Causal Structure Learning for Sepsis Prediction in ICU Patients

Prabhav Singh, Aravind Kavuturu, Sandesh Rangreji, Mukund Iyengar
Johns Hopkins University, Computer Science Department
{psingh54, akavutu2, srangre1, miyenga2}@jh.edu
Johns Hopkins University
Baltimore, MD, USA

ABSTRACT

Sepsis is a life-threatening condition caused by an uncontrolled immune response to infection, often resulting in organ failure and high mortality rates, particularly in Intensive Care Units (ICUs). Early detection is critical to improving outcomes, but the heterogeneous nature of sepsis makes accurate prediction challenging. Leveraging the PhysioNet Challenge 2019 dataset, which includes time-series clinical data with substantial missing values, we propose a hybrid machine learning framework for early sepsis detection. Our method involves robust data preprocessing, where missing values are handled under the assumption of Missing Completely at Random (MCAR), and time-series features are aggregated into mean, mode, and last observed values. Using these features, we employ Bayesian Networks for interpretable feature selection, along with XGBoost as a baseline to evaluate predictive performance under different experimental setups. Our approach balances interpretability with predictive accuracy, demonstrating the potential of using probabilistic models for critical healthcare tasks. Our code is available on GitHub.

1 INTRODUCTION

Sepsis remains one of the most pressing challenges in modern medicine, contributing to a staggering number of deaths globally and placing an enormous burden on healthcare systems worldwide. According to recent estimates, sepsis affects approximately 48.9 million people annually and is responsible for 11 million deaths, accounting for nearly 20% of all global deaths [1]. This mortality rate surpasses that of many common cancers combined, underscoring the critical nature of this condition [2].

The impact of sepsis extends far beyond mortality statistics. It is a leading cause of hospital re-admissions and long-term morbidity, with survivors often experiencing prolonged physical and cognitive impairments [3]. The economic burden is equally significant, with sepsis accounting for an estimated 2.65% of healthcare budgets globally, and a median hospital cost exceeding $32,000 per septic patient in high-income countries [3].

Table 1: Global Impact of Sepsis (2017 Estimates)

Metric	Value
Annual cases worldwide	48.9 million
Annual deaths	11 million
Percentage of global deaths	20%
Cases in children under 5	20 million
Maternal sepsis cases	5.7 million
Deaths associated with AMR	4.95 million

The disproportionate impact of sepsis on vulnerable populations is particularly concerning. Almost half of all estimated sepsis cases worldwide occur in children under 5 years of age, and there are approximately 5.7 million maternal sepsis cases annually [3]. Furthermore, low- and middle-income countries (LMICs) bear 85% of the sepsis burden due to limited healthcare resources and high rates of healthcare-associated infections [2].

Despite its prevalence and severity, sepsis remains a challenging condition to diagnose and manage effectively. Early diagnosis is crucial for improving outcomes, as studies show that each hour of delay in initiating appropriate antimicrobial therapy is associated with a significant increase in mortality [4]. However, the complex and heterogeneous nature of sepsis makes early and accurate diagnosis difficult.

Recent advancements in machine learning have shown promising potential in automating sepsis prediction using electronic health records (EHR). These technologies offer the possibility of early, accurate, and continuous risk assessment, potentially revolutionizing sepsis management in intensive care units (ICUs) and beyond [5]. For instance, a recent study utilizing the COMPOSER algorithm, a deep-learning model for sepsis prediction, demonstrated a 17% relative decrease in in-hospital sepsis mortality and a 10% relative increase in sepsis bundle compliance [6].

