Predicting Sepsis Onset in ICU Patients Using Machine Learning and Feature Selection: A Case Study of MIMIC-IV Data

Wenhong Shan
School of Control Science and Engineering
Shandong University, Jinan, Shandong 250061, China

Duanchen Sun
School of Mathematics
Shandong University
Jinan, Shandong 250100, China

Zhi-Ping Liu*
School of Control Science and Engineering
Shandong University
Jinan, Shandong 250061, China
zpliu@sdu.edu.cn

Abstract—**This study aims to predict the onset of sepsis in ICU patients using multiple machine-learning models with data from the MIMIC-IV database. We employed seven prediction models—Logistic Regression, Gaussian Naive Bayes, Random Forest, Artificial Neural Network, SVM, XGBoost, and Gradient Boosting Decision Tree—using 81 features extracted from routine checks conducted within 12 hours before and 4 hours after ICU admission from 46,530 patients. The features were represented by key values such as maximum, minimum, and mean, referencing the official MIMIC derived daily data tables. XGBoost achieved the best performance with an AUC of 0.81 and an accuracy of 0.729. We then applied Recursive Feature Elimination (RFE) to identify the optimal feature subset for each model, finding 13 features commonly selected across all models. This overlap highlights their importance in predicting sepsis and suggests potential for model simplification without losing predictive power. Notably, 6 of these common features overlap with the top 20 SHAP features from the XGBoost model, validating their critical role. Training models with these 18 common features demonstrated that this feature selection process can lead to a simplified, yet effective, predictive model. Thus, we developed a simple, rapid, and practical tool for early detection and clinical intervention of sepsis.**

Keywords—Sepsis Prediction, Intensive Care Unit, MIMIC-IV database, Machine Learning, Feature Selection.

I. Introduction

Sepsis is a major cause of morbidity and mortality in the Intensive Care Unit (ICU) [1]. A study on the prevalence of sepsis in critically ill patients in China reported that 42.5% of patients developed sepsis or septic shock either at ICU admission or within the first 48 hours of ICU stay [2]. Additionally, data from the global Intensive Care Over Nations (ICON) audit revealed that 29.5% of patients experienced sepsis during their ICU stay, with incidence rates ranging from 13.6% to 39.3% [3].The complexity of sepsis makes rapid and accurate diagnosis challenging [4]. However, early prediction of sepsis onset is crucial, as each hour of delay in effective antibiotic administration after the onset of sepsis increases the mortality rate [5]. Currently, the clinical standard for determining the timing of sepsis infection is blood culture, but this method is time-consuming. It can take up to 48 hours to identify the bacteria in the blood sample, potentially missing the optimal treatment window [6].

Here, we aim to develop a sepsis prediction model suitable for clinical environments. We employed sepsis annotations based on the Sepsis-3 criteria [7] and referred to relevant literature. Currently, the available sepsis prediction studies primarily rely on time series data (such as RNN, LSTM, Transformer) [8][9], analyzing physiological indicators at different time points to make predictions. These studies can capture the dynamic changes in a patient’s condition, thus improving prediction accuracy. However, these methods require a large amount of continuous data and involve complex data processing and model training [10]. The significant demand for a large amount of data restricts their clinical application in the early identification of sepsis through intelligent medical systems.

Considering that in actual clinical treatment, there is significant data missing, making it challenging to obtain complete continuous physiological data for each patient. The prediction need to be effective and easy to be conducted. To address this challenge, we selected key values from the abundant data obtained within a time window from patient monitors (e.g., heart rate and frequently measured blood gas scores) to make predictions. By using key values such as maximum, minimum, and average, we can reduce the model's complexity. This approach effectively filters critical information (e.g., temperature spikes due to infection in sepsis), while reducing data complexity. This selection makes the model training process more efficient and ensures rapid decision-making in emergency situations. Additionally, the smaller data volume reduces the burden of data availability, storage and processing. Furthermore, since it does not require handling complex time series data except the values in a short-time window, the model training and optimization process is simpler and faster [11].

