

Prediction of Sepsis Mortality in ICU Patients Using Machine Learning Methods

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

Abstract

Problem: Sepsis, a life-threatening condition, accounts for the deaths of millions of people worldwide. Accurate prediction of sepsis outcomes is crucial for effective treatment and management. Previous studies have utilized machine learning for prognosis, but have limitations in feature sets and model interpretability.

Aim: This study aims to develop a machine learning model that enhances prediction accuracy for sepsis outcomes using a reduced set of features, thereby addressing the limitations of previous studies and enhancing 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 BigQuery, 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. Data preprocessing included handling missing values, regrouping categorical variables, and using the Synthetic Minority Over-sampling Technique (SMOTE) to balance the data. We evaluated several machine learning models: Decision Trees, Gradient Boosting, XGBoost, LightGBM, Multilayer Perceptrons (MLP), Support Vector Machines (SVM), and Random Forest. The Sequential Halving and Classification (SHAC) algorithm was used for hyperparameter tuning, and both train-test split and cross-validation methodologies were employed for performance and computational efficiency.

Results: The Random Forest model was the most effective, achieving an area under the receiver operating characteristic curve (AUROC) of 0.94 with a confidence interval of ± 0.01 . This significantly outperformed other models and set a new benchmark in the literature. The model also provided detailed insights into the importance of various clinical features, with the Sequential Organ Failure Assessment (SOFA) score and average urine output being highly predictive. SHAP (Shapley Additive Explanations) analysis further enhanced the model's interpretability, offering a clearer understanding of feature impacts.

Conclusion: This study demonstrates significant improvements in predicting sepsis outcomes using a Random Forest model, supported by advanced machine learning techniques and thorough data preprocessing. Our approach provided detailed insights into the key clinical features impacting sepsis mortality, making the model both highly accurate and interpretable. By enhancing the model's practical utility in clinical settings, we offer a valuable tool for healthcare professionals to make data-driven decisions, ultimately aiming to minimize sepsis-induced fatalities.

Keywords: Sepsis, Machine Learning, MIMIC-IV, ICU Mortality

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 [1–5]. 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 [6]. Among those, 1.7 million adults develop sepsis each year in the United States, which causes around 270,000 deaths [7]. In a 2020 study, Suveges and other examine [8] analyzed 110,204 hospital admissions, revealing a direct correlation between the length of hospital stay and survival, with an average stay of 10 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 [9–11].

Traditionally, various scoring systems (i.e. SOFA score) were used to predict in-hospital mortality for critically ill patients with sepsis [12–15]. Such systems, while effective, are often limited in the range of features they examine [16–18]. For example, these scoring systems typically focus on a narrow set of clinical parameters, which might not capture the full complexity of sepsis. This limitation can lead to incomplete assessments of a patient’s condition and subsequently, less accurate predictions. Other studies, such as retrospective analysis, are also popular methods for evaluating relationships between a specific feature and mortality. For instance, Bi’s study [19] demonstrates a correlation between PaO₂/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 at a time. This approach does not account for the multifactorial nature of sepsis, where multiple physiological and biochemical parameters interact in complex ways. As a result, these studies often miss critical interactions between features that could improve the predictive accuracy of sepsis outcomes. The inability to integrate and analyze multiple features simultaneously poses a significant barrier to developing more comprehensive and precise predictive models for sepsis. Furthermore, the reliance on retrospective data means these models are often not adaptive to the dynamic and rapidly changing clinical status of sepsis patients, further limiting their real-time applicability and effectiveness.

To overcome the limitations of traditional methods, recent studies have pivoted towards Machine Learning (ML) and Deep Learning (DL) approaches [20–30]. In Bao’s study [6], they presented the efficacy of the Light GBM algorithm in predicting sepsis patient mortality, suggesting its integration into clinical tools. Similarly, Shifang et al. [31] 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. [32–40], 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. These models often lacked comprehensive feature selection strategies and advanced data preprocessing techniques, which limited their accuracy and practicality in clinical settings. Additionally, the use of numerous features complicated the models, leading to overfitting and inefficiency, making them less suitable for real-time application. In the following, we will list the main contributions of our work:

- Advanced data preprocessing techniques were employed to address missing or duplicate values and to regroup categorical variables, significantly enhancing data quality and model performance.
- A thorough review of academic literature and recommendations from clinical experts guided our feature selection process, leading to more accurate predictions using a smaller, more relevant feature set.
- The use of SHAP (SHapley Additive exPlanations) analysis improved the interpretability of our model’s predictive outcomes, providing granular insights into the factors affecting sepsis mortality.
- The Synthetic Minority Over-sampling Technique (SMOTE) was used to address data imbalance, significantly improving the robustness of our model.
- Our proposed model, particularly the Random Forest model, achieved an AUROC of 0.94 with a narrower confidence interval, representing a 6.3% improvement compared to the best existing study.

This paper sets a new benchmark in the field, significantly improving model accuracy and efficiency, and making our model a practical tool for healthcare professionals. The use of machine learning methods in medicine provides an immediate and accurate second opinion, serving

99 as an alternate source of confirmation for medical professionals. Mortality predictions derived from
100 these models are valuable assets for resource management in hospitals, allowing for the refactoring
101 of resources to prioritize patients in more desperate conditions. Additionally, predictive models
102 facilitate more efficient use of healthcare services by enabling urgent treatment for patients at
103 greater risk of death, ultimately helping to save more lives. These advancements enhance clinical
104 decision-making and improve patient outcomes.

