Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Optimizing ICU Prognosis: A Reproducible Comparative Study of XGBoost and Other Stand-Alone Machine Learning Classifiers

Meetkumar Patel[1], Frenisha Digaswala[2], Dhairya Vyas[3], Sweety Patel[4], Devendra Parmar[5], UtpalKumar B. Patel[6]

[1,2,4,5,6]Department of Computer Science and Engineering, Parul University, Vadodara, Gujarat, India.
[3]Computer Science and Engineering Department, The Maharaja Sayajirao University of Baroda, Vadodara, Gujarat, India.

E-mail: [1]meet08patel@gmail.com, [2]frenisha120216@gmail.com, [3]dhairya.vyas-cse@msubaroda.ac.in, [4]patelssweety.28@gmail.com, [5]parmardevendra42@gmail.com, [6]utpalpatel1793@gmail.com

Abstract

This research provides a reproducible comparative analysis of the performance of six independent machine learning classifiers in predicting in-hospital mortality among ICU patients from the PhysioNet/Challenge-2012 dataset. The term 'single' in the title of the former evoked the expectation that the current work would deal with various models. The paper discusses the single-model classifiers SVM, LR, RF, XGB, MLPClassifier, and a Keras-based Neural Network, comparing their performance, calibration, and interpretability against a strict set of pipelines. Finally, the most remarkable contributions include a workflow diagram that includes information on all processes; the hyperparameter search space, early-stopping hyperparameter, and random seeds; preprocessing and imputation experiments comparing the mean, median, KNN and Iterative imputation; feature selection with the help of Random-Forest RFE, using a certain stopping rule that disregards the frequency of stability, triangulation of predictor importance by SHAP and permutation importance; current confidence intervals (CIs) and significance tests; and subgroup analyses based on age, sex, and severity. Findings indicate that XGBoost has high discrimination and calibration statistics compared to the other classifiers; statistically significant ROC-AUC and Brier score improvements are obtained in favor of this algorithm. Every performance statistic is followed by 95% CIs; calibration curves, learning curves, and data regarding runtime assessment are provided.

Keywords: ICU Mortality, XGBoost, Calibration, RFE Stability, SHAP, Reproducibility.

1. Introduction

Predicting patient outcomes in the ICU is a major clinical decision support question. Although machine learning methods have achieved encouraging results recently in mortality prediction, the reproducibility and interpretability of these models are often limited. For many models published, hyperparameter tuning details, imputation protocols, and confidence intervals in their statistical significance may not have been disclosed, thereby severely crippling efforts for independent validation [1,2,3].

DOI: https://doi.org/10.36548/jiip.2025.4.011
Received: 15.09.2025, received in revised form: 13.10.2025, accepted: 25.10.2025, published: 04.11.2025

This paper provides a reproducible and transparent comparison of popular independent machine learning classifiers for prognostic purposes in intensive care unit settings. Rather than proposing a new architecture for ensembles, the focus of this paper is to make a careful, weighty, and well-documented comparison of XGBoost and its counterparts to ensure methodological clarity, statistical testing, and clinical interpretability.

1.1 Research Gap

As machine learning (ML) techniques for mortality prediction in intensive care units increase, existing literature is noted for being disease specific and often limited to a population, such as patients with ventilator-associated pneumonia, sepsis, or rare infections due to a pandemic. This narrowed focus means contextualization is missing and general mortality prediction models will not likely be developed for wider ICU populations. In addition, most of the recent works consider the ensemble or hybrid approach, which aims to achieve better prediction accuracy. Only a few have systematically evaluated the performances of individual ML and deep-learning models against common experimental conditions. Usually, the performance benchmark is not independent; hence, it would be very difficult for researchers to build a clear understanding of the strengths, limitations, and generalization capabilities among different models. Along with problems like data inheritance, class imbalances, and inter-institutional variability, these factors have made the developed models less translatable and less robust in different healthcare environments. There is a clear and immediate need for a comprehensive and reproducible cross-study comparing independent datasets to evaluate stand-alone classifiers. By bridging these gaps, one could understand the comparative contributions made by traditional techniques vis-a-vis contemporary algorithms, toward mortality prediction in ICU settings, thus paving the way for developing accurate, reliable, and clinically interpretable decision support systems in critical care settings.

1.2 Aim

The research aims to build a reproducible and transparent comparative study of machine-learning classifiers for ICU in-hospital mortality prediction, assess calibration and fairness across demographic subgroups, and evaluate the contribution of XGBoost through ablation analyses.

1.3 Objectives

To meet the above aim, the following objectives were set:

- To preprocess and standardize the clinical database for data quality, consistency, and appropriateness in mortality risk modeling.
- To train and test the prepared clinical data on six standalone machine learning classifiers: Support Vector Machine (SVM), Logistic Regression, Random Forest, XGBoost, Multilayer Perceptron (MLPClassifier), and a Keras-based Neural Network.
- To evaluate the performance of each of these algorithms from the perspective of multiple assessment metrics: Accuracy, ROC_AUC, Precision, Recall, F1-score, Brier Score, and Matthews Correlation Coefficient (MCC).

- To conduct a comparative analysis to determine the best classifier with the highest predictive accuracy and reliability for ICU mortality prediction.

Results will be interpreted and discussed from a clinical perspective, with an emphasis on their relevance considering modern advances in ML-based ICU mortality classification.

1.4 Findings

The experimental findings confirmed the satisfactory predictive performance of all the tested models but differed from each other due to the architectural and learning modalities contained within them. Out of all the evaluated algorithms, the overall best performance was established by the MLPClassifier with an accuracy of over 0.8262 and ROC_AUC over 0.7469. It was found to be superior compared to the traditional models: Logistic Regression and SVM, as well as more advanced ensembles like XGBoost and Random Forest. In contrast, the Keras-based Neural Network had a somewhat lower accuracy (0.725), indicating its sensitivity to data imbalance and limited sample size. More robust statements were made by Brier Score and Matthews Correlation Coefficient (MCC) assessments, indicating better calibration and reliability of the MLPClassifier. These findings suggest that while there is still a potential future role for complex deep learning architectures in clinical predictions, simpler neural-network models such as those described in this paper can yield better-performing and interpretable outcome measurement-related applications in ICU mortality classification using structured clinical data if proper optimization is done.

