EARLY SEPSIS DETECTION USING ENSEMBLE MACHINE LEARNING MODELS

Dr. P. Eshawariah[1], Nagula Padma Sharadha[2], B. Niharika[3], B. Keerthana[4], D. Jhansi Lakshmi[5]

1.Assistant Professor, Department of Computer Science and Engineering (Data Science), Vignan's Institute of Management and Technology for Women

2,3,4,5 B-Tech Student, Department of Computer Science and Engineering (Data Science), Vignan's Institute of Management and Technology for Women
Email: padmasharadha6@ gmail.com

ABSTRACT

Sepsis is a critical medical condition characterized by the body's extreme response to infection, which can lead to organ failure and death if not detected at an early stage. Timely identification of sepsis in intensive care unit (ICU) patients remains a significant challenge due to the complexity and variability of clinical symptoms. This project presents a machine learning-based approach for early prediction of sepsis using ensemble learning techniques such as Random Forest and Bagging Classifier. The system utilizes patient clinical data, including vital signs and laboratory measurements, to build predictive models. The implementation involves essential preprocessing steps such as handling missing values and feature scaling to improve model performance. The trained models are evaluated using standard performance metrics including accuracy, precision, recall, and F1-score, with a focus on reducing false negatives to enhance patient safety. The final model is deployed using a Streamlit-based web application, allowing users to input patient data and obtain real-time predictions. Additionally, the project incorporates basic MLOps practices, including model saving, configuration management, and structured workflow, ensuring scalability and reproducibility. The developed system aims to assist healthcare professionals in early detection of sepsis, thereby improving decision-making, reducing mortality rates, and enhancing overall patient outcomes.

Keywords: Sepsis, Machine learning, Ensemble Learning, Random Forest, ICU, Clinical data, Early detection. Healthcare Analytics, Streamlit, MLOps, Bagging.

I. INTRODUCTION

Sepsis is a critical medical condition that arises when the body's immune response to infection leads to systemic inflammation, tissue damage, and potential organ failure. It is one of the leading causes of mortality worldwide, particularly in intensive care units (ICUs). Early detection is essential, as even a small delay in diagnosis and treatment can significantly increase the risk of death.

Traditional methods for detecting sepsis rely on manual monitoring of patient vital signs and laboratory results. These approaches are often time-consuming, error-prone, and unable to capture complex patterns present in large-scale clinical data. Due to the non-specific nature of sepsis symptoms, early diagnosis becomes highly challenging using conventional techniques.

With the advancement of data-driven healthcare, machine learning techniques have emerged as powerful tools for analysing clinical data and identifying

early signs of diseases. These techniques can process large datasets efficiently and provide accurate predictions, assisting healthcare professionals in decision-making.

In this work, an intelligent sepsis prediction system is developed using ensemble machine learning techniques. Algorithms such as Random Forest and Bagging Classifier are utilized due to their ability to handle high-dimensional data, reduce overfitting, and improve prediction accuracy. The system incorporates preprocessing steps to handle missing data and enhance model performance.

Furthermore, the system integrates Explainable Artificial Intelligence (XAI) methods to improve transparency and trust in predictions. Basic MLOps practices are also implemented to ensure scalability, reproducibility, and efficient deployment. A user-friendly interface is developed using Streamlit, enabling real-time prediction of sepsis based on patient input.

The proposed system aims to provide a reliable, efficient, and scalable solution for early detection of sepsis, thereby supporting timely medical intervention and improving patient care outcomes.

II. RELATED WORK

Several studies have explored the application of machine learning techniques for early detection of sepsis. Traditional approaches relied on statistical methods and rule-based systems, which often lacked accuracy and adaptability. With the emergence of machine learning, researchers have developed predictive models using clinical datasets such as electronic health records (EHRs).

Some studies have utilized deep learning models such as Artificial Neural Networks (ANN) and Recurrent Neural Networks (RNN) to capture complex temporal patterns in patient data. While these models demonstrate strong learning capabilities, they often require large datasets and high computational resources, making them less practical for real-time deployment.

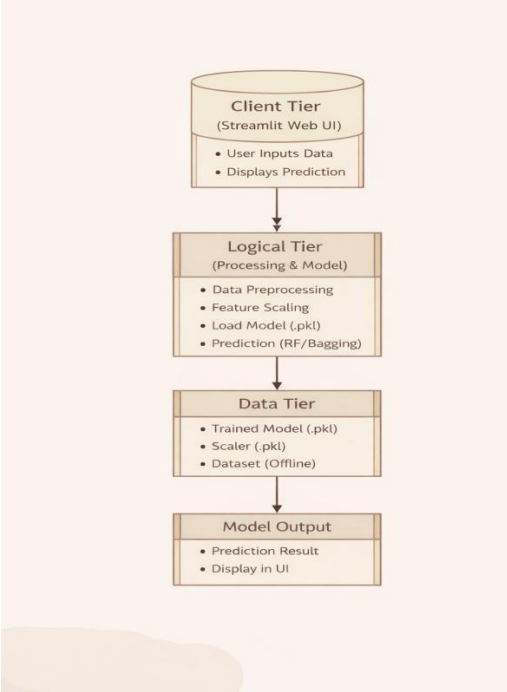
Ensemble learning techniques, such as Random Forest and Bagging, have gained popularity due to their ability to improve prediction accuracy and reduce model variance. These models combine multiple weak learners to produce robust predictions and are less prone to overfitting compared to individual models.

