EARLY DETECTION OF SEPSIS FROM CLINICAL DATA USING MACHINE LEARNING

JOSMAN ESNEIDER RICO TORRES
DEISY TATIANA TORRES PEDRAZA

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍA FÍSICOMECÁNICAS
ESCUELA DE INGENIERÍA ELÉCTRICA, ELECTRÓNICA Y DE TELECOMUNICACIONES
INGENIERÍA ELECTRÓNICA
BUCARAMANGA
2025

EARLY DETECTION OF SEPSIS FROM CLINICAL DATA USING MACHINE LEARNING

JOSMAN ESNEIDER RICO TORRES
DEISY TATIANA TORRES PEDRAZA

Degree work presented as a requirement to qualify for the title of Electronic Engineer

Advisor
MSc. Camilo Andres Santos Ortiz

Co-advisor
MSc(c). Harold Hernando Rodriguez Rodriguez
Carlos Augusto Fajardo Ariza, Ph.D.

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍA FÍSICOMECÁNICAS
ESCUELA DE INGENIERÍA ELÉCTRICA, ELECTRÓNICA Y DE TELECOMUNICACIONES
INGENIERÍA ELECTRÓNICA
BUCARAMANGA
2025

DEDICATION

I dedicate this work to my mother, my greatest inspiration, for teaching me by her example to fight for my dreams, for never leaving me alone, and for reminding me that I can trust myself. This achievement is the result of her love, strength, and dedication. To my family, for believing in me, supporting me in every situation, and always being there for me at every stage of my life. It is an honor to be a source of pride for them and an example of perseverance. Finally, to those who walked alongside me on this journey, sharing the days of hard work, laughs, achievements, and difficult moments. They taught me not to give up, to be stronger, and were there for dreams that once seemed far away.

Josman Esneider Rico Torres

I dedicate this achievement to my mother, for being my guide, my role model, and my greatest source of inspiration. Her unconditional love, dedication, and constant support gave me the strength to reach this goal and successfully complete this stage. To my brother, for his constant support, his advice, for being a voice of encouragement in difficult times, and for always motivating me to keep going. To all the people who accompanied me in this process, thank you for your support and for being there when I needed it most. To my friends, for your constant companionship, for making the journey more bearable, and for being a source of encouragement and motivation at every moment.

Deisy Tatiana Torres Pedraza

ACKNOWLEDGMENTS

With gratitude in our hearts, we recognize those who made this journey a truly meaningful experience. Firstly, to our families for their unconditional love, for believing in us, and for being the support that encouraged us to continue, especially our mothers, for being the driving force in our lives and the reason that pushes us forward.

We would also like to extend our gratitude to our advisor Camilo Santos and co-advisors Harold Rodriguez and Carlos Fajardo for their guidance, commitment, and valuable contributions to the development of this research. Likewise, we would like to thank the CPS research group for opening their doors to us, allowing us to work with them, and for being a constant source of learning and professional inspiration.

Finally, we would like to express our sincere gratitude to all those who made this process an unforgettable memory, for accompanying us with commitment and dedication, to our friends, who with every gesture, advice, and smile left an indelible mark on our lives.

TABLE OF CONTENTS

LIST OF FIGURES

page.

LIST OF TABLES

page.

LIST OF APPENDICES

page.

RESUMEN

TÍTULO DETECCIÓN TEMPRANA DE SEPSIS A PARTIR DE DATOS CLÍNICOS USANDO APRENDIZAJE AUTOMÁTICO [*]

AUTOR: Josman Esneider Rico Torres y Deisy Tatiana Torres Pedraza [**]

PALABRAS CLAVE: Detección temprana, unidad de cuidados intensivos, aprendizaje automático, PhysioNet Challenge, segmentación temporal.

DESCRIPCIÓN: La sepsis es una disfunción orgánica potencialmente mortal causada por una respuesta desregulada del paciente ante una infección, lo que representa una prioridad de salud mundial con una estimación de 11 millones de muertes al año. Su detección temprana sigue siendo un reto fundamental en las unidades de cuidados intensivos (UCI), donde los síntomas iniciales inespecíficos, la gran heterogeneidad de los datos y los frecuentes valores faltantes en los registros médicos electrónicos a menudo retrasan la intervención y empeoran los resultados de los pacientes. En respuesta a ello, este estudio propone una novedosa metodología orientada al paciente para la identificación anticipada de sepsis mediante el aprendizaje automático en la base de datos de PhysioNet Challenge. Nuestra principal contribución radica en el desarrollo de un proceso de estructuración de datos temporales que incorpora una ventana de observación única de 21 horas alineada con el inicio de la sepsis, procesada a través de ventanas deslizantes de 6 horas para captar la evolución clínica, y que emplea la imputación multivariable utilizando LightGBM y optimización bayesiana. Entre los modelos de refuerzo, que alcanzaron colectivamente un AUC-ROC$>$0.93, CatBoost demostró el rendimiento más equilibrado, alcanzando la mayor sensibilidad (0,720) y puntuación de utilidad (0,668). Nuestros modelos no solo lograron un rendimiento competitivo frente a los métodos más avanzados, sino que también proporcionaron información clínica significativa a través del análisis SHAP, identificando la duración de la estancia en UCI, los niveles BUN y los patrones respiratorios como predictores clave. Estos resultados demuestran que nuestro flujo de preprocesamiento sistemático y sensible al tiempo ofrece un marco robusto para los sistemas de apoyo a la toma de decisiones clínicas.

---

[*] Tesis de pregrado

[**] Facultad de Ingenieria Fisicomecánicas. Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones. Director: MSc. Camilo Andrés Santos Ortiz. Codirectores: MSc(c). Harold Hernando Rodriguez Rodriguez, Carlos Augusto Fajardo Ariza. PhD.

ABSTRACT

TITLE: EARLY DETECTION OF SEPSIS FROM CLINICAL DATA USING MACHINE LEARNING [*]

AUTHOR: Josman Esneider Rico Torres and Deisy Tatiana Torres Pedraza [**]

Keywords: Early detection, intensive care units, machine learning, PhysioNet Challenge, temporal segmentation.