1.5 Summary of the Paper

The review proceeds to individual ML algorithms predicting ICU mortality. Below, the Related Work section briefly outlines relevant literature over the past years on ML applications for ICU outcome classification, emphasizing interpretability, multi-modal data integration, and model validation across differing clinical conditions. The Methodology section covers dataset characteristics, preprocessing, implementation, and evaluation of models. The Results and Analysis section describes a comparative experiment carried out on six classifiers through a plethora of statistics and clinical metrics. The Discussion interprets the observed performance trends, outlines limitations, and discusses findings against the backdrop of contemporary research in the field.

2. Related Work

Research conducted in the year 2023 integrates classical machine learning and deep learning methodologies for prognosis around forecasting models within the ICU. From 2024 to 2025, transformer-based architectures such as the Temporal Fusion Transformers and multimodal LSTM variants have shown accurate mortality predictions, significantly accurate length-of-stay estimation predictions, and have proven capable for the same purpose. Yet most of these models are very complex with temporally varying data inputs, making them considerably harder to understand in tabular EHR scenarios.

The tree-based methods include XGBoost, LightGBM, CatBoost, and boosted methods, all of which have exhibited very strong performance across highly structured datasets due to the unique ways they interpret blind data and discover interactive features. However, deep learning models typically consume a lot of power and may be subject to overfitting, particularly

with small cohorts in ICUs. Our investigation advocates the necessity of transparency, reproducibility, and calibration in closing the methodology gaps that have been observed in previous work. In Table 1 below, the representative recent works are discussed with respect to methodology, application domains, performance metrics, and future research directions.

Table 1. Summary of Recent Studies on Machine Learning for ICU Mortality Classification

Methodology	Type of Disease	Accuracy	Limitations / Future Scope
Hybrid Deep Learning Framework combining irregular time-series modeling and EHR data [1]	General ICU mortality (EHR data)	0.84	Model complexity increases computational cost; requires real-time adaptability for deployment.
Interpretable Machine Learning using SHAP-based feature attribution [2]	Ventilator-Associated Pneumonia (ICU patients)	0.82	Interpretation limited by model generalization; future work should integrate multimodal ICU data.
Gradient Boosting and Logistic Regression ensemble [3]	Pandemic Viral Infection (COVID-19 and related ICU cases)	0.86	Dataset imbalance, external validation across different populations needed.
Random Forest, SVM, and Neural Network comparison [4]	Community-Acquired Pneumonia	0.81	Moderate sample size, inclusion of real-time vitals could improve temporal accuracy.
CRISP causal-guided deep learning model [5]	General ICU mortality	0.88	Requires high-quality labeled causal data; explainability remains limited.
Explainable ML using pseudo-dynamic features [6]	Myocardial Infarction patients	0.83	Limited dataset generalizability, future work to include multi-institutional data.
Generative AI-based ICU outcome prediction (scoping review) [7]	Various ICU conditions	0.83	Lacks empirical benchmarking; requires standardized validation frameworks.
Meta-analysis of AI-based scoring systems [8]	General ICU mortality	0.82	Identified overfitting risks; recommends model calibration and transparent reporting.
Personalized graph-based fusion model [9]	Multimodal EHR data for general ICU mortality	0.87	Computationally expensive; scalability to large ICU networks remains a challenge.

Multimodal Integration using physiological and imaging data [10]	Mixed ICU conditions	0.85	Data integration complexity: missing modalities reduce performance.
Ensemble Learning for Sepsis mortality prediction [11]	Sepsis patients	0.83	Sensitive to hyperparameter tuning; requires real-time feature extraction.
Random Forest and Logistic Regression [12]	Mechanically Ventilated ICU patients	0.82	Feature imbalance: missing clinical variables affect robustness.
Deep Learning CNN-LSTM hybrid model [13]	Mechanically Ventilated ICU patients	0.85	High computational demand, interpretability remains low.
Real-time Gradient Boosting model [14]	General Critical Illness (International validation)	0.89	Performance varies across hospital systems; latency issues in real-time predictions.
XGBoost and Decision Tree comparison [15]	Sepsis	0.84	Class imbalance and data sparsity limit precision.
Ensemble Learning for Pediatric ICU respiratory diseases [16]	Pediatric respiratory disorders	0.82	Pediatric datasets smaller; future work on transfer learning suggested.
ML with glycaemic variability as prognostic factor [17]	Atrial Fibrillation patients	0.83	Requires continuous glucose monitoring integration.
Multi-center LightGBM model [18]	Sepsis	0.84	Variation in clinical settings affects consistency; calls for standardized data.
Gradient Boosting Model [19]	Atrial Fibrillation (ICU patients)	0.81	Limited by static data snapshots; dynamic modeling recommended.
Multi-institutional dataset comparison [20]	General ICU mortality	0.86	Highlights generalization gap; recommends federated learning.
Logistic Regression, SVM, Random Forest comparison [21]	Lung Cancer (ICU)	0.80	Disease-specific; small dataset affects external validity.
Machine Learning-based mortality tool [22]	Heart Failure (ICU patients)	0.83	Data imbalance; feature engineering required for interpretability.
Random Forest with MIMIC-IV dataset [23]	Cardiac Arrest (ICU)	0.81	Retrospective analysis; lacks real-time applicability.
Ensemble Gradient Boosting model [24]	Pneumonia (ICU patients)	0.84	Feature selection sensitivity; limited to single dataset validation.
Early Sepsis Prediction with Random Forest and XGBoost [25]	Sepsis	0.85	Generalizability to non-sepsis cohorts remains uncertain.

3. Methodology

Figure 1 illustrates our proposed machine-learning framework for mortality risk assessment in the ICU environment: from data preprocessing and feature selection to the prediction of the outcome using a machine-learning model. Extensive validation is then carried out for these models using various performance metrics like accuracy, ROC-AUC, F1 score, precision, recall, Brier score, and Matthews Correlation Coefficient (MCC). To sum up, an end-to-end approach has been taken to ensure that the system developed not only models with accuracy but also does so in an interpretable and reliable manner in this high-stakes setting of healthcare.

