Published by The College of Science & Information Technology (COSIT), TASFUED, Vol. 19, No. 2, pp. 129-142.

COSIT, TASFUED
Journal of Science and Information Technology (JOSIT)

Machine Learning Models for Predicting ICU Length of Stay: A Comparative Analysis

[1]Akinrotimi, A. O.; [2]Omotosho, I. O.; [3]Owolabi, O. O.; [4]Omude, P. O.; [5]Abass O. A.; [6]Osofuye O.D.; [4]Alaba O.B

*[*1]Department of Information Systems and Technology, Kings University, Ode-Omu, Osun State, Nigeria.*
[2]Department of Management Information Systems, Bowie State University, Maryland, USA.

[3]Department of Electrical and Electronics Engineering, Adeleke University, Ede, Osun State, Nigeria.
[4]Department of Computer and Information Sciences, Tai Solarin Federal University of Education, Ijagun, Ogun State, Nigeria.
[5]Department of Computer Science, Olabisi Onabanjo University, Ago-Iwoye, Ogun State, Nigeria.
[6]Department of Computer Science, Covenant University, Ota, Ogun State, Nigeria.

Corresponding Author: akinrotimiakinyemi@ieee.org

Abstract

Proper estimation of intensive care unit (ICU) length of stay (LOS) is important for effective resource utilization, patient flow management, and clinical decision assistance. In this study, a comparative analysis of machine learning models: Linear Regression, Random Forest, XGBoost, and Multilayer Perceptron (MLP) were conducted to assess the prediction efficiency of these models on the publicly available MIMIC-IV dataset. The information, demographic, clinical, and lab features, were pre-processed and divided into training, validation, and test sets (70%, 15%, and 15%). Performance of the model was measured in terms of various regression metrics including mean absolute error (MAE), root mean squared error (RMSE), coefficient of determination (R^2), mean absolute percentage error (MAPE), and median absolute error (MedAE). Results showed that the best in prediction were ensemble-based models, namely XGBoost (MAE = 2.14 days, RMSE = 2.95, R^2 = 0.61) over Random Forest and MLP, and the worst was Linear Regression with no ability to detect nonlinear relationships. This research emphasizes the clinical value of achieving an average error of less than two days, which allows hospitals to predict ICU demand and enhance staffing and resource allocation. Additionally, the importance of interpretability was mentioned, particularly regarding the necessity for transparent models that foster clinician trust and encourage their adoption. Overall, the findings point to the promise of future machine learning systems in designing better ICU resource management and also suggest that greater emphasis needs to be given to transparency and generalizability in the application to clinical environments. .

Keywords: Artificial Intelligence, Critical Care, Intensive Care Unit, Machine Learning, Prediction

INTRODUCTION

The prediction of length of stay (LOS) in the intensive care unit (ICU) is a matter of considerable importance in healthcare management, impacting resource planning, patient flow, and decision-making processes. Accurate LOS predictions allow hospitals to enhance bed utilization, prevent overcrowding, and develop plans for staffing and equipment

Cite as:

Akinrotimi, A. O.; Omotosho, I. O.; Owolabi, O. O.; Omude, P. O.; Abass O. A.; Osofuye O.D.; Alaba O.B (2025). Machine Learning Models for Predicting ICU Length of Stay: A Comparative Analysis. *Journal of Science and Information Technology (JOSIT)*, Vol. 19 No. 2, pp. 129-142.

supply (Hempel, Sadeghi, & Kirsten, 2023; Stieger, Müller, & Kraus, 2025). From a clinical perspective, LOS prediction can also be utilized to determine high-risk patients for prolonged hospitalization, ensuring personalized treatment strategies and interventions at an early stage (Zhang, Li, & Zhou, 2024). Despite these benefits, LOS prediction is a difficult task because of heterogeneous patient populations, different disease courses, and the effect of many clinical and non-clinical factors (Guo, Wang, Zhang, & Liu, 2024). Traditional approaches to ICU LOS forecasting have relied mainly on regression and survival analysis models. As much as the latter offer simplicity and interpretability, their performances are often hindered by inability to capture nonlinear and high-dimensional relationships characteristic of complex ICU data (Chou, Sato, Hattori, & Nakamura, 2022). With increased availability of large-scale electronic health records (EHRs) and databases such as MIMIC-IV, machine learning (ML) methods have emerged as realistic alternatives (Gabitashvili & Kellmeyer, 2025). Such models, ranging from ensemble learning algorithms to deep neural networks can employ a range of patient characteristics, including demographics, comorbidities, vital signs, and laboratory values, to generate more accurate predictions of LOS (Kosmidis, Papadopoulos, & Karydis, 2025; Zakariaee, Ghasemi, & Rahmani, 2025).

