Proceedings of CONF-APMM 2025 Symposium: Simulation and Theory of Differential-Integral Equation in Applied Physics
DOI: 10.54254/2753-8818/2025.DL27346

Performance Comparison of Models for Short-Term Mortality Prediction in ICU Patients

Junhe Wang

Department of Mathematics, University of Manchester, Manchester, the United Kingdom
wangjunhe1031@gmail.com

Abstract. Early prediction of mortality risk in intensive care unit (ICU) patients is critical for clinical decision-making, but existing prediction models still have room for improvement in analyzing short-term physiological indicators. This study, based on the MIMIC-IV Demo dataset, examines the performance differences of machine learning and deep learning models in short-term mortality prediction. Dynamic physiological indicators and static clinical data from the first 24 hours of admission were used to compare the predictive performance of XGBoost, Random Forest, and LSTM models for 7-day mortality in ICU patients. The results showed that XGBoost significantly outperformed other models in key metrics such as precision, AUC, and recall. However, LSTM's prediction performance was relatively weak due to limited data and a short observation window. The study suggests that XGBoost is more suitable for predicting short-term ICU mortality risk in current small-scale clinical data. This research provides a reference for the development of real-time clinical early warning systems. It also suggests that further improvements are needed in the future, including larger sample sizes and more optimized time series modeling methods, to enhance prediction accuracy.

Keywords: XGBoost., Random Forest, LSTM, ICU Mortality Prediction, Clinical Medicine.

1. Introduction

Critical care medicine has advanced significantly. Electronic medical records and bedside monitoring systems are now common in hospitals worldwide. These systems produce massive amounts of clinical data [1]. This data includes static admission information and dynamic physiological indicators. Together, they form a solid foundation for assessing patient condition and predicting prognosis [1]. The combination of dynamic and static data from early hospitalization is particularly valuable. It provides a rich information source for building predictive models.

Traditional medical tools like SOFA and APACHE II scores rely on manually selected indicators [2]. They offer some clinical interpretability. However, they struggle to capture the complex patterns within multidimensional time series data. Machine learning and deep learning provides new methods [3]. These methods can automatically learn patterns from large-scale data without manual intervention. They excel at modeling time series and nonlinear features.

Different algorithms have unique strengths for the same task. For example, long short-term memory networks (LSTMs) effectively capture long-term dependencies. Random Forest models

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2025.DL27346

handle data with complex interactions robustly. XGBoost is highly capable of modeling nonlinear relationships. Therefore, comparing different models on a single clinical task offers a crucial reference for model selection.

This study used the MIMIC-IV Demo dataset [4]. It combined dynamic measurements and static features from the first 24 hours of an ICU stay. Three prediction models: LSTM, Random Forest, and XGBoost have been built and compared. For the LSTM model, the original time series to capture temporal patterns has been used. For the tree-based models, statistical summaries of the data has been used to model nonlinear relationships. Finally, the performance of the three models was evaluated and compared on the task of predicting 7-day mortality.

Accurately identifying high-risk patients early is very important for critical care. This capability helps doctors create personalized treatment plans. It also supports the wise use of medical resources and can improve survival rates. This study makes three main contributions. First, it provides a systematic comparison of models for predicting 7-day mortality. Second, it identifies the approach with the highest clinical applicability. Third, it generates evidence to support early warning systems. These systems can facilitate timely intervention and guide clinical decisions. The findings offer clear value for clinicians and help guide future research.

2. Methodology

2.1. Extreme Gradient Boosting model

Extreme Gradient Boosting (XGBoost) is an algorithm based on the Gradient Boosting Decision Tree (GBDT) framework [5]. GBDT itself is an additive model. It is built upon the boosting ensemble concept. The training process uses a forward stagewise approach. This approach optimizes the model step by step [5]. Each iteration in this process fits a new CART regression tree. This new tree specifically targets the residuals from the previous model. The residuals represent the differences between the predictions from the prior t-1 iterations and the actual target values. By focusing on these errors, the model continuously reduces its overall prediction error. This method leads to a gradual and significant improvement in the model's predictive performance.