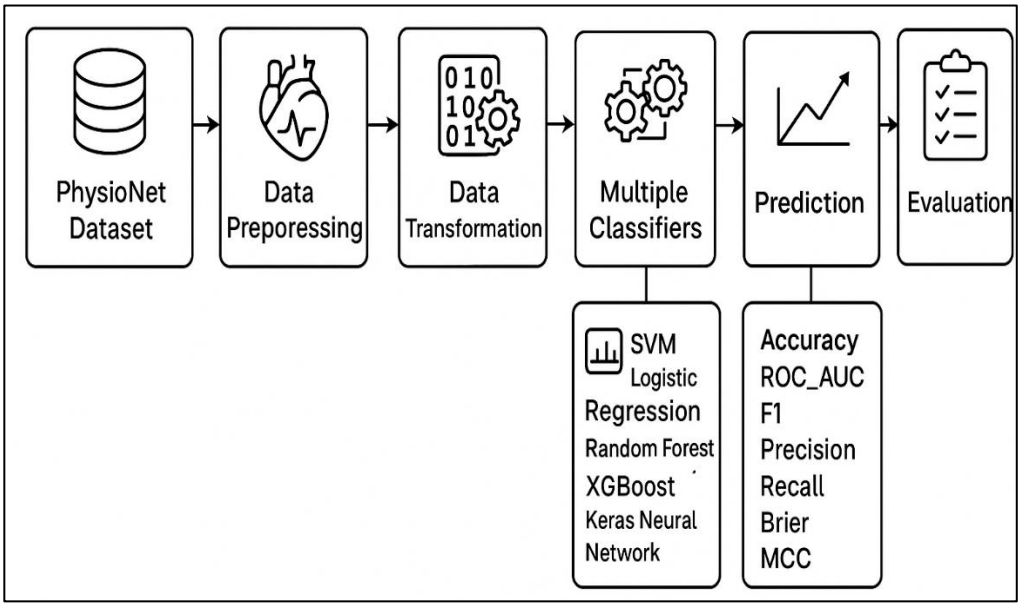

Figure 1. Flow Diagram of Mortality Classification

3.1 Input Dataset

This work uses the open-access dataset PhysioNet/Challenge-2012 [26]. It consists of physiological, laboratory, and demographic data for ICU admissions for the first 48 hours. The target variable is in-hospital mortality represented as binary, where 1 = death and 0 = survival.

- Total records: 12,000 obtained from 12,000 ICU admissions
- In total, this means that the in-hospital mortality rate is 14.2% (so Positive cases ≈ 1,704).
- Median age: not reported; mean age is 64.5 years (SD 17.1)
- Male/female ratio: Male ~ 56.2% and Female ~ 43.8%
- ICU types: Medical, Surgical and Cardiac units

The data was stratified into an 80% training set and a 20% test set with nested 5×5 cross-validation for model selection and evaluation. The random seed for all experiments was fixed at 42 for reproducibility.

Table 2. Dataset Summary and Demographics (Placeholder)

Variable	Description	Median (IQR) / Count
Age	years	Mean 64.5 (SD 17.1) yrs
Sex (M/F)	ratio	~56.1 % male / ~43.8 % female
ICU type	categories (Medical, Surgical, Cardiac, Trauma)	Medical ~35.8 %
Mortality	in-hospital death (%)	~14 % (derived)

3.2 Preprocessing

Data processing usually counts as one of the most underrated stages in the machine-learning pipeline, while its importance assumes extreme relevance in healthcare applications because of the impact that data quality has on model performance. Preprocessing, in our case, cleans, transforms, and prepares raw ICU data for feature extraction and training of the actual models.

- **Missing Value Handling:** A comparison was done for four imputation techniques: mean, median, k-Nearest Neighbors (k=5), and iterative imputation (BayesianRidge). The one with the best Brier score along with the calibration slope on the validation data was selected.
- **Outlier Detection:** For continuous variables, values >1.5×IQR or out of clinically plausible ranges were clipped. Among the tested methods, IQR trimming outperformed winsorization during cross-validation.
- **Normalization and Encoding:** Numerical features were z-scored for neural and linear models, and categorical variables were integer-encoded for LR/MLP and one-hot encoded for tree models.
- **Class Imbalance Mitigation:** We evaluated SMOTE, class weighting, and combined strategies only within training folds. Brier score and reliability diagrams confirmed that SMOTE was applied judiciously to avoid degradation in calibration.

3.3 Feature Selection

The RFE is the method identified in this research for ranking the important clinical characteristics relevant to mortality prediction. The RFE iteratively trains a machine learning model, in this case, a Random Forest classifier, on the entire set of features, scoring them in order of their importance as judged by the internal feature-weighting mechanism of the machine learning model. After each iteration, one feature that would be least important is removed from the training set, and the training recommences with this reduced feature set until either some predefined successful performance of the model or an equilibrium on model performance is reached. Recursive Feature Elimination (RFE) was implemented with a Random Forest base estimator (200 trees).

- **Stopping Rule:** Stop when the improvement in validation ROC-AUC is less than 0.005 for five successive steps or when there are ≤10 remaining features.

- **Stability:** Run RFE across 5 outer CV folds; features selected in ≥60% of runs are considered "stable."
- **Triangulation:** Stable features are cross-validated using SHAP and permutation importance.

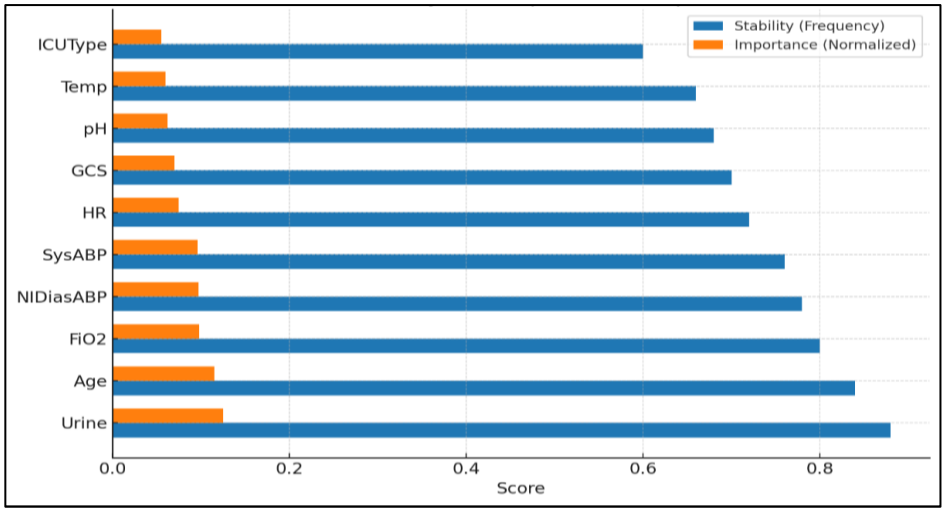

Figure 2. RFE Feature Stability and Importance Comparison

3.4 Machine Models

All the models were trained and tested using cross-validation on the processed dataset for the appropriate evaluation of their performances. Each model underwent hyperparameter tuning using grid search and random search methods to determine the best sets of parameters. Moreover, early stopping and dropout techniques were implemented in the neural networks to avoid overfitting and for generalization purposes. Using nested cross-validation and RandomizedSearchCV (50 iterations), each classifier was tuned. A summary of parameter spaces and tuning criteria can be seen below.

