Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Explainable time-series machine learning for early sepsis prediction and risk stratification: a multi-dataset computational study

Agostino Aiezzo

agostino.aiezzo@youbiquo.eu

Youbiquo S.r.l.

Martina Falasca

University of Salerno

Silvia Spoto

Fondazione Policlinico Universitario Campus Bio-Medico

Alessandro Loppini

Università Campus Bio-Medico

Simonetta Filippi

Università Campus Bio-Medico

Silvia Angeletti

Università Campus Bio-Medico

Maria Vittoria Ristori

Fondazione Policlinico Universitario Campus Bio-Medico

Gianmarco Cafaro

University of Salerno

Antonio Portuesi

DEC Digital Equipment Communication S.r.l.

Mario Garofano

DEC Digital Equipment Communication S.r.l.

Adolfo Santoro

Youbiquo S.r.l.

Silvia Scarpetta

University of Salerno

Letizia Chiodo

Università Campus Bio-Medico

Research Article

Keywords: Sepsis, Intensive care, Early prediction, Deep learning, Long short-term memory, Time-series clustering, Explainable machine learning, Feature importance, Clinical decision support

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11164498/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Explainable time-series machine learning for early sepsis prediction and risk stratification: a multi-dataset computational study

Agostino Aiezzo[1*], Martina Falasca[2], Silvia Spoto[3], Alessandro Loppini[4], Simonetta Filippi[5], Silvia Angeletti[4,6], Maria Vittoria Ristori[6], Gianmarco Cafaro[2], Antonio Portuesi[7], Mario Garofano[7], Adolfo Santoro[1], Silvia Scarpetta[2,8], Letizia Chiodo[5*]

[1]Youbiquo S.r.l., Cava de' Tirreni, Salerno, Italy.
[2]Department of Physics "E.R. Caianiello", University of Salerno, Fisciano, Salerno, Italy.
[3]Department of Diagnostic and Therapeutic Medicine, Fondazione Policlinico Universitario Campus Bio-Medico, Rome, Italy.
[4]Department of Medicine, Surgery and Dentistry, Università Campus Bio-Medico di Roma, Rome, Italy.
[5]Department of Engineering, Università Campus Bio-Medico di Roma, Rome, Italy.
[6]Operative Research Unit of Laboratory, Fondazione Policlinico Universitario Campus Bio-Medico, Rome, Italy.
[7]DEC Digital Equipment Communication S.r.l., Milan, Italy.
[8]INFN, Sezione di Napoli, group collaborativo Salerno, Fisciano, Salerno, Italy.

*Corresponding author(s). E-mail(s): agostino.aiezzo@youbiquo.eu; l.chiodo@unicampus.it;

Abstract

Background: Sepsis is a critical, time-dependent disease, and its early detection is essential to improve patient outcomes. In this work, we propose an integrated computational framework that combines phenotypic data exploration, temporal machine-learning modeling, and dynamic feature relevance to provide a comprehensive toolkit for early warning and risk stratification. The system was evaluated on two complementary datasets: SepsisExp, which includes expert-validated severity labels, and MIMIC-III, a widely used benchmark database for intensive care research.

Methods: The framework comprises a phenotypic exploration phase, in which outlier detection and correlation analysis are integrated with time-series k-means (ts-KMeans) clustering to map the heterogeneous clinical trajectories of the disease. This foundational analysis informed the development of Long Short-Term Memory (LSTM) models optimized to capture subtle temporal dependencies hours before sepsis onset. To bridge the gap between high-performance computing and bedside utility, model predictions were interpreted through permutation-based feature importance (PermFit), which reveals a dynamic shift in biomarker relevance as clinical events approach.

Results: The proposed methodology establishes a robust framework for early sepsis prediction, achieving an in-depth characterization of the phenotypes characteristic of sepsis and stable predictive performance across different time horizons. In addition, the feature-importance analysis reveals a consistent temporal shift in key biomarkers that reflects clinically plausible disease progression.

Conclusions: By transforming complex clinical data into interpretable insights, the proposed approach provides a transparent decision-support tool that can support timely clinical evaluation in time-critical settings.

Keywords: Sepsis, Intensive care, Early prediction, Deep learning, Long short-term memory, Time-series clustering, Explainable machine learning, Feature importance, Clinical decision support

List of abbreviations

AUC-ROC: Area Under the Receiver Operating Characteristic curve; BUN: Blood Urea Nitrogen; CC BY: Creative Commons Attribution; CRP: C-Reactive Protein; CUP: Unique Project Code (Codice Unico di Progetto); FiO_2: Fraction of Inspired Oxygen; HIS: Hospital Information System; ICC: Integrated Critical Care (Philips IntelliSpace Critical Care and Anesthesia); ICU: Intensive Care Unit; INR: International Normalized Ratio; IQR: Interquartile Range; LSTM: Long Short-Term Memory; MCC: Matthews Correlation Coefficient; NRRP: National Recovery and Resilience

Plan; PFI: Permutation Feature Importance; PT: Prothrombin Time; qSOFA: quick Sequential Organ Failure Assessment; SMOTE: Synthetic Minority Over-sampling Technique; SOFA: Sequential Organ Failure Assessment; SVRI: Systemic Vascular Resistance Index; TCN: Temporal Convolutional Network; ts-KMeans: Time-Series k-Means.

1 Background

Sepsis is a life-threatening disease caused by a dysregulated immune response to infection [1], resulting in approximately 49 million cases and 13 million deaths annually [2–4]. Survivors often face long-term deficits and elevated 90-day readmission risk [5], imposing a substantial burden on healthcare systems. Consequently, early diagnosis is essential to improve outcomes. However, recognizing sepsis at an initial stage is complex: there are no specific blood tests, symptoms are highly variable, and physicians' assessments rely on the monitoring of vital signs and laboratory tests, which is time-consuming and potentially biased. Nevertheless, the growing availability of clinical data has enabled machine learning approaches to predict disease onset with useful accuracy.

Data-driven approaches have been increasingly applied to personalize sepsis risks, diagnoses, treatments, and prognoses [6]. Recent literature demonstrates that ensemble learning can improve performance and protect privacy against membership inference attacks, while compensating for the phenotypic heterogeneity of sepsis through an implicit regularization mechanism that reduces variance and improves robustness [7]. Other hybrid early predictors handle missing or asynchronous data using Gaussian processes or Dynamic Time Warping, making them flexible to the varying temporal "rhythms" of the disease [8]. Machine learning has also been applied to identify, post hoc, the most informative vital and metabolic signs, an approach that requires large datasets, careful feature selection, and a meaningful patient cohort [9]. In addition,

natural language processing techniques can interpret subjective clinical observations (e.g., "disoriented patient") and integrate them with objective data, predicting sepsis significantly earlier than standard clinical criteria [10]. Although these approaches are promising, challenges remain in explainability, clinical applicability, and the effective exploitation of temporal information [11].

Motivated by these considerations, we propose a unified computational framework for the early prediction of sepsis structured around three pillars: (i) phenotypic data exploration to characterize patient heterogeneity; (ii) temporal machine-learning modeling for early warning; and (iii) dynamic feature importance analysis to ensure clinical interpretability. Methodologically, the phenotypic phase integrates outlier detection and correlation analysis with ts-KMeans clustering to map diverse clinical trajectories. This understanding guides the development of LSTM models optimized to capture subtle temporal dependencies hours before the onset of sepsis. Finally, model predictions are interpreted through permutation-based feature importance (PermFit) [12, 13], which reveals a dynamic change in biomarker relevance as the clinical event approaches. This information not only corroborates model predictions but also provides clinically meaningful insight into the temporal progression of the disease, thereby enhancing interpretability and supporting clinical decision-making. The main contributions of this work are:

- We propose a framework that integrates phenotypic analysis, temporal deep learning, and dynamic feature importance analysis to model the progression of sepsis; the pipeline is readily applicable to other longitudinal clinical datasets.
- We demonstrate that time-series clustering identifies clinically significant patient phenotypes, enabling risk stratification.
- We show that the models maintain high predictive stability across multiple temporal horizons, providing clinicians with a reliable window for early intervention.

