computation

MDPI

Article

Advancements in Predictive Analytics: Machine Learning Approaches to Estimating Length of Stay and Mortality in Sepsis

Houssem Ben Khalfallah [1,2], Mariem Jelassi [1], Jacques Demongeot [2,*] and Narjès Bellamine Ben Saoud [1]

[1] RIADI Laboratory, Ecole Nationale des Sciences de l'Informatique, Manouba University, La Manouba 2010, Tunisia; houssem.ben-khalfallah@univ-grenoble-alpes.fr (H.B.K.); mariem.jelassi@ensi-uma.tn (M.J.); narjes.bellamine@ensi.rnu.tn (N.B.B.S.)
[2] AGEIS Laboratory, University Grenoble Alpes, 38700 La Tronche, France
* Correspondence: jacques.demongeot@univ-grenoble-alpes.fr; Tel.: +33-615960440

Abstract: Sepsis remains a major global health concern, causing high mortality rates, prolonged hospital stays, and substantial economic burdens. The accurate prediction of clinical outcomes, such as mortality and length of stay (LOS), is critical for optimizing hospital resource allocation and improving patient management. The present study investigates the potential of machine learning (ML) models to predict these outcomes using a dataset of 1492 sepsis patients with clinical, physiological, and demographic features. After rigorous preprocessing to address missing data and ensure consistency, multiple classifiers, including Random Forest, Extra Trees, and Gradient Boosting, were trained and validated. The results demonstrate that Random Forest and Extra Trees achieve high accuracy for LOS prediction, while Gradient Boosting and Bernoulli Naïve Bayes effectively predict mortality. Feature importance analysis identified ICU stay duration (ICU_DAYS_OBS) as the most influential predictor for both outcomes, alongside vital signs, white blood cell counts, and lactic acid levels. These findings highlight the potential of ML-driven clinical decision support systems (CDSSs) to enhance early risk assessment, optimize ICU resource planning, and support timely interventions. Future research should refine predictive features, integrate advanced biomarkers, and validate models across larger and more diverse datasets to improve scalability and clinical impact.

Keywords: sepsis; machine learning; length of stay (LOS); mortality prediction; clinical decision support systems (CDSSs); intensive care unit (ICU); feature importance; predictive analytics; vital signs; mortality

Academic Editor: Shengkun Xie

Received: 16 November 2024
Revised: 25 December 2024
Accepted: 30 December 2024
Published: 1 January 2025

Citation: Ben Khalfallah, H.; Jelassi, M.; Demongeot, J.; Bellamine Ben Saoud, N. Advancements in Predictive Analytics: Machine Learning Approaches to Estimating Length of Stay and Mortality in Sepsis. *Computation* **2025**, *13*, 8. https://doi.org/10.3390/computation13010008

1. Introduction

Sepsis is a severe, life-threatening medical condition marked by a dysregulated response to infection, leading to organ dysfunction. It has been recognized as a major global health concern, resulting in substantial mortality and morbidity. In 2017, there were an estimated 48.9 million sepsis cases and 11 million sepsis-related deaths worldwide, representing nearly 20% of all global deaths [1]. This devastating disease not only places an immense emotional burden on families and communities, but also strains healthcare systems and economies.

Sepsis can manifest from preventable infections in community settings or from opportunistic or nosocomial (more generally healthcare-associated) infections acquired in healthcare environments like hospitals, affecting hundreds of millions of patients annually [1]. Despite established diagnostic criteria for sepsis, the disease's complex presentation and the non-specific nature of its symptoms often mimic other illnesses, making diagnosis challenging and leading to potential treatment delays, which can significantly worsen patient prognosis [2].

The repercussions of delayed sepsis diagnosis and treatment are severe, often leading to progression to septic shock, multiple organ failure, and, in extreme cases, death. Sepsis is also a significant contributor to long-term morbidity, with survivors often requiring ongoing treatment and support for complete recovery [1]. The economic implications of sepsis are substantial. It imposes a growing economic burden on healthcare systems globally, including the direct costs of sepsis management and indirect costs of the burden of illness imposed by sepsis [3].

Given these challenges, there is a crucial need for advanced diagnostic and prognostic tools for sepsis. Traditional diagnostic methods often involve laboratory tests to identify the specific pathogen causing the infection. However, these techniques often lead to treatment delays due to the time required for tests to return results [2].

The advent of artificial intelligence (AI), and particularly machine learning (ML), offers a promising solution to these challenges. ML algorithms can analyze large-scale patient datasets, identifying variables strongly associated with outcomes such as mortality and length of hospital stay. These variables can encompass a wide range of data, including demographic information, clinical data such as vital signs and laboratory values, comorbidities, and other relevant factors.

Clinical decision support systems (CDSSs) using ML algorithms could significantly aid clinicians in identifying high-risk patients, prioritizing their treatment, and monitoring their response to therapy. Predictive systems capable of forecasting sepsis outcomes can provide valuable insights to guide clinical decision-making, potentially leading to improved patient outcomes and more efficient utilization of healthcare resources.

However, despite the promising potential of ML-based CDSSs, there is a paucity of research comparing the performance of different ML models in predicting sepsis outcomes. Various ML models, including support vector machines (SVMs), XG Boost, and regression models, each have their own strengths and weaknesses, and it remains unclear which model performs best in the context of sepsis prediction. This study aims to address this gap in the literature by evaluating and comparing the performance of several ML models in predicting sepsis mortality and length of hospital stay. Our goal is to inform the future development and implementation of ML-based CDSSs for sepsis prediction, contributing to a better understanding of the predictive power of various clinical and demographic variables in sepsis outcomes. This information can be used to refine existing ML models and develop more accurate and efficient tools for sepsis prediction, ultimately leading to improved patient outcomes and a more efficient use of healthcare resources.

2. Background

2.1. Medical Background

Sepsis, a life-threatening condition caused by the body's response to an infection, presents a significant challenge for healthcare providers due to its complexity and high mortality rates. Sepsis manifests in various stages, each defined by a set of clinical criteria and laboratory results, and the transition between these stages is multifaceted and influenced by various factors, including the patient's health status, the nature of the infection, and the administered treatment [Pediatric SIRS, sepsis, and septic shock criteria].

At the onset of the condition, an individual may initially present no symptoms of Systemic Inflammatory Response Syndrome (No SIRS), which represents the baseline or normal state. Progression to Systemic Inflammatory Response Syndrome (SIRS) occurs when the individual exhibits two or more abnormal clinical readings related to temperature, heart rate, respiratory rate, or white blood cell count, one of which must be abnormal temperature or leukocyte count. It is worth noting that while SIRS can be a precursor to sepsis, it can also be triggered by other non-infectious conditions, emphasizing the need for careful clinical interpretation [Pediatric SIRS, sepsis, and septic shock criteria].

The term "Probable SIRS" (PS) is used to refer to a scenario where a patient meets some but not all of the criteria for SIRS, suggesting a possible early stage of the condition that warrants monitoring. However, more research is needed to clearly define this state and its associated criteria. When a patient with SIRS is confirmed to have a source of infection, the condition is then defined as sepsis. The presence of organ dysfunction, such as cardiovascular dysfunction, Acute Respiratory Distress Syndrome (ARDS), or two other organ dysfunctions, indicates progression to severe sepsis. Should cardiovascular dysfunction persist despite adequate fluid resuscitation, a state of septic shock is reached, representing the most severe and life-threatening stage of sepsis [Pediatric SIRS, sepsis, and septic shock criteria].

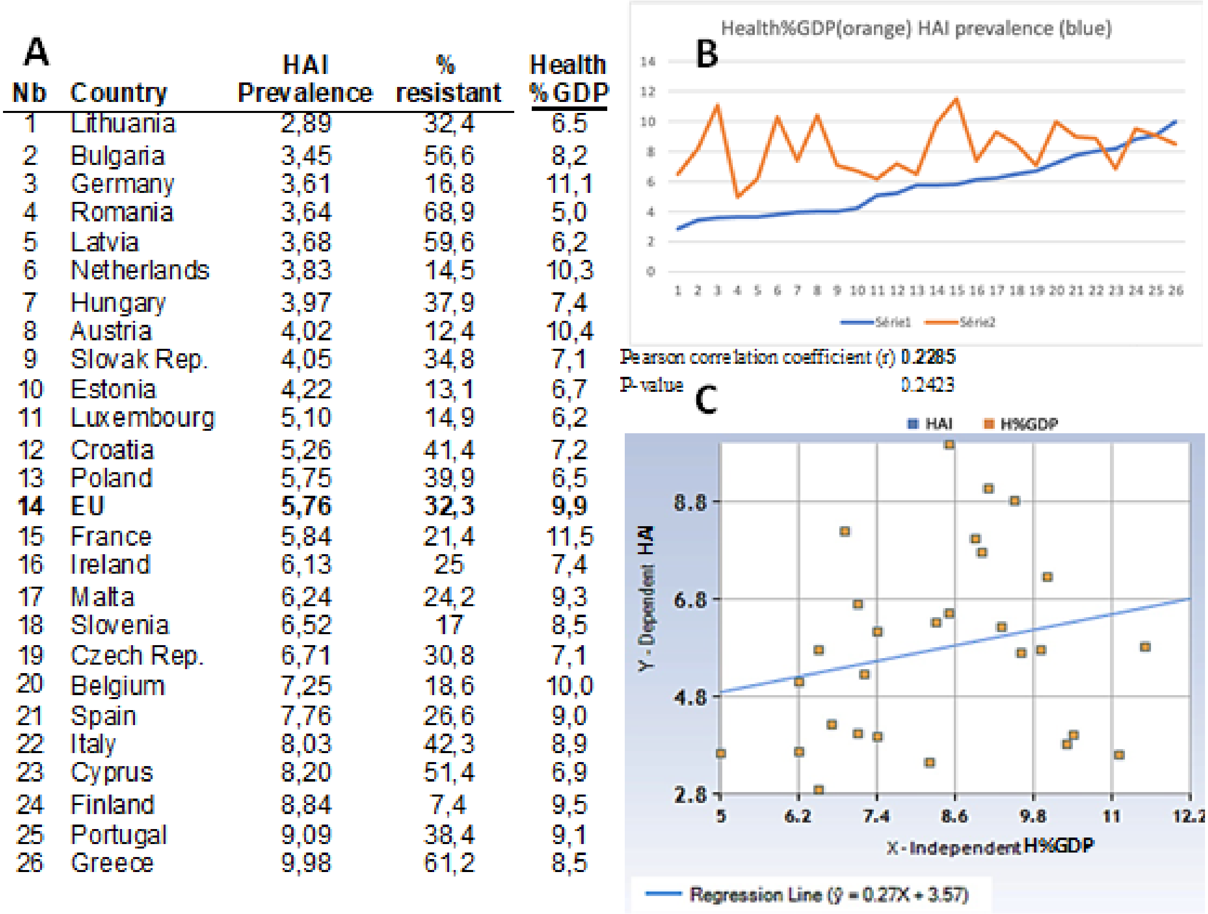
The terms "Bacteremia", "Bacteremia Probable SIRS" (BPS), and "Probable Septic Shock" (PSS) are used when the presence of bacteria in the blood is detected, which can potentially lead to sepsis if the body's immune response becomes dysregulated. BPS refers to a scenario where a patient with bacteremia exhibits symptoms suggesting the onset of SIRS. PSS refers to a situation where a patient with sepsis shows signs that could potentially progress to septic shock, such as refractory hypotension or severe organ dysfunction. Understanding these stages and transitions, explained in Table 1, is vital for the early recognition, appropriate management, and development of predictive models for sepsis, which has become a critical focus in research due to the significant morbidity, mortality, and economic burden associated with this condition. A clear delineation of these stages can also assist in stratifying patients for clinical trials, evaluating the effectiveness of interventions, and benchmarking the quality of care across different healthcare settings.

Table 1. Sepsis states and their definitions [4].

