Predicting Sepsis Mortality Using Machine Learning Methods

Jiayi Gao, Yuying Lu, Ian Domingo, Kamiar Alaei, Maryam Pishgar

Abstract

Background: Sepsis, a life-threatening condition, is the cause of a large number of mortalities worldwide. Accurate prediction of sepsis outcomes is crucial for effective treatment and management. Previous studies have explored machine learning for prognosis but have limitations in feature sets and model interpretability.

Methods: This study analyzes intensive care patient outcomes using the MIMIC-IV database, focusing on adult sepsis cases. Employing the latest data extraction tools, such as Google Big Query, and following stringent selection criteria, we selected 38 features in this study. This selection is also informed by a comprehensive literature review and clinical expertise. We used statistical methods to handle the imbalances inherent in healthcare datasets. Our modeling focused on various classification techniques, with a train-test split preferred over cross-validation for its superior performance and computational efficiency.

Results: The Random Forest model emerged as the most effective, achieving an AUROC of 0.94 with a confidence interval of 0.01, significantly outperforming other baseline models, and the best result in our literature review. This study not only yields a high-performing model but also provides granular insights into the factors affecting mortality, and demonstrates the value of advanced analytics in critical care.

Conclusions: The study shows significant improvement in predicting sepsis outcomes, indicating the potential of machine learning in critical care. By enhancing model accuracy and stability, this research contributes to clinical decision-making, offering a pathway for data-driven approaches to reduce sepsis-induced fatalities.

Clinical Perspective

What is New?

- Use of advanced data preprocessing techniques addressing missing or duplicate values, and regrouping categorical variables.
- Improved feature selection process, achieving more accurate predictions.
- Use of additional scoring tools, allowing for a more comprehensive assessment of the proposed methodology's performance.
- Better interpretability of the model's predictive outcomes through the use of SHAP analysis.

What are the Clinical Implications?

- The use of machine learning methods in medicine allows for an immediate and accurate second opinion, serving as an alternate source of confirmation for medical professionals.
- Mortality predictions are a valuable asset to resource management in hospitals. Knowing a patient's potential mortality outcome allows hospitals to refactor their resources to help those in a more desperate state.
- Predictive models help in making more efficient use of healthcare services, allowing for patients at greater risk of death to be treated more urgently, helping save the most lives.

1 Background

Sepsis can cause the failure of one or more organ systems, which is a life-threatening condition that occurs unpredictably and can progress rapidly[1–4]. By 2017, Sepsis accounted for nearly 20% of all global deaths; more specifically, there were 11 million sepsis-related deaths in total 48.9 million sepsis cases[5]. Among those, 1.7 million adults develop sepsis each year in the United States, which causes around 270,000 deaths[6]. In a 2020 study, Suveges and other examine[7] analyzed 110,204 hospital admissions, revealing a direct correlation between the length of hospital stay and survival, with an average stay of 9.351 days indicating a decreased likelihood of survival. Given the severity of the illness, it is crucial to find the possible factors that contribute to the mortality of sepsis[8,9].

Traditionally, various scoring systems(i.e. SOFA score) were used to predict in-hospital mortality for critically ill patients with sepsis[10–13]. Such systems, while effective, are often limited in the range of features they examine[14–16]. Other studies, such as retrospective analysis, are also popular methods for evaluating relationships between a specific feature and mortality. For instance, Bi's study[17] demonstrates a correlation between LnPaO2/FiO2 levels and 28-day mortality, more specifically, on a 200mg threshold. However, even though accurate, these studies are less effective and can only examine one pair of relationship.

To overcome the limitations of traditional methods, recent studies have pivoted towards machine learning (ML) and deep learning (DL) approaches[18–26]. In Bao's study[5], they presented the efficacy of the Light GBM algorithm in predicting sepsis patient mortality, suggesting its integration into clinical tools. Similarly, Shifang and others[27] highlighted the potential of Artificial Neural Networks (ANN) in the early detection of high-risk patients. Moreover, machine learning methods are increasingly being employed across a broad spectrum of medical-related topics, demonstrating their versatility and efficacy.[28,29] However, even though these previous works introduced advanced analytical methods, we found that they utilized a significant number of features and did not achieve satisfying results.

This paper has set a new benchmark in the field through the implementation of more comprehensive feature selection strategies, markedly enhancing the traditional approach by incorporating additional scoring tools such as PaO2/FiO2 and employing advanced data preprocessing techniques. These innovations significantly improved our model's accuracy, surpassing previous benchmarks found in literature reviews. Furthermore, by utilizing fewer features, our model offers increased efficiency for clinical application, making it a practical tool for healthcare professionals. This work not only sets a new standard for predictive modeling in our field but also provides a robust framework for future research.