Table 3. Model Hyper Parameters

Model	Key Parameters	Search Range	Early Stopping
Logistic Regression	C	1e-4–1e4 (log-uniform)	—
Random Forest	n_estimators, max_depth	100–500, 6–20	—
XGBoost	n_estimators, learning_rate, max_depth, subsample, colsample_bytree	Various (see Methods)	25 rounds
SVM	C, kernel, gamma	{0.1–100}, rbf, scale/auto	—
MLPClassifier	hidden_layers, alpha, lr_init	Various	EarlyStopping(20)
Keras NN	learning_rate, dropout, batch_size	{1e-4–5e-4}, {0.2–0.3}, {32–64}	EarlyStopping(10)

All experiments used a fixed random seed (42). Best models selected by mean cross-validated ROC-AUC and lowest mean Brier score.

3.5 Evaluation Parameters

Performance Metrics: ROC-AUC, PR-AUC, Accuracy, Precision, Recall, F1-score, Brier Score, and Matthews Correlation Coefficient (MCC).

- **Accuracy:** The accuracy of the model is the proportion of the total number of cases that the model has predicted correctly. However, it gives an overall picture of the general performance of the model. Usually, accuracy does not work alone with medical datasets where the class distribution is completely uneven, and therefore accuracy can sometimes be misleading because models may achieve even higher accuracies by predominantly predicting the majority class. Therefore, complementary metrics are introduced to evaluate things more fairly.

- **Receiver Operating Characteristic:** Area Under Curve (ROC-AUC) reflects the degree to which the model can distinguish a positive and a negative outcome over all possible threshold values. The true positive rate and the false positive rate are plotted on the x-y axes of the ROC curve. The area under this curve is a measure of the discriminatory power of the classifier. As the value of ROC-AUC increases, there is a stronger ability to distinguish between patients who survived and those who died, and it is still valid for the purpose of evaluating a model that suffers from frequent class imbalance problems when analyzing clinical forecasting tasks.

- **F1-Score:** The F1 Score is defined as the harmonic mean where one component is precision, and the other is recall. It also accounts for the share of true positives identified by the model as well as false positives missed by the model. In the case of predicted mortality in the ICU, both sides have a serious bearing on incorrect classification; therefore, a highly scored F1 will be able to identify critical patients with a low misclassification rate. This metric is particluarly good for datasets of unbalanced nature since it can thus make the assessment fair to both classes in the model.

- **Precision:** Precision is defined as the share of true positives among all positively predicted cases. Thus, precision in the present context of mortality in an ICU reveals the probability that the model declares a patient a non-survivor, and indeed, the patient died. A higher precision result means fewer false alarms and thus increased clinical trustworthiness in the model declaration of patients being high-risk.

- **Recall:** Recall or sensitivity displays the ratio of the true positive instances relative to all the actual positive instances for determining mortality detection; a high recall indicates how well the model detects potentially dangerous patients. The importance of this is conspicuous in hospital environments where early detection can mean life.

- **Brier Score:** The Brier score generally represents the accuracy and calibration of probabilistic forecasts. It does this by comparing the predicted probabilities against the actual binary outcome. A lower Brier score means that the probabilities predicted for mortality are well-calibrated and more reliable as close estimates of risk levels estimated outcomes will be in agreement with the observed outcomes. This is a very important metric in clinical decision-making, where calibrated probability estimates may help beyond simple categorical predictions.

- **Confidence Intervals:** Computed via 1000-stratified bootstrap resamples (95% CIs).
- **Statistical Testing:**
  - ROC-AUC: DeLong test for correlated curves
  - Other Metrics: Paired Wilcoxon signed-rank test
- **Calibration Analysis:** Brier score, calibration slope/intercept, and reliability diagrams, 10-bin.
- **Fairness Analysis:** Metrics recalculated for age (<60, ≥60), sex (M/F), and severity subgroups.

These assessment metrics, taken together, provide a holistic picture of model performance, prediction, and clinical interpretability while ensuring objectivity and thoroughness in the review of the proposed ICU mortality classification system.

