Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

A Predictive Machine Learning Framework for Early-Stage Sepsis Detection

Bhavya R A
Department of CSE
SJC Institute of Technology
Chickaballapur, India

Likhitha R
Department of CSE
SJC Institute of Technology
Chickaballapur, India

Shashank S R
Department of CSE
SJC Institute of Technology
Chickaballapur, India

Megha Sri B N
Department of CSE
SJC Institute of Technology
Chickaballapur, India

Mrudhula Varsha Shri A
Department of CSE
SJC Institute of Technology
Chickaballapur, India

Abstract - Sepsis represents a critical medical condition that ensues when the body's reaction to an infection leads to damage to tissues and failure of organs. The prompt and precise identification of sepsis is crucial, as timely intervention can significantly decrease mortality rates among patients in the Intensive Care Unit (ICU). This paper introduces a machine learning-driven system capable of forecasting the onset of sepsis several hours before it is clinically diagnosed by examining the physiological parameters of patients and their electronic health records. The study incorporates techniques for data preprocessing, feature selection, and model optimization, utilizing algorithms such as Random Forest, Logistic Regression, Gradient Boosting, and Decision Tree. The Random Forest model proposed achieved exceptional outcomes, with an accuracy of 99.01%, an F1-score of 99%, and an AUCROC value of 99.99%. This system highlights the potential of artificial intelligence in aiding clinical decision-making, facilitating early detection, and improving patient outcomes.

Index Terms - Sepsis, Machine Learning, Early Detection, Random Forest, Artificial Intelligence.

I. INTRODUCTION

Sepsis is a critical condition that poses a significant threat to life, marked by an inappropriate response of the host to infection, which results in organ failure and elevated mortality rates in intensive care units (ICUs)[1][2]. Despite the progress made in critical care, sepsis continues to be a leading cause of death globally, primarily due to delays in both diagnosis and treatment. Research studies have consistently verified that recognizing sepsis early and intervening promptly can greatly enhance patient outcomes, shorten the duration of ICU stays, and reduce mortality rates. Conventional rule-based scoring methods and manual evaluations in clinical settings frequently struggle to identify sepsis in its initial phases because of the intricate, fluctuating, and diverse characteristics of physiological data in critically ill individuals[3][4].

Contemporary intensive care units produce significant amounts of high-frequency data, which includes continuous vital signs, physiological waveforms, laboratory test results, and electronic health record (EHR) information[5]. This data-intensive setting offers a chance to leverage enhanced machine learning methods to detect subtle patterns and temporal trends that may specify imminent clinical decline. Deep learning algorithms, in specific, have shown exceptional ability in capturing nonlinear relationships and long-term temporal dependencies in multivariate time-series data, making them mostly effective for the primary prediction of sepsis[6].

II.RELATED WORK

[7] Numerous works have used Random Forest models for the initial detection of sepsis, owing to their effectiveness in handling noisy and diverse clinical data. These models analyze vital signs and laboratory results gathered from electronic medical records to pinpoint initial physiological changes linked to the onset of sepsis. Random Forest classifiers have shown impressive predictive capabilities, with reported AUROC values between 0.85 and 0.91, surpassing traditional clinical scoring systems. Furthermore, their built-in feature importance mechanisms improve interpretability, making them well-suited for clinical decision support systems.

[8] Gradient boosting methods, such as XGBoost and LightGBM, have been extensively investigated for predicting sepsis. These models are adept at identifying intricate nonlinear relationships and interactions among clinical variables, achieving impressive predictive accuracy in the initial detection of sepsis. Numerous studies indicate that AUROC values surpass 0.90 on standard ICU datasets. Nevertheless, their dependence on extensive hyperparameter optimization and sensitivity to data imbalance and outliers may restrict their scalability and real-time use in critical care environments.

[9] Deep learning architectures like Long Short-Term Memory networks, convolutional neural networks, and Transformer-based models have been suggested for the initial detection of sepsis using multivariate time-series information. These models excel at capturing temporal dependencies and intricate physiological patterns, frequently attaining AUROC

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

values as high as 0.95 in controlled experimental settings. However, despite their inspiring performance, deep learning models require extensive labeled datasets, considerable computational power, and face challenges regarding interpretability. These issues create difficulties for implementation at the bedside and getting in clinical settings.