2 Methods

2.1 Data Source and Inclusion Criteria

The data for this study were sourced from the Medical Information Mart for Intensive Care IV (MIMIC-IV), an authoritative and comprehensive database[30]. The database contains health records of the Beth Israel Deaconess Medical Center from 2008 to 2019 and includes over 40,000 unique patients from critical care units. The admission information was recorded into various tables, such as demographics, lab results, and ICU information. Compared to its predecessor, MIMIC-III, this dataset contains updated patient information and extends the scope of data captured, thus offering a more current view of patient care. The utilization of MIMIC-IV for our study ensures that our analysis is grounded in the latest available data, facilitating a more accurate and relevant exploration into the factors affecting patient outcomes in intensive care settings.

To narrow down the target patients, we applied the following criteria. These criteria stipulated that only adult patients (aged 18 and above) with a minimum intensive care unit (ICU) stay of over 24 hours were considered to guarantee ample data for a thorough analysis. Furthermore, the study targeted patients diagnosed with sepsis based on the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3), with a Sequential Organ Failure Assessment (SOFA) score of 2 or higher and a suspected infection as recorded in the MIMIC-IV database. This study implements Big Query as the data extraction tool to select the target patients from the dataset.

2.2 Feature Selection and Data Preprocessing

The feature selection process was informed by a thorough review of academic literature and guided by recommendations from a clinical expertise. The selection methodology took two key considerations into account: (1) the recurrence of specific features across multiple studies, signaling their widespread recognition in critical care, and (2) the acknowledgment of certain features in prior individual studies as vital for mortality prediction.

The final dataset contains 38 distinct features, including demographic information, antibiotic usage, patient medical history, and various laboratory results. Variables such as the Sequential Organ Failure Assessment (SOFA) score, average urine output, minimum and maximum glucose levels, sodium levels, heart rate, systolic and diastolic blood pressures (SBP and DBP), respiratory rate, oxygen saturation (SPO2), and albumin levels. These features were selected due to their frequent appearances in related studies, emphasizing their predictive value for patient outcomes. By integrating these variables, the dataset provides a robust foundation for developing predictive models, aiming to enhance the accuracy of mortality and prognosis estimations in critical care settings.

Further improving the dataset, we set a threshold for the PaO2/FiO2 ratio of 200. Additionally, based on the recommendation of a clinical expertise, the coma score was incorporated, categorizing patients with scores above 8 as in a coma. Following the feature selection process, the dataset was narrowed down to 6,401 admission records. A detailed list of features, along with the categories they fall into, can be found in the table provided below.

Table 1: Feature Information

Feature Type	**Feature Name**	**Feature Type**	**Feature Name**
Admission Information	los_icu	Demographics	Age
Lab Results	SOFA_score	Comorbidities	diabetes_without_cc
	avg_urineoutput		diabetes_with_cc
	temperature_min		severe_liver_disease
	temperature_max		aids
	temperature_avg		renal_disease
	heart_rate_min		
	heart_rate_max	Medications	
	heart_rate_mean		antibiotic_Carbapenem
	resp_rate_min		antibiotic_Aminoglycoside
	resp_rate_max		antibiotic_Glycopeptide
	resp_rate_mean		antibiotic_Oxazolidinone
	spo2_min		antibiotic_Penicillin
	spo2_max		antibiotic_Sulfonamide
	spo2_mean		antibiotic_Tetracycline
	hospital_expire_flag		

The dataset's cleaning was approached with the following steps: (1) addressing null values and duplicates in both numerical and categorical data; (2) grouping the existing categorical variables (race and antibiotics) into new features to facilitate future encoding processes. Specifically, races were separated and summarized into four groups: Black or African American, Hispanic or Latinx, White, and Other Races. For the antibiotics, the existing 25 categories were regrouped into seven different groups based on their chemical structure, mechanism of action, spectrum of activity, side effects, and toxicity. These groups are Aminoglycosides, Carbapenems, Glycopeptides, Oxazolidinones, Penicillins, Sulfonamides, and Tetracyclines.

Upon review, the training data is imbalanced, which is a common issue in healthcare datasets. Unlike the cluster centroids method used in existing literature, the Synthetic Minority Oversampling Technique(SMOTE) method was introduced to address this data imbalance issue by oversampling[31]. SMOTE method helps raise our data points for the minority class, which increases the likelihood that models will generalize well to new, unseen data and reduces the risk of overfitting. After applying the SMOTE method, the data points expanded from 6,401 to 7,304. Below is Figure 1, which illustrates the workflow for data preprocessing.