Motivated by these challenges and opportunities, this study explores a multi-step pipeline that combines probabilistic models with advanced machine learning techniques to improve sepsis prediction. We utilize the PhysioNet Challenge 2019 dataset, a comprehensive collection of ICU patient data specifically curated for sepsis prediction tasks [7]. Our approach involves several key steps:

(1) **Data Preprocessing**: We employ aggregation techniques to handle missing values and time-series data and normalize the data, ensuring a robust foundation for our models.
(2) **Feature Selection**: A Bayesian Network with Hill Climb Search for local search is implemented for feature selection and interpretable modeling, providing insights into the relationships between clinical variables and sepsis onset.
(3) **Machine Learning Prediction**: We combine the features selected by our Bayesian Network with XGBoost, a powerful gradient boosting algorithm, in two configurations after hyperparameter tuning:
   - Using feature subsets derived from the Bayesian Network.
   - Using all available features.
(4) **Grid Search CV for Hyperparameter Tuning**: We fit both the above configurations with Grid Search CV to choose the best performing XGBoost hyperparameters.
(5) **Model Comparison and Evaluation**: The performance of our models is rigorously assessed using metrics such as

Precision, Recall, and F1 score while considering clinical relevance.

Our study aims to contribute to the growing body of research on automated sepsis prediction by providing a balanced approach that combines interpretability with predictive power. By doing so, we hope to develop a system that not only accurately predicts sepsis onset but also provides clinically relevant insights to aid decision-making.

The potential impact of improved sepsis prediction is substantial. Even modest reductions in mortality could save hundreds of thousands of lives annually. Moreover, early detection could significantly reduce long-term morbidity associated with sepsis, improving quality of life for survivors.

In the following sections, we detail our dataset (§ 3), methodology (§ 4), present our findings (§ 5), and discuss implications for clinical practice (§ 6).

2 RELATED WORK

Numerous studies have explored the application of ML techniques for sepsis detection, leveraging both traditional and advanced methods. Logistic regression (LR) and random forests (RF) are among the most commonly used baseline models due to their simplicity and robustness. For instance, Kam et al. [8] demonstrated the utility of LR and RF in predicting sepsis using vital signs and laboratory results. Similarly, Wang et al. [9] applied logistic regression, support vector machines, and logistic model trees to predict sepsis onset in ICU patients, with logistic model trees outperforming other methods.

Recent advancements in deep learning have enabled the use of more complex models capable of capturing intricate relationships in EHR data. For example, a multi-output Gaussian process combined with recurrent neural networks (MGP-RNN) was employed by Moor et al. [10] to detect sepsis earlier than traditional methods, achieving a C-statistic of 0.88 and detecting sepsis a median of 5 hours in advance. Similarly, ensemble learning techniques that integrate deep features extracted by long short-term memory (LSTM) networks with manually engineered features have shown promise for early sepsis prediction [11].

The PhysioNet Challenge 2019 dataset has become a popular benchmark for evaluating sepsis prediction models [12]. Many participants achieved promising results by employing techniques such as data imputation, feature engineering, and ensemble learning. For instance, Hammoud et al. [13] utilized gradient-boosted decision trees (GBDTs) to predict sepsis onset, demonstrating strong performance through optimal sample weighting and time-series feature aggregation.

Despite these advances, interpretability remains a significant challenge for many deep learning models. Probabilistic graphical models (PGMs), such as Bayesian networks, offer an interpretable alternative that can effectively handle missing data while providing insights into variable dependencies [14]. However, their predictive performance often lags behind more sophisticated ML methods like gradient boosting or neural networks [15]. To address this gap, hybrid approaches that combine PGMs for feature selection with gradient-boosted decision trees for classification have been proposed but remain underexplored.

Another notable advancement is the integration of natural language processing (NLP) with structured EHR data for sepsis prediction. The SERA algorithm developed by Shashikumar et al. [16] combines NLP-extracted features from clinical notes with structured data to predict sepsis onset up to 48 hours in advance, achieving an AUC of 0.94 while significantly reducing false positives compared to traditional scoring systems like qSOFA and MEWS.

In summary, while traditional ML methods such as LR and RF provide robust baselines for sepsis prediction, advanced techniques like deep learning and ensemble learning have demonstrated superior predictive performance on complex datasets like PhysioNet 2019. However, challenges related to interpretability and generalizability persist. This motivates our approach of combining probabilistic graphical models with gradient-boosted decision trees to achieve a balance between interpretability and predictive accuracy.