To align with clinical realities, our study is based on the public resource of the public MIMIC-IV database. There is a severe issue of class imbalance in the actual data, with many studies finding lower accuracy for positive cases [12]. Our study particularly addresses this problem and gives it significant consideration during the model development process.

In this work, we conduct seven classification algorithms for early detection of sepsis using the MIMIC-IV data. The algorithms include Artificial Neural Network (ANN), Gaussian Naïve Bayes (GNB), Gradient Boosting Decision Tree (GBDT), Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM) and XGBoost. Each feature is represented by several key values selected within a specified time window. Our findings indicate that XGBoost exhibited the best performance among all the classifiers, outperforming the other models. We also conduct feature selection for prioritizing important features using recursive feature elimination. The feature importance is evaluated by Shapley values. The predictive model is then be simplified and optimized via these selected features.

The rest of the paper is organized as follows. Section II briefly introduces the data and methods used in our study. Section III presents the detailed prediction results and feature selection. Finally, Section IV presents a conclusion of this study.

II. MATERIALS AND METHODS

A. *Data*

In this study, we utilized the MIMIC-IV database, which contains extensive clinical data of ICU patients. The MIMIC-IV database determines the suspicion of infection time by recording and analyzing the antibiotic usage time and microbiological culture time. These data are automatically extracted and organized from patients' electronic health records and are determined according to the Sepsis-3 criteria [13]. We applied and obtained the scientific research license of MIMIC-IV strictly according to its protocol.

To refine our dataset, we used PostgreSQL software to extract data from the MIMIC-IV database. We first removed patients with more than 50% missing data. The analysis was based on the official derived tables in the MIMIC-IV database, using the suspected infection time for sepsis 3 as the criterion for sepsis infection. We excluded patients who were infected between ICU admissions and those who developed sepsis after 24 hours of ICU admission. Given that most sepsis infections occur within the first day of ICU admission, our study specifically focuses on predicting sepsis infection within this critical timeframe.

The dataset extracted from MIMIC-IV includes 84 columns and 46,530 rows, encompassing key physiological parameters related to sepsis prediction, such as vital signs, laboratory results, etc. In this study, we utilized the clinical data from 12 hours before and 4 hours after ICU admission as input features for our models. For practical clinical application scenarios, we only used the first single measurement value for each index. The observation window totals 16 hours of data used to extract features to predict whether the patient will be suspected of having a sepsis infection within 24 hours after ICU admission. The prediction labels are based on whether the patient is suspected of sepsis infection within 24 hours after ICU admission.

B. *Data Preprocessing*

In this work, we utilized the K-Nearest Neighbors (KNN) method to impute missing values. Figure 1a depicts the distribution of sepsis labels in the database, revealing a significant imbalance in the sample distribution. Considering that sample imbalance is common in practical predictive applications, we opted not to use oversampling to balance the samples. Instead, we adjusted the Youden Index during the model training process to address this imbalance. Figure 1b shows the distribution of heart rate and SBP before and after KNN imputation.

