Bridging the Sepsis Prediction Gap: Explainable Deep Learning Models for Early Detection Using Time Series XAI Techniques

Abstract

Sepsis is one of the most deadly illnesses with a high risk of mortality. Consequently, identifying it at the beginning of illness symptoms is crucial and plays a key role in improving patient outcomes. This study presents a customized solution for the early detection of sepsis with an emphasis on the use of interpretability and explainability techniques, utilizing a range of machine learning approaches and interpretable artificial intelligence methods. The database on which this research study is based has many problems; the main ones being large data gaps and class disparities. Employing robust methods, precise categorizations, and rigorous computations, Approximately 12 diverse models were developed and optimized. With ROC-AUC indicators of 0.9566 and 0.9595 and F1 scores of 0.85 and 0.85 respectively, Bidirectional Long Short-Term Memory (BiLSTM) and Temporal Convolutional Network (TCN) models performed better than conventional models in terms of sepsis prediction. These two approaches have shown remarkable progress in detecting clinical patterns while avoiding false negative resultsan essential aspect of the medical field. To assess model performance and offer clear insights into model predictions, interpretation-based techniques were employed. This improved clinical confidence and facilitated well-informed decisions in crucial medical diagnoses.

1 Introduction

SEPSIS is a significant global health concern, causing approximately 11 million deaths annually and accounting for about 20% of all global mortality [1]. When a body encounters an infection, the immune system releases cytokines and other inflammatory mediators. While this response aims to eliminate pathogens, excessive or uncontrolled activation can cause tissue damage, multiple organ failure, and even death [1]. There is no single definitive test to diagnose sepsis [2]. Clinicians depend on a combination of clinical signs, laboratory tests, and biomarkers to evaluate whether a patient is septic or non-septic [3]. Early detection of these symptoms enables preventative measures to reduce mortality [4].

Despite the urgent need for early detection, current diagnostic approaches are restricted by complexity and variability inherent in sepsis progression [5]. Traditional methods may not capture changing patterns in patient data that can signal sepsis onset. As highlighted in [2], sepsis presents a diagnostic challenge because it is not directly recognizable. These challenges are compounded by substantial data quality and reliability issues. For instance, a widely used dataset is the PhysioNet 2019 Challenge dataset, which has more than 90% missing values and demonstrates severe class imbalance [3].

Many studies have investigated various Machine Learning (ML) and Deep Learning (DL) models for early detection of sepsis, but several gaps persist. A notable limitation in several studies, including those by Lyra et al. [6], Wu et al. [7], and Liao et al. [8], is the absence of model interpretability. Algorithms like Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Extreme Gradient Boosting (XGBoost) achieved moderate success in sepsis prediction but did not implement interpretable techniques such as Shapley Additive exPlanations (SHAP) or Permutation feature importance (PFI) to clarify their decision-making processes. This lack of transparency limits interpretability and impedes model improvement.

Although some research utilized feature selection techniques, few studies have compared the effects of different selection methods on model performance. Furthermore, management of missing data and class imbalance is inconsistent, with limited assessments of imputation methods or balancing strategies. This problem is especially apparent in the PhysioNet 2019 dataset, where about 90% of records belong to the non-septic class, and over 90% of values are missing for many laboratory variables due to the intermittent nature of lab testing rather than regular hourly sampling [3].

This study directly addresses the critical need for reliable early detection of sepsis, a task often hindered by highly incomplete clinical data and severe class imbalance. To overcome these challenges and improve the credibility of predictive models in clinical practice, an integrated framework is proposed that combines advanced machine learning with Explainable AI (XAI) techniques.

The key contributions of this research are summarized as follows:

1. Development of a robust data preprocessing pipeline specifically designed to address extreme missingness and class imbalance inherent in sepsis datasets, ensuring the generation of more balanced and informative training data.
2. Enhancement of model interpretability for early sepsis prediction through the application of diverse XAI methods, including SHAP and Permutation Feature Importance, offering comprehensive insights into the decision-making processes of machine learning models.
3. Provision of clinically meaningful explanations that not only identify the most influential features driving predictions but also pinpoint the temporal significance of these features, thereby fostering clinical confidence and enabling more informed decision-making in critical care environments.

The rest of this paper is organized as follows: Section 2 reviews related work on sepsis prediction and machine learning in healthcare. Section 3 outlines the methodology, in-

cluding data preprocessing, model development, and explainability techniques. Section 4 presents and discusses the experimental results. Finally, Section 5 summarizes the findings and suggests directions for future research and clinical application.

2 Literature Review

To contextualize this research, twelve key studies applying machine learning (ML) and deep learning (DL) techniques for early sepsis prediction have been reviewed. These studies were specifically selected based on their relevance to the progression of sepsis prediction methods, their methodological diversity across traditional ML and DL, and their focus on addressing critical challenges such as data imbalance, missingness, and temporal modeling. This curated selection ensures a comprehensive understanding of the current landscape and its limitations.

Traditional ML methods have been extensively utilized for sepsis prediction. Lyra et al. (2019) [6] employed Random Forest (RF) classification to address clinical data imbalance, achieving a mean Area Under the Receiver Operating Characteristic Curve (AUROC) of 0.81. Wu et al. (2019) [7] proposed a customized down-sampling approach combined with a dynamic sliding window, resulting in an AUC of 0.89. Similarly, Nirgudkar and Ding (2019) [9] applied imputation strategies alongside a weak ensembler technique, obtaining an internal validation accuracy of 93.45%. Kong et al. (2019) [11] developed Gradient Boosting Machine (GBM) and RF models, reporting superior performance for GBM at 84.5% accuracy, compared to other models ranging between 77% and 83%. More recently, Vandana and Chhikara (2024) [10] evaluated various ML models with feature selection techniques for ICU sepsis prediction, achieving a peak accuracy of 95% using a Decision Tree model combined with Information Gain.

The adoption of DL methods has expanded significantly, particularly due to their effectiveness in capturing temporal dependencies within clinical data. Lipton et al. (2016) [12] addressed the challenge of missing data by employing Recurrent Neural Networks (RNNs) to impute missing values in clinical time series. Liu et al. (2019) [13] introduced a Heterogeneous Event Aggregation (HEA) framework enhanced with LSTM to model complex temporal interactions within Electronic Health Records (EHRs), achieving an AUC of 0.8224 with their 16-head HEA-LSTM model. Apalak and Kiasaleh (2020) [14] utilized Temporal Convolutional Networks (TCNs) on the MIMIC-III dataset, focusing on Heart Rate Variability (HRV) features, and reported an AUROC of 0.82. Oei et al. (2021) [15] applied a BiLSTM model to the same dataset, reaching an AUROC of 0.85. Liao et al. (2022) [8] conducted a comparative study on the PhysioNet 2019 dataset, benchmarking DL models such as InceptionTime and TCN against traditional ML methods like RF and XGBoost. Their results demonstrated that DL models outperformed traditional approaches, with TCN and InceptionTime achieving AUROC scores of 0.92 and 0.90, respectively.

Despite these advancements, several critical gaps persist within the existing literature. One prominent limitation is the lack of model interpretability; many studies have not incorporated explainability tools such as SHapley Additive exPlanations (SHAP) or Permutation feature importance (PFI), which are essential for building trust in clinical settings. Additionally, there remains an inconsistent treatment of missing data and class imbalance, which hinders model reliability and generalizability. The limited comparative evaluation of feature selection methods and the underutilization of ensemble techniques further restrict the robustness and optimization of predictive models.

3 Methodology

3.1 Research Gap & Proposed Approach

A detailed review of existing work highlights three persistent limitations that reduce clinical viability and trustworthiness of current ML-based systems for sepsis prediction.

First, many models lack interpretability, which poses a serious challenge in medical settings. Although models including XGBoost, LSTM, and GRU have achieved strong predictive performance in studies by Lyra et al. [6], Wu et al. [7], and Liao et al. [8], they generally do not integrate XAI tools. As a result, these models function as black boxes, limiting clinicians ability to understand or validate their outputs. This lack of transparency undermines clinical trust and makes it difficult to meet regulatory standards for artificial intelligence (AI) in healthcare.

Second, the issue of data availability and quality remains largely unresolved. Specifically, this includes challenges related to missing data, noisy measurements, and incomplete clinical records which are common in real-world healthcare datasets. In the PhysioNet 2019 dataset [3], over 90% of values for some lab features are missing, and non-septic patients account for an overwhelming majority of cases. Such gaps in data can impair model learning, increase bias, and diminish predictive reliability. Prior studies apply different preprocessing methods for imputation and resampling [12, 23], but very few offer comparative analyses to determine which strategies are most effective. This inconsistency raises concerns about the reproducibility and reliability of the resulting models.

Third, while feature selection is frequently mentioned [10], most existing work does not rigorously compare multiple selection methods. Without evaluating different strategies side by side, opportunities to improve both model performance and interpretability may be missed. An optimized feature set can simplify models while retaining essential clinical insight.

To tackle the challenges of early sepsis detection and the inherent complexities of clinical data, the proposed framework combines advanced temporal modeling with a comprehensive suite of explainability techniques. Specifically, it incorporates a range of Explainable AI (XAI) methods, including SHAP, PFI, Permutation Feature Importance (PFI), Integrated Gradients, and Accumulated Local Effects (ALE). This approach provides detailed, multi-dimensional insights into the reasoning behind model predictions. This framework, which leverages XAI alongside machine learning, is designed to transform complex neural network de-

cisions into clear, trustworthy insights that clinicians can readily interpret and act upon.

Methodological approach emphasizes evaluation of BiLSTM and TCN models due to their superior capacity for capturing temporal dependencies in sequential clinical dataa critical requirement for early sepsis detection, where temporal evolution of patient parameters contains essential predictive information [14, 21]. By combining these advanced temporal modeling capabilities with comprehensive explainability tools, this study aims to bridge the gap between computational sophistication and clinical practicality, ultimately advancing the development of trustworthy AI systems for critical medical diagnostics.

3.2 Data Preparation

3.2.1 Used Dataset

This study employs the PhysioNet 2019 Challenge dataset [3] as the primary source of multivariate time series data for sepsis prediction. The dataset includes comprehensive clinical records of ICU patients with rich temporal features crucial for sepsis detection. Focusing on the PhysioNet 2019 dataset because it is widely regarded as the state-of-the-art benchmark for sepsis prediction tasks. The dataset is extensively used in the research community due to its richness, diversity of clinical variables, which fosters standardized evaluation across studies. Leveraging such a comprehensive dataset ensures that our findings are comparable to existing work and applicable to real-world clinical scenarios.