XGBoost is an optimized implementation of gradient boosting that improves upon GBDT by incorporating regularization and second-order gradient information [6]. Its objective function is optimized iteratively by adding trees that fit the gradients of the loss function. The objective function consists of two parts: loss function and regularization term.

$$L\left(\phi\right) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega\left(f_k\right), \tag{1}$$

where

$$\Omega\left(f\right) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2. \tag{2}$$

In the objective function, a convex loss function ℓ is used here. It is differentiable. This function quantifies the prediction error. Specifically, it measures how much the predicted value $\hat{y}_i$ differs from the true value y_i. The second term Ω governs the model complexity. It specifically determines

DOI: 10.54254/2753-8818/2025.DL27346

the architecture of the regression trees [6]. The objective function optimization leverages both the first- and second-order derivatives of the loss function, improving accuracy and convergence. XGBoost supports parallelization, which is a major strength. Although trees are built sequentially, nodes at the same level can be processed in parallel, and candidate split gains for a given node are computed using multithreading, resulting in faster training.

2.2. Random forest model

Random forest is an ensemble learning method. Its core concept is to improve predictive performance and stability by combining multiple decision trees [7]. A random forest constructs multiple decision trees. When a prediction is needed for a particular sample, the predictions of each tree in the forest are tallied, and a final result is selected from these predictions through voting. Suppose there are N decision trees, and the prediction of the i -th tree for an input x is $\hat{y}_i(x)$. The final output of a random forest is obtained through majority voting.

$$\hat{y}_{RF}(x) = mode(\hat{y}_1(x), \hat{y}_2(x), \ldots, \hat{y}_N(x)), \tag{3}$$

where $\mathrm{mode}()$ denotes the class that appears most frequently.

Random forests incorporate two types of randomness. Sample randomness involves sampling the training data with replacement to generate different subsets for training multiple decision trees. Feature randomness involves randomly selecting only a subset of features for decision-making when splitting a node in each tree [7]. This increases tree diversity and model robustness.

2.3. Long Short-Term Memory model

The Long Short-Term Memory (LSTM) network is a specialized type of Recurrent Neural Network (RNN). RNNs belong to a specific category of neural networks. They are designed to process sequential data. This design allows them to model time series or other ordered information. RNNs have a recurrent structure. This structure transfers hidden information from one time step to the next. Consequently, RNNs can retain historical context. LSTMs represent an improvement over simple RNNs. They introduce a concept known as the gating mechanism [8]. Each LSTM unit contains three specific gates. These gates are the input gate, the forget gate, and the output gate [8]. The three gates in an LSTM manage information flow. The forget gate:

$$\mathrm{f_t} = \sigma(\mathrm{W_f} \cdot [\mathrm{h_{t-1}}, \mathrm{x_t}] + \mathrm{b_f}), \tag{4}$$

determines how much of the prior cell state Ct-1 to keep. It makes this decision using the previous hidden state ht-1 and the current input Xt. Weights Wf, bias bf, and the sigmoid function σ are used in this process. The input gate

$$\mathrm{i_t} = \sigma(\mathrm{W_i} \cdot [\mathrm{h_{t-1}}, \mathrm{x_t}] + \mathrm{b_i}), \tag{5}$$

controls how much new information is written into the cell state. The output gate

DOI: 10.54254/2753-8818/2025.DL27346

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), \tag{6}$$

controls the amount of information that flows from the cell state to the hidden state.

These gates decide what data should be retained and determine what should be discarded. LSTMs control the flow of information through these gates, allowing them to better handle long-term dependencies. This effectively avoids the vanishing or exploding gradient problem in RNNs [9].