- We present a post-hoc study of the most relevant biomarkers and of how their relevance varies across predictive horizons, which helps verify the reasoning of "black-box" models and supports clinicians' understanding of disease progression.

The remainder of this paper is organized as follows. Section 2 describes the two datasets and the core pillars of the computational framework. Section 3 presents the experimental findings of each phase of the pipeline. Section 4 synthesizes the main findings and their clinical implications, including the limitations of the study. Section 5 outlines future research directions for the early prediction of sepsis and the optimization of healthcare resources.

2 Methods

The aim of this work is to develop and externally evaluate a computational framework for the early prediction of sepsis that is both accurate and clinically interpretable. The study design is retrospective and computational: no new patient recruitment was performed, and all analyses were conducted on secondary data. The setting comprises two publicly available intensive care unit (ICU) cohorts from different institutions and healthcare systems (see Section 2.1). This section describes the public datasets used to evaluate the framework and the definition of the patient cohorts, the techniques adopted for data exploration, the architecture of the predictive model, the evaluation metrics, and the procedure used to estimate the importance of clinical features.

2.1 Datasets: MIMIC-III and SepsisExp

This work is based on two publicly available datasets. The first is MIMIC-III (v1.4) [14, 15], the de facto standard database for intensive care research. It contains more than 58,000 hospital admissions of more than 45,000 anonymized patients treated in critical care units, and incorporates diverse clinical features such as physiological waveforms and laboratory measurements.

To assess generalization, we complement MIMIC-III with a second dataset, SepsisExp [7], which contains the temporal trends of 42 clinical parameters together with patient ages and discharge dates. Its distinguishing feature is the assignment of sepsis severity levels at each time stamp, defined through clinical surveys conducted by experts on a five-level scale; in this way, the granularity of the disease is not limited to a simple binary division. The overall size and class imbalance of the two cohorts are summarized in Fig. 1.

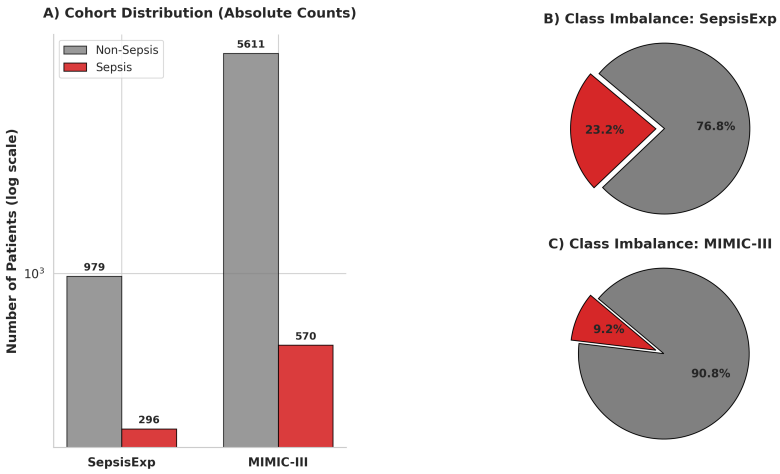


Fig. 1 Dataset distribution and class imbalance analysis. (A) Absolute counts of sepsis and non-sepsis patients in the SepsisExp and MIMIC-III cohorts (logarithmic scale). (B, C) Visual representation of the prevalence of sepsis cases, highlighting the inherent class imbalance in both datasets.

2.1.1 Cohort definition

In the context of early sepsis prediction, the data must be structured so as to be as informative as possible for the predictive model while reflecting real-world use cases; it is therefore essential to identify the onset of the disease. For MIMIC-III (v1.4), we followed the Sepsis-3 criteria [16]: after a suspected infection is identified, the moment at which organ failure manifests is determined by computing the SOFA score [8, 9, 17] (Fig. 2a). The input time window for the model is then constructed: given a

chosen prediction margin, historical data preceding the onset of the disease, plus the prediction margin, are provided to the model. The choice of prediction margin and of the amount of input data determines the predictive capacity of the model.

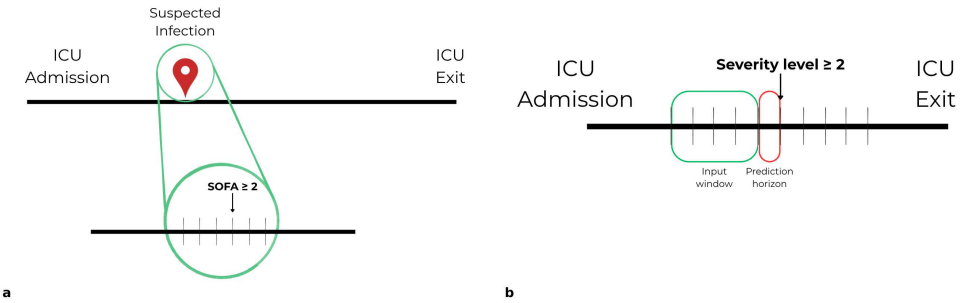


Fig. 2 Definition of the cohorts for MIMIC-III and SepsisExp. (a) Once a suspected infection has been identified, the SOFA score is computed hourly in its proximity. (b) Using the severity levels, the onset is defined in order to identify the prediction horizon and, consequently, the input window.

The cohort for SepsisExp was structured analogously (Fig. 2b). Because information on possible infections and the parameters required to compute the SOFA score are unavailable in this dataset, we proceeded as follows: given the severity level at each time stamp and the binary labeling provided by the original authors, the onset is defined once the severity level exceeds a given threshold. The input window for the predictive model can then be circumscribed as described above.

2.2 Data Analytics: Objectives and Variable Understanding

Once the datasets are identified, it is important to assess their quality and to investigate preliminary information that may be useful from both a clinical and a predictive standpoint. The proposed analytical pipeline consists of three essential stages:

- *Correlation study.* The objective is to measure statistical interdependence and to eliminate redundant correlations on the basis of medical feedback. To this end, we

adopted the Spearman correlation (1):

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{1}$$

where $d_i = \text{rank}(x_i) - \text{rank}(y_i)$ represents the difference between the ranks of each observation and n is the total number of samples.

- *Outlier detection.* The objective is to mitigate anomalies and ensure the integrity of the datasets. For this purpose, we used the Interquartile Range (IQR) (2):

$$\text{IQR} = Q_3 - Q_1 \tag{2}$$

As a non-parametric measure, the IQR is particularly robust against skewed distributions and the distortions caused by outliers that are typical of physiological parameters in intensive care.

- *Clustering analysis.* The objective is to facilitate targeted analysis and classification by identifying phenotypes within the longitudinal data. Accordingly, the most suitable technique is Time-Series k-Means (ts-KMeans) with Euclidean distance (3):

$$J = \sum_{k=1}^{K} \sum_{X_i \in C_k} \|X_i - \mu_k\|^2 \tag{3}$$

where K is the number of clusters, C_k is the k-th cluster, and μ_k is the centroid of the cluster. The aim is not only to discover the characteristic patterns contained in the dataset, but also to perform an initial positioning of patients within the clusters.

2.3 Predictive Modeling: Architecture and Design

The choice of the LSTM architecture for the predictive model was a direct consequence of the analytical requirements that emerged from the exploration phase. The LSTM

is the reference architecture for this class of problems because it can capture the evolution of physiological variables, a crucial aspect of early prediction. The adopted model architecture is illustrated in Fig. 3.