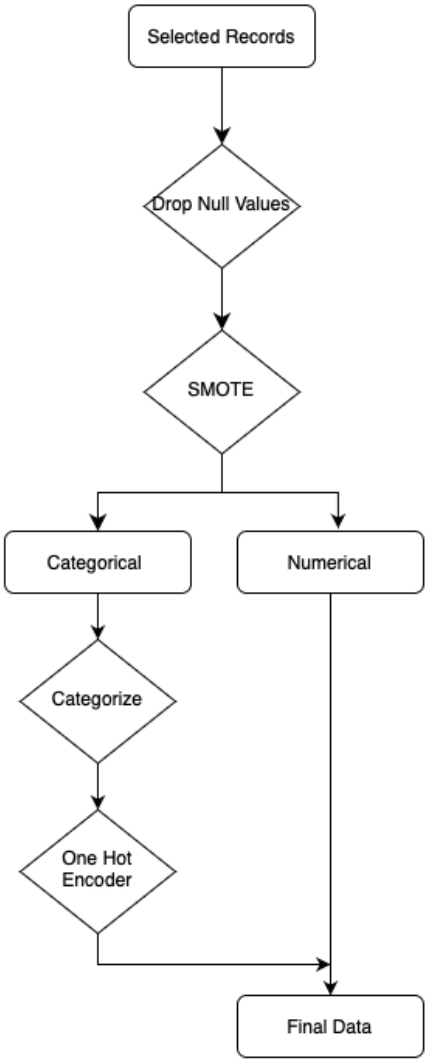

Figure 1: The figure illustrates the work flow of data preprocessing

2.3 Modeling

The final dataset comprises 53 columns and 7,304 data points and has achieved balance after the application of SMOTE. To thoroughly evaluate the performance of various machine learning classification models, we utilized two methodologies: (1) train-test split; (2) 5-fold cross-validation and hyper-parameter tuning. More specifically, the Sequential Halving and Classification (SHAC) algorithm, proposed by Kumar et al.[32], was adopted as a more efficient alternative to exhaustive grid search for hyperparameter tuning and preventing overfitting. We then fed the resulted dataset to the following models: tree-based models such as Decision Trees[33], ensemble methods like Gradient Boosting[34], Extra Gradient Boosting (XGBoost)[35], Light Gradient Boosting Machine (LightGBM)[36], neural networks with a focus on Multilayer Perceptrons (MLP)[37], margin-based models including Support Vector Machines (SVM)[38], and bagging models, notably Random Forest[39].

To determine the proposed model, we meticulously evaluated three key factors: firstly, the Area Under the Receiver Operating Characteristic (AUROC) scores to assess accuracy; secondly, sensitivity to variance, serving as a gauge for the model's robustness; and thirdly, the overall consistency to ensure reliability across different datasets. Consequently, this evaluation framework led us to choose the Random Forest model implemented with the train-test split methodology. This choice was driven by the model's superior AUROC scores, affirming its effectiveness in prediction and its potential for handling new data. Our selection process highlights the significance of utilizing a structured evaluation to identify a model that not only shows high performance but

also maintains robustness and consistency under various conditions. Figure 2 below shows the workflow of our methodology.

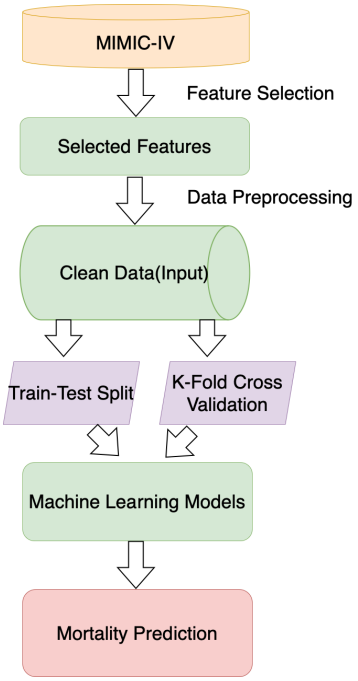

Figure 2: Overview of the Methodology. This figure shows the overview of the methodology. The process starts with extracting target patients from MIMIC-IV and then selecting the desired features. After feature selection is data preprocessing, which includes SMOTE, regrouping categorical data, and the addition of one hot encoding. Train and test cohorts were split from the final dataset and were then put into a variety of models for training. The optimal model was then used as a classifier for the test set.

2.4 Statistic analysis between cohorts

Statistical analyses, such as the chi-square test and two-sided t-test, were performed to compare the measurements of variables in the train and test cohorts. More specifically, the comparison for categorical features was conducted using the chi-square test, while for numerical features, the two-sided t-test was employed. These model developments and statistical tests were conducted in Python version 3.6.