State	Feature
No SIRS	Normal state
SIRS	- Heart rate (HR) > 90 beats per minute - Respiratory rate (RR) > 20 beats per minute (or partial pressure of arterial $CO_2 < 32$) - Temperature either >38 °C or <36 °C - White blood cell count (WBC) either >12,000 or <4000 cells/mm^3 (or >bands)
Sepsis	- SIRS and Infection (blood test result)
Severe Sepsis	- Sepsis and organ failure (shown in ICL code)
Septic Shock	- Sepsis and hypotension (systolic blood pressure (SBP) is below 90 mm Hg, below 40 mm Hg, below baseline, or the mean arterial pressure (MAP) is below 70 mm Hg)
PS	- No infection—no vitals to determine SIRS
Bacteremia	- Infection (blood test result) and no SIRS
BPS	- Infection—no vitals to determine SIRS (thus sepsis)
PSS	- Hypotension, positive blood test, and no vitals to determine SIRS (thus sepsis)

2.2. Healthcare-Associated Infections and Economic Impact

Sepsis imposes a significant economic burden globally, not only due to treatment costs but also through long-term morbidity and resource consumption. In the U.S., sepsis represents a significant economic burden. A study by the Healthcare Cost and Utilization Project (HCUP) and the Agency for Healthcare Research and Quality found that, in 2013, sepsis treatment costs in hospitals reached nearly USD 24 billion, making it the most expensive condition to treat under the U.S. healthcare system. The average cost per sepsis-related hospital stay was over USD 18,000, which was 70% more expensive than the average hospital stay. In fact, hospital spending for sepsis care rose by 19% from 2011 to 2013, a rate more than double that of all hospitalizations [5]. The mean daily hospital costs in 2013 increased with sepsis severity: USD 1830 for sepsis, USD 2193 for severe sepsis, and USD 3087 for septic shock [6]. More recently, in 2021, the annual costs of sepsis were estimated at nearly USD 27 billion. Amid the COVID-19 pandemic, the financial impact of sepsis on the nation's hospitals, due to its frequent occurrence as a complication of the disease, was estimated to have exceeded USD 200 billion [7]. This financial impact could be completely justified, if it contributed to the elimination of Hospital-Associated Infections (HAIs). Unfortunately, Figure 1 shows a weak correlation between health expenditure as a percentage of GDP (H%GDP) in European Community countries and HAI prevalence and resistance to antibiotics. This justifies the present approach aimed at improving their diagnosis and prevention.



Nb	Country	HAI Prevalence	% resistant	Health %GDP
1	Lithuania	2,89	32,4	6.5
2	Bulgaria	3,45	56,6	8,2
3	Germany	3,61	16,8	11,1
4	Romania	3,64	68,9	5,0
5	Latvia	3,68	59,6	6,2
6	Netherlands	3,83	14,5	10,3
7	Hungary	3,97	37,9	7,4
8	Austria	4,02	12,4	10,4
9	Slovak Rep.	4,05	34,8	7,1
10	Estonia	4,22	13,1	6,7
11	Luxembourg	5,10	14,9	6,2
12	Croatia	5,26	41,4	7,2
13	Poland	5,75	39,9	6,5
14	**EU**	**5,76**	**32,3**	**9,9**
15	France	5,84	21,4	11,5
16	Ireland	6,13	25	7,4
17	Malta	6,24	24,2	9,3
18	Slovenia	6,52	17	8,5
19	Czech Rep.	6,71	30,8	7,1
20	Belgium	7,25	18,6	10,0
21	Spain	7,76	26,6	9,0
22	Italy	8,03	42,3	8,9
23	Cyprus	8,20	51,4	6,9
24	Finland	8,84	7,4	9,5
25	Portugal	9,09	38,4	9,1
26	Greece	9,98	61,2	8,5

Figure 1. Correlation study between Hospital-Associated Infections (HAIs, with their percentage of resistance to antibiotics) and health expenditure as a percentage of GDP (H%GDP) in European Community countries (data from [8]).

2.3. Technical Background

The application of machine learning in sepsis study and prediction involves the use of algorithms to analyze large and complex datasets, identify patterns, and make predictions about future outcomes. After the first attempts to perform medical decision-making with expert systems [9,10], machine learning models have been increasingly utilized in healthcare research for their ability to efficiently analyze vast quantities of data and their potential for predictive accuracy [11,12]. One of the primary uses of machine learning in sepsis study is for mortality prediction. Machine learning models can be trained to recognize patterns in patient data that indicate a higher risk of mortality due to sepsis. Various types of machine learning models have been employed for this purpose, including logistic regression, support vector machines, and deep learning models such as convolutional neural networks and recurrent neural networks. These models can utilize a range of patient data, including demographic information, comorbidities, vital signs, and laboratory test results, to make their predictions. Machine learning models for mortality prediction can serve several important functions in the management of sepsis. Firstly, they can assist in the early identification of patients at high risk of mortality, allowing for timely intervention and potentially improving patient outcomes. Secondly, they can provide valuable information for decision-making in clinical practice, such as guiding the allocation of resources and the selection of treatment strategies [13]. In addition to mortality prediction, machine learning models have also been used to predict length of stay for patients with sepsis. Length of stay is an important outcome measure in healthcare, as it is associated with the severity of illness, the effectiveness of treatment, and healthcare costs. Machine learning models can analyze patterns in patient data to predict length of stay, helping to optimize resource allocation and care planning [13]. Despite the potential of machine learning models for sepsis study and prediction, there are several challenges and limitations that need to be addressed. The performance of these models can be affected by the quality and completeness of the data, the selection of relevant features, the choice of the machine learning algorithm, and the model's ability to generalize to different patient populations. Furthermore, the interpretability of machine learning models is often limited, which can make it difficult for clinicians to understand and trust the model's predictions [11].

3. Methodology

3.1. Data Sources Description

The meticulous selection of a robust and representative dataset is a pivotal step in the development of an accurate and reliable machine learning model, especially when it comes to predicting complex medical conditions such as sepsis. The chosen dataset should ideally encompass a considerable number of sepsis patients, and incorporate relevant clinical variables that could influence the predictive model. Such variables can include, but are not limited to, vital signs, laboratory results, and results from various imaging studies (Figure 2). It is also of great importance that the dataset embodies the heterogeneity of the target population, thereby capturing a wide array of patient characteristics and diverse clinical outcomes. This ensures that the derived model is generalizable and can cater to a broad spectrum of potential patients.

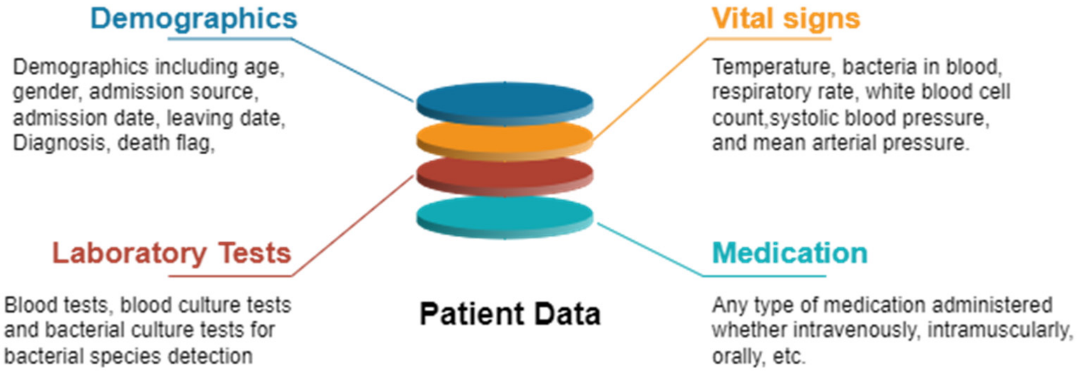


Figure 2. Patient data structure.

For our analyses, we employed a database composed of Electronic Health Records (EHRs) from the University of California Davis Health System (UCDHS). The database included records from 1492 adult patients (aged 18 years and above) who have been admitted either to emergency rooms or to the intensive care unit (ICU). All personal health information within the records was anonymized to preserve patient privacy. Prior to inclusion in this study and subsequent analyses, informed consent was obtained from all subjects or their legally authorized representatives. The research protocol was thoroughly reviewed and approved by the institutional review board of the University of California, Davis.

The patient cohort comprised a balanced gender distribution with 45.0% female patients. The average length of hospital stay was 17.0 days (SD = 36.7 days), which reflects substantial variability in hospitalization duration, including a subset of patients with prolonged stays. Notably, 38.0% of the patients were admitted to the ICU directly from the emergency department, reflecting the acuity and severity of their conditions at the time of admission. The dataset is composed of 74.8% recovered patients, and 25.2% succumbed to sepsis death. All data were abstracted retrospectively from the EHRs via structured query language interrogation of a de-identified relational database. Patients were included in the database if they were hospitalized and discharged between 1 January 2010 and 31 December 2010. The following six variables were used as observation variables in our model: temperature, respiratory rate (RR), white blood count (WBC), mean arterial pressure (MAP), systolic blood pressure (SBP), and blood culture results, with explicit mention of the bacterial species that were present in the culture. The first five variables are measurements of the patient's condition recorded over time: temperature, RR, WBC, and MAP are part of the SIRS criteria. A patient's state evolves during his hospital stay, thereby he can have several sepsis states. The heterogeneity of patient states and their vital signs distribution are, respectively, represented by Figures 3 and 4. The comprehensive nature of this dataset provides a solid foundation for developing a machine learning model aimed at predicting the patient's length of stay as well as mortality prediction, thereby contributing to timely and effective patient management.

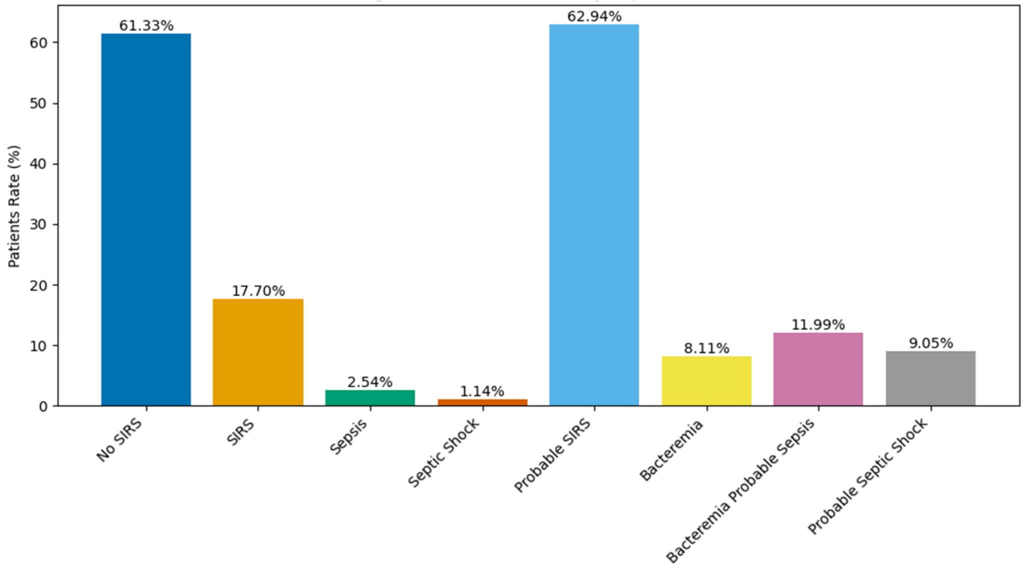


Figure 3. Rate of patients by sepsis state.