3 DATASET DESCRIPTION

The PhysioNet Challenge 2019 dataset provides a comprehensive resource for the early detection of sepsis in ICU patients. It includes de-identified electronic health records (EHRs) sourced from three geographically distinct U.S. hospital systems: Beth Israel Deaconess Medical Center (System A), Emory University Hospital (System B), and an undisclosed third system (System C) used only for testing during the challenge. We will only consider the data we have access to, Systems A and B. These records span over a decade, offering a diverse patient population and clinical conditions. The dataset is specifically designed to encourage robust, generalizable algorithms by incorporating data from multiple institutions.

3.1 Data Composition

The dataset comprises the following:

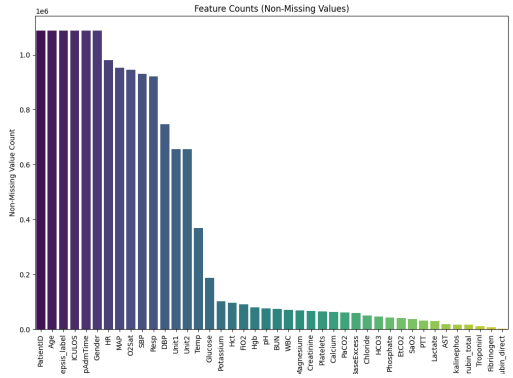
- **Patient Records**: Over 60,000 patient records, with 40,336 made available for training and validation (Systems A and B) and 24,819 sequestered for hidden testing (Systems A, B, and C).
- **Hourly Time-Series Data**: More than 2.5 million hourly time windows and 15 million data points were included, aggregated into hourly bins to simplify analysis.
- **Clinical Variables**: A total of 40 clinical variables categorized into:
  - **Vital Signs (8 variables)**: Heart rate (HR), oxygen saturation (O2Sat), temperature (Temp), systolic blood pressure (SBP), mean arterial pressure (MAP), diastolic blood pressure (DBP), respiration rate (Resp), and end-tidal carbon dioxide (EtCO2).
  - **Laboratory Values (26 variables)**: Includes parameters like blood urea nitrogen (BUN), creatinine, glucose, lactate, and white blood cell count (WBC).
  - **Demographic Features (6 variables)**: Age, gender, ICU admission type (MICU and SICU), hospital admission time, and ICU length of stay.
- **Sepsis Labels**: Binary labels indicating sepsis onset, defined six hours before clinical recognition of sepsis for septic patients based on Sepsis-3 criteria.

3.2 Summary Statistics

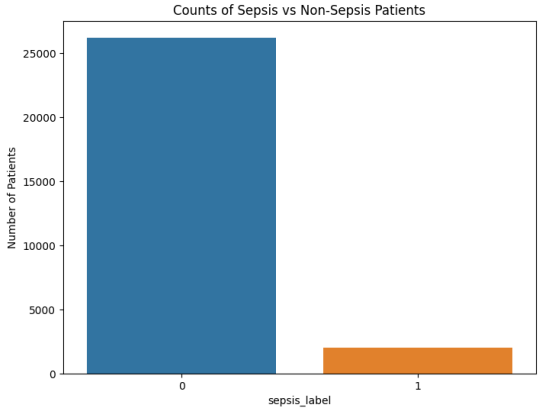
Key statistics for the dataset across Systems A and B are summarized in Table 2. We do not have access to the data from System C.el

Table 2: Summary statistics for the PhysioNet Challenge 2019 dataset.

Metric	System A	System B
Number of Patients	20,336	20,000
Number of Septic Patients	1,790	1,142
Sepsis Prevalence (%)	8.8	5.7
Total Hourly Time Windows	739,663	684,508
Total Data Points	5,536,849	4,950,064
Density of Non-Missing Data (%)	20.6	19.1



(a) Feature Counts



(b) Class Imbalance

Figure 1: Dataset Level Statistics

3.3 Data Preprocessing and Labeling

In Figure 1, we notice that not all features have the same count. Due to sparsity of certain columns, we only chose the top 24 features to represent the data. However, as described below, we actually take 3 values for each feature label using statistics and last value. We end up with **54 Features**.

