Early Prediction of Sepsis Using Machine Learning and Explainable AI

1 Introduction

SEPSIS is a **significant global health concern**, causing about 11 million deaths annually. Sepsis is responsible for approximately **20% of all global deaths**. When the body responds to an infection, the immune system releases cytokines and other inflammatory mediators. As this response aims to eliminate the pathogen, an excessive or uncontrolled reaction can cause **tissue damage, multiple organ failure**, and even death [1]. Sepsis cannot be directly detected as there is no single test to diagnose it. Doctors rely on a combination of **clinical signs, laboratory tests, and biomarkers** to assess whether a patient is septic or non-septic [2]. If these symptoms are detected earlier, this would lead to **preventative steps** to avoid death [3].

Several machine learning (ML) algorithms have been proposed to predict sepsis early, classifying septic and non-septic patients. Due to the **complexity and variability of sepsis progression**, various approaches have been attempted, such as conventional machine learning classifiers, ensemble learning methods, and deep learning models. Across multiple studies that focused on detecting sepsis using ML algorithms [4], a **common dataset** has been utilized to develop and evaluate predictive models.

The **PhysioNet/Computing in Cardiology Challenge 2019 dataset** has been one of the most widely used datasets as it provides a large, diverse, and well-structured collection of **electronic health records (EHRs)** from **intensive care units (ICUs)**. The data contains over 60,000 ICU patients, and each patient has a separate file that contains the hourly measurements of a specific patient (each row represents one-hour measurements for that patient). The data were collected from three geographically distinct U.S. hospital systems as follows:

- **Beth Israel Deaconess Medical Center (Hospital A)**
- **Emory University Hospital (Hospital B)**
- **A third unidentified hospital system (Hospital C)**

The training dataset contains 40,336 patient records from Hospitals A and B that were publicly available, while the other 24,819 patient records (from all three hospitals) were sequestered as hidden test sets for unbiased model evaluation. A common evaluation matrix used on these datasets was **Normalized Utility Score (NUS), AUROC, AUPRC, and F1 Score** [2].

One of the robust models used in sepsis detection was the **XGBoost model**. S. Lyra et al. [4] trained it on 80% of the training dataset, which included 107 features (including the original and derived features). The model was evaluated on the remaining 20% of the dataset. They validated the model on prospective data entirely unseen during the training phase. As a result, they achieved a **normalized utility score** (a widely recognized scoring methodology for sepsis detection) of 0.494 on test data and 0.378 on prospective data at threshold 0.3. Additionally, the model achieved F1 scores of 80.8% and 67.1%, respectively, for the test data and the prospective data for the same threshold [5].

As the **laboratory measurements** in the dataset provided by the PhysioNet 2019 challenge have more than **90% values missing** [6], Wu et al. [7] applied a **down-sampling process** to the retrospective data for training their XGBoost model. During the testing stage, a **dynamic sliding window** and the trained XGBoost model were used to predict sepsis on both the PhysioNet dataset and the retrospective datasets. As a result, they achieved **80.74% accuracy** (77.90% sensitivity and 84.42% specificity) and 83.95% (84.82% sensitivity and 82.00% specificity) on the test set of PhysioNet-A and PhysioNet-B, respectively. The AUC score was 0.89 for both datasets [7].

Similarly, Vandana and R. Chhikara explored various machine learning models along with feature selection approaches in order to predict sepsis accurately. Of the models used in their study, the **Decision Tree model combined with Information Gain** provided the highest performance with an accuracy of 95% across a range of evaluation metrics, making it the best-performing model of all models tested in the study with Sensitivity (Recall) 36.7% [8]. Similarly, Kong et al developed models such as **Gradient boosting machine (GBM)**

and Random Forest (RF) but GBM was the best performance since GBM is ensemble-learning model that builds sequence of decision trees which this branches used to minimize the residual errors of the previous trees in this study there are 86 predictor variables covering laboratory test results. The GBM model outperformed the other models by achieving 84.5% but the other model was from 77 to 83. However, The GMB is the lowest interpretability. So, using it gets more accuracy but low interpretability but Random forest is more interpretable[9]. Furthermore, Apalak and Kiasaleh applied the **TCNs** on **MIMIC-III** focusing on Heart Rate Variability (HRV) features connected with sepsis and the mode TCN analtes time pattern in HRV to predict sepsis. The result of the model here is 82% AUROC in predicting 6 hours for sepsis early [10].