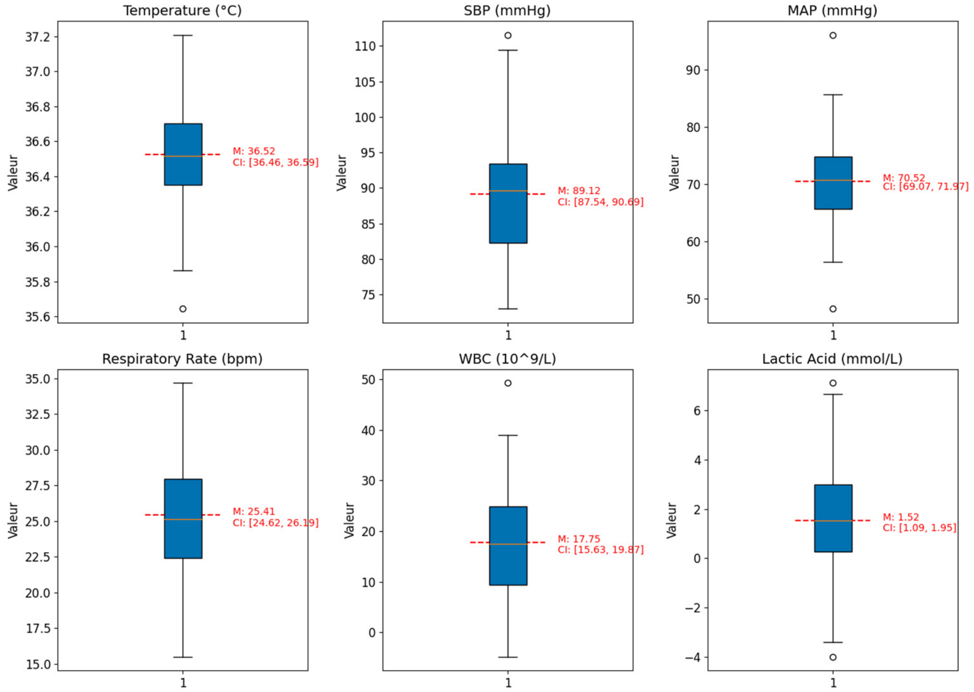


Figure 4. Vital sign distribution.

3.2. Data Preprocessing

The preprocessing stage is a critical aspect in the development of predictive models, particularly in healthcare applications where raw data are often incomplete, noisy, or inconsistent. This process generally entails data cleaning, the imputation of missing values, and normalization to ensure homogeneity across different measures. In our research, we employed the Electronic Health Records (EHRs) of 1492 patients diagnosed with sepsis to construct a decision support tool from predictive classifiers for patient outcomes and length of stay.

These data included biomarker measurements along with their respective sampling times, patient admission and discharge times, and outcomes (survived or deceased). For the purpose of this study, we considered data samples from all days of each patient's hospital stay for model training and testing.

Each patient's data comprised multiple entries corresponding to readings taken on different days, with some days featuring several readings recorded at various times. We amalgamated all entries pertaining to the same day into a single data point, representing a unique day for each patient, e.g., we averaged the body temperature if it has been measured several times during a day in order to retain only the mean body temperature of the day, and that for each patient. For those days with fewer recorded features at different times (from one to three), we filled the missing data with the most recent recorded value from that specific day. This step was crucial as it enabled the model to learn the features that predict mortality as early as possible.

3.3. Machine Learning Pipeline

Figure 5 depicts the overall pipeline used in this study for performing the mortality and LOS prediction task. All the models have been trained on samples from all days included in this dataset for mortality prediction, irrespective of the number of days to outcome, but for the case of the length of stay prediction, we first split the patients in equally sized bins based on their LOS distribution, thus maximizing entropy and avoiding bias for the training data. In order to predict survivability results given a patient's vital signs, classification methods were used, taking into account five features (temperature, respiratory rate, WBC, MAP, and lactate levels) and the mortality state of each patient, for both binary and multi-class classification. We evaluated the performance of the classifiers by performing cross-validation (5-fold cross validation) and calculating the Receiver-Operator Characteristic (ROC) curves, the Area Under the Curve (AUC), accuracy, balanced accuracy, and F1 score. The optimum combination of features thus obtained was then used to train various machine learning classification models. The trained models were then tested based on three different methods, with each having its own strengths. Five-fold cross validation was utilized to assess the predictive ability and statistical significance of the models. Assessment of the developed models was performed based on different metrics, whose mean and standard deviation are reported in Figure 5. A further detailed account of the step-by-step procedure is presented below.

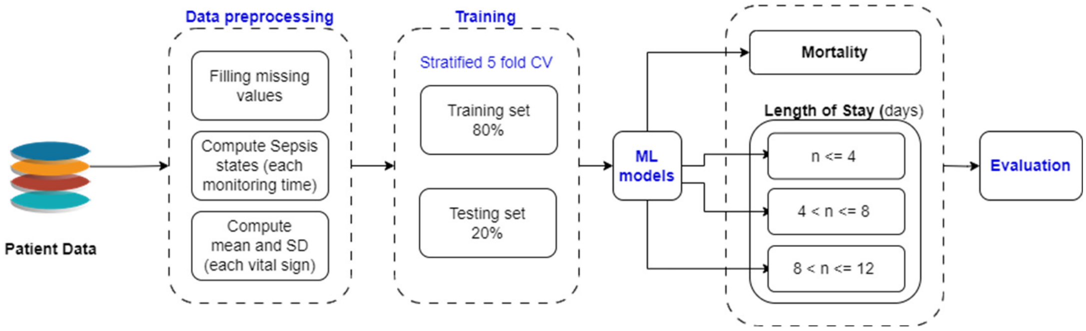


Figure 5. Machine learning process.

3.4. Training Models

Our research aimed to demonstrate the potential for leveraging machine learning to predict critical clinical outcomes. Specifically, we designed models to predict patient mortality and length of hospital stay (LOS), two key indicators of patient prognosis and resource utilization.

To predict mortality, our model incorporated a set of five clinically relevant features: body temperature, respiratory rate, white blood cell count (WBC), mean arterial pressure (MAP), and lactate levels. The choice of these parameters was guided by previous studies, indicating their significant role in the pathophysiology of sepsis and their correlation with patient outcomes. The dataset, composed of these features and the final outcome for each of the 1492 patients, was summarized by calculating the mean and standard deviation (STD) for each measurement variable over the course of each patient's hospital stay.

Turning our attention to the prediction of LOS, we augmented our feature set with two additional variables: the incidence of a positive blood culture and the count of prescribed drug administrations. Both factors have been associated with prolonged hospital stays in prior studies. We then categorized patients into three groups, using various LOS thresholds (4, 8, and 12 days) derived from the median LOS in our data (10.4 days).

In terms of model validation, we employed a 5-fold cross-validation scheme, a robust method that ensures the reliability and generalizability of our models. Additionally, we investigated a multi-class classification strategy where we grouped the LOS into multiple bins (0–4, 5–8, 9–12, and 12+ days). This strategy produced results consistent with the binary classification approach.

Finally, our quest for model optimization led us to examine various machine learning algorithms. We sought to identify the most suitable models for both mortality and LOS prediction in sepsis patients. These investigations are expected to inform the development of reliable, high-performing predictive models, paving the way for future advancements in clinical decision support systems using classical or recent methods:

- 'CalibratedClassifierCV(LinearDiscriminantAnalysis())': Linear Discriminant Analysis (LDA) is a classifier with a linear decision boundary, generated by fitting class-conditional densities to the data and using Bayes' rule. The model fits a Gaussian density to each class. 'CalibratedClassifierCV' is a probability calibration with isotonic regression or sigmoid.
- 'CalibratedClassifierCV(QuadraticDiscriminantAnalysis())': Similar to LDA, but Quadratic Discriminant Analysis (QDA) allows for the decision boundary to be quadratic.
- 'DecisionTreeClassifier()': A decision tree classifier builds a model in the form of a tree structure, making decisions based on feature values.
- 'ExtraTreeClassifier()': An extremely randomized tree classifier, this algorithm randomizes certain decisions and thresholds more than a regular decision tree, adding additional randomness to the model.
- 'DummyClassifier()': This is a classifier that makes predictions using simple rules, and is useful as a simple baseline to compare with other (real) classifiers.
- 'RandomForestClassifier()': A forest of randomized decision trees, often yielding highly accurate predictions.
- 'BaggingClassifier()': Bagging (Bootstrap Aggregating) is a way to decrease the variance of the prediction by generating additional data for training from the dataset using combinations with repetitions to produce multisets of the same size as the original.
- 'ExtraTreesClassifier()': Extra Trees is like a Random Forest in that it builds multiple trees and splits nodes using random subsets of features, but with two key differences: it does not bootstrap observations (meaning it samples without replacement), and nodes are split on random splits, not the best splits.
- 'AdaBoostClassifier()': The core principle of AdaBoost is to fit a sequence of weak learners (i.e., models that are only slightly better than random guessing) on repeatedly modified versions of the data.
- 'GradientBoostingClassifier()': Gradient Boosting builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions.
- 'LogisticRegression()': Logistic regression, despite its name, is a linear model for classification rather than regression. It is also used in the last layer of Neural Network models.
- 'SGDClassifier()': Linear classifiers (SVMs, logistic regression, etc.) with SGD training. SGD stands for Stochastic Gradient Descent: the gradient of the loss is estimated one sample at a time, and the model is updated along the way with a decreasing strength schedule (also known as the learning rate).
- 'RidgeClassifier()': A classifier using Ridge regression. This classifier first converts the target values into $\{-1, 1\}$ and then treats the problem as a regression task (multi-output regression in the multi-class case).
- 'RidgeClassifierCV()': Ridge classifier with built-in cross-validation.
- 'PassiveAggressiveClassifier()': Passive Aggressive algorithms are online learning algorithms. Such an algorithm remains passive for a correct classification outcome, and turns aggressive in the event of a miscalculation, updating and adjusting.
- 'LogisticRegressionCV()': Logistic regression CV (Cross-validation) classifier.
- 'Perceptron()': The Perceptron is another simple classification algorithm suitable for large-scale learning.
- 'BernoulliNB()': Naive Bayes classifier for multivariate Bernoulli models.
- 'GaussianNB()': Gaussian Naive Bayes (GaussianNB) classifier. Can perform online updates to model parameters via the 'partial_fit' method (for details on algorithms

used to update feature means and variance online, see Stanford CS tech report STAN-CS-79-773).
- 'KNeighborsClassifier()': Classifier implementing the k-nearest neighbors vote.
- 'NearestCentroid()': Nearest Centroid is a simple algorithm representing each class by the centroid of its members. It also has no parameters to choose, making it a good baseline classifier.
- 'MLPClassifier()': Multi-layer Perceptron classifier. This model optimizes the log-loss function using LBFGS or stochastic gradient descent.
- 'LabelPropagation()': Label Propagation classifier is a semi-supervised learning method that propagates labels from the labeled to the unlabeled data.
- 'SVC(probability = True)': C-Support Vector Classification. Its implementation is based on libsvm. The fit time complexity is more than quadratic with the number of samples, which makes it hard to scale to datasets with more than 20,000 samples.
- 'LinearSVC()': Linear Support Vector Classification. Similar to SVC with parameter kernel = 'linear', but implemented in terms of liblinear rather than libsvm, so it has more flexibility in its choice of penalties and loss functions and should scale better to large numbers of samples.