Description: Sepsis is a life-threatening organ dysfunction caused by a dysregulated host response to infection, representing a global health priority with an estimated 11 million deaths annually. Its early detection remains a critical challenge in intensive care units (ICUs), where nonspecific early symptoms, high data heterogeneity, and frequent missing values in electronic health records often delay intervention and worsen patient outcomes. In response, this study proposes a novel patient-oriented methodology for early sepsis detection using machine learning on the PhysioNet Challenge database. Our main contribution lies in the development of a comprehensive temporal data structuring pipeline that incorporates a unique 21-hour observation window aligned with sepsis onset, processed through 6-hour sliding windows to capture clinical evolution, and employs multivariate imputation using LightGBM and Bayesian optimization. Among the boosting models, which collectively achieved AUC-ROC $>$0.93, CatBoost demonstrated the most balanced performance, attaining the highest sensitivity (0.720) and utility score (0.668). Our models not only achieved competitive performance against state-of-the-art methods but also provided clinically meaningful insights through SHAP analysis, identifying ICU length of stay, BUN levels, and respiratory patterns as key predictors. These results demonstrate that our systematic and temporally-aware preprocessing pipeline offers a robust framework for clinical decision-support systems.

---

[*] Undergraduate Thesis

[**] Facultad de Ingenieria Fisicomecánicas. Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones. Advisor: MSc. Camilo Andrés Santos Ortiz. Co-advisors: MSc(c). Harold Hernando Rodriguez Rodriguez, Carlos Augusto Fajardo Ariza. PhD

INTRODUCTION

The conceptualization of sepsis has evolved through international consensus definitions. Sepsis-1 (1991) introduced the systemic inflammatory response syndrome (SIRS), while Sepsis-2 (2001) retained it despite its low specificity, adding clinical and hemodynamic variables[1]. Sepsis-3 (2016) redefined sepsis as life-threatening organ dysfunction from a dysregulated host response to infection[2], eliminating SIRS and severe sepsis, and adopting the Sequential Organ Failure Assessment (SOFA) score, which assesses six organ systems; an increase of ≥ 2 points indicates significant dysfunction[3].

Early detection is particularly difficult, as the initial symptoms are often nonspecific and can be confused with other common diseases, delaying medical intervention and worsening the patient's prognosis[4]. In 2020, there were 48.9 million cases of sepsis and 11 million associated deaths worldwide, accounting for nearly 20 % of global mortality[5]. The economic impact is also considerable: in high-income countries, costs can exceed

---

1 Roger C. Bone, Robert A. Balk, Frank B. Cerra et al.: *Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis*. En: *Chest* 101.6 (1992), págs. 1644-1655. DOI: 10.1378/chest.101.6.1644.

2 Mervyn Singer, Clifford S. Deutschman, Christopher W. Seymour et al.: *The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)*. En: *JAMA* 315.8 (2016), págs. 801-810. DOI: 10.1001/jama.2016.0287.

3 Simon Lambden et al.: *The SOFA score—development, utility and challenges of accurate assessment in clinical trials*. En: *Critical Care* 23.1 (nov. de 2019), pág. 374. DOI: 10.1186/s13054-019-2663-7.

4 Health Catalyst: *Optimizing Sepsis Care Improves Early Recognition and Outcomes*. Health Catalyst Success Story. 2025.

5 World Health Organization: *Sepsis*. https://www.who.int/news-room/fact-sheets/detail/sepsis. 2020.

$32,000 per patient, especially when sepsis is not diagnosed in time[6]. In Colombia, the situation is equally worrying: a study analyzing ICU admissions in 2019 reported that 10.36 % of patients developed sepsis and 36.7 % died[7]. Constant monitoring is hampered by the large volume of clinical data and variability between patients, as well as the presence of inconsistencies and missing values in records. These difficulties highlight the need for decision support tools that can process complex information and assist physicians in the early detection of sepsis. In this context, artificial intelligence, and in particular machine learning, has emerged as a promising alternative. These models can process large volumes of clinical data, detect subtle patterns, and generate timely predictions that often remain unnoticed by traditional methods.

A key driver of research in this field has been the PhysioNet/Computing in Cardiology Challenge 2019[8], which provided a publicly available dataset derived from the MIMIC-III clinical database and comprises a total of 40,336 ICU patient records collected from two hospital systems in the United States, including up to 40 variables such as vital signs, laboratory results, and demographic information. Of these, 2,932 correspond to septic cases and 37,404 to non-septic cases, labeled according to the Sepsis-3 definition.

The aim of this study is to develop a model for early sepsis detection by implementing a rigorous, clinically-grounded data preprocessing pipeline. This approach prioritizes careful patient selection, advanced handling of missing values, and temporal alignment to create

---

6 Elyse Ladbrook et al.: *A systematic review of the cost-impact of sepsis care bundles*. En: *Journal of Hospital Infection* (2025). Available under a Creative Commons license. ISSN: 0195-6701. DOI: 10.1016/j.jhin.2025.08.006.

7 Henry Oliveros et al.: *One-year survival of patients admitted for sepsis to intensive care units in Colombia*. En: *BMC Infectious Diseases* 24.1 (jul. de 2024), pág. 678. ISSN: 1471-2334. DOI: 10.1186/s12879-024-09584-7.

8 Matthew A. Reyna et al.: *A large, annotated dataset of sepsis patients from the PhysioNet/Computing in Cardiology Challenge 2019*. En: *PhysioNet* (2019). Available at https://physionet.org/challenge/2019/.

an analysis-ready dataset. This foundation enables the subsequent selection of a machine learning model using a broad set of metrics, including the utility score. This work focuses on overlooked aspects of data quality and evaluation rigour in order to provide a more reliable and clinically relevant predictive framework.