[10] Dealing with absent and inconsistent clinical data continues to be a significant hurdle for AI-driven sepsis prediction systems. Previous research has utilized methods like mean imputation, k-nearest neighbor imputation, and forward-filling techniques, which could introduce bias if outliers are present. Recent research highlights the need for robust imputation techniques paired with interpretable models to enhance dependability and trust in clinical settings. Tree-based models, especially Random Forests combined with explainable AI methods like SHAP, present a well-rounded solution by offering both excellent predictive capabilities and clear decision-making transparency.

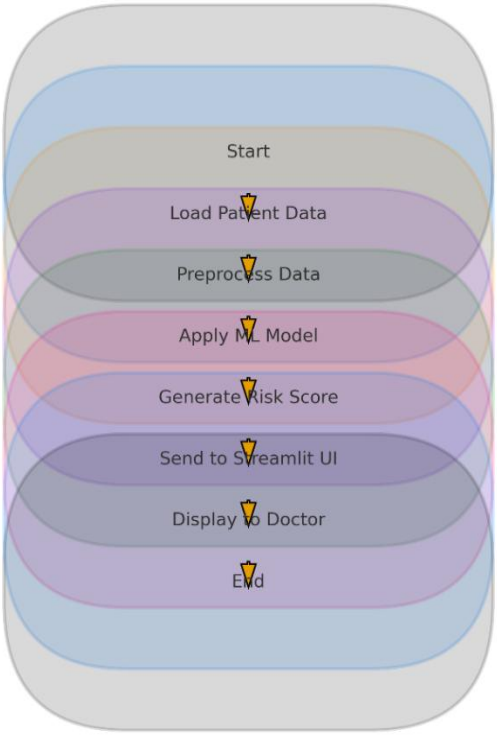


Figure.1 System Architecture and Workflow for AI-Based Early Sepsis Detection.

The above figure. 1 depicts the complete workflow of the proposed system, starting with the collection and processing of patient data, followed by machine learning-based risk assessment and real-time visualization to aid clinical decision-making.

III.PROBLEM STATEMENT

The diagnosis of sepsis often occurs only after a significant decline in physiological function has already taken place. This delay is largely attributed to the limitations of conventional diagnostic scoring systems like SIRS,SOFA, and qSOFA. These traditional approaches rely on fixed threshold values that may not accurately detention the dynamic and often rapid evolution of physiological changes associated with sepsis.

As a result, healthcare providers may overlook critical early cautionary signals, which can lead to a delayed response in treatment and an increase in patient mortality rates. The failure to recognize initial symptoms and subtle indicators of sepsis can be particularly concerning, as timely intervention is crucial for improving outcomes.

Given these challenges, there is a insistent need for an automated monitoring system that controls machine learning techniques. Such a system could continuously analyze patient data in real-time, enabling the identification of sepsis at an earlier stage. By integrating advanced algorithms and predictive analytics, this technology could significantly improve clinicians' ability to detect the onset of sepsis, leading to prompt intervention and potentially saving lives. Consequently, creating and executing automated systems for identifying sepsis signifies an important progress in the effort to enhance patient care and outcomes.

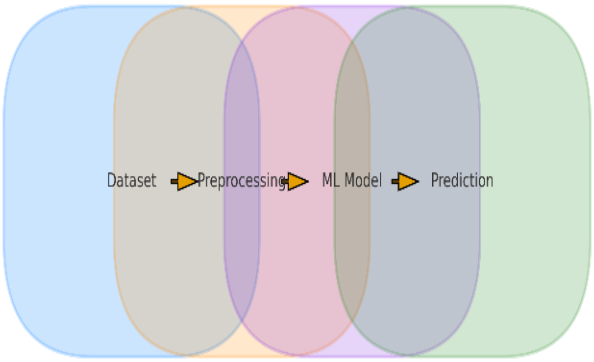


Figure .2 Overall Machine Learning Pipeline for Early Sepsis Prediction.

Figure 2 illustrates the step-by-step machine learning pipeline, emphasizing the progression from gathering and processing the dataset to training the model and ultimately predicting the risk of sepsis.

IV.OBJECTIVES

The objectives of the proposed system are:

- To create a machine learning model aimed at the early diagnosis of sepsis.
- To preprocess and examine ICU patient datasets for precise feature selection.
- To evaluate the performance of different classifiers to determine the most effective model.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

- To develop a real-time visualization dashboard for monitoring ICU patients.
- To aid in clinical intervention and decision-making to enhance patient outcomes.