In the effort of further analysis, Liao et al. [11] conducted an **experimental study** comparing the performance of deep learning (DL) models such as **LSTM and Temporal Convolutional Networks (TCN)**, with traditional machine learning (ML) models, including **Random Forest (RF) and XGBoost**, using the PhysioNet 2019 dataset. The results of their study showed that the DL models outperformed the traditional ML models with LSTM and TCN achieving AUROC scores (0.85–0.89), while XGBoost and RF demonstrated comparable performance with AUROC scores between 0.84 and 0.88. Although the deep learning (DL) models outperformed the machine learning (ML) models in AUROC scores, they achieved nearly similar accuracies, ranging from 0.94 to 0.957 [11].

While these models compared on PhysioNet 2019 dataset, Oei et al. [12] applied a deep learning approach using a **Bidirectional LSTM (Bi-LSTM)** on a different dataset, **MIMIC-III**, achieving an AUROC of 0.85 [12]. As the MIMIC-III dataset is widely used for sepsis detection, Hong et al. [13] proposed a **novel method** on this dataset by aggregating **heterogeneous clinical events** over time and feeding them into a deep learning model. Diverse clinical variables, including vital signs, lab results, and demographic information, were aggregated into **time windows** to capture both short-term and long-term temporal dependencies. They used a **Gated Recurrent Unit (GRU)-based model**. The model was trained and evaluated on the MIMIC-III dataset, achieving an AUROC of 0.85 [13].

2 Research Gaps and Study Aim

Although many studies have explored different ML and DL models for early sepsis detection, several common gaps remain unaddressed. A common limitation across many of these studies, including those by Lyra et al. [4], Wu et al. [7], and Liao et al. [11], is the **lack of model interpretability**. Cutting-edge models such as XGBoost, LSTM, and GRU achieved high performance in predicting sepsis but did not provide or employ any **interpretable techniques** like **SHapley Additive exPlanations (SHAP)** or **Local Interpretable Model-Agnostic Explanations (LIME)** to provide insights on how these models reached their decision, which limits our interpretation of these models and also hinders any future improvement.

Furthermore, although some studies applied **feature selection techniques**, they did not compare the impact of different selection methods on model performance. Additionally, handling of **missing data and class imbalance** still remains inconsistent with limited evaluation of imputation methods or balancing strategies, especially on the PhysioNet 2019 dataset, as it contains about 90% of the data belonging to the non-septic class and over 90% of values are missing for many lab variables, as lab tests are performed intermittently and not at regular hourly intervals [2]. Moreover, most of the studies focused on **standalone models** without taking into account the potential benefits of **ensemble models** that could improve robustness and generalizability.

The aim of this study is to enhance model interpretability and explainability in the context of early sepsis detection by addressing the mentioned limitations. This study seeks to bridge this gap by applying different **XAI methods** such as **SHAP, LIME**, and **Permutation Feature Importance** along with other different methods to evaluate and explain decision-making processes, strengthening trust in model predictions.

3 Data Overview and Preprocessing

The two training datasets were merged together, making sure of the **generalizability of the trained models** across the two hospitals. The raw merged dataset consisted of **1.5 million ICU observations** across over **40,000 patients**.

After applying **exploratory data analysis (EDA)**, the following were found:

Severe Class Imbalance: Only ∼7% of patients were labeled as septic (`SepsisLabel = 1`), with the majority being non-septic.

Patient-level distribution: Sepsis patients: 2,805
Non-sepsis patients: 37,087

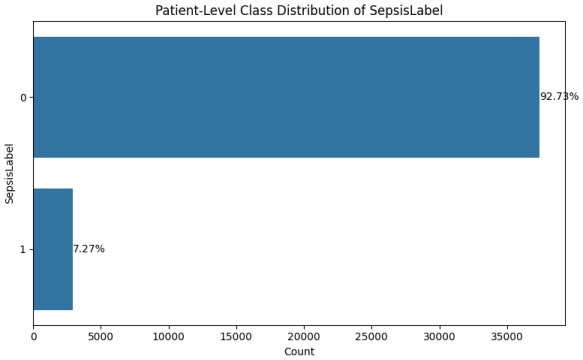

Figure 1: Distribution of classes per patient

High Percentage of Nulls Across Most Features: About **25 features** have more than **89% nulls**. Over **10 lab-related features** had >95% missingness (e.g., `Bilirubin_direct`, `Fibrinogen`, `TroponinI`).