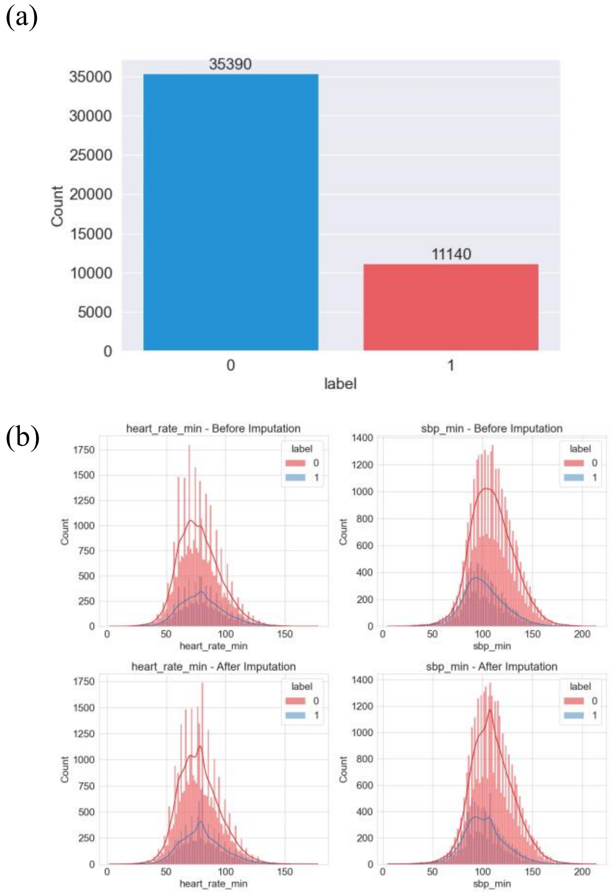

Fig. 1: Data Overview. (a) Sepsis Label Distribution; (b) Heart Rate and SBP Distribution Before and After KNN Imputation.

C. *Exploratory Data Analysis and Feature Engineering*

Table 1 categorizes our 81 features into 10 groups, each represented by a single feature.

TABLE I. Feature Categories and Descriptions.

Feature type	Category name	Representative feature
I	Heart Indicators	heart_rate_mean
II	Blood Pressure Indicators	sbp_mean
III	Respiratory Indicators	resp_rate_mean
IV	Temperature Indicators	temperature_mean
V	Oxygen Indicators	spo2_mean
VI	Blood Sugar Indicators	glucose_vitalsign_mean
VII	Blood Gas Indicators	ph_min
VIII	Blood Indicators	wbc_min
IX	Coagulation Function Indicators	inr_min
X	Glasgow Coma Scale (GCS)	gcs_motor

Next, we implemented summary statistics for exploring the features and their relationships, which are presented in Figures 2 and 3, respectively. Figure 2 illustrates the relationships between the typical features in the ten categorizes. Figure 2a is a heatmap showing the Pearson's correlation coefficients between these representative features. It indicates the linear relationships between each pair of features. Figure 2b refers to a heatmap of the Mutual Information (MI) between the representative features and the target variable. It helps identify which features share the most information with the target. The heatmap of the Spearman Rank Correlation coefficients between the representative features is shown in Figure 2c. It reveals the monotonic relationships between these features.

Figure 3 shows the distribution of these features before and after data imputations. Figure 3a refer to the boxplots of the representative features, illustrating the distribution, median, quartiles, and potential outliers of each feature. Similarly, Figure 3b illustrates the violin plots of the representative features, combining boxplot characteristics with density information to show the distribution shape and density of the data.

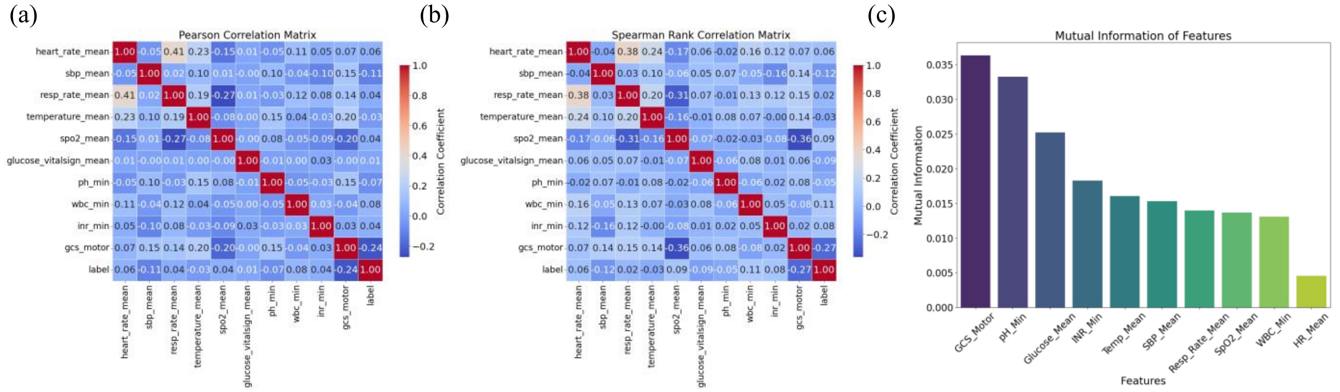

Fig. 2: Relationships Between Features. (a) Heatmap of Pearson Correlation Coefficients; (b) Heatmap of Spearman Rank Correlation Coefficients; (c) Histogram of Mutual Information.