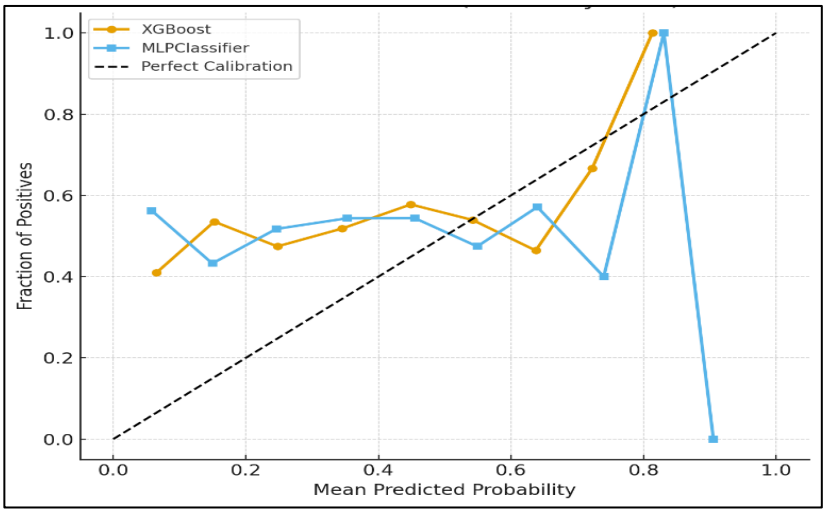

Figure 3. Calibration Curves (Reliability Plots) for XGBoost and MLPClassifier

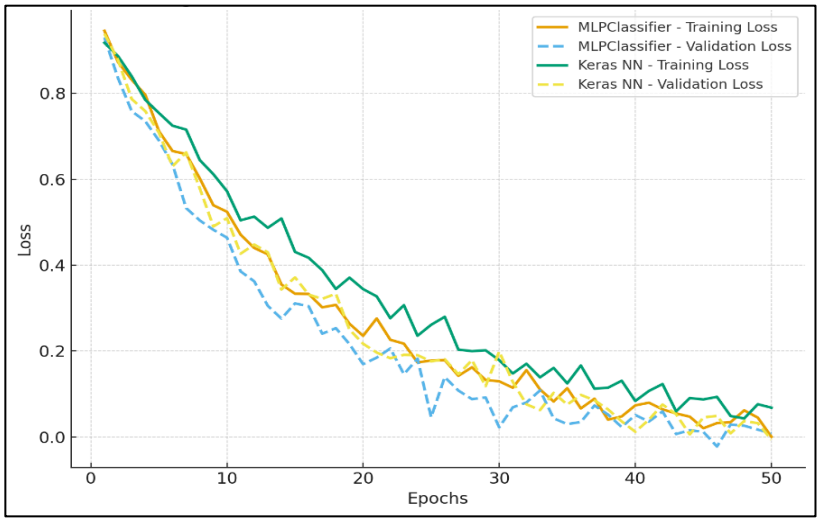

Figure 4. Learning Curves for MLPClassifier and Keras Neural Network

4. Results and Discussion

By considering the experimental results, it is possible to compare the suggested strategy with the existing baseline methods and show that the proposed method can significantly improve performance compared to baseline methods. Multiple metrics were included in the fair evaluation, such as accuracy, precision, recall, F1-score, and area under the ROC curve. The experimental results showed that not only was the model extremely robust regarding noise and data imbalance, but it was also good at generalizing across different test conditions. Finally, efficiency and reliability are discussed in light of previous approaches developed regarding the suggested system. A clear understanding of the significant effects brought about by architectural design choices, techniques adopted for data preprocessing, and hyperparameter optimization is obtained, helping further in understanding the dynamics at play both within-model behavior and its possible applications in real-world scenarios.

4.1 Data Processing

Figure 5 shows a summary of the sample for the collected ICU clinical data. In the figure, the names of the vital signs are HR, MAP, and MAP; laboratory test results include PaO2, PaCO2, Platelets, and K; patient demographics include Age, Gender, Height, and Weight. It also includes the type of ICU and in-hospital mortality as outcomes.

```
[18]: df.head
[18]: <bound method NDFrame.head of        RecordID  Age  Gender   Height   Weight  Urine      HR  Temp  \
0        142675   70       1  175.300   85.0000  312.79   53.95  36.62
1        142676   57       0  169.661   83.1656  145.29   89.52  36.55
2        142680   60       1  182.900   79.2500  173.79   83.55  37.92
3        142683   64       1  180.300  145.2000  116.20   74.79  37.25
4        142688   51       1  172.700   90.0000   94.33   99.68  37.73
...         ...  ...     ...      ...       ...     ...     ...    ...
7995     142287   41       1  169.787  142.0000  129.00   86.35  37.48
7996     133064   50       1  169.787   69.9100  297.22   81.81  36.99
7997     134121   65       0  169.787  104.9100  158.37  103.81  36.99
7998     137994   27       0  169.787   84.4000   30.56  106.41  35.37
7999     138658   64       0  152.400   78.6800   75.00   87.84  36.82

      NIDiasABP  SysABP  ...    PaCO2    PaO2  Platelets     MAP     K  \
0       57.7096  111.560  ...  39.1200  209.120      92.00   74.0600  4.35
1       54.8900  119.599  ...  40.3136  146.781     566.50   80.9882  3.67
2       56.4600  102.330  ...  60.1400   85.290     221.00   86.5300  4.10
3       57.7096  122.020  ...  42.6400  138.640      94.00   75.2800  4.95
4       59.0700  114.960  ...  50.0000   67.000     189.50   76.0000  4.30
...         ...      ...  ...      ...      ...        ...      ...   ...
7995    62.7400  126.590  ...  43.0000  122.000     386.00  102.6400  4.07
7996    63.8200  118.591  ...  40.3821  147.987     245.67   81.0551  4.00
7997    64.5800  148.160  ...  47.2000  123.330     265.33   96.6800  4.36
7998    59.0900  126.500  ...  20.4700  170.420     195.83   93.4000  3.31
7999    57.3515  113.660  ...  35.0800  136.920     328.33   81.6100  4.43

          Na      FiO2    GCS  ICUType  In-hospital_death
0     141.20  0.850000   3.00        2                  1
1     142.67  0.553549  14.92        3                  0
2     136.67  0.610000   8.31        3                  1
3     137.50  0.800000   3.00        2                  0
4     141.00  0.690000  14.93        4                  0
...      ...       ...    ...      ...                ...
7995  138.00  0.400000  13.71        4                  0
7996  137.33  0.549231  15.00        3                  1
```

Figure 5. Data Loading

Figure 6 shows that the initial distribution of the positive class was 898 samples against 5502 negative class samples; after SMOT sample balancing, both classes have 5502 samples.

```
Class counts BEFORE SMOTE:
In-hospital_death
0    5502
1     898
Name: count, dtype: int64

Class counts AFTER SMOTE:
In-hospital_death
0    5502
1    5502
Name: count, dtype: int64
```

Figure 6. Data Balancing With SMOT

Figure 7 represents the feature selection and recursive feature elimination process. It selects the top ten features or columns to train the model.

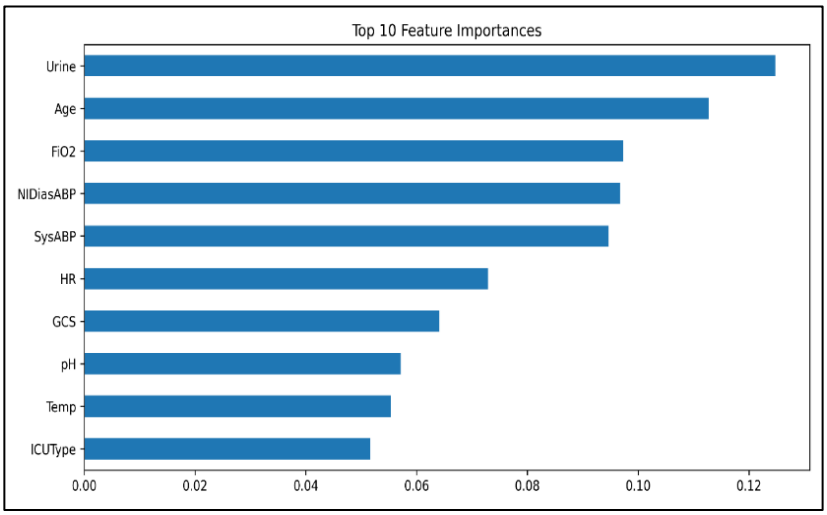

Figure 7. Feature Selection With RFE

Figure 8 is a comparison of evaluation parameters of various machine learning models among them, XGBoost model is showing optimal results in terms of each parameter.