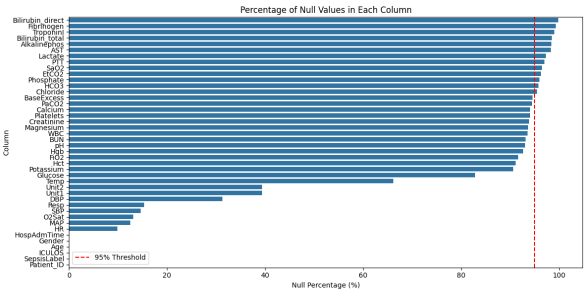

Figure 2: Distribution of null percentage per column

Variable Stay Length: A major issue is that each patient does not have a fixed length of stay, as it ranges from **8 to 336 hours**, with an average of **38.48 hours** after filtering. This variability is problematic because it prevents the trained models from having a dynamic input length, which is a key consideration in ICU patient monitoring.

4 Handling Missing Data

The dataset categorized missing features into **three groups** based on their missingness percentage:

Features with >95% missingness (e.g., `Bilirubin_direct, Fibrinogen, Lactate`) were excluded from direct use with their real values in modeling but were **not dropped entirely**. Instead, their appearance was preserved by adding **binary flags** (e.g., `TroponinI_measured`) to indicate whether the test was performed or not. This decision is based on the understanding that high missingness in ICU lab values is not due to noise or data corruption. These lab tests are only performed when requested by doctors, which is often requested in response to a specific concern.

In contrast, **vital signs** (e.g., `HR, SBP, etc.`) are measured **frequently**, often **every hour** [2]. Infrequently collected lab data is often **informative**; even its rare appearance can be more informative than its regular appearance, as it can hold more clinical meaning than its actual value [14].

Keeping these tests' presence as **binary flags** allows models to learn from the **implicit decision-making process of clinicians**. As an example of how a model can know whether these features are important or not: in clinical practice, doctors don't order lab tests randomly or on a fixed schedule. They decide what to measure based on the patient's condition. So, if the patient is not showing signs of liver failure, the doctor might not order a `bilirubin` test. But if a patient has abnormal heart activity, the doctor might order a `troponin` test to check for heart damage. So, when the test is performed, that means the doctor had a reason to order it based on the patient's condition. This can be called **"informative missingness"**, which is the idea of allowing the models to learn from the absence or appearance of the data [14].

Features with 35–94% missing values (e.g., `Glucose, BUN, WBC`) were retained, and binary `"_observed"` columns were added to indicate if the value was originally present. Most of the mid-range features are **lab tests**, which are, as mentioned above, not frequently measured or, in many cases, rarely measured. The imputation of this range is done based on our assumption of **less time sensitivity** for that range due to being infrequently measured [15]. The **MICE algorithm** [16] is used for imputation as it uses **relationships between variables**. MICE builds a predictive model for each feature using the other features. This leads to more accurate and realistic imputations. **Gaussian noise** was added to avoid overfitting, also to add some randomness and low error to the data, as clinical data cannot correctly be imputed.

The binary columns of these features were added to preserve the originally measured. These columns will also be used for the **custom weighting loss function**. This weighting mechanism assigns **higher importance to truly measured values** and **lower importance to imputed ones**, helping the model avoid over-reliance on potentially inaccurate imputations and reducing the risk of misleading predictions. This loss function is a modification of the standard **binary cross-entropy loss** designed specifically for sepsis prediction. It integrates two weighting components:

- **Class Weighting** — to handle class imbalance, particularly the rarity of Sepsis = 1.
- **Trust Weighting** — to handle data reliability, by giving more weight to samples with a higher proportion of truly measured (non-imputed) fea-

tures.

$$\text{loss}(x_i) = \text{BCE}(y_i, \hat{y}_i) \cdot (1 + \alpha \cdot y_i) \cdot (1 + \beta \cdot \text{measured\_score}_i)$$

Where:

- y_i is the true label (0 or 1)
- $\hat{y}_i$ is the predicted probability from the model
- measured_score_i is a float in $[0, 1]$ representing the fraction of observed (non-imputed) features for that sample
- α is a hyperparameter to increase the importance of positive samples (Sepsis = 1)
- β is a hyperparameter to increase the importance of reliable, observed data

After applying the **MICE algorithm**, it was found that **Diastolic Blood Pressure (DBP)** remains entirely missing for 7,411 patients. This happens because MICE fails when a variable is completely missing within a patient group (patient rows). To solve this issue, **physiological imputation strategies** were followed. First, we used **forward and backward fill per patient**, and then, for rows where DBP was still missing but `Systolic BP (SBP)` and `Mean Arterial Pressure (MAP)` were available, we used this clinical formula [17]:

$$\text{DBP} = \frac{3 \cdot \text{MAP} - \text{SBP}}{2}$$

Features with <35% missingness (e.g., `HR`, `O2Sat`): most of them represent **vital signs** that are measured frequently, often each hour. The imputation strategy for this range follows the clinical assumption that hourly measured signs do not vary significantly within short time intervals, such as one hour [18]. Therefore, missing values were imputed using **forward and backward filling per patient**.