3. Result

This study used the MIMIC-IV Demo dataset to build a prediction task. The goal was to predict whether a patient would die within 7 days, based on dynamic physiological measurements from the first 24 hours after admission and static demographic information. Data preprocessing included filtering hospital records and vital signs, standardizing timestamps, dealing with incomplete data, and converting categorical variables. The dynamic time series features consisted of 6 vital signs and 4 laboratory tests, aggregated hourly to form a 24-hour × 10-feature matrix [4]. Missing values were filled using forward and backward filling, with remaining gaps replaced by zeros. Static features included demographic variables and admission information.

For the traditional machine learning approach, XGBoost and Random Forest were used with optimized hyperparameters. For deep learning, a two-layer LSTM network was implemented. All models were evaluated on an independent test set using AUC, Accuracy, Precision, Recall, and the optimized F1-score. Thresholds were adjusted based on the maximum F1-score. The following results section presents the performance and comparison of the three models.

3.1. Xgboost model

The XGBoost model demonstrated perfect classification performance on the test set, with both an AUC and F1 score of 1.0000. In Figure1, the confusion matrix shows that all 83 test cases were correctly classified, with no false positives. The model predicted "death" for 81 cases on the test set, all true negatives. It also predicted "survival" for two cases, all true positives. The ROC curve shows an AUC value of 1.0000, indicating perfect classification performance. The curve goes from the lower left corner (0,0) directly to (0,1) and then to (1,1), forming a perfect right angle. This demonstrates that the model is able to perfectly distinguish between positive and negative classes.

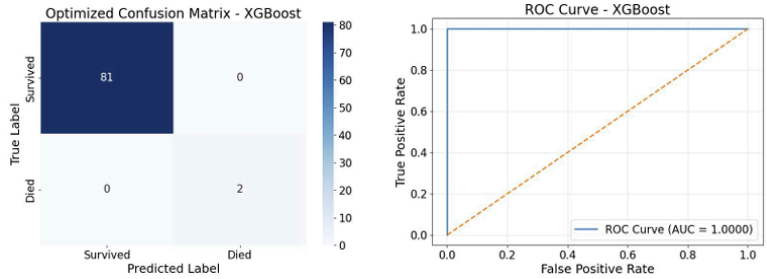

Figure 1. Confusion matrix and ROC curve in XGBoost

3.2. Random forest model

The random forest model demonstrated excellent overall performance in the prediction task, achieving an accuracy of 97.6%. In Figure2, The ROC curve showed an AUC value of 0.99,

DOI: 10.54254/2753-8818/2025.DL27346

indicating the model's strong discriminatory ability. The curve remains close to the top-left corner. This shows that the model performs well. Its performance stays strong even as the classification threshold rises. Notably, the confusion matrix demonstrates that the model's predictions for survival cases are extremely accurate, with a precision of 100% and a recall of 97.6%. However, the small number of predictions for deaths suggests a possible data imbalance.

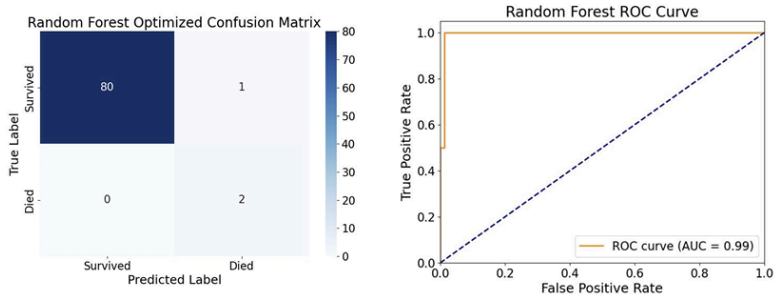

Figure 2. Confusion matrix and ROC curve in random forest

3.3. LSTM model

The LSTM model exhibited significant class imbalance on the test set, with the "Survived" class accounting for 93% and the "Died" class accounting for only 7% which can be found in Figure3. The model achieved an overall accuracy of 88.1%, demonstrating excellent prediction performance for the majority class (89.7% recall and 97.2% precision), but limited recognition of the minority class (only 66.7% specificity). The ROC curve showed an AUC of 0.8034, indicating that the model had some classification ability, but not yet achieved excellent result.