105
106 The rest of the paper is organized as follows: Section 2 Methods describes the data source and
107 inclusion criteria, feature selection and data preprocessing, modeling, statistical analysis between
108 cohorts, and variables impacts. Section 3 Results presents the cohort characteristics, evaluation
109 metrics, and Shapley value analysis. Section 4 Discussion interprets the findings and their signifi-
110 cance. Section 5 Limitations addresses the study’s constraints and potential weaknesses. Section
111 6 Future Work suggests directions for enhancing predictive capabilities and research extensions.
112 Finally, Section 7 Conclusion summarizes the key contributions and clinical impact of the study.

113 2 Methods

114 2.1 Data Source and Inclusion Criteria

115 The data for this study were sourced from the Medical Information Mart for Intensive Care IV
116 (MIMIC-IV), an authoritative and comprehensive database [41]. The database contains health
117 records of the Beth Israel Deaconess Medical Center from 2008 to 2019 and includes over 40,000
118 unique patients from critical care units. The admission information was recorded into various
119 tables, such as demographics, lab results, and ICU information. Compared to its predecessor,
120 MIMIC-III, this dataset contains updated patient information and extends the scope of data cap-
121 tured, thus offering a more current view of patient care. The utilization of MIMIC-IV for our study
122 ensures that our analysis is grounded in the latest available data, facilitating a more accurate and
123 relevant exploration into the factors affecting patient outcomes in intensive care settings.

124
125 To narrow down the target patients, we applied the following criteria. These criteria stipulated
126 that only adult patients (aged 18 and above) with a minimum intensive care unit (ICU) stay of
127 over 24 hours were considered to guarantee ample data for a thorough analysis. Furthermore,
128 the study targeted patients diagnosed with sepsis based on the Third International Consensus
129 Definitions for Sepsis and Septic Shock (Sepsis-3), with a Sequential Organ Failure Assessment
130 (SOFA) score of 2 or higher and a suspected infection as recorded in the MIMIC-IV database.
131 This study implements BigQuery as the data extraction tool to select the target patients from the
132 dataset.

133 2.2 Feature Selection and Data Preprocessing

134 The feature selection process was informed by a thorough review of academic literature and guided
135 by recommendations from a clinical expertise. The selection methodology took two key consider-
136 ations into account: (1) the recurrence of specific features across multiple studies, signaling their
137 widespread recognition in critical care, and (2) the acknowledgment of certain features in prior
138 individual studies as vital for mortality prediction. This selection was based on their prevalence
139 in existing literature, clinical importance, and statistical validation to ensure their relevance and
140 predictive power.

141
142 The final dataset contains 38 distinct features, including demographic information, antibiotic
143 usage, patient medical history, and various laboratory results. Variables such as the Sequential
144 Organ Failure Assessment (SOFA) score, average urine output, minimum and maximum glucose
145 levels, sodium levels, heart rate, systolic and diastolic blood pressures (SBP and DBP), respiratory
146 rate, oxygen saturation (SPO2), and albumin levels. These features were selected due to their fre-
147 quent appearances in related studies, emphasizing their predictive value for patient outcomes. By
148 integrating these variables, the dataset provides a robust foundation for developing predictive mod-
149 els, aiming to enhance the accuracy of mortality and prognosis estimations in critical care settings.

150
151 Further improving the dataset, we set a threshold for the PaO2/FiO2 ratio of 200 [17]. Ad-
152 ditionally, based on the recommendation of a clinical expertise, the coma score was incorporated,
153 categorizing patients with scores above 8 as in a coma. Following the feature selection process,
154 the dataset was narrowed down to 6,401 admission records. A detailed list of features, along with

155
156

the categories they fall into, can be found in the Table 1 provided below.

Table 1: Detailed Overview of 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	Medications	antibiotic_Carbapenem
	heart_rate_max		antibiotic_Aminoglycoside
	heart_rate_mean		antibiotic_Glycopeptide
	resp_rate_min		antibiotic_Oxazolidinone
	resp_rate_max		antibiotic_Penicillin
	resp_rate_mean		antibiotic_Sulfonamide
	spo2_min		antibiotic_Tetracycline
	spo2_max		
spo2_mean			
hospital_expire_flag			

157

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

166
167 Upon review, the training data is imbalanced, which is a common issue in healthcare datasets.
168 Unlike the cluster centroids method used in existing literature, the Synthetic Minority Over-
169 sampling Technique(SMOTE) method was introduced to address this data imbalance issue by
170 oversampling [42]. SMOTE method helps raise our data points for the minority class, which
171 increases the likelihood that models will generalize well to new, unseen data and reduces the
172 risk of overfitting. After applying the SMOTE method, the data points expanded from 6,401
173 to 7,304. By doing so, SMOTE helps balance the dataset, which is crucial for training models
174 that generalize well to new, unseen data and reduces the risk of overfitting. This method ensures
175 that our predictive models are more robust and reliable. Below is Figure 1, which illustrates the
176 workflow for data preprocessing.

177 2.3 Modeling

178 The final dataset comprises 53 columns and 7,304 data points and has achieved balance after
179 the application of SMOTE. To thoroughly evaluate the performance of various machine learning
180 classification models, we utilized two methodologies: (1) train-test split; (2) 5-fold cross-validation
181 and hyper-parameter tuning. More specifically, the Sequential Halving and Classification (SHAC)
182 algorithm, proposed by Kumar et al. [43], was adopted as a more efficient alternative to exhaus-
183 tive grid search for hyperparameter tuning and preventing overfitting. We then fed the resulted
184 dataset to the following models: tree-based models such as Decision Trees [44], ensemble methods
185 like Gradient Boosting [45], Extra Gradient Boosting (XGBoost) [46], Light Gradient Boosting
186 Machine (LightGBM) [47], neural networks with a focus on Multilayer Perceptrons (MLP) [48],
187 margin-based models including Support Vector Machines (SVM) [49], and bagging models, notably
188 Random Forest [50].