However, existing systems face several challenges, including class imbalance in medical datasets, lack of generalization across different patient populations, and difficulty in handling missing data. Additionally, many models lack interpretability, which limits their acceptance in clinical environments.

To address these limitations, the proposed system combines ensemble learning techniques with preprocessing methods, explainable AI tools, and MLOps practices. This integrated approach enhances prediction accuracy, improves model transparency, and ensures practical usability in healthcare settings.

III. SYSTEM ARCHITECTURE

The proposed Early Sepsis Prediction System is designed using a modular and scalable architecture that ensures efficient data processing, accurate prediction, and real-time usability. The system follows a three-tier architecture, consisting of the Client Layer, Processing Layer, and Data Layer.



A. Client Layer (User Interface)
The client layer provides an interactive interface for users, primarily healthcare professionals. A web-based dashboard is developed using Streamlit, allowing users to input patient clinical parameters such as heart rate, temperature, respiration rate, and laboratory values.

The interface is designed to be simple and user-friendly, enabling quick data entry and instant visualization of prediction results. The output is displayed as a binary classification indicating whether the patient is likely to develop sepsis or not.

B. Processing Layer (Business Logic)
This layer is responsible for the core functionality of the system and includes data preprocessing, model execution, and prediction generation.

Data Preprocessing Module
Clinical data often contains missing or inconsistent values. Therefore, preprocessing steps such as handling missing values, normalization, and feature scaling are applied to improve data quality and model performance.
Model Module
The system utilizes ensemble learning algorithms, specifically Random Forest and Bagging Classifier. These models are trained on historical clinical data and stored as serialized files. During prediction, the processed input data is passed to the trained model to generate results.
Prediction Engine
The prediction engine processes the input data and produces an output label (Sepsis / Non-Sepsis). It ensures fast and accurate inference suitable for real-time applications.

C. Data Layer
The data layer manages storage and retrieval of essential components required for the system. Trained machine learning models stored as .pkl files. Preprocessing scaler for feature transformation. Historical dataset used during training. This separation ensures scalability and easy maintenance of the system.

D. System Workflow
The overall workflow of the system is as follows:

- User inputs patient data through the interface
- Data is pre-processed and validated
- Pre-trained model is loaded
- Prediction is generated
- Result is displayed to the user

This structured workflow ensures smooth data flow and efficient prediction.

IV. IMPLEMENTATION AND DATASET

The system is implemented using modern machine learning tools and follows a structured development approach to ensure scalability and reproducibility.

A. Environment Setup
The system is developed using Python due to its extensive support for data analysis and machine learning. Libraries such as Pandas, NumPy, and Scikit-learn are used

for preprocessing and model development. Streamlit is used to build the web-based interface. Development is carried out using tools like Jupyter Notebook and Visual Studio Code. The trained models are saved using Joblib for efficient reuse and deployment.

B. User Interface Implementation

A Streamlit-based dashboard is developed to provide an interactive user experience. The interface allows users to:

- Enter patient clinical data
- Submit data for prediction
- View results instantly

The interface is designed with simplicity and clarity to ensure usability in clinical environments.

C. Model Implementation

The system uses ensemble learning techniques for prediction:

- Random Forest Classifier: Builds multiple decision trees and combines their outputs to improve accuracy and reduce overfitting.
- Bagging Classifier: Enhances stability and generalization by training multiple models on different subsets of data.

Both models are trained using pre-processed clinical data and evaluated using standard performance metrics.

D. Dataset Description

The dataset used in this project is derived from publicly available clinical data sources and contains patient records with vital signs and laboratory measurements.

Key features include:

- Heart rate
- Temperature
- Respiration rate
- Blood pressure
- Oxygen saturation
- Other clinical parameters

The dataset includes a target variable indicating whether a patient developed sepsis, enabling supervised learning.

E. Data Preprocessing and Validation

Data preprocessing is a critical step in improving model performance. The following techniques are applied:

- Handling missing values using imputation methods
- Feature scaling using standardization
- Balancing data using resampling techniques

Additionally, data validation is performed using automated frameworks to ensure data consistency and reliability.

F. Deployment

The trained model is deployed using a Streamlit-based web application. The application allows real-time prediction by accepting user input and generating results instantly.

The deployment ensures accessibility, scalability, and ease of use without requiring complex setup.

V. RESULTS AND DISCUSSION

The proposed system is evaluated using multiple performance metrics to assess its effectiveness in predicting sepsis.