5 Data Balancing Strategy

The dataset exhibited significant **class imbalance**, with only approximately **7% of patients labeled as septic**. This imbalance can lead to huge bias in model performance. To mitigate this problem, a **two-stage data balancing strategy** was applied.

The first stage was to **downsample the majority class** (non-sepsis patients) to a close range to the minority class (septic patients). The majority class was rebalanced to **7,000 patients** using a **resampling method** from `scikit-learn`.

The second stage was to **upsample the minority class** to **3,000 patients** using **SMOTE**. The new ratio of **3:7** was not random, but it was chosen to help the models understand the **rareness of septic patients**.

After completing the preprocessing steps, a histogram of the modified data was plotted to check each column's distribution and validate the modified data. The graph shows that the moderate missing null percentage was similar to the original data, as follows:

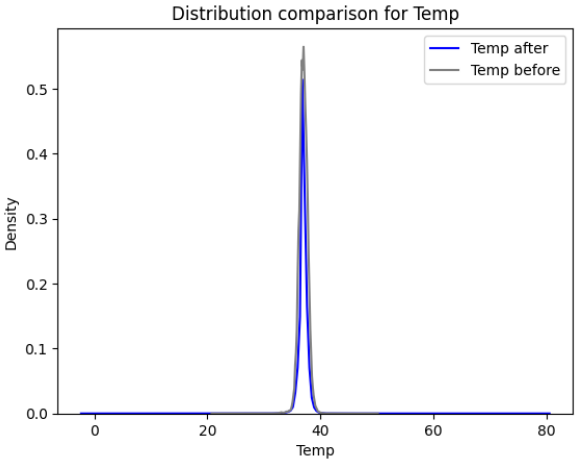

Figure 3: Histogram of Temp feature distribution after preprocessing

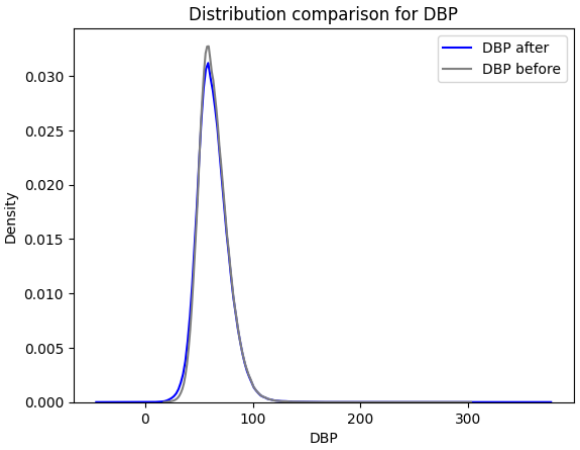

Figure 4: Histogram of DBP feature distribution after preprocessing

After applying the preprocessing steps, the data distribution of the imputed values was analyzed to ensure the integrity and the clinical validity of the transformed dataset.

6 Classification Model and Training

6.1 XGBoost and Bayesian Optimization

We selected XGBoost (Extreme Gradient Boosting) [20] for its efficiency and scalability with gradient

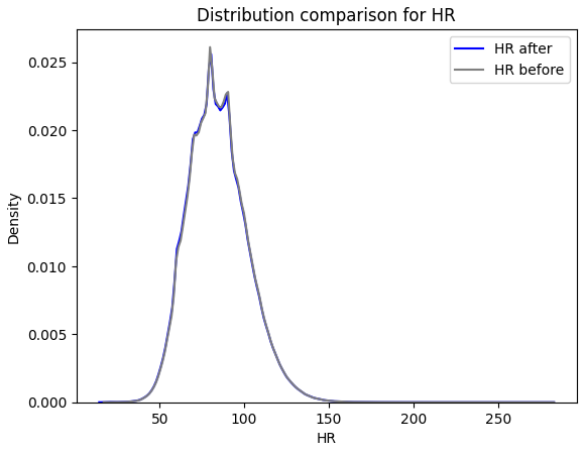

Figure 5: Distribution of the Heart Rate (HR) feature after preprocessing.

boosted decision trees. Hyperparameter tuning is crucial; we employed Bayesian optimization (TPE algorithm [21]) to find optimal values for key hyperparameters:

Table 1: XGBoost Hyperparameters Tuned

Parameter	Description
`max_depth`	Max tree depth.
`learning_rate`	Step size shrinkage.
`colsample_bytree`	Column subsample ratio per tree.
`lambda`, `alpha`	L2 and L1 regularization.