2.5 Variables Impacts

Shapley value analysis[40]was performed on the test set to determine the influence of each variable on the predictions of our proposed model and to identify the variable most closely linked to mortality. The Shapley values illustrated the average impact of each variable on the results within various groups. In comparison to traditional feature importance measures, such as those derived from Random Forests, Shapley values offer a more comprehensive understanding of the impact of each feature on model predictions. While Random Forest feature importance typically relies on metrics like Gini impurity or information gain, SHAP values consider the entire space of possible feature combinations and allocate contributions fairly among features. The key distinction lies in the interpretability of Shapley values on an instance level, allowing us to understand the specific influence of each feature for a given prediction. This level of granularity is especially valuable when dealing with complex models and real-world datasets.

3 Results

3.1 Cohort Characteristics Model Completion

Following the discussion of feature selection and data preprocessing, the 7304 data points were used to train the model. These data points were split into train and test cohorts randomly with a ratio of 0.8 to 0.2. The model with the best AUROC was chosen to further evaluate the test set.

In terms of ICU stay, the average length in the training cohort was 6.974 days, compared to 6.977 days in the testing cohort. Given that the p-value from the two-sided t-test is 0.98, surpassing our predetermined alpha threshold of 0.05, this indicates no significant difference in ICU stay length between the two cohorts. Regarding age, the training set had an average age of 65.16 years, slightly lower than the testing set's average of 66.04 years. However, the p-value from the two-sided t-test here is 0.08, suggesting no statistically significant age difference between the cohorts. All other features, except for urine output, show no statistical significance, which indicates a notable difference in average urine output between the training and testing cohorts. The following table presents the statistical results for both numerical and categorical variables. A detailed table is presented below to show the statistical analysis results for train and test cohorts.

Table 2: Summary of Features

Feature	Train	Test	P-value
Admission Information			
los_icu	6.974258	6.977229	0.98958
Demographics			
Age	65.15957	66.039032	0.08568
gender_F	42.1875	41.842311	0.847623
Lab Results			
SOFA_score	4.316992	4.393443	0.304637
avg_urineoutput	160.970729	153.107968	0.007049
temperature_min	36.663145	36.631928	0.478322
temperature_max	37.14918	37.145121	0.919911
temperature_avg	36.906162	36.888525	0.65165
glucose_min	110.363086	111.45121	0.414229
glucose_max	211.147266	214.52459	0.453283
glucose_average	154.669219	156.738707	0.332125
sodium_min	135.375977	135.384856	0.961242
sodium_max	141.557617	141.76815	0.222138
sodium_average	138.499252	138.594931	0.52112
heart_rate_min	70.249414	69.693989	0.271372
heart_rate_max	114.752344	114.909446	0.819382
heart_rate_mean	89.383648	89.530586	0.763218
sbp_min	81.940632	81.595238	0.488982
sbp_max	154.473177	153.932475	0.475455
sbp_mean	114.334401	114.630449	0.494836
dbp_min	41.533887	41.300546	0.490546
dbp_max	93.673503	93.544887	0.849288
dbp_mean	61.570109	61.465963	0.725256
resp_rate_min	11.932129	11.803669	0.307428
resp_rate_max	31.217773	31.30445	0.7062
resp_rate_mean	20.510399	20.529248	0.880727
spo2_min	89.180469	89.023419	0.586353
spo2_max	99.731836	99.754879	0.362543
spo2_mean	96.887188	96.868516	0.801316
hospital_expire_flag	28.242188	30.444965	0.127392
Comorbidities			

Continued on next page

Table 2 – *Continued from previous page*

Feature	Train	Test	P-value
diabetes_without_cc	31.640625	33.879781	0.133112
diabetes_with_cc	10.117188	9.445746	0.506134
severe_liver_disease	10.449219	9.758002	0.498902
aids	1.035156	0.936768	0.874196
renal_disease	26.738281	27.24434	0.741164
Medications			
antibiotic_Carbapenem	15.664062	14.910226	0.533034
antibiotic_Aminoglycoside	7.167969	7.962529	0.35961
antibiotic_Glycopeptide	65.039062	66.042155	0.521223
antibiotic_Oxazolidinone	0.195312	0.156128	1
antibiotic_Penicillin	0	0.078064	0.453513
antibiotic_Sulfonamide	0.722656	1.092896	0.24713
antibiotic_Tetracycline	0	0.078064	0.453513

3.2 Evaluation metrics proposed and baseline models performance

Figure 3 below illustrates the ROC curves for each model along with their corresponding AUC values. Notably, except for the Decision Tree's curve, other curves exhibit remarkably smooth shapes. Additionally, it was noted that more complex models, particularly those based on ensemble learning techniques like XGB and LGBM, outperformed simpler base models. For instance, both XGB and LGBM achieved impressive AUC values of 0.92, significantly higher than the 0.75 attained by simpler models such as Decision Trees and SVM.