A. Functional Results

All core functionalities of the system are successfully implemented, including:

- Data preprocessing
- Model prediction
- Real-time interface
- Output visualization

The system operates efficiently and produces accurate predictions.

B. Model Performance Analysis

The performance of different models is compared using evaluation metrics such as accuracy, precision, recall, F1-score, and AUC-ROC.

- Random Forest achieved the highest accuracy (approximately 96%)
- Bagging Classifier also performed well with accuracy above 95%
- Deep learning models showed lower accuracy (~76%)

The high recall values of ensemble models indicate their effectiveness in detecting sepsis cases, which is crucial in healthcare applications.

C. Comparative Analysis

A comparison between traditional approaches and the proposed system shows significant improvements:

Table-1: Comparison between Traditional system and Proposed system.

System Type	Accuracy	Scalability	Reliability
Traditional Methods	Low	Low	Moderate
Deep Learning Models	Medium	High	Moderate
Proposed System	High	High	High

The proposed system demonstrates superior performance in terms of accuracy and reliability.

D. Explainable AI Analysis

To improve interpretability, Explainable AI techniques are integrated:

- SHAP: Identifies feature importance and impact on predictions

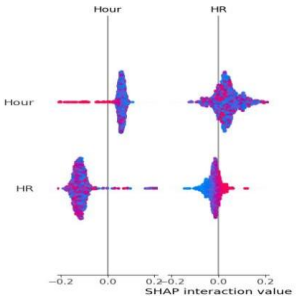


Figure-2: SHAP Interaction Plot Showing Feature Impact on Sepsis Prediction

- LIME: Explains individual predictions

Figure-3: LIME Explanation for Individual Prediction

These techniques help healthcare professionals understand the reasoning behind predictions, increasing trust in the system.

E. Performance Discussion

The results clearly indicate that ensemble models outperform deep learning models in this application. The system achieves high accuracy, low error rates, and fast response time.

The integration of explainability and validation further enhances the reliability of the system, making it suitable for real-world healthcare deployment.

VI. CONCLUSION

This paper presents an efficient machine learning-based system for early prediction of sepsis using ensemble learning techniques. The proposed approach utilizes Random Forest and Bagging Classifier models to achieve high accuracy and reliability in prediction.

The system successfully addresses challenges such as data preprocessing, model generalization, and real-time deployment. Experimental results demonstrate that ensemble models outperform deep learning approaches in terms of accuracy, recall, and stability.

The integration of Explainable AI techniques improves transparency, while MLOps practices ensure scalability and reproducibility. The deployment of the system using a web-based interface makes it practical and accessible for healthcare professionals.

Overall, the proposed system provides a reliable and efficient solution for early detection of sepsis, enabling timely medical intervention and improving patient outcomes.

Future Work :

Future enhancements may include:

- Integration with real-time hospital systems
- Implementation of advanced deep learning models
- Development of mobile-based applications
- Improved data handling and feature engineering

REFERENCES

[1] David W Shimabukaro, Christopher W Bardaon, Mitchell D Feldman, "Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial," *Journal of Critical Care*, vol. 36, pp. 123-131, 2016.

[2] Michael J. Pettinati, Geng Bo Chen, Kuldeep Singh Rajput, “Practical Machine Learning based Sepsis Prediction” , Computer Methods and Programs in Biomedicine, vol 194, pp. 105480, 2020.

[3] Xian Chuan Wang, Zhiyi Wang, Jie Weng, "A New Effective Machine Learning Framework for Sepsis Diagnosis," Multimedia Tools and Applications, vol. 81, pp. 9511-9530, 2022.

[4] Heather M. Ginestra, Jennifer C. MD, "A Machine Learning Algorithm to Predict Severe Sepsis and Septic Shock: Development, Implementation, and Impact on Clinical Practice," Critical Care Medicine, vol. 44, no. 3, pp. 368-374, 2016.

[5] Lucas M. Fleuren, Thomas L. T. Klausch, “Machine Learning Algorithm for the prediction of Sepsis: a systematic review and meta-analysis of diagnostic test accuracy”. Critical Care, vol. 24, no. 1, pp. 1-18, 2020.

[6] Dong Wang, Jinbo Li, Yali Sun, "A Machine Learning Model for Accurate Prediction of Sepsis in ICU Patients," Journal of Critical Care, vol. 64, pp. 1-8, 2021.

[7] Wenqian Shen, Guanjun Wang, “Early Prediction of Sepsis based on Machine Learning Algorithm”, Journal of Biomedical Informatics, vol. 112, p. 103519, 2020.

[8] Daniel Roberto, Alessio Singnori, “Early Detection of Sepsis with Machine Learning Techniques: AA brief Clinical Perspective”, Journal of Critical Care, vol. 54, pp. 184-188, 2019.

[9] Pankaj Chaudhary, Deepak Kumar, "Outcome Prediction of Patients for Different Stages of Sepsis Using Machine Learning Models," Health Information Science and Systems, vol. 9, no. 1, pp. 1-11, 2021.