189 To determine the proposed model, we meticulously evaluated three key factors: firstly, the
190 Area Under the Receiver Operating Characteristic (AUROC) scores to assess accuracy; secondly,

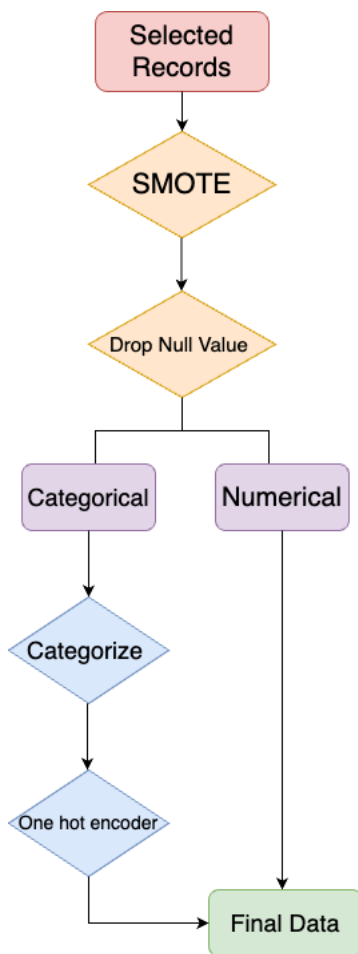


Figure 1: Work flow of data preprocessing. The process begins with selected records (red), followed by initial preprocessing steps like SMOTE and dropping null values (orange). Data is then divided into categorical and numerical types for further processing (purple). Categorical data is categorized and encoded (blue). Finally, the processed data is ready for model fitting (green).

191 sensitivity to variance, serving as a gauge for the model’s robustness; and thirdly, the overall
 192 consistency to ensure reliability across different datasets. Consequently, this evaluation frame-
 193 work showed that the Random Forest model outperformed other models using both the train-test
 194 split and cross-validation methodologies. This choice was driven by the model’s superior AUROC
 195 scores, affirming its effectiveness in prediction and its potential for handling new data. Our se-
 196 lection process highlights the significance of utilizing a structured evaluation to identify a model
 197 that not only shows high performance but also maintains robustness and consistency under various
 198 conditions. Figure 2 below shows the workflow of our methodology.
 199

200 2.4 Statistic analysis between cohorts

201 Statistical analyses, such as the chi-square test and two-sided t-test, were performed to compare
 202 the measurements of variables in the train and test cohorts. More specifically, the comparison
 203 for categorical features was conducted using the chi-square test, while for numerical features, the
 204 two-sided t-test was employed. These model developments and statistical tests were conducted in
 205 Python version 3.6.
 206

207 2.5 Variables Impacts

208 Shapley value analysis [51] was performed on the test set to determine the influence of each vari-
 209 able on the predictions of our proposed model and to identify the variable most closely linked
 210 to mortality. The Shapley values illustrated the average impact of each variable on the results

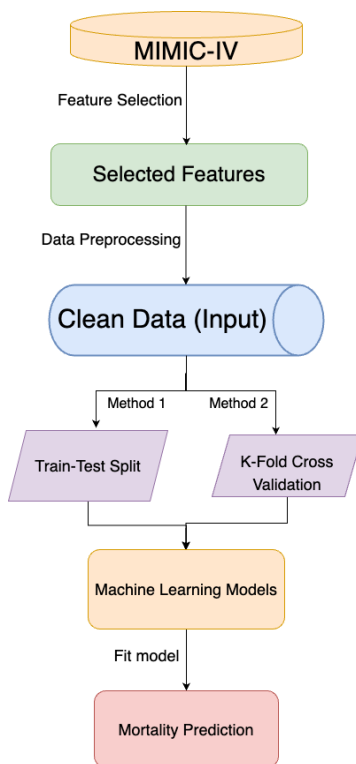


Figure 2: Overview of the Methodology. The process starts with extracting target patients from MIMIC-IV, followed by feature selection and data preprocessing (green). The input data (blue) is split for model fitting (purple). The optimal model (yellow) is then used for predictions (red).

211 within various groups [52]. In comparison to traditional feature importance measures, such as
 212 those derived from Random Forest, Shapley values offer a more comprehensive understanding of
 213 the impact of each feature on model predictions. While Random Forest feature importance typ-
 214 ically relies on metrics like Gini impurity or information gain, SHAP values consider the entire
 215 space of possible feature combinations and allocate contributions fairly among features [53]. The
 216 key distinction lies in the interpretability of Shapley values on an instance level, allowing us to
 217 understand the specific influence of each feature for a given prediction. This level of granularity
 218 is especially valuable when dealing with complex models and real-world datasets.
 219

220 3 Results

221 3.1 Cohort Characteristics Model Completion

222 Following the discussion of feature selection and data preprocessing, the 7304 data points were
 223 used to train the model. The number of data points was determined by setting thresholds
 224 for LnPaO₂/FiO₂ at 200 mg, based on a literature review of Bi’s study, which highlighted
 225 LnPaO₂/FiO₂ as an important factor influencing sepsis mortality. Additionally, the SMOTE
 226 method was employed to balance the dataset [19]. These data points were split into train and
 227 test cohorts randomly with a ratio of 0.8 to 0.2. The model with the best AUROC was chosen to
 228 further evaluate the test set.
 229