The structure of this document is organized as follows: this introductory section has presented the context, problem description, and general approach. The Objectives section outlines both the general and specific goals of the study. The Literature Review summarizes key research on machine learning approaches for early sepsis prediction. The Methods section describes the data preprocessing pipeline and methodological framework. The Results section reports the main experimental findings, while the Discussion and Conclusions section analyzes their implications, highlights study limitations, and provides concluding remarks. Finally, the Bibliography and Appendices include the references and supplementary material supporting this work.

1. OBJECTIVES

1.1. GENERAL OBJECTIVE

To develop a Machine Learning model for the early detection of sepsis based on clinical data.

1.2. SPECIFIC OBJECTIVES

To preprocess a public dataset to train a Machine Learning model for the early detection of sepsis using clinical data.

To select a Machine Learning model for the early detection of sepsis based on the study of the state of the art.

To evaluate the performance of the model using metrics such as AUC, ROC, Utility Score (PhysioNet 2019 Challenge metric), F1-score, and recall.

2. LITERATURE REVIEW

Incomplete and heterogeneous patient records remain one of the most persistent challenges in building reliable predictive models. The DACMI Challenge at IEEE ICHI 2019[9] provided a benchmark for addressing this issue through the imputation of ICU time-series data from MIMIC-III. Among the strongest contributions, Zhang et al.[10] proposed a LightGBM-based framework that captured temporal dependencies, Jesson et al.[11] introduced 3D-MICE combining multiple imputation with Gaussian processes, and Sun et al[12] developed MICE-DA, a deep autoencoder with temporal attention. These methods highlighted the importance of robust imputation and explicit modeling of missingness patterns.

Subsequent studies explored simpler strategies. Strickler et al.[13] used simple imputation strategies (mean, last valid entry, or zero). They reported a utility score of 0.83 and a sensitivity of 0.56 without applying changes in temporal labeling. Afterwards, they implemented label shifting to evaluate predictive performance at 1, 6, and 12 hours before diagnosis. Similarly, An Ensemble Machine Learning Model for the Early Detection of Sepsis imputed values using carry-forward or variable means, applied a 12-hour label shifting strategy, and retained 34 features with derived statistics, obtaining a utility score of 0.558 and an

---

[9] Yuan Luo y IEEE ICHI Organizers: *Data Analytics Challenge on Missing data Imputation (DACMI)*. `https://ewh.ieee.org/conf/ichi/2019/challenge.html`. 2019.

[10] et al. Zhang: *Evaluating the state of the art in missing data imputation for clinical data*. En: *Briefings in Bioinformatics* 23.1 (2022), bbab489. DOI: `tpi10.1093/bib/bbab489`.

[11] et al. Jesson: *3D-MICE: Multiple Imputation with Gaussian Processes for Clinical Time Series*. En: *IEEE ICHI 2019 Proceedings* (2019). Presented at DACMI Challenge.

[12] P. Sun: «MICE-DA: A MICE method with Data Augmentation for missing data imputation». En: *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. 2019. DOI: `10.1109/ichi.2019.8904724`.

[13] Ethan A. T. Strickler et al.: *Exploring a global interpretation mechanism for deep learning networks when predicting sepsis*. En: *Scientific Reports* 13.3067 (2023). DOI: `10.1038/s41598-023-30091-3`.

area under the curve (AUC) of 0.792.

The PhysioNet 2019 Challenge suffered from incomplete and heterogeneous patient records, which influenced the design and evaluation of predictive models. It nevertheless highlighted the methodological diversity in early sepsis prediction. Morrill et al.[14] achieved the best performance by applying mathematical signatures to represent time-series dynamics, combined with a gradient boosting model, reporting a utility score of approximately 0.433. Du et al.[15] employed gradient boosting trees with engineered clinical features, yielding a utility score of about 0.402. Zabihi et al[16] explicitly incorporated missingness patterns into an ensemble of XGBoost models, achieving a utility score of 0.400. All these studies, as well as the present work, are based on the PhysioNet/Computing in Cardiology Challenge 2019 dataset, providing a consistent benchmark that facilitates fair comparison across models. However, despite this shared foundation, important gaps remain.

Most works adopt imputation strategies based on simple criteria, such as forward-fill or mean replacement, often without medical justification and despite high levels of missingness. Patient selection frequently includes incomplete records, which undermines the robustness of early predictions. Furthermore, model evaluation is typically limited to a single algorithm or small ensembles, with restricted metrics that hinder a comprehensive assessment of performance.

---

[14] James Morrill et al.: «The Signature-Based Model for Early Detection of Sepsis From Electronic Health Records in the Intensive Care Unit». En: *Computing in Cardiology*. Vol. 46. 2019, págs. 1-4. DOI: 10.22489/CinC.2019.014.

[15] John Anda Du, Nadi Sadr y Philip de Chazal: «Automated Prediction of Sepsis Onset Using Gradient Boosted Decision Trees». En: *Computing in Cardiology*. Vol. 46. 2019, págs. 1-4.

[16] M. Zabihi, S. Kiranyaz y M. Gabbouj: «Sepsis Prediction in Intensive Care Unit Using Ensemble of XGBoost Models». En: *2019 Computing in Cardiology (CinC)*. Singapore, 2019, Page 1-Page 4. DOI: 10.22489/CinC.2019.238.

3. METHODS

3.1. TEMPORAL SEGMENTATION PER PATIENT

The PhysioNet Challenge provides two distinct datasets (A and B) derived from different hospital systems. As a first step, we combined both datasets into a single analysis-ready dataset. We then applied forward fill to laboratory variables, limited to 12 hours based on clinical guidance from a specialist and consistent with clinical practice where the median interval between measurements is 10–14 hours[17]. Next, we extracted 21-hour observation windows specifically designed to align with the PhysioNet Challenge's evaluation framework, thereby ensuring consistency with its reward–penalty scheme. For septic patients, windows were aligned with the first positive *SepsisLabel* value, including the 11 hours prior (always present) and 9 hours after. If a patient had at least 6 hours of data after the septic event (i.e., at least 18 hours of actual data in total), we completed the 21 hours by filling the remaining 1-3 hours hours with NaN. The same criteria were applied to non-septic patients, from whom we additionally selected the window with the highest data density.