3.2.2 Data Preprocessing & Handling Imbalance

Data preprocessing involved handling missing values using advanced imputation techniques such as k-nearest neighbors (KNN) imputation and forward filling methods. To address the severe class imbalance, SMOTE (Synthetic Minority Over-sampling Technique) and class weighting strategies were employed. These methods ensure a balanced representation of septic and non-septic cases, enhancing the robustness and generalizability of predictive models.

3.3 Predictive Models

In this research, several DL and ML models are implemented to identify the most effective approach for early sepsis prediction using multivariate clinical time series data. The selection of models is driven by their ability to capture complex temporal relationships, represent diverse clinical variables, and maintain computational efficiency and robustness against data irregularities. This section provides a detailed overview of the models explored in this study, categorized into traditional ML models and DL time series models.

3.3.1 Traditional Machine Learning Models

RF is an ensemble method that aggregates the predictions from a multitude of decision trees trained on bootstrapped samples of the training set [11, 16]. Each tree is grown based on a randomly selected subset of features to reduce inter-tree correlation. In this study, the maximum depth was controlled and Gini impurity was employed as the splitting criterion. The model exhibited high interpretability and robustness against overfitting, making it particularly suitable for clinical tabular data. RFE was further applied to enhance feature selection performance [11, 16], with the complete configuration summarized in Table 1.

XGBoost, a gradient boosting framework, was also employed. This method builds models sequentially while applying regularization to mitigate overfitting. In the present implementation, a binary logistic objective was used and hyperparameters were optimized through Bayesian search. Generalization was further supported via early stopping across five-fold cross-validation. The final model demonstrated high predictive performance, particularly in imbalanced settings [17–19]. The configuration details are outlined in Table 1.

3.3.2 Traditional Machine Learning Models

Table 1: Model Configurations

Model	Key Parameters
RF	n_est=100, bootstrap, Gini min_split=10, min_leaf=5 max_depth=None
XGBoost	depth=7, lr=0.124 subsample=0.97, colsample=0.84 gamma=0.50, min_child=8 alpha=5.74, lambda=2.45 n_est=1000
Logistic Reg.	L-BFGS, max_iter=1000 Standard scaling, 80-20 split

In addition, RF was utilized with 100 estimators, optimized splitting thresholds, and restricted tree depths to minimize variance. The model used characteristic importance scores to identify clinically influential predictors of sepsis [6]. The set of engineered features used in these models is detailed in Table 2.

Table 2: Engineered Features

Feature	Formula
Pulse Pressure	$SBP - DBP$
Cardiac Output	$(SBP - DBP) \times HR$
Shock Index	HR/SBP
Modified SI	HR/MAP
Z-scores	SBP, DBP, HR, MAP

3.3.3 Deep Learning Time Series Models

Several DL architectures were evaluated to capture the temporal dynamics of multivariate clinical time series data. These models are grouped into recurrent-based and convolution-based frameworks.

Recurrent Models incorporate hidden states and temporal feedback mechanisms. A baseline RNN with 64 hidden units, global average pooling, and a sigmoid output layer was developed, employing dropout regularization to reduce overfitting [12]. LSTM networks were constructed by stacking LSTM layers with dropout and global average pooling, optimized using the Adam optimizer [12]. BiLSTM was then introduced to process sequences in both forward and backward directions, thereby enhancing context modeling. This architecture included stacked BiLSTM layers, batch normalization, and dropout regularization [20]. Finally, GRU networks were employed for their parameter efficiency, utilizing simplified gating structures. GRUs incorporated L2 regularization, dropout, and early stopping strategies [11, 16]. The architectural details and regularization strategies of these recurrent models are provided in Tables 3 and 4, respectively.

Table 3: Recurrent Models: Architecture

Model	Architecture Description
RNN	64 hidden units, global average pooling, dense sigmoid output
LSTM	Stacked LSTM layers with dropout, global average pooling
BiLSTM	Two BiLSTM layers, batch normalization, dropout, global average pooling
GRU	GRU layers with simplified gating, global average pooling

Table 4: Recurrent Models: Regularization Techniques

Model	Regularization and Optimization
RNN	Dropout regularization to prevent overfitting
LSTM	Dropout, Adam optimizer
BiLSTM	Dropout, Batch normalization
GRU	L2 regularization, Dropout, Early stopping

Convolutional Models employ convolutional operations to extract local and hierarchical temporal patterns. TCNs utilize dilated causal convolutions with expanding receptive fields and global max pooling to capture long-range dependencies efficiently [7, 14]. The architectural details of these convolutional models are summarized in Table 5.

InceptionTime networks apply multiple convolution kernels of different sizes in parallel (inception modules), combined with dropout and global average pooling for regularization [8, 21]. 1D-CNNs use conventional convolution and pooling layers to extract local features, offering computational efficiency albeit with limited long-range modeling capacity [7, 9]. Finally, ResNet-1D incorporates residual skip connections into 1D convolutional layers to facilitate deeper temporal representations and improved gradient flow [15, 20]. Table 5 summarizes the architectural details of these convolutional models.

Table 5: Convolutional Models: Detailed Architecture

Model	Architecture Details
TCN	**Conv1D Layers:** 3 layers with filters [128, 256, 512], dilation rates [1, 2, 4], ReLU activation **Regularization:** Dropout (0.3) after each Conv1D layer **Output:** Dense layer (1 unit, sigmoid activation) **Optimizer:** Adam, Loss: Binary cross-entropy
InceptionTime	**Inception Module:** 3 parallel Conv1D paths with kernel sizes [10, 20, 40] **Regularization:** Dropout layer after each parallel path **Output:** Dense layer (1 unit, sigmoid activation) **Optimizer:** Adam, Loss: Binary cross-entropy
1D-CNN	**Conv1D Layers:** 2 layers - Layer 1: 64 filters (kernel=3), Layer 2: 128 filters (kernel=3), ReLU activation **Dense Layers:** Hidden layer (128 units, ReLU), Output layer (1 unit, sigmoid) **Optimizer:** Adam, Loss: Binary cross-entropy
ResNet-1D	**Residual Blocks:** 2 blocks - Block 1: 64 filters, Block 2: 128 filters, ReLU activation **Regularization:** Batch normalization within blocks, Dropout (0.3) **Output:** Dense layer (1 unit, sigmoid activation) **Optimizer:** Adam, Loss: Binary cross-entropy

3.4 Data Overview and Preprocessing

The training dataset consisted of 1.5 million ICU observations from over 40,000 patients [3]. The class distribution per patient, as shown in Figure 1, highlights the proportion of septic and non-septic cases.

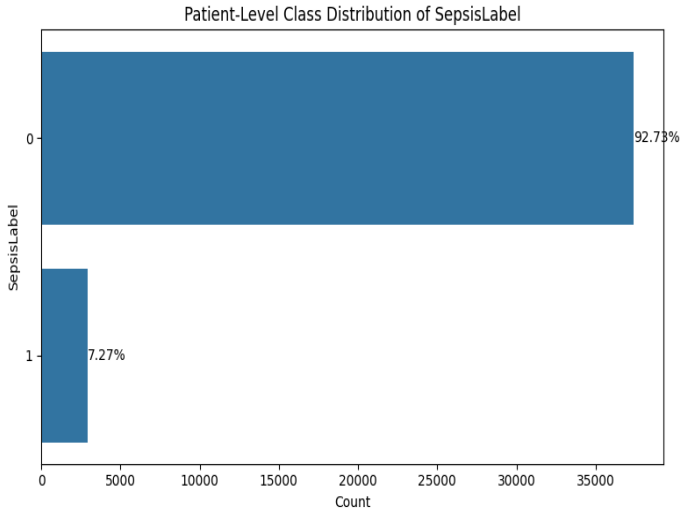

Figure 1: Distribution of classes per patient

After applying exploratory data analysis (EDA), the following key findings were observed:

- **Severe Class Imbalance:** Only ∼7% of patients were labeled as septic (`SepsisLabel = 1`), with the majority being non-septic. This imbalance is clearly visualized in Figure 1.
- **Patient-Level Distribution [3]:**
  - Sepsis patients: 2,805
  - Non-sepsis patients: 37,087
- **High Percentage of Nulls Across Most Features:** About 25 features have more than 89% nulls. Over

10 lab-related features had >95% missingness (e.g., `Bilirubin_direct`, `Fibrinogen`, `TroponinI`). This high level of missingness is depicted in Figure 2, which shows the distribution of null percentages across the dataset's features [3, 12].

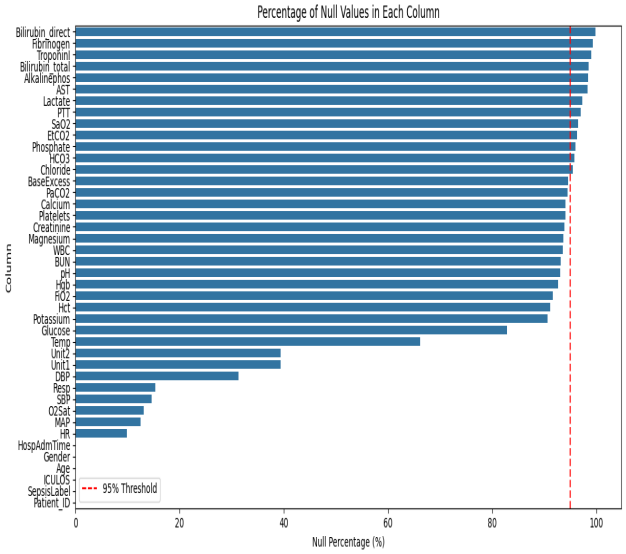

Figure 2: Distribution of null percentage per column

A major issue is that each patient does not have a fixed length of stay, as it ranges from 8 to 336 hours, with an average of 38.48 hours after filtering. This variability is problematic because it prevents trained models from having dynamic input length, which is a key consideration in ICU patient monitoring. While no specific figure illustrates this variability, it is a critical factor affecting the dataset's structure as analyzed alongside Figure 1 and Figure 2 [3, 20].

