Early Prediction of Sepsis and ICU Length of Stay using MIMIC-IV Data: A Critical Analysis

Nome Cognome
Department of Computer Science and Engineering
University of Bologna
Student Id number: 0001117305
Email: giuseppe.budano@studio.unibo.it

Abstract—**This project explores the application of Machine Learning techniques to predict the risk of sepsis and the Length of Stay (LOS) in Intensive Care Units (ICU) using data from the MIMIC-IV database. I specifically focused on the first 6 hours of admission to develop an early warning system. Both manual feature engineering with Random Forest and Automated Machine Learning (AutoML) approaches were evaluated. While the models achieved high accuracy ($> 83\%$), my results highlight critical challenges related to class imbalance and clinical definitions. Specifically, the analysis reveals that using discharge ICD codes as ground truth introduces temporal ambiguity, suggesting that future implementations must adopt the dynamic Sepsis-3 criteria to ensure clinical validity.**

1. Introduction

Sepsis is a life-threatening organ dysfunction caused by a dysregulated host response to infection, representing one of the most significant challenges in modern medicine. Accounting for approximately 20% of all global deaths, it remains a leading cause of mortality in Intensive Care Units (ICUs) where the clinical trajectory can deteriorate from stable to critical shock within hours. Time is the most critical factor in sepsis management; clinical evidence suggests that for every hour of delay in administering appropriate antibiotics and fluid resuscitation, patient survival rates decrease by approximately 8%. Consequently, there is an urgent clinical demand for automated "Early Warning Systems" (EWS) capable of stratifying patient risk immediately upon hospital admission, overcoming the trade-off between sensitivity and specificity often found in traditional manual scoring systems like SIRS or qSOFA.

The primary objective of this project is to leverage Machine Learning to address this gap by analyzing the complex, non-linear interactions between physiological signals. I aimed to answer two fundamental questions within the first few hours of a patient's ICU stay: firstly, whether a patient is at high risk of developing sepsis (Risk Stratification), and secondly, to estimate the expected Length of Stay (Resource Planning) to optimize bed management.

A distinct feature of my study is its strict temporal constraint; unlike retrospective studies that utilize the entire hospitalization history (which includes data not available at decision time), I restricted the observation window strictly to the first **6 hours** from admission. This design choice simulates a real-time deployment scenario where the model must operate with partial, noisy, and initially sparse data, mirroring the uncertainty faced by clinicians during the triage phase and prioritizing operational realism over theoretical maximum accuracy.

This project was developed individually by the author, encompassing the entire research pipeline from data extraction to critical evaluation. My workflow began with the engineering of a large-scale cohort from the MIMIC-IV database, requiring the complex merging of demographic, administrative, and clinical event tables. Subsequently, I implemented a robust preprocessing pipeline to handle time-series windowing, statistical aggregation of vital signs, and median imputation for missing physiological values. The modeling phase involved the development of baseline Random Forest models compared against an Automated Machine Learning (AutoML) approach using FLAML to efficiently explore hyperparameter spaces. Finally, I subjected the results to a rigorous evaluation focusing not only on numerical metrics but on clinical validity, identifying key methodological limitations such as the potential data leakage inherent in using discharge ICD codes for early prediction.

2. Related Work

The prediction of sepsis is one of the most active research areas in medical informatics. Existing approaches can be broadly categorized into clinical rule-based scores and data-driven machine learning models. Traditionally, clinicians rely on scoring systems such as SIRS, SOFA, and qSOFA. While specific, SOFA requires laboratory results that may not be immediately available (e.g., Bilirubin, Platelets). Conversely, qSOFA is faster but has shown low sensitivity in recent validation studies, often missing early warning signs. This limitation motivates the need for data-driven models capable of leveraging non-linear relationships between vital signs.

The release of the MIMIC-III and subsequently MIMIC-IV databases has spurred significant research in this domain. Desautels et al. developed the "Insight" algorithm utilizing minimal vital sign data, demonstrating that simple physiological signals like heart rate and mean arterial pressure are predictive of sepsis onset. More recently, Moor et al. applied Temporal Convolutional Networks (TCNs) to capture long-term dependencies in clinical time series, effectively outperforming static baselines.

Recent literature strongly supports the trend towards ultra-early prediction, validating the restricted time-window approach adopted in this project. For instance, **Wang et al. [3]** demonstrated that high-performance AI models can be constructed using clinical information from as early as the first hour of admission, suggesting that immediate stratification is feasible even with limited data. Furthermore, the effectiveness of Ensemble Learning strategies, which form the basis of my Random Forest and AutoML approach, has been corroborated by **Akinduyite et al. [4]**. Their work highlights how combining multiple weak learners often outperforms single sophisticated models in sepsis prediction, supporting my choice of Random Forest and FLAML as robust baselines.