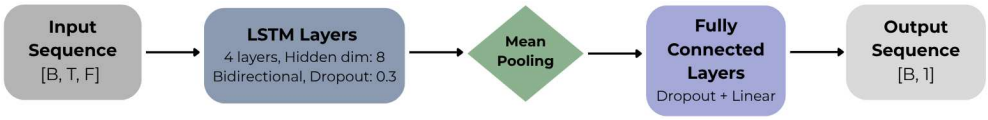


Fig. 3 Lightweight LSTM model architecture for sepsis forecasting.

2.3.1 Technical and training details

Two distinct predictive models were trained, one for each dataset. The models share the same parameters and mechanisms because they address problems common to both clinical cohorts. Dropout and early stopping are employed to mitigate data scarcity and to make the predictive capability of the models as generalizable as possible. In addition, a limited number of hidden layers and hidden states is used to avoid unnecessary model complexity. The full hyperparameter configuration is reported in Table 1.

Regarding the training phase, the models were optimized using the Adam algorithm with weight decay to constrain the network weights. For the SepsisExp dataset, we avoided synthetic balancing techniques because the initial class distribution was sufficiently representative to allow the model to capture the underlying patterns. We observed that, when the class imbalance is not extreme, introducing methods such as SMOTE can create artificial samples that do not correspond to actual physiological trajectories, deteriorating performance especially in specificity and false-positive rate. In contrast, the MIMIC-III dataset exhibited a more pronounced class imbalance; consequently, to prevent bias toward the majority class and to enhance sensitivity to rare septic events, an oversampling strategy was implemented. The binary cross-entropy loss with logits (BCEWithLogitsLoss) was adopted as the training objective because it guarantees numerical stability during backpropagation.

Table 1 Hyperparameter configuration of the LSTM models.

Hyperparameter	**Value (common to both datasets)**
Model Architecture	LSTM
Hidden Dimension	8
Number of Layers	4
Pooling Strategy	Mean
Dropout Rate	0.3
Learning Rate	5×10^{-3}
Weight Decay	0
Batch Size	64
Max Epochs	100
Early Stopping Patience	10
Dataset	**Imbalance handling**
SepsisExp	None
MIMIC-III	Oversampling

2.3.2 Performance indicators and validation metrics

We report the evaluation metrics used to validate the performance and predictive capability of the models. Because this is a binary classification problem, we refer to the standard metrics reported in the literature, namely Accuracy, AUC-ROC, and the triad of Precision, Recall, and F1-score, whose formulas are given below.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$\text{AUC-ROC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(t))\, dt \tag{5}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN} \tag{7}$$

$$\text{F1} = \frac{2TP}{2TP + FP + FN} \tag{8}$$

In addition to the metrics above, two further measures were considered:

- *Brier score*, which measures the accuracy of the predicted binary probabilities:

$$\text{Brier Score} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2 \tag{9}$$

- *Matthews Correlation Coefficient (MCC)*, which provides a balanced measure of binary classification quality even under class imbalance:

$$\text{MCC} = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{10}$$

Together, these metrics provide a comprehensive overview of model performance: Accuracy and AUC-ROC offer a general assessment of discriminatory power, whereas the F1-score and MCC ensure that the models are effectively evaluated even in the presence of class imbalance. Furthermore, inclusion of the Brier score allows us to validate the reliability of the predicted probabilities, an essential requirement for clinical decision-support tools.

2.4 Post-hoc Analysis: Permutation Feature Importance

Neural networks are a class of machine-learning models that are particularly difficult to interpret. In the medical context, explainability of model outputs is of paramount importance: for clinicians, it is essential to verify which physiological and metabolic factors determined a prediction, both for validation purposes and for understanding. Several solutions have been proposed over the years, including one of the most effective and easily adaptable to any model: Permutation Feature Importance (PFI) [12].

The procedure is illustrated in Fig. 4 and is as follows. Once a model has been trained, a baseline score is computed on a test set. The choice of the evaluation metric is critical because it must reflect the primary objective of the model. For instance,

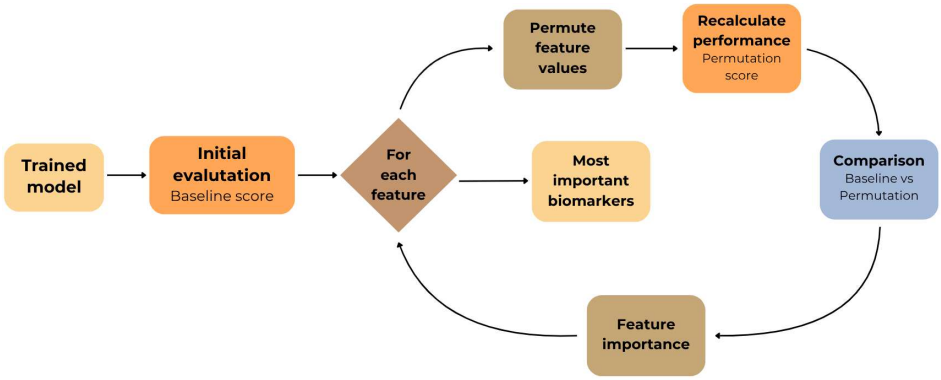


Fig. 4 Workflow of permutation feature importance.

in the clinical domain, which is typically characterized by class imbalance, metrics such as MCC or AUROC are preferred over simple accuracy because they provide a more stringent assessment of the model's ability to identify true-positive cases without inflating performance.

The next step is to permute one feature at a time and to re-evaluate the model's performance on the permuted data. This process yields a new value, the permutation score, which is compared with the baseline: if the score does not decrease, the permuted feature is not relevant; conversely, a decrease in the score indicates that the tested feature is particularly important for the model. The advantages of this methodology lie not only in the possibility of quantitatively computing the importance of each feature, but also in its independence from the architecture used, since the analysis is based solely on model performance.

By integrating this explainability layer, we complete the description of the three pillars of the framework, which jointly cover the typical development phases of medical applications. In the next section, we present the results of each step applied to the two clinical cohorts identified above, detailing the achieved performance and the key physiological drivers identified by the PFI analysis.

2.5 Use of generative AI and AI-assisted technologies

During the preparation of this work the authors used generative AI tools (ChatGPT, Gemini, and Anthropic Claude) solely to improve the language, readability, and organization of the manuscript, and to assist with code refactoring. The authors reviewed and edited all AI-assisted output as needed, verified its correctness, and take full responsibility for the content of the published article. No generative AI tool was used to produce, analyze, or interpret the scientific data, or to generate any scientific conclusion.

3 Results

This section presents the results of the data and phenotype exploration, followed by the performance of the LSTM models on both datasets, and finally the most influential biomarkers across prediction horizons.

3.1 Statistical Validation: Data Quality Assessment and Variable Selection

In the SepsisExp dataset, statistically abnormal variables primarily reflect organ dysfunction, systemic inflammation, and circulatory instability (e.g., SVRI, stroke volume, C-reactive protein). These extreme values are consistent with the clinical literature and reflect the heterogeneous physiological responses typical of septic patients (Fig. 5).

In MIMIC-III, by contrast, abnormal values predominantly characterize respiratory and metabolic instability, with markers such as pCO_2, FiO_2, and tidal volume showing extreme variations that are expected in sepsis (Fig. 6).

In agreement with the clinicians, rather than removing these anomalies we retained them as informative signals for distinguishing between normal and abnormal disease presentations; their total removal would have biased the model toward "average" patients, reducing sensitivity to critical, high-risk cases.