230 In terms of ICU stay, the average length in the training cohort was 6.974 days, compared to
 231 6.977 days in the testing cohort. Given that the p-value from the two-sided t-test is 0.989, sur-
 232 passing our predetermined alpha threshold of 0.05, this indicates no significant difference in ICU
 233 stay length between the two cohorts. Regarding age, the training set had an average age of 65.160
 234 years, slightly lower than the testing set’s average of 66.039 years. However, the p-value from the
 235 two-sided t-test here is 0.086, suggesting no statistically significant age difference between the co-
 236 horts. All other features, except for urine output, show no statistical significance, which indicates
 237 a notable difference in average urine output between the training and testing cohorts. Table 2

238
239
240
241

below presents the statistical results of train and test cohorts for both numerical and categorical variables, where numerical feature values represent the average measurements, and categorical feature values represent the percentage of individuals in each category.

Table 2: Statistical Analysis Results of Train and Test Cohorts for Numerical and Categorical Variables

Feature	Train	Test	P-value
<i>Admission Information</i>			
los_icu	6.974	6.977	0.989
<i>Demographics</i>			
Age	65.160	66.039	0.086
gender_F	42.188	41.842	0.848
<i>Lab Results</i>			
SOFA_score	4.317	4.393	0.305
avg_urineoutput	160.971	153.108	0.007
temperature_min	36.663	36.632	0.478
temperature_max	37.149	37.145	0.920
temperature_avg	36.906	36.888	0.652
glucose_min	110.363	111.451	0.414
glucose_max	211.147	214.524	0.453
glucose_average	154.669	156.739	0.332
sodium_min	135.376	135.385	0.961
sodium_max	141.558	141.768	0.222
sodium_average	138.499	138.595	0.521
heart_rate_min	70.250	69.694	0.271
heart_rate_max	114.752	114.909	0.819
heart_rate_mean	89.384	89.530	0.763
sbp_min	81.941	81.595	0.489
sbp_max	154.473	153.932	0.475
sbp_mean	114.334	114.630	0.495
dbp_min	41.534	41.300	0.490
dbp_max	93.673	93.545	0.849
dbp_mean	61.570	61.466	0.725
resp_rate_min	11.932	11.804	0.307
resp_rate_max	31.218	31.304	0.706
resp_rate_mean	20.510	20.529	0.881
spo2_min	89.180	89.023	0.586
spo2_max	99.732	99.755	0.362
spo2_mean	96.887	96.868	0.801
hospital_expire_flag	28.242	30.445	0.127
<i>Comorbidities</i>			
diabetes_without_cc	31.641	33.880	0.133
diabetes_with_cc	10.117	9.446	0.506
severe_liver_disease	10.449	9.758	0.499
aids	1.035	0.937	0.874
renal_disease	26.738	27.244	0.741
<i>Medications</i>			
antibiotic_Carbapenem	15.664	14.910	0.533
antibiotic_Aminoglycoside	7.168	7.962	0.360
antibiotic_Glycopeptide	65.039	66.042	0.521
antibiotic_Oxazolidinone	0.195	0.156	1.000
antibiotic_Penicillin	0.000	0.078	0.453
antibiotic_Sulfonamide	0.723	1.093	0.247

Continued on next page

Table 2 – Continued from previous page

Feature	Train	Test	P-value
antibiotic_Tetracycline	0.000	0.078	0.453

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.

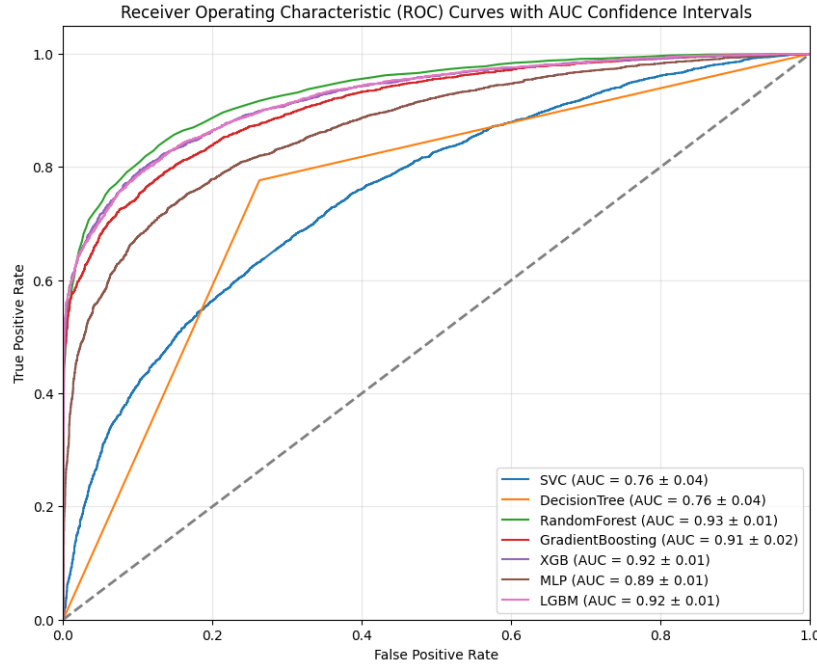


Figure 3: The figure illustrates the ROC AUC scores and the confidence intervals of mortality prediction machine learning models. It shows that the model with the highest AUROC score and the most stable confidence interval is RandomForest.