6.2 Model Training

The model training involved:

- **Cross-Validation:** 5-fold CV on training set for robustness and handling imbalance (post-balancing). 96 features were retained for the final model training.
- **Early Stopping:** Prevented overfitting by stopping if validation AUC didn't improve.
- **Objective Function:** Binary logistic for sepsis probabilities.
- **Ensemble Method:** Averaged predictions from 5 CV models for the final output.

6.3 Model Evaluation Metrics

Performance was assessed using:

- **AUROC:** Area Under the ROC Curve (discrimination).
- **Accuracy (ACC):** Overall correct classifications: (TP+TN)/Total.
- **Utility Score (U-Score):** A custom metric reflecting clinical utility, balancing TP, FP, and FN with specific weights (normalized per challenge rules, often per septic patient).

$$U = \frac{\text{Weighted Sum of TP, TN, FP, FN}}{\text{Normalization Factor}}$$

An optimal probability threshold was found by maximizing the U-Score on the test set.

7 Results and Discussion

The 5-fold cross-validation yielded consistent validation AUC scores, with individual folds peaking around 0.885, 0.872, 0.898, 0.873, and 0.876 respectively, as shown in the training logs. The final ensemble model achieved a **Train ROC AUC of 0.8804** and a **Test ROC AUC of 0.8950**.

The Utility Score was evaluated across various thresholds to determine the optimal operating point for clinical application (Figure 7).

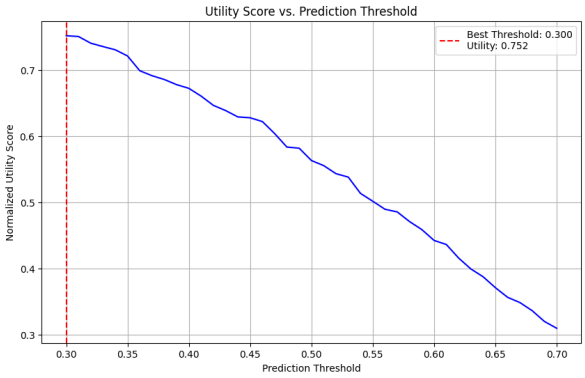

Figure 6: Utility Score vs. Prediction Threshold (Test Set). Max score achieved at a specific threshold.

An optimal probability threshold (e.g., 0.300, based on the plot) yielded a maximum Utility Score of 0.752. This threshold represents the best trade-off between detecting sepsis cases early and minimizing false alarms according to the specified utility function. Performance at this optimal threshold is detailed in Table 2.

Table 2: Evaluation Metrics at Optimal Threshold (0.752 U-Score)

	Precision	Recall	F1-Score
Class 0 (Non-Sepsis)	0.92	0.81	0.86
Class 1 (Sepsis)	0.65	0.83	0.73
Accuracy		0.81	
Macro Avg	0.78	0.82	0.79
Weighted Avg	0.84	0.81	0.82

The model demonstrates excellent sensitivity (recall 0.83) for sepsis detection, which is crucial for

timely intervention, while maintaining reasonable precision (0.65). Feature importances (Figure **??**) highlight the most critical predictors.

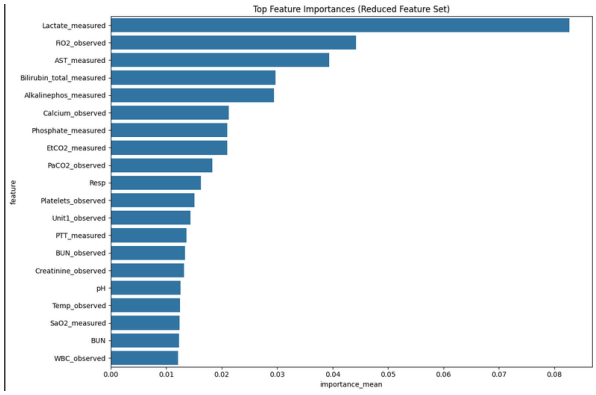

Figure 7: Top 10 Mean Feature Importances (Ensemble). Features include Lactate, FiO2, AST, Bilirubin, Alkaline Phosphatase.

