SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

A Unified Weighted Ensemble Framework for Simultaneous Early Sepsis Prediction and ICU Intervention Requirement Forecasting

Dr. T. Rajesh[1], Dr. B. Sashidhar[2], Vaishnavi Kalancha[3], D. Naga Swetha[4], K Gnana Prasuna[5], Siva Sankar Namani[6]

[1]Department of Computer Science and Engineering,G. Narayanamma Institute of Technology and Science, Hyderabad.
E-mail: rajesh@gnits.ac.in
[2]Department of Computer Science and Engineering (AI & ML),G. Narayanamma Institute of Technology and Science, Hyderabad.
E-mail: bolasasi@gnits.ac.in
[3]Department of Computer Science and Engineering,G. Narayanamma Institute of Technology and Science, Hyderabad.
E-mail: vaishnavikalancha7@gmail.com
[4]Department of Computer Science and Engineering,G. Narayanamma Institute of Technology and Science, Hyderabad.
E-mail: swetha@gnits.ac.in
[5]Department of Computer Science and Engineering,G. Narayanamma Institute of Technology and Science, Hyderabad.
E-mail: gnanaprasuna@gnits.ac.in
[6]Department of Computer Science and Engineering (AI & ML),G. Narayanamma Institute of Technology and Science, Hyderabad.
E-mail: siva@gnits.ac.in

Abstract
Sepsis is estimated to impact 48.9 million people globally annually and responsible for almost 11 million deaths, and is the leading cause of death in hospitalised patients in the UK and Ireland. Most existing machine learning and deep learning methods are dedicated to predicting a single outcome, have a learning strategy that considers only a single row, can possibly lead to inter-patient information leakage, and need a different model for each intervention in the ICUs, which limits clinical applicability. To overcome these drawbacks, a unified patient-level multi-task weighted ensemble approach for the simultaneous prediction of sepsis, mechanical ventilation, vasopressor therapy and renal replacement therapy is proposed in this study, based on the PhysioNet Computing in Cardiology Challenge 2019, containing 20,336 ICU admissions and 790,215 hourly observations. The proposed framework employs the first 24 hours of data from the ICU to represent a patient-level summary of their data, adds clinical feature engineering, task-specific weighted ensemble learning of XGBoost–LightGBM, and optimises the F2 score threshold to better balance precision and recall and to prevent inter-patient information leakage.
The mean AUROC and mean accuracy of experimental evaluation were 0.948 and 0.8946%, respectively, where AUC were 0.841, 0.980, 0.979, and 0.991 for sepsis, mechanical ventilation, vasopressor therapy, and renal replacement therapy, respectively. Comparative evaluation showed that the proposed framework achieved better results than six of the baseline machine learning and deep learning models, with a 0.133 AUROC improvement over the best-performing deep learning model and an AUROC standard deviation of < ±0.006 over 5-fold cross validation. The results show that the proposed framework has demonstrated its capabilities to accurately classify the risk level of a patient in the early phase of hospitalization in the ICU and guide interventions, with relevance for clinical use, being computationally efficient, clinically accurate, and clinically deployable.

Keywords: Sepsis prediction; ICU intervention forecasting; Multi-task learning; Gradient boosting ensemble; XGBoost; LightGBM; Feature engineering; PhysioNet 2019; Class imbalance; Early warning scores

1. Introduction

Sepsis is an uncontrolled response to infection that causes multi-organ dysfunction, is a major cause of preventable death in intensive care units (ICU) and is considered a major public health issue. It is a disease that is estimated to cause around 11 million deaths each year and impact on nearly 48.9 million people worldwide every year [1]. Given these facts, it should be clear that fast recognition and treatment could make a significant difference to the mortality outcome and that automated prediction systems are essential to enable the early clinical recognition and intervention of patients and, consequently, improve the clinical outcome. The methods of traditional ICU risk assessment are based on clinical scores, such as Systemic Inflammatory Response Syndrome (SIRS) [4], Quick Sequential Organ Failure Assessment (qSOFA) [8], Sequential Organ Failure Assessment (SOFA) [27] and National Early Warning Score 2 (NEWS2) [23]. These methods, which rely on a set threshold for a single point in time, have been broadly employed and do not consider physiology changes with time. Furthermore, they are only general clinical risk predictors, but not critical intervention risk predictors such as mechanical ventilation, vasopressor therapy, renal replacement therapy, and so on, limiting their utility for making comprehensive clinical decisions in the ICU [5].

1.1Existing System

Early prognosis of sepsis with the ability to identify complex relationships in EHRs is accurately possible with

machine learning and deep learning models. Another important component technology is the recurrent neural network (RNN). Boosting algorithms like XGBoost, LightGBM and transformer models are a part of them which are found to be effective in the proactive modeling.There are, however, three important aspects of the current techniques which remain limiting. Most studies have been directed to predicting sepsis as a task problem rather than predicting various interventions in the intensive care unit (ICU) [7]; thus, various models are required to predict each respective intervention in the ICU. Secondly, several existing approaches employ row-level learning, where each hourly ICU observation is considered an independent sample. As a result, multiple observations from the same patient may appear in both the training and testing partitions, leading to inter-patient information leakage and inflated performance estimates due to overfitting [8].Third, deep learning models can well capture the temporal dependencies, they tend to need a lot of computational resources and are not consistently better I found that it outperforms gradient boosting techniques in highly irregularly sampled structured ICU datasets with lots of missing and zero values. missingness [10]. These constraints underscore the importance of having a comprehensive and loss-less prediction tool that predicts multiple important clinical outcomes in the ICU, as well as Efficiently performs numerical calculations.

1.2 Proposed Framework

To tackle this requirement, this study introduces a multi-task The weighted ensemble model was used to predict sepsis, mechanical ventilation (MV) and vasopressors (VP). All children who received both therapy (VT) and renal replacement therapy (RRT) at the same time within the data set of PhysioNet Computing in Cardiology Challenge 2019 which has 20,336 ICU admissions and 790,215 hourly observations. The framework combines the first 24 hours of observations in the ICU into the representation of a patient, providing a patient-level representation without information leakage across patients and with the potential to predict four clinically relevant outcomes [7]. Both the predictive accuracy and robustness and computation efficiency are improved by ensemble method of the task-specific weighted XGBoost–LightGBM. This work is unique in that it proposes a single predictive model to accurately predict sepsis and three key interventions required in the ICU in one system without any leakage [28]. Contrary to previous models which build separate models for each clinical outcome, the proposed framework builds an accurate and clinically deployable multi-task prediction by combining patient-level learning, clinically informed feature engineering, and task-specific weighted ensemble learning.