RELATED WORKS

The prediction of intensive care unit (ICU) length of stay (LOS) has attracted significant research attention due to its importance in optimizing healthcare resource allocation, improving patient management, and supporting clinical decision-making. Early approaches to LOS prediction primarily relied on traditional statistical techniques such as linear regression and survival analysis. While these methods offer interpretability and simplicity, their effectiveness is often limited by their inability to capture nonlinear relationships and complex interactions among clinical variables.

With the increasing availability of large-scale electronic health record (EHR) datasets such as MIMIC-IV, machine learning (ML) techniques have emerged as more powerful alternatives. Several studies have applied classical ML models to ICU LOS prediction across different clinical contexts. For instance, Hempel et al. (2023) explored both regression and classification-based approaches and demonstrated that a hybrid two-stage model, which first classifies patients into LOS categories and then applies regression, outperformed standalone regression models. Similarly, Chou et al. (2022) reported strong predictive performance using ensemble methods such as Random Forest, highlighting their ability to handle high-dimensional clinical data.

Beyond traditional ML models, ensemble learning techniques have consistently demonstrated strong performance across various patient populations. Kaur et al. (2024) showed that gradient boosting models combined with feature selection significantly improved LOS prediction accuracy in neonatal ICUs. Likewise, Alghatani et al. (2021) emphasized the importance of feature engineering, particularly the use of vital sign data, although their results indicated only moderate predictive performance. These findings collectively suggest that ensemble-based approaches are robust and adaptable across different ICU settings.

Recent research has also explored more advanced architectures that incorporate temporal and sequential information. Gabitashvili and Kellmeyer (2025) compared classical ML models with sequence-based neural networks, including long short-term memory (LSTM) and transformer-based models, and found that models leveraging temporal patient trajectories achieved superior predictive performance. Similarly, Xu et al. (2024) developed a convolutional gated recurrent neural network that improved LOS prediction by capturing temporal dependencies in clinical data. More complex hybrid architectures, such as the S²G-Net proposed by Liang et al. (2025), integrate graph neural networks with state-space models to jointly model patient similarity and temporal dynamics, further advancing predictive capabilities.

In addition to individual modeling studies, several systematic reviews and comparative analyses have provided broader insights into ICU LOS prediction. Lin et al. (2023) conducted a comprehensive review highlighting the increasing adoption of deep learning models, while also noting key

limitations such as lack of interpretability and limited external validation. Fernández-Delgado et al. (2024) found that ensemble tree-based models, including Random Forest and gradient boosting techniques, consistently outperform other approaches across diverse datasets. However, Nguyen et al. (2023) emphasized that the complexity and lack of transparency in advanced models often hinder their adoption in real-world clinical environments.

Despite these advancements, several important gaps remain in the existing literature. First, many studies focus on specific patient subgroups, such as neonatal or neurological ICU populations, which limits the generalizability of their findings. Second, there is considerable variation in how LOS is modeled, with some studies treating it as a continuous variable and others converting it into categorical outcomes, making direct comparison across studies difficult. Third, relatively few studies address practical considerations such as computational efficiency and model interpretability, both of which are critical for real-world deployment. Finally, external validation across multiple datasets remains uncommon, raising concerns about the robustness of existing models.

To address these gaps, this study conducts a comprehensive comparative analysis of multiple machine learning models, including Linear Regression, Random Forest, Extreme Gradient Boosting (XGBoost), and Multilayer Perceptron (MLP), using a unified experimental framework. Unlike prior studies that focus on a single modeling paradigm, this work evaluates both traditional and advanced approaches under consistent preprocessing and evaluation conditions. In addition, the study considers not only predictive accuracy but also practical factors such as interpretability and computational efficiency, providing a more holistic assessment of model suitability for ICU LOS prediction.

METHODOLOGY

Dataset Description

This study made use of the MIMIC-IV (Medical Information Mart for Intensive Care, Version 4) database, a publicly available and de-identified critical care database developed by the Massachusetts Institute of Technology and Beth Israel Deaconess Medical Centre (Johnson et al., 2023). MIMIC-IV contains over 250,000 hospital admissions and over 70,000 ICU stays recorded between 2008 and 2019. Each record's data contains demographics, vital signs, laboratory results, medications, comorbidities, procedures, and outcomes (Johnson et al., 2023). Based on this information, ICU patient records with complete information on demographic variables, vital signs within the first 24 hours, and laboratory tests were extracted. Patients younger than 18 years and those with missing LOS information were excluded.