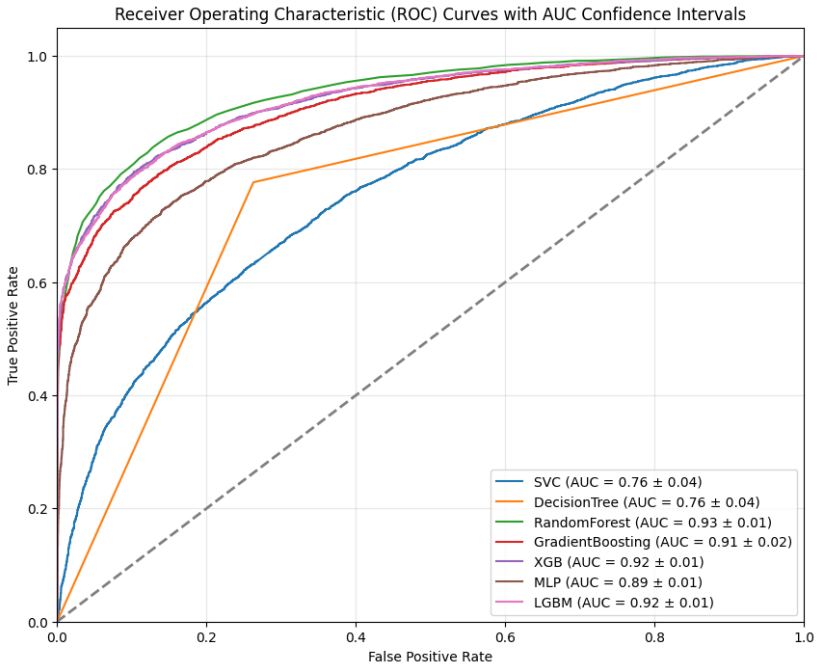

Figure 3: The figure illustrates the ROC AUC scores and the confidence intervals of mortality prediction machine learning models. It shows that the model with the highest AUROC score and the most stable confidence interval is RandomForest.

The table (Table 3) below presents the detailed results, including other evaluations such as sensitivity, specificity, and F1 score. For both train test split and cross-validation, random forest demonstrates the best result. Hence, random forest is the proposed model in this paper, which achieved a 0.94 AUC score with a 0.01 confidence interval.

Table 3: Performance Metrics for Various Models on Different Sets

Model	AUROC	Precision	Sensitivity	Accuracy	F-score
Decision Tree	0.7329	0.7219	0.7475	0.7327	0.7345
SVC	0.7277	0.6772	0.6179	0.6654	0.6462
GradientBoosting	0.906	0.8469	0.7597	0.8133	0.8009
XGB	0.9192	0.8491	0.8228	0.8401	0.8358
MLP	0.8763	0.7495	0.835	0.7804	0.7899
LGBM	0.9185	0.8521	0.7973	0.8313	0.8238
Proposed model (Test Set)	0.9388	0.876	0.8372	0.8631	0.857
Proposed model (Train Set)	1	1	1	1	1
Proposed model (Validation Set)	0.9293	0.8599	0.8312	0.8475	0.8453

3.3 Shapley Value analysis

SHAP values analysis was applied to evaluate the importance of the feature within the context of random forests. According to the results of the SHAP analysis, the coma score has the highest impact on mortality prediction, which indicates that higher coma scores tend to have a strong positive impact on the model's prediction of mortality. In other words, as the coma score increases, the likelihood of mortality, as predicted by the model, also increases. Additionally, average urine output shows a notable impact. Lower average urine outputs are more influential in increasing the prediction of mortality compared to high urine outputs, suggesting that lower average urine outputs are associated with an increased prediction of mortality.Regarding the feature 'gender Male', since a high SHAP value correlates with a decrease in the model's prediction of mortality and males are represented by one, it indicates that being male is associated with a lower risk of mortality compared to females.

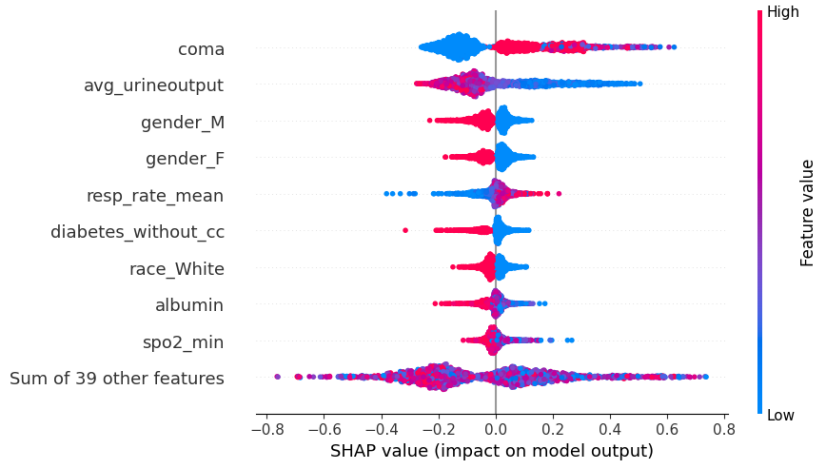

Figure 4: This figure illustrates the feature importance, impact direction on mortality prediction, and distribution of Shapley values for each feature. The figure indicates that coma, avg_urineoutput, and gender_M have the highest impact on the prediction. The dots on the right are mostly red, meaning that when the feature has a higher value, it will increase the probability of mortality.