1.3 Major Contributions

This work has the following main contributions:

1. Patient-level multi-task framework that simultaneously predicts sepsis, mechanical ventilation, vasopressor therapy and renal replacement therapy with a mean AUROC of 0.948 that is leakage-free.

2. Patient-level feature representation using a combination of temporal, physiological and clinically relevant features to predict risks of admission to the ICU at the early stage.

3. An AUROC of 0.948 was achieved by the task-specific weighted ensemble framework outperforming the best-performing deep learning baseline by 0.133 AUROC. Full evaluation and deployment through a real-time. Streamlit clinical decision support interface of six base models.

The rest of this paper is organized as follows: In Section 2, the relevant work is reviewed, Section 3 presents the methodology proposed, in Section 4, the experimental results are discussed and finally in Section 5, suggestions are made for future work.

2. Literature Survey

Early stratification based on clinical scoring systems was mainly rule-based systems in the ICU. The Systemic Inflammatory Response Syndrome (SIRS) was one of the earliest proposed systematic inflammation criteria to define sepsis related systematic inflammation, proposed by Bone et al. [4] which were, however, poorly specific in the heterogeneous setting of the critical care [13]. The introduction of Sequential Organ Failure Assessment (SOFA) score redefines Sepsis-3 consensus [27] which gave clarity about organ dysfunction, but needed laboratory tests that are difficult to obtain on the bedside. In response, the simple bedside screening tools of the Quick Sequential Organ Failure Assessment (qSOFA) [8] and the National Early Warning Score 2 (NEWS2) [23] were developed. These scoring systems are all threshold, reactive and fail to predict multiple requirements of interventions in the ICU at the same time [5]. As a result, the clinical scores are increasingly used as informative features in recent machine learning frameworks instead of as prediction tools.The prediction of sepsis early has been significantly improved by machine learning techniques which learn complex nonlinear relationships from EHRs. Desautels et al. [6] created a model named with an AUROC of 0.88, and Henry et al. [11] introduced a random forest model, TREWScore, to develop an early warning system with an AUROC of 0.83 and with an average warning time of 28 hours before septic shock. Ensemble tree-based methods on imbalanced clinical datasets, as shown by Mani et al. [18]have proven to be effective. Reyna et al. [19] launched the PhysioNet/Computing in Cardiology Challenge to spur research on early sepsis detectionwith predictive models on standardized benchmark datasets.

Gradient boosting algorithms, such as XGBoost [3] and LightGBM [14] have demonstrated strong predictive capability to model nonlinear relationships and their performance when dealing with missing data and clinical variables with mixed data types.To predict AKI in septic patients, Yue et al.[30] proved that XGBoost performed well with an AUROC of 0.817. Despite these advances, most of the current studies are limited to predicting a single outcome, and use a row-wise formulation that could leak information across patients.To model temporal dependency in physiological time-series, the models of deep learning have been extended to predict ICU. To date, LSTM networks have been extensively used for early sepsis prediction, and Scherpf et al. [25] showed that they can achieve competitive performance, but they also discussed the difficulty to generalise the model across institutions. Transformer-based architectures have additionally enhanced sequential modeling abilities for clinical time-series evaluation [19] and temporal probabilistic profile approaches from Moor et al. [26] have shown that temporal patterns offer valuable deterioration signals in addition to static clinical features. The attention mechanism and temporal feature fusion in critical care prediction have also been investigated in recent hybrid deep learning models [20,21].

Despite these improvements, deep learning models tend to be expensive to train, are not well understood, and have not always been shown to outperform gradient boosting models on structured ICU data with irregular sampling and high missingness rates [10]. Multiple models are included in the ensemble learning to enhance the prediction accuracy and generalization, especially in the complex prediction problems in the healthcare domain. Chen and Guestrin [3] and Ke et al. [14] set up gradient boosting frameworks for structured clinical data, which are now the state-of-the-art. Futoma et al. [9] showed how multi-task learning can achieve better clinical prediction results by sharing feature representations across the different prediction tasks, and Harutyunyan et al. [17] proposed a benchmark for the multitask mortality prediction, decompensation, and length of hours in clinical time-series data. Recent research also has demonstrated that the strength of the predictions can be improved by combining the decision boundaries of various classifiers as ensemble learning [18,22]. Most of the current methods are however based on a single-model architecture or on a naive averaging method without any task-specific weights tailored to the specific prediction task. Recently, there have been efforts to apply machine learning to prediction of future ICU interventions, going beyond sepsis prediction. They proposed a graph convolutional network to predict mechanical ventilation and vasopressor need by using temporal and inter-variable modeling in predicting the mechanical ventilation, with an accuracy of 91.9% [29]. Wijnberge et al. [28] showed that machine learning predictive capabilities of vasopressors can reduce the duration of intraoperative hypotension, and Kellum et al. [15] defined KDIGO criteria which form the basis of the staging of acute kidney injury and prediction of renal interventions. Physiological time-series and electronic health records have also been used to develop intervention-specific prediction models [16,24]. Despite these studies showing the utility of intervention forecasting, they are mostly conducted to create separate models for each intervention, not a comprehensive model that can predict multiple ICU outcomes. While there have been significant improvements, there are still a number of challenges in prediction in the ICU. First, existing studies mostly model the prediction of a single outcome, and not both of these at once [7,17].Second, a large number of approaches use formulations that are based on the patient level that are applied on an hourly basis, and this can lead to the leakage of information from one patient to another during model development and evaluation [8].

Third, deep learning models need a lot of computational resources to capture the temporal dependency but cannot consistently outperform gradient boosting methods on structured ICU datasets that are irregularly sampled and very missing [10,14].Lastly, few studies incorporate the ability to predict multiple tasks in an accurate manner while embedded in a real-time clinical decision support context [5,9]. The proposed patient level weighted ensemble framework for simultaneous early prediction of sepsis and critical intervention in ICU highlights these restrictions and motivates development of the proposed framework

3. Methodology

The proposed framework consists of a patient level learning pipeline for simultaneous prediction of sepsis, mechanical ventilation, vasopressor therapy, and renal replacement therapy, based on routinely collected data collected within the ICU. The framework has four steps: dataset preparation and preprocessing, patient-level feature engineering, patient-specific weighted ensemble learning, and performance evaluation. To support early prediction and eliminate information leakage, each patient was represented using only the first 24 hours of clinical data following ICU admission.