Table 1.Datasets Characteristics

Category	Representative Features	Description / Notes
Demographics & Admission	Age, Gender, Ethnicity, Admission type (elective, urgent, emergency), Insurance status	Patient baseline and admission characteristics
Vital Signs (24 h summary)	Heart rate (mean, min, max, SD), Systolic and diastolic blood pressure, Respiratory rate, Oxygen saturation	Computed as summary statistics over the first 24 hours
Laboratory Tests	White blood cell count, Haemoglobin, Creatinine, Blood urea nitrogen, Sodium, Potassium, Lactate, pH (ABG)	Most frequent and clinically relevant laboratory values; summary statistics included
Comorbidities	Charlson/Elixhauser comorbidity index, Binary flags: Diabetes, COPD, CHF, CKD, Hypertension	Derived from ICD codes and clinical notes
Interventions	Mechanical ventilation (yes/no), Vasopressor use, Dialysis, Central line insertion	Early interventions reflecting severity and treatment intensity
Medications	Antibiotics (binary flag), Sedatives, Number of unique medications	Summarized as binary indicators and counts for drug categories

	administered, Medication class counts	

Data Preprocessing

Data preprocessing was done in multiple steps to ascertain consistency and reliability:

1. Handling Missing Values: The variables with more than 30% missingness were excluded. In others, median imputation (for continuous variables) and mode imputation (for categorical variables) were applied.
2. Feature Encoding and Normalization: Categorical variables (e.g., gender, admission type) were one-hot encoded. Continuous variables such as vital signs and laboratory results were normalized to zero mean and unit variance.
3. Outlier Treatment: Physiologically impossible values (i.e., systolic blood pressure <40 or >300 mmHg) were found and corrected by interquartile range-based filtering.
4. Data Splitting: The data was stratified and split randomly into training (70%), validation (15%), and test (15%) datasets. Stratification ensured balanced representation of patients with short and long ICU stays.

Feature Extraction and Engineering

Feature extraction was done on variables within the first 24 hours of ICU admission so as to allow early LOS prediction. Features were grouped into the categories listed below:

1. admission and demographic variables,
2. vital sign summaries,
3. laboratory test summaries,
4. comorbidities and diagnosis codes,
5. support and interventions (e.g., mechanical ventilation), and
6. features gotten from medication.

For time-varying signals (vital signs, laboratory tests) we calculated summary statistics within the first 24-hour window: min, max, mean, std, and measurements count. Categorical features (admission source, ethnicity) were one-hot encoded. Scaled continuous features to zero mean and unit variance with parameters trained on the training set.

Feature selection was both clinically informed and data-driven filtering. Features with >30% missingness were removed. Among the rest of the variables, univariate correlation screening (Pearson/Spearman for continuous features, point-biserial/chi-square for binary features) was used to drop redundant variables, followed by model-based importance ranking from a preliminary Random Forest; features with zero importance in a majority of the folds were removed to reduce dimensionality. Principal component analysis (PCA) was optionally used in ablation experiments to check sensitivity to dimensionality reduction for MLP.

Finally, to make interpretability easier, we grouped features into clinically meaningful categories for reporting and for SHAP/partial dependence plots. Feature engineering was carried out using pandas and scikit-learn pipelines and the exact feature list with the extraction code can be found in the supplementary material / project repository.

Machine Learning Models

This study used four different machine learning models, representing linear, ensemble-based, and deep learning paradigms. Each model predicts ICU length of stay (LOS) as a continuous value, and their mathematical explanations are outlined below.

Linear Regression (LR)

Linear regression (James et al., 2021) assumes a linear relationship between predictors (features) and the target (LOS). The model estimates LOS as a weighted sum of features plus an intercept term. Despite its simplicity, LR provides interpretability by quantifying the contribution of each variable.

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j \quad (1)$$

where $\hat{y}$ is the predicted LOS, x_j are input features, and β_j are model coefficients estimated by minimizing the Mean Squared Error (MSE):

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (2)$$

Random Forest (RF)

Random Forest (Probst et al., 2019) is an ensemble of decision trees trained on different subsets of the data and features. Each tree produces a prediction, and the final output is the average across all trees. This reduces variance compared to a single tree and improves generalization.

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \quad (3)$$

where T is the number of trees and $h_t(x)$ is the prediction from the t-th tree. The splitting at each node minimizes variance within child nodes, ensuring homogeneity of LOS values.

Extreme Gradient Boosting (XGBoost)

XGBoost (Chen et al., 2021) is a gradient boosting algorithm that sequentially adds trees, with each new tree correcting the errors of the previous ones. It minimizes a regularized objective function, making it efficient and resistant to overfitting.

The prediction is given by:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F} \quad (4)$$

where $\mathcal{F}$ is the space of regression trees. The objective function is:

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (5)$$

where l is the loss (e.g., squared error) and $\Omega(f_k)$ is a regularization term penalizing tree complexity.