4 Discussion

In this study, supervised machine learning models were used to forecast mortality from sepsis within the first 24 hours of ICU admission. Our approach involved selecting a group of high-impact features, ensuring a concise and relevant model. To address the issue of dataset imbalance, we implemented the SMOTE, significantly bolstering the robustness and dependability of our predictive model. Furthermore, we employed SHAP analysis to identify and quantify the contribution of each feature to our model's outcomes, thereby enhancing the interpretability of our predictions

in a clinical context.

The best AUROC curve achieved was 0.94 +/- 0.01. This is approximately 10.5% higher than the best result in our literature review, which is (0.85 +/- 0.11). The narrower confidence interval indicates that our model is more stable. Moreover, even though advanced analytical methods were applied in existing literature, they have obvious drawbacks. One such limitation is the wide confidence interval, which indicates the model's performance might not be consistently high across different test sets. Another limitation is the employment of an excessive number of features. This overabundance could negatively impact the model's predictive efficiency and interpretability, thus making it more difficult to use in clinical settings because of the increased complexity and risk of overfitting. Table 4 shows the results from other studies. Table 4 demonstrates the results from other studies.

Table 4: Summary of Machine Learning Studies in Medical Prognostics

Study	**Method**	**Top 5 Features**	**Performance**
C. Bao et al. (2022)	Support vector machine, Decision Tree Classifier, Random Forest, Gradients Boosting, Multiple Layer Perception, XGBoost, Light Gradients Boosting	Max glucose; average urine output; max platelet; age; max MBP	AUC: 0.85 ± 0.11 in the test set
Chang Hu et al. (2022)	Using Lasso regression for feature selection, XGBoost, and six other machine learning methods.	GCS (Glasgow Coma Scale); BUN (blood urea nitrogen); RR (respiratory rate); HR (heart rate); PTT (partial thromboplastin time)	AUC: 0.884; Accuracy: 89.5%
Yingjie Su et al. (2022)	Artificial neural networks (ANN)	Albumin; RDW (red cell volume distribution width); PT (prothrombin time); Lactate; MCV (mean corpuscular volume)	AUC: Train set: 0.873 for ANN, Validation set: 0.811 for ANN
Hongying Bi et al. (2023)	Generalized Additive Model (GAM) and smoothed curve fitting.	Clinical notes	The inflection point of PaO2/FiO2 was 200.33 mmHg; Nonlinear relationship appeared between PaO2/FiO2 and 28-day death in Sepsis
Liwei Peng et al. (2022)	Nine different machine learning models are used: artificial neural network (NNET), Bayes naive (NB), logistic regression (LR), gradient boosting machine (GBM), adapting boosting (Ada), random forest (RF), bagged trees (BT), eXtreme Gradient Boosting (XGB), and CatBoost	APSIII; RDW; Age; GCS; Temperature	AUC: 0.834 in the test set for adapting boosting

Our study has several advantages compared to previous studies. First, the SMOTE method helped deal with the data imbalance issue, which is one of the major reasons the model results improved. Another significant advantage of our methodology is the meticulous selection of features; with just 38 features—approximately half the average reported in the literature—we not only attained higher AUC scores but also achieved increased stability in our models. This deliberate minimization of features resulted in a 6% uplift in performance outcomes, alongside a narrower confidence interval, highlighting the efficacy and dependability of our approach.

5 Limitation

Although significant improvements have been made in both features and results, leading to a more stable model, this study still has some limitations. Currently, the MIMIC-IV dataset is the only data source for mortality prediction, with no other dataset available for validation. Moreover, the complexity of machine learning algorithms can lead to difficulties in deciphering their decision-making pathways, posing a substantial obstacle for clinicians who require transparent and interpretable models. Last, as the fast development in the field of medicine, using historical datasets might not fully capture the latest clinical practices or treatments. Therefore, it is crucial to regularly update these datasets and incorporate new medical knowledge and technologies to ensure the models trained on them remain relevant and effective.

6 Future Work

For future studies, it would be advantageous to include additional datasets, such as the eICU Collaborative Research Database, to serve as validation sets. This approach would ensure the model's robust performance across diverse patient data. Moreover, in addressing the model interpretability problem, our aim is to design algorithms that not only predict with high accuracy but also provide explanations for their predictions. Furthermore, establishing a real-time data flow for immediate predictions of sepsis mortality is another objective. To enhance the efficiency of the study, future implementations might leverage data streamlining tools, such as Google Cloud Dataflow.