3.5. Model Evaluation and Comparative Analysis

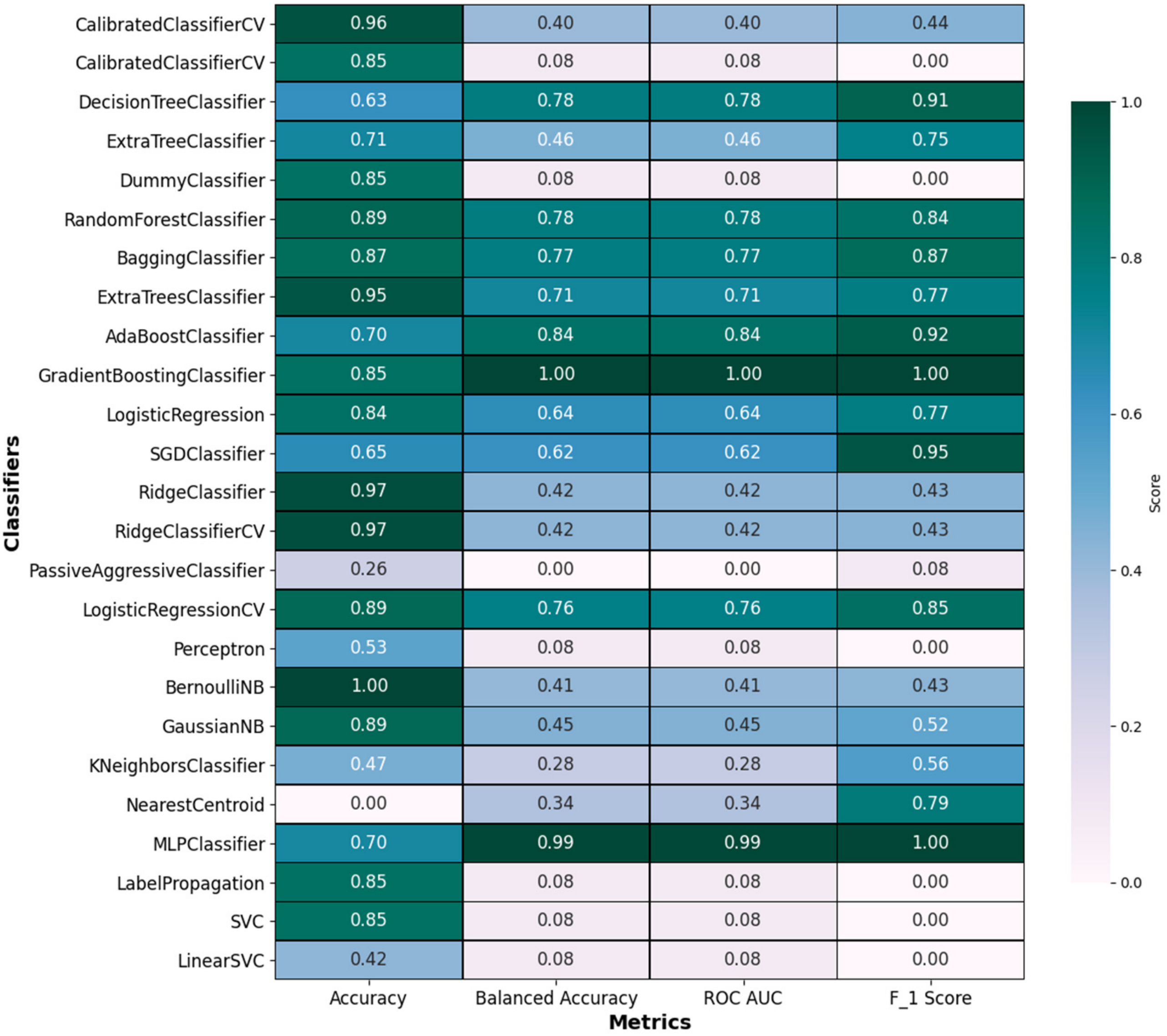
Evaluating the performance of machine learning models is a critical step in the development process, informing us about their effectiveness and potential for generalization. Several metrics are typically utilized for this purpose, including accuracy, sensitivity, specificity, F1 score, and the Receiver Operating Characteristic Area Under the Curve (ROC AUC). These metrics collectively provide a holistic view of a model's performance across multiple dimensions. ROC AUC, for example, is an aggregate measure of a model's performance across all possible classification thresholds. The F1 score is a harmonious balance between precision and recall, providing an understanding of how well the model maintains this balance. Balanced accuracy is especially useful when dealing with unbalanced class instances, as it relies on both sensitivity (true-positive rate) and specificity (true-negative rate).

Comparisons were made based on the aforementioned metrics to identify the strengths and weaknesses of our machine learning model relative to these existing systems. This comparative analysis is critical to not only understand our model's effectiveness, but also to identify opportunities for further research and improvements. Our predictive modeling task involves predicting the length of stay in a hospital, broken down into three different categories: less than 4 days, between 4 and 8 days, and between 8 and 12 days. We assessed numerous classification algorithms and evaluated them based on four metrics: accuracy, balanced accuracy, ROC AUC, and F1 score.

4. Results

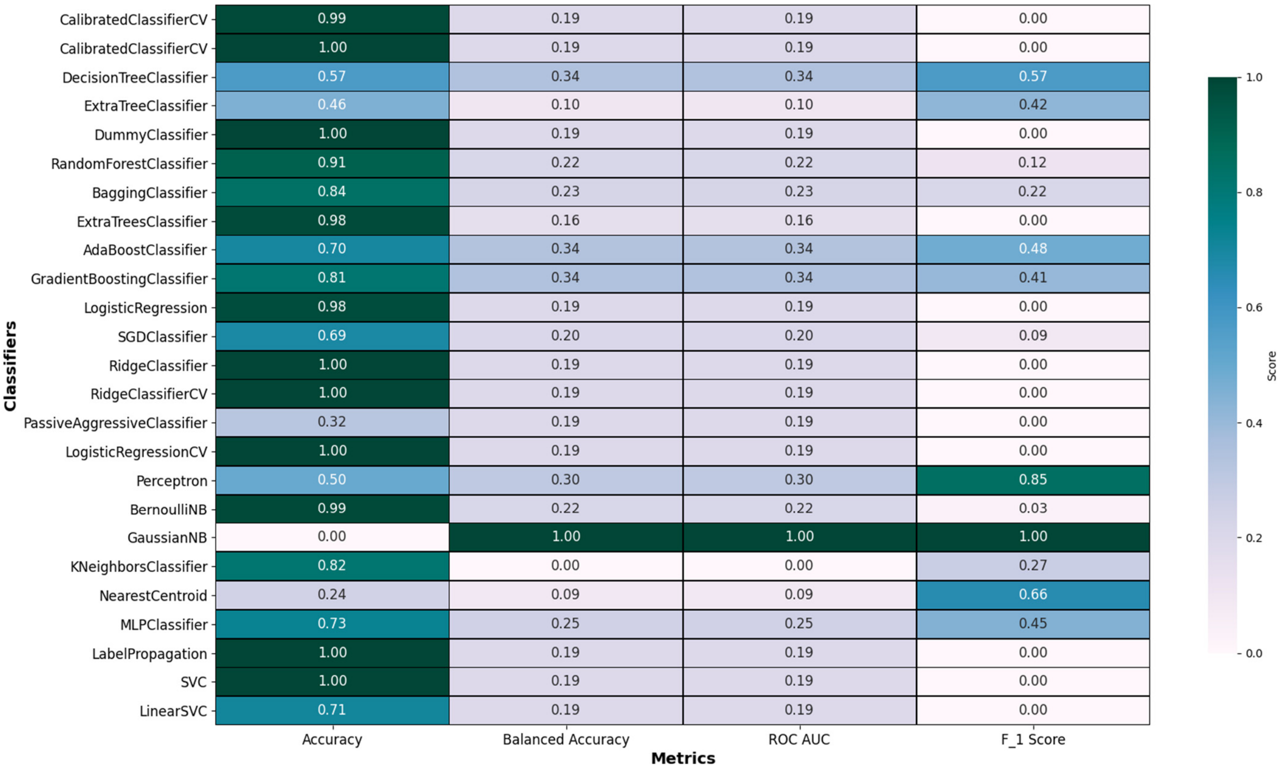
4.1. Length of Stay Prediction

For predicting stays of fewer than 4 days (Figure 6), the Gradient Boosting Classifier outperforms other classifiers in terms of the F1 score, which is 0.50, indicating a relatively balanced performance on precision and recall. This classifier also presents a high balanced accuracy and ROC AUC of approximately 0.66. This suggests it performs fairly well on both negative and positive classes, and is good at distinguishing between them. However, the overall highest accuracy was achieved by the Bernoulli NB classifier with a score of 0.73. It is noteworthy that while it achieved the highest accuracy, it did not perform as well as the other metrics. When it comes to predicting stays between 4 and 8 days (Figure 7), the performance of classifiers has diminished. The Gaussian NB classifier outperforms other classifiers in terms of balanced accuracy and ROC AUC with a score of approximately 0.64, demonstrating a decent ability to distinguish classes. The F1 score for this classifier is also the highest, at 0.51. This means that in terms of the balance between precision and recall, Gaussian NB performs best. The classifier with the highest accuracy was Logistic Regression CV with a score of approximately 0.74. However, it is important to note that while this model had high accuracy, it was not balanced in its ability to predict positive and negative outcomes. For predicting stays between 8 and 12 days (Figure 8), the Random Forest Classifier outperformed other classifiers in terms of accuracy (0.83) and F1 score (0.63), demonstrating that it has both high precision and recall. Logistic Regression also performs quite well across all metrics. For ROC AUC and balanced accuracy, the Gaussian NB classifier slightly outperforms others, with a score of approximately 0.73 (see Figures 6–8).



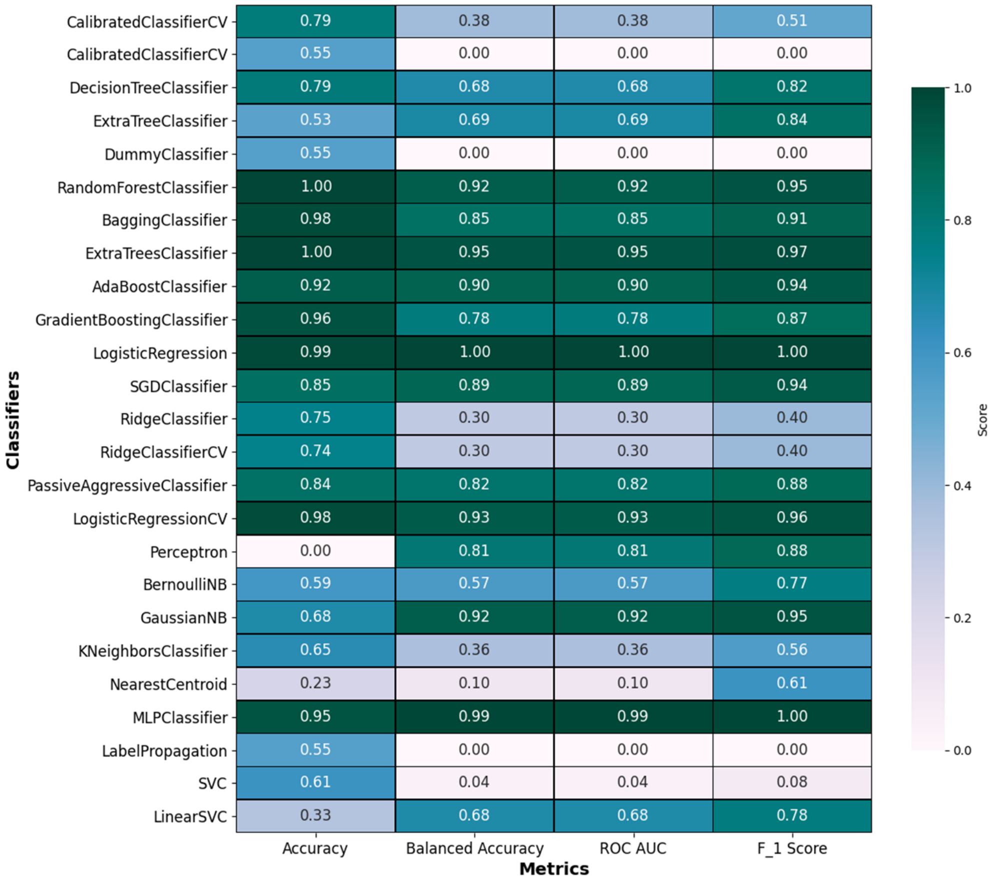
	Accuracy	Balanced Accuracy	ROC AUC	F_1 Score
CalibratedClassifierCV	0.96	0.40	0.40	0.44
CalibratedClassifierCV	0.85	0.08	0.08	0.00
DecisionTreeClassifier	0.63	0.78	0.78	0.91
ExtraTreeClassifier	0.71	0.46	0.46	0.75
DummyClassifier	0.85	0.08	0.08	0.00
RandomForestClassifier	0.89	0.78	0.78	0.84
BaggingClassifier	0.87	0.77	0.77	0.87
ExtraTreesClassifier	0.95	0.71	0.71	0.77
AdaBoostClassifier	0.70	0.84	0.84	0.92
GradientBoostingClassifier	0.85	1.00	1.00	1.00
LogisticRegression	0.84	0.64	0.64	0.77
SGDClassifier	0.65	0.62	0.62	0.95
RidgeClassifier	0.97	0.42	0.42	0.43
RidgeClassifierCV	0.97	0.42	0.42	0.43
PassiveAggressiveClassifier	0.26	0.00	0.00	0.08
LogisticRegressionCV	0.89	0.76	0.76	0.85
Perceptron	0.53	0.08	0.08	0.00
BernoulliNB	1.00	0.41	0.41	0.43
GaussianNB	0.89	0.45	0.45	0.52
KNeighborsClassifier	0.47	0.28	0.28	0.56
NearestCentroid	0.00	0.34	0.34	0.79
MLPClassifier	0.70	0.99	0.99	1.00
LabelPropagation	0.85	0.08	0.08	0.00
SVC	0.85	0.08	0.08	0.00
LinearSVC	0.42	0.08	0.08	0.00

Figure 6. Heatmap of model performance metrics for length of stay prediction less than 4 days.