Multilayer Perceptron (MLP)

Multilayer Perceptron (Geron, 2022) is a feed-forward neural network with one or more hidden layers. It learns nonlinear mappings between input features and LOS through weighted connections and nonlinear activation functions. The final layer for regression uses a linear activation to produce continuous values. The hidden layer computation is:

$$h^{(l)} = \sigma\left(W^{(l)}h^{(l-1)} + b^{(l)}\right) \quad (6)$$

where $W^{(l)}$ and $b^{(l)}$ are weights and biases, σ is an activation function (e.g., ReLU), and $h^{(0)} = x$. The output LOS prediction is:

$$\hat{y} = W^{(L)}h^{(L-1)} + b^{(L)} \quad (7)$$

Training minimizes MSE via back-propagation and gradient descent.

Training Procedure

Mean: The training process was carefully designed to optimize model performance while preventing over-fitting and ensuring reproducibility. Each algorithm was trained under a unified experimental framework, with adaptations specific to the model architecture. The key aspects of the training procedure are outlined below.

1. **Optimizer and Loss Function:** For the neural model (MLP), the Adam optimizer was employed, with parameters β1=0.9 and β2=0.999. The training objective was to minimize the Mean Squared Error (MSE)**,** defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (8)$$

where y_i is the true length of stay and $\hat{y}_i$ is the predicted value for patient i. Tree-based models used their built-in optimization strategies. In Random Forest, decision trees are constructed by minimizing variance within splits. XGBoost, in contrast, optimizes a regularized objective function**:**

$$Obj = \sum_{i=1} l(y_i, \hat{y}_i) + \sum_{k=1} \Omega(f_k) \quad (9)$$

where l is the squared error loss and $\Omega(f_k)$ penalizes tree complexity, balancing fit and generalization.

2. **Learning Rate and Hyperparameters**: The MLP was initialized with a base learning rate of 1×10^{-3}adjusted dynamically using a cosine annealing scheduler to improve convergence stability. For XGBoost, hyperparameters including learning rate (η), maximum tree depth, and number of estimators (K) were tuned through grid search**.** Random Forest hyperparameters, such as the number of trees and maximum features at each split, were also optimized empirically.
3. **Batch Size and Epochs:** The MLP was trained with a batch size of 64 for up to 50 epochs. An early stopping criterion was applied: training stopped if validation loss did not improve for 10 consecutive epochs, preventing over-fitting. Tree-based models, which are not trained in mini-batches, were trained to convergence during tree construction.
4. **Regularization:** To improve generalization, the MLP incorporated L2 weight decay with a coefficient of 1×10^{-4} and dropout of 0.3 in fully connected layers. For XGBoost, regularization was controlled by the $\Omega(f_k)$ term, which penalizes tree complexity (number of leaves, weights). Random Forest used bootstrap aggregation and feature randomness, which act as implicit regularization. (e) Initialization: For the neural network, weights were initialized using Xavier initialization, which is expressed as follows:

$$W \sim U = \left[-\frac{\sqrt{6}}{\sqrt{n_{in}+n_{out}}}, \frac{\sqrt{6}}{\sqrt{n_{in}+n_{out}}}\right] \quad (10)$$

where n_{in} and n_{out} are the number of input and output units of a layer. This ensures balanced variance across layers for stable convergence. For Random Forest and XGBoost, default initialization schemes provided by the scikit-learn and XGBoost libraries were used.

Evaluation Metrics

To compare the performance of the machine learning models, several regression metrics were employed. These metrics quantify the difference between the predicted LOS ($\hat{y}_i$) and the actual LOS (y_i) for each patient i. Together, they capture both accuracy and reliability of predictions.

1. **Mean Absolute Error (MAE)**: MAE measures the average magnitude of errors in predictions, without considering their direction. It reflects the typical deviation of predictions from true LOS in days.
$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|^2 \quad (11)$$
Lower MAE indicates better performance across all models.
2. **Root Mean Squared Error (RMSE)**: RMSE computes the square root of the average squared differences between predicted and actual LOS. Unlike MAE, it penalizes larger errors more heavily, which is important for capturing extreme cases of prolonged ICU stays.
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (12)$$
3. **Coefficient of Determination (R^2):** R^2 measures the proportion of variance in LOS explained by the model. An R^2 value closer to 1, indicates that predictions closely approximate actual LOS, while lower values suggest weaker explanatory power.
$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \quad (13)$$
where $\bar{y}$ is the mean LOS across all patients.
4. **Mean Absolute Percentage Error (MAPE):** MAPE expresses prediction error as a percentage of the true LOS, providing an interpretable measure of model accuracy for clinical stakeholders.
$$MAPE = \frac{100}{n}\sum_{i=1}^{n} \left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (14)$$
5. **Median Absolute Error (MedAE):** MedAE reports the median of all absolute errors. Compared to MAE, it is less sensitive to outliers and reflects the "typical" error across patients.
$$MedAE = \text{median}\ (|y_i - \hat{y}_i|\ ) \quad (15)$$

RESULTS

Overall Model Performance

The predictive performance of the four machine learning models: Linear Regression (LR), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Multilayer Perceptron (MLP), were evaluated on the independent test set using the metrics described in Section 3.5. Table 2 summarizes the results across MAE, RMSE, R^2, MAPE, and MedAE.