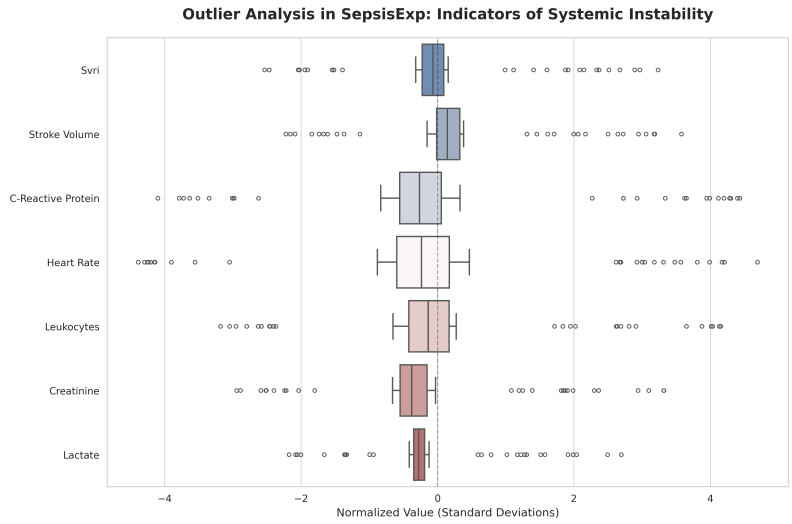


Fig. 5 Distribution and outlier analysis for key physiological variables in the SepsisExp dataset.

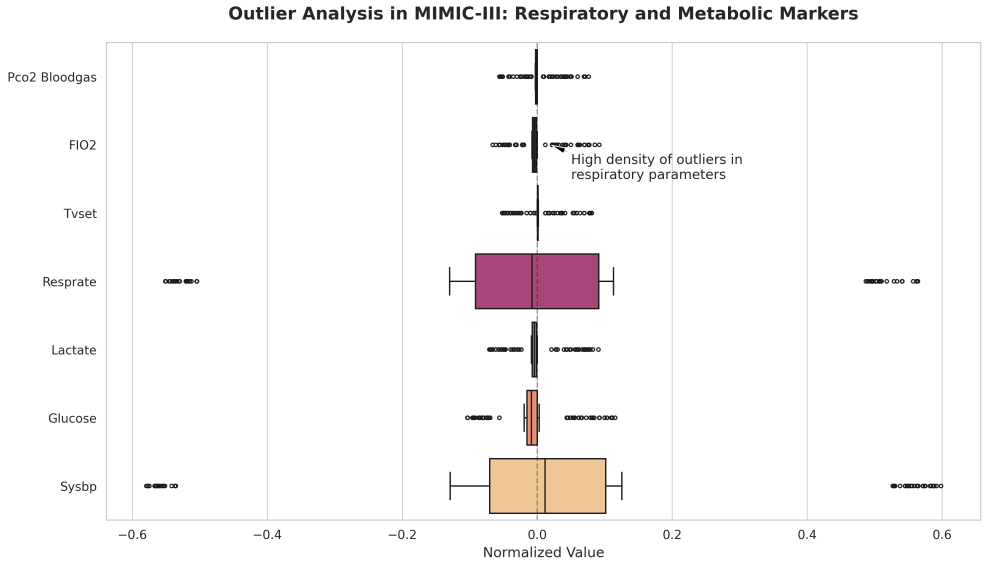


Fig. 6 Distribution and outlier analysis for key physiological variables in the MIMIC-III dataset.

Once the clinical integrity of the individual parameters had been confirmed, correlation analysis was performed to identify redundant parameters and to avoid collinearity issues that could impair model learning. The analysis revealed markedly different

structural profiles between the two datasets. SepsisExp showed few but strong correlations that were consistent with the clinical literature (e.g., cardiac output–stroke volume, lactate–acid–base markers, blood pressure measurements), while most variables remained essentially independent (Fig. 7). Accordingly, the experts decided not to pursue a substantive reduction of the feature set, preferring a more conservative approach.

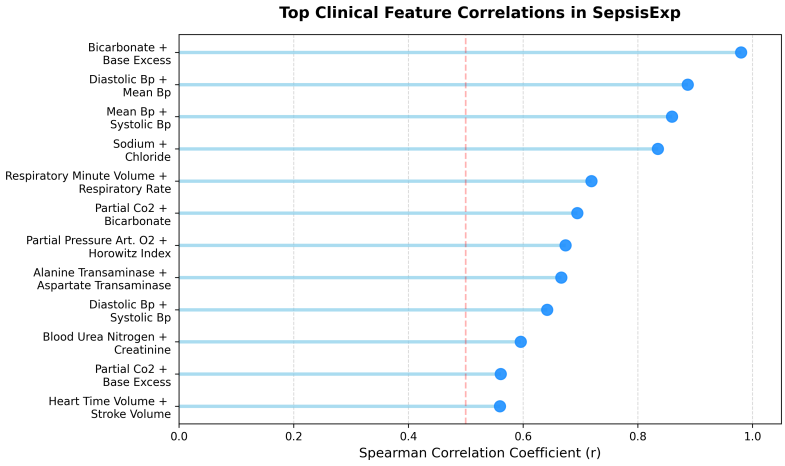


Fig. 7 Ranking of the most significant Spearman correlations ($r > 0.5$) in the SepsisExp dataset.

In contrast, MIMIC-III exhibited extensive redundancy, with many expected clinical correlations (INR–PT, hematocrit–hemoglobin, blood pressure) and consistent physiological groups (renal, cardiac, hepatic, respiratory) (Fig. 8). Consequently, with the support of the physicians, the feature set was reduced from 47 to 34 variables.

3.2 Patient Phenotyping: Sepsis Risk Stratification

Having established the structure and quality of the data, we proceeded to a more detailed exploration of the patient subpopulations. Using ts-KMeans with elbow-method optimization, we stratified both cohorts into distinct clinical phenotypes. In SepsisExp, five clusters revealed three main risk profiles (Fig. 9):

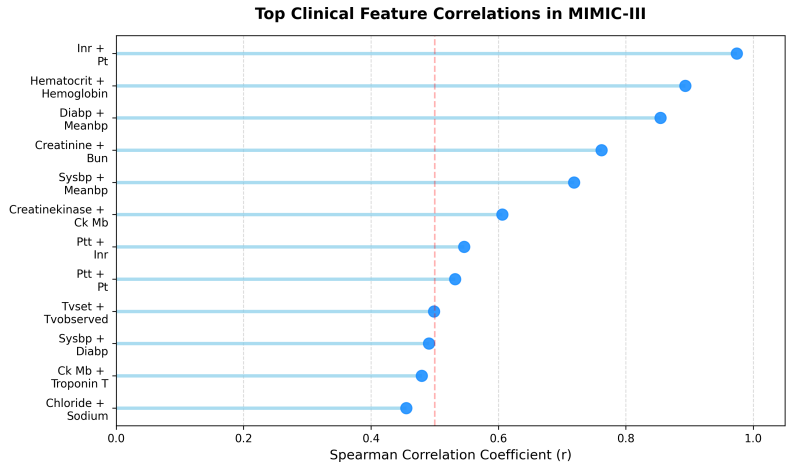


Fig. 8 Ranking of the top correlations in the MIMIC-III dataset.

- *High-risk septic phenotype (C2–C3):* C2 (59% prevalence) exhibits typical sepsis characteristics with inflammatory response and hemodynamic decline; C3 (24%) presents milder risk and instability. These clusters represent the primary targets for early intervention.
- *Critical non-septic outliers (C0):* the smallest group, with extreme metabolic and renal dysfunction but only 16% sepsis prevalence, suggesting organ failure of other causes.
- *Stable baseline (C1–C4):* the largest groups, with minimal physiological changes and low sepsis prevalence (3.6–16%), ensuring robustness against false positives.

Whereas the previous case focused primarily on hemodynamics and inflammation, in the heterogeneous MIMIC-III cohort the focus shifts to renal function and coagulation as primary risk factors across five clusters (Fig. 10):

- *High-risk renal phenotype (C0):* the highest-risk group, with the largest percentage of positive cases (23%). Its distinctive features are severe renal dysfunction (elevated creatinine and BUN levels) and coagulopathy (prolonged INR/PT). Combined with the presence of hyperkalemia (elevated potassium levels), these findings suggest acute kidney injury.