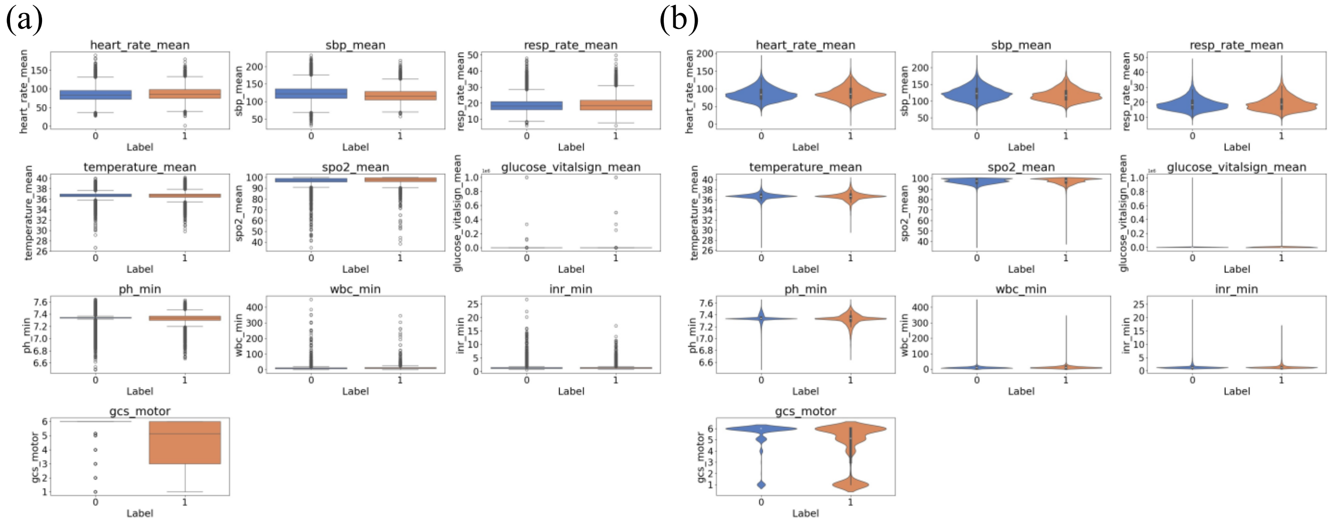

Fig. 3: Differences in Categorical Features. (a) Box Plot; (b) Violin Plot.

D. Training and Testing data

In this study, we first loaded the preprocessed dataset and split it into training and testing subsets in a 70:30 ratio. To further evaluate the model's performance, we used Stratified Shuffle Split cross-validation, which divides the data into 5 cross-validation sets while maintaining the consistency of the positive and negative sample ratio in each validation set.

The models we selected, including XGBoost, Random Forest, and Logistic Regression, are widely used and effective for binary classification tasks in medical problems. As highlighted by Vellido et al, these models excel in handling complex medical data, addressing class imbalance issues, and improving predictive accuracy, making them suitable choices for our study [15].

For hyperparameter tuning across all models, we utilized GridSearchCV. By defining appropriate parameter spaces—such as max_depth, n_estimators, and learning_rate for XGBoost—and combining them with K-fold cross-validation, we systematically searched for the optimal parameter combinations to enhance model performance. GridSearchCV evaluated the performance of each parameter combination, ultimately identifying the configuration that maximized the ROC-AUC score.

E. Prediction Model Construction

In this study, we utilized seven different machine learning classifiers for sepsis prediction. The training time varied among the models, with Random Forest and GBDT having relatively longer training times compared to the others.

F. Model Evaluation and Hyperparameter Optimization

To ensure the performance and reliability of our models, we conducted comprehensive model evaluation and hyperparameter optimization using various methods. Initially, we measured model performance with common classification metrics such as Accuracy, Precision, Recall, F1-score, Log loss, and AUC. To further understand the model's predictive abilities and feature importance, we generated detailed classification reports, confusion matrices, and plotted the ROC curves.