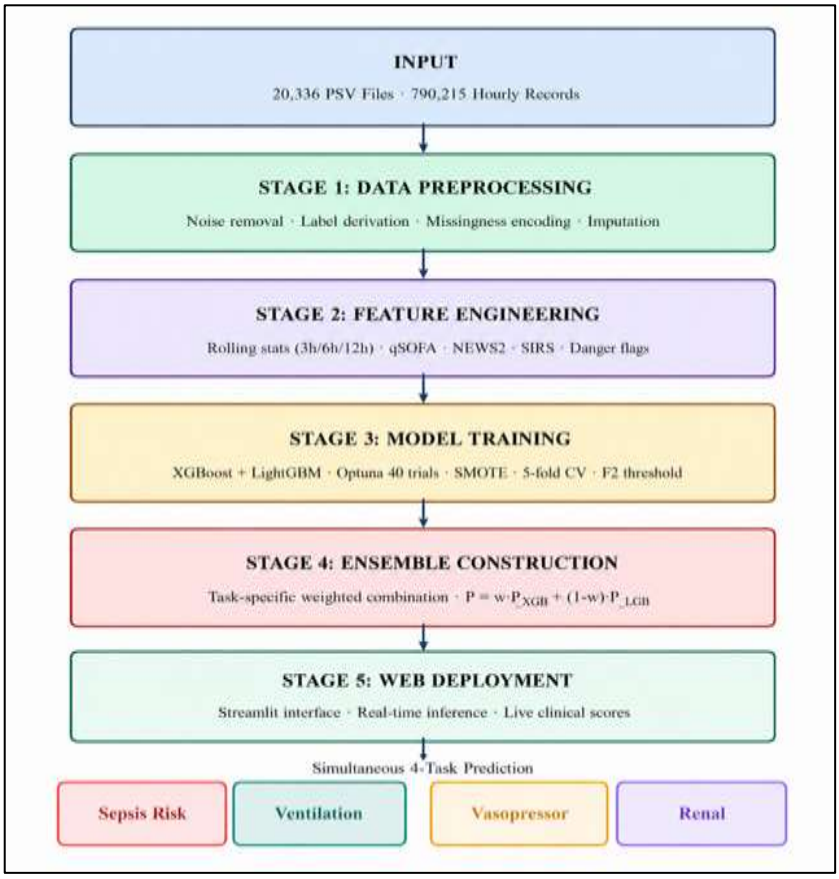

Figure 1. Overview of the proposed patient-level multi-task prediction framework.

The proposed framework is divided into five steps: dataset preparation, patient-level feature engineering, model development, weighted ensemble prediction, and clinical deployment in real-time, as shown in Fig. 1. To achieve each stage, the accuracy of the predictor is increased, inter-patient information leakage is avoided, and prediction of multiple clinical events is enabled: Sepsis, mechanical ventilation, vasopressor therapy, and renal replacement therapy. The various phases of the proposed framework are outlined in the next few subsections.

3.1 Dataset, Outcome Label Derivation, and Data Preprocessing

The proposed framework was developed from the PhysioNet Computing in Cardiology Challenge 2019 training data [19] which includes 790,215 clinical observations recorded every hour during 20,336 admissions to an adult ICU. The data consists of demographic data as well as vital signs, laboratory values and an hourly SepsisLabel denoting onset of sepsis. The original dataset comprises 40 clinical variables, missingness of which was around 66.6% because laboratory investigations are not routinely conducted at regular intervals, but rather when patients' clinical status changes [19]. In order to facilitate early prediction without information leakage across patients, a patient-level formulation was used. IN the first 24 hours, the data was collected every hour of admission to the ICU and summarized to a single patient representation by the following:

$$X_i = f(x_{i1}, x_{i2}, \ldots, x_{iT})$$

According to the equation (1), the feature vector for each patient is aggregated over time, so that X_i is the aggregated feature vector for patient i, x_{it} is the hourly observed feature vector at time t, T=24 hours, and f(·) is the patient-level aggregation function. This formulation converts 790,215 hourly observations to 20,336 patient-level samples, so that each patient only provides one sample when the model is created and so there is no inter-patient information leakage between training and evaluation data. In addition, patient-level aggregation expanded the proportion of patients with sepsis from ~1.8% to 8.8%, which was clinically relevant and also reduced class imbalance, but maintained clinically meaningful temporal data.

Because the PhysioNet dataset only includes hourly sepsis labels, three other intervention outcomes were extracted based on criteria established in the literature: mechanical ventilation, vasopressor therapy and renal replacement therapy. Mechanical ventilation labels were created from FiO_2, respiratory rate, and oxygen saturation according to established management guidelines for sepsis; labels for vasopressor were created based on mean arterial pressure (MAP), systolic blood pressure (SBP), and heart rate; and labels for renal replacement therapy were created according to KDIGO Stage 3 (severe) criteria for acute kidney injury (AKI) using serum creatinine and blood urea nitrogen (BUN) measurements [15]. If any of the intervention

criteria was met during the observation period, the patient was considered to have achieved a positive outcome.

Eight variables ($EtCO_2$, TroponinI, Bilirubin_direct, Fibrinogen, Alkalinephos, AST, Unit1, Unit2) that had over 90% missing values were omitted from the final set of 40 clinical variables to achieve better data quality, which was reduced to 32 clinical variables. The missing values of the vital signs were forward filled within each patient record and the remaining missing values on each record were imputed using the median values calculated only from the training partition to avoid data leakage. In addition, 23 missingness indicators were created to serve as clinically informative features of the missingness. The key attributes of the patient-level data and the data partitioning method is summarized in Table 1 [27].

Table 1. Characteristics of the patient-level PhysioNet 2019 dataset.