3.4.1 Handling Missing Data

Figure 3 illustrates the complete pipeline for handling missing data in the dataset, where features are systematically divided into three distinct groups based on their missingness percentage. Each group is then subjected to a tailored imputation or handling strategy, justified by the nature of the data and clinical considerations.

These groups are:

1. **Features with >95% Missingness: Informative Missingness and Binary Flags**
   This category includes features such as `Bilirubin_direct`, `Fibrinogen`, and `Lactate`, which exhibit a very high percentage of missing values. Instead of outright exclusion, these features were handled by dropping their original values from direct use in modeling but retaining their implicit information. Specifically, binary flags were introduced to indicate whether the corresponding test was performed or not. This approach is rooted in the concept of "informative missingness," which posits that the absence or presence of certain clinical measurements can itself be a valuable piece of information for predictive models [12, 22]. In an ICU setting, high missingness in lab values often does not signify data corruption but rather reflects deliberate clinical decision-making. Doctors typically order these specific lab tests only when there are particular clinical concerns or symptoms, making their infrequent appearance highly significant [12]. For instance, a `bilirubin` test might only be ordered if liver dysfunction is suspected, and its presence (or absence) provides diagnostic insight. This allows models to learn from the context of why a test was (or was not) performed, rather than solely from its measured value.

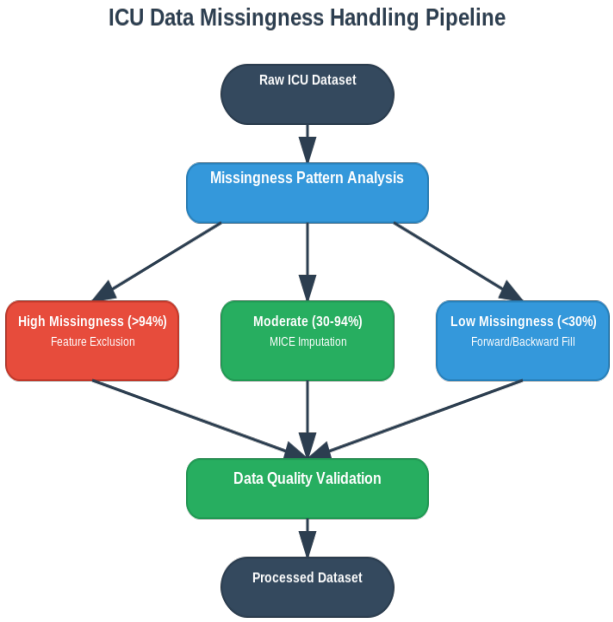

Figure 3: Nulls handling pipeline.

2. **Features with 35–94% Missingness: Multiple Imputation by Chained Equations (MICE)**
   Features falling within this intermediate range of missingness, such as `Glucose`, Blood Urea Nitrogen (BUN), and White Blood Cell (WBC), were retained for imputation. These lab tests are generally not measured as frequently as vital signs, and their values are assumed to have less time-sensitive variability, making them suitable for more sophisticated imputation techniques [20]. The Multiple Imputation by Chained Equations (MICE) algorithm [23] was employed for this purpose. MICE works by building a separate predictive model for each feature with missing values, using all other features in the dataset as predictors. This iterative process generates multiple complete datasets, leading to more robust and realistic imputations compared to single imputation methods.To account for clinical data uncertainty and prevent overfitting, Gaussian noise was added during the imputation process.
   A specific challenge arose with Diastolic Blood Pressure (DBP), which remained entirely missing for 7,411 patients even after MICE imputation. This scenario typically occurs when MICE cannot impute values due to complete missingness within a patient's record. To address this, a physiological imputation strategy was implemented: forward and backward filling per patient. Furthermore, if DBP was still missing but Systolic Blood Pressure (SBP) and Mean Arterial Pressure (MAP) were available, the following clinically derived

formula was applied to estimate DBP [24]:

$$\text{DBP} = \frac{3 \cdot \text{MAP} - \text{SBP}}{2} \tag{1}$$

3. **Features with <35% Missingness: Forward and Backward Filling**
   This group primarily comprises vital signs, such as Heart Rate (HR) and Oxygen Saturation (O2Sat), which are frequently and regularly measured in the ICU. Given the clinical assumption that vital signs do not fluctuate significantly over short time intervals [25], missing values in this category were imputed using a straightforward forward and backward filling approach per patient. This method effectively propagates the last known valid observation forward and the next known valid observation backward to fill gaps, reflecting the relative stability of these measurements within short periods.

3.4.2 Data Balancing Strategy

The dataset was initially organized at the patient level by combining hourly clinical values for each person to efficiently model sepsis predictions, as illustrated in Figure 4. This organization is crucial for capturing the temporal dynamics associated with sepsis.

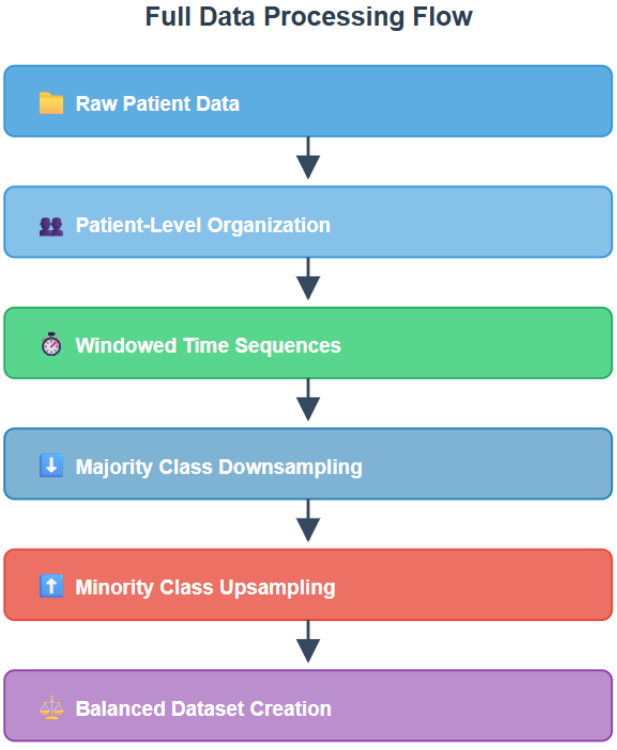

Figure 4: Pipeline for balancing class distribution in patient data.

To address the challenge of class imbalance while preserving essential temporal information, a window-based sampling approach was implemented, as shown in Figure 5. Each patient's timeline was divided into fixed-length overlapping intervals. If any time step within a window indicated a positive sepsis diagnosis, that entire window was classified as sepsis-positive. This method ensures consistency across patients, regardless of their hospital stay duration.

Due to the inherent class imbalance in the datasetwhere non-sepsis instances significantly outnumber sepsis cases, a customized two-stage sampling framework was implemented. This strategy ensures that the classifier receives a balanced and informative representation of both classes during training. The full pipeline is outlined in Figure 5 and comprises the following steps:

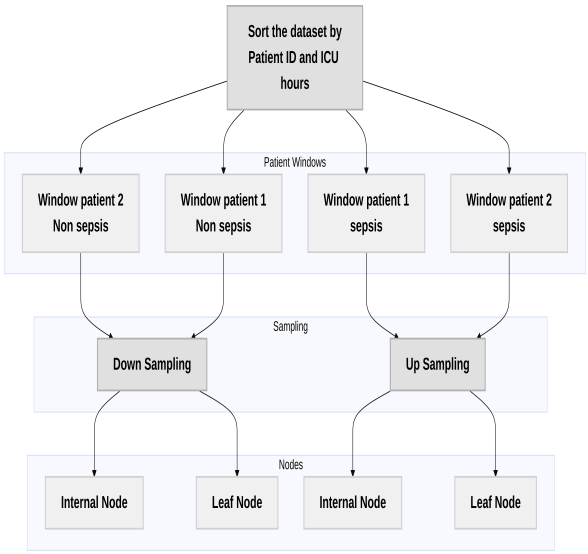

Figure 5: Window-based sampling of patient data to capture temporal dynamics.

1. **Decision Tree-Based Undersampling:** To mitigate the dominance of the majority class (non-sepsis), a targeted undersampling strategy was employed. Initially, a random subset (maximum of 50,000) of majority-class samples was flattened and labeled across all time steps. A decision tree classifier (maximum depth of 5) was then trained to learn key discriminative features. Using this model, a representative sample of 35,000 sequences was drawn, preserving feature diversity and temporal context. This intelligent reduction decreased training time while maintaining clinical relevance.
2. **Decision Tree-Guided Synthetic Oversampling:** To increase representation of the minority class (sepsis), biologically-informed synthetic sequences were generated. Specifically, the previously trained decision tree's feature importance scores guided the perturbation process: synthetic samples were created by injecting Gaussian noise into time-aligned features, with noise magnitudes weighted by their relative importance (only features with importance > 0.01 were modified). This process expanded the sepsis class to 27,000 sequences, enhancing the models sensitivity to rare but critical patterns indicative of sepsis onset.
3. **Stratified Train-Test Splitting:** The resulting balanced dataset was partitioned into training and testing subsets. A total of 28,000 non-sepsis and 21,600 sepsis sequences were allocated to the training set, while the remaining samples were reserved for evaluation. This stratified split ensures unbiased performance assessment and preserves the effects of the prior balancing steps.

This hybrid sampling strategycombining intelligent data reduction and biologically-aware augmentationwas essential for training a robust and generalizable model under realistic

clinical data constraints.

This controlled approach to class balancing minimizes the risk of overfitting to the majority class while ensuring the model learns meaningful patterns from both sepsis and non-sepsis examples, as demonstrated in Figure 6.

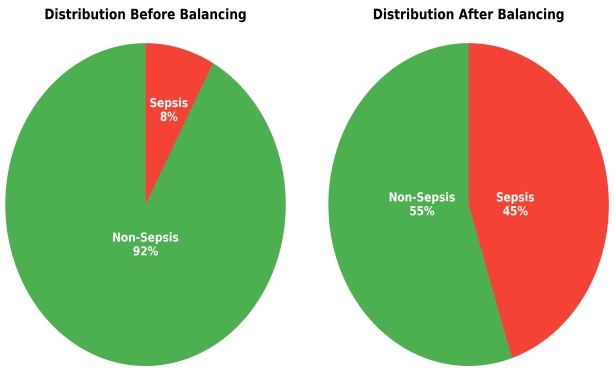

Figure 6: Class distribution before and after applying the two-stage balancing strategy.