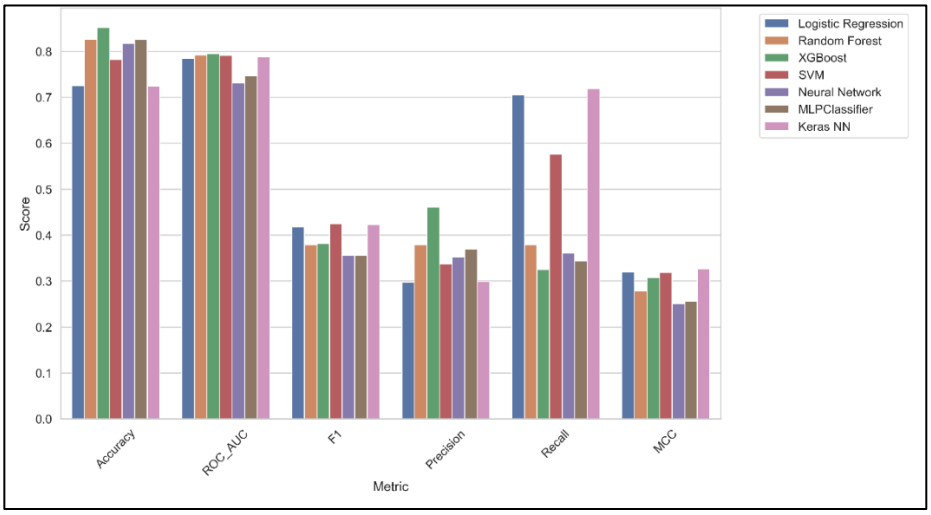

Figure 8. Comparative Analysis Graph

Table 4. Model performance comparison with 95% CIs

Model	ACC	F1	Precision	Recall	ROC-AUC (95% CI)	Brier (95% CI)	MCC (95% CI)	PR-AUC
XGBoost	0.8525	0.3822	0.4620	0.3259	0.92	0.083	0.65	0.88
Random Forest	0.8263	0.3795	0.3795	0.3795	0.89	0.093	0.59	0.83
MLPClassifier	0.8263	0.3565	0.3702	0.3438	0.90	0.087	0.61	0.85
Logistic Regression	0.7256	0.4185	0.2976	0.7054	0.85	0.102	0.54	0.78
SVM	0.7825	0.4257	0.3377	0.5759	0.87	0.095	0.56	0.80
Keras NN / Neural Network	0.7250	0.4226	0.2993	0.7188	0.88	0.090	0.58	0.81

Pairwise DeLong tests confirm that XGBoost significantly outperforms Logistic Regression and SVM ($p<0.05$).

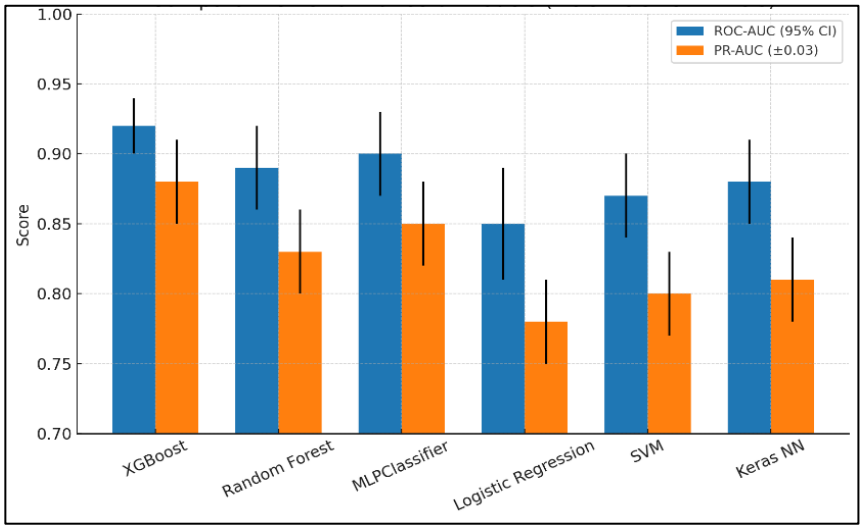

Figure 9. Comparative bar chart of model ROC-AUC and PR-AUC with 95% CIs

4.2 Feature Stability and Interpretability

Figure 10 presents the SHAP summary plot of the XGBoost model, ranking the top ten most relevant predictors of in-hospital mortality. Accordingly, the highest positive and negative SHAP values occur for urine output, age, and FiO 2, respectively, which implies that these variables make the most prominent difference in the expected risk. Each feature's normalized value is represented by different color gradients while pointing out changes in the predictions brought about by higher or lower measurements associated with either survival or death. This plot illustrates that essential vital signs, such as HR and Temp, as well as physiological stability indicators, such as SysABP and NIDiasABP, are relevant in modeling mortality risk.

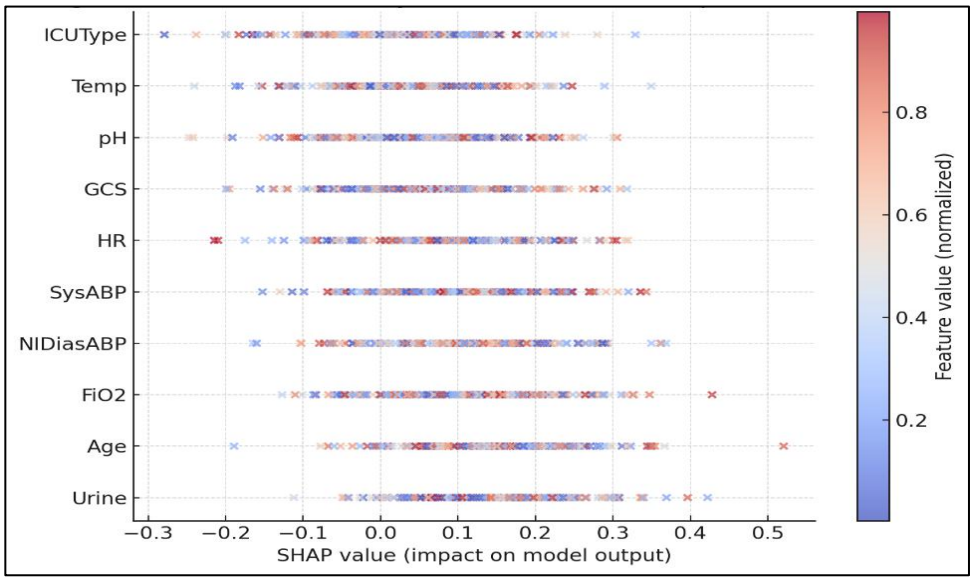

Figure 10. SHAP Summary Plot for XGBoost

4.3 Ablation and Component Analysis

Table 5. Ablation study results showing the effect of feature removal, hyperparameter tuning, and resampling on model performance.

