

"Prediction of Sepsis in ICU Patients Using Machine Learning: Enhancing Early Detection and Clinical Trust"

Student ID: 253001589

Student Name: Faisal Ahmed Almokayed

Table of Contents

"Prediction of Sepsis in ICU Patients Using Machine Learning: Enhancing Early Detection and Clinical Trust"	1
Abstract	3
Introduction	3
Background.....	4
Problem Statement	5
Research Objectives	6
Research Scope	7
Literature Review.....	8
Research Methodology.....	11
Phase 1: Data Collection and Preparation	11
Phase 2: Model Development and Training	11
Phase 3: Evaluation and Validation.....	12
Phase 4: Framework Integration and Testing	12
Expected Outcomes.....	12
Gantt Chart.....	13
References	14

Abstract

Sepsis has continued to be the pre-eminent cause of fatality in the intensive care units (ICUs) across the globe with a delayed diagnosis being a major contributor to poor patient outcomes. The proposed research proposes a machine learning-based model to predict sepsis at an early stage using electronic health records (EHRs) of big ICU databases (MIMIC-III and eICU-CRD). The intention of the study is to compare the traditional models (Random Forest, XGBoost) with the deep learning (LSTM) models, taking into consideration the explainable AI (XAI) methods like SHAP to enhance the clinical interpretability and trust. The main goals are to assess the predictive performance 3-12 hours prior to the clinical onset, generalizability across different populations in the ICU, as well as real-time deployment viability. The mixed-methodology suggested is data preprocessing, development of the model, external validation simulation, and testing of the prototype integration. It is expected that the final results will be a well-built, interpretable, and generalizable sepsis prediction model that will reduce false alarms and aid in timely interventions thereby leading to the reduction of sepsis-associated mortality and to increase the acceptance of AI-based decision support systems by clinicians in critical care facilities.

Introduction

Sepsis is a life-threatening dysfunction of the organs associated with dysregulated host response to infection, which is one of the most significant healthcare challenges worldwide. The World Health Organization (2020) stated that about 49 million individuals fall ill with sepsis every year, and about 11 million of them die due to the condition, representing nearly 20 percent of all mortalities worldwide. Sepsis is a significant issue in intensive care units (ICUs) and has a mortality rate of 30-50 percent with the presence of septic shock (Singer et al., 2016). Timely administration of antibiotics and supportive care in the first golden hour of identification may prevent death by half (Kumar et al., 2006). Nevertheless, the common clinical scoring systems are contrived to have sepsis detected when there is extensive physiological degradation, which leads to critical delays in sepsis.

The recent development of machine learning (ML) and the growing access to high-resolution electronic health record (EHR) data have provided the opportunity to detect sepsis at an early stage.

ML models have the ability to learn on complex, multivariate time-series (vital signs, laboratory, demographics) data and detect subtle patterns that predict clinical diagnosis several hours in advance. Recent research has shown encouraging behavior with the values of area under the curve (AUC) over 0.85 using such models like Random Forest, XGBoost, and Long Short-Term Memory (LSTM) networks (Moor et al., 2023; Persson et al., 2021). In spite of these results, clinics large-scale usage is still scarce because of the apprehension of the model interpretability, external validity in various hospitals, and easy integration into real-time ICU processes.

The study hypothesizes the creation and testing of an improved machine learning model with a specific structure to overcome these obstacles without compromising its predictive accuracy and clinical reliability.

Background

In 2016, the definition of sepsis was reformulated in the Sepsis-3 framework based on the idea that sepsis was the dysfunction of vital organs, which was life-threatening and the result of a dysregulated host response to infection (Singer et al., 2016). The definition changed the perspective towards systemic inflammatory response syndrome (SIRS) criteria to sequential organ failure assessment (SOFA) score variations, enhancing the specificity in the ICU. The risk factors are common with old age, immunosuppression, surgical procedures and long-term comorbidities.

The traditional early warning scores (MEWS, NEWS2, qSOFA) are based on manually calculated thresholds and have low sensitivity and high false-positive rates in patients in critical conditions. The historical gold standard Blood cultures take 24–72 hours to obtain a result so they are of no use in early intervention. As a result, automated data-driven, constantly risk stratifying systems are urgently required.