Classifiers	Accuracy	Balanced Accuracy	ROC AUC	F_1 Score
CalibratedClassifierCV	0.99	0.19	0.19	0.00
CalibratedClassifierCV	1.00	0.19	0.19	0.00
DecisionTreeClassifier	0.57	0.34	0.34	0.57
ExtraTreeClassifier	0.46	0.10	0.10	0.42
DummyClassifier	1.00	0.19	0.19	0.00
RandomForestClassifier	0.91	0.22	0.22	0.12
BaggingClassifier	0.84	0.23	0.23	0.22
ExtraTreesClassifier	0.98	0.16	0.16	0.00
AdaBoostClassifier	0.70	0.34	0.34	0.48
GradientBoostingClassifier	0.81	0.34	0.34	0.41
LogisticRegression	0.98	0.19	0.19	0.00
SGDClassifier	0.69	0.20	0.20	0.09
RidgeClassifier	1.00	0.19	0.19	0.00
RidgeClassifierCV	1.00	0.19	0.19	0.00
PassiveAggressiveClassifier	0.32	0.19	0.19	0.00
LogisticRegressionCV	1.00	0.19	0.19	0.00
Perceptron	0.50	0.30	0.30	0.85
BernoulliNB	0.99	0.22	0.22	0.03
GaussianNB	0.00	1.00	1.00	1.00
KNeighborsClassifier	0.82	0.00	0.00	0.27
NearestCentroid	0.24	0.09	0.09	0.66
MLPClassifier	0.73	0.25	0.25	0.45
LabelPropagation	1.00	0.19	0.19	0.00
SVC	1.00	0.19	0.19	0.00
LinearSVC	0.71	0.19	0.19	0.00

Figure 7. Heatmap of model performance metrics for length of stay prediction for n between 4 and 8 days.



Classifiers	Accuracy	Balanced Accuracy	ROC AUC	F_1 Score
CalibratedClassifierCV	0.79	0.38	0.38	0.51
CalibratedClassifierCV	0.55	0.00	0.00	0.00
DecisionTreeClassifier	0.79	0.68	0.68	0.82
ExtraTreeClassifier	0.53	0.69	0.69	0.84
DummyClassifier	0.55	0.00	0.00	0.00
RandomForestClassifier	1.00	0.92	0.92	0.95
BaggingClassifier	0.98	0.85	0.85	0.91
ExtraTreesClassifier	1.00	0.95	0.95	0.97
AdaBoostClassifier	0.92	0.90	0.90	0.94
GradientBoostingClassifier	0.96	0.78	0.78	0.87
LogisticRegression	0.99	1.00	1.00	1.00
SGDClassifier	0.85	0.89	0.89	0.94
RidgeClassifier	0.75	0.30	0.30	0.40
RidgeClassifierCV	0.74	0.30	0.30	0.40
PassiveAggressiveClassifier	0.84	0.82	0.82	0.88
LogisticRegressionCV	0.98	0.93	0.93	0.96
Perceptron	0.00	0.81	0.81	0.88
BernoulliNB	0.59	0.57	0.57	0.77
GaussianNB	0.68	0.92	0.92	0.95
KNeighborsClassifier	0.65	0.36	0.36	0.56
NearestCentroid	0.23	0.10	0.10	0.61
MLPClassifier	0.95	0.99	0.99	1.00
LabelPropagation	0.55	0.00	0.00	0.00
SVC	0.61	0.04	0.04	0.08
LinearSVC	0.33	0.68	0.68	0.78

Figure 8. Heatmap of model performance metrics for length of stay prediction for n between 8 and 12 days.

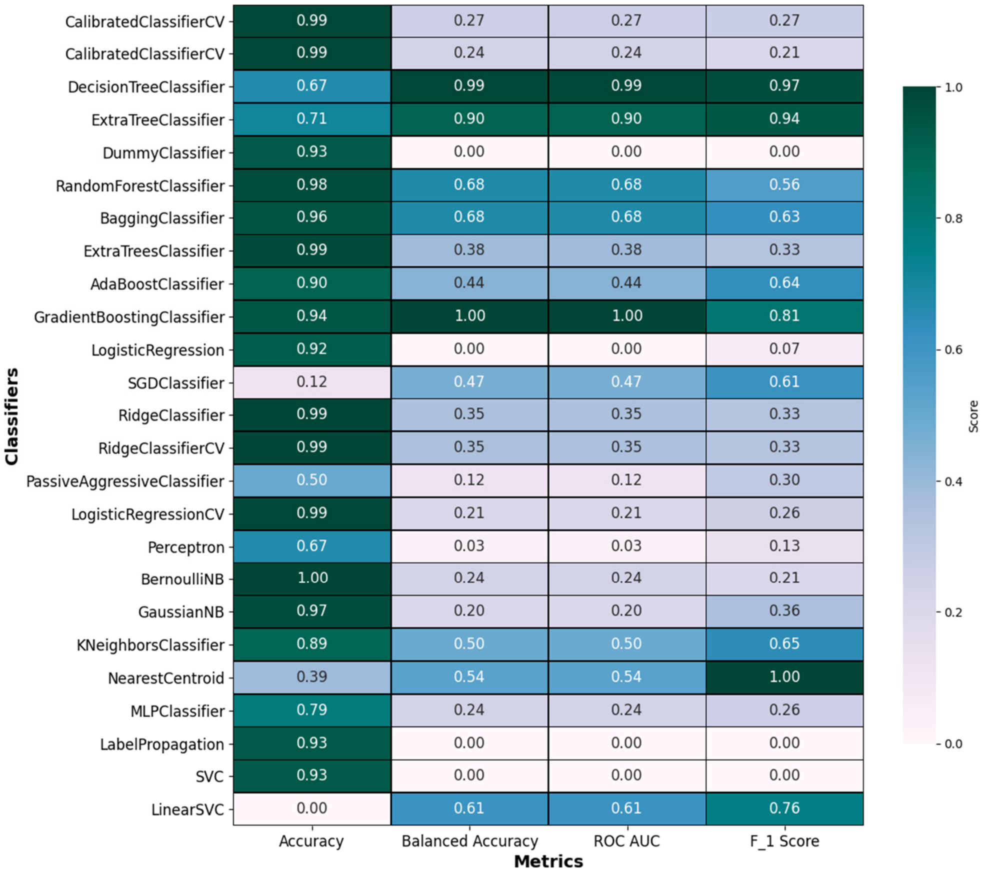
4.2. Mortality Prediction

Among the different classification models used (Figure 8), Bernoulli NB achieved the highest accuracy of 0.776, closely followed by Ridge Classifier and Ridge Classifier CV, both with accuracies of 0.774. In contrast, Linear SVC yielded the lowest accuracy score of 0.37.

When considering balanced accuracy, an essential metric for imbalanced classes, the models demonstrated more varied performance. Nearest Centroid achieved the highest balanced accuracy of 0.54, albeit with a modest overall accuracy of 0.528. This was followed by Linear SVC and the SGD Classifier with balanced accuracies of 0.546 and 0.535, respectively.

For the ROC AUC score, which assesses the tradeoff between true-positive rate and false-positive rate, Nearest Centroid once again achieved the highest score of 0.54. Linear SVC and the SGD Classifier followed closely, both yielding ROC AUC scores above 0.53.

The F1 score, measuring the balance between precision and recall, was led by Nearest Centroid (0.365), Gradient Boosting Classifier (0.297), and Decision Tree Classifier (0.354). While Nearest Centroid showed relatively modest performance in other areas, it demonstrated superior performance for balancing precision and recall (Figure 9).



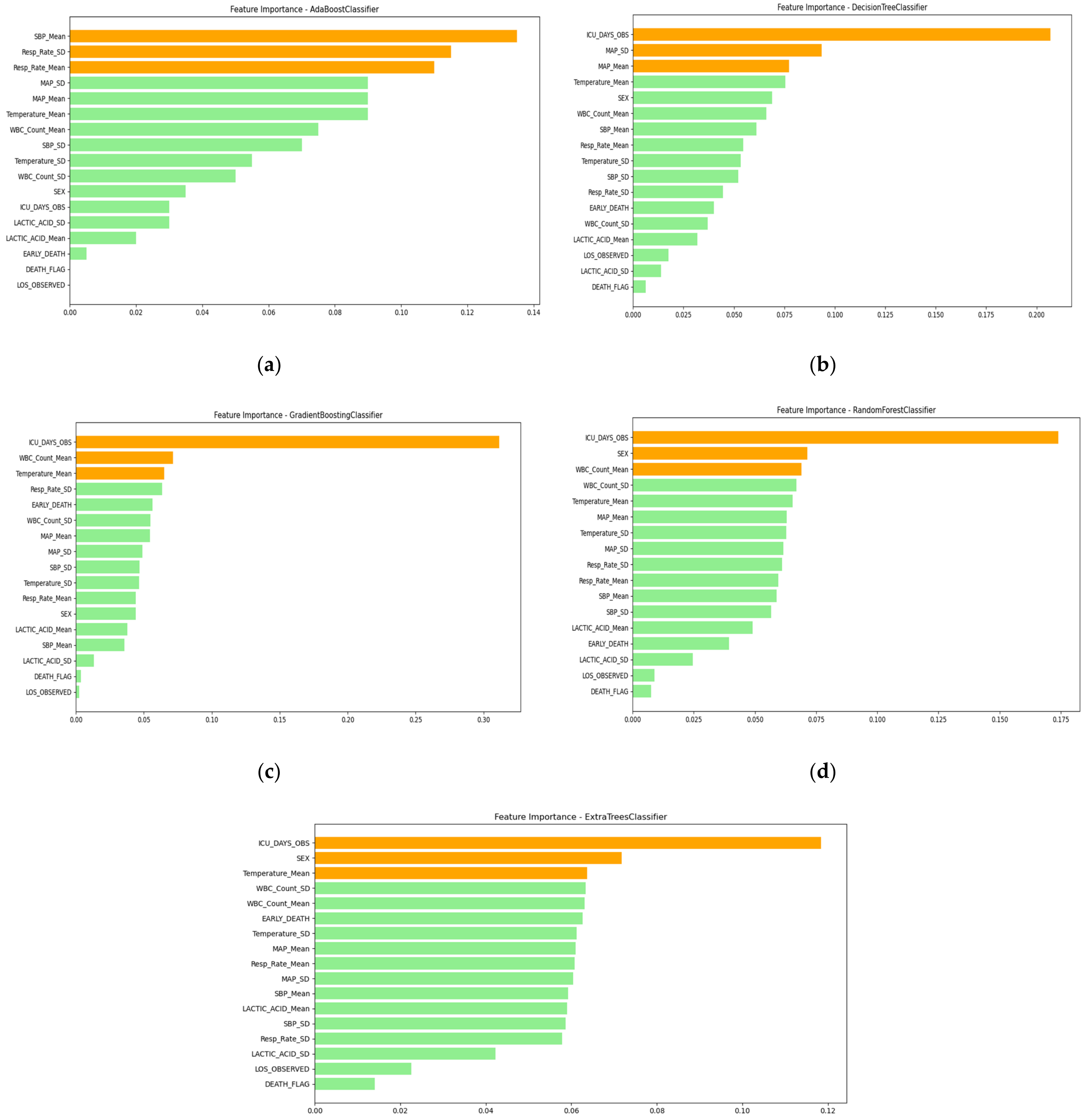
Classifiers	Accuracy	Balanced Accuracy	ROC AUC	F_1 Score
CalibratedClassifierCV	0.99	0.27	0.27	0.27
CalibratedClassifierCV	0.99	0.24	0.24	0.21
DecisionTreeClassifier	0.67	0.99	0.99	0.97
ExtraTreeClassifier	0.71	0.90	0.90	0.94
DummyClassifier	0.93	0.00	0.00	0.00
RandomForestClassifier	0.98	0.68	0.68	0.56
BaggingClassifier	0.96	0.68	0.68	0.63
ExtraTreesClassifier	0.99	0.38	0.38	0.33
AdaBoostClassifier	0.90	0.44	0.44	0.64
GradientBoostingClassifier	0.94	1.00	1.00	0.81
LogisticRegression	0.92	0.00	0.00	0.07
SGDClassifier	0.12	0.47	0.47	0.61
RidgeClassifier	0.99	0.35	0.35	0.33
RidgeClassifierCV	0.99	0.35	0.35	0.33
PassiveAggressiveClassifier	0.50	0.12	0.12	0.30
LogisticRegressionCV	0.99	0.21	0.21	0.26
Perceptron	0.67	0.03	0.03	0.13
BernoulliNB	1.00	0.24	0.24	0.21
GaussianNB	0.97	0.20	0.20	0.36
KNeighborsClassifier	0.89	0.50	0.50	0.65
NearestCentroid	0.39	0.54	0.54	1.00
MLPClassifier	0.79	0.24	0.24	0.26
LabelPropagation	0.93	0.00	0.00	0.00
SVC	0.93	0.00	0.00	0.00
LinearSVC	0.00	0.61	0.61	0.76