The top 10 most important features identified were: Lactate_measured, FiO2_observed, AST_measured, Bilirubin_total_measured, Alkalinephos_measured, Calcium_observed, Phosphate_measured, EtCO2_measured, PaCO2_observed, and Resp. These align well with clinical indicators of hypoperfusion, respiratory compromise, liver function, and overall patient status relevant to sepsis. This explainability aids clinical interpretation.

Table 3 summarizes the key overall performance metrics for the ensemble model on the test set.

Table 3: Ensemble Model Performance Summary (Test Set)

Metric	Value
Accuracy (@ Optimal Thr)	0.81
Test ROC AUC	0.8950
Max Utility Score	0.752

8 Comparison with Yang et al. (2019)

To provide context for our findings, we compare the performance of our final ensemble model against the results reported by Yang et al. [22], who also utilized an XGBoost ensemble approach on the PhysioNet 2019 dataset. Table 4 summarizes this comparison using key metrics reported in both studies: Accuracy (ACC), Area Under the ROC Curve (AUROC), and the Utility Score (U-Score). The results for Yang et al. [22] are taken from their reported ensemble model performance (Table 3 in their paper).

It is important to note a potential difference in evaluation context. Our model's metrics, particularly the U-Score, are reported at the optimal probability threshold (yielding 0.752) determined by maximizing the U-Score on our test set (as shown in Figure 7). Yang et al. [22] initially reported metrics at a fixed threshold of 0.50 and presented a U-Score of 0.425 for their ensemble, which might correspond to that fixed threshold or another threshold derived from their specific analysis (potentially related to their Figure 3 showing U-Score variation). This difference in threshold optimization strategy likely contributes significantly to the variation observed in the U-Scores.

Table 4: Comparison of Ensemble Model Performance

Metric	**Our Ensemble Model (Optimal Threshold)**	**Yang et al. [22] (Ensemble Result)**
Accuracy (ACC)	0.81	0.818
AUROC	0.895	0.847
U-Score	0.752	0.425

As shown in Table 4, our ensemble model achieves a higher AUROC (0.895 vs. 0.847), indicating superior overall discrimination between septic and non-septic cases across all possible thresholds compared to the model reported by Yang et al. [22]. The accuracy scores are very close (0.81 vs. 0.818), suggesting similar overall correctness at their respective operating points. While the U-Scores differ substantially (0.752 vs. 0.425), our model's score reflects performance optimized specifically for this clinical utility metric at its best threshold. In contrast, the score from Yang et al. [22] may represent performance at a potentially non-optimal or fixed threshold, highlighting the importance of threshold selection for maximizing clinical utility.

9 Random Forest Model Evaluation

Random Forest (RF) was chosen for its ability to handle high-dimensional datasets with many features. It can effectively evaluate feature importance by leveraging an ensemble of decision trees to improve predictive performance.

Before feeding the data to the model, we plotted the correlation between features to identify the most correlated ones. A threshold of 90% correlation was set, and features exceeding this threshold were dropped to reduce multicollinearity and improve model stability.

The data, after dropping the highly correlated features, was fed to an RF model with the following parameters:

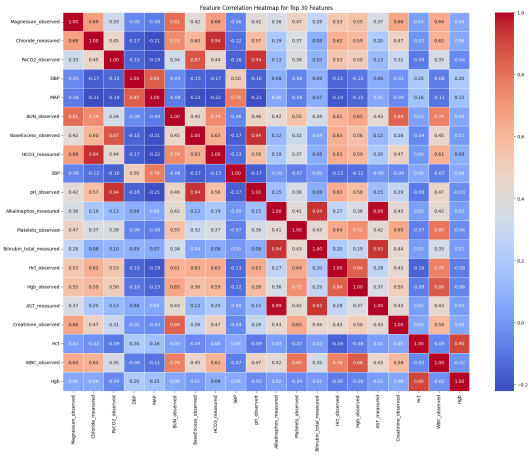

Figure 8: Correlation Matrix of Features

Table 5: Random Forest Hyperparameters

Parameter	Description
`n_estimators`	50 trees
`max_depth`	7 to prevent overfitting
`min_samples_split`	10
`min_samples_leaf`	5
`max_features`	Square root of the total features
Bootstrap Sampling	Enabled for diversity

This configuration was selected to ensure that the model would generalize well while avoiding overfitting, especially given the high-dimensionality of the dataset.

After training the model, a 5-fold cross-validation was performed with various metrics.

9.1 Cross-validation Results

Mean ROC-AUC Score: 0.8577: The ROC-AUC score measures performance across all classification thresholds. A higher AUC indicates better model performance. The RF model with these hyperparameters performed well with an AUC of 0.8577, highlighting its ability to distinguish between septic and non-septic classes.