3. Proposed Method

To ensure a rigorous and structured approach to the data science lifecycle, I adopted the **CRISP-DM** (Cross-Industry Standard Process for Data Mining) methodology. This framework guided my project through the phases of Business Understanding, Data Understanding, Data Preparation, Modeling, and Evaluation.

3.1. Business Understanding & Data Understanding

The primary "business" goal was identified as the early detection of sepsis to reduce mortality and the estimation of ICU stay to improve resource allocation. To achieve this, I utilized the **MIMIC-IV (v2.2)** database, a comprehensive repository of de-identified health-related data associated with patients admitted to the Beth Israel Deaconess Medical Center. To ensure a clinically homogeneous cohort and avoid confounding factors, I applied strict inclusion criteria: adult patients (age ≥ 18) admitted to the Intensive Care Unit for their **first stay**.

Readmissions were excluded as they often represent a biased population with different physiological baselines. Furthermore, I removed admissions with a Length of Stay (LOS) shorter than six hours, as they would not provide a complete observation window. This filtering process resulted in a final cohort of approximately 85,000 unique ICU stays. The ground truth for the target variables was derived as follows: sepsis status was identified by mapping ICD-9 and ICD-10 codes from the `diagnoses_icd` table to explicit sepsis codes, while the Length of Stay was computed as the precise fractional difference between discharge and admission time.

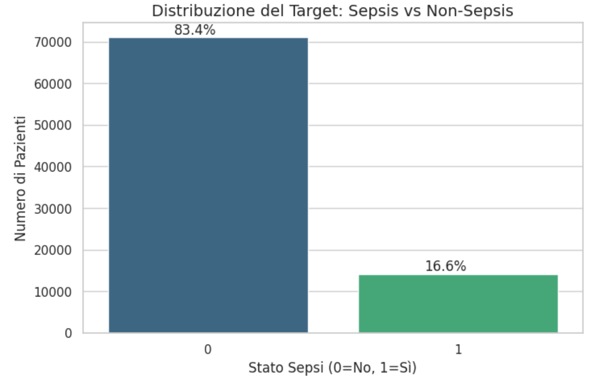

Figure 1. **Class Imbalance Analysis.** The dataset shows a prevalence of sepsis of approximately 16.5%. This imbalance poses a significant challenge for the classification task, biasing the model towards the majority class.

3.2. Data Preparation

The transformation of raw, asynchronous clinical logs into a structured format represented the core engineering challenge of this work. My pipeline was designed to strictly prevent *temporal leakage*, ensuring the model only has access to information available at the time of prediction.

Feature Selection and Extraction. Based on clinical literature (SIRS and SOFA criteria), I specifically selected a subset of physiological signals highly indicative of systemic infection. These include: **Heart Rate**, **Systolic and Diastolic Blood Pressure**, **Respiratory Rate**, **Body Temperature**, and **Oxygen Saturation** (SpO_2). Additionally, critical laboratory markers such as **Lactate** and **White Blood Cell count** were included when available within the first 6 hours.

Temporal Windowing and Statistical Aggregation. The raw data in MIMIC-IV consists of timestamped events. I applied a strict filtering mask to retain only measurements recorded within the interval $[t_{in}, t_{in} + 6h]$. Since the frequency of measurements varies by patient, I flattened these time-series into a fixed-size feature vector using statistical aggregation. For each vital sign and laboratory result, I computed the Mean, Median, Standard Deviation, Minimum, and Maximum. This approach allows the model to capture not just the "average" state of the patient, but also the stability (via std dev) and the presence of acute extremes (min/max), which are often the earliest signs of septic shock.

Outlier Management and Data Quality. A critical aspect of clinical data preparation involves the management of outliers. Physiological signals in ICU settings are prone to sensor artifacts. I adopted a conservative approach to outlier removal: in the context of sepsis prediction, extreme values (e.g., severe hypotension) are often the signal of interest rather than noise. Therefore, I retained statistical outliers that were biologically plausible, removing only obvious artifacts. This decision was further supported by the choice of Tree-based models, which are inherently robust to outliers.