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 demonstrated the best result. Hence, Random Forest (RF) is the proposed model in this paper, which achieved a 0.94 AUC score with a 0.01 confidence interval representing a 6.3% improvement compared to the best existing study. It is important to note that, although we typically expect cross-validation to provide a more robust estimate of model performance, in this particular study, the train-test split yielded slightly better results by chance. By using a random seed of 42 to split our dataset into training and test sets, we achieved an AUROC of 0.93, which was 1% higher than the cross-validation results. Both methods demonstrated high AUROC scores and narrow confidence intervals, indicating the stability and robustness of the dataset and features selected. This consistency across different data partitioning methods validates the reliability of our chosen approach and ensures that the model’s performance is not significantly affected by the method of data partitioning.

3.3 Shapley Value analysis

SHAP values analysis was applied to evaluate the importance of the feature within the context of Random Forest. According to the results of the SHAP analysis, the coma score has the highest

Table 3: Metrics for Baseline Models on Test Set and Random Forest 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
RF (Test Set)	0.9388	0.876	0.8372	0.8631	0.857
RF (Trainig Set)	1.0000	1.0000	1.0000	1.0000	1.0000
RF (Cross Validation)	0.9293	0.8599	0.8312	0.8475	0.8453

267 impact on mortality prediction, which indicates that higher coma scores tend to have a strong positive
 268 impact on the model’s prediction of mortality. In other words, as the coma score increases,
 269 the likelihood of mortality, as predicted by the model, also increases. Additionally, average urine
 270 output shows a notable impact. Lower average urine outputs are more influential in increasing
 271 the prediction of mortality compared to high urine outputs, suggesting that lower average urine
 272 outputs are associated with an increased prediction of mortality. Regarding the feature ‘gender
 273 Male’, since a high SHAP value correlates with a decrease in the model’s prediction of mortality
 274 and males are represented by one, it indicates that being male is associated with a lower risk of
 275 mortality compared to females. Figure 4 demonstrates the detailed result of the SHAP values
 276 analysis. The ‘Sum of 39 features’ in the SHAP plot represents the combined SHAP values of the
 277 remaining less influential features. This aggregation provides a holistic view of their cumulative
 278 impact on the model’s predictions, highlighting that while individually these features may not
 279 have a significant impact, together they can still influence the model’s outcomes.

280
 281 After reviewing other articles about sepsis mortality rates, we found that there is some overlap
 282 between the top features. For example, the top five features in Bao’s study are glucose max,
 283 urine output, platelets max, age, and MBP max. This overlaps with our result on urine output,
 284 indicating that urine output is an important feature to be considered in real-life situations. These
 285 insights highlight key factors affecting mortality. Understanding that higher coma scores and lower
 286 average urine outputs significantly increase the risk of mortality, while being male is associated
 287 with a lower risk, can help in developing targeted interventions to mitigate mortality risks.

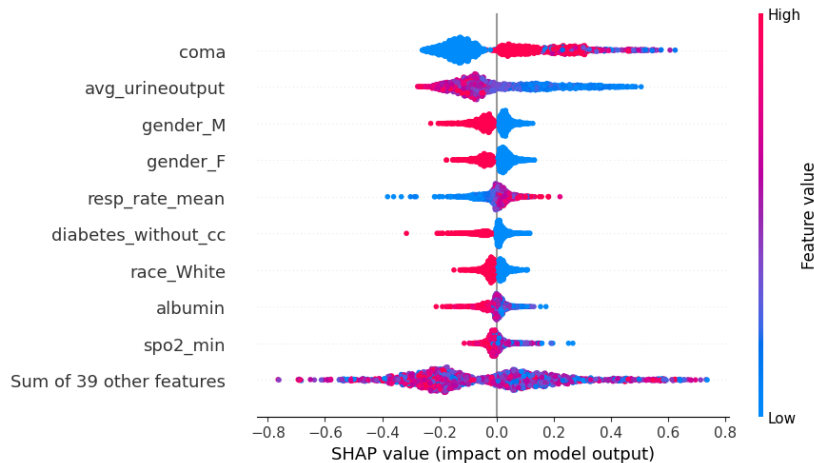


Figure 4: This figure illustrates the feature importance, impact direction on mortality prediction, and distribution of Shapley values for each feature. The figure indicates that coma, avg_urineoutput, and gender_M have the highest impact on the prediction. The dots on the right are mostly red, meaning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

4 Discussion

289

290

291

292

293

294

295

296

297

298

299

300

301

302

303

304

305

306

307

308

309

310

311

312

313

314

315

316

317

318

319

320

321

322

323

324

325

In this study, supervised machine learning models were used to forecast mortality from sepsis over 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 6.3% higher than the best result in our literature review, which is 0.884. The narrower confidence interval indicates that our model is more stable. The higher AUROC means that the model is good at accurately predicting the mortality of patients by successfully discriminating between positive cases (hospital exp flag = 1, the patient died) and negative cases (hospital exp flag = 0, the patient does not die). The stability of the model is important for providing consistent prediction results and ensures more robustness against changes in input data, making the model’s outcomes more reliable. Furthermore, the study’s approach contributed to reducing sepsis-induced fatalities by providing personalized suggestions for each patient through model fitting. The model aids clinicians in making early identifications of sepsis, ensuring more attention is given to patients at higher risk of sepsis mortality. This early detection and targeted intervention enhance healthcare efficiency and effectiveness, ultimately helping to reduce sepsis-induced fatalities. 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.