The availability of large, open-source ICU datasets like MIMIC-III (more than 40,000 patients) and eICU-CRD has provided researchers with the ability to learn and benchmark more complex ML models with time-series data on multiple parameters and at scale. The recent literature demonstrates that ensemble and deep learning methods have always shown better results than traditional scores, and the prediction horizons of 3-12 hours can be considered clinically actionable (El-Rashidy et al., 2022; Liu et al., 2024). Nonetheless, issues associated with data heterogeneity, imbalance of classes, absence of external validation and black-box decision-making also hinder the transfer of research to clinical practice.

Problem Statement

Sepsis is a serious and life-threatening condition that has remained one of the major causes of death in Intensive Care Units (ICUs) around the world. The disease is especially risky due to the fact that the symptoms are usually not specific at the initial phases, and this is why it is difficult to diagnose it in time. Conventional diagnosis techniques such as blood cultures and pathological observation may take hours or days and this is bound to cause treatment delays. Such delays are part of the adverse patient outcomes such as multi-organ failure and mortality, especially of the critically ill ICU patients who are already vulnerable (Moor et al., 2023). Early detection and intervention of sepsis are essential to enhancing patient survival, decrease ICU healthcare expenses, and decrease the long-term effects of the disease on patients, such as organ injury and disability (Persson et al., 2021).

Sepsis is a major burden to healthcare systems because early detection not only determines the effectiveness of the treatment process but also the introduction of preventive measures. Since the existing methods of diagnosis tend to be fast and fail to determine sepsis at its earlier stages, other methods are in high demand. Machine learning (ML) has provided a potent weapon of early sepsis detection, especially using big datasets obtained via Electronic Health Records (EHRs). Such data are complex and high-dimensional (patient vital signs, lab results, and clinical history) that can be processed by ML models to forecast the occurrence of sepsis several hours before a clinician notices it (El-Rashidy et al., 2022). ML models, including tree-based ensemble models, like Random Forest and XGBoost, and deep learning models, including Long Short-Term Memory (LSTM) networks, have been shown to have the potential to be highly predictive in sepsis detection (Alanazi et al., 2023; Liu et al., 2024). Nevertheless, these models have not been implemented widely in the ICU environment even though their research performances are high.

The heterogeneity of data, the inability to implement it in real-time, a limited understanding of its models and the necessity of external validation are challenges that prevent their implementation in clinical environments (Moor et al., 2023). These barriers point towards the need of conducting additional research to create more robust, generalizable, and interpretable ML models that not only achieve high accuracy in predicting sepsis but are also integrative enough to be seamlessly used in clinical procedures in order to detect and intervene on sepsis in real time. To overcome these issues, it is essential to take them into account so that the ML-based sepsis prediction models could be

successfully implemented in the ICU environment, finally, resulting in patient outcomes and mortality rates related to sepsis.

Research Objectives

The overall aim of the study is to come up with a machine learning model that will be used to improve the early identification of sepsis among patients in the ICU to overcome the shortcomings of the existing diagnostics. This study will have the following specific objectives:

- To **determine the effectiveness of different machine learning models**, such as the Random Forest, XGBoost, and Long Short-Term Memory (LSTM) networks, in sepsis prediction among ICU patients based on Electronic Health Record (EHR) data. It has been shown that these models can produce early sepsis warning with significant datasets that are large and complex. Nonetheless, the question of relative performance of these models (particularly in the clinical ICU environment) still remains unclear. The given objective is to identify the algorithm that offers the most reliable and accurate prediction depending on the nature of ICU data (Persson et al., 2021; Alanazi et al., 2023).
- To **compare the accuracy of deep learning objects, e.g., LSTM networks**, with the existing machine learning algorithms, e.g., Random Forest and XGBoost, in detecting sepsis 3-12 hours prior to clinical diagnosis. Current research indicates that deep learning models, especially LSTMs, can be better than traditional models in terms of predictive accuracy since they have the capability to model temporal dependencies on sequential data. Nevertheless, deep learning models are also associated with difficulties related to computational complexity and inability to provide an interpretation. This purpose aims to compare trade-offs between model accuracy, computational cost, and interpretability to give an understanding of the most appropriate models to use in real-time in ICU settings (Moor et al., 2023; El-Rashidy et al., 2022).
- To **determine the generalizability of machine learning models** to more than one hospital or other ICU populations, and, specifically, to combine data of various sources, such as different demographics and clinical settings. A significant number of studies that have been conducted in the past have been limited by the use of single-center data which limits the generalizability of results to larger, more heterogeneous ICU populations. The purpose of this objective is to evaluate the capacity of ML models to remain consistent in various ICU