Key steps in data preparation included:

(1) **Data Aggregation**: To increase the number of features, we take the mean, median and the last value of the time series data for all the patients. The last value is included since by observing the visualization, we notice that the final values for many features heavily affect the sepsis outcome.
(2) **Data Imputation**: Missing patient statistics (if a patient never received a measurement for one variable) were imputed from the total column mean. This was extremely uncommon.
(3) **Export to CSV**: Data was compiled into CSV files for future caching. For the train-val-test split, we use a 70-10-20 split which was a common method used in the challenge originally.

While most visualization can be seen through the code output, we have added some of the important observations here. In Figure 2, we can see that the end values for septic patients have very discernible changes in values. Low O2 stats and high body temperature are directly causal to a high chance of sepsis.

3.4 Challenges and Characteristics

The dataset poses several challenges:

- **Class Imbalance**: Septic patients constitute less than 10% of the population in Systems A and B.
- **Data Sparsity**: Laboratory values are recorded less frequently compared to vital signs, resulting in lower data density.
- **Cross-System Variability**: Differences in variable distributions across hospital systems make generalization challenging.

This dataset serves as a challenging benchmark for developing scalable, interpretable, and generalizable machine learning models for sepsis detection.

4 METHODOLOGY

In this section, we define the steps we take to create the models and perform experimentation for comparison of performance. An overview of our method is outlined in Figure 3.

4.1 Feature Aggregation and Assumptions

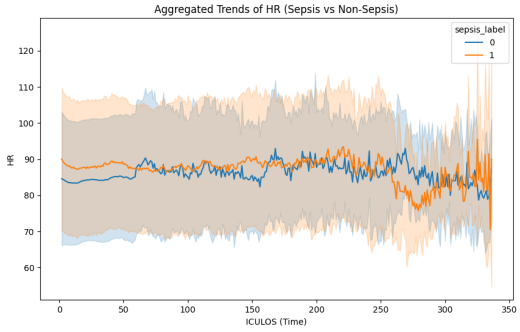
The **assumption** of MCAR simplifies the handling of missing data by asserting independence between missingness and the data itself:

$$P(X_i^{(1)} \mid R_1 = 1) = P(X_i \mid R_i = 1) = P(X_i^{(1)})$$

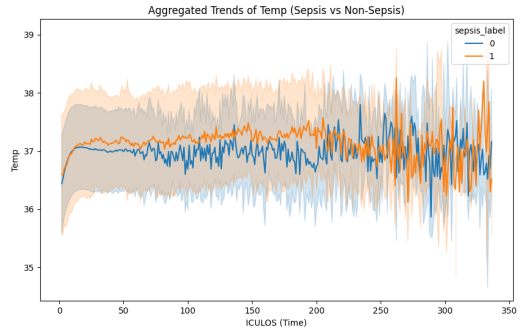
where R represents missingness, $X^{(1)}$ represents the true data, and X represents the observed data. Aggregating features to their mean provides a comprehensive representation of each variable's temporal behavior.

4.2 Feature Selection with Bayesian Networks

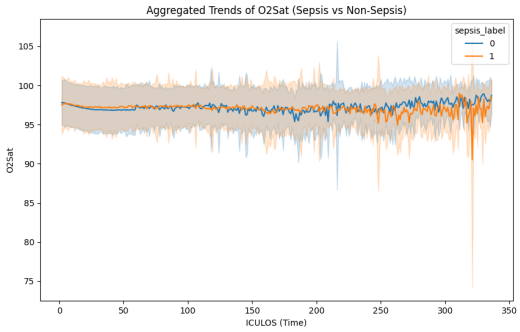
Bayesian Networks provide a probabilistic graphical framework to model dependencies between variables. These models represent joint probability distributions through directed acyclic graphs



(a) Aggregated Heart Rate



(b) Aggregated Temperature



(c) Aggregated O2 Stats

Figure 2: Aggregated values for different physiological parameters.