Mean Accuracy: 0.8046: The model achieved a mean accuracy of 0.8046, indicating that around 80% of the time, the model made correct predictions, which is a good result.

9.2 Classification Report

The classification report includes precision, recall, and F1-score for both septic (Class 1) and non-septic (Class 0) classes:

The F1-score is the harmonic mean of precision and recall, providing a balance between the two:

- **F1-score for Class 0 (Non-Septic):** 0.88

Table 6: Classification Report

Class	Metrics
Class 0 (Non-Septic)	Precision: 0.83, Recall: 0.94, F1-score: 0.88, Support: 1400
Class 1 (Septic)	Precision: 0.79, Recall: 0.54, F1-score: 0.64, Support: 600

- **F1-score for Class 1 (Septic):** 0.64

9.3 Confusion Matrix

The confusion matrix provides a breakdown of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN):

Table 7: Confusion Matrix

Actual/Predicted	Value
Actual: Non-Septic, Predicted: Non-Septic	1313 (TN)
Actual: Non-Septic, Predicted: Septic	87 (FP)
Actual: Septic, Predicted: Non-Septic	274 (FN)
Actual: Septic, Predicted: Septic	326 (TP)

This matrix shows that the model correctly identified 1313 non-septic cases (TN) and 326 septic cases (TP), while incorrectly classifying 87 non-septic cases as septic (FP) and 274 septic cases as non-septic (FN).

9.4 ROC Curve and AUC Score

The ROC Curve graphically represents the trade-off between the True Positive Rate (sensitivity) and False Positive Rate (1-specificity). The AUC Score indicates the model's overall ability to discriminate between classes:

AUC Score: 0.8670: This demonstrates that the model performs well in distinguishing between septic and non-septic cases.

9.5 Feature Importance

As shown in the graph, the top features were:

- **FiO2 (Fraction of Inspired Oxygen):** This feature is crucial in detecting respiratory failure, a key complication in sepsis. The PaO2/FiO2 ratio is an important indicator of ARDS, critical for identifying patients at risk of severe respiratory decline [23].
- **PaCO2 (Partial Pressure of Carbon Dioxide):** Elevated PaCO2 can indicate respiratory acidosis, often seen in septic shock, underscoring its importance in sepsis detection [23].

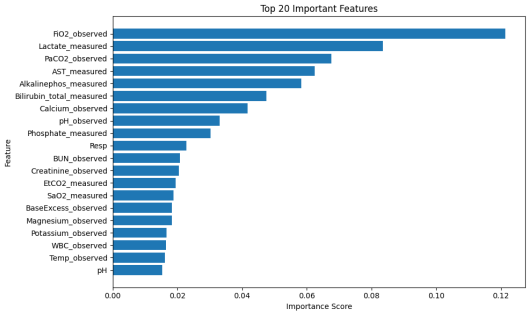

Figure 9: Top Important Features

- **Bilirubin_total:** High bilirubin levels indicate liver dysfunction in sepsis. It is defined as TBIL ≥ 2 mg/dL and is linked to poorer outcomes in septic shock, reflecting liver impact [24].
- **Lactate:** Known for reflecting tissue hypoxia, lactate is a reliable marker of sepsis severity. High levels are associated with worse outcomes, especially in patients with hepatic dysfunction [24].

Acknowledgments

References

[1] World Health Organization, "WHO calls for global action on sepsis—cause of 1 in 5 deaths worldwide," Sep. 8, 2020. [Online]. Available: https://www.who.int/news/item/08-09-2020-who-calls-for-global-action-on-sepsis---cause-of-1-in-5-deaths-worldwide.

[2] M. A. Reyna et al., "Early prediction of sepsis from clinical data: The PhysioNet/Computing in Cardiology Challenge 2019," *Crit. Care Med.*, vol. 48, no. 2, pp. 210–217, Feb. 2020. DOI: 10.1097/CCM.0000000000004145.

[3] A. Kumar et al., "Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock," *Crit. Care Med.*, vol. 34, no. 6, pp. 1589–1596, Jun. 2006. DOI: 10.1097/01.CCM.0000217961.75225.EF.

[4] S. Lyra, S. Leonhardt, and C. H. Antink, "Early prediction of sepsis using random forest classification for imbalanced clinical data," in *Computing in Cardiology*, 2019.

[5] F. Mohammad et al., "Early prediction of onset of sepsis in a clinical setting," *arXiv preprint*, arXiv:2402.03486, Feb. 2024. [Online]. Available: https://arxiv.org/abs/2402.03486.