Figure 9. Heatmap of model performance metrics for mortality prediction.

From this analysis, it is evident that no single model was the best across all metrics. Each classifier demonstrated its own strengths and weaknesses that may be relevant depending on the specific clinical context and the relative importance of true-positive, false-positive, true-negative, and false-negative predictions. Further research is required to investigate the potential benefits of combining these models into an ensemble or adjusting their parameters for optimal performance. Additionally, the data's preprocessing and potential feature engineering might be areas worth exploring for improved model performance.

4.3. LOS Feature Importance

In this study of data for lengths of stay (LOSs) less than or equal to 4 days (Figure 10), different machine learning models have been used to analyze the importance of specific features in relation to the length of stay in a hospital setting. ICU_DAYS_OBS (number of days observed in ICU) was observed to have the highest importance across all the models, which makes sense, as a longer stay in the ICU logically relates to a longer total length of stay in the hospital, i.e., the longer someone is in the ICU, the more severe their ailments are.



(**a**) (**b**)

(**c**) (**d**)

(**e**)

Figure 10. Feature importance study for length of stay (LOS) prediction less than 4 days by various classifiers (**a**) adaboost; (**b**) decision tree; (**c**) gradient boosting; (**d**) random forest; (**e**) extratrees. The color orange represents the first three best predictors and the green the others.

Moving away from this expected correlation, other biological factors were identified as significant contributors in predicting the length of hospital stay.

One significant finding is the importance of the feature Temperature_Mean across all models. An elevated body temperature, or fever, is often a sign of an infection or other medical condition, which could require prolonged hospitalization. Similarly, the standard deviation of the temperature (Temperature_SD), representing the variability in a patient's temperature, also showed a meaningful contribution across all models. High variability could indicate an unstable health condition, needing a longer stay.

Mean arterial pressure (MAP_Mean and MAP_SD), representing the average blood pressure of an individual, was another significant feature identified by our models. Both the mean and standard deviation of MAP were seen to be crucial across all classifiers. Blood pressure is an important indicator of cardiovascular health, and fluctuations can suggest potential health complications requiring extended hospital care.

Interestingly, sex appeared to be a significant predictor in most models, suggesting possible biological differences in how males and females respond to certain treatments or their vulnerability to certain conditions.

Features like the white blood cell (WBC) count (both mean and standard deviation) also showed a high level of importance. An elevated WBC count is a general response to infection or immune system stress, which could prolong hospital stays. Variability in the WBC count might imply recurrent or persistent infection, explaining its role in predicting the length of stay.

The standard deviation and mean values of the patient's respiratory rate and systolic blood pressure were also shown to be significant across most models. These could be indicative of the patient's condition severity and the body's response to treatment.

Interestingly, early death (EARLY_DEATH) showed significance in the Extra Trees Classifier and Gradient Boosting Classifier models. This could suggest that patients with certain high-risk characteristics that may lead to early death are likely to have longer hospital stays.

It is also important to highlight that while the models' accuracy ranged between 67% and 72%, the F1 score (a measure that considers both precision and recall) ranged between 34% and 50%, suggesting the models may be more adept at predicting true positives but less effective at identifying true negatives.

Future research should aim to refine these models further, exploring potential interactions between features and optimizing model parameters to increase overall performance. It is also necessary to validate these findings on different datasets and consider other relevant features, which may help enhance our understanding of factors contributing to the length of hospital stay.

The data for lengths of stay (LOSs) between 4 and 8 days (Appendix A Figure A1) presents a different trend when compared to LOSs of fewer than 4 days. In the Decision Tree Classifier, the top three features in terms of importance are WBC_Count_Mean, Resp_Rate_SD, and MAP_SD. Notably, WBC_Count_Mean decreased in importance compared to the previous dataset. In contrast, the role of Resp_Rate_SD slightly dropped, while the importance of MAP_SD increased. This suggests that for patients staying 4 to 8 days, these physiological indicators become more relevant. In the Random Forest Classifier, the SEX feature becomes more important, indicating that gender plays a more significant role in LOSs of 4 to 8 days. ICU_DAYS_OBS is also more critical in this model compared to the previous one, suggesting that the length of stay in the ICU might have a more substantial impact on overall LOS in the 4–8 day range. When studying Extra Trees Classifier, as well as Ada Boost Classifier, similarly to the Random Forest Classifier, the feature importance of SEX and ICU_DAYS_OBS rose, suggesting similar trends, which indicate a consistent model interpretation irrespective of the change in LOS. While in the Gradient Boosting Classifier, the importance of ICU_DAYS_OBS increased, and WBC_Count_Mean decreased. This trend is somewhat similar to what we have seen with the Decision Tree, Random Forest, and Extra Trees models.

Looking at the performance metrics, the accuracy scores for all models are generally lower for this dataset, except for Extra Trees Classifier, where accuracy is higher. This suggests that the models may be less successful in correctly classifying LOS for the 4–8-days group compared to the less-than-4-days group. Moreover, the ROC_AUC and F1 scores for these models decreased when compared to the previous dataset. The balanced accuracy scores are also relatively lower, which indicates that the models are less capable of dealing with imbalanced classes for this dataset.

In conclusion, the trend of feature importance seems to differ between the two datasets. Notably, physiological measures become more important for longer LOSs, and the sex of the patient also becomes a more crucial factor. However, these models perform less accurately and are less capable of handling class imbalance for the 4–8 day LOS dataset compared to the less-than-4-day dataset.

Analyzing the data for LOSs between 8 and 12 days (Appendix A Figure A2), we can see that the features have varying levels of importance depending on the machine learning algorithm used. However, some variables stand out as being consistently important across multiple models.

The ICU_DAYS_OBS feature continues to be the most significant feature, according to Decision Tree Classifier, Random Forest Classifier, Extra Trees Classifier, and Gradient Boosting Classifier. Ada Boost Classifier, however, puts much less emphasis on this feature.

The importance of the SEX feature fluctuates across different models, ranking high in Ada Boost Classifier and the various tree-based classifiers, yet was deemed less impactful in Gradient Boosting Classifier.

Regarding other variables, notably WBC_Count_SD, WBC_Count_Mean, and Temperature_Mean features, these demonstrate importance to a greater or lesser extent across different models. Other features, such as Resp_Rate_Mean, SBP_SD, MAP_Mean, and Resp_Rate_SD, are also seen as important by the models, but their significance levels vary.

It is crucial to remember that each algorithm works differently, leading to varying interpretations of feature importance. For example, Ada Boost's results are based on a combination of weak learners (typically decision stumps), meaning it might identify different feature importance compared to tree-based methods that involve more complex splits.

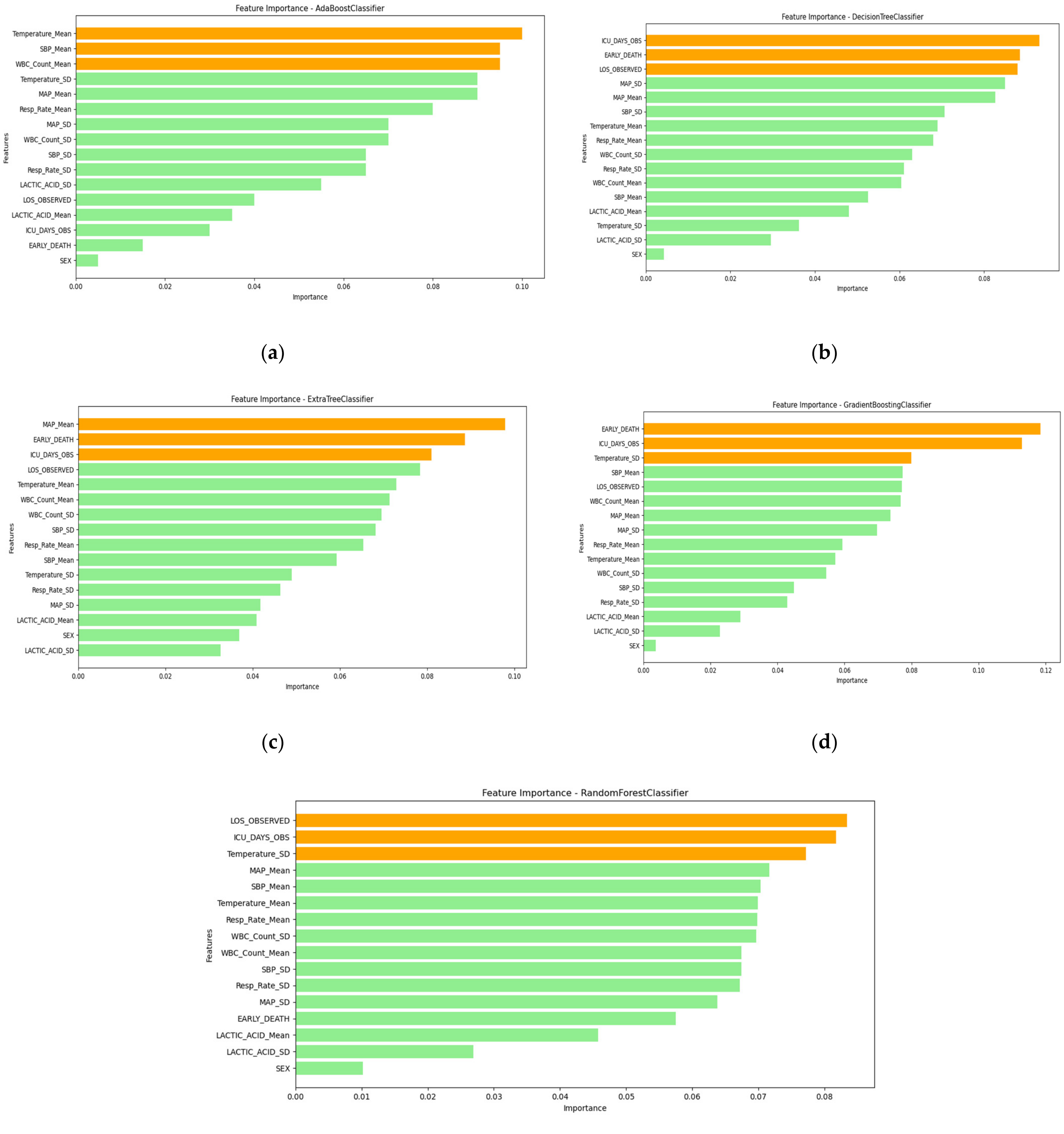
Lastly, model performance, as evaluated by accuracy, balanced accuracy, ROC_AUC, and F1 score, varies between models and is slightly better than previous data where the length of stay was up to 8 days. Random Forest Classifier and Extra Trees Classifier outperform other models in terms of accuracy, balanced accuracy, and F1 score. The ROC_AUC is highest for Extra Trees Classifier. Ada Boost Classifier and Gradient Boosting Classifier also show good results but are slightly outperformed by the tree-based models.