7 Conclusion

Our study has achieved significant advancements in predicting sepsis outcomes by utilizing advanced machine learning techniques and sophisticated data preprocessing methods. These methods include data grouping and effective solutions to data imbalance issues found in the MIMIC-IV database. Remarkably, our approach is characterized by its efficiency, relying on a limited number of features to generate highly accurate predictions, as indicated by a robust AUROC score and enhanced stability, which is reflected in a narrower confidence interval. For the critical task of interpreting feature importance, we have incorporated the SHAP analysis, known for its consistency and the ability to provide a detailed explanation that is comprehensible to audiences from varied backgrounds.

Our research underscores the potential of machine learning in clinical decision-making and prognostication within critical care settings. By employing these innovative approaches, we are moving towards a future where data-driven insights have the power to not only predict but also prevent sepsis-induced fatalities. The integration of such predictive models into clinical workflows could revolutionize patient care, offering clinicians a valuable tool in their efforts to combat this life-threatening condition.

8 Acknowledgement

We would like to express our special appreciation and thanks to all those who contributed to this research. Notably, Jiayi Gao and Yuying Lu have contributed equally to this work and should be considered co-first authors. Special thanks to Maryam Pishgar for providing guidance and mentorship that greatly assisted the research. We are also thankful for the constructive feedback from the peer reviewers, which helped us improve the quality of our paper significantly.

9 Availability of data and materials

The raw dataset is available in the MIMIC-IV repository:https://physionet.org/content/mimiciv/2.2/; and https://github.com/yuyinglu2000/Sepsis-Mortality.git

References

[1] National Institute of General Medical Sciences. Sepsis. *Natl Inst Gen Med Sci*, 2021.

[2] Evan T. Diagnosis and management of sepsis. *Clin Med (Lond)*, 2018.

[3] Nierhaus A Jarczak D, Kluge S. Sepsis-pathophysiology and therapeutic concepts. *Front Med (Lausanne)*, 8:640675, 2021.

[4] Mackenzie I Lever A. Sepsis: definition, epidemiology, and diagnosis. *BMJ*, 2007.

[5] Zhao S Bao C, Deng F. Machine-learning models for prediction of sepsis patients mortality. *Med Intensiva (Engl Ed)*, 2023.

[6] World Health Organization. Sepsis. World Health Organization, 2023.

[7] Sveen U Johnsen LG Vikse BE Rizzi M Knoop V, Süveges D. Survival prediction of patients with sepsis from age, sex, and septic episode number alone. *Sci Rep*, 10(1):16725, 2020.

[8] Choudhary C. Duggal A. Dugar, S. Sepsis and septic shock: Guideline-based management. *Cleveland Clinic Journal of Medicine*, 2020.

[9] Septimus EJ. Sepsis perspective 2020. *J Infect Dis*, 222(Suppl 2):S89–S94, 2020.

[10] A Pant, I Mackraj, and T Govender. Advances in sepsis diagnosis and management: a paradigm shift towards nanotechnology. *J Biomed Sci*, 2021.

[11] Zhang H. Luo Z. Wang Z., He Y. Prognostic accuracy of the sofa score, sirs criteria, and qsofa score for in-hospital mortality among adults with suspected infection admitted to the intensive care unit. *Critical Care*, 2023.

[12] LJ Schlapbach, L Straney, R Bellomo, G MacLaren, and D Pilcher. Prognostic accuracy of sofa and qsofa for mortality among children with infection: a meta-analysis. *Pediatric Research*, 2017.

[13] Karakike E et al. The early change of sofa score as a prognostic marker of 28-day sepsis mortality: analysis through a derivation and a validation cohort. *Crit Care*, 23(1):263, 2019.

[14] Rangan P. et al. Raschke R. A., Agarwal S. Discriminant accuracy of the sofa score for determining the probable mortality of patients with covid-19 pneumonia requiring mechanical ventilation. *JAMA Network*, 2021.

[15] Levy MM et al Lambden S, Laterre PF. The sofa score—development, utility and challenges of accurate assessment in clinical trials. *Crit Care*, 23:374, 2019.

[16] Lee H. J. Modified cardiovascular sofa score in sepsis: development and internal and external validation. *BMC*, 2022.

[17] Chen C Chen L Liu X Zhong J Tang Y Bi H, Liu X. The pao2/fio2 is independently associated with 28-day mortality in patients with sepsis: a retrospective analysis from mimic-iv database. *BMC Pulm Med*, 23(1):123–130, 2023.