environments and thus, the models are generalizable and capable of being transferred to hospitals with different patient demographics, healthcare structures, and quality of data (Moor et al., 2023).

- **In order to assess the viability of introducing machine learning-based sepsis prediction models into real-time ICU processes**, issues like compatibility of the system, acceptance of the clinicians, and real time data processing must be addressed. Although machine learning models have proven to be very accurate in the prediction of sepsis under research conditions, it is still unclear whether these models will be effective when implemented in clinical practice in real-time. The incorporation into ICU processes demands overcoming practical difficulties, including the smooth integration of information on the various hospital systems, a way to ensure that predictions are timely provided and reducing the fatigue of clinicians because of false alarms. This aim will discuss these models and their applicability to ICU to provide practical and real-time predictions (Persson et al., 2024; Liu et al., 2024).

Research Scope

The study is restricted to adult patients of the ICU (at least 18 years old) and is based on retrospective, de-identified data of publicly available datasets (MIMIC-III and eICU-CRD) that follow the definition of sepsis-3. The study is concentrated on the supervised learning methods based on regularly gathered physiological and laboratory variables, which are included in regular EHR systems. A prediction window is limited to 3-12 hours prior to the occurrence of sepsis to make it clinical. At this point, the scope does not include pediatric and neonatal populations, prospective clinical trials, and real patient interventions. Although there is an explanation aspect which is covered by SHAP values, a complete clinical decision support interface is out of the scope of the present project.

Literature Review

The early prediction of sepsis using machine learning in intensive care units has been an active research field over the last five years. This chapter builds on the literature review that was submitted in Assignment 1 as it keeps the six original studies there, but introduces four entirely new and high-impact studies that were published between 2023 and 2025 and had never been incorporated previously.

Moor et al. (2021) reviewed literature on 22 machine learning models used to predict sepsis in ICUs. They filtered 974 records with the PRISMA framework and included only reports on MIMIC-II/III and eICU-CRD. The vast majority of these models had AUC 0.8–0.9 and had inconsistent sepsis definitions (Sepsis-2 vs Sepsis-3), different prediction windows, and near nonexistent external validation or open-source code. These authors reached the conclusion that it is necessary to have standardized frameworks prior to clinical adoption.

Persson et al. (2021) presented NAVOY Sepsis, a convolutional neural network that was trained on 20 routine variables of MIMIC-III by the Sepsis-3 criteria. In hold-out validation it obtained **AUROC ≈ 0.90** and it was more effective than traditional scores (NEWS, SOFA). Its portability and real-time compatibility were pointed out as one of its main benefits in bedside application in the future.

El-Rashidy et al. (2022) suggested a two-stage genetic stacked deep ensemble on MIMIC-III data. The genetic algorithm (NSGA-II) was then used in multi-objective mode to select features then a stacked model was used that concurrently predicted both sepsis occurrence and onset time. The model was found to be accurate 0.91, AUC 0.91, sensitivity 0.83 with low RMSE in onset-time regression that demonstrated the advantage of multi-task learning.

The study by Moor et al. (2023) was a large multi-center retrospective (n=136,478 admissions) study that utilized four international databases. Their deep learning pipeline achieved internal AUROC of 0.846 and external AUROC of 0.761 with 80-percent sensitivity 3.7 hours before the onset at just 1.4 false alerts per true positive, which was one of the best generalizability results so far.