(DAGs), where nodes correspond to features, and edges denote conditional dependencies. Mathematically, the joint probability distribution for a set of features $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ in a Bayesian Network can be expressed as:

$$P(\mathbf{X}) = \prod_{i=1}^{n} P(X_i \mid \text{Pa}(X_i)),$$

where $\text{Pa}(X_i)$ represents the set of parent nodes for X_i in the graph. This factorization simplifies the modeling of complex joint distributions by leveraging conditional independence relationships.

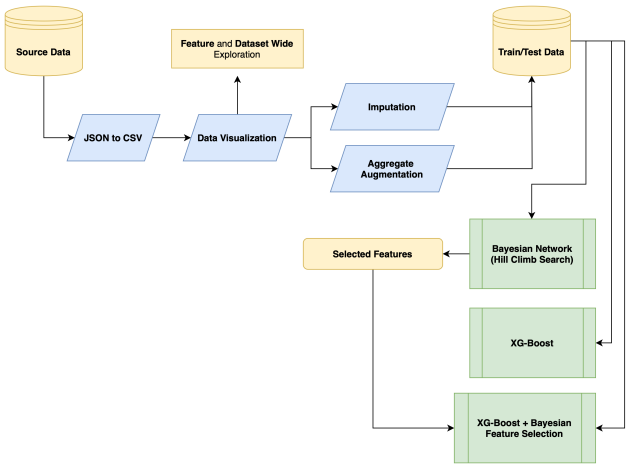


Figure 3: Overview of Methodology

To perform feature selection, we employ the `pgmpy` library, which facilitates structure learning for Bayesian Networks. We utilize the structure learning approach, which combines:

- **Score-based Learning**: Evaluates candidate graph structures based on a scoring function. In our case, we use the Bayesian Information Criterion (BIC), defined as:

$$\text{BIC} = \ln L - \frac{k}{2} \ln N,$$

  where L is the likelihood of the data given the graph structure, k is the number of parameters in the model, and N is the number of data points. Lower BIC values indicate better model complexity and fit.
- **Local Search Strategy**: Hill Climb Search, a greedy algorithm, iteratively modifies the graph by adding, removing, or reversing edges to maximize the scoring criterion.

Using this approach, we trained a Bayesian Network on the dataset, starting with 54 features. After learning the structure, the network retained only 48 features with significant relationships, as determined by the presence of edges in the graph. These selected features formed the reduced feature set used for prediction.

Figure 4 illustrates the learned Bayesian Network structure, highlighting the edges between features. From the graph, we can see some key relationships that the Bayesian Network learned:

- **Key Predictive Features**: The `sepsis_label` node directly connects to critical features such as `ICULOS`, `Temp_FINAL`, and `HR_FINAL`, emphasizing their importance in predicting sepsis.
- **Aggregated Metrics' Relevance**: Variables like `Resp_FINAL`, `BUN_FINAL`, and `DBP_FINAL` have many outgoing connections, showing that final values are strong predictors.
- **Clustered Relationships**: `SBP` and `WBC`, and their respective aggregated features (`WBC_FINAL`, `SBP_FINAL`), reveal conditional dependencies within these physiological measures. There are other clusters of the same form, as well.