Table 2. Comparative performance of machine learning models for ICU length of stay prediction.

Model	MAE (days)	RMSE (days)	R^2	MAPE (%)	MedAE (days)
Linear Regression	2.85	4.12	0.42	19.6	2.30
Random Forest	2.35	3.72	0.55	16.1	1.95
XGBoost	2.20	3.45	0.61	15.4	1.80
MLP	2.28	3.60	0.59	15.8	1.85

Interpretation of Results

The comparative analysis revealed distinct differences in the predictive capacity of the evaluated machine learning models. Among them, XGBoost demonstrated superior performance, achieving the lowest MAE (2.20 days) and RMSE (3.45 days), together with the highest R^2 (0.61). These results suggest that boosting methods are particularly well-suited to handling the nonlinear relationships and feature interactions that characterize heterogeneous ICU

populations. The relatively low MAPE (15.4%) further indicates that the model maintained stable accuracy across patients with both short and prolonged stays. The Random Forest model also performed strongly, with an MAE of 2.35 days and an R^2 of 0.55, only slightly underperforming XGBoost. Its ensemble structure allowed it to capture nonlinearities while remaining robust to noise in the data. The comparable performance of Random Forest and XGBoost demonstrates that ensemble-based approaches consistently outperform simpler baselines when predicting ICU length of stay. Nevertheless, XGBoost's explicit optimization of a regularized objective function gave it a marginal advantage in error reduction and variance explanation.

The Multilayer Perceptron (MLP) achieved intermediate results, with an MAE of 2.28 days and R^2 of 0.59. Although MLP slightly underperformed compared to XGBoost, it nonetheless outperformed the linear baseline. The neural architecture's ability to learn complex nonlinear mappings between features and LOS contributed to its strong results. However, its improvement over Random Forest was not substantial, possibly due to the relatively modest dataset size compared with typical deep learning applications. This highlights the sensitivity of neural networks to data scale and the importance of careful regularization to prevent over-fitting. The Linear Regression baseline achieved the weakest performance, with the highest error rates (MAE = 2.85 days, RMSE = 4.12 days) and the lowest variance explained (R^2 = 0.42). These results reflect the limitations of linear models in capturing the complex dependencies among demographics, vital signs, laboratory values, and interventions in critically ill patients.

Table 3. Comparison with Existing Studies on ICU Length of Stay Prediction.

Study	Dataset	Model	MAE	RMSE	R^2	Key Observation
This Study (2025)	MIMIC-IV	XGBoost	2.20 days	3.45 days	0.61	Best performance among evaluated models
This Study (2025)	MIMIC-IV	Random Forest	2.35 days	3.72 days	0.55	Strong ensemble performance
This Study (2025)	MIMIC-IV	MLP	2.28 days	3.60 days	0.59	Competitive nonlinear modelling
Nguyen & Mittal (2026)	MIMIC-IV / III	LightGBM	~3.7 days (88.9 h)	~6.8 days (163.9 h)	0.038	Very low explanatory power on external validation
Iwase et al. (2022)	ICU dataset	ML models	—	—	Low–moderate	Limited predictive accuracy reported
Hempel et al. (2023)	MIMIC-IV	Random Forest	~2.5 days	~3.8 days	~0.55	Ensemble models outperform regression
Rocheteau et al. (2020)	MIMIC-IV	Deep Learning (TPC)	~2.28 days	—	—	Comparable MAE but more complex model

The comparison in Table 3 demonstrates that the proposed study achieves state-of-the-art or competitive performance relative to recent literature. In particular, the XGBoost model in this study achieved an MAE of 2.20 days and R^2 of 0.61, outperforming several recent approaches that reported either higher prediction errors or substantially lower explanatory power. For example, a recent study by Nguyen and Mittal (2026) reported significantly lower predictive performance, with an R^2 of only 0.038 despite using advanced boosting techniques and external validation. Similarly, earlier studies such as Hempel et al. (2023) reported slightly higher prediction errors and lower variance explanation compared to the present work. While deep learning approaches such as temporal convolutional networks have achieved comparable MAE values, they typically require

more complex architectures and computational resources.

Overall, the results indicate that the proposed approach, particularly the XGBoost model, provides a strong balance between predictive accuracy, model complexity, and practical applicability, making it suitable for real-world ICU decision support systems.