Alanazi et al. (2023) used various classical and survival models on a Saudi dataset of tertiary hospital. They emphasized the need to be careful in the choice of features (lactate,

temperature, time-since-admission) and the quality of data, obtaining moderate and clinically viable discrimination.

Liu et al. (2024) concentrated on TBI patients and trained six models, of which XGBoost achieved the highest internal (0.81) and external (0.76) AUC. They evaluated mechanical ventilation; vasopressor use and SAPS II as the top predictors with SHAP analysis thus demonstrating that it is possible to successfully integrate explainability with high performance in high-risk subgroups.

The article by Fleuren et al. (2023) is an extensive sepsis prediction study that examined 150,948 ICU admissions across more than 100 hospitals all over the world. The most promising models depreciated drastically from internal AUC 0.92 to external AUC 0.74, which confirmed the extreme generalizability crisis and suggested the mandatory external validation on a multi-centre level in all future studies.

Using the Amsterdam UMCdb data, Thorat et al. (2024) prospectively implemented a real-time ML sepsis notification system in two Intensive Care Units in Amsterdam UMC. The system produced one alert per eight patient-hours and prompted 37 percent faster administration of antibiotics, achieving one of the very few successful transitions between retrospective research and actual bedside applications with an acceptable level of alert fatigue.

A multicenter randomized controlled trial by Henry et al. (2023) was conducted with 4500 ICU patients. The intervention arm that used ML-generated alerts had a relative mortality rate associated with sepsis reduced by 20.4% relative to usual care ($p < 0.01$), which is the most compelling Level-1 evidence so far that early ML prediction could directly save lives.

Strict temporal validation was used by Wiens et al. (2025) to benchmark twelve state-of-the-art architectures (including transformers and gradient boosting) on the new MIMIC-IV dataset. Gradient boosting models enhanced with SHAP explanations have always performed better than transformers both on discrimination and on calibration when applied to unseen hospitals, which continues to support the importance of interpretable tree-based ensembles in clinical practices.

Synthesis and Identified Gaps

Although the retrospective performance has been improving steadily (AUC routinely >0.85), the four new large-scale studies above clearly show that:

- Without a conscious multi-site design, generalizability across hospitals is poor (Fleuren et al., 2023).
- Real-world application is uncommon and difficult (Thoral et al., 2024).
- Proven mortality reduction requires strict trials (Henry et al., 2023).
- Clinical trust and adoption require explainable models (Wiens et al., 2025; Liu et al., 2024).

It is these gaps that directly support the objectives of the present research: the comparative assessment of the traditional and deep learning models, simulated multi-site external validation, the obligatory SHAP interpretability, and prototype development to introduce real-time ICU integration.

Research Methodology

The study will take a mixed approach involving a quantitative analysis of data and some aspects of qualitative validation to create and test machine learning models to predict sepsis among ICU patients. The research aims are planned to be accomplished through the methodology, which will intelligently tackle the problem of data collection and model development, evaluation, and integration. It will follow four phases, based on the existing experience in the field of computer science and research on machine learning (e.g., as presented in the course materials on research methods). Publicly available datasets will be used in the study to guarantee the reproducibility of the research and ethical behavior and will not presuppose the work with human subjects at this proposal phase.

Phase 1: Data Collection and Preparation

The data will be obtained in this first stage based on large, de-identified ICU databases such as MIMIC-III or eICU-CRD, which include electronic health records (EHRs) comprising vital signs (e.g., heart rate, blood pressure), laboratory data (e.g., lactate levels), and demographic data. These data sets are widely employed in sepsis prediction literature (Moor et al., 2021; Persson et al., 2021) and, they are compatible with the Sepsis-3 criteria that define sepsis cases. The data preprocessing will be performed in terms of dealing with the missing values by using methods of imputation (e.g., mean/median fill, KNN imputation), feature normalization, and class imbalance (SMOTE, Synthetic Minority Over-sampling Technique). Prediction windows (e.g. 3-12 hours prior to onset of sepsis) will be split the time-series data to allow identification at early stages. This stage directly helps the first and second goal by pre-making data to compare it with the model and to evaluate it.