4.4. Mortality Feature Importance

The importance of features in the Decision Tree, Extra Trees, Random Forest, Ada Boost, and Gradient Boosting classifiers provide integral insights into the critical factors influencing the prediction of sepsis-related mortality.

Prominently, the length of the hospital stay, denoted as LOS_OBSERVED, and the duration spent in the Intensive Care Unit (ICU_DAYS_OBS), consistently surfaced as significant variables across all classifier models. This underlines the inherent importance of these time-dependent factors in the accurate prediction of sepsis mortality. These findings align with clinical observations, where a more extended hospital stay or ICU admission often correlates with a more severe disease course, thereby augmenting the risk of mortality. This highlights the complex relationship between time, disease progression, and mortality in sepsis.

Furthermore, the occurrence of early death (EARLY_DEATH) was another dominant feature identified across all models, excluding Ada Boost (Figure 11). The prominence of this variable reiterates the potential value of temporal markers in predicting severe outcomes, including mortality in sepsis.



(**a**) (**b**)

(**c**) (**d**)

(**e**)

Figure 11. Feature importance study for mortality prediction by using various classifiers (**a**) adaboost; (**b**) decision tree; (**c**) extratrees; (**d**) gradient boosting; (**e**) random forest. The color orange represents the first three best predictors and the green the others.

Among the physiological parameters analyzed, mean temperature (Temperature_Mean) and its standard deviation (Temperature_SD) consistently emerged as significant variables across all classifiers. Particularly, the Gradient Boosting Classifier underscored the role of temperature variability as an essential component, thereby highlighting the potential influence of temperature fluctuations in predicting sepsis mortality. Other critical factors included the mean Systolic Blood Pressure (SBP_Mean) and its variability (SBP_SD), Mean Arterial Pressure (MAP_Mean) and its variability (MAP_SD), and the mean Respiratory Rate (Resp_Rate_Mean) and its variability (Resp_Rate_SD). The regular appearance and substantial importance of these vital sign parameters underscore their essential role in assessing the severity and progression of sepsis, thereby reinforcing their value in prognostic modeling.

Interestingly, white blood cell count parameters (WBC_Count_Mean and WBC_Count_SD), often viewed as indicators of the body's immune response, consistently emerged as key features across all classifiers, except Ada Boost. This finding suggests a significant role for the host immune response in sepsis outcomes and the potential utility of these features in mortality prediction. Metabolic markers, such as lactic acid levels (LACTIC_ACID_Mean and LACTIC_ACID_SD), were identified as influential features by all classifiers, albeit with lower importance values. The identification of these markers suggests that metabolic alterations could play a significant part in sepsis progression and underline the role of biochemical parameters in predictive modeling. Interestingly, the gender of the patient (SEX) had minimal to no impact on the prediction of sepsis mortality, according to the investigated models. This could indicate that biological sex may not play a significant role in the outcome of sepsis, at least within the confines of the current models.

The variability observed in the ranking of feature importance across different classifiers reiterates the multifaceted nature of sepsis. It suggests that various predictive models might capture different aspects of the disease process, reflecting the complex interaction of multiple factors in sepsis pathophysiology. This complexity prompts the need for further investigation to enhance the predictive accuracy and clinical utility of these models in the management of sepsis. It is important that future works in this area consider these factors and explore their integration in a multimodal predictive framework for improved sepsis outcomes.

5. Discussion

Across all models and LOS categories, the ICU_DAYS_OBS variable consistently ranks as the most influential factor. This is a logical finding, given that ICU_DAYS_OBS likely correlates with LOS directly. Patients who spend more time in the ICU naturally have a longer overall hospital stay.

Sex is another variable that has consistently high feature importance across all LOS categories and most models. However, its ranking varies somewhat depending on the model and LOS category. For example, Ada Boost ranks SEX as the most influential variable, whereas the tree-based models rank it as second or third in importance. Certain vital signs, such as Temperature_Mean, Resp_Rate_Mean, SBP_SD, MAP_Mean, and Resp_Rate_SD, are repeatedly identified as important by the models across all LOS categories. Their ranking, however, changes with the length of stay and the model used. These fluctuations might be attributed to the complexity of the human body's response to illness and the wide-ranging impacts of different conditions on these vital signs. Moreover, both the mean and standard deviation of the white blood cell (WBC) count (WBC_Count_SD and WBC_Count_Mean) are consistently highlighted as important factors by the models across LOS categories. These variables could be indicative of the body's response to infection or inflammation, which may have a significant bearing on a patient's hospital stay duration. LACTIC_ACID_Mean and LACTIC_ACID_SD also fluctuate in their importance across the models and LOS categories. High lactic acid levels can indicate tissue hypoxia, sepsis, or liver dysfunction, among other serious conditions, which might significantly affect the length of a patient's stay. Lastly, the death flag variables (DEATH_FLAG and EARLY_DEATH) generally rank low in feature importance across the models and LOS categories. This suggests that whether a patient ultimately survives or not might be less predictive of their LOS compared to other factors. Regarding the models' performance, Random Forest Classifier and Extra Trees Classifier consistently outperform others in terms of accuracy, balanced accuracy, and F1 score. The ROC_AUC is highest for Extra Trees Classifier. Ada Boost Classifier and Gradient Boosting Classifier also show good results, but are slightly outperformed by the tree-based models. To summarize, the length of stay in the hospital is influenced by various factors, ranging from the number of days in ICU to the sex of the patient, vital signs, and certain lab results. Different machine learning models weigh these factors differently but highlight the aforementioned variables as consistently influential.

Considering mortality prediction, feature importance provides also a useful measure of the impact that each input variable has on the prediction of the model. However, to be meaningful, the assigned feature importance must be thoroughly evaluated and interpreted within the context of the model and the dataset at hand. The listed importance of the classifiers we used to predict sepsis-related mortality can be evaluated in terms of relevance, consistency, and practical applicability.

Evaluating the relevance of the features pertains to whether they make intuitive and practical sense in the context of sepsis management. Key predictors such as LOS_OBSERVED and ICU_DAYS_OBS, which represent the length of overall hospital stay and time spent in the ICU, respectively, are clinically relevant. Their significance in the models aligns with the clinical understanding that a longer duration of hospital or ICU stay might imply a more severe disease course, thereby leading to a higher likelihood of mortality. Other variables such as Temperature_Mean and blood pressure parameters, which are well-known markers of critical illness, emerge as significant features. This underscores the relevance of these factors in our models, as they reflect the physiological responses often seen in severe sepsis.

Considering consistency involves examining whether the same features are deemed important across different predictive models. Variables like ICU_DAYS_OBS, LOS_OBSERVED, and Temperature_Mean consistently appear as important predictors across various classifiers, validating their value in predicting outcomes. However, some inconsistencies are observed, which can be attributed to the distinct mechanisms and assumptions embedded in different models. For example, EARLY_DEATH is a feature with significant importance in the Gradient Boosting Classifier, while it is less emphasized in other models. This discrepancy may be due to the model's specific handling of feature interactions and non-linear relationships.

Practical applicability is another crucial dimension for evaluation. In sepsis management, it is important for the identified features to be modifiable or actionable to a certain extent. Variables such as WBC_Count_Mean or LACTIC_ACID_Mean, which can be influenced through medical interventions, stand out for their practical implications. On the other hand, while LOS_OBSERVED is an important predictor, it may not be directly actionable. Its value is potentially more applicable for risk stratification and prognosis, informing healthcare professionals about the potential trajectory of a patient's illness, rather than immediate clinical management.

Moreover, while the low importance of the SEX feature across models aligns with the principle of gender equity in healthcare, it also urges the consideration of other demographic factors such as age, ethnicity, or comorbidities, which might have been omitted from the current dataset but could influence sepsis outcomes.

In summary, the presented feature importance indeed carries meaningful information for sepsis prediction. However, careful interpretation and further validation are necessary. Future research should aim to refine these models, possibly integrating more complex or indirect markers of sepsis severity, to improve their predictive accuracy and clinical utility.

6. Future Research and Conclusions

The insights obtained from the analysis of length of stay (LOS) and sepsis-related mortality prediction provide a promising foundation for future research in the domain of healthcare and machine learning. Several potential avenues for subsequent investigation emerge from this study's findings. Firstly, a deeper exploration into the interplay of the highlighted features in determining LOS and sepsis mortality could be instrumental. While ICU_DAYS_OBS and SEX have been identified as significant predictors across LOS categories, the reasons for their varying levels of importance across different models need further examination. The exploration could also be extended to other influential variables, such as vital signs and lab results, to better understand the nature of these fluctuations and to refine our models accordingly. Additionally, future work could delve into the integration of additional demographic and clinical features that were not included in the current study, but could provide valuable predictive power. For instance, incorporating variables like age, ethnicity, and comorbidities may offer more nuanced and personalized predictions, thereby enhancing the models' clinical applicability. Moreover, the low ranking of death flag variables in predicting LOS suggests that survival may not be a primary determinant of hospital stay duration. This opens up interesting questions about what other factors might play a role, suggesting further investigations into other potential predictors of LOS are needed. Similarly, the seemingly insignificant role of SEX in predicting sepsis mortality also warrants further research. On a methodological level, comparative performances of Random Forest Classifier, Extra Trees Classifier, Ada Boost Classifier, and Gradient Boosting Classifier reveal the distinct strengths and limitations of each model. Future research could explore hybrid or ensemble models that leverage the strengths of multiple classifiers to potentially enhance predictive accuracy and robustness. Lastly, although this study provides valuable insights into predicting LOS and sepsis mortality, it is important to emphasize that machine learning models are tools that complement, rather than replace, clinical judgment. Future work should focus on how these models can be integrated effectively into clinical workflows. This involves ensuring that predictions are interpretable and actionable for healthcare professionals, thus facilitating decision-making in patient care. In conclusion, while this study has made significant strides in understanding and predicting LOS- and sepsis-related mortality, it has also uncovered new areas for further exploration. The pursuit of these research directions will contribute to the continued evolution of machine learning in healthcare, ultimately aiming for improved patient outcomes.