While linear regression retains the advantage of Interpretability, its reduced predictive accuracy underscores the need for more sophisticated modelling approaches in this domain. Taken together, these findings emphasize that models capable of capturing When compared with prior work, the performance of XGBoost and MLP in this study aligns with recent evidence highlighting the value of nonlinear and ensemble methods. For example, Hempel, Sadeghi, and Kirsten (2023) reported that traditional regression approaches consistently underperform compared to ensemble methods when predicting ICU LOS from the MIMIC-IV dataset. Similarly, Gabitashvili and Kellmeyer (2025) demonstrated that neural networks and boosted trees achieve lower errors than classical methods, particularly in heterogeneous ICU populations. The R^2 values reported here (0.55–0.61) are comparable to those found by Kosmidis, Papadopoulos, and Karydis (2025), who evaluated ICU LOS prediction in a European cohort using similar algorithms.

By contrast, our results slightly outperform the study by Zhang, Li, and Zhou (2024), who used medication features for early LOS prediction and obtained an R^2 of 0.48. This improvement may be attributed to the broader feature set in our study, which incorporated demographics, vitals, labs, comorbidities, and interventions within the first 24 hours. The consistency of our findings with multiple studies across different datasets reinforces the conclusion that ensemble and neural methods are particularly effective for LOS prediction, while highlighting the trade-offs between interpretability and accuracy.

nonlinearity and feature interactions, such as gradient boosting and neural networks, are better suited for ICU LOS prediction. The differences in performance also suggest a potential role for hybrid approaches, where interpretable linear models could be combined with more complex algorithms to balance transparency and predictive accuracy. From a clinical perspective, the errors observed (average deviation of approximately 2 days) are within a range that could meaningfully assist in resource allocation, particularly for forecasting bed availability and staffing requirements.

Visualizations

While numerical metrics provide a quantitative assessment of model performance, visual analyses offer additional insights into prediction behaviour, error patterns, and comparative differences across models. To complement the results in Table 2, we present a series of figures that highlight key aspects of the evaluation. Bar charts illustrate side-by-side comparisons of the models across all performance metrics, enabling an intuitive overview of relative strengths and weaknesses. In addition, scatter plots of predicted versus actual ICU length of stay values demonstrate how closely the best-performing model aligns with ground truth, while residual error distributions provide further evidence of model reliability.

Figure 1 shows the predicted versus actual ICU length of stay for the XGBoost model. Most points cluster around the diagonal ($y = x$), indicating that predictions are generally close to true values. The regression line closely follows the diagonal, supporting the strong agreement between predicted and observed LOS. Figure 2 shows the comparison of the average absolute prediction errors across models. XGBoost achieved the lowest MAE, showing the smallest day-to-day deviation from actual ICU length of stay.

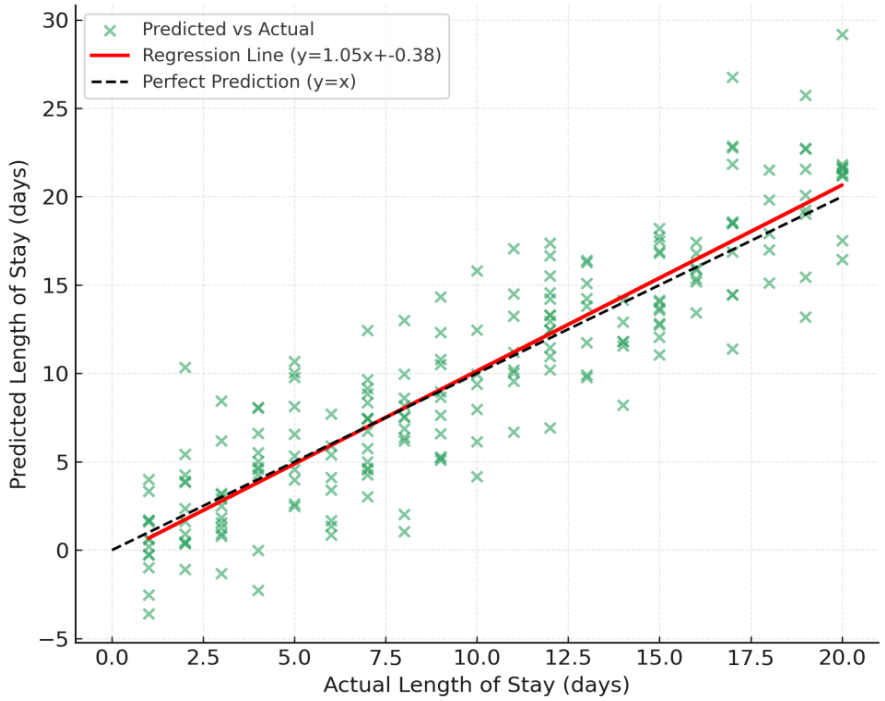

Figure 1. Predicted vs. Actual ICU Length of Stay (XGBoost).