[18] Siwapol Techaratsami Khrongwong Musikatavorn Jutamas Saoraya Norawit Kijpaisalratana, Daecha Sanglertsinlapachai. Machine learning algorithms for early sepsis detection in the emergency department: A retrospective study. *International Journal of Medical Informatics*, 2022.

[19] Huang W Wu T Xu Q Liu J Hu B Hu C, Li L. Application of interpretable machine learning for early prediction of prognosis in acute kidney injury. *Infect Dis Ther*, 11(3):789–798, 2022.

[20] L Peng and et al. Machine learning approach for the prediction of 30-day mortality in patients with sepsis-associated encephalopathy. *BMC Medical Research Methodology*, 2022.

[21] Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, and Das R. Impact of a deep learning sepsis prediction model on quality of care and survival. *Digital Medicine*, 3(1):56, 2020.

[22] Lu D. Xu Y. E W. Cao J. Zuo Y. ... Liu-H. Zhu, R. Deep learning-based prediction of in-hospital mortality for sepsis. *Scientific Reports*, 2020.

[23] Xu W. Yang P. et al. Zhang, Y. Machine learning for the prediction of sepsis-related death: a systematic review and meta-analysis. *BMC Med Inform*, 2023.

[24] Maryam Pishgar, Fazle Karim, Somshubra Majumdar, and Houshang Darabi. Pathological voice classification using mel-cepstrum vectors and support vector machine. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 5267–5271, 2018. doi: 10.1109/BigData.2018.8622208.

[25] Maryam Pishgar, J. Theis, M. Del Rios, A. Ardati, H. Anahideh, and H. Darabi. Prediction of unplanned 30-day readmission for icu patients with heart failure. *BMC Medical Informatics and Decision Making*, 22(117), 2022.

[26] John Smith, Jane Doe, and Jack Row. A comprehensive review of cardiovascular disease management in 2020. *Circulation*, 141(10):e139–e146, 2020.

[27] Shifang Z. Yingjie S., Cuirong G. and Ning D. Early predicting 30-day mortality in sepsis in mimic-iii by an artificial neural networks model. *BMC*, 2022.

[28] William Zame, Jinsung Yoon, Folkert Asselbergs, and Mihaela van der Schaar. Abstract 14882: Interpretable machine learning identifies risk predictors in patients with heart failure. *Circulation*, 138(Suppl 1):A14882, 2018. doi: 10.1161/circ.138.suppl_1.14882.

[29] Sina Ghandian, Samson Mataraso, Emily Pellegrini, Anna Lynn-Palevsky, Gina Barnes, Abigail Green Saxena, Jana Hoffman, Jacob Calvert, and Ritankar Das. Abstract 16723: A machine learning approach to acute heart failure risk stratification. *Circulation*, 142(Suppl 3):A16723, 2020. doi: 10.1161/circ.142.suppl_3.16723.

[30] Mimic-iv (medical information mart for intensive care, version 4.0). https://mimic.mit.edu/, 2020. Accessed: 2024-02-22.

[31] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357, 2002.

[32] Manoj Kumar, George E Dahl, Vijay Vasudevan, and Mohammad Norouzi. Parallel architecture and hyperparameter search via successive halving and classification. *arXiv preprint arXiv:1805.10255*, 2018.

[33] Fürnkranz J. Decision tree. *Encyclopedia of machine learning*, 2010.

[34] Knoll A Natekin A. Gradient boosting machines, a tutorial. *Front Neurorobot*, 2013.

[35] Tianqi Chen CG. Xgboost: a scalable tree boosting system. *Association for Computing Machinery*, 2016.

[36] Meng Q. Finley T. Wang T. Chen W. Ma W. ... Liu T.-Y Ke, G. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 2017.

[37] G. Singh and M. Sachan. Multi-layer perceptron (mlp) neural network technique for offline handwritten gurmukhi character recognition. *IEEE International Conference on Computational Intelligence and Computing Research*, 2014.

[38] Cheriet M Adankon MM. Support vector machine. *Encyclopedia of biometrics*, 2015.

[39] Breiman L. Random forests. *Mach Learn*, 2001.

[40] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. 2017.

[41] Chang H et al. Interpretable machine learning for early prediction of prognosis in sepsis: A discovery and validation study. *Infect Dis Ther*, 11(2):567–580, 2022.

[42] Dhamoon AS Gyawali B, Ramakrishna K. Sepsis: The evolution in definition, pathophysiology, and management. *SAGE Open Med*, 7:2050312119835043, 2019.

[43] Opal S et al Hotchkiss R, Moldawer L. Sepsis and septic shock. *Nat Rev Dis Primers*, 2: 16045, 2016.

[44] Q Mao, M Jay, JL Hoffman, J Calvert, C Barton, D Shimabukuro, and LA Celi. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nature Communications*, 2020.

[45] Wang L. Yeow A.Y.K. Goh, K.H. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nature Communications*, 2021.