Comorbidity Extraction. Recognizing that patient history significantly influences outcomes, I extracted preexisting conditions using the Charlson Comorbidity Index definitions based on discharge codes (ICD-9/10). Binary flags were created for major chronic conditions such as Diabetes, Hypertension, and Cancer. I acknowledge a limitation here: in this retrospective analysis, this constitutes a form of data leakage. In a prospective deployment, this module would need to query historical electronic health records (EHR) instead.

Handling Missing Data. Clinical data is inherently sparse. I addressed missing values using **Median Imputation**. I selected the median over the mean or complex interpolation because clinical variables often follow skewed distributions. Furthermore, in an emergency setting, the "missingness" of a value often carries information itself (indicating the clinician did not deem the test necessary).

3.3. Modeling

I formulated the problem as two distinct supervised learning tasks: a binary classification task to predict sepsis risk and a regression task to estimate the number of days in ICU. To ensure robustness, the dataset was split into training (80%) and testing (20%) sets.

Random Forest (Baseline). I selected Random Forest as the baseline model due to its ensemble nature, which makes it robust to overfitting and capable of handling non-linear interactions. The model was trained with manually selected hyperparameters, using **Stratified Cross-Validation** to maintain the class distribution across folds.

AutoML (FLAML). To challenge the baseline, I employed the **FLAML** (Fast and Lightweight AutoML) library. FLAML utilizes a novel *Cost-Frugal Search* algorithm to efficiently explore the hyperparameter space of gradient boosting learners (XGBoost, LightGBM, CatBoost). This method optimizes the trade-off between model performance and computational cost. For the classification task, the optimization metric was set to **ROC-AUC** to maximize discriminative power, while for the regression task, **Mean Absolute Error (MAE)** was minimized.

4. Results

4.1. Task A: Sepsis Prediction Analysis

The dataset, consisting of 85,222 unique ICU stays, was partitioned into a training set (80%) and a hold-out testing set (20%). To ensure statistical validity, I applied Stratified Sampling for the classification task. The quantitative performance is summarized in Table 1.

TABLE 1. MODEL PERFORMANCE METRICS BY TASK

Task & Model	Recall	ROC-AUC	MAE (Days)
Task A: Classification (Sepsis)			
Random Forest	0.68	0.82	*N/A*
AutoML (Best)	0.48	0.84	*N/A*
Task B: Regression (LOS)			
Random Forest	*N/A*	*N/A*	2.87
AutoML (Best)	*N/A*	*N/A*	**2.33**

For the classification task, both models achieved an Accuracy exceeding 83%. However, this metric proves misleading in this context due to the class imbalance; a naive classifier predicting zero positives would achieve similar accuracy but zero utility. The critical metric, Recall (Sensitivity), reveals the limitations of the 6-hour observation window. My manual Random Forest achieved a Recall of 0.68, meaning it successfully identified 68% of septic patients but missed the remaining 32% (False Negatives). This indicates that two-thirds of the septic population exhibit clear physiological distress in the first hours, while one-third presents with "silent" symptoms that aggregated statistics fail to capture.

Interestingly, while AutoML slightly improved the overall discriminative power (ROC-AUC 0.84 vs 0.82), it prioritized specificity over sensitivity, dropping Recall to 0.48. This phenomenon highlights a common pitfall in automated optimization: unless explicitly constrained to maximize Recall, the algorithm will optimize for the global AUC, effectively "playing it safe" to reduce False Positives. In a clinical context, this behavior is suboptimal, as missing a diagnosis is far more costly than a false alarm.

The Confusion Matrix presented in **Fig. 2** visually confirms this conservative bias. Feature Importance analysis (SHAP) indicated that Lactate, Minimum Blood Pressure, and Max Heart Rate were the top drivers of prediction, validating that the model is leveraging correct physiological signals despite the recall limitations.

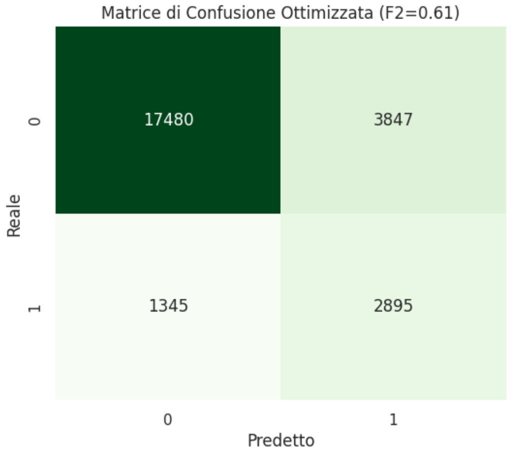

Figure 2. **Confusion Matrix (Random Forest).** The model tends to be conservative, resulting in a significant number of False Negatives (bottom-left quadrant), favoring the majority class.