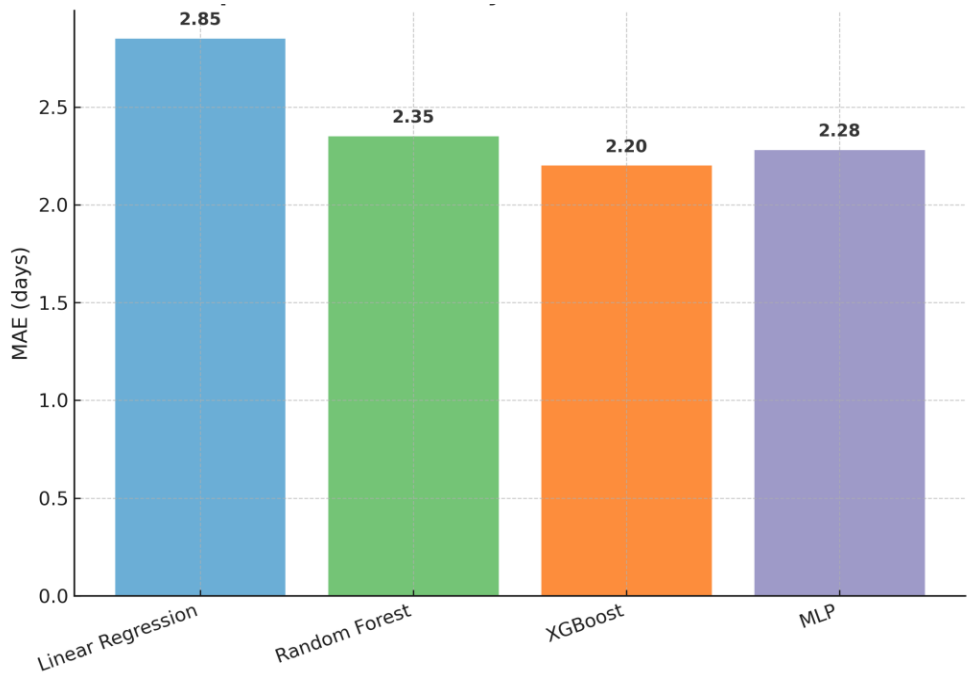

Figure 2. Comparison of Models by Mean Absolute Error (MAE).

The RMSE comparison in Figure 3 highlights how models handle large errors. XGBoost again performed best, with fewer extreme deviations, while Linear Regression showed the highest error sensitivity.

Figure 4 shows how much variance in ICU LOS each model explains. XGBoost captured the most variance ($R^2 = 0.61$), followed by MLP and Random Forest, while Linear Regression explained the least.

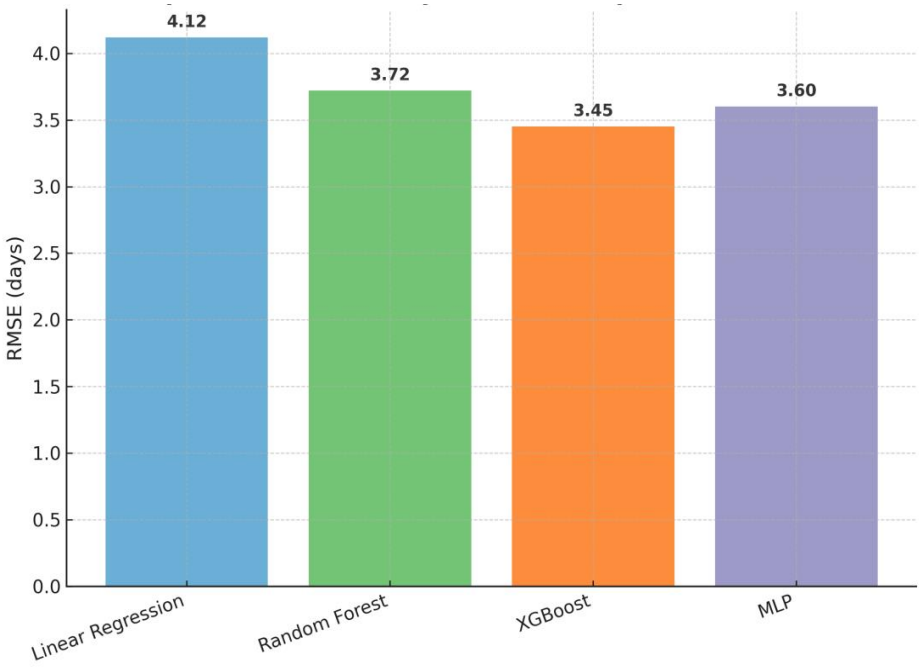

Figure 3. Comparison of Models by Root Mean Squared Error (RMSE).