Finally, we derived the *shock index* (HR/SBP), a cardiovascular marker widely studied in the medical literature, demonstrating clinical utility in the prognosis of sepsis[18].

---

[17] Walid Alali et al.: *Impact of laboratory test result availability on clinical decision-making: evidence from routine data*. En: *Scientific Reports* 12.1 (2022), pág. 22191. DOI: 10.1038/s41598-022-25961-1.

[18] Mohamed Y. Rady et al.: *Shock index as a marker of severity in patients with sepsis*. En: *Medicina Intensiva* 40.7 (2016), págs. 399-404. DOI: 10.1016/j.medin.2016.03.005.

3.2. TEMPORAL DYNAMICS OF FEATURES

To capture the temporal evolution of clinical variables, we implemented a 6-hour sliding window with a 1-hour stride, generating at each step a new record that summarized patient data over time. Figure 1 illustrates this approach within the complete processing pipeline, which generates at each step a new record that summarized patient data over time. Within each window, vital signs were characterized by minimum, maximum, mean, variance, and last recorded value, while laboratory variables were represented by their minimum and maximum. ICU stay was captured at the end of each interval, and constant variables (age, gender) remained unchanged.

The first 6-hour window (hours 0–5) created the initial record, and each subsequent shift added a new instance, resulting in 16 records per patient. Thus, each individual has multiple records, but data processing and splitting are always performed per patient, avoiding the mixing of records between different individuals. Each statistic was represented as a separate column, and all records were merged into a single structured dataset for analysis.

3.3. DATA CLEANING AND IMPUTATION

Patients with entirely missing vital signs or laboratory data, features with over 20 % missing values, and those failing to meet the 21-hour observation window criteria were excluded. As a result, a total of 4,760 patient records were discarded, yielding a final dataset of 35,576 patients (1,965 septic and 33,607 non-septic). Table 1 illustrates the 13 clinical variables selected for analysis, which were subsequently transformed into the 43 statistical features detailed in the previous subsection.

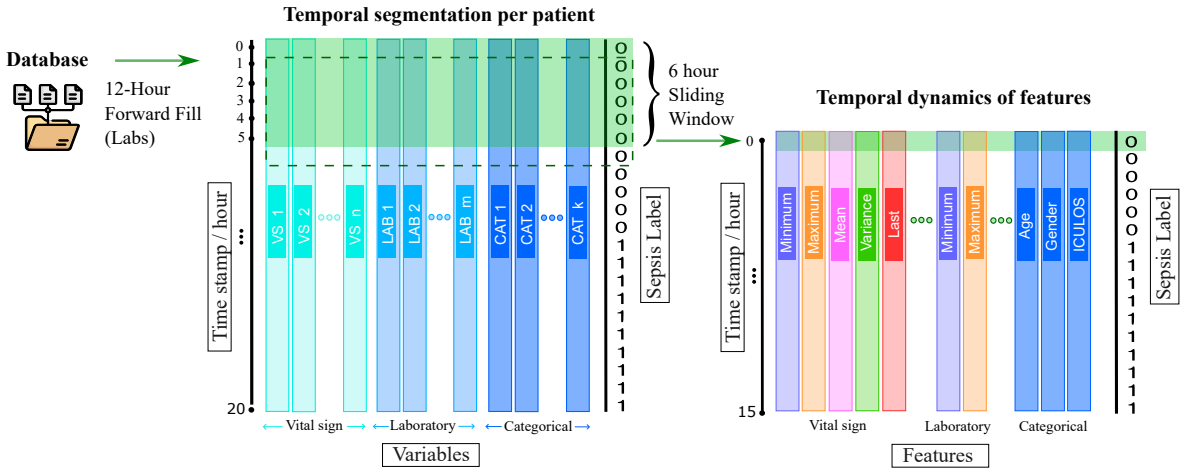

Figure 1. Illustration of the patient data analysis pipeline: for each patient, a 21-hour period is analyzed using a 6-hour sliding window with a 1-hour stride, from which statistical features of clinical variables are extracted, transforming temporal data into a structured feature set for predictive modeling.

Multivariate clinical data were imputed using a LightGBM estimator following the strategy of the DACMI challenge winner[19], with hyperparameters tuned via HalvingRandomSearchCV, a successive halving random search that allocates more resources to promising combinations while discarding underperforming ones. We standardized the data through z-score normalization, transforming each variable to have zero mean and unit standard deviation.

[19] Xiao Xu et al.: *A Multi-directional Approach for Missing Value Estimation in Multivariate Time Series Clinical Data*. En: *Journal of Healthcare Informatics Research* 4.4 (2020), págs. 365-382. DOI: 10.1007/s41666-020-00076-2.

Table 1. Clinical variables included in the study after preprocessing and selection: from the original 40 variables, multiple derived features were generated, of which 43 were retained after removing columns with more than 20 % missing values.

Variable	**Unit**	**Type**
HR / SBP	Dimensionless	Derived
HR	bpm	Vital sign
O2Sat	%	Vital sign
Temp	°C	Vital sign
SBP	mmHg	Vital sign
MAP	mmHg	Vital sign
DBP	mmHg	Vital sign
Resp	breaths/min	Vital sign
BUN	mg/dL	Laboratory
Age	years	Demographic
Gender	Categorical	Demographic
ICULOS	Days	Length of stay
SepsisLabel	Binary	Target

4. RESULTS

The dataset was first split into 85 % training and 15 % testing subsets. Figure 2 presents the Receiver Operating Characteristic (ROC) curves with 95 % confidence intervals for eleven machine learning models evaluated using 5-fold cross-validation with patient-level grouping on the training set. This grouping ensures that all records from the same patient remain within a single fold, preventing information leakage between training and validation subsets. Tree-based boosting models demonstrated superior performance, achieving the highest AUC scores and significantly outperforming simpler classifiers.