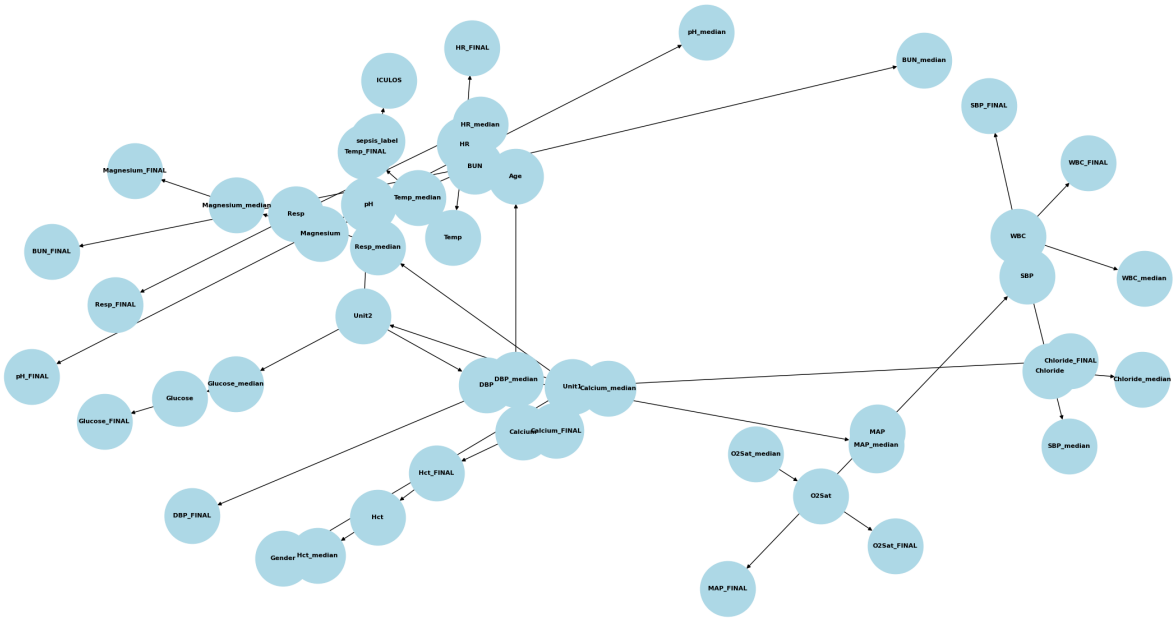


Figure 4: Graphical representation of the learned Bayesian Network structure.

4.3 Experiments

Finally, we choose to run 3 experiments with the two different datasets we end up with:

- Original dataset with 54 features.
- Bayesian dataset with 48 features.

We choose to establish baselines by running the trained Bayesian Graph over the test and validation data. We then use XGBoost, hyperparameter-tuned using 5 Fold CV over it's parameters, to find the best model on both datasets. For reference, the parameters selected for our XGBoost classifier are listed below:

```
Without Feature Selection:
{'learning_rate': 0.2, 'max_depth': 3,
'n_estimators': 300}

With Feature Selection:
{'learning_rate': 0.2, 'max_depth': 3,
'n_estimators': 100}
```

We can draw the conclusion that with less features, the XGB-Classifier did not need a high number of estimators to predict on the dataset.

5 RESULTS

We analyze our results on the validation and test data for the three models, choosing the Bayesian Model as our baseline.

The results presented in Table 4 highlight the trade-offs between interpretability and predictive performance across the models. The Bayesian Model, while interpretable, shows limited effectiveness in identifying sepsis label instances. This is reflected in its high validation accuracy (94.32%) but low recall (31.83%) and a test F1 score of 25.17%. The low recall suggests that the model struggles with sensitivity to the sepsis class, likely due to imbalance in the sepsis vs non-sepsis labels in the dataset.

XGBoost trained on the Bayesian-selected features demonstrates a significant improvement in both recall and F1 score, with a validation recall of 60.13% and a test F1 score of 63.91%. This indicates that the feature selection process effectively captures the most predictive features, reducing noise while preserving relevant information. The improved recall and F1 score suggest a better balance between precision and recall, making it a strong candidate for practical deployment where interpretability is also important.