Additionally, we conducted a detailed analysis of three key parameters: Test Score, Validation Score, and the Number of Selected Features. The Test Score and Validation Score represent the model's performance on the test and validation datasets, respectively. The Number of Selected Features corresponds to the optimal set of features identified through Recursive Feature Elimination (RFE). Specifically, we systematically explored various feature combinations and selected the one that maximized the area under the ROC curve (AUC) for each model. This approach allowed us to rigorously assess both the model's performance and the effectiveness of feature selection.

Model interpretability was achieved through the application of SHAP (SHapley Additive exPlanations) values, which quantified the contribution of each feature to the prediction results. This method allowed us to identify the most important features for the model's predictions and highlighted key features selected during the feature selection process.

For enhancing model performance, we employed both grid search and random search strategies for hyperparameter optimization. By fine-tuning the hyperparameters, we identified the optimal model configuration and saved the best model for subsequent predictions and evaluations. Additionally, RFE was performed to select features based on their impact on the AUC curve, ensuring that only the most significant features were used in the final model.

TABLE II. Performance Metrics for Different Machine Learning Models.

Model	Accuracy	Precision	Recall	F1 Score	Test Score	Validation Score	Selected Features
ANN	0.697	0.426	0.691	0.528	69.740	71.465	45
GNB	0.675	0.401	0.666	0.501	67.541	67.572	40
GBDT	0.716	0.452	0.754	0.565	71.617	78.232	71
LR	0.683	0.414	0.722	0.526	68.264	68.294	73
RF	0.755	0.499	0.668	0.571	75.528	76.424	81
SVM	0.694	0.418	0.648	0.508	69.417	83.221	55
XGBoost	0.729	0.465	0.729	0.568	72.928	79.559	61

III. RESULTS AND DISCUSSION

Table II presents the performance metrics for seven machine learning models predicting sepsis onset in ICU patients: ANN, GNB, GBDT, LR, RF, SVM, and XGBoost.

Upon examining the results, the RF classifier emerged as the top performer with the highest accuracy and validation scores, indicating its strong capability in handling complex data and providing reliable predictions. The use of all 81 features in RF likely contributed to its superior performance by capturing nuanced patterns in the data.

XGBoost also showed robust performance across various metrics, making it a strong contender. Its ability to balance precision and recall highlights its effectiveness in managing false positives and false negatives, crucial in clinical settings.

SVM demonstrated the highest validation score, indicating potential in specific scenarios despite moderate overall performance. LR and ANN provided moderate performance, with ANN slightly outperforming LR in recall, suggesting it might be better at identifying true positive cases of sepsis, though at the cost of lower precision.

GNB demonstrated the lowest performance, likely due to its assumptions about data distribution not aligning with the complexity of the medical dataset. Despite this, GNB's simplicity and speed can offer value in scenarios requiring quick, less resource-intensive models.

Finally, the GBDT performed well, excelling in recall, making it effective in identifying actual sepsis cases. This model's strong recall, coupled with good accuracy and precision, makes it a viable option for sepsis prediction.

Overall, in our study, RF and XGBoost emerged as the top models for predicting sepsis onset. These results align with findings by Camacho-Cogollo et al. [14], who highlighted XGBoost's superior performance in handling sparse data and large healthcare datasets. Their study demonstrated that XGBoost, through tree learning techniques and algorithmic optimizations, achieved higher accuracy, recall, F1, and AUC, further supporting its robustness in early sepsis prediction.

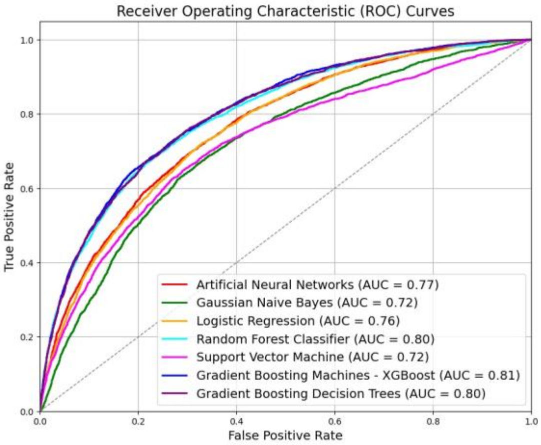

Fig. 4: ROC Curves for Different Predictive Models.