326

327

328

329

330

331

332

333

334

335

336

337

338

339

340

341

342

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 6.3% uplift in performance outcomes, alongside a narrower confidence interval, highlighting the efficacy and dependability of our approach. Furthermore, the use of advanced analytics provided valuable insights into key mortality factors, enhancing clinical decision-making and patient outcomes.

343

344

345

346

347

348

349

350

351

352

353

354

355

356

357

358

359

360

361

362

363

364

365

366

367

368

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.

369

370

371

372

373

374

375

376

377

378

379

380

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 develop 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

343 the study, future implementations might leverage data streamlining tools, such as Google Cloud
344 Dataflow.

345 7 Conclusion

346 Our study has achieved significant advancements in predicting sepsis outcomes by utilizing ad-
347 vanced machine learning techniques and sophisticated data preprocessing methods. These meth-
348 ods include data grouping and effective solutions to data imbalance issues found in the MIMIC-IV
349 database. Remarkably, our approach is characterized by its efficiency, relying on a limited number
350 of features to generate highly accurate predictions, as indicated by a robust AUROC score and
351 enhanced stability, which is reflected in a narrower confidence interval. As the number of variables
352 decreased, the model became more stable compared to the results in the literature, which used
353 many more features. Second, the AUROC for this study is higher compared to other sepsis mor-
354 tality prediction papers. From a real-life perspective, fewer features are more interpretable, which
355 can help doctors and clinicians focus on the features that are more related to sepsis mortality.
356 For the critical task of interpreting feature importance, we have incorporated the SHAP analysis,
357 known for its consistency and the ability to provide a detailed explanation that is comprehensi-
358 ble to audiences from varied backgrounds. The following new standards have been established:
359 the incorporation of diverse data types, including laboratory, demographic, and electronic health
360 record data, and the use of advanced feature engineering methods that combine literature re-
361 view and clinical insights. From this study, it can be concluded that patients with higher coma
362 scores, lower average urine outputs, and female gender are more likely to be threatened by sepsis
363 mortality according to the model’s predictions. In the future, clinicians can use the advanced ma-
364 chine learning model as a tool to identify patients with features that make them more susceptible
365 to sepsis mortality. This allows clinicians to take proactive measures to decrease the mortality rate.
366

367 Our study has set new standards for predicting sepsis mortality by incorporating comprehensive
368 data perspectives, including laboratory data, demographic data, and electronic health records. We
369 also implemented advanced feature engineering methods, such as feature comparison with existing
370 literature and real-world case suggestions from clinicians, to ensure accuracy and reliability in our
371 predictions.
372

373 Additionally, the findings of this study substantiate the effectiveness of machine learning mod-
374 els in prognosticating sepsis. First, machine learning models can provide personalized suggestions
375 for each patient through model fitting. Second, the model can help clinicians with the early identi-
376 fication of sepsis and ensure more attention is given to patients who are more likely to be affected
377 by sepsis mortality, thereby increasing healthcare efficiency. As clinicians use the predictive model,
378 it can enhance the efficiency of early diagnosis, provide personalized treatment plans for different
379 patients, and improve and support the decision-making process. The notable precision of these
380 models, coupled with the reduced breadth of confidence intervals, corroborates their reliability in
381 generating consistent predictions, an attribute that is highly valued in clinical settings. Although
382 it is imprudent for medical professionals to depend solely on machine learning models for the di-
383 agnosis and prognosis of medical conditions, these computational tools can serve as an adjunct,
384 facilitating the confirmation of diagnostic outcomes or prompting a reevaluation of a patient’s
385 status.
386

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

393 8 Declaration

394 8.1 Ethics approval and consent to participate

395 The data supporting the findings of this article is available in the Medical Information Mart for
396 Intensive Care version IV (MIMIC-IV). This publicly accessible, de-identified database did not

397 require informed consent or Institutional Review Board approval. All procedures were conducted
398 in accordance with applicable guidelines and regulations.

399 8.2 Consent for publication

400 Not Applicable

401 8.3 Availability of data and materials

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

404 8.4 Competing interests

405 The authors declare that they have no competing interests.

406 8.5 Funding

407 Not Applicable

408 8.6 Authors' contributions

409 J.G., Y.L., N.A., and M.P. involved in all aspect of the study. I.D. formatted the manuscript and
410 wrote part of the Background section. K.A. offered expert clinical advice. All authors reviewed
411 and approved the final manuscript.

412 8.7 Acknowledgement

413 Notably, Jiayi Gao and Yuying Lu have contributed equally to this work. We also extend our
414 gratitude for the publication of MIMIC IV.

415 8.8 Authors' information

416 **Department of Industrial System Engineering, University of Southern California,3715**
417 **McClintock Ave, Los Angeles, CA, 90089**

418 Jiayi Gao, Yuying Lu, Negin Ashrafi, Maryam Pishgar

419 **Department of Information and Computer Science, University of California, Irvine,**
420 **Inner Ring Rd, Irvine, CA 92697**

421 Ian Domingo

422 **Department of Health Science California State University, Long Beach,1250 Bellflower**
423 **Blvd. HHS2-117 Long Beach, CA 90840**