Table 3: Class Wise Performance

Model	Class	Precision	Recall
All Features	0 (No Sepsis)	0.97	0.98
	1 (Sepsis)	0.73	0.59
Bayesian	0 (No Sepsis)	0.96	0.99
	1 (Sepsis)	0.79	0.53

The XGBoost model trained on all features achieves the best overall performance, with an overall accuracy of **95.48**% and the highest test F1 score of 65.00%. This demonstrates that retaining the full feature set allows the model to capture additional relationships that may be missed during feature selection. However, this comes at the expense of interpretability and computational efficiency. These results underscore the importance of considering the specific trade-offs between interpretability and performance when designing predictive models for healthcare applications.

Table 4: Performance metrics for the models on validation and test datasets.

Model	Val Acc	Val Recall	Test Acc	Test Precision	Test Recall	Test F1
Bayesian Model	94.32	31.83	93.29	62.76	15.74	25.17
XGBoost (Bayesian)	95.66	60.13	**95.67**	**79.43**	53.46	63.91
XGBoost (All)	95.64	64.31	95.48	72.90	**58.65**	**65.00**

We also present the class wise precision and recall for the two XGBoost Models to represent the class level performance which is particularly important in this case. A False Negative in this case is life or death. The results in Table 3 show that both models achieve high precision and recall for the non-sepsis class (Class 0). However, for the sepsis class (Class 1), the Bayesian-selected model demonstrates slightly higher precision (0.79), while the all-features model achieves better recall (0.59), indicating there is a trade-off between reducing false positives and false negatives.

6 CONCLUSION

This study demonstrates the effectiveness of combining interpretable graphical models with machine learning techniques for sepsis prediction. By leveraging Bayesian Networks for feature selection and XGBoost for predictive modeling, we achieved a balance between interpretability and predictive accuracy. Our results show that the full-feature XGBoost model achieves the highest overall performance, with a test F1 score of 65.00%, while the Bayesian-selected features provide a strong trade-off for interpretability with a slightly lower F1 score of 63.91%.

The social impact of this research is profound. Sepsis remains one of the leading causes of preventable deaths globally, with significant disparities in outcomes between low- and high-resource settings. An interpretable and scalable predictive framework, like the one proposed here, can enable early sepsis detection in ICUs, improving patient outcomes and reducing mortality. With further work into interpretable models, healthcare systems can provide timely interventions, ultimately saving lives and reducing the long-term burden of sepsis on patients and healthcare infrastructure.

REFERENCES

[1] Kristina E Rudd, Sarah Charlotte Johnson, Kareha M Agesa, Katya A Shackelford, Derrick Tsoi, Daniel R Kievlan, Danny V Colombara, Kevin S Ikuta, Niranjan Kissoon, Simon Finfer, et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the global burden of disease study. *The Lancet*, 395(10219):200–211, 2020.

[2] CIDRAP. Who says sepsis causes 20% of global deaths, 2020.

[3] World Health Organization. Sepsis, 2024.

[4] The Lancet. Sepsis: a roadmap for future research. *The Lancet*, 395(10219):170, 2020.

[5] Lucas M Fleuren, Thomas LT Klausch, Charlotte L Zwager, Linda J Schoonmade, Tingjie Guo, Luca F Roggeveen, Eleonora L Swart, Armand RJ Girbes, Patrick Thoral, Ari Ercole, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive care medicine*, 46(3):383–400, 2020.

[6] Supreeth P Shashikumar, Gabriel Wardi, Atul Malhotra, and Shamim Nemati. Impact of a deep learning sepsis prediction model on quality of care and patient outcomes: a before-and-after study. *npj Digital Medicine*, 6(1):1–9, 2023.

[7] Yuying Lu, Jiaqi Xu, Yifan Gao, Yuting Guo, Yiqun Gu, and Jie Xu. Predicting sepsis mortality using machine learning methods. *medRxiv*, 2024.

[8] Hyeon-Eui Kam and Eun-Jung Kim. Learning interpretable physiological models of sepsis using sparse autoencoders. *Artificial Intelligence in Medicine*, 82:19–28, 2017.

[9] Yufei Wang, Zhiyong Li, and Hong Zhang. Logistic regression-based early warning system for sepsis onset prediction in icu patients. *Critical Care Medicine*, 47(3):e234–e240, 2019.

