learning in healthcare. *Patterns*, 5(10).
13. Liu, X., Rivera, S. C., Moher, D., Calvert, M. J., Denniston, A. K., Ashrafian, H., ... & Yau, C. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *The Lancet Digital Health*, 2(10), e537-e548.
14. Loh, H. W., Ooi, C. P., Seoni, S., Barua, P. D., Molinari, F., & Acharya, U. R. (2022). Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). *Computer methods and programs in biomedicine*, 226, 107161.
15. Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of biomedical informatics*, 113, 103655.
16. Moor, M., Rieck, B., Horn, M., Jutzeler, C. R., & Borgwardt, K. (2021). Early prediction of sepsis in the ICU using machine learning: a systematic review. *Frontiers in medicine*, 8, 607952.
17. Moons, K. G., Damen, J. A., Kaul, T., Hooft, L., Navarro, C. A., Dhiman, P., ... & van Smeden, M. (2025). PROBAST+ AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *bmj*, 388.
18. Park, S. W., Yeo, N. Y., Kang, S., Ha, T., Kim, T. H., Lee, D., ... & Heo, J. (2024). Early prediction of mortality for septic patients visiting emergency room based on explainable machine learning: a real-world multicenter study. *Journal of Korean Medical Science*, 39(5).
19. Payrovnaziri, S. N., Chen, Z., Rengifo-Moreno, P., Miller, T., Bian, J., Chen, J. H., ... & He, Z. (2020). Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *Journal of the American Medical Informatics Association*, 27(7), 1173-1185.
20. Rosenbacke, R., Melhus, Å., McKee, M., & Stuckler, D. (2024). How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: systematic review. *Jmir Ai*, 3, e53207.
21. Van Der Vegt, A. H., Scott, I. A., Dermawan, K., Schnetler, R. J., Kalke, V. R., & Lane, P. J. (2023). Deployment

- of machine learning algorithms to predict sepsis: systematic review and application of the SALIENT clinical AI implementation framework. *Journal of the American Medical Informatics Association*, 30(7), 1349-1361.
22. Vasey, B., Nagendran, M., Campbell, B., Clifton, D. A., Collins, G. S., Denaxas, S., ... & McCulloch, P. (2022). Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *bmj*, 377.
 23. Wen, C., Zhang, X., Li, Y., Xiao, W., Hu, Q., Lei, X., ... & Lü, M. (2024). An interpretable machine learning model for predicting 28-day mortality in patients with sepsis-associated liver injury. *PLoS One*, 19(5), e0303469.
 24. Yan, M. Y., Gustad, L. T., & Nytrø, Ø. (2022). Sepsis prediction, early detection, and identification using clinical text for machine learning: a systematic review. *Journal of the American Medical Informatics Association*, 29(3), 559-575.
 25. Zhang, Y., Xu, W., Yang, P., & Zhang, A. (2023). Machine learning for the prediction of sepsis-related death: a systematic review and meta-analysis. *BMC Medical Informatics and Decision Making*, 23(1), 283.