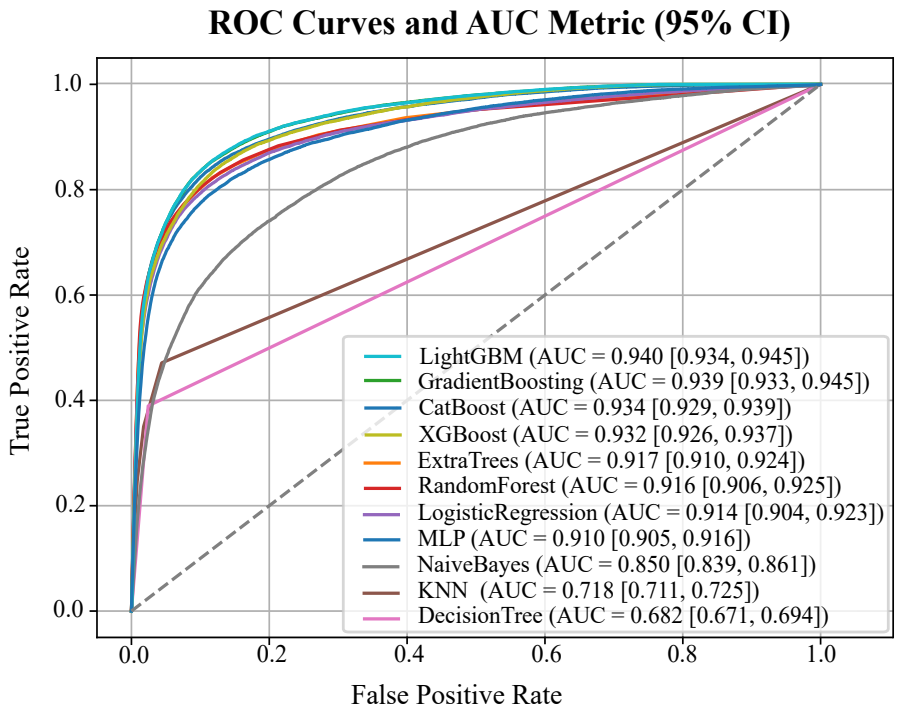

Figure 2. ROC curves of the evaluated machine learning models. Eleven algorithms are compared, showing the True Positive Rate against the False Positive Rate. The area under the curve (AUC) with its 95 % confidence interval is indicated for each model.

As detailed in the appendix .1, the top-performing models —LightGBM, Gradient Boosting, CatBoost, and XGBoost— whose AUC-ROC values fell within the 95 % confidence interval of the best classifier, were selected to proceed with Bayesian optimization of their hyper-parameters, which was carried out by exploring specific ranges for each algorithm. To address class imbalance, we incorporated class weights into model training. Equation (1) presents the standard binary cross-entropy loss modified to account for these weights.

$$L = -\sum_{i=1}^{N} \Big[w_1 \cdot y_i \cdot \log(p_i) \; + \; w_0 \cdot (1 - y_i) \cdot \log(1 - p_i) \Big] \tag{1}$$

where:

- $y_i \in 0, 1$ is the true class label of instance i
- p_i is the predicted probability of sepsis for instance i
- w_1 and w_0 are class weights: $w_0 = 1$ for non-septic patients and $w_1 = 0{,}5 \times \frac{N_{\text{non-septic}}}{N_{\text{septic}}}$ for septic patients, where the factor $0{,}5$ attenuates the class weight ratio to avoid excessive corrections and maintain training stability.
- N is the total number of training samples.

In addition, the decision threshold was dynamically optimized to emphasize clinical criteria and avoid bias, maximizing a composite function of F1-score and sensitivity, as shown in Equation (2). A weighting factor of $\alpha = 0{,}6$ was assigned to sensitivity. We applied this procedure to the selected models, and in the case of CatBoost it yielded an optimal threshold of $\tau^* = 0{,}12$.

$$\tau^* = \arg\max_{\tau \in [0,1]} \Big[\alpha \cdot \text{Sensitivity}(\tau) + (1 - \alpha) \cdot F1(\tau) \Big] \tag{2}$$

We used the the optimized models to predict the test set and evaluated its performance using the metrics detailed in Table 2. These included the clinical utility score, introduced in the PhysioNet Challenge and formalized in Equation (3) is specifically designed to assess the clinical relevance of early sepsis prediction, rewarding correct predictions made within a window close to six hours before onset and penalizing false positives or delayed/missed detections.

$$U_{\text{norm}} = \frac{\sum_{i=1}^{N} U_{\text{obs},i} - \sum_{i=1}^{N} U_{\text{inaction},i}}{\sum_{i=1}^{N} U_{\text{best},i} - \sum_{i=1}^{N} U_{\text{inaction},i}} \tag{3}$$

where $U_{\text{obs},i}$ is the accumulated utility from the model's predictions, $U_{\text{best},i}$ corresponds to the optimal early detection within the predefined window, $U_{\text{inaction},i}$ represents the case of no detection, and N is the total number of patients[20].

Table 2. Performance of the four best-performing gradient boosting models for early sepsis prediction after advanced hyperparameter optimization. Results on the independent test set demonstrate consistently high discriminative capability (all AUC $>$ 0.94)

Metric	**LightGBM**	**GradientBoosting**	**CatBoost**	**XGBoost**
Accuracy	0.9624	0.9654	0.9589	0.9617
F1 Score	0.5538	0.5755	0.5473	0.5548
Precision	0.4688	0.4996	0.4414	0.4638
AUC ROC	0.9476	0.9577	0.9484	0.9472
Sensitivity	0.6764	0.6784	0.7201	0.6903
Specificity	0.9726	0.9757	0.9674	0.9715
Utility Score	0.6347	0.6313	0.6686	0.6440

[20] Reyna et al. 2019.