Figure 4 displays the ROC curves and AUC values for seven machine learning models in sepsis prediction. XGBoost (AUC = 0.81) demonstrates the best performance, followed closely by RF and GBDT, both with an AUC of 0.80. ANN also shows strong performance with an AUC of 0.77. LR follows with an AUC of 0.76. Both SVM and Gaussian Naive Bayes have lower performance, each with an AUC of 0.72. The high AUC for XGBoost, combined with its consistent true positive rate across varying false positive rates, highlights its superior predictive power.

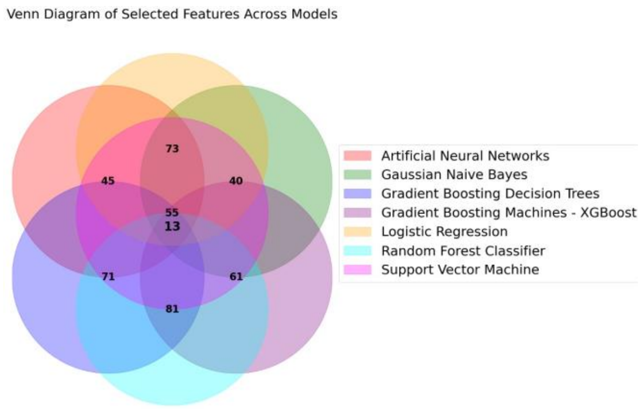
Venn Diagram of Selected Features Across Models
73
45
40
55
13
71
61
81
Artificial Neural Networks
Gaussian Naive Bayes
Gradient Boosting Decision Trees
Gradient Boosting Machines - XGBoost
Logistic Regression
Random Forest Classifier
Support Vector Machine

Fig. 5: Overlapping Features Selected by Seven Models Using RFE.

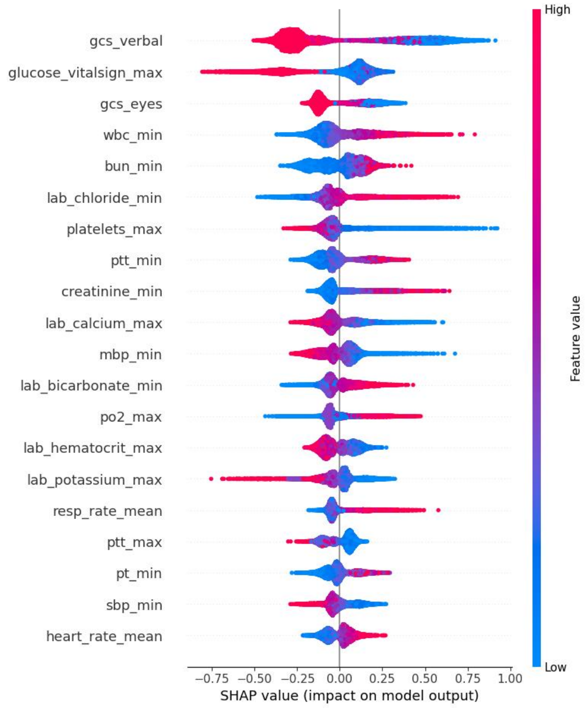

Fig. 6: Top 20 SHAP Features in the XGBoost Model.

Recursive Feature Elimination (RFE) was employed to train the former seven models and identify the optimal feature subset for each model. By comparing the features selected by each model, we found 13 features that were commonly selected by all models (as shown in Fig. 5). The overlap of these features across different models indicates their significant importance in predicting sepsis.

To further reduce the complexity of prediction model, we tried to train using these 18 common features, aiming to optimize model performance and enhance interpretability. Additionally, we observed that 6 of these 13 common features overlap with the top 20 SHAP features from the XGBoost model (as shown in Fig. 6), further validating their critical role in sepsis prediction.