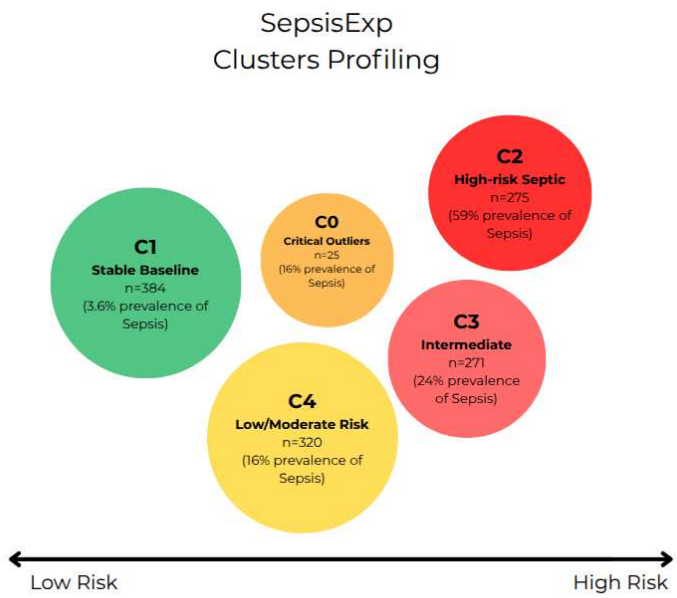


Fig. 9 Cluster profiling in SepsisExp according to different levels of risk.

- *Respiratory and vital instability group (C2):* a group with a still elevated level of risk (15%) whose patients are characterized by compensation for their state of distress, as evidenced by significant signs of tachypnea, tachycardia, and elevated FiO_2, likely indicating an early stage of the disease or respiratory failure.
- *Stable, low-risk baseline (C3–C4):* the healthiest group, with the lowest risk (2.5–8%). Patients in group C4 are characterized by normal kidney function, above-average blood pressure, and generally stable vital signs; in group C3, despite slight hemodynamic fluctuations, no obvious signs of systemic infection are found.

Although cluster C1 shows extreme physiological values consistent with mechanical ventilation, it is too small for robust generalization.

At the end of this phase, clustering revealed latent, clinically valuable dynamics in both cohorts, enabling a transition from a "one-size-fits-all" approach to phenotype-targeted monitoring and therapy.

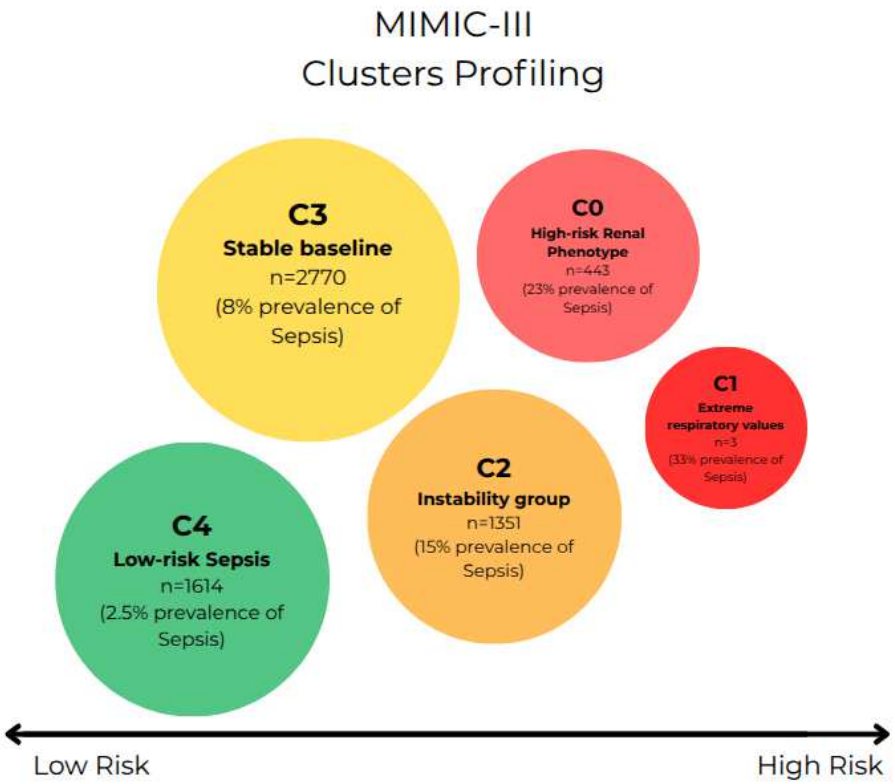


Fig. 10 Cluster profiling in MIMIC-III according to different levels of risk.

3.3 Comparative Performance Analysis: Stability and Predictive Horizons

This section presents the results on the predictive capability of the models for the two populations under study. For both cohorts, we used a fixed 24-hour input window and prediction horizons $T \in \{1, 2, 3, 6, 9, 12\}$ hours, with attention to the balance between sensitivity and specificity (Tables 2 and 3).

For the SepsisExp cohort, peak performance is reached at $T = 3$ (ACC = 84.7%, AUC = 0.884), with well-balanced recall (0.648) and precision (0.678)—a clinically favorable window for timely therapeutic intervention. This strong performance stems from the LSTM's ability to capture the discriminative physiological patterns

Table 2 Detailed performance metrics for the SepsisExp dataset across different prediction horizons (T). Values for Precision, Recall, and F1-score refer specifically to the sepsis class (class 1).

T (h)	ACC	AUC	MCC	Brier	Prec (S)	Rec (S)	F1 (S)
1	0.842	0.877	0.560	0.133	0.672	0.655	0.659
2	0.835	0.882	0.550	0.141	0.652	0.668	0.652
3	0.847	0.884	0.564	0.132	0.678	0.648	0.663
6	0.821	0.868	0.516	0.154	0.617	0.652	0.629
9	0.832	0.873	0.527	0.146	0.641	0.631	0.635
12	0.838	0.870	0.541	0.141	0.654	0.638	0.641

already identified during clustering. Across all horizons, the model exhibits remarkable resilience: the AUC remains stable around 0.87–0.88 and precision never drops below 61%, demonstrating that the architecture learns robust temporal features that persist even at 12-hour forecasts. The balanced precision–recall trade-off minimizes alarm fatigue, while the Brier score (0.132 at $T = 3$) confirms good probability calibration. Finally, an MCC of 0.564 attests to robust predictive power well beyond chance, despite class imbalance.

For the MIMIC-III cohort, peak performance occurs at $T = 1$ (ACC = 88.9%, AUC = 0.809, Brier = 0.096), reflecting strong discriminatory power near disease onset. From a clinical standpoint, however, the $T = 12$ results are equally relevant: despite lower accuracy, the model still identifies 60.9% of septic patients, enabling preventive action well in advance. Across horizons, the AUC remains stable (0.76–0.81), confirming consistent discrimination between patient types. Beyond $T = 9$, precision drops to approximately 0.25 while recall increases and stabilizes, revealing a long-term "sensitive" bias that is strategically valuable in ICU settings, where a missed alarm carries greater cost than a false positive. This trade-off is further mitigated by the very high performance on the non-septic class (F1 $\cong$ 0.88), which keeps false alarms among non-septic patients to a minimum.

The comparative evaluation across both cohorts underscores the robustness of the LSTM architecture and its ability to balance diagnostic accuracy with the clinical

Table 3 Detailed performance metrics for the MIMIC-III dataset across different prediction horizons (T). Note the higher accuracy compared with SepsisExp, contrasted by lower precision and MCC due to the more pronounced class imbalance.