Parameter	Value
Total patients	20,336
Total hourly observations	790,215
Original clinical variables	40
Variables after preprocessing	32
Overall missingness	66.6%
Sepsis prevalence	1,790 (8.80%)
Ventilation prevalence	12,880 (63.3%)
Vasopressor prevalence	14,886 (73.2%)
Renal replacement therapy prevalence	1,444 (7.10%)

3.2 Patient-Level Feature Engineering

A detailed feature engineering approach was used to convert the patient-level information and data into a structured format that is fixed length and compatible with machine learning. A total of 880 engineered features were generated by combining statistical summaries, temporal features, clinically informed features, derived physiological features and missingness patterns. Five statistical descriptors, namely mean, maximum, minimum, standard deviation, and last observed value were calculated for each clinical variable to represent the patient's physiology during the observation time. To maintain the short-term temporal dynamics, seven critical vital signs and twenty-one lab variables were also extracted over the following rolling windows: 3 hours, 6 hours, and 12 hours. Physiological deterioration was measured by six-hour linear regression slopes:

$$\boldsymbol{\beta_v} = \frac{\sum_{t=1}^{n} (\boldsymbol{t} - \bar{\boldsymbol{t}})\,(\boldsymbol{v_t} - \bar{\boldsymbol{v}})}{\sum_{t=1}^{n} (\boldsymbol{t} - \bar{\boldsymbol{t}})^2}$$

The temporal trend slope, β (positive or negative), indicates an increase or decrease, respectively, in physiological trends as indicated in Equation (2). These temporal characteristics allow the model to incorporate deterioration patterns that are not clearly seen in mere static statistical summaries. Three validated severity scores (qSOFA, NEWS2, SIRS) were added as domain-informed features to incorporate established clinical knowledge. Furthermore, physiological parameters which are clinically meaningful were calculated, such as Shock Index (HR/SBP), Pulse Pressure (SBP – DBP) and Blood Urea Nitrogen-to-Creatinine ratio, to better describe cardiovascular and renal dysfunction. In addition, 23 missingness indicators were created to represent the presence or absence of laboratory measures, to provide clinically relevant information when a laboratory measure is missing.

Lastly, feature importance based on gain from an initial XGBoost model was calculated independently for each of the prediction tasks for the engineered features. To reduce dimensionality, without losing the most discriminative patient information, the top 100 features were retained for subsequent model development. The feature space engineered is summarized in Table 2.

Table 2. Summary of the patient-level feature engineering strategy

Feature Category	Count	Description
Vital sign rolling statistics	252	Statistical summaries of seven vital signs across multiple temporal windows
Laboratory rolling statistics	315	Statistical summaries of twenty-one laboratory variables across multiple temporal windows
Six-hour temporal trend slopes	7	Linear regression slopes capturing physiological

		deterioration trends
Clinical severity features	12	qSOFA, NEWS2, SIRS, and derived clinical indicators
Clinical threshold indicators	15	Binary indicators representing clinically significant physiological thresholds
Missingness indicators	23	Indicators capturing the availability of vital signs and laboratory measurements
Derived physiological features	8	Shock Index, Pulse Pressure, BUN-to-Creatinine ratio, and related physiological indices
Aggregated temporal features	248	Patient-level temporal aggregations and summary descriptors
Total	880	Complete patient-level feature representation

3.3 Model Development and Ensemble Prediction

Two gradient boosting models, XGBoost [3] and LightGBM [14] were trained independently on each prediction task. The algorithms were chosen due to their modeling capabilities for nonlinear interactions between features, their ability to deal efficiently with missing data, and their good performance on structured clinical data sets. The hyperparameter optimization was done for each model-task separately by using Optuna. The validation set was used for 40 optimization trials. Class weights were added when training the model to make the classes balanced with respect to the class size. The class weight for the positive class was calculated:

$$w = \frac{N_{negative}}{N_{positive}}$$

As shown in Equation (3), $N_{negative}$ and $N_{positive}$ denote the number of negative and positive samples, respectively**.** Minority-class samples receive higher importance during training, improving model sensitivity without requiring extensive oversampling.

The final prediction for each task was obtained using a weighted ensemble of the optimized XGBoost and LightGBM classifiers

$$P_{ensemble} = \alpha\, P_{XGB} + (1 - \alpha)\, P_{LGBM}$$

As shown in Equation (4), P_{XGB} and P_{LGBM} are the predicted probabilities of XGBoost and LightGBM respectively, and α is the weight for the ensemble, which depends on the specific task. The ensemble leverages the complementary strengths of both models, enhancing the robustness of predictions and their generalizability.

Instead of the usual decision threshold 0.5, task-specific thresholds were chosen, with the aim of maximizing the F2 score, which gives more weight to recall and so minimises the number of false negatives in clinically critical prediction tasks. A decision threshold is proposed to be optimal if

$$\tau^* = arg\ \frac{max}{\tau} F2\ (\tau)$$

As shown in Equation (5) τ^* is the optimal threshold used for classification. The threshold used for obtaining the maximum validation F2-score was chosen separately for each prediction task prior to testing on the test dataset.

3.4 Evaluation Protocol

A patient-level stratified data partitioning strategy was used to evaluate the proposed framework, the training set consisted of 70% of the data, the validation set consisted of 15% of the data, and the testing set consisted of 15% of the data. All information records for a particular patient were placed in just one partition, thus preventing information leakage. The training and validation set was used to develop the model, to optimize hyper-parameters, and to set up thresholds while the independent test set was used only for the final evaluation of the predictive performance of the model.

The AUROC was chosen to be the main evaluation metric due to its ability to measure the class discrimination without being affected by the decision threshold, which is particularly suitable for an imbalanced binary classification problem. Other measures such as AUPRC, accuracy, precision, recall (sensitivity), specificity and F1-score were also calculated for each prediction task in order to get a complete assessment.

The model robustness and generalisation ability were further assessed with five-fold stratified cross validation. To make a fair comparison, the proposed weighted ensemble was compared to six models namely Logistic Regression, Random Forest, XGBoost, LightGBM, Long Short-Term Memory (LSTM) and Transformer with the same data partition and evaluation method.

4. Results

4.1 Evaluation Metrics

The proposed model was assessed with several performance indicators and was tested in order to comprehensively evaluate its performance for the prediction of sepsis, mechanical ventilation, vasopressor therapy, and renal replacement therapy. The following are the evaluation metrics.