Next, we plan to train models using both these 13 common features and the 20 SHAP features from the XGBoost model and evaluate their predictive performance. By comparing the performance of these two feature sets, we aim to identify the most valuable features for prediction, thus providing practical guidance for clinical practice. The final selected features will play a crucial role in the early prediction and clinical intervention of sepsis and should be given priority attention by clinicians.

TABLE III. Prediction Results with Selected Features.

Feature Set	Accuracy	Precision	Recall	F1 Score	AUC
13(Overlapping)	0.71	0.77	0.71	0.73	0.76
20 (SHAP)	0.73	0.78	0.73	0.74	0.78
6 (Common)	0.66	0.76	0.67	0.68	0.74

The results presented in TABLE III illustrate the performance of our models using different sets of selected features. For each feature set, we evaluated the model based on several metrics including accuracy, precision, recall, F1 score, and AUC. For the Overlapping Features set, the model achieved an accuracy of 0.71. The SHAP Features set resulted in the highest accuracy of 0.73, while the Common Features set yielded an accuracy of 0.66.

The performance metrics from TABLE III demonstrate that these selected features significantly contribute to the accurate prediction of sepsis, supporting their importance in clinical assessments and early detection strategies. The results validate the utility of these physiological parameters in developing robust predictive models for sepsis, potentially aiding in timely and effective clinical interventions.

TABLE IV. Selected Features.

Category	Features
Lab Values	wbc_min, lab_hemoglobin_min, lab_hematocrit_max, lab_calcium_max, lab_hematocrit_min
Vital Signs	pco2_max, heart_rate_min, mbp_min, temperature_min
GCS Scores	gcs_eyes, gcs_verbal, gcs_min
Other Clinical Measures	bun_min

As shown in Table IV, these 13 features are categorized into four groups: Lab Values, Vital Signs, Glasgow Coma Scale (GCS) Scores, and Other Clinical Measures.

Lab Values include key biochemical and hematological indicators such as the minimum hemoglobin level (lab_hemoglobin_min). These indicators help assess the patient’s physiological state and disease progression. Vital Signs cover the minimum temperature (temperature_min), minimum heart rate (heart_rate_min), and minimum mean arterial pressure (mbp_min), all of which reflect the patient's current health status.

GCS Scores include GCS eye response (gcs_eyes), GCS verbal response (gcs_verbal), and total GCS score (gcs_min), which are used to assess the patient's level of consciousness. In the XGBoost model, GCS scores have very high feature importance, indicating that the patient's coma state is a critical predictor of sepsis and should be closely monitored in clinical practice.

Other Clinical Measures, such as the minimum blood urea nitrogen level (bun_min), provide crucial information about the patient's infection status and kidney function, aiding in the comprehensive assessment of the patient's health state.

It is noteworthy that the GCS indicators show very high feature importance when interpreted with SHAP values in the XGBoost model and are also selected by multiple models. This indicates that, clinically, the patient's coma indicators are significant parameters for identifying infections and should be closely monitored.

IV. Conclusion

In this study, we employed various machine learning models to early detection of sepsis based on MIMIC-IV database data. We extracted 81 features from routine checks within 12 hours before and 4 hours after ICU admission from 46,530 patients, using key values such as maximum, minimum, and mean measurements for each feature in the time windows. The aim was to assess these features' predictive ability for sepsis onset within 24 hours of ICU admission. Seven prediction methods were considered, including ANN, GBDT, GNB, LR, RF, SVM and XGBoost. Recursive feature elimination and Shapley values were used for feature selection and importance quantification. Our results demonstrate that XGBoost yielded the optimal performance. The selected features and the predicted performance using their key measurement values suggest that our proposed approach is both effective and suitable for accurate prediction of sepsis onset. Furthermore, the proposed methods need to be implemented on more independent datasets, and additional clinical trials validating the proposed intelligent system are very necessary for practical clinical applications.

ACKNOWLEDGMENTS

This work was partially supported by the National Key Research and Development Program of China under grant numbers 2022YFA1004800 and 2020YFA0712402; the National Natural Science Foundation of China (NSFC) under grant number 62373216; the Fundamental Research Funds for the Central Universities (No. 2022JC008). We would like to thank the PhysioNet team for providing access to the MIMIC-IV database. The use of this database was approved under the PhysioNet Credentialed Health Data Access Committee (CHDAC) with the required CITI training completed and data use agreement signed.