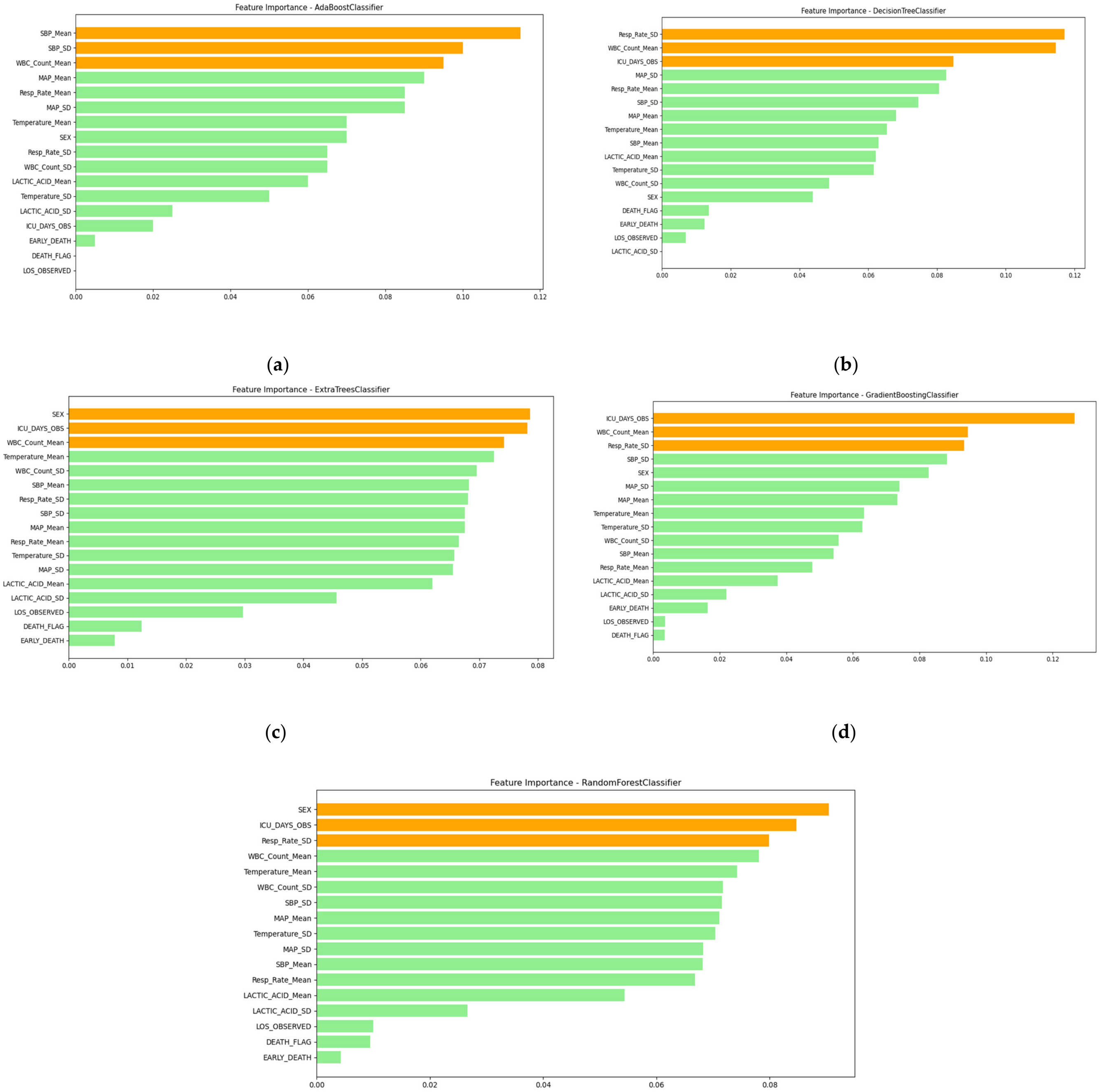
Author Contributions: Conceptualization, H.B.K., M.J., J.D. and N.B.B.S.; Methodology, H.B.K. and M.J.; software, H.B.K.; validation, H.B.K., M.J., J.D. and N.B.B.S.; formal analysis, H.B.K. and M.J.; investigation, H.B.K., M.J., J.D. and N.B.B.S.; resources, M.J. and N.B.B.S.; data curation, H.B.K. and M.J.; writing—original draft preparation, H.B.K. and M.J.; writing—review and editing, H.B.K., M.J., J.D. and N.B.B.S.; visualization, H.B.K., M.J., J.D. and N.B.B.S.; supervision, M.J.; project administration, M.J.; funding acquisition, N.B.B.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created. Data used come from the database composed of Electronic Health Records (EHRs) from the University of California Davis Health System (UCDHS) used by the study [4].

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

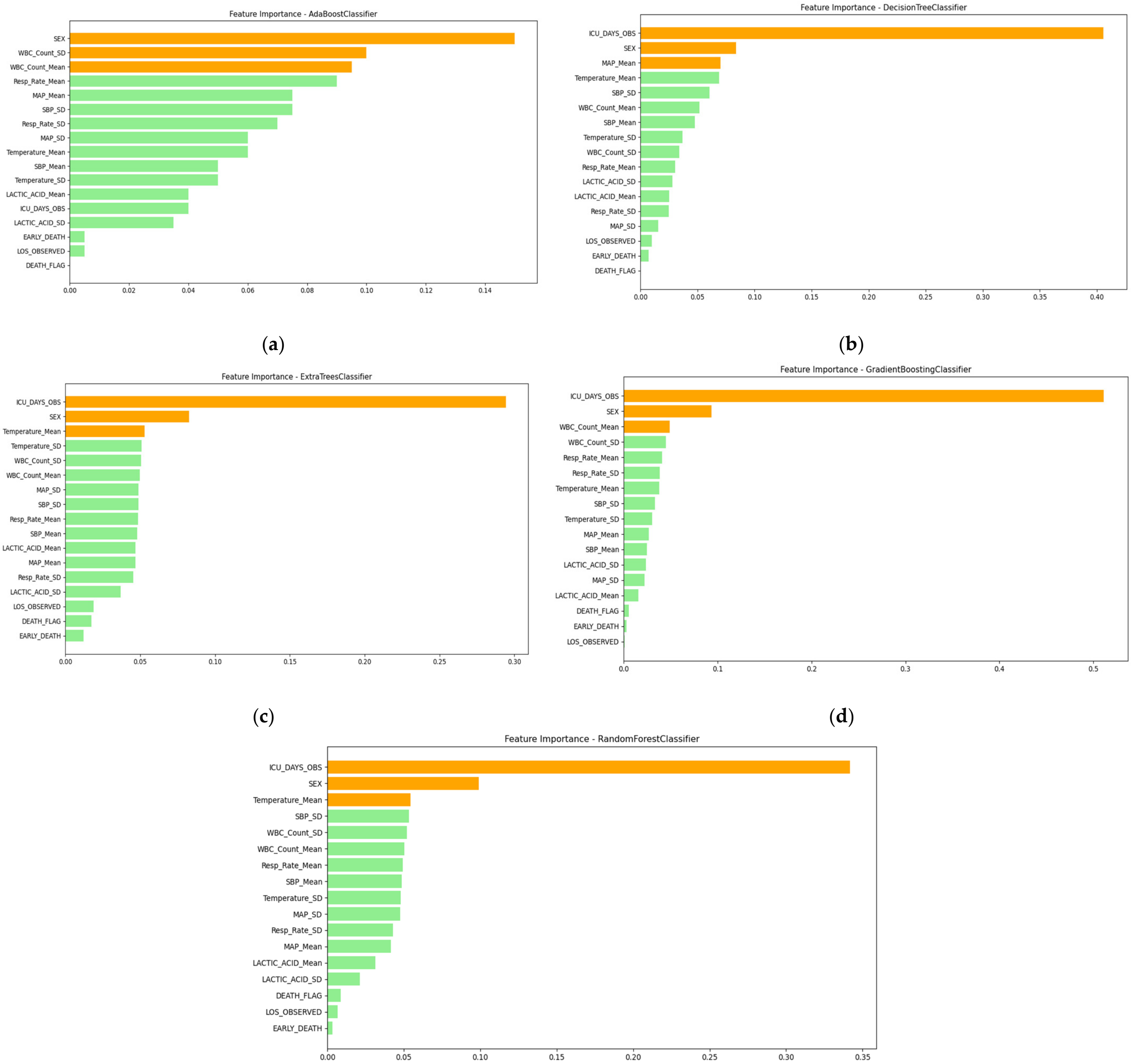


(**a**) (**b**)

(**c**) (**d**)

(**e**)

Figure A1. Feature importance study for length of stay prediction for LOS between 4 and 8 days, by using various classifiers (**a**) adaboost; (**b**) decision tree; (**c**) extratrees; (**d**) gradient boosting; (**e**) random forest. The color orange represents the first three best predictors and the green the others.



(**a**) (**b**)

(**c**) (**d**)

(**e**)

Figure A2. Feature importance study for length of stay prediction for LOS between 8 and 12 days. by using various classifiers (**a**) adaboost; (**b**) decision tree; (**c**) extratrees; (**d**) gradient boosting; (**e**) random forest. The color orange represents the first three best predictors and the green the others.

References

1. Guarino, M.; Perna, B.; Cesaro, A.E.; Maritati, M. 2023 update on sepsis and septic shock in adult patients: Management in the emergency department. *J. Clin. Med.* **2023**, *12*, 3188. [CrossRef]
2. Duncan, C.F.; Youngstein, T.; Kirrane, M.D. Diagnostic challenges in sepsis. *Curr. Infect. Dis. Rep.* **2021**, *23*, 65. [CrossRef] [PubMed]
3. Chalupka, A.N.; Talmor, D. The economics of sepsis. *Crit. Care Clin.* **2012**, *28*, 57–76. [CrossRef] [PubMed]
4. Tsoukalas, A.; Albertson, T.; Tagkopoulos, I. From data to optimal decision making: A data-driven, probabilistic machine learning approach to decision support for patients with sepsis. *JMIR Med. Inform.* **2015**, *3*, e3445. [CrossRef] [PubMed]
5. Martin, A.B.; Hartman, M.; Benson, J.; Catlin, A.; The National Health Expenditure Accounts Team. National Health Spending In 2014: Faster Growth Driven By Coverage Expansion And Prescription Drug Spending. *Health Aff.* **2016**, *35*, 150–160. [CrossRef] [PubMed]

6. Paoli, C.J.; Reynolds, M.A.; Sinha, M.; Gitlin, M.; Crouser, E. Epidemiology and Costs of Sepsis in the United States—An Analysis Based on Timing of Diagnosis and Severity Level*. *Crit. Care Med.* **2018**, *46*, 1889. [CrossRef] [PubMed]
7. Kulick, D.; Vaghela, S.; Bina, E. Sepsis Poses a Cost-Containment Challenge in the Face of the COVID-19 Pandemic. Healthcare Financial Management Association (HFMA). 20 July 2020. Available online: https://www.hfma.org/topics/financial-sustainability/article/sepsis-poses-a-cost-containment-challenge-in-the-face-of-the-cov.html (accessed on 19 July 2023).
8. European Centre for Disease Prevention and Control. Healthcare-associated infections: Surgical site infections. In *ECDC. Annual Epidemiological Report for 2016*; ECDC: Stockholm, Sweden, 2018.
9. François, P.; Crémilleux, B.; Robert, C.; Demongeot, J. Meninge: A medical consultation system for child meningitis. *Artif. Intell. Med.* **1992**, *4*, 19–30. [CrossRef]
10. François, P.; Robert, C.; Crémilleux, B.; Bucharles, C.; Demongeot, J. Variables processing in expert system building: Application to the aetiological diagnosis of infantile meningitis. *Med. Inform.* **1990**, *15*, 115–124. [CrossRef] [PubMed]
11. Fleuren, L.M.; Klausch, T.L.; Zwager, C.L.; Schoonmade, L.J.; Guo, T.; Roggeveen, L.F.; Swart, E.L.; Girbes, A.R.J.; Thoral, P.; Ercole, A.; et al. Machine learning for the prediction of sepsis: A systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Med.* **2020**, *46*, 383–400. [CrossRef] [PubMed]
12. Khalfallah, H.B.; Jelassi, M.; Demongeot, J.; Saoud, N.B.B. Decision support systems in healthcare: Systematic review, meta-analysis and prediction, with example of COVID-19. *AIMS Bioeng.* **2023**, *10*, 27–52. [CrossRef]
13. Iwase, S.; Nakada, T.A.; Shimada, T.; Oami, T.; Shimazui, T.; Takahashi, N.; Yamabe, J.; Yamao, Y.; Kawakami, E. Prediction algorithm for ICU mortality and length of stay using machine learning. *Sci. Rep.* **2022**, *12*, 12912. [CrossRef] [PubMed]