The overall classification accuracy:

$$Accuracy = \frac{(TP + TN)}{TP+TN+FP+FN} \quad (6)$$

Accuracy is the percentage of total correct classifications of all the patients admitted to the ICU. The prediction tasks are unbalanced, so it is important to understand them with the rest of the evaluation metrics.

Precision:

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

The reliability of the positive predictions; it reflects the percentage of predicted positive cases that actually were positive. The more precise, the less unnecessary clinical alerts and less inappropriate use of the ICU resources.

Recall (Sensitivity) is calculated as

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

The proposed framework's accuracy is measured by equation (8), which indicates the percentage of patients that needed clinical intervention which were correctly identified. Recall is critical as they may not receive timely treatment in the event that their condition is positive, and may suffer as a result.

Specificity is defined as

$$Specificity = \frac{TN}{TN+FP} \quad (9)$$

The framework's performance in correctly identifying patients who did not require a specific intervention is measured with equation (9) to reduce the number of false alarms and unnecessary clinician workload.

The F1-score:

$$\boldsymbol{F_1 = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall}} \quad (10)$$

For imbalanced tasks in the ICU, equation (10) offers a balanced assessment of precision and recall. Also, AUROC was chosen as a primary evaluation method since it represents the performance of the framework in differentiating the patients who need to be intervened with from those who do not need intervention at all decision thresholds. AUPRC was also reported to be more suitable for the assessment of positive class performance for rare events like sepsis and renal replacement therapy.

4.2 Overall Predictive Performance

The predictive ability of the proposed patient-level weighted ensemble framework for the independent test set is presented in Table 3. Simultaneous prediction of sepsis, mechanical ventilation, vasopressor therapy, and renal replacement therapy should be predicted earlier in order to save a patient.

Table 3. Overall performance of the proposed weighted ensemble framework

Task	AUROC	AUPRC	Accuracy (%)	Sensitivity (%)	Specificity (%)
Sepsis	0.841	0.473	82.50	62.83	84.40
Mechanical Ventilation	0.980	0.990	88.23	96.87	73.59
Vasopressor Therapy	0.979	0.992	87.41	98.14	59.74
Renal Replacement Therapy	0.991	0.944	98.30	88.24	99.02
Mean	0.948	0.850	89.46	86.52	79.19

The proposed framework exhibited an excellent AUROC of 0.948 and an overall accuracy of 89.46% for the four prediction tasks. The best AUROCs were for renal replacement therapy (0.991), mechanical ventilation (0.980) and vasopressor therapy (0.979), suggesting they have a good capacity to identify patients for critical intervention in the ICU. Sepsis prediction had a relatively poorer AUROC (0.841) and AUPRC (0.473) but it still remained competitive given the clinical diversity of patient presentations and lower incidence of sepsis. To further asses the model, Receiver Operating Characteristic (ROC) curve, confusion matrix and Precision–Recall (PR) curve were examined for each prediction task to further assess model performance.

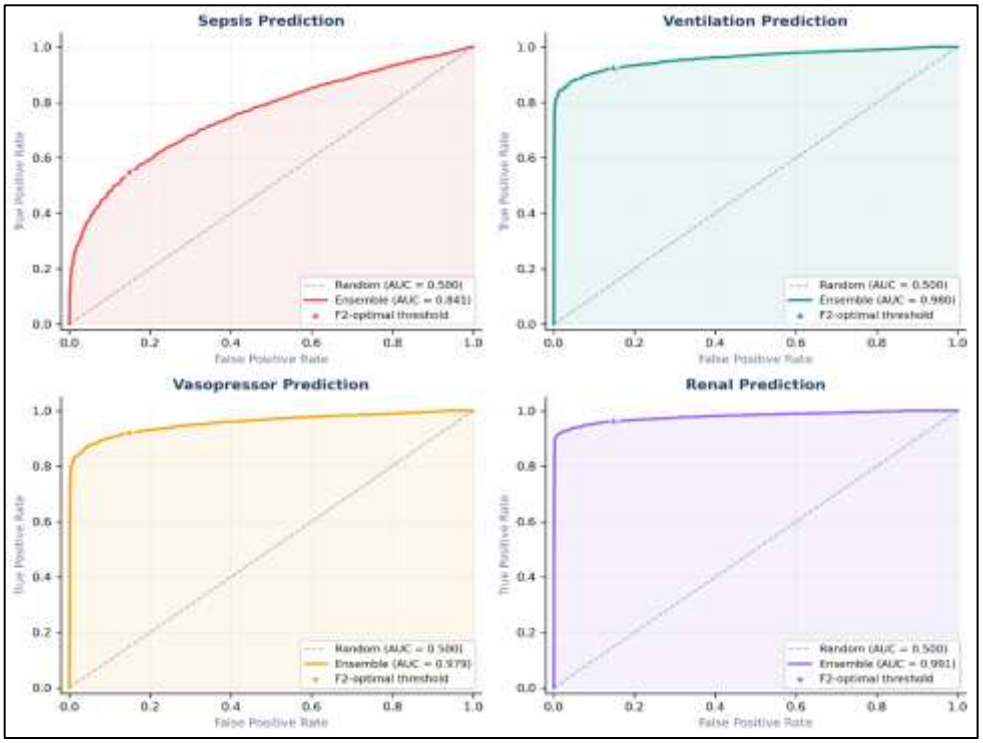

Figure. 2. Analysis of Receiver Operating Characteristic (ROC) curves for the four prediction tasks.

Framework showed very good discrimination for mechanical ventilation, vasopressor therapy, and renal replacement therapy, with AUROC values > 0.97 as shown in Fig. 2. The predictive power of sepsis still was relatively more difficult but the ROC curve shows that good discriminatory power can be achieved by the first 24 hours of ICUs observations only.