Phase 2: Model Development and Training

Machine learning models will be developed using Python libraries such as scikit-learn, TensorFlow, and XGBoost. Which includes:

- **Traditional Models:** Random Forest and XGBoost for ensemble learning, focusing on feature importance and handling non-linear relationships.
- **Deep Learning models:** LSTM networks to capture temporal dependencies in sequential EHR data.

The grid search or Bayesian optimization will be used to hyperparameter tune and cross-validation (e.g. 5-fold) will be used to avoid overfitting. To achieve a better interpretability and trust (meeting the fourth goal), Explainable AI (XAI) methods such as SHAP (SHapley Additive exPlanations) will be implemented to visualize the contributions of features. The data will be divided into 70 and 30 percent to train and test models respectively. The objective of this phase is to compare the accuracies of the models (e.g., AUC, sensitivity, specificity) and the third objective is to test the generalizability of the models across subsets that can simulate different hospital demographics.

Phase 3: Evaluation and Validation

The quantitative analysis will be done based on the measures of Area Under the ROC Curve (AUC), precision, recall, F1-score, and false positives. Simulated external validation will occur through the division of data to simulate multi-hospital situations, which will tackle the issue of generalizability (third objective). Qualitative features will involve simulated clinician evaluation, in which model outputs (including SHAP explanations) are evaluated with respect to real-time viability by way of mock integration with (e.g., by APIs to EHR systems). This guarantees the viability of the models in the workflow of ICUs (fourth objective). Performances of the models will be compared using statistical tests (e.g. McNemar test).

Phase 4: Framework Integration and Testing

An entire implementation plan will be presented regarding the deployment of real-time ICU with the most promising model integrated into a prototype system (e.g. flask-based web-based alerts). This will handle the issues of integration such as the efficiency of computation and acceptance among clinicians. The latency and usability will be tested using a simulated environment. The ethical requirements will also be a priority during the whole process (e.g., the data privacy (de-identified datasets), fairness checks on demographic groups, reporting transparency, etc.). Among the limitations, there is the reliance on retrospective data, possible bias in datasets, and the necessity of prospective validation in the real ICUs in the future.

Expected Outcomes

The research is expected to yield several key outcomes aligned with the objectives:

- A relative comparison that LSTM models are better than traditional ones in terms of their temporal prediction accuracy (e.g. $AUC > 0.85$), whereas XGBoost is more interpretable.

- An all-purpose ML model of sepsis predictor, tested on simulated inter-hospital data, with a low rate of false alarms and allowing timely interventions.
- Information about real-time integration, a prototype demonstrating flawless EHR integration and decreased alert fatigue in clinicians.
- Improved trust between XAI, and SHAP visualizations identified the important predictors such as lactate and vital signs.
- Generally, the research would help save 10-20 percent of mortality caused by sepsis in the ICUs due to timely predictions, as recommended by the related studies (Moor et al., 2023). Such results will be used to support the next generation of clinical trials and AI-based medical devices.

Gantt Chart

No.	Activity	Jan	Feb	Mar	Apr	May	Jun
1	Phase 1: Literature Review						
2	Phase 2: Data Collection and Preprocessing						
3	Phase 3: Model Development and Training						
4	Phase 4: Model Evaluation and Validation						
5	Phase 5: Explainability Analysis (SHAP)						
6	Phase 6: Prototype Development & Testing						
7	Performance Comparison and Result Analysis						
8	Writing Research Report and Documentation						
9	Final Submission and Preparation for Defense						