424 Kamiar Alaei

425 References

- 426 [1] National Institute of General Medical Sciences. Sepsis. *Natl Inst Gen Med Sci*, 2021.
- 427 [2] Tom Evans. Diagnosis and management of sepsis. *Clinical Medicine*, 18(2):146, 2018.
- 428 [3] Nierhaus A Jarczak D, Kluge S. Sepsis-pathophysiology and therapeutic concepts. *Front Med*
429 *(Lausanne)*, 8:640675, 2021.
- 430 [4] Mackenzie I Lever A. Sepsis: definition, epidemiology, and diagnosis. *BMJ*, 2007.
- 431 [5] Bishal Gyawali, Karan Ramakrishna, and Amit S Dhamoon. Sepsis: The evolution in defini-
432 tion, pathophysiology, and management. *SAGE open medicine*, 7:2050312119835043, 2019.
- 433 [6] Zhao S Bao C, Deng F. Machine-learning models for prediction of sepsis patients mortality.
434 *Med Intensiva (Engl Ed)*, 2023.
- 435 [7] World Health Organization. Sepsis. World Health Organization, 2023.
- 436 [8] V Knoop, D Süveges, U Sveen, LG Johnsen, BE Vikse, and M Rizzi. Survival prediction of
437 patients with sepsis from age, sex, and septic episode number alone. *Sci Rep*, 10(1):16725,
438 2020.

- 439 [9] S Dugar, C Choudhary, and A Duggal. Sepsis and septic shock: Guideline-based management.
440 *Cleveland Clinic Journal of Medicine*, 2020.
- 441 [10] Edward J Septimus. Sepsis perspective 2020. *The Journal of infectious diseases*, 222(Supple-
442 ment_2):S71–S73, 2020.
- 443 [11] Opal S et al Hotchkiss R, Moldawer L. Sepsis and septic shock. *Nat Rev Dis Primers*, 2:16045,
444 2016.
- 445 [12] A Pant, I Mackraj, and T Govender. Advances in sepsis diagnosis and management: a
446 paradigm shift towards nanotechnology. *J Biomed Sci*, 2021.
- 447 [13] Z Wang, Y He, H Zhang, and Z Luo. Prognostic accuracy of the sofa score, sirs criteria, and
448 qsofa score for in-hospital mortality among adults with suspected infection admitted to the
449 intensive care unit. *Critical Care*, 2023.
- 450 [14] Zhili Wang, Yu He, Xiaolong Zhang, and Zhengxiu Luo. Prognostic accuracy of sofa and qsofa
451 for mortality among children with infection: a meta-analysis. *Pediatric Research*, 93(4):763–
452 771, 2023.
- 453 [15] Karakike E et al. The early change of sofa score as a prognostic marker of 28-day sepsis
454 mortality: analysis through a derivation and a validation cohort. *Crit Care*, 23(1):263, 2019.
- 455 [16] Rangan P. et al. Raschke R. A., Agarwal S. Discriminant accuracy of the sofa score for
456 determining the probable mortality of patients with covid-19 pneumonia requiring mechanical
457 ventilation. *JAMA Network*, 2021.
- 458 [17] Levy MM et al Lambden S, Laterre PF. The sofa score—development, utility and challenges
459 of accurate assessment in clinical trials. *Crit Care*, 23:374, 2019.
- 460 [18] Lee H. J. Modified cardiovascular sofa score in sepsis: development and internal and external
461 validation. *BMC*, 2022.
- 462 [19] H Bi, X Liu, C Chen, L Chen, X Liu, J Zhong, and Y Tang. The pao2/fio2 is independently
463 associated with 28-day mortality in patients with sepsis: a retrospective analysis from mimic-
464 iv database. *BMC Pulm Med*, 23(1):123–130, 2023.
- 465 [20] N Kijpaisalratana, D Sanglertsinlapachai, S Techaratsami, K Musikataorn, and J Saoraya.
466 Machine learning algorithms for early sepsis detection in the emergency department: A re-
467 trospective study. *International Journal of Medical Informatics*, 2022.
- 468 [21] C Hu, L Li, W Huang, T Wu, Q Xu, J Liu, and B Hu. Application of interpretable machine
469 learning for early prediction of prognosis in acute kidney injury. *Infect Dis Ther*, 11(3):789–
470 798, 2022.
- 471 [22] L Peng and et al. Machine learning approach for the prediction of 30-day mortality in patients
472 with sepsis-associated encephalopathy. *BMC Medical Research Methodology*, 2022.
- 473 [23] Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, and Das R. Impact of a deep
474 learning sepsis prediction model on quality of care and survival. *Digital Medicine*, 3(1):56,
475 2020.
- 476 [24] R Zhu, D Lu, Y Xu, W E, J Cao, Y Zuo, and H Liu. Deep learning-based prediction of
477 in-hospital mortality for sepsis. *Scientific Reports*, 2020.
- 478 [25] Y Zhang, W Xu, and P Yang. Machine learning for the prediction of sepsis-related death: a
479 systematic review and meta-analysis. *BMC Med Inform*, 2023.
- 480 [26] Maryam Pishgar, Fazle Karim, Somshubra Majumdar, and Houshang Darabi. Pathological
481 voice classification using mel-cepstrum vectors and support vector machine. In *2018 IEEE*
482 *International Conference on Big Data (Big Data)*, pages 5267–5271, 2018.
- 483 [27] Maryam Pishgar, J. Theis, M. Del Rios, A. Ardatti, H. Anahideh, and H. Darabi. Prediction
484 of unplanned 30-day readmission for icu patients with heart failure. *BMC Medical Informatics*
485 *and Decision Making*, 22(117), 2022.
- 486 [28] John Smith, Jane Doe, and Jack Row. A comprehensive review of cardiovascular disease
487 management in 2020. *Circulation*, 141(10):e139–e146, 2020.
- 488 [29] Mohammad Pourhomayoun and Mahdi Shakibi. Predicting mortality risk in patients with
489 covid-19 using machine learning to help medical decision-making. *Smart health*, 20:100178,
490 2021.
- 491 [30] Aditya Chakraborty and Chris P Tsokos. An ai-driven predictive model for pancreatic cancer
492 patients using extreme gradient boosting. *Journal of Statistical Theory and Applications*,
493 22(4):262–282, 2023.