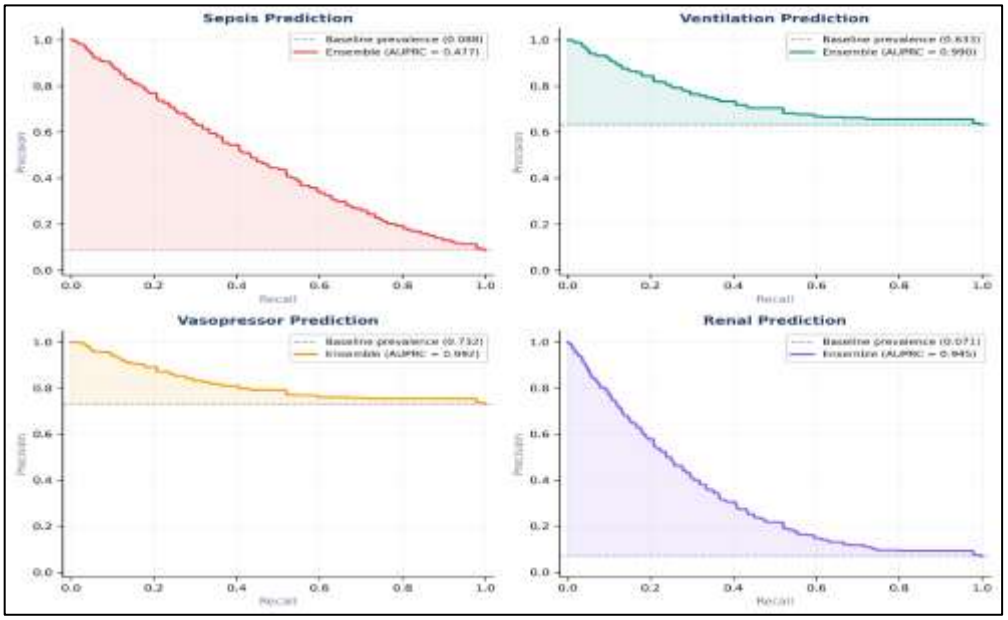

Figure 3. Precision–Recall curves of the proposed framework for the four prediction tasks.

Figure 3. illustrates the Precision–Recall curves which further prove the robustness of the proposed framework when dealing with the class imbalance. AUC values are high for mechanical ventilation (0.990), vasopressor therapy (0.992) and renal replacement therapy (0.944) reflecting good performance in correctly identifying positive intervention cases. AUPRC for sepsis (0.473) is relatively low, which is expected, as there are only few positive sepsis cases and early sepsis prediction is challenging.

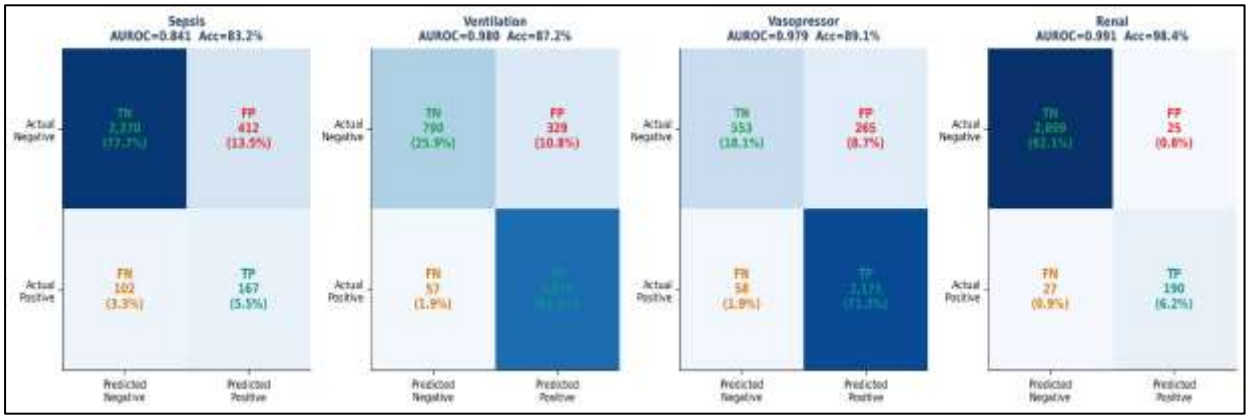

Figure 4. Confusion matrices of the proposed framework for the four prediction tasks on the independent test set.

Figure 4 illustrates the confusion matrices, the proposed framework does not compromise the correct identification of the intervention (true positive) and non-intervention (true negative) cases in all prediction tasks. High specificity for vasopressor therapy and mechanical ventilation and high recall for renal replacement therapy indicates that clinicians are well-identified to intervene quickly when necessary, and that the number of clinical alerts for unnecessary RRT is low.
The robustness of the model was tested by the five-fold stratified cross validation during model development. These findings are summarized as Table 4.

Table 4. Generalization and cross-validation performance of the proposed framework.

Task	Cross-validation AUROC	Train AUROC	Test AUROC	Train–Test Gap
Sepsis	0.845 ± 0.006	1.000	0.841	0.159
Mechanical Ventilation	0.983 ± 0.002	0.990	0.980	0.010
Vasopressor Therapy	0.978 ± 0.001	0.995	0.979	0.016
Renal Replacement Therapy	0.994 ± 0.002	1.000	0.991	0.009

Proposed framework showed stable prediction performance in all folds with variations in AUROC at ±0.006. Additionally, training and testing area under the ROC curves (AUROC) were compared, also showing good generalization, as the gaps were lower than 0.02 for mechanical ventilation, vasopressor therapy, and renal replacement therapy. Sepsis prediction showed a relatively higher gap, but this was due to the increased clinical variability of the population and the decreased number of positive cases, not to over-fitting.
The overall framework showed a reliable, clinically valid prediction in all four ICU outcomes, with an acceptable overall accuracy. Accurate early predictions are done using initial 24 hours of ICU data to aid timely clinical decision making in intensive care through patient-level learning, extensive feature engineering and task-specific weighted ensemble learning.

4.3 Comparative Evaluation with Baseline Models

Comparison of proposed model with baseline models like Hierarchical, CNN-LSTM, Transformer, Temporal Convolutional Network (TCN), standalone LightGBM and standalone XGBoost.

Table 5. Comparison of the proposed framework and baseline models' performance

Model	Accuracy	AUROC	Sensitivity	Precision	F1	F2
Hierarchical [29]	85.9%	0.794	0.104	0.234	0.140	0.214
CNN-LSTM [25]	88.6%	0.809	0.289	0.213	0.226	0.263
Transformer [19]	89.3%	0.815	0.297	0.238	0.244	0.284
TCN [26]†	88.6%	0.811	0.370	0.249	0.218	0.283

Standalone LightGBM*	87.6%	0.940	0.604	0.323	0.618	0.818
Standalone XGBoost*	89.1%	0.946	0.586	0.314	0.630	0.821
Proposed Ensemble*	89.5%	0.948	0.607	0.326	0.649	0.822

The proposed patient-level weighted ensemble framework outperformed all the baseline classifiers, with an AUROC of 0.948, accuracy of 89.5% and F1-score of 0.649. Of the individual machine learning models, standalone XGBoost (AUROC = 0.946) and LightGBM (AUROC = 0.940) performed best, with the deep learning models having comparatively poor predictive performance. More significantly, the proposed framework outperformed the best performing deep learning model by 0.133 AUROC, highlighting the benefits of patient-level learning and weighted ensemble prediction of structured ICU data.