V.DATA COLLECTION AND INTEGRATION

The system utilizes the PhysioNet 2019 Challenge dataset, which contains greater than 40,000 patient records from ICUs across various hospitals. Each record features 41 attributes, including heart rate, blood pressure, respiratory rate, temperature, and blood markers. The data preprocessing steps included addressing missing values, normalizing features, and employing SMOTE to balance the classes. A MySQL database was implemented to structure the data for training and testing, allowing for efficient access through model execution.

VI.PROPOSED SYSTEM ARCHITECTURE

The suggested architecture adopts a modular outline that includes data preprocessing, feature extraction, model training, and prediction components. The process is structured into separate phases: Dataset → Preprocessing → ML Model → Prediction.

TABLE I
SYSTEM ARCHITECTURE MODULES

Module	Description
Data Acquisition	Collects ICU patient vitals and lab data.
Data Preprocessing	Cleans and normalizes input data, handling missing data, and outlier removal.
Feature Selection	Identifies variables crucial for sepsis onset, such as ICU length of stay (ICULOS), heart rate (HR), respiration rate (Resp), temperature, and SpO levels.
Model Training	Applies ML algorithms to classify the risk of sepsis.
Prediction Module	Generates risk alerts based on the trained model.
Visualization Dashboard	Displays the risk score and vital sign trends in real time using a Streamlit-based web application.

This layered approach ensures accuracy, scalability, and adaptability to new datasets.

VII.METHODOLOGY

The methodology involves five key stages:

- Data preprocessing using adaptive imputation.
- Feature selection through importance ranking.
- Model training with multiple ML classifiers.
- Validation using stratified k-fold cross-validation.
- Performance evaluation.

The suggested imputation algorithm adapts by choosing mean or median values according to outlier thresholds, which enhances the reliability of the data. Random Forest was chosen as the most suitable classifier because of its excellent interpretability and ability to avoid overfitting. The models were developed in Python utilizing the Scikit-learn libraries.

A. Mathematical Model

Let D be the dataset of patient records:

$$D = \{(X_1,y_1),(X_2,y_2),\ldots,(X_n,y_n)\}$$

Where X represents feature vectors and y represents labels (sepsis or non-sepsis). The Random Forest classifier generates k decision trees:

$$\mathrm{RF}(X) = \text{majority vote}(T_1(X),T_2(X),\ldots,T_k(X))$$

Sepsis probability:

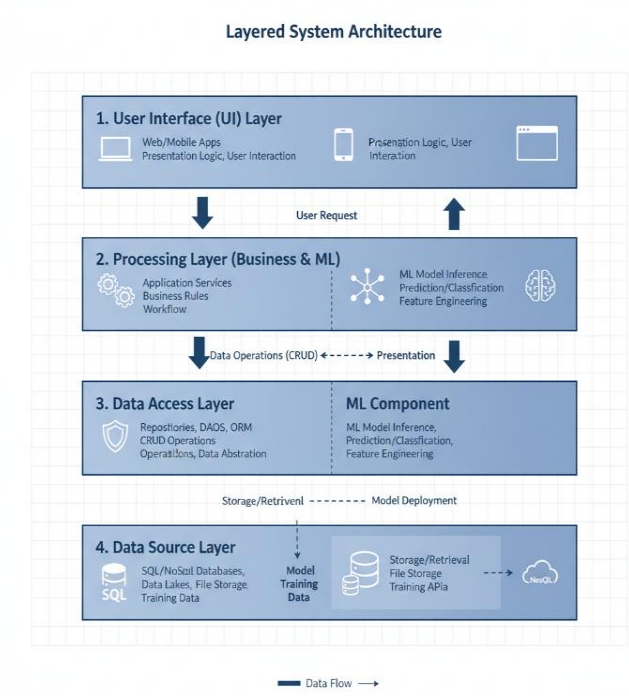
$$P(\text{sepsis}|X) = \frac{\text{Number of positive tree outputs}}{k}$$

VIII. EXPERIMENTAL RESULTS AND DISCUSSION

Experiments were conducted using a setup with an Intel Core i5 processor and 16GB of RAM, operating within a Python environment. The dataset was split into three parts: 75% designated for training, 12.5% for validation, and 12.5% for testing. The metrics utilized for evaluation included Accuracy , Precision , Recall, F1-Score , and AUC-ROC. The Random Forest model obtained an accuracy of 99.01%, markedly surpassing other classifiers.