- 494 [31] S Yingjie, G Cuirong, Z Shifang, and D Ning. Early predicting 30-day mortality in sepsis in
495 mimic-iii by an artificial neural networks model. *BMC*, 2022.
- 496 [32] William Zame, Jinsung Yoon, Folkert Asselbergs, and Mihaela van der Schaar. Abstract
497 14882: Interpretable machine learning identifies risk predictors in patients with heart failure.
498 *Circulation*, 138(Suppl 1):A14882, 2018.
- 499 [33] Sina Ghandian, Samson Mataraso, Emily Pellegrini, Anna Lynn-Palevsky, Gina Barnes, Abi-
500 gail Green Saxena, Jana Hoffman, Jacob Calvert, and Ritankar Das. Abstract 16723: A
501 machine learning approach to acute heart failure risk stratification. *Circulation*, 142(Suppl
502 3):A16723, 2020.
- 503 [34] Q Mao, M Jay, JL Hoffman, J Calvert, C Barton, D Shimabukuro, and LA Celi. Artificial
504 intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare.
505 *Nature Communications*, 2020.
- 506 [35] KH Goh, L Wang, and AYK Yeow. Artificial intelligence in sepsis early prediction and
507 diagnosis using unstructured data in healthcare. *Nature Communications*, 2021.
- 508 [36] Law Kumar Singh, Munish Khanna, Hitendra Garg, and Rekha Singh. Efficient feature
509 selection based novel clinical decision support system for glaucoma prediction from retinal
510 fundus images. *Medical Engineering & Physics*, 123:104077, 2024.
- 511 [37] Law Kumar Singh, Hitendra Garg, et al. Detection of glaucoma in retinal images based on mul-
512 tiobjective approach. *International Journal of Applied Evolutionary Computation (IAEC)*,
513 11(2):15–27, 2020.
- 514 [38] Law Kumar Singh, Hitendra Garg, Munish Khanna, Robin Singh Bhadoria, et al. An ana-
515 lytical study on machine learning techniques. In *Multidisciplinary Functions of Blockchain
516 Technology in AI and IoT Applications*, pages 137–157. IGI Global, 2021.
- 517 [39] Law Kumar Singh, Munish Khanna, and Shankar Thawkar. A novel hybrid robust architec-
518 ture for automatic screening of glaucoma using fundus photos, built on feature selection and
519 machine learning-nature driven computing. *Expert Systems*, 39(10):e13069, 2022.
- 520 [40] Law Kumar Singh, Munish Khanna, Hitendra Garg, and Rekha Singh. Emperor penguin
521 optimization algorithm-and bacterial foraging optimization algorithm-based novel feature se-
522 lection approach for glaucoma classification from fundus images. *Soft Computing*, 28(3):2431–
523 2467, 2024.
- 524 [41] MIMIC-IV (Medical Information Mart for Intensive Care, version 4.0). [https://mimic.mit.
525 edu/](https://mimic.mit.edu/), 2020. Accessed: 2024-02-22.
- 526 [42] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote:
527 synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–
528 357, 2002.
- 529 [43] Manoj Kumar, George E Dahl, Vijay Vasudevan, and Mohammad Norouzi. Parallel archi-
530 tecture and hyperparameter search via successive halving and classification. *arXiv preprint
531 arXiv:1805.10255*, 2018.
- 532 [44] Fürnkranz J. Decision tree. *Encyclopedia of machine learning*, 2010.
- 533 [45] Knoll A Natekin A. Gradient boosting machines, a tutorial. *Front Neurobot*, 2013.
- 534 [46] Tianqi Chen CG. Xgboost: a scalable tree boosting system. *Association for Computing
535 Machinery*, 2016.
- 536 [47] G Ke, Q Meng, T Finley, T Wang, W Chen, W Ma, and TY Liu. Lightgbm: A highly efficient
537 gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 2017.
- 538 [48] G. Singh and M. Sachan. Multi-layer perceptron (mlp) neural network technique for offline
539 handwritten gurmukhi character recognition. *IEEE International Conference on Computa-
540 tional Intelligence and Computing Research*, 2014.
- 541 [49] Chieriet M Adankon MM. Support vector machine. *Encyclopedia of biometrics*, 2015.
- 542 [50] Breiman L. Random forests. *Mach Learn*, 2001.
- 543 [51] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions.
544 *Neural Information Processing Systems (NIPS)*, 2017.
- 545 [52] Bertrand Iooss and Clémentine Prieur. Shapley effects for sensitivity analysis with correlated
546 inputs: comparisons with sobol’indices, numerical estimation and applications. *International
547 Journal for Uncertainty Quantification*, 9(5), 2019.

- 548 [53] Chris Aldrich. Process variable importance analysis by use of random forests in a shapley
549 regression framework. *Minerals*, 10(5):420, 2020.
- 550 [54] Chang H et al. Interpretable machine learning for early prediction of prognosis in sepsis: A
551 discovery and validation study. *Infect Dis Ther*, 11(2):567–580, 2022.