In Table 2, Gradient Boosting achieved the highest accuracy (0.9654), F1-score (0.5755) and AUC-ROC (0.9577), while LightGBM and XGBoost yielded competitive results but did not outperform the other models in any metric. CatBoost stood out by attaining the highest sensitivity (0.7201) and the best clinical utility score (0.6686), which more adequately reflects the clinical cost of errors. These findings suggest that CatBoost is the most suitable model for early sepsis detection, as it prioritizes minimizing false negatives.

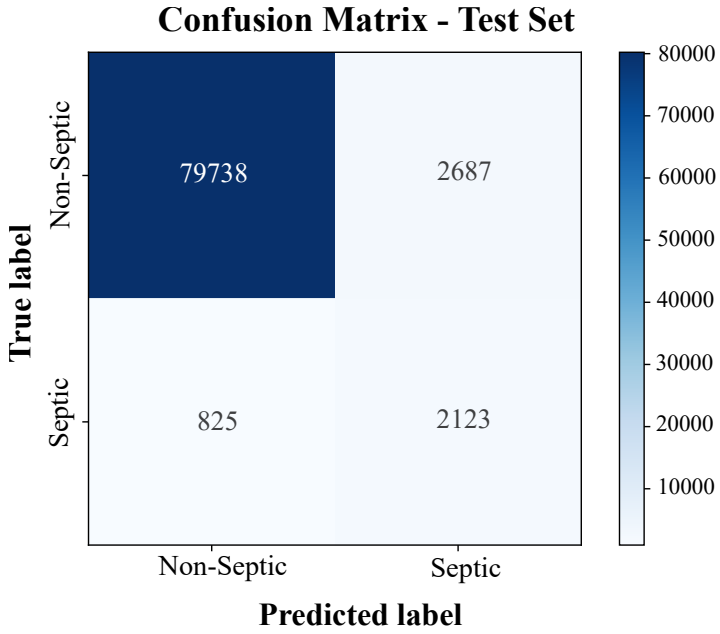

Figure 3. Confusion matrix of the CatBoost model for sepsis prediction. The horizontal axis corresponds to the predicted classes and the vertical axis to the actual classes.

As shown in Figure 3, the confusion matrix on the test set confirmed that CatBoost correctly identified 2,123 septic cases, while missing 825, thereby reducing the risk of undetected diagnoses in critically ill patients.

Finally, Figure 4 presents the SHAP-based feature importance analysis revealed that ICULOS (ICU length of stay) was the most influential predictor. This finding aligns with clinical knowledge, as prolonged ICU stays are often associated with increased complication risk.

Similarly, the importance of BUN levels reflects established renal dysfunction patterns in sepsis progression. Other relevant variables included physiological parameters such as temperature and respiratory rate. These results demonstrate that the model is capturing clinically meaningful patterns, leveraging variables that are routinely considered in critical care settings. The integration of temporal evolution, laboratory markers, and vital signs appears central to improving early sepsis prediction in the ICU.

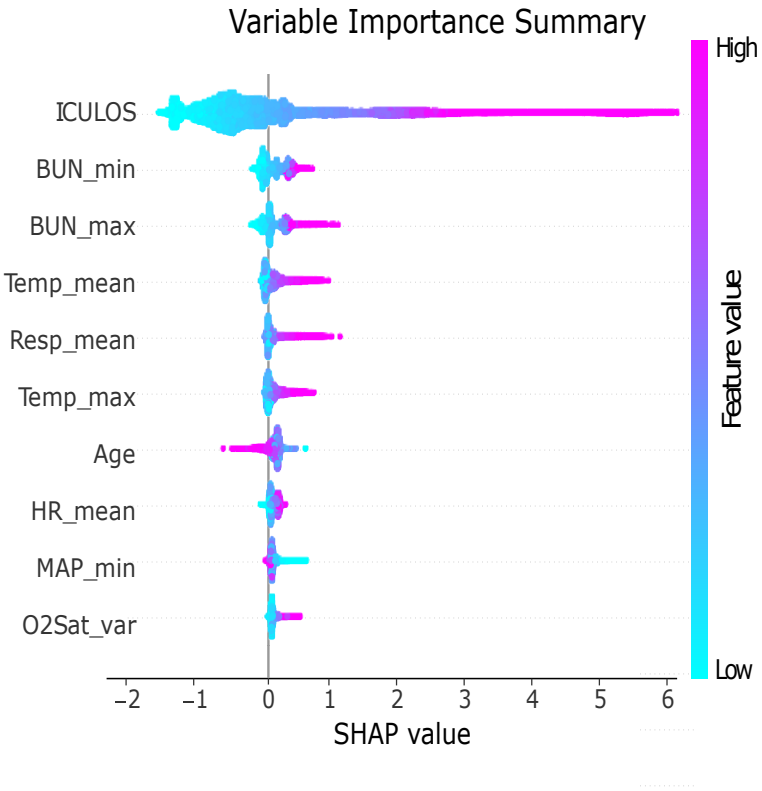

Figure 4. Relative variable importance obtained through SHAP analysis for the CatBoost model. The bars indicate the percentage contribution of each feature to the model's predictive performance.

The appendix .2 provides the implementation code, including data preprocessing, model training, and hyperparameter optimization, along with the corresponding GitHub repository.

5. DISCUSSION AND CONCLUSIONS

This study offers a different approach by combining rigorous variable preprocessing, imputation based on the LightGBM estimator, and explicit optimization geared toward sensitivity and clinical utility. These methodological decisions, which are rarely detailed in many state-of-the-art studies, made it possible to reduce noise, preserve nonlinear relationships, and align the model's objectives with the real needs of the ICU.

Our results show competitive performance: most models achieved AUC values above 0.90 even with initial configurations, suggesting that the quality of preprocessing was decisive. Following optimization, a characteristic trade-off between metrics was observed: while sensitivity showed substantial improvements, other metrics such as precision experienced a reduction. Notably, while our best model CatBoost achieved a utility score 0.6686 below the 0.83 reported by Strickler et al., it demonstrated a substantially higher sensitivity (0.7201 vs. 0.56), which is crucial in the clinical environment. Furthermore, our model's utility score surpassed the results achieved by the top-ranking teams in the PhysioNet Challenge on the public test sets. However, these comparisons should be interpreted with caution, as we did not evaluate on the Challenge's hidden set.