- Logistic Regression: 95.3% accuracy
- Gradient Boosting: 97.4% accuracy
- Decision Tree: 96.2% accuracy

The superior classification performance was confirmed by figures representing ROC and Precision-Recall curves. The model successfully predicted sepsis onset up to 6 hours earlier than clinical diagnosis, demonstrating its potential for real-time ICU deployment.



International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Figure. 3 Layered Architecture of the Proposed AI-Based Sepsis Detection System.

Figure.3 depicts the multi-tier architecture that shows how the user interface, processing and machine learning elements, data access methods, and foundational data sources work together for comprehensive sepsis risk prediction.

IX.ADVANTAGES AND LIMITATIONS

A. Advantages

- Early detection before clinical deterioration.
- Reduces ICU mortality rates.
- Interpretable model with feature importance rankings.
- Scalable and adaptable to new patient data.

B. Limitations

- Requires continuous data monitoring infrastructure.
- Dataset variation across hospitals affects generalization.
- Performance depends on the accuracy of input vital signs.

CONCLUSION AND FUTURE ENHANCEMENTS

The study demonstrates the efficacy of machine learning in early detection of sepsis, offering a reliable and interpretable solution for clinical use. By integrating advanced preprocessing, adaptive imputation, and Random Forest classification, the proposed system achieves high predictive accuracy. Future work will focus on:

- Integration with IoT wearable monitoring devices.
- Use of deep learning and real-time sequential modelling (LSTM).
- Automated doctor alert system via mobile notification.
- Real-time integration with hospital information systems.
- Expansion to multi-hospital federated learning dataset.

ACKNOWLEDGMENT

The authors would like to acknowledge the providers of the PhysioNet 2019 Challenge dataset for making the data publicly available.

REFERENCES

[1] W. Song, S. Y. Jung, H. Baek, C. W. Choi, Y. H. Jung, and S. Yoo, "A predictive model based on machine learning for the early detection of late-onset neonatal sepsis: development and observational study," *JMIR Med. informatics*, vol. 8, no. 7, p. e15965, 2020.

[2] J. Li, M. Zhu, and L. Yan, "Predictive models of sepsis-associated acute kidney injury based on machine learning: a scoping review," *Ren. Fail.*, vol. 46, no. 2, p. 2380748, 2024.

[3] N. Kijpaisalratana, D. Sanglertsinlapachai, S. Techaratsami, K. Musikatavorn, and J. Saoraya, "Machine learning algorithms for early sepsis detection in the emergency department: a retrospective study," *Int. J. Med. Inform.*, vol. 160, p. 104689, 2022.

[4] R. J. van Wijk, S. B. Nagaraj, J. C. Ter Maaten, and H. R. Bouma, "Early sepsis prediction in the emergency department using machine learning," *Am. J. Emerg. Med.*, 2025.

[5] A. Gedikci Ondogan, M. Sargin, and K. Canoz, "Use of electronic. medical records in the digital healthcare system and its role in communication and medical information sharing among healthcare professionals," *Informatics Med. Unlocked*, vol. 42, p. 101373, 2023, doi: https://doi.org/10.1016/j.imu.2023.101373.

[6] S. M. Lauritsen *et al.*, "Early detection of sepsis utilizing deep learning on electronic health record event sequences," *Artif. Intell. Med.*, vol. 104, p. 101820, 2020.

[7] D. Wang *et al.*, "A machine. learning model for accurate prediction of sepsis in ICU patients," *Front. public Heal.*, vol. 9, p. 754348, 2021.

[8] C. Hu *et al.*, "Interpretable machine. learning for early prediction of prognosis in sepsis: a discovery and validation study," *Infect. Dis. Ther.*, vol. 11, no. 3, pp. 1117–1132, 2022.

[9] F. Li *et al.*, "Harnessing artificial intelligence in sepsis care: advances in early detection, personalized treatment, and real-time monitoring," *Front. Med.*, vol. 11, p. 1510792, 2025.

[10] H. Yilmaz Başer, T. Evran, and M. A. Cifci, "Machine Learning-Augmented Triage for Sepsis: Real-Time ICU Mortality Prediction Using SHAP-Explained Meta-Ensemble Models," *Biomedicines*, vol. 13, no. 6, p. 1449, 2025.