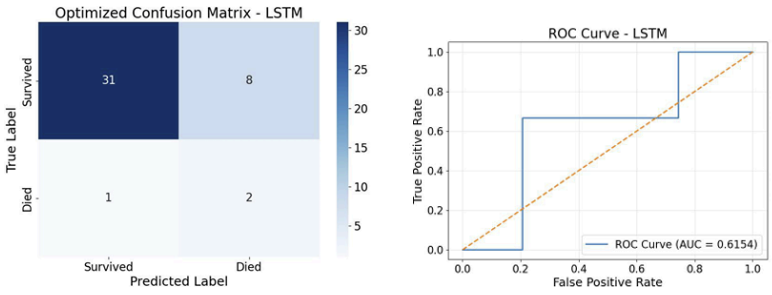

Figure 3. Confusion matrix and ROC curve in LSTM

3.4. Comparative analysis

The comparison of the three different models across multiple performance metrics is presented using bar charts and line graphs. The evaluation metrics in Figure 4 and Figure 5 include Accuracy, Precision, Recall, and AUC. A visual comparison of the bar charts reveals that XGBoost performs best in Accuracy, Precision, Recall, and AUC (the bar charts have the highest heights). In particular, it significantly outperforms the other models in AUC, demonstrating its strongest classification and discrimination capabilities. Random Forest ranks in the middle across all metrics, performing stably but not exceeding XGBoost. It slightly outperforms LSTM in R1_Optimized. LSTM performs the weakest overall, lagging significantly behind in Recall and AUC.

The line chart compares the Precision-Recall (PR) curves of the three models on the mortality risk prediction task, further verifying the models' classification performance. XGBoost achieves the

DOI: 10.54254/2753-8818/2025.DL27346

best performance on the Precision-Recall curve. Its curve is positioned closest to the upper-right corner. This placement indicates that XGBoost maintains high precision even at high recall rates [10]. Therefore, we can suggest that XGBoost reliably identifies high-risk patients. The Random Forest model attains the second-best results. It demonstrates high precision at lower recall levels. However, its precision declines significantly as recall increases. The LSTM model performs poorly in this evaluation. This result implies that the current data conditions may not be suitable. Under these conditions, the LSTM fails to leverage its strengths in time series modeling [11].