The dataset was customized to control the number of samples from each class, reducing bias caused by class imbalance. This preprocessing step ensures more balanced training and testing sets, thereby promoting fairer model evaluation and improved generalization performance across both classes, as shown in Figure 7.

Moreover, this class balancing strategy enhances the models sensitivity to minority class instances, which is essential for the timely and accurate detection of sepsis. It also enables a more reliable assessment of the models clinical effectiveness under realistic deployment scenarios, where early identification of rare cases can be critical.

Figure 7: Class distribution after applying the balancing strategy.

This approach further minimizes the risk of overfitting to the majority classa common challenge in medical datasets with a low prevalence of positive outcomes. By enforcing controlled class proportions, the model is better equipped to learn meaningful and generalizable patterns from both sepsis and non-sepsis examples.

The distribution of the temperature (Temp) feature before and after preprocessing is shown in Figure 8. The gray line represents the original distribution, while the blue line depicts the distribution after applying imputation and normalization. As seen, both lines follow a nearly identical trajectory, with peaks occurring at approximately the same value (37.5řC). This overlap confirms that the preprocessing preserved the essential characteristics of the original temperature data without introducing significant distortion.

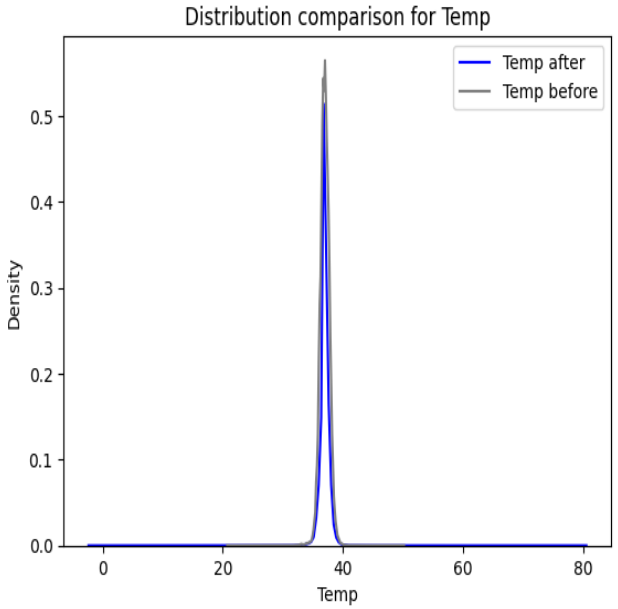

Figure 8: Histogram of Temperature (Temp) feature distribution before (gray) and after (blue) preprocessing.

Similarly, Figure 9 illustrates the distribution of the Diastolic Blood Pressure (DBP) feature before and after preprocessing. The alignment between the gray and blue curves is again prominent, particularly around the central region (75 mmHg). This consistency suggests that the imputation and scaling steps applied to the DBP feature successfully retained the original distribution's structure, thus preserving the clinical integrity of the measurements.

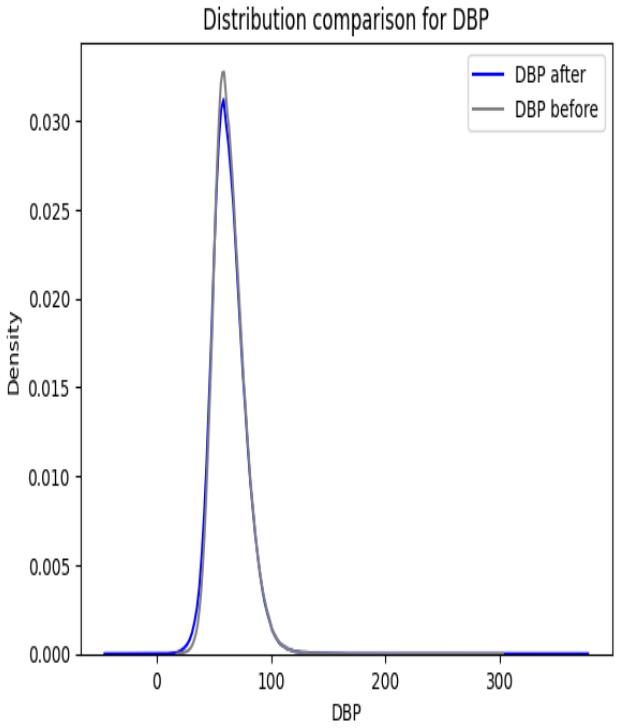

Figure 9: Histogram of Diastolic Blood Pressure (DBP) feature distribution before (gray) and after (blue) preprocessing.

In continuation, the Heart Rate (HR) feature displayed in Figure 10 further confirms the stability of the preprocessing pipeline. Despite slight skewness in the heart rate values, the preprocessed (blue) and original (gray) curves demonstrate strong alignment across most of the range. This agreement reinforces the effectiveness of the preprocessing steps in preserving key physiological patterns across multiple vital signs.

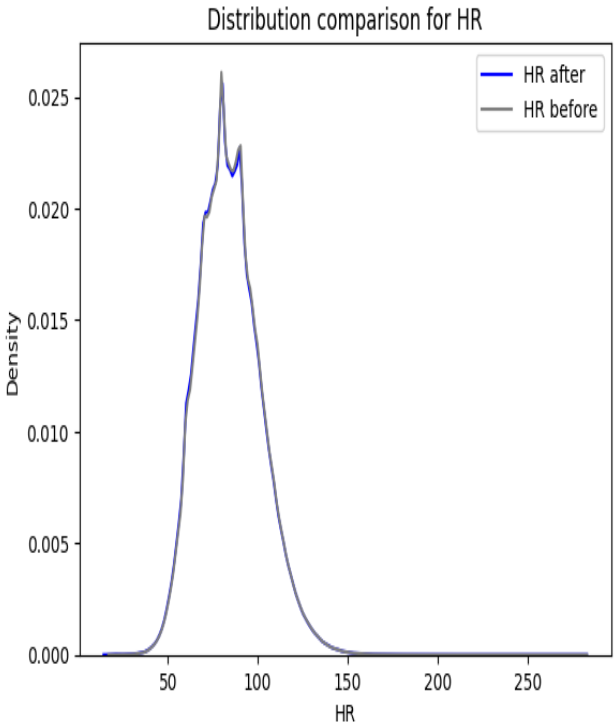

Figure 10: Distribution of Heart Rate (HR) feature before (gray) and after (blue) preprocessing.

Together, these results indicate that the preprocessing methodology maintained the essential distributional properties of critical features, ensuring data fidelity and minimizing bias introduced through imputation.

4 Experimental Results & Discussion

4.1 Experimental Setup

All models were implemented in Python 3.9 and executed in a high-performance computing environment provided by Kaggle, leveraging GPU acceleration for efficient training and evaluation.

Software and Frameworks. The experimental framework utilized Python 3.9 as the primary programming language with TensorFlow 2.13.0 (`tensorflow.keras.layers`) serving as the deep learning framework. The implementation incorporated NumPy for numerical computations, Pandas for data manipulation, Scikit-learn for machine learning utilities, Matplotlib for visualization, and SHAP and PFI for model interpretability analysis.

Execution Environment. All training and evaluation procedures were conducted on the Kaggle Kernel (Free Tier) platform, equipped with an NVIDIA Tesla P100 GPU featuring 16 GB HBM2 memory. Hardware acceleration was enabled throughout the experimental process, with the runtime environment configured as a GPU-enabled notebook to ensure optimal computational performance.

Development System. Initial model development and testing were performed on a local workstation prior to deployment on the Kaggle GPU environment. The local system specifications comprised an Intel Core i7-13650HX processor (13th generation, 14-core), NVIDIA GeForce RTX 3050 graphics card (6 GB GDDR6), 16 GB DDR5 RAM (2Œ8 GB, 4800 MHz), 512 GB M.2 PCIe NVMe SSD storage, and 15.6-inch FHD display (1920Œ1080) operating at 120 Hz refresh rate. This configuration provided an optimal balance between development flexibility and computational efficiency during data preprocessing and hyperparameter tuning phases.

Reproducibility. Experimental reproducibility was ensured through systematic implementation of fixed random seeds using NumPy and TensorFlow random state initialization protocols. Model evaluation employed stratified 5-fold cross-validation methodology to guarantee robust and statistically reliable performance assessment across heterogeneous data partitions.

4.2 Model Performance Results

Among time-series models, BiLSTM, TCN, and ResNet exhibit superior performance with F1 (S) scores of 0.85, 0.85, and 0.77, respectively, as shown in Table 6.

To test and evaluate the models, 23,000 data points out of a total of 85,000 were kept hidden as an internal test set.

BiLSTM demonstrates strong performance with an F1 (S) score of 0.85, particularly in recall (0.95), indicating effective identification of positive instances (Table 6). Its precision (0.77) is comparable to that of TCN. The ROC-AUC score of 0.9566 further supports its robust discriminative ability. The models false negative (FN) count is 377 out of 23,000, highlighting its reliability in medical contexts where minimizing missed positive cases is critical [26]. The model benefits from bidirectional processing, which captures temporal context from both past and future data.

TCN achieves an identical F1 (S) score of 0.85, with precision of 0.77 and recall of 0.94 (Table 6). It records the highest ROC-AUC score of 0.9595 among all time-series models, reflecting excellent class separation. The FN count is slightly higher than BiLSTM at 480, yet still shows strong performance in minimizing undetected positives.

ResNet yields a lower F1 (S) score of 0.77 compared to TCN and BiLSTM, with precision at 0.75 and recall at 0.78 (Table 6). Its ROC-AUC is 0.8908, indicating relatively reduced discriminative capability. Moreover, the FN count stands at 1,350 out of 23,000, suggesting diminished effectiveness in capturing positive cases.

Overall, both TCN and BiLSTM stand out as the most effective time-series models. TCN offers a slight edge in ROC-AUC, whereas BiLSTM leads in recall, making it particularly suited for clinical tasks requiring high sensitivity.