The selection of a classification threshold such as that obtained with equation (2), determined by a sweep that prioritized sensitivity and the F1-score, is consistent with clinical practice: a false alarm is preferable to overlooking a true case. Although the default threshold is 0.5, as it suggests good class separation, its application in our unbalanced dataset would result in clinically unacceptable sensitivity. This decision, however, has an inherent consequence: a low threshold increases the false positive rate. The threshold must be recalibrated for each new clinical setting to preserve its actual utility. The optimal balance between metrics is directly influenced by factors specific to each ICU, such as the prevalence of sepsis and the operational capacity to manage alerts.

The superiority of boosting-based models confirms their ability to handle complex clinical data. However, in contrast to approaches that operate as “black boxes” —providing little explanation for the reasoning behind their predictions— our SHAP analysis yields clinically interpretable insights. It identified variables beyond traditional scores (SOFA, SIRS)—such as ICU length of stay (ICULOS), BUN levels, and respiratory dynamics—as key predictors This alignment with established clinical knowledge, where prolonged ICU stay and renal dysfunction are recognized risk factors for complications, confirms that machine learning can effectively complement conventional clinical criteria by capturing recognizable pathophysiological patterns.

Although the results are promising, significant limitations remain. The main limitation was data quality: class imbalance and a high proportion of missing values forced the exclusion of variables and the implementation of imputation, which, although it preserved the clinical consistency of the dataset, may have introduced biases. Furthermore, the models were trained and evaluated solely on the PhysioNet Challenge database, without external validation in independent cohorts, which limits the evidence for their generalization. The imputation and optimization pipeline also requires significant computational effort, which could make it difficult to replicate in other settings. Finally, the prioritization of sensitivity and utility score favored early detection, but at the expense of lower precision. This was reflected in the confusion matrix, with an increase in false positives that, in a real clinical setting, could translate into alert fatigue. Nevertheless, false negatives, although less frequent, pose a relevant clinical risk by leaving septic patients undetected.

This study demonstrates that a methodology centered on meticulous clinical data management, combined with boosting models optimized for clinically-relevant outcomes, provides a robust framework for early sepsis detection. The effectiveness of our proposal stems from a systematic and reproducible preprocessing pipeline: observation windows with sepsis onset, capturing short-term clinical dynamics using sliding segments, and in-

corporating advanced multivariate imputation. This temporal structuring and explicit optimization ensured that the models prioritized early detection of true positives, critical for ICU patient outcomes.

The impact of this design is reflected in the results: all boosting models achieved AUC-ROC values above 0.93, confirming the discriminative power of the pipeline. CatBoost stood out with a sensitivity of 0.7201 and a clinical utility score of 0.6686, surpassing the performance of top-ranking teams in the PhysioNet Challenge on public test sets. Beyond specific comparisons, our approach stands out from the state of the art as a methodological alternative, designed to prioritize critical clinical criteria in the early prediction of sepsis.

In summary, this work contributes a replicable framework that addresses critical gaps in prior studies—such as simplistic imputation, incomplete patient selection, and narrow evaluation metrics—while providing interpretable clinical insights through SHAP analysis. The proposed pipeline demonstrates strong potential for deployment in ICUs with similar data structures, with future validation required in multi-center cohorts and real-time implementations to confirm its generalizability and clinical impact.

BIBLIOGRAPHY

Alali, Walid et al.: *Impact of laboratory test result availability on clinical decision-making: evidence from routine data*. En: *Scientific Reports* 12.1 (2022), pág. 22191. DOI: 10.1038/s41598-022-25961-1.

Bone, Roger C., Robert A. Balk, Frank B. Cerra et al.: *Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis*. En: *Chest* 101.6 (1992), págs. 1644-1655. DOI: 10.1378/chest.101.6.1644.

Catalyst, Health: *Optimizing Sepsis Care Improves Early Recognition and Outcomes*. Health Catalyst Success Story. 2025.

Du, John Anda, Nadi Sadr y Philip de Chazal: «Automated Prediction of Sepsis Onset Using Gradient Boosted Decision Trees». En: *Computing in Cardiology*. Vol. 46. 2019, págs. 1-4.

Fu, Mengsha et al.: «An Ensemble Machine Learning Model For the Early Detection of Sepsis From Clinical Data». En: *Computing in Cardiology 2019*. Vol. 46. PhysioNet / Computing in Cardiology Challenge. Nanjing University of Aeronautics y Astronautics, China, 2019, págs. 1-4. DOI: 10.22489/CinC.2019.317.

Hao, Pei-Yi: «Early Prediction of Sepsis Utilizing Multi-branches Multi-tasks Hybrid Deep Learning Model». En: *Proceedings of the Third International Conference on Innovations in Computing Research (ICR'24)*. Vol. 1058. Lecture Notes in Networks and Systems. Springer, 2024, págs. 63-74. DOI: 10.1007/978-3-031-65522-7_6.

He, YiRan et al.: *A machine-learning approach for prediction of hospital mortality in cancer-related sepsis*. En: *Clinical eHealth* 6 (2023), págs. 17-23. DOI: 10.1016/j.ceh.2023.06.003.

Hou, Nianzong et al.: *Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost*. En: *Journal of Translational Medicine* 18.462 (2020). DOI: 10.1186/s12967-020-02620-5.

Islam, Khandaker Reajul et al.: *Machine Learning-Based Early Prediction of Sepsis Using Electronic Health Records: A Systematic Review*. En: *Journal of Clinical Medicine* 12.17 (2023), pág. 5658. DOI: 10.3390/jcm12175658.

Jesson, et al.: *3D-MICE: Multiple Imputation with Gaussian Processes for Clinical Time Series*. En: *IEEE ICHI 2019 Proceedings* (2019). Presented at DACMI Challenge.