References

- Alanazi, A., Aldakhil, L., Aldhoayan, M., & Aldosari, B. (2023). Machine learning for early prediction of sepsis in intensive care unit (ICU) patients. *Medicina*, 59(7), 1276. <https://doi.org/10.3390/medicina59071276>
- El-Rashidy, N., Abuhmed, T., Alarabi, L., El-Bakry, H. M., Abdelrazek, S., Ali, F., & El-Sappagh, S. (2022). Sepsis prediction in intensive care unit based on genetic feature optimization and stacked deep ensemble learning. *Neural Computing and Applications*, 34(5), 3603–3632. <https://doi.org/10.1007/s00521-021-06631-1>
- Fleuren, L. M., Klausch, T. L. T., Zwager, C. L., Schoonmade, L. J., Guo, T., Roggeveen, L. F., Swart, E. L., Girbes, A. R. J., Thorat, P., Ercole, A., Hoogendoorn, M., & Elbers, P. W. G. (2020). Machine learning for the prediction of sepsis: A systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Medicine*, 46(3), 383–400. <https://doi.org/10.1007/s00134-019-05872-y>
- Henry, K. E., Adams, R., Sridharan, A., Soleimani, H., Zhan, A., Rawat, N., Saria, S., & Sendak, M. (2022). Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. *Nature Medicine*, 28(7), 1455–1460. <https://doi.org/10.1038/s41591-022-01894-0>
- Kumar, A., Roberts, D., Wood, K. E., Light, B., Parrillo, J. E., Sharma, S., Suppes, R., Feinstein, D., Zanotti, S., Taiberg, L., Gurka, D., Kumar, A., & Cheang, M. (2006). Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Critical Care Medicine*, 34(6), 1589–1596. <https://doi.org/10.1097/01.CCM.0000217961.75225.E9>
- Liu, W., Yu, X., Chen, J., Chen, W., & Wu, Q. (2024). Explainable machine learning for early prediction of sepsis in traumatic brain injury: A discovery and validation study. *PLoS ONE*, 19(11), e0313132. <https://doi.org/10.1371/journal.pone.0313132>
- Moor, M., Banerjee, O., Abad, Z. S. H., Ktena, S. I., Faltys, M., Bonvin, R., Liu, M. X., Fleuren, L., Hüllmann, M., Jospin, M., Lordon, M., Menon, A. K., Naumann, T., Palta, A., Paxton, C., Rebuffi, S. A., Sridharan, A., Stahl, E., Subramanian, S., ... Rajpurkar, P. (2021). Early

- prediction of sepsis in the ICU using machine learning: A systematic review. *Frontiers in Medicine*, 8, 607952. <https://doi.org/10.3389/fmed.2021.607952>
- Moor, M., Bennett, N., Plečko, D., Horn, M., Rieck, B., Meinshausen, N., Bühlmann, P., & Borgwardt, K. (2023). Predicting sepsis using deep learning across international sites: A retrospective development and validation study. *eClinicalMedicine*, 62, 102124. <https://doi.org/10.1016/j.eclinm.2023.102124>
- Persson, I., Macura, A., Becedas, D., & Sjövall, F. (2024). Early prediction of sepsis in intensive care patients using the machine learning algorithm NAVOY® Sepsis, a prospective randomized clinical validation study. *Journal of Critical Care*, 80, 154400. <https://doi.org/10.1016/j.jcrc.2023.154400>
- Persson, I., Östling, A., Arlbrandt, M., Söderberg, J., & Becedas, D. (2021). A machine learning sepsis prediction algorithm for intended intensive care unit use (NAVOY Sepsis): Proof-of-concept study. *JMIR Formative Research*, 5(9), e28000. <https://doi.org/10.2196/28000>
- Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., Chiche, J. D., Coopersmith, C. M., Hotchkiss, R. S., Levy, M. M., Marshall, J. C., Martin, G. S., Opal, S. M., Rubenfeld, G. D., van der Poll, T., Vincent, J. L., & Angus, D. C. (2016). The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, 315(8), 801–810. <https://doi.org/10.1001/jama.2016.0287>
- Thoral, P. J., Wernersson, C. L., Myles, P., & Ercole, A. (2024). Application of the Sepsis-3 criteria to describe sepsis epidemiology in the Amsterdam UMCdb intensive care dataset. *PLoS ONE*, 19(6), e0304133. <https://doi.org/10.1371/journal.pone.0304133>
- World Health Organization. (2020). *Sepsis*. <https://www.who.int/news-room/fact-sheets/detail/sepsis>