Table 6: Model Performance Comparison

Model Name	F1 (S)	F1 (NS)	Prec (S)	Rec (S)	ROC-AUC	FN
Random Forest	0.69	0.88	0.77	0.62	0.886	227/2000
Bootstrap RF	0.62	0.86	0.72	0.54	0.834	275/2000
XGBoost	0.72	0.86	0.64	0.82	0.810	107/2000
Logistic Regression	0.65	0.81	0.57	0.74	0.813	153/2000
TCN	0.85	0.90	0.77	0.94	0.9595	480/23000
BiLSTM	0.85	0.91	0.77	0.95	0.9566	377/23000
ResNet	0.77	0.82	0.75	0.78	0.8908	1350/23000
LSTM	0.80	0.84	0.78	0.82	0.8957	953/23000
InceptionTime	0.76	0.91	0.68	0.79	0.9014	1115/23000
1D CNN	0.77	0.86	0.72	0.82	0.9000	1419/23000
GRU	0.70	0.79	0.60	0.83	0.8386	1360/23000
RNN	0.73	0.81	0.63	0.87	0.8700	1022/23000

FN - False Negative; F1 (S) - F1 score for Sepsis class; F1 (NS) - F1 score for Non-Sepsis class

4.3 XAI Techniques

the top five modelsTCN, BiLSTM, ResNet, LSTM, and InceptionTimewere selected for explainability analysis due to their strong overall performance in sepsis detection, as shown in Table 6.

4.3.1 TCN Model Explainability

As summarized in Table 6, TCN achieves an F1 (S) score of 0.85 and an F1 (NS) score of 0.90, with precision (S) of 0.77, recall (S) of 0.94, and ROC-AUC of 0.9595. It recorded 480 FNs, reflecting a strong ability to identify positive sepsis cases while performing well on negative cases.

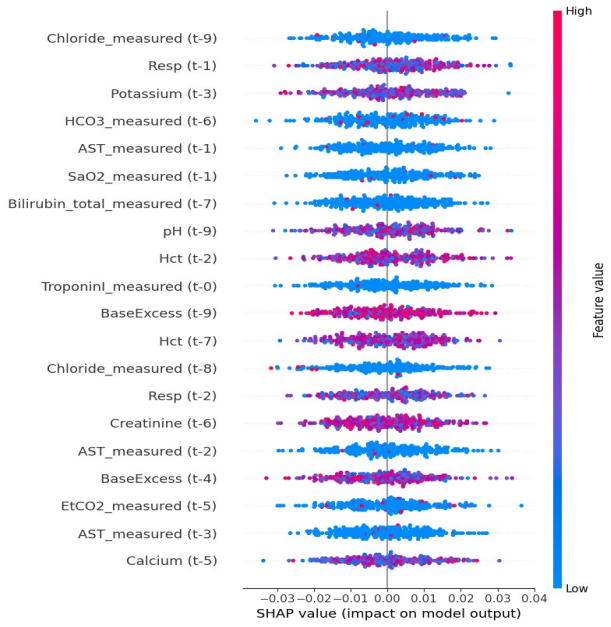

Figure 11: SHAP summary plot for TCN model showing clinical feature contributions to sepsis prediction

SHAP employs game theory principles to fairly assign contribution values to input features for model outputs. For complex time-series models such as TCNs, SHAP quantifies the influence of each feature value at individual time steps on the final prediction, providing granular, instance-level explanations that are critical for clinical decision-making and real-time monitoring.

Figure 11 presents the SHAP summary plot for the TCN model. Each point corresponds to a prediction instance, with the x-axis indicating contribution magnitude and direction. Features are ranked from top to bottom by their importance. Colors represent the actual feature values (red: high, blue: low), and the feature names are tagged with their corresponding time step relative to the prediction time (e.g., t-1, t-9).

Figure 11 confirms that specific clinical measurements at certain time steps have a strong influence on the models sepsis prediction. Notably, **Chloride measured at time step t-9**, **Respiration at t-1**, **Potassium at t-3**, and **HCO_3 measured at t-6** are among the key contributors. High Chloride (t-9) and high Respiration rate (t-1) are associated with an increased predicted risk of sepsis, while high Potassium levels (t-3) tend to decrease the risk.

These findings align well with established clinical knowledge: elevated respiration rate is a hallmark of systemic inflammatory response, and abnormal chloride levels are often seen in patients with metabolic imbalance linked to sepsis. Potassium levels, while variable, may reflect different physiological states, with lower values often linked to critical illness. Therefore, the SHAP plot not only explains the models decision process but also validates it against several medically known sepsis indicators and risk patterns.

Gradient-based methods compute the sensitivity of model output to input feature perturbations, revealing important time windows and key features influencing predictions for temporal models such as TCNs.

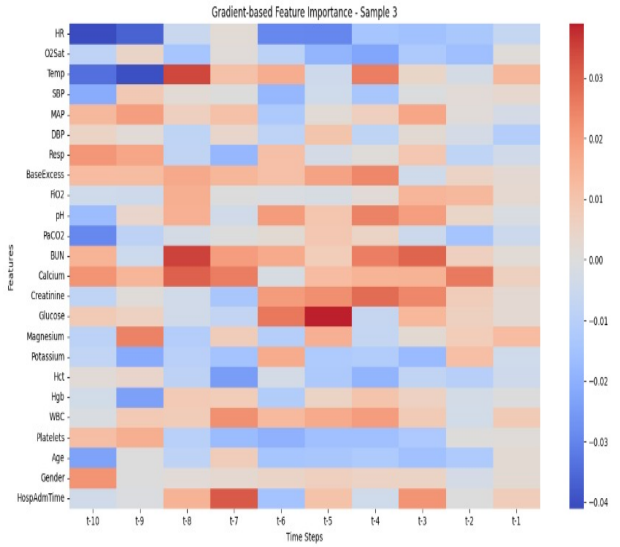

Figure 12: Gradient-based feature importance across time steps for a sample from the test set (TCN model)

Figure 12 shows a gradient-based heatmap illustrating feature sensitivity across time steps for a test sample. Warmer colors indicate positive influence, cooler colors indicate negative influence, and paler areas indicate low sensitivity. Vital signs, including temperature, O2Sat, HR, SBP, DBP, and respiration, show high sensitivity at times t-8, t-5, and t-2. Laboratory values such as BUN, creatinine, potassium, and hematocrit are influential around t8 and t3. This visualization highlights the temporal dynamics of clinical features affecting sepsis risk prediction.

4.3.2 BiLSTM Model Explainability

As summarized in Table 6, BiLSTM achieves an F1 (S) score of 0.85, a recall of 0.95, a precision of 0.77, and ROC-AUC of 0.955, demonstrating strong positive case identification capacity.

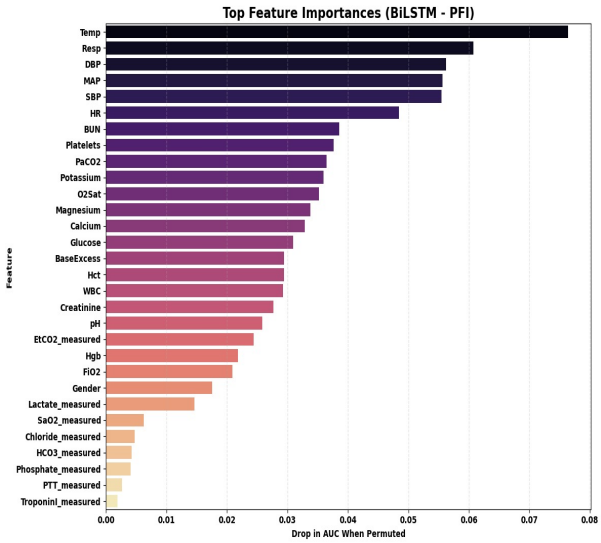

Figure 13: PFI analysis for BiLSTM model showing clinical feature contributions to sepsis prediction

SHAP was not applied to the BiLSTM model due to its complex sequential memory structure, which makes instance-level attribution less interpretable in a global context. Instead, PFI was selected to capture the aggregate influence of each feature on model predictions. PFI analysis in Figure 13 ranks clinical features by their impact on BiLSTM predictions, with temperature as most important, followed by mean arterial pressure, respiration rate, and SBP.

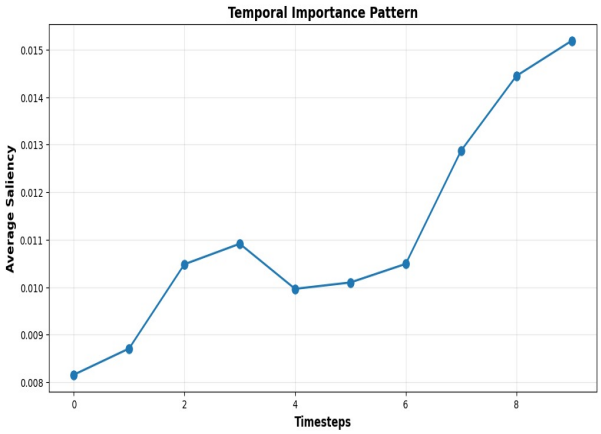

Figure 14: Integrated Gradients visualization showing prediction changes across time steps for a sample from the test set

Integrated Gradients visualization Figure 14, based on 100 high-risk samples at critical time steps, reveals BiLSTMs capacity to detect early predictive signals for sepsis. Saliency increases notably around time step 3, indicating model sensitivity to mid-sequence changes rather than relying solely on late deterioration.

4.3.3 ResNet Model Explainability

As summarized in Table 6, ResNet attains an F1 (S) score of 0.77, precision of 0.75, recall of 0.78, and ROC-AUC of 0.8908, reflecting moderate performance in positive sepsis identification.