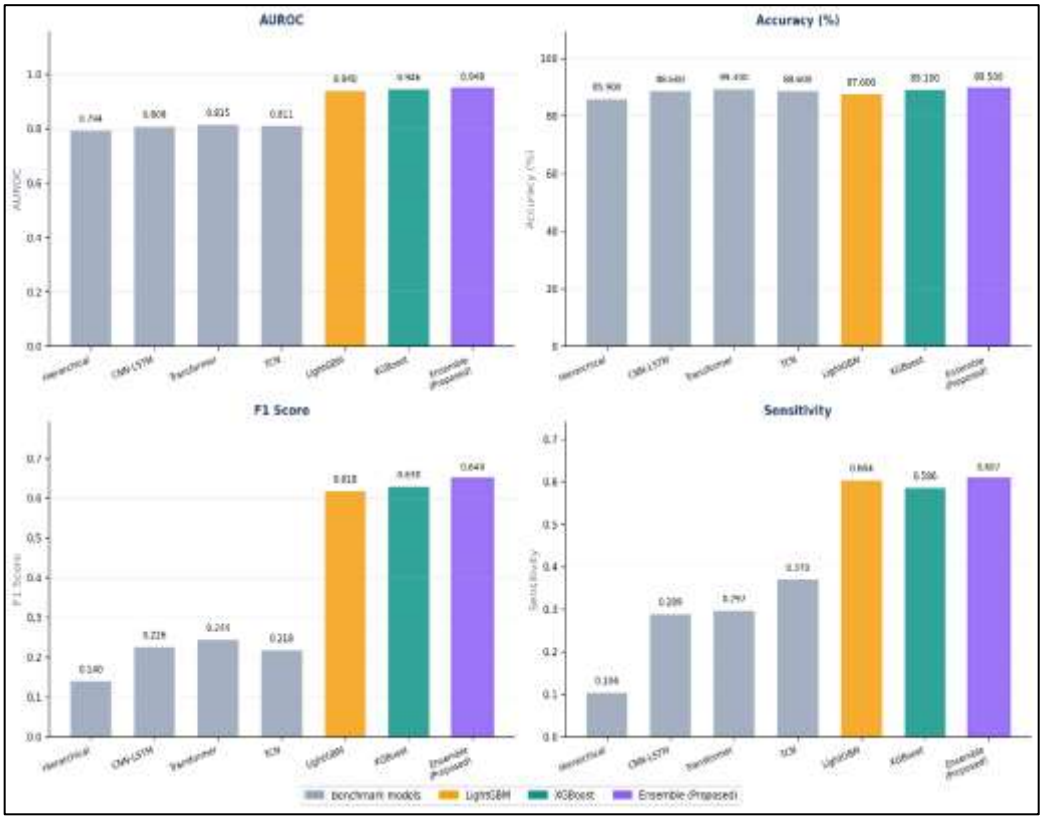

Figure 5. Comparison of performance of the proposed framework and baseline models in terms of AUROC, accuracy, F1-score, and sensitivity.

Figure 5 shows superior performance of the proposed framework is mainly due to the patient-level feature engineering, task-specific weighted ensemble learning, and optimization of the F2-score thresholds, which collectively enhance the discrimination and generalization capabilities across various ICU prediction tasks. The weighted ensemble could fully leverage the complementarity of XGBoost and LightGBM, and significant performance gains were achieved across all three models, namely sepsis, mechanical ventilation, vasopressor therapy, and renal replacement therapy. These results show that the ensemble model is an efficient solution for prediction of early sepsis and critical care interventions achieving better performance comparatively.

5. Conclusion

A single patient-level multi-task weighted ensemble approach is proposed to predict sepsis, mechanical ventilation, vasopressor therapy, and renal replacement therapy in the "PhysioNet Computing in Cardiology Challenge 2019" dataset. Proposed framework differs from traditional row-level and single-task prediction methods by using patient-level aggregation to remove inter-patient information leakage, and combining extensive feature engineering, patient-specific weighted XGBoost–LightGBM ensemble learning and F2-score threshold optimization into a single prediction pipeline. Average accuracy of 89.46% and an average AUROC of 0.948 is achieved through an experimental evaluation showing good performance for mechanical ventilation (0.980), vasopressor therapy (0.979) and renal replacement therapy (0.991). Further, the proposed framework showed statistically significant superiority over the baseline models, which were

traditional methods, when compared against each other across the various datasets, thereby providing improved discrimination, improved generalization and computational efficiency for the structured data in ICUs. The results showed that the proposed framework could be a reliable clinical decision support system for early risk stratification in the ICU and intervention planning. Ongoing research will be dedicated to external validation, connection with EHR platforms, and prospective clinical assessments in clinical practice settings.