[6] S. Nirgudkar and T. Ding, "Early detection of sepsis using ensembles," presented at *PhysioNet/Computing in Cardiology Challenge*, 2019.

[7] Q. Wu et al., "A customized down-sampling machine learning approach for sepsis prediction," presented at *PhysioNet/Computing in Cardiology Challenge*, 2019.

[8] Vandana and R. Chhikara, "Comparative Analysis of ML Models with Selection Methods for Early Predictive Analytics of Sepsis in ICU," *Indian Journal of Science and Technology*, vol. 17, no. 41, pp. 4338–4348, Nov. 2024. DOI: 10.17485/IJST/v17i41.1925.

[9] G. Kong, X. Lin, and H. Hu, "Using machine learning methods to predict sepsis," *Int. J. Med. Inform.*, vol. 131, p. 103940, 2019. DOI: 10.1016/j.ijmedinf.2019.103940.

[10] K. Apalak and K. Kiasaleh, "Early Detection of Sepsis With Temporal Convolutional Networks," *IEEE Access*, vol. 8, pp. 132437-132449, 2020. DOI: 10.1109/ACCESS.2020.3010445.

[11] J. Liao et al., "Does deep learning REALLY outperform non-deep machine learning for clinical prediction on physiological time series?" *arXiv preprint*, arXiv:2211.06034, Nov. 2022. [Online]. Available: https://arxiv.org/abs/2211.06034.

[12] S. P. Oei et al., "Towards early sepsis detection from measurements at the general ward through deep learning," *Intelligence-Based Medicine*, vol. 5, p. 100042, 2021. DOI: 10.1016/j.ibmed.2021.100042.

[13] S. Hong et al., "Early prediction of sepsis from clinical data via heterogeneous event aggregation," *arXiv preprint*, arXiv:1910.06792, Oct. 2019. [Online]. Available: https://arxiv.org/abs/1910.06792.

[14] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Modeling missing data in clinical time series with RNNs," *arXiv preprint arXiv:1606.04130*, 2016. [Online]. Available: https://arxiv.org/abs/1606.04130.

[15] H. Harutyunyan, H. Khachatrian, D. C. Kale, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific Data*, vol. 6, no. 1, p. 96, 2019. DOI: 10.1038/s41597-019-0103-9.

[16] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, "Multiple imputation by chained equations: what is it and how does it work?" *International Journal of Methods in Psychiatric Research*, vol. 20, no. 1, pp. 40-49, 2011. DOI: 10.1002/mpr.329.

[17] R. E. Klabunde, *Cardiovascular Physiology Concepts*, 2nd ed. Lippincott Williams & Wilkins, 2011. ISBN: 978-1451113846.

[18] A. E. Johnson, T. J. Pollard, L. Shen, et al., "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, p. 160035, 2016. [Online]. Available: https://www.nature.com/articles/sdata201635.

[19] G. Stochero, E. Rojas, R. Ribeiro, and A. P. Santos, "Missing data imputation techniques for wireless continuous vital signs monitoring," *Journal of Clinical Monitoring and Computing*, vol. 37, no. 6, pp. 1503-1518, 2023. DOI: 10.1007/s10877-023-01044-4.

[20] T. Chen, G. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, New York, NY, USA, 2016:785–794.

[21] J. Bergstra, R. Bardenet, Y. Bengio, B. Kégl, "Algorithms for hyper-parameter optimization," *Advances in Neural Information Processing Systems 24*, Curran Associates, Inc., 2011:2546–2554.

[22] M. Yang, X. Wang, H. Gao, Y. Li, X. Liu, J. Li, and C. Liu, "Early prediction of sepsis using Multi-Feature fusion based XGBoost learning and Bayesian optimization," in *Computing in Cardiology (CinC)*, vol. 46, Sep. 2019. DOI: 10.22489/CinC.2019.071. [Online]. Available: https://ieeexplore.ieee.org/document/8966345

[23] C. Nickson, "PaO2/FiO2 Ratio (P/F Ratio)," Life in the Fast Lane (LITFL), Nov. 2020. [Online]. Available: https://litfl.com/pa02-fi02-ratio/

[24] N. Takahashi, T.-A. Nakada, K. R. Walley, and J. A. Russell, "Significance of lactate clearance in septic shock patients with high bilirubin levels," *Sci. Rep.*, vol. 11, no. 1, p. 6313, Mar. 2021, doi: 10.1038/s41598-021-85700-w. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7970876/