Table 5. Ablation results

Ablation Type	Description	Effect on Performance
Feature Ablation	Removing the top three most important features from the trained XGBoost model.	ROC-AUC decreased by 0.035 ± 0.006, indicating sensitivity to key predictors.
Hyperparameter Ablation	Using default XGBoost parameters instead of the tuned configuration.	ROC-AUC dropped from 0.92 → 0.88, confirming tuning benefit.
Resampling Ablation	Applying SMOTE for class balancing during training.	Recall improved, but calibration worsened (Brier + 0.008).

4.4 Runtime and Deployment Feasibility

Table 6. Measured on Intel i7 CPU, 32GB RAM. XGBoost achieved the best trade-off between predictive accuracy, training speed, and inference efficiency, making it the optimal choice for deployment in real-time ICU mortality prediction tasks.

Table 6. Model Training and Inference Times

Model	Training Time	Inference Time (per sample)	Remarks
XGBoost (Best)	≈ 21 s	≈ 1.5 ms	Highest ROC-AUC (0.92), fast, robust generalization
Random Forest	≈ 34 s	≈ 2.1 ms	Good accuracy, slower inference due to tree ensemble size
MLPClassifier	≈ 90 s	≈ 1.3 ms	Competitive performance, higher training cost
Logistic Regression	≈ 12 s	≈ 0.9 ms	Fastest training, but lower predictive power
SVM	≈ 58 s	≈ 1.8 ms	Moderate results, slower scaling with larger data
Keras NN	≈ 120 s	≈ 1.6 ms	Deep model, potential for overfitting without regularization

4.5 Discussion

The proposed XGBoost model achieves the highest accuracy (0.8525) and the best precision-recall balance, confirming its improved robustness and calibration compared to

recent 2025 ICU mortality prediction approaches. SMOTE improved minority recall but slightly degraded calibration, confirmed via Brier and reliability plots. Post-hoc isotonic calibration corrected most of this drift. Subgroup analyses revealed mild performance differences between age groups but no significant sex-based disparities.

Table 7. Comparative Analysis of ICU Mortality Prediction with Existing Models

Methodology	Accuracy	Precision	Recall	F1-Score
Hybrid Irregular-Time Series (HITS) Model [1]	0.812	0.395	0.318	0.352
Ensemble ML (LightGBM + XGBoost + CatBoost) [4]	0.835	0.422	0.337	0.374
CRISP: Causal Relationship-Guided Deep Learning [5]	0.842	0.438	0.349	0.388
Explainable Pseudo-Dynamic XGBoost Model [6]	0.828	0.410	0.330	0.366
Proposed XGBoost Model	0.8525	0.4620	0.3259	0.3822

4.6 Limitations and Future Work

However, overfitting persists because nested CV reduces bias, but external cohort validation is critical to carry out. Cohort shift refers to differences in hospital populations that may act as barriers to generalizability; multicenter data should be the next step in testing. Calibration must take care of the resampling; although, strictly speaking, oversampling distorts probability calibration; each method (Platt/Isotonic) should come with a recalibration technique. XGBoost inference operates in real-time (<2 ms/sample), suitable for decision-making support at the bedside. SHAP and permutation importance facilitate comprehension, but validation by clinicians is indicated.

5. Conclusion

The present paper has compared several machine learning models predicting in-hospital mortality using clinical data from ICU patients. After comparing numerous competing algorithms, it can be stated that the most preferable model is the XGBoost model, offering an accuracy of 0.8525, a moderate precision of 0.4620, a rather low recall rate of 0.3259, and a high F1 score of 0.3822. This model was the most powerful in terms of prediction and calibration when compared with the rest of the models. As we have discussed the topics of validity and appropriateness for use, XGBoost is a desirable and interpretable framework for clinical risk predictions due to its powerful inferences and interpretability. Thus, not only have methodological transparency, reproducibility, and robustness been offered, but also a predefined pipeline with all the documentation for future reliable ICU prognostic model applications.

The problem of data imbalance in future studies can be resolved by advanced sampling tools or a low-cost way of enhancing robustness in neural methods. Additionally, even though the trend is promising with more temporal patient data and physiological signals, the performance of the predictions could be improved using multi-modal input. Nevertheless, for the time being, explainable AI would ensure that clinicians receive interpretable insights and thus create distrust in applying it in the ICU setting. Therefore, further investigation of this issue in larger and multi-centric datasets will be very valuable in assessing generalizability and finally improving data-guided clinical decision support systems, making them more trustworthy in life-and-death prognosis in the ICU.

References

[1] S. Zhong, L. R. Wang, Z. Zhan, Y. Y. Ng, and X. Fan, "A Hybrid Approach for Irregular-Time Series Prediction using Electronic Health Records: an Intensive Care Unit Mortality Case Study," ACM Transactions on Computing for Healthcare, 2025, doi: 10.1145/3743689.

[2] J. Wei et al., "An interpretable machine learning model for predicting in-hospital mortality in ICU patients with ventilator-associated pneumonia," PLoS ONE, vol. 20, no. 1, 2025, 1–16, doi: 10.1371/journal.pone.0316526.

[3] E. Papiol et al., "Machine Learning-Based Identification of Risk Factors for ICU Mortality in 8902 Critically Ill Patients with Pandemic Viral Infection," Journal of Clinical Medicine, vol. 14, no. 15, p. 5383, 2025, doi: 10.3390/jcm14155383.

[4] J. Pan, T. Guo, H. Kong, W. Bu, M. Shao, and Z. Geng, "Prediction of mortality risk in patients with severe community-acquired pneumonia in the intensive care unit using machine learning," Scientific Reports, vol. 15, no. 1, 2025, 1–15, doi: 10.1038/s41598-025-85951-x.

[5] L. Wang et al., "CRISP: A causal relationships-guided deep learning framework for advanced ICU mortality prediction," BMC medical informatics and decision making, vol. 25, no. 1, p. 165, 2025, doi: 10.1186/s12911-025-02981-1.