References

[1] Calsavara, A.J.C., Costa, P.A., Nobre, V., & Teixeira, A.L. (2018). Factors associated with short and long term cognitive changes in patients with sepsis. Scientific Reports, vol. 8, pp. 4509. doi: 10.1038/s41598-018-22754-3.

[2] Wang, M., Jiang, L., Zhu, B., Li, W., Du, B., et al. (2020). The Prevalence, Risk Factors, and Outcomes of Sepsis in Critically Ill Patients in China: A Multicenter Prospective Cohort Study. Frontiers in Medicine, vol. 7. doi: 10.3389/fmed.2020.593808.

[3] Sakr, Y., Jaschinski, U., Wittebole, X., Szakmany, T., Lipman, J., Namendys-Silva, S.A., et al. (2018). Sepsis in intensive care unit patients: worldwide data from the intensive care over nations audit. Open Forum Infectious Diseases, vol. 5, pp. ofy313. doi: 10.1093/ofid/ofy313.

[4] Liu, V.X., et al. (2017). The Timing of Early Antibiotics and Hospital Mortality in Sepsis. American Journal of Respiratory and Critical Care Medicine, vol. 196, no. 7, pp. 856-863. doi: https://doi.org/10.1164/rccm.201609-1848oc.

[5] Kumar, A., et al. (2006). Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Critical Care Medicine, vol. 34, no. 6, pp. 1589-1596. doi: https://doi.org/10.1097/01.ccm.0000217961.75225.e9.

[6] Osthoff, M., Gürtler, N., Bassetti, S., et al. (2017). Impact of MALDI-TOF-MS-based identification directly from positive blood cultures on patient management: a controlled clinical trial. Clinical Microbiology and Infection, vol. 23, pp. 78-85.

[7] Singer, M., Deutschman, C.S., Seymour, C.W., et al. (2016). The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA, vol. 315, pp. 801-810.

[8] Yang, Z., Cui, X., & Song, Z. (2023). Predicting sepsis onset in ICU using machine learning models: a systematic review and meta-analysis. BMC Infectious Diseases, vol. 23, pp. 635. doi: https://doi.org/10.1186/s12879-023-08614-0.

[9] Tang, Y., Zhang, Y., & Li, J. (2024). A time series driven model for early sepsis prediction based on transformer module. BMC Medical Research Methodology, vol. 24, pp. 23. doi: https://doi.org/10.1186/s12874-023-02138-6.

[10] Wang, Y., Zhao, Y., Callcut, R., et al. (2022). Integrating physiological time series and clinical notes with transformer for early prediction of sepsis. arXiv preprint arXiv:2203.14469.

[11] Yang, Z., Cui, X., & Song, Z. (2023). Predicting sepsis onset in ICU using machine learning models: a systematic review and meta-analysis. BMC Infectious Diseases, vol. 23, pp. 635. doi: https://doi.org/10.1186/s12879-023-08614-0.

[12] Ying, T.X., & Abu-Samah, A. (2022). Early Prediction of Sepsis for ICU Patients using Gradient Boosted Tree. In 2022 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS), pp. 78-83. IEEE.

[13] Johnson, A.E.W., Pollard, T.J., Shen, L., Li-wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., & Mark, R.G. (2020). MIMIC-IV (version 1.0). PhysioNet. doi: https://doi.org/10.13026/a3wn-hq05.

[14] Camacho-Cogollo, J.E., Bonet, I., Gil, B., & Iadanza, E. (2022). Machine Learning Models for Early Prediction of Sepsis on Large Healthcare Datasets. *Electronics*, vol. 11, pp. 1507. doi: https://doi.org/10.3390/electronics11091507.

[15] Vellido, A., Ribas, V., Morales, C., et al. (2018). Machine learning in critical care: state-of-the-art and a sepsis case study. *Biomedical Engineering Online*, vol. 17, pp. 1-18.