4.2. Task B: Length of Stay Prediction Analysis

The regression task yielded more distinct insights into the benefits of automated tuning. The baseline Random Forest predicted the Length of Stay with a Mean Absolute Error

(MAE) of 2.87 days. The AutoML engine, after exploring gradient boosting architectures, reduced this error to **2.33 days**.

As shown in **Fig. 3**, this improvement of approximately 0.5 days per patient is operationally significant for hospital resource planning. It suggests that while manual heuristics can capture linear relationships, the AutoML algorithms successfully modeled the complex, non-linear dependencies between admission physiology and long-term stay duration. However, a residual error of over 2 days persists. This "irreducible error" likely stems from the fact that LOS is not solely determined by patient physiology at admission; it is heavily confounded by logistic factors (e.g., bed availability in wards, discharge delays) and late-stage complications (e.g., hospital-acquired infections) which occur days after the prediction point and are therefore invisible to any model relying solely on admission data.

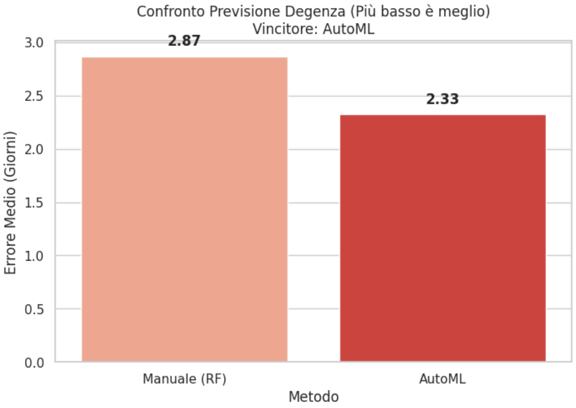

Figure 3. **Length of Stay Prediction Error.** AutoML reduced the prediction error by approx. 0.5 days compared to the manual baseline, demonstrating superior optimization.

5. Conclusions

This project demonstrated the feasibility of processing raw MIMIC-IV data to create a tabular dataset for admission screening, establishing robust accuracy baselines using both manual and automated machine learning techniques. However, the critical analysis highlights significant methodological areas that differentiate this academic exercise from a deployable clinical tool.

The primary limitation identified lies in the use of ICD codes as the ground truth. Since ICD codes represent a billing summary at discharge, they do not capture the onset time of sepsis. A patient developing sepsis on day 5 is labeled "positive" in the dataset, confusing a model that only observes data from day 0. This temporal mismatch is the likely cause of the limited Recall observed. Furthermore, relying on discharge codes for comorbidities introduces Data Leakage, as this information is not known at admission time in a real-world scenario.

Beyond the technical metrics, this study underscores the inherent complexity of translating AI into intensive care. The results suggest that while automated machine learning (AutoML) can optimize numerical predictions—as evidenced by the superior performance in the Length of Stay regression task—it cannot automatically correct for fundamental deficits in data definition, such as the ambiguity of sepsis onset. Consequently, the role of such models should be envisioned not as autonomous diagnostic agents, but as "Triage Support Systems" that flag high-risk physiological patterns for human review.

To bridge the gap between this prototype and clinical utility, future work must prioritize the implementation of dynamic **Sepsis-3 criteria**. This would involve defining sepsis based on the suspicion of infection and a real-time change in SOFA score, allowing for the labeling of the exact hour of onset. Additionally, moving from statistical aggregation to **Time-Series Modeling** using Recurrent Neural Networks (LSTM or GRU) would allow the model to learn from the trajectory of vital signs rather than just static values, potentially capturing subtle trends of deterioration missed by the current approach. Ultimately, the integration of these advanced temporal definitions represents the necessary step to transform a retrospective analysis into a prospective clinical tool.

References

[1] A. L. Goldberger et al., “PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals,” *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[2] M. Singer et al., “The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3),” *JAMA*, vol. 315, no. 8, pp. 801–810, 2016.

[3] H. Wang et al., “Early Sepsis Prediction Using Publicly Available Data: High-Performance AI/ML Models with First-Hour Clinical Information,” *Diagnostics*, vol. 15, no. 2727, 2025.

[4] O. C. Akinduyite et al., “Early Prediction of Sepsis Using Ensembled Learning,” in *Dept. of Mathematical and Physical Sciences, Afe Babalola University*, Nigeria.