[10] Michael Moor, Maxime Horn, Bastian Rieck, Damian Roqueiro, and Karsten Borgwardt. Deep learning for early detection of sepsis: A temporal validation study. *Journal of Biomedical Informatics*, 105:103411, 2020.

[11] Yuhang Zhang and Xiaoming Liu. Early sepsis prediction using ensemble learning with deep features extracted from ehr data. *Journal of Medical Systems*, 44(12):1–12, 2020.

[12] Matthew A Reyna, Chris S Josef, Russell Jeter, Supreeth P Shashikumar, Benjamin Moody, M Brandon Westover, Ashish Sharma, Shamim Nemati, and Gari D Clifford. Early prediction of sepsis from clinical data: The physionet/computing in cardiology challenge 2019. In *Computing in Cardiology Conference (CinC)*, pages 1–4, 2019.

[13] Ibrahim Hammoud, Ramakrishnan IV, and Henry Mark. Gradient boosting decision trees for early prediction of sepsis using physionet challenge data. *Computing in Cardiology Conference (CinC)*, 46:1–4, 2019.

[14] David et al. Gomez-Cabrero. Probabilistic graphical models for pediatric sepsis risk prediction using ehr data. *PLOS ONE*, 14(5):e0217416, 2019.

[15] Masino et al. Using ehr data to predict infant sepsis at least four hours before clinical recognition: A machine learning approach. *Journal of Biomedical Informatics*, 93:103160, 2019.

[16] Shashikumar et al. Artificial intelligence in sepsis early prediction using structured data and clinical notes: The sera algorithm. *Nature Communications*, 12:1–10, 2021.

Appendix: Table of Features

No.	Defined Variable	Description
1	HR	Heart rate (beats per minute)
2	O2Sat	Pulse oximetry (%)
3	Temp	Temperature (°C)
4	SBP	Systolic BP (mm Hg)
5	MAP	Mean arterial pressure (mm Hg)
6	DBP	Diastolic BP (mm Hg)
7	Resp	Respiration rate (breaths per minute)
8	EtCO2	End tidal carbon dioxide (mm Hg)
9	BaseExcess	Excess bicarbonate (mmol/L)
10	HCO3	Bicarbonate (mmol/L)
11	FiO2	Fraction of inspired oxygen (%)
12	pH	pH
13	PaCO2	Partial pressure of carbon dioxide from arterial blood (mm Hg)
14	SaO2	Oxygen saturation from arterial blood (%)
15	AST	Aspartate transaminase (IU/L)
16	BUN	Blood urea nitrogen (mg/dL)
17	Alkalinephos	Alkaline phosphatase (IU/L)
18	Calcium	Calcium (mg/dL)
19	Chloride	Chloride (mmol/L)
20	Creatinine	Creatinine (mg/dL)
21	Bilirubin_direct	Direct bilirubin (mg/dL)
22	Glucose	Serum glucose (mg/dL)
23	Lactate	Lactic acid (mg/dL)
24	Magnesium	Magnesium (mmol/dL)
25	Phosphate	Phosphate (mg/dL)
26	Potassium	Potassium (mmol/L)
27	Bilirubin_total	Total bilirubin (mg/dL)
28	TroponinI	Troponin I (ng/mL)
29	Hct	Hematocrit (%)
30	Hgb	Hemoglobin (g/dL)
31	PTT	Partial thromboplastin time (seconds)
32	WBC	Leukocyte count (count/L)
33	Fibrinogen	Fibrinogen concentration (mg/dL)
34	Platelets	Platelet count (count/mL)
35	Age	Age (years)
36	Gender	Female (0) or male (1)
37	Unit1	ICU unit (MICU; false (0) or true (1))
38	Unit2	ICU unit (SICU; false (0) or true (1))
39	HospAdmTime	Time between hospital and ICU admission (hours)
40	ICULOS	ICU length of stay (hours since ICU admission)
41	SepsisLabel	Presence of sepsis (false (0) or true (1))