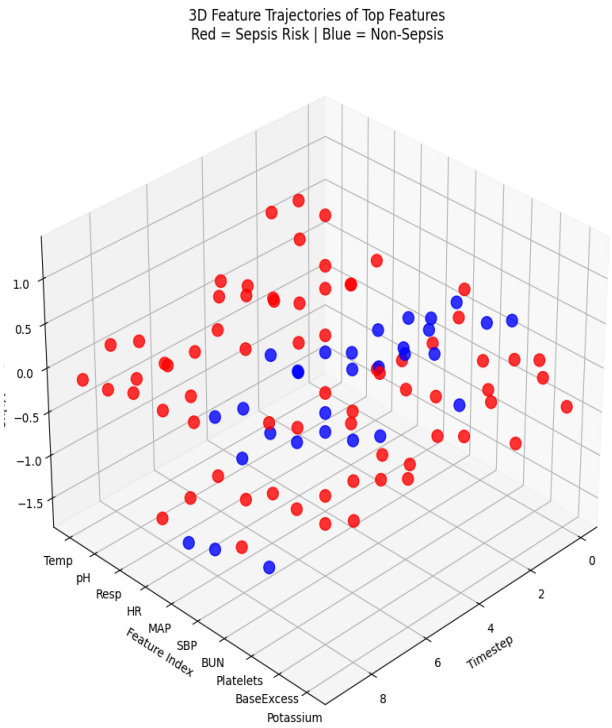

Figure 15: 3D temporal feature trajectories in sepsis prediction (Red: Sepsis Cases, Blue: Non-Sepsis Controls)

Figure 15 illustrates a three-dimensional visualization of temporal evolution of clinical features used in sepsis prediction. The x-axis represents time-step progression (0-8), the y-axis shows various features including temperature, respiratory rate, MAP, HR, pH, DBP, hematocrit, hemoglobin, platelets, and SBP. The z-axis corresponds to normalized feature importance values (-1.5 to +1.5). Red and blue points denote sepsis cases and non-sepsis controls, respectively. This visualization reveals distinct temporal patterns and feature trajectories that differentiate clinical outcomes, emphasizing the dynamic nature of sepsis progression and time-dependent relationships among variables. By capturing how the predictive value of biomarkers evolves, this 3D representation supports identification of critical feature combinations and lead-lag relationships, enhancing predictive modeling and early warning systems.

4.3.4 LSTM Model Explainability

As shown in Table 6, the LSTM model achieves an F1 (S) score of 0.80, with a precision of 0.78, a recall of 0.82, and an ROC-AUC of 0.8957, indicating solid performance in identifying positive sepsis cases.

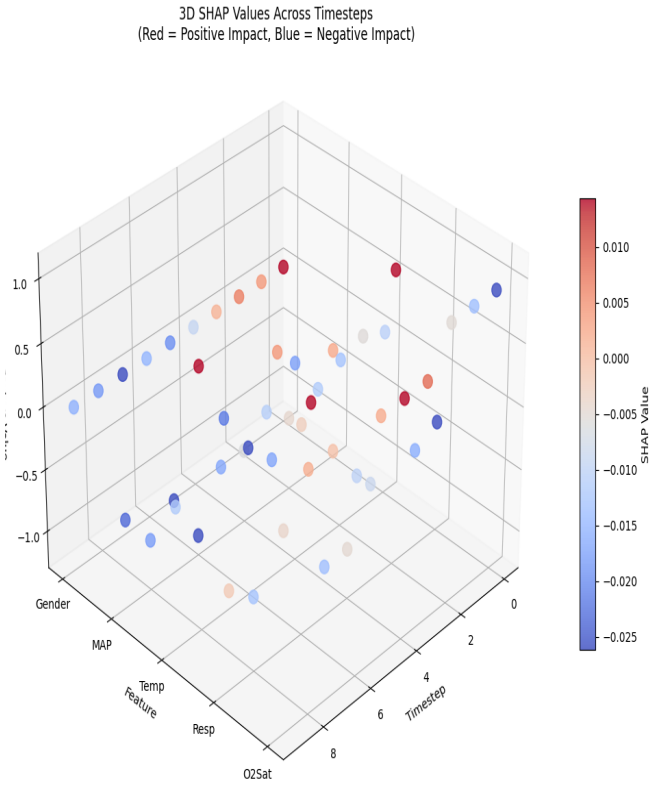

Figure 16: 3D SHAP values across timesteps for LSTM model (Red: Positive Impact, Blue: Negative Impact)

Figure 16 visualizes the temporal evolution of feature importance across multiple timesteps (08) for the LSTM model using 3D SHAP analysis. This visualization illustrates how critical featuresHR among themcontribute to the model's prediction. O2Sat, Temperature, Respiratory Rate (Resp), and Platelets contribute to sepsis prediction over time. Red points indicate positive SHAP values (features pushing toward sepsis prediction), while blue points represent negative SHAP values (features pushing the prediction away from sepsis).

The color intensity, ranging from -0.025 to +0.015, reflects the magnitude of each feature's contribution. This temporal perspective reveals how the predictive importance of vital signs evolves throughout the monitoring period, enabling clinicians to understand which biomarkers become more critical at different stages of patient observation and how the LSTM model weighs these dynamic relationships for early sepsis detection.

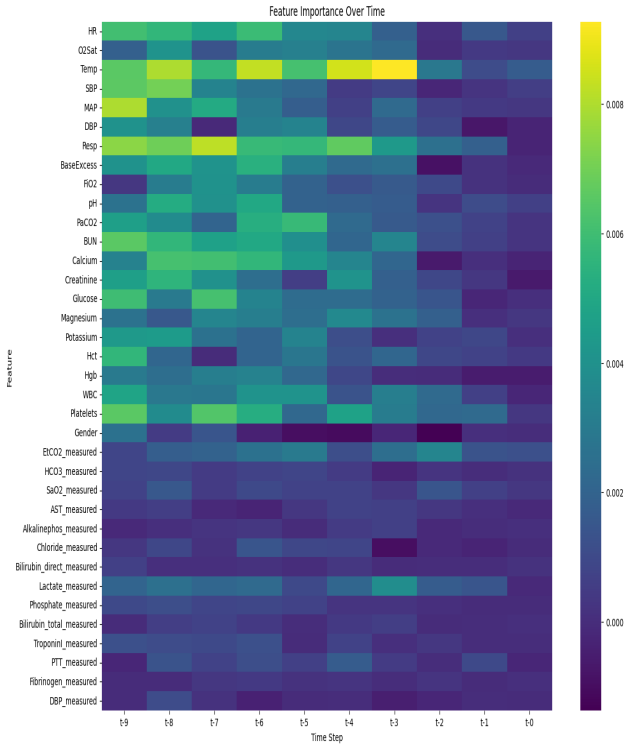

Figure 17: Feature importance over time for LSTM model

The heatmap in Figure 17 displays the temporal evolution of feature importance across timesteps (t-9 to t-0) for LSTM model in sepsis prediction. This visualization reveals how predictive significance of clinical features changes over time, with importance values ranging from 0.000 to 0.010 shown in a color gradient from dark blue to bright yellow. HR, Temperature, and Respiratory Rate (Resp) emerge as critical features, demonstrating high importance in early time steps, while O2Sat and blood pressure parameters maintain consistent relevance throughout the monitoring period. Temporal patterns reveal that vital signs demonstrate peak importance during middle timesteps (t-7 to t-5). Laboratory parameters, particularly electrolytes and biomarkers, demonstrate diverse temporal importance patterns. Capturing these dynamics is critical for optimizing the performance of LSTM model as it helps identify which clinical parameters become most predictive at different stages of patient monitoring and supports development of time-sensitive early warning systems for sepsis detection.

4.3.5 InceptionTime Model Explainability

As shown in Table 6,The InceptionTime model achieves an F1 (S) score of 0.76, with a recall of 0.86, a precision of 0.68, and an ROC-AUC of 0.9014 on a test set of 23,000 samples.

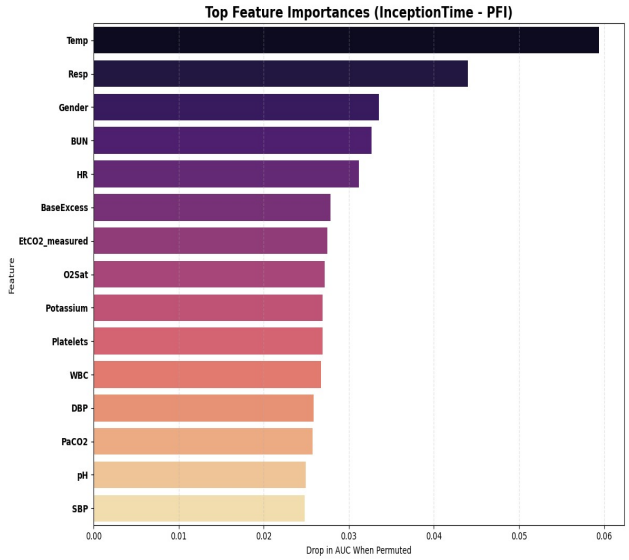

Figure 18: PFI analysis for InceptionTime model showing clinical feature contributions to sepsis prediction

PFI analysis in Figure 18 displays the contribution of clinical features to the model's sepsis prediction performance, with temperature as the most critical feature, followed by respiration rate and gender.

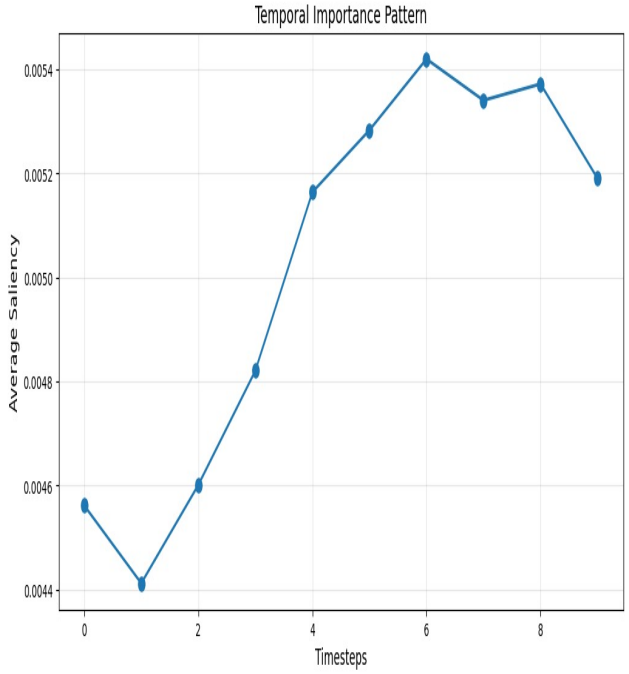

Figure 19: Gradient-based prediction changes across time steps for a sample from the test set (InceptionTime model)

Integrated Gradients visualization in Figure 19, based on 100 high-risk samples in their critical time steps, demonstrates that model successfully captures meaningful predictive signals for sepsis onset throughout temporal sequence. Saliency values begin to rise noticeably around timestep 3, indicating model detects important early-stage changes rather than relying solely on late-stage deterioration.