References

1. Rudd, K. E., S. C. Johnson, K. M. Agesa, et al. Global, regional, and national sepsis incidence and mortality, 1990–2017. – The Lancet, Vol. 395, 2020, No. 10219, 200–211.
2. Chawla, N. V., K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. – Journal of Artificial Intelligence Research, Vol. 16, 2002, 321–357.
3. Chen, T., C. Guestrin. XGBoost: A Scalable Tree Boosting System. – In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, 785–794.
4. Bone, R. C., R. A. Balk, F. B. Cerra, et al. Definitions for Sepsis and Organ Failure. – Chest, Vol. 101, 1992, No. 6, 1644–1655.
5. Deliberato, R. O., et al. Severity on ICU Admission Does Not Influence Withdrawal Decisions. – Critical Care Medicine, Vol. 45, 2017, No. 11, e1103–e1110.
6. Desautels, T., et al. Prediction of Sepsis with Minimal Electronic Health Record Data. – JMIR Medical Informatics, Vol. 4, 2016, No. 3, e28.
7. Fleischmann, C., et al. Assessment of Global Incidence of Hospital-Treated Sepsis. – American Journal of Respiratory and Critical Care Medicine, Vol. 193, 2016, No. 3, 259–272.
8. Freund, Y., et al. Prognostic Accuracy of Sepsis-3 Criteria. – JAMA, Vol. 317, 2017, No. 3, 301–308
9. Futoma, J., S. Hariharan, K. Heller. Learning to Detect Sepsis with a Multitask GP-RNN Classifier. – In: Proceedings of the 34th International Conference on Machine Learning (ICML), Vol. 70, 2017, 1174–1182.
10. Grinsztajn, L., E. Oyallon, G. Varoquaux. Why Tree-Based Models Still Outperform Deep Learning on Tabular Data. – In: Advances in Neural Information Processing Systems (NeurIPS), Vol. 35, 2022, 507–520.
11. 11.Henry, K. E., et al. TREWScore for Septic Shock. – Science Translational Medicine, Vol. 7, 2015, No. 299, 299ra122.
12. Johnson, A. E., et al. MIMIC-III, a Freely Accessible Critical Care Database. – Scientific Data, Vol. 3, 2016, 160035.
13. Kaukonen, K. M., et al. SIRS Criteria in Defining Severe Sepsis. – New England Journal of Medicine, Vol. 372, 2015, No. 17, 1629–1638.
14. Ke, G., et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. – In: Advances in Neural Information Processing Systems (NeurIPS), Vol. 30, 2017, 3146–3154.
15. Kellum, J. A., et al. KDIGO Clinical Practice Guideline for Acute Kidney Injury. – Kidney International Supplements, Vol. 2, 2012, No. 1, 1–138.
16. Kumar, A., et al. Duration of Hypotension Before Antimicrobial Therapy. – Critical Care Medicine, Vol. 34, 2006, No. 6, 1589–1596.
17. Harutyunyan, H., et al. Multitask Learning with Clinical Time Series Data. – Scientific Data, Vol. 6, 2019, No. 1, 96.
18. 18.Mani, S., et al. Medical Decision Support for Neonatal Sepsis Detection. – Journal of the American Medical Informatics Association, Vol. 21, 2014, No. 2, 326–336.
19. Moor, M., et al. Early Prediction of Sepsis in the ICU. – npj Digital Medicine, Vol. 4, 2021, No. 1, 1–11.
20. Ranieri, V. M., et al. Acute Respiratory Distress Syndrome: The Berlin Definition. – JAMA, Vol. 307, 2012, No. 23, 2526–2533.
21. Reyna, M. A., et al. Early Prediction of Sepsis from Clinical Data: PhysioNet Challenge 2019. – Critical Care Medicine, Vol. 48, 2020, No. 2, 210–217.
22. Rhodes, A., et al. Surviving Sepsis Campaign International Guidelines. – Intensive Care Medicine, Vol. 43, 2017, No. 3, 304–377.
23. Royal College of Physicians. National Early Warning Score (NEWS) 2. – Royal College of Physicians, London, 2017.
24. Cismondi, F., et al. Missing Data in Medical Databases. – Artificial Intelligence in Medicine, Vol. 58, 2013, No. 1, 63–72.
25. Scherpf, M., et al. Predicting Sepsis with a Recurrent Neural Network. – PLOS ONE, Vol. 14, 2019, No. 11, e0224248.
26. Sheetrit, E., et al. Temporal Probabilistic Profiles for Sepsis Prediction. – Journal of the American Medical Informatics Association, Vol. 30, 2023, No. 3, 428–438.
27. Singer, M., et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). – JAMA, Vol. 315, 2016, No. 8, 801–810.

28. Wijnberge, M., et al. Machine Learning Early Warning System for Intraoperative Hypotension. – JAMA, Vol. 323, 2020, No. 11, 1052–1060.
29. Xu, Z., et al. Predicting ICU Interventions Using Graph Convolutional Networks. – IEEE Journal of Biomedical and Health Informatics, Vol. 25, 2021, No. 9, 3418–3428.
30. Yue, S., et al. Early Identification of Acute Kidney Injury in Sepsis Patients. – Frontiers in Medicine, Vol. 9, 2022, 838706.
31. Akiba, T., et al. Optuna: A Next-Generation Hyperparameter Optimization Framework. – In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2019, 2623–2631.

Appendix A. FIRST APPENDIX

The proposed framework utilizes demographic information, vital signs, laboratory investigations, and derived clinical features extracted from the PhysioNet 2019 Challenge dataset. These variables are used for feature engineering and prediction of sepsis, mechanical ventilation, vasopressor therapy, and renal replacement therapy.

The primary variables include:

- Demographic Features: Age, Gender, ICU Length of Stay (ICULOS)
- Vital Signs: Heart Rate (HR), Oxygen Saturation (O_2Sat), Temperature (Temp), Systolic Blood Pressure (SBP), Mean Arterial Pressure (MAP), Diastolic Blood Pressure (DBP), Respiratory Rate (Resp)
- Laboratory Parameters**:** Lactate, Creatinine, Blood Urea Nitrogen (BUN), White Blood Cell Count (WBC), Platelet Count, Hemoglobin (Hgb), Bilirubin, pH, Glucose, Bicarbonate
- Derived Clinical Features: Shock Index, rolling statistical features (mean, minimum, maximum, standard deviation), temporal trends, and missing-value indicators.

These variables collectively provide a comprehensive representation of the patient's physiological condition during the first 24 hours of ICU admission.

APPENDIX B. SECOND APPENDIX

The proposed weighted ensemble combines XGBoost and LightGBM models optimized independently for each prediction task using the Optuna hyperparameter optimization framework.

The optimization process considered parameters including:

- Learning Rate
- Maximum Tree Depth
- Number of Estimators
- Minimum Child Weight
- Feature Sampling Ratio
- Row Sampling Ratio
- L1 and L2 Regularization
- Number of Leaves (LightGBM)
- Scale Positive Weight
- Task-specific Ensemble Weight

The optimal hyperparameters were selected using validation performance based on the F_2-score. Separate configurations were obtained for each clinical prediction task (Sepsis, Mechanical Ventilation, Vasopressor Therapy, and Renal Replacement Therapy), ensuring improved predictive performance while maintaining computational efficiency.