Johnson, Alistair E.W., Tom J. Pollard, Lu Shen et al.: *MIMIC-III, a freely accessible critical care database*. En: *Scientific Data* 3 (2016), pág. 160035. DOI: 10.1038/sdata.2016.35.

Ladbrook, Elyse et al.: *A systematic review of the cost-impact of sepsis care bundles*. En: *Journal of Hospital Infection* (2025). Available under a Creative Commons license. ISSN: 0195-6701. DOI: 10.1016/j.jhin.2025.08.006.

Lambden, Simon et al.: *The SOFA score — development, utility and challenges of accurate assessment in clinical trials*. En: *Critical Care* 23.1 (2019), pág. 374. DOI: 10.1186/s13054-019-2663-7.

Lambden, Simon et al.: *The SOFA score—development, utility and challenges of accurate assessment in clinical trials*. En: *Critical Care* 23.1 (nov. de 2019), pág. 374. DOI: 10.1186/s13054-019-2663-7.

Liu, Zheng et al.: *Interpretable machine learning for predicting sepsis risk in emergency triage patients*. En: *Scientific Reports* 15.1 (2025), págs. 1-13. DOI: 10.1038/s41598-025-85121-z.

Luo, Yuan y IEEE ICHI Organizers: *Data Analytics Challenge on Missing data Imputation (DACMI)*. https://ewh.ieee.org/conf/ichi/2019/challenge.html. 2019.

Morrill, James et al.: «The Signature-Based Model for Early Detection of Sepsis From Electronic Health Records in the Intensive Care Unit». En: *Computing in Cardiology*. Vol. 46. 2019, págs. 1-4. DOI: 10.22489/CinC.2019.014.

Morrill, James et al.: «Utilization of the Signature Method to Identify the Early Onset of Sepsis From Multivariate Physiological Time Series in Critical Care Monitoring». En: *Critical Care Medicine*. Vol. 48. 1. 2020, pág. 441. DOI: 10.1097/CCM.0000000000004510.

Oliveros, Henry et al.: *One-year survival of patients admitted for sepsis to intensive care units in Colombia*. En: *BMC Infectious Diseases* 24.1 (jul. de 2024), pág. 678. ISSN: 1471-2334. DOI: 10.1186/s12879-024-09584-7.

Rady, Mohamed Y. et al.: *Shock index as a marker of severity in patients with sepsis*. En: *Medicina Intensiva* 40.7 (2016), págs. 399-404. DOI: 10.1016/j.medin.2016.03.005.

Reyna, Matthew A., Clifford Josef, Seyed et al.: «Early Prediction of Sepsis from Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019». En: *Computing in Cardiology Conference (CinC)*. 2019, págs. 1-4. DOI: 10.1097/CCM.0000000000004145.

Reyna, Matthew A. et al.: *A large, annotated dataset of sepsis patients from the PhysioNet/Computing in Cardiology Challenge 2019*. En: *PhysioNet* (2019). Available at https://physionet.org/challenge/2019/.

Singer, Mervyn, Clifford S. Deutschman, Christopher W. Seymour et al.: *The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)*. En: *JAMA* 315.8 (2016), págs. 801-810. DOI: 10.1001/jama.2016.0287.

Singer, Mervyn et al.: *The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)*. En: *JAMA* 315.8 (2016), págs. 801-810. DOI: 10.1001/jama.2016.0287.

Strickler, Ethan A. T. et al.: *Exploring a global interpretation mechanism for deep learning networks when predicting sepsis*. En: *Scientific Reports* 13.3067 (2023). DOI: 10.1038/s41598-023-30091-3.

Su, Yingjie et al.: *Early predicting 30-day mortality in sepsis in MIMIC-III by an artificial neural networks model*. En: *European Journal of Medical Research* 27.294 (2022). DOI: 10.1186/s40001-022-00925-3.

Sun, P.: «MICE-DA: A MICE method with Data Augmentation for missing data imputation». En: *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. 2019. DOI: 10.1109/ichi.2019.8904724.

World Health Organization: *Sepsis*. https://www.who.int/news-room/fact-sheets/detail/sepsis. 2020.

Xu, Xiao et al.: *A Multi-directional Approach for Missing Value Estimation in Multivariate Time Series Clinical Data*. En: *Journal of Healthcare Informatics Research* 4.4 (2020), págs. 365-382. DOI: 10.1007/s41666-020-00076-2.

Yang, Meicheng et al.: «SSP: Early Prediction of Sepsis Using Fully Connected LSTM-CNN Model». En: *Computing in Cardiology*. Vol. 46. 2019, págs. 1-4. https://www.cinc.org/archives/2019/pdf/CinC2019-280.pdf.

Zabihi, M., S. Kiranyaz y M. Gabbouj: «Sepsis Prediction in Intensive Care Unit Using Ensemble of XGBoost Models». En: *2019 Computing in Cardiology (CinC)*. Singapore, 2019, Page 1-Page 4. DOI: 10.22489/CinC.2019.238.

Zhang, et al.: *Evaluating the state of the art in missing data imputation for clinical data*. En: *Briefings in Bioinformatics* 23.1 (2022), bbab489. DOI: tpi10.1093/bib/bbab489.

*

APPENDICES

Anexo A. Hyperparameter Configuration

Table 3 presents the hyperparameter ranges explored for each algorithm during Bayesian optimization. These intervals were defined based on preliminary experiments and recommendations from related literature.

Table 3. Hyperparameter ranges explored for each algorithm during Bayesian optimization.

Algorithm	**Parameter**	**Range**
LightGBM	n_estimators	100–2000
	depth	20–150
Gradient Boosting	n_estimators	100–2000
	depth	3–15
XGBoost	n_estimators	100–2000
	depth	3–12
CatBoost	n_estimators	800–2000
	depth	6–12

Anexo B. GitHub Repository

The complete implementation code, including preprocessing scripts and training routines, is publicly available at:

https://github.com/Lilith022/Early-Detection-of-Sepsis-from-Clinical-Data-using-a-Machine-Learning-Model

This repository ensures reproducibility and facilitates further research in early sepsis detection.