4.4 Discussion

Recent studies demonstrate that innovative approaches can be found for early diagnosis of sepsis by applying advanced computer algorithms [3, 11, 14]. Research indicates that intelligent systems are more adept at evaluating medical data across complex timescales [12, 14, 20]. Advanced neural network models have shown promising results in this regard. BiLSTM and TCN models have achieved remarkable results, with F1 ratios reaching approximately 0.85, surpassing traditional methods such as logistic regression (0.65) and RF (0.69) [8, 14].

What distinguishes these intelligent systems is their exceptional ability to accurately detect true cases, with recall ratios ranging between 0.94 and 0.95, reflecting high efficiency in detecting truly infected patients [16]. This superiority is due to fundamental technical advances in computer architecture, as these systems are capable of understanding complex temporal patterns present in electronic patient records. Bidirectional model analyzes preceding and subsequent temporal indicators, while temporal analysis networks rely on advanced algorithms to extract temporal patterns efficiently without need for expensive and complex computational operations.

Preliminary data processing methods played a key role in significantly improving performance. Researchers used advanced techniques to balance samples and generate synthetic data that mimics real biomarkers [6, 7], as well as intelligent strategies to address missing data while preserving original clinical context [23]. Unlike traditional methods that ignore missing data, the research team considered these gaps to be clinically meaningful indicators that reflect real-world diagnostic decisions made by clinicians in practice [12].

New methods used in this research demonstrate significant progress compared to previous work, achieving a 13% increase in accuracy compared to studies that relied on gradient boosting algorithms [17, 19].

This work contributes to development of an integrated diagnostic system that combines high scientific accuracy with easy interpretation of results.

New analytical approach, powered by multiple XAI techniques, effectively revealed complex and time-dependent relationships between a variety of diagnostic indicators. For instance, SHAP analysis revealed significant temporal influence of key vital signs such as Temp, HR, SBP, MAP, and DBP at various time steps leading up to prediction (Figure 11). Similarly, gradient-based heatmaps identified critical time windows where model's sensitivity to fluctuations in these same vital signs peaked, which helped distinguish sepsis from non-sepsis trajectories (Figure 12). These granular, time-aware insights consistently ranked vital signsTemperature (Temp), HR, SBP, MAP, and DBPas top global predictors across models (Figures 13, 18), enabling design of more intelligent medical monitoring systems responsive to dynamic and subtle evolution of patient's condition.

The effectiveness of these vital signs in sepsis detection is well established in the literature. Temperature alterations, both hyperthermia (>38řC) and hypothermia (<36řC), are fundamental sepsis criteria with odds ratios (OR) of 2.126 for elevated temperature [27, 28]. HRV analysis has demonstrated remarkable predictive capabilities, with studies showing that HR combined with temperature achieves

area under the curve values of 0.94 for sepsis prediction [29]. Additionally, heart rate characteristics monitoring has proven effective for early detection, particularly in neonatal sepsis, where decreased HRV punctuated by transient decelerations serves as a reliable biomarker [30, 31].

Blood pressure components provide crucial hemodynamic insights for sepsis assessment. SBP maintained at approximately 140 mmHg during the first 10 hours of hospitalization has been associated with decreased mortality risk [32, 33]. MAP serves as a critical perfusion indicator, with studies demonstrating that MAP below 70 mmHg carries an OR of 3.874 for sepsis development [27]. DBP has emerged as particularly significant, with values below 59 mmHg associated with increased 28-day mortality (OR 1.915) and serving as a marker of arterial tone and coronary perfusion adequacy [34, 35].

A combination of these vital signs provides synergistic diagnostic value. Research has shown that a minimal feature set yielding maximal predictability combines HR and temperature, achieving sensitivity of 0.85 and specificity of 0.90 [29]. Furthermore, shock index (HR/SBP) significantly correlates with sepsis severity, with increased values strongly influencing Sequential Organ Failure Assessment (SOFA) scores [27, 36].

Despite these promising developments, the study faces several fundamental challenges that need to be addressed. First, a lack of diversity in training samples may affect generalizability of results to other medical conditions. This shortcoming highlights the urgent need for in-depth research to expand the database and improve efficiency of the proposed diagnostic system. A deeper clinical evaluation of the interpretation mechanisms used in study is also required. Next steps should focus on application aspects, including integrating these systems with existing medical software and developing interactive interfaces that meet requirements for use in clinical settings.

Practically, this strategy represents an important advance in the development of effective early warning systems in intensive care units, which may contribute to reducing mortality rates resulting from sepsis [1, 4]. Furthermore, this strategy presents an innovative model for applying AI technologies in healthcare by balancing scientific rigor with practical relevance, a critical factor for successful application of these technologies in intensive care settings.

Effect of Meta-Heuristic Optimization on ResNet

Table 7 summarizes the impact of each meta-heuristic optimization stage relative to the pure (non-optimized) ResNet baseline, including both performance metrics and the final selected hyperparameters for each method. The results show that Tabu Search mainly stabilizes training and reduces validation loss while keeping accuracy comparable to the baseline model. When Simulated Annealing is used to refine Tabu Search parameters, the model achieves a modest yet consistent gain in both accuracy and F_1 score. The Grey Wolf Optimizer configuration provides the largest improvement, notably boosting ROC–AUC and further lowering validation loss compared with all other variants.

Table 7: Comparison between the pure ResNet baseline and each optimized variant. Values are taken from the corresponding confusion matrices and ROC–AUC evaluations.

Model	Accuracy	F_1 (Sepsis)	ROC–AUC	Val. Loss
Pure ResNet (no optimizer)	0.89	0.87	0.8868	0.4629
ResNet + Tabu Search (optimized)	0.89	0.88	0.8868	0.4890
ResNet + SA-optimized Tabu	0.90	0.89	0.89	0.4890
ResNet + GWO (final model)	0.91–0.92	0.91	0.9798	0.4828

Pure ResNet Baseline

The baseline ResNet, trained without meta-heuristic optimization, uses a dropout rate of 0.30 and a learning rate of 0.001. It achieves an accuracy of 89% and F_1-score of 0.87 for the sepsis class, with a ROC–AUC of 0.8868 and validation loss of 0.4629.

ResNet with Tabu Search

Tabu Search is first used to tune the ResNet hyperparameters. The search is carried out with:

- Tabu iterations: $n_{\text{iter}} = 3$
- Tabu list size: tabu_size $= 2$
- Step size: step_size $= 0.0442$

This procedure refines the learning rate and dropout while maintaining a validation loss of 0.4890 and accuracy around 89%, providing a more stable training behavior compared with the pure ResNet.

ResNet with Simulated AnnealingOptimized Tabu

Simulated Annealing is then applied as a meta-optimizer over the Tabu Search parameters themselves. The SA scheme uses:

- Initial temperature: $T_0 = 3.0$
- Cooling rate: $\alpha = 0.9$
- Number of SA iterations: 5

Within SA, candidate Tabu configurations such as $(n_{\text{iter}} = 3, \text{tabu\_size} = 2, \text{step\_size} \in \{0.0442, 0.0567, 0.0700\})$ are evaluated, and worse configurations can be accepted with probability $\exp(-\Delta E/T)$. The best setting keeps $n_{\text{iter}} = 3$, tabu_size $= 2$, and step_size $= 0.0442$, leading to a ResNet model with:

- Dropout rate: 0.2396
- Learning rate: 0.001064
- Test accuracy: 0.90
- F_1 (sepsis): 0.89

while preserving a validation loss of 0.4890.

ResNet with Grey Wolf Optimizer (GWO)

Finally, the Grey Wolf Optimizer directly tunes the ResNet hyperparameters within the following bounds:

- Dropout rate: $[0.10, 0.50]$
- Learning rate: $[0.0001, 0.01]$

The GWO configuration employs:

- Number of wolves: 10
- Number of iterations: 5

GWO converges to:

- Optimized dropout: 0.2567

- Optimized learning rate: 0.000501

which yields the best-performing model with test accuracy of 91–92%, F_1-score of 0.91, ROC–AUC of 0.9798, and a reduced validation loss of 0.4828, clearly outperforming the pure ResNet baseline and the Tabu-based variants.

5 Conclusion & Future Work

This work offers a comprehensive system for predicting sepsis in its early stages by integrating interpretable artificial intelligence (AI) methods alongside sophisticated machine learning (ML) algorithms. By improving system performance for solving common problems in clinical datasets, such as missing data and imbalanced classes [3, 25], we developed preprocessing workflows tailored to real-world clinical irregularities. Temporal deep learning models BiLSTM and TCN perform best with F1 scores of 0.84 and 0.85, and ROC-AUC scores of 0.955 and 0.966, respectively (Table 6). These models bolster performance in capturing the dynamic nature of physiological signals, critically reducing false negatives in life-threatening situations where early detection is paramount [1, 4]. The temporal modeling capability of these architectures enables identification of subtle time-based patterns that traditional methods often miss [8, 14], providing clinicians with more reliable early warning systems for sepsis onset. To enhance interpretability and clinician trust, a variety of interpretable AI methods, including SHAP, PFI, and other gradient and cumulative local effect methods were used to reveal not only which features were most predictive but also when they became most critical in a patient's timeline, as shown in our XAI analyses (Figures 11–19).

Looking ahead, testing Selected Models within real intensive care unit (ICU) environments will be a crucial step in confirming their validity and robustness across a wide range of clinical scenarios. Predictive models could also be incorporated into clinical workflows through intuitive dashboards that provide explanations and forecasts, thus improving decision-making at the bedside. Including longitudinal patient records as well as comprehensive medical histories could improve the accuracy of predictions and generalizability. Other temporal models could also enhance these improvements using ensemble learning strategies that draw on multiple temporal models. The ability to provide proactive interventions could be possible with real-time adaptive monitoring systems that continuously revise predictions based on new patient information [12]. Addressing ethical issues like algorithmic bias and patient privacy will need to be attended to in order for AI technologies to be responsibly utilized in the healthcare sector [38, 39]. With these considerations, the goal is to develop reliable, transparent, and clinically actionable AI systems for early sepsis detection that can meaningfully improve patient outcomes in critical care settings [40, 41].