T (h)	ACC	AUC	MCC	Brier	Prec (S)	Rec (S)	F1 (S)
1	0.889	0.809	0.418	0.096	0.424	0.540	0.474
2	0.880	0.799	0.379	0.106	0.388	0.512	0.440
3	0.873	0.796	0.345	0.111	0.359	0.477	0.410
6	0.862	0.775	0.361	0.116	0.345	0.547	0.422
9	0.798	0.763	0.280	0.154	0.247	0.577	0.345
12	0.801	0.788	0.303	0.154	0.260	0.609	0.363

need for early warnings. Although the performance differences stem from varying class imbalances—most pronounced in MIMIC-III—the overall stability of the models highlights their potential as reliable decision-support tools that can help clinicians manage sepsis more timely and effectively.

3.4 Temporal Dynamics of Predictors: Shifting Biomarker Relevance Across Horizons

In this final section, we analyze feature importance across prediction horizons in order to map how risk factors evolve from the early stages to sepsis onset.

Analysis within the SepsisExp cohort clearly reveals a temporal shift among the key predictors (Fig. 11), with one notable exception: CRP remains the dominant predictor across all horizons. Beyond this constant, feature importance redistributes into three phases:

- *Short term (1–2 h):* the model primarily observes signs of stress compensation. This is evident from the increased importance assigned to respiratory rate and cardiac volume, as well as to the quick Sequential Organ Failure Assessment (qSOFA) score and potassium.
- *Mid term (3–6 h):* the model places greater weight on organ status and function. Leukocytes and hemoglobin increase in importance, indicating that the model

attends significantly to the immune response at this stage. Furthermore, there is particular interest in proper kidney function, as also indicated by the increased relevance of the BUN and urine-output variables. Many features exhibit similar high values ($\sim$ 0.10), demonstrating that the model simultaneously captures multiple aspects of the disease, which partly explains the high AUC observed at the 6-hour horizon.

- *Long term (9–12 h):* the model looks for signs of chronic hepatic, pancreatic, or metabolic distress, as demonstrated by the increasing importance assigned to bilirubin, pancreatic lipase, and hemoglobin. Furthermore, net balance and glucose also rank highly, suggesting greater importance given to hydration status and metabolic stress at this stage. Variables such as minute respiratory volume and lymphocytes are effectively ignored by the model, as they exhibit negative importance values.

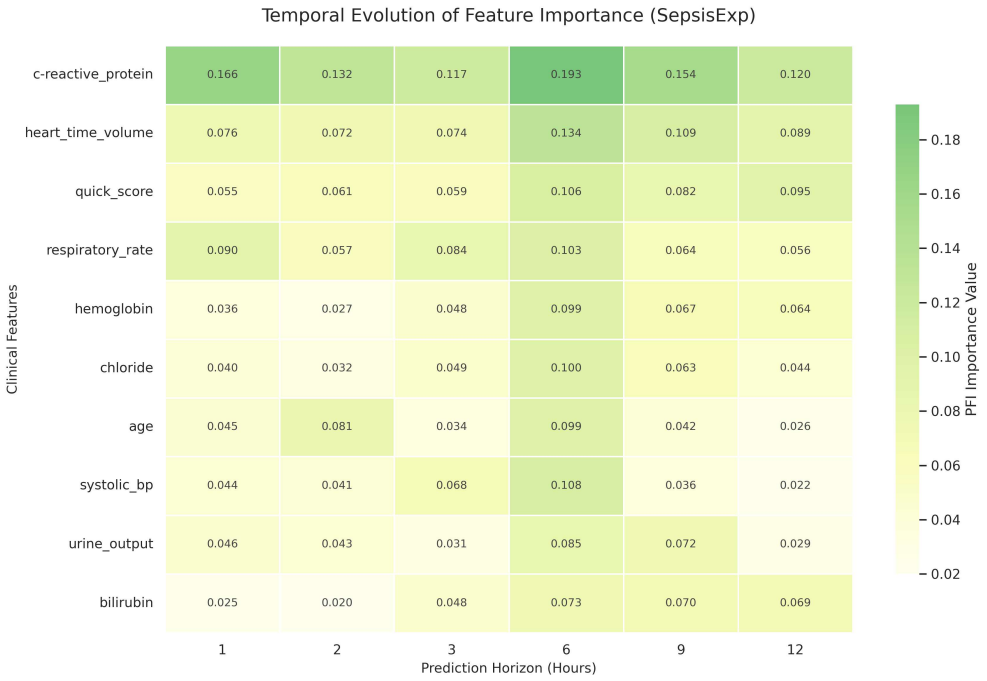


Fig. 11 Heat map illustrating how feature importance changes in SepsisExp over different prediction horizons. C-reactive protein is the predominant feature.

Feature-relevance analysis using the MIMIC-III trained model reveals dynamics somewhat different from the previous case (Fig. 12), with a stronger focus on hemodynamic and respiratory parameters. Systolic blood pressure (sysbp) serves as the dominant pivot for monitoring shock and hypotension. The temporal dynamics again follow three phases:

- *Short term (1–2 h):* the model focuses primarily on identifying systemic collapse. In addition to systolic blood pressure, variables such as hematocrit and hemoglobin emerge as indicators of hemoconcentration or of the need for transfusion. The model also looks for possible indicators of respiratory failure, such as FiO_2 and lactate, which signal possible tissue hypoxia.
- *Mid term (3–6 h):* the model primarily considers acid–base balance and renal function, as demonstrated by the importance assigned to blood pH and BUN. Furthermore, significant weights are assigned to diastolic blood pressure (diabp) and mean arterial pressure (meanbp), reconstructing the patient's complete blood pressure profile.
- *Long term (9–12 h):* the model attempts to identify signs of organ instability and distress. Hemodynamic variables are completely dominant at this point, far from the onset of the disease. Interestingly, bilirubin becomes more prominent, suggesting that the model can detect liver dysfunction many hours before the onset of the disease.

This analysis of feature importance proves useful not only in overcoming the limitations imposed by the "black-box" nature of neural networks, but also as a reliable tool for measuring disease progression. Providing clinicians not only with the model's predictions but also with a real-time indication of which variables influence risk offers significant added value.

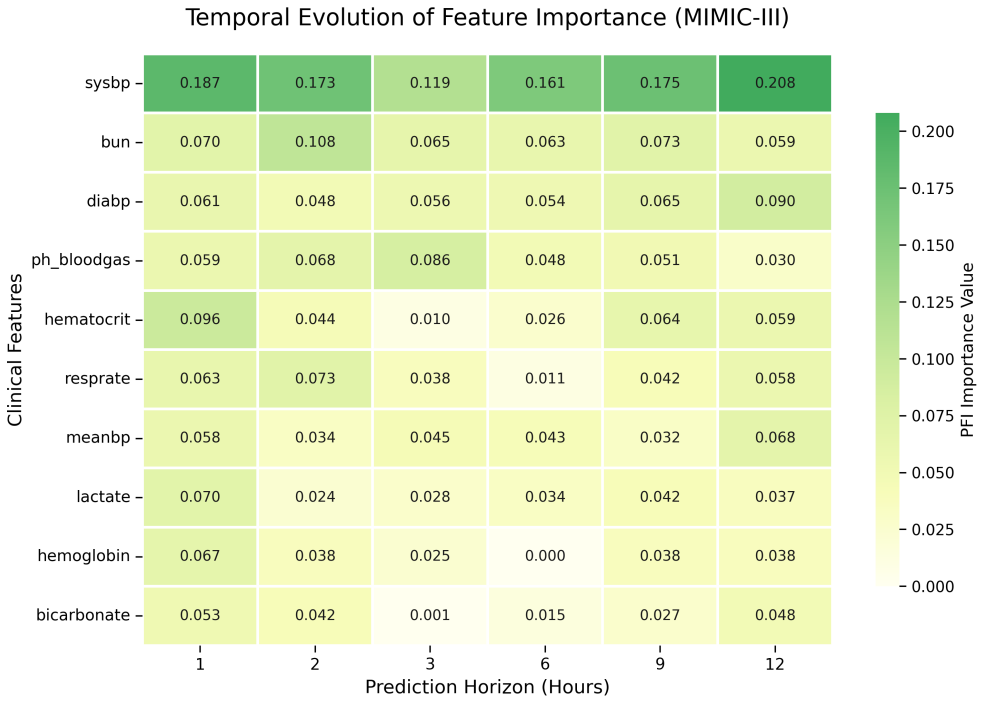


Fig. 12 Heat map illustrating how feature importance changes in MIMIC-III over different prediction horizons. Sysbp is the predominant feature.