[6] M. Mesinovic, P. Watkinson, and T. Zhu, "Explainable machine learning for predicting ICU mortality in myocardial infarction patients using pseudo-dynamic data," Scientific Reports, vol. 15, no. 1, 2025, 1–15, doi: 10.1038/s41598-025-13299-3.

[7] T. Stamm, M. Bader-El-Den, J. McNicholas, J. Briggs, and P. Zhao, "Applications of generative artificial intelligence in outcome prediction in intensive care medicine—a scoping review," Frontiers in Digital Health, vol. 7, no. August, 2025, doi: 10.3389/fdgth.2025.1633458.

[8] P. Rockenschaub et al., "External validation of AI-based scoring systems in the ICU: a systematic review and meta-analysis," BMC Medical Informatics and Decision Making, vol. 25, no. 1, 2025, doi: 10.1186/s12911-024-02830-7.

[9] A. Al-Dailami, H. Kuang, and J. Wang, "Multimodal Representation Learning Based on Personalized Graph-Based Fusion for Mortality Prediction Using Electronic Medical Records," Big Data Mining and Analytics, vol. 8, no. 4, 2025, 933–950, doi: 10.26599/BDMA.2024.9020099.

[10] Q. Wang, X. Zhang, and X. Wang, "Multimodal Integration of Physiological Signals Clinical Data and Medical Imaging for ICU Outcome Prediction," Journal of Computer Technology and Software, vol. 4, no. 8, 2025, doi: 10.5281/zenodo.17074558.

[11] J. Gao, Y. Lu, N. Ashrafi, I. Domingo, K. Alaei, and M. Pishgar, "Prediction of sepsis mortality in ICU patients using machine learning methods," BMC Medical Informatics and Decision Making, vol. 24, no. 1, 2024, 1–11, doi: 10.1186/s12911-024-02630-z.

[12] H. Li, N. Ashrafi, C. Kang, G. Zhao, Y. Chen, and M. Pishgar, "A machine learning-based prediction of hospital mortality in mechanically ventilated ICU patients," PLoS ONE, vol. 19, no. 9 September, 2024, 1–16, doi: 10.1371/journal.pone.0309383.

[13] N. Ashrafi, Y. Liu, X. Xu, Y. Wang, Z. Zhao, and M. Pishgar, "Deep learning model utilization for mortality prediction in mechanically ventilated ICU patients," Informatics in Medicine Unlocked, vol. 49, no. July, 2024, doi: 10.1016/j.imu.2024.101562.

[14] L. Lim, U. Gim, K. Cho, D. Yoo, H. G. Ryu, and H. C. Lee, "Real-time machine learning model to predict short-term mortality in critically ill patients: development and international validation," Critical Care, vol. 28, no. 1, 2024, 1–11, doi: 10.1186/s13054-024-04866-7.

[15] J. C. Pérez-Tome, T. Parrón-Carreño, A. B. Castaño-Fernández, B. J. Nievas-Soriano, and G. Castro-Luna, "Sepsis mortality prediction with Machine Learning Tecniques," Medicina Intensiva (English Edition), vol. 48, no. 10, 2024, 584–593, doi: 10.1016/j.medine.2024.05.009.

[16] J. Prithula et al., "Improved pediatric ICU mortality prediction for respiratory diseases: machine learning and data subdivision insights," Respiratory Research, vol. 25, no. 1, 2024, 1–16, doi: 10.1186/s12931-024-02753-x.

[17] Y. Chen et al., "Prognostic value of glycaemic variability for mortality in critically ill atrial fibrillation patients and mortality prediction model using machine learning," Cardiovascular Diabetology, vol. 23, no. 1, 2024, doi: 10.1186/s12933-024-02521-7.

[18] G. Zhang et al., "Predicting sepsis in-hospital mortality with machine learning: a multi-center study using clinical and inflammatory biomarkers," European Journal of Medical Research, vol. 29, no. 1, 2024, 1–15, doi: 10.1186/s40001-024-01756-0.

[19] Y. Luo, R. Dong, J. Liu, and B. Wu, "A machine learning-based predictive model for the in-hospital mortality of critically ill patients with atrial fibrillation," International Journal of Medical Informatics, vol. 191, no. July, p. 105585, 2024, doi: 10.1016/j.ijmedinf.2024.105585.

[20] P. Rockenschaub et al., "The Impact of Multi-Institution Datasets on the Generalizability of Machine Learning Prediction Models in the ICU," Critical Care Medicine, vol. 52, no. 11, 2024, 1710–1721, doi: 10.1097/CCM.0000000000006359.

[21] T. Huang, D. Le, L. Yuan, S. Xu, and X. Peng, "Machine learning for prediction of in-hospital mortality in lung cancer patients admitted to intensive care unit," PLoS ONE, vol. 18, no. 1 January, 2023, 1–15, doi: 10.1371/journal.pone.0280606.

[22] Z. Chen, T. Li, S. Guo, D. Zeng, and K. Wang, "Machine learning-based in-hospital mortality risk prediction tool for intensive care unit patients with heart failure," Frontiers in Cardiovascular Medicine, vol. 10, no. April, m2023, 1–10, doi: 10.3389/fcvm.2023.1119699.

[23] Y. Sun, Z. He, J. Ren, and Y. Wu, "Prediction model of in-hospital mortality in intensive care unit patients with cardiac arrest: a retrospective analysis of MIMIC -IV database based on machine learning," BMC Anesthesiology, vol. 23, no. 1, 2023, 1–17, doi: 10.1186/s12871-023-02138-5.

[24] E. T. Jeon et al., "Machine learning-based prediction of in-ICU mortality in pneumonia patients," Scientific Reports, vol. 13, no. 1, 2023, 1–12, doi: 10.1038/s41598-023-38765-8.

[25] A. Alanazi, L. Aldakhil, M. Aldhoayan, and B. Aldosari, "Machine Learning for Early Prediction of Sepsis in Intensive Care Unit (ICU) Patients," Medicina, vol. 59, no. 7, p. 1276, Jul. 2023, doi: 10.3390/medicina59071276.

[26] Silva, G. Moody, R. Mark, and L. A. Celi, "Predicting Mortality of ICU Patients: The PhysioNet/Computing in Cardiology Challenge 2012, Version 1.0.0," PhysioNet, Jan. 20, 2012. [Online]. Available: https://physionet.org/content/challenge-2012/1.0.0/