References

[1] World Health Organization, "WHO calls for global action on sepsis—cause of 1 in 5 deaths worldwide," Sep. 8, 2020. [Online]. Available: https://www.who.int/news/item/08-09-2020-who-calls-for-global-action-on-se

[2] R. L. Gauer, M. L. Braun, and M. C. Ott, "Sepsis: Diagnosis and Management," *Am. Fam. Physician*, vol. 101, no. 7, pp. 409–418, Apr. 2020. [Online]. Available: https://www.aafp.org/pubs/afp/issues/2020/0401/p409.html

[3] M. A. Reyna, C. B. Josef, R. Jeter, *et al.*, "Early prediction of sepsis from clinical data: The PhysioNet/Computing in Cardiology Challenge 2019," *Crit. Care Med.*, vol. 48, no. 2, pp. 210–217, Feb. 2020. doi: 10.1097/CCM.0000000000004145

[4] A. Kumar, D. Roberts, K. E. Wood, *et al.*, "Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock," *Crit. Care Med.*, vol. 34, no. 6, pp. 1589–1596, Jun. 2006. doi: 10.1097/01.CCM.0000217961.75225.EF

[5] T. W. van der Vaart, J. F. Huguenard, and A. J. Rogers, "Diagnostic Challenges in Sepsis," *Curr. Infect. Dis. Rep.*, vol. 23, no. 12, pp. 1–9, Oct. 2021. doi: 10.1007/s11908-021-00765-y

[6] S. Lyra, S. Leonhardt, and C. H. Antink, "Early prediction of sepsis using random forest classification for imbalanced clinical data," in *Computing in Cardiology*, vol. 46, Sep. 2019. doi: 10.22489/CinC.2019.071

[7] Q. Wu, Y. Li, X. Liu, *et al.*, "A customized downsampling machine learning approach for sepsis prediction," in *Computing in Cardiology*, vol. 46, Sep. 2019. doi: 10.22489/CinC.2019.071

[8] J. Liao, Y. Li, X. Liu, *et al.*, "Does deep learning REALLY outperform non-deep machine learning for clinical prediction on physiological time series?" *arXiv preprint*, arXiv:2211.06034, Nov. 2022. [Online]. Available: https://arxiv.org/abs/2211.06034

[9] S. Nirgudkar and T. Ding, "Early detection of sepsis using ensembles," in *Computing in Cardiology*, vol. 46, Sep. 2019. doi: 10.22489/CinC.2019.071

[10] Vandana and R. Chhikara, "Comparative Analysis of ML Models with Selection Methods for Early Predictive Analytics of Sepsis in ICU," *Indian J. Sci. Technol.*, vol. 17, no. 41, pp. 4338–4348, Nov. 2024. doi: 10.17485/IJST/v17i41.1925

[11] G. Kong, X. Lin, and H. Hu, "Using machine learning methods to predict sepsis," *Int. J. Med. Inform.*, vol. 131, p. 103940, Nov. 2019. doi: 10.1016/j.ijmedinf.2019.103940

[12] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Modeling missing data in clinical time series with RNNs," *arXiv preprint*, arXiv:1606.04130, Jun. 2016. [Online]. Available: https://arxiv.org/abs/1606.04130

[13] L. Liu, Y. Li, X. Liu, *et al.*, "Early prediction of sepsis from clinical data via heterogeneous event aggregation," *arXiv preprint*, arXiv:1910.06792, Oct. 2019. [Online]. Available: https://arxiv.org/abs/1910.06792

[14] K. Apalak and K. Kiasaleh, "Early Detection of Sepsis With Temporal Convolutional Networks," *IEEE Access*, vol. 8, pp. 132437–132449, 2020. doi: 10.1109/ACCESS.2020.3010445

[15] S. P. Oei, M. J. Schuijs, *et al.*, "Towards early sepsis detection from measurements at the general ward through deep learning," *Intelligence-Based Medicine*, vol. 5, p. 100042, 2021. doi: 10.1016/j.ibmed.2021.100042

[16] F. Mohammad, Y. Li, X. Liu, *et al.*, "Early prediction of onset of sepsis in a clinical setting," *arXiv preprint*, arXiv:2402.03486, Feb. 2024. [Online]. Available: https://arxiv.org/abs/2402.03486

[17] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785

[18] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Advances in Neural Information Processing Systems*, vol. 24, 2011, pp. 2546–2554.

[19] M. Yang, X. Wang, H. Gao, Y. Li, X. Liu, J. Li, and C. Liu, "Early prediction of sepsis using Multi-Feature fusion based XGBoost learning and Bayesian optimization," in *Computing in Cardiology*, vol. 46, Sep. 2019. doi: 10.22489/CinC.2019.020.

[20] H. Harutyunyan, H. Khachatrian, D. C. Kale, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific Data*, vol. 6, no. 1, p. 96, Jun. 2019. doi: 10.1038/s41597-019-0103-9

[21] Y. Hong, J. Zhou, S. Ji, and Y. Fang, "Sepsis prediction using InceptionTime architecture," in *Proceedings of the 2021 International Conference on Smart City and Ubiquitous Computing*, 2021.

[22] N. Takahashi, T.-A. Nakada, K. R. Walley, and J. A. Russell, "Significance of lactate clearance in septic shock patients with high bilirubin levels," *Sci. Rep.*, vol. 11, no. 1, p. 6313, Mar. 2021. doi: 10.1038/s41598-021-85700-w

[23] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, "Multiple imputation by chained equations: What is it and how does it work?" *Int. J. Methods Psychiatr. Res.*, vol. 20, no. 1, pp. 40–49, Mar. 2011. doi: 10.1002/mpr.329

[24] R. E. Klabunde, *Cardiovascular Physiology Concepts*, 2nd ed. Philadelphia, PA: Lippincott Williams & Wilkins, 2012. ISBN: 9781451113846.

[25] A. E. Johnson, T. J. Pollard, L. Shen, *et al.*, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, p. 160035, May 2016. doi: 10.1038/sdata.2016.35

[26] J. Futoma, S. Hariharan, K. Heller, M. Sendak, N. Brajer, M. Clement, A. Bedoya, and C. O'Brien, "An Improved Multi-Output Gaussian Process RNN with Real-Time Validation for Early Sepsis Detection," *arXiv preprint arXiv:1708.05894*, 2017. Available: https://arxiv.org/abs/1708.05894

[27] T. Kenzaka, A. Okayama, S. Kuroki, *et al.*, "Importance of vital signs to the early diagnosis and severity of sepsis: Association between vital signs and Sequential Organ Failure Assessment score in patients with sepsis," *Intern. Med.*, vol. 51, no. 8, pp. 871–876, 2012. doi: 10.2169/internalmedicine.51.6951

[28] Sepsis Alliance, "Symptoms - Sepsis Alliance," Jan. 2023. [Online]. Available: https://www.sepsis.org/sepsis-basics/symptoms/

[29] E. S. Rangan, A. J. Rogers, *et al.*, "Performance effectiveness of vital parameter combinations for early warning of sepsis," *J. Am. Med. Inform. Assoc.*, vol. 29, no. 10, pp. 1596–1606, Oct. 2022. 10.1093/jamiaopen/ooac080

[30] K. D. Fairchild, L. E. Schelonka, *et al.*, "Complex signals bioinformatics: Evaluation of heart rate characteristics monitoring as a novel risk marker for neonatal sepsis," *J. Clin. Monit. Comput.*, vol. 28, no. 4, pp. 329–339, Aug. 2014. doi: 10.1007/s10877-013-9530-x

[31] S. L. Kausch, A. J. Rogers, *et al.*, "A cardiorespiratory signature of sepsis in 3 NICUs," *medRxiv*, Sep. 2022. [Online]. Available: https://doi.org/10.1101/2022.09.28.22280469

[32] Infectious Disease Advisor, "Systolic Blood Pressure Trajectory May Influence In-Hospital Mortality Risk in Sepsis," May 2024. [Online]. Available: https://www.infectiousdiseaseadvisor.com/news/systolic-blood-pressure-trajectory-may-influ

[33] J. L. Zhu, Y. Li, X. Liu, *et al.*, "Influence of systolic blood pressure trajectory on in-hospital mortality in patients with sepsis," *BMC Infect. Dis.*, vol. 23, no. 1, p. 90, Feb. 2023. doi: 10.1186/s12879-023-08054-w

[34] S. Y. Liu, Y. Li, X. Liu, *et al.*, "Association between diastolic blood pressure during the first 24 h and 28-day mortality in patients with septic shock: A retrospective observational study," *BMC Emerg. Med.*, vol. 23, no. 1, p. 97, Sep. 2023. doi: 10.1186/s12873-023-00864-0

[35] S. Ranjit and R. Natraj, "Hemodynamic assessment and management of septic shock in children," *J. Pediatr. Crit. Care*, vol. 11, no. 10, pp. 186–194, Nov. 2024. [Online]. Available: 10.5005/jp-journals-10071-24246

[36] T. Berger, J. Green, T. Horeczko, Y. Hagar, N. Garg, A. Suarez, E. Panacek, and N. Shapiro, "Shock Index and Early Recognition of Sepsis in the Emergency Department: Pilot Study," *West J. Emerg. Med.*, vol. 14, no. 2, pp. 168–174, Mar. 2013. doi: 10.5811/westjem.2012.8.11546. PMCID: PMC3628475. PMID: 23599863.

[37] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Scientific Reports*, vol. 8, no. 1, pp. 6085, 2018. Available: https://arxiv.org/abs/1606.01865

[38] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard Journal of Law & Technology*, vol. 31, no. 2, pp. 841–887, 2017. Available: https://arxiv.org/abs/1711.00399

[39] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)," in *Proc. 35th Int. Conf. Machine Learning (ICML)*, pp. 2668–2677, 2018. Available: https://arxiv.org/abs/1711.11279

[40] H. Harutyunyan, H. Khachatrian, D. C. Kale, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific Data*, vol. 6, no. 1, p. 96, 2019. Available: https://www.nature.com/articles/s41597-019-0103-9

[41] A. E. W. Johnson, T. J. Pollard, L. Shen, et al., "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, p. 160035, 2016. DOI: 10.1038/sdata201635

[42] G. D. Rubenfeld, E. R. Caldwell, E. Peabody, J. Hudson, R. Thompson, D. Moses, and L. Martin, "Incidence and outcomes of acute lung injury," *New England Journal of Medicine*, vol. 353, no. 16, pp. 1685–1693, 2005. DOI: 10.1056/NEJMoa050333

[43] J. L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. K. Reinhart, P. M. Suter, and L. G. Thijs, "The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure," *Intensive Care Medicine*, vol. 22, pp. 707–710, 1996. DOI: 10.1007/BF01709751