4 Discussion

We validated an LSTM-based framework for early sepsis prediction across two heterogeneous clinical cohorts, SepsisExp and MIMIC-III. The approach rests on three pillars: exploratory data analysis, architecture optimization, and permutation-based feature importance assessment. We identified distinct patient phenotypes reflecting varying risk levels, ranging from silent metabolic stress to acute hemodynamic collapse. Furthermore, the LSTM networks demonstrated robust, stable performance across multiple prediction horizons, offering clinicians a valuable window for proactive intervention. Finally, the retrospective analysis revealed that the models dynamically shift their focus over time, capturing early organ-specific signals before concentrating on systemic deterioration markers as sepsis onset approaches.

The consistency of the results across two structurally different datasets supports the generality of the pipeline: the same design principles—conservative feature reduction guided by clinical feedback, a deliberately lightweight architecture, and imbalance handling calibrated to the severity of the imbalance—yielded stable discrimination in both cohorts. Moreover, the observed temporal re-weighting of biomarkers is clinically plausible: short-horizon predictions rely on acute compensation signals (respiratory rate, blood pressure, lactate), whereas long-horizon predictions attend to slower organ-level derangements (renal and hepatic markers). This behavior is coherent with the pathophysiology of sepsis and with the aim of making "black-box" predictions auditable at the bedside [12, 13].

4.1 Comparison with published benchmarks

To situate the PRISM framework, we compare it against published sepsis-prediction benchmarks on the same two cohorts, matching dataset and metric wherever possible (Table 4). The reference studies differ in evaluation protocol, prediction horizon, and control population, so the comparison is indicative rather than a strict head-to-head; we flag the relevant differences below.

Table 4 Comparison of PRISM against published sepsis-prediction benchmarks on the same cohorts. For SepsisExp the PRISM value is computed with the *same* all-timesteps protocol as Schamoni et al. (pooling over the prediction interval and the preceding hospital stay); for MIMIC-III, PRISM values are peak AUC/AUPRC. Benchmark values are the reported ranges (AUROC and AUC denote the same metric).

Cohort	Benchmark	Benchmark metric	PRISM
SepsisExp	Schamoni et al. [7]	AUROC 0.68–0.78 (all-timesteps)	AUROC 0.70–0.71 (aligned)
MIMIC-III	Moor et al. (GP-TCN) [8]	AUPRC 0.35–0.51	AUPRC 0.47 ($T = 1$ h)
MIMIC-III	Goh et al. (SERA) [10]	AUC 0.87–0.94	AUC 0.809 ($T = 1$ h)
MIMIC-III	Camacho-Cogollo et al. [9]	AUC 0.77–0.92	AUC 0.809 ($T = 1$ h)

SepsisExp. To make the comparison with Schamoni et al. [7] a fair head-to-head, we recompute PRISM's AUROC with their exact protocol: the risk score is pooled

over *all* timesteps of the prediction interval *and* the preceding hospital stay (their "all-timesteps" convention), with a per-timestep label indicating whether sepsis onsets within the lead time. Pooling over the patient's full ICU stay—the most conservative reading of their "preceding time of hospital stay"—PRISM attains an AUROC of 0.70–0.71 across 4/8/12 h lead times. Against the *full model* of Schamoni et al. (0.68–0.71, their Table 4) this is in line at 4 h and 8 h and ahead at 12 h; against their *best ensemble* (0.71–0.78) PRISM is competitive, within 0.05 at 4 h/8 h and within 0.001 at 12 h. Notably, PRISM is a far more compact model (a 4-layer bidirectional LSTM with 4 hidden units) than the reference (a 2-layer LSTM with 200 units), so matching the full model under an identical protocol is a meaningful result. The two are directly comparable: same cohort (1,275 patients) and same 24 h input window.

MIMIC-III. On MIMIC-III, the most informative comparison is on the area under the precision–recall curve, the appropriate metric for this strongly imbalanced cohort. On this metric PRISM is competitive with the Gaussian-process TCN family of Moor et al. [8]: PRISM's AUPRC (0.36–0.47 across horizons, peaking at 0.47 at $T = 1$ h; computed on the aligned evaluation) exceeds the MGP-RNN baseline (0.25) and the MGP-TCN (0.35) reported at 7 h, and is comparable to the best reported DTW-KNN (0.40).

Where PRISM trails. Two aspects, where PRISM trails the strongest references, are noted briefly. On MIMIC-III the peak AUC (0.809 at $T = 1$ h) is below the SERA algorithm of Goh et al. [10] (0.87–0.94) and at the lower end of the range reported by Camacho-Cogollo et al. [9] (0.77–0.92); correspondingly, PRISM's sepsis-class recall (0.48–0.61) is lower than SERA's 0.87 sensitivity. We regard this as a design trade-off—PRISM favours a stable, interpretable, long-horizon signal over peak discrimination—and a concrete target for the SMOTE-based imbalance handling outlined in the Conclusion, rather than a limitation of the present framework.

From a clinical perspective, the framework is intended as a decision-support tool rather than an autonomous diagnostician. The combination of phenotype stratification, stable early-warning scores, and time-resolved explanations can prioritize patients for closer monitoring, reduce alarm fatigue through balanced precision–recall trade-offs, and help clinicians interpret model output in terms of familiar physiological variables.

Several limitations should be acknowledged. First, both datasets are retrospective and derive from specific institutional contexts; MIMIC-III in particular reflects a single-center US ICU population, and the generalizability of the models to other healthcare settings remains to be demonstrated. Second, sepsis labels depend on consensus definitions and expert annotation, which are known sources of label noise and of heterogeneity in onset determination. Third, the class imbalance in MIMIC-III constrains precision at longer horizons, and although the oversampling strategy mitigates majority-class bias, more advanced imbalance-handling and modeling strategies (e.g., SMOTE-ENN, SMOTE-Tomek, Transformers, or Temporal Convolutional Networks) may further improve performance. Fourth, the present evaluation reports aggregate metrics; a dedicated assessment of calibration curves, decision-curve analysis, and alarm rates in a simulated clinical workflow would be required before deployment. Prospective, multicenter validation is the necessary next step before any clinical use.

5 Conclusion

Despite the promising results, addressing dataset limitations and class imbalance could further enhance performance. While we employed basic resampling, future work could leverage advanced strategies such as SMOTE, SMOTE-ENN, or SMOTE-Tomek. Architecturally, exploring alternative paradigms—such as Transformers or Temporal

Convolutional Networks (TCNs)—could yield further gains. Additionally, multi-horizon disease-progression modeling and ensemble techniques represent promising directions.

Crucially, this work is designed to augment and support clinical judgment rather than supplant it, complementing physician expertise through data-driven tools. In time-critical conditions such as sepsis, integrating phenotypic mapping, early warning, and dynamic biomarker monitoring can improve therapeutic administration. Ultimately, this approach transforms vague alarms into actionable clinical insights, enabling effective resource allocation and personalized therapeutic strategies from the earliest stages of disease progression.