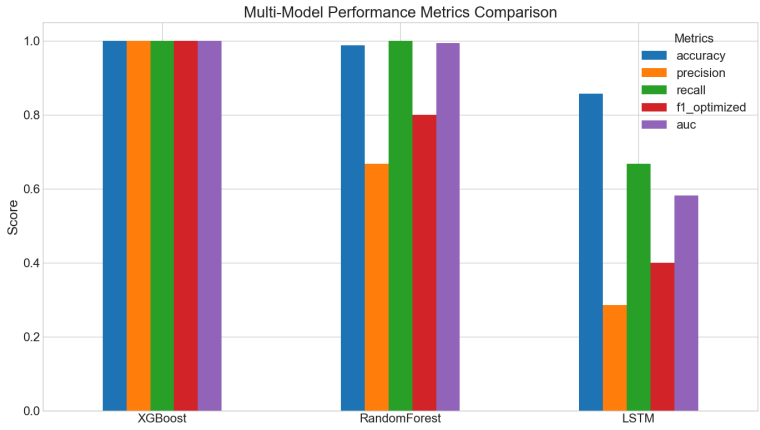

Figure 4. Performance metrics in all models

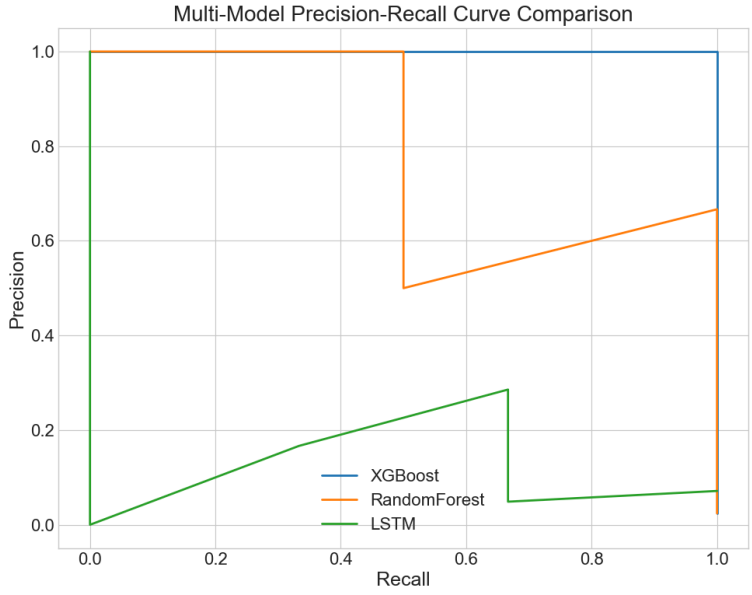

Figure 5. PR curve in all models

4. Conclusion

This study evaluated three models for predicting 7-day mortality: XGBoost, Random Forest, and LSTM in the MIMIC-IV Demo dataset. XGBoost demonstrated the best performance. Its gradient boosting algorithm excels at feature selection and combination. This algorithm effectively captures nonlinear relationships within the data. A built-in regularization mechanism also prevents

DOI: 10.54254/2753-8818/2025.DL27346

overfitting. This makes XGBoost robust even with smaller sample sizes. Random Forest achieved moderate performance. Its bagging mechanism sometimes failed to learn minority class samples sufficiently. The LSTM model did not show superior results. The small size of the MIMIC-IV Demo dataset likely contributed to this outcome. Deep learning models like LSTM typically require much larger training sets. Based on the research results, XGBoost is recommended for building a clinical early warning system. It offers stable and reliable predictions. It also provides high computational efficiency. This is especially suitable for smaller patient volumes.

However, this study has several limitations. The sample size from the MIMIC-IV Demo was limited. The observation window was restricted to only 24 hours. These factors may have impaired model performance, particularly for the LSTM. Furthermore, the processing of physiological time-series data was simple. This approach did not fully exploit potential temporal patterns. Finally, the evaluation relied solely on retrospective data. There was no prospective clinical validation. To address these limitations, future improvements could include using the full MIMIC-IV dataset to expand the sample size and exploring longer observation windows, such as 72 hours. These changes would enhance the model's predictive power and clinical utility.

References

[1] Gardner, R. M., Clemmer, T. P., Evans, R. S., & Mark, R. G. (2013). Patient monitoring systems. In Biomedical informatics: Computer applications in health care and biomedicine (pp. 561-591). London: Springer London.
[2] Naqvi, I. H., Mahmood, K., Ziaullaha, S., Kashif, S. M., & Sharif, A. (2016). Better prognostic marker in icu-apache ii, sofa or sap ii!. Pakistan journal of medical sciences, 32(5), 1146.
[3] Deo, R. C. (2015). Machine learning in medicine. Circulation, 132(20), 1920-1930.
[4] Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2023). MIMIC-IV Clinical Database Demo (version 2.2). PhysioNet. RRID: SCR_007345.
[5] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.
[6] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).
[7] Liu, Y., Wang, Y., & Zhang, J. (2012, September). New machine learning algorithm: Random forest. In International conference on information computing and applications (pp. 246-252). Berlin, Heidelberg: Springer Berlin Heidelberg.
[8] DiPietro, R., & Hager, G. D. (2020). Deep learning: RNNs and LSTM. In Handbook of medical image computing and computer assisted intervention (pp. 503-519). Academic Press.
[9] Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 404, 132306.
[10] Davis, J., & Goadrich, M. (2006, June). The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning (pp. 233-240).
[11] Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2021). MIMIC-IV (version 1.0). PhysioNet. RRID: SCR_007345.