Declarations

Ethics approval and consent to participate. This study used only publicly available, fully de-identified datasets and did not involve any direct recruitment of human or animal subjects. MIMIC-III was accessed under the terms of the PhysioNet Credentialed Health Data License; the original data-collection studies were approved by the Massachusetts Institute of Technology (MIT) Committee on the Use of Humans as Experimental Subjects (COUHES) and the Institutional Review Board (IRB) of Beth Israel Deaconess Medical Center (BIDMC), with a waiver of informed consent granted for the de-identified data [14]. For the SepsisExp dataset, the original data collection (Ground Truth for Sepsis Questionnaire survey and clinical data extraction at the University Medical Centre Mannheim, Germany) was approved by the Medical Ethics Commission II of the Medical Faculty Mannheim, Heidelberg University, Germany, with a waiver of patient consent (reference number 2016-800R-MA) [18]. All methods were carried out in accordance with relevant guidelines and regulations.

Consent for publication. Not applicable.

Availability of data and materials. MIMIC-III (v1.4) is publicly available on PhysioNet (https://physionet.org/content/mimiciii/1.4/) under a credentialed-access license. SepsisExp is publicly available under a Creative Commons Attribution 4.0 International (CC BY 4.0) license at https://www.cl.uni-heidelberg.de/statnlpgroup/sepsisexp/ and is described in [7, 18].

Code availability. The code and analysis scripts supporting the findings of this study are available in the GitHub repository https://github.com/silvia1970/ml-prism and are archived with a persistent identifier at https://doi.org/10.5281/zenodo.22789461 [19].

Competing interests. The authors declare that they have no competing interests.

Funding. This work was supported by the European Union – NextGenerationEU – National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, through the Research Program "National Centre for HPC, Big Data and Quantum Computing", Project CN00000013, Spoke 6, CUP B83C22002940006 (Cascade Funding Spoke 6, Project "PRISM – Predictive Risk Identification of Sepsis with Machine Learning"). The funding body had no role in the design of the study, in the collection, analysis, and interpretation of the data, or in the writing of the manuscript.

Acknowledgements. Not applicable.

Authors' contributions. AA: methodology, software, formal analysis, investigation, resources, data curation, writing—original draft, writing—review and editing, visualization. MF: software, formal analysis, investigation, resources, data curation, writing—review and editing. SS (Spoto): conceptualization, investigation, data curation, writing—review and editing, supervision. AL: conceptualization, formal analysis, investigation, resources, data curation, writing—review and editing, supervision. SF: conceptualization, investigation, resources, writing—review and editing, supervision.

SA (Angeletti): conceptualization, investigation, writing—review and editing, supervision. MVR: investigation, writing—review and editing. GC: investigation, writing—review and editing. AP: conceptualization, writing—review and editing, supervision. MG: conceptualization, writing—review and editing, funding acquisition. AS: conceptualization, methodology, software, formal analysis, investigation, resources, writing—original draft, writing—review and editing, supervision, funding acquisition. SS (Scarpetta): conceptualization, investigation, resources, writing—review and editing, supervision, funding acquisition. LC: conceptualization, investigation, resources, writing—review and editing, supervision, project administration, funding acquisition. All authors read and approved the final manuscript.

References

[1] Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016;315(8):801–810. https://doi.org/10.1001/jama.2016.0287.

[2] Prescott HC, et al. Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock. Critical Care Medicine. 2026;54(4):725–812. https://doi.org/10.1097/CCM.0000000000007075.

[3] Rudd KE, Johnson SC, Agesa KM, Shackelford KA, Li T, Ko D, et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. The Lancet. 2020;395(10219):200–211. https://doi.org/10.1016/S0140-6736(19)32989-7.

[4] Fleischmann C, Scherag A, Adhikari NKJ, Hartog CS, Tsaganos T, Schlattmann P, et al. Assessment of Global Incidence and Mortality of Hospital-treated Sepsis. Current Estimates and Limitations. American Journal of Respiratory

and Critical Care Medicine. 2016;193(3):259–272. https://doi.org/10.1164/rccm.201504-0781OC.

[5] Rhee C, Murphy MV, Li L, Platt R, Klompas M. Comparison of trends in sepsis incidence and coding using administrative claims versus objective clinical data. Clinical Infectious Diseases. 2015;60(1):88–95. https://doi.org/10.1093/cid/ciu710.

[6] Liu AC, Patel K, Vunikili RD, Johnson KW, Abdu F, Belman SK, et al. Sepsis in the era of data-driven medicine: personalizing risks, diagnoses, treatments and prognoses. Briefings in Bioinformatics. 2020;21(4):1182–1195. https://doi.org/10.1093/bib/bbz059.

[7] Schamoni S, Hagmann M, Riezler S. Ensembling Neural Networks for Improved Prediction and Privacy in Early Diagnosis of Sepsis. Proceedings of Machine Learning Research. 2022;182:1–23.

[8] Moor M, Horn M, Rieck B, Roqueiro D, Borgwardt K.: Early Recognition of Sepsis with Gaussian Process Temporal Convolutional Networks and Dynamic Time Warping. arXiv preprint arXiv:1902.01659.

[9] Camacho-Cogollo JE, Bonet I, Gil B, Iadanza E. Machine Learning Models for Early Prediction of Sepsis on Large Healthcare Datasets. Electronics. 2022;11(9):1507. https://doi.org/10.3390/electronics11091507.

[10] Goh KH, Wang L, Yeow AYK, Poh H, Li K, Yeow JJL, et al. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nature Communications. 2021;12(1):711. https://doi.org/10.1038/s41467-021-20910-4.

[11] Hernandez B, Ming DK, Rawson TM, Bolton W, Wilson R, Vasikasin V, et al. Advances in diagnosis and prognosis of bacteraemia, bloodstream infection, and sepsis using machine learning: a comprehensive living literature review. Artificial Intelligence in Medicine. 2025;160:103008. https://doi.org/10.1016/j.artmed.2024.103008.

[12] Mi X, Zou B, Zou F, Hu J. Permutation-based identification of important biomarkers for complex diseases via machine learning models. Nature Communications. 2021;12(1):3008. https://doi.org/10.1038/s41467-021-22756-2.

[13] Staniek M, Fracarolli M, Hagmann M, Riezler S.: Early Prediction of Causes (not Effects) in Healthcare by Long-Term Clinical Time Series Forecasting. arXiv preprint arXiv:2408.03816.

[14] Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific Data. 2016;3:160035. https://doi.org/10.1038/sdata.2016.35.

[15] Khaled A, Sabir M, Qureshi R, Caruso CM, Guarrasi V, Xiang S, et al.: Leveraging MIMIC Datasets for Better Digital Health: A Review on Open Problems, Progress Highlights, and Future Promises. arXiv preprint arXiv:2506.12808.

[16] Seymour CW, Liu VX, Iwashyna TJ, Brunkhorst FM, Rea TD, Scherag A, et al. Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016;315(8):762–774. https://doi.org/10.1001/jama.2016.0288.

[17] Solís-García J, Vega B, Nepomuceno J, Riquelme J, Nepomuceno-Chamorro I. Comparing artificial intelligence strategies for early sepsis detection in the ICU: an experimental study. Applied Intelligence. 2023;53. https://doi.org/10.1007/

s10489-023-05124-z.

[18] Lindner HA, Schamoni S, Kirschning T, Worm C, Hahn B, Centner FS, et al. Ground truth labels challenge the validity of sepsis consensus definitions in critical illness. Journal of Translational Medicine. 2022;20(1):27. https://doi.org/10.1186/s12967-022-03228-7.

[19] Aiezzo A.: ml-prism: code and analysis scripts for the PRISM sepsis prediction framework. Zenodo.