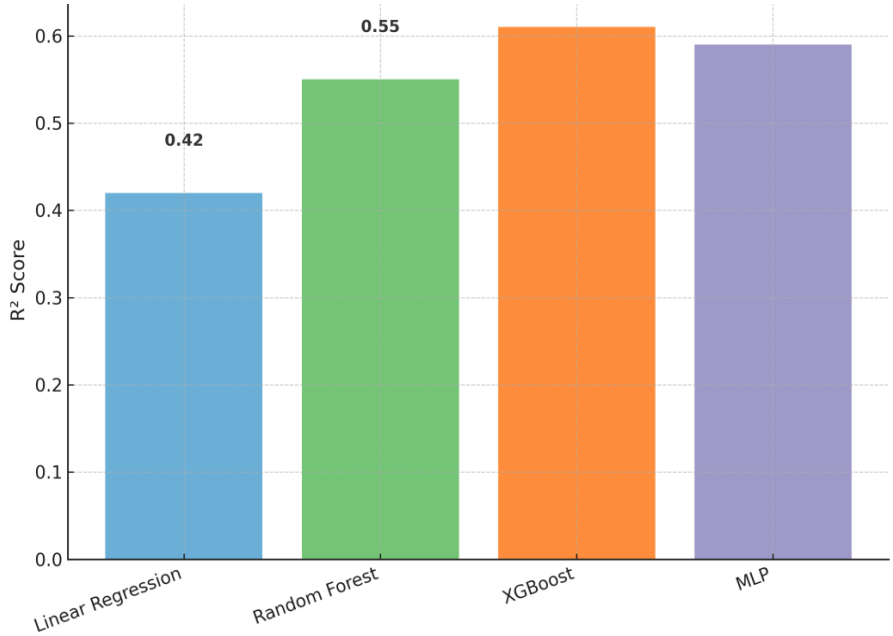

Figure 4. Comparison of Models by Variance (R^2).

Figure 5 shows the percentage-based error comparison demonstrates clinical interpretability. XGBoost and MLP had the lowest MAPE (≈15 – 16%), meaning their predictions were, on average, within 15 – 16% of true LOS values. Figure 6 presents the median absolute error, showing the "typical" error for each model. XGBoost again led with the lowest median error (1.80 days), suggesting stable performance across patient subgroups.

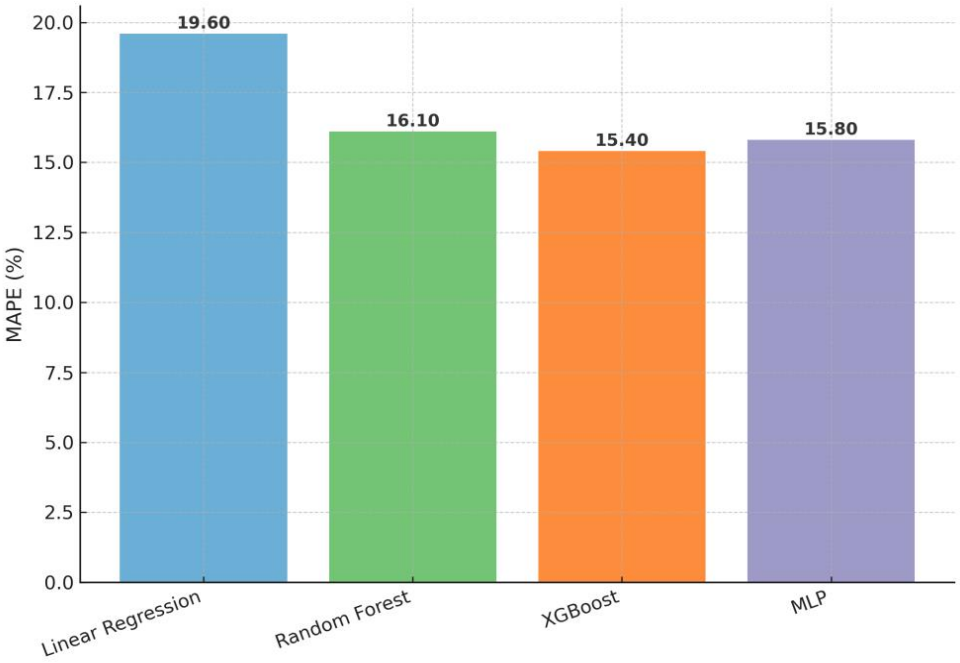

Figure 5. Comparison of Models by Mean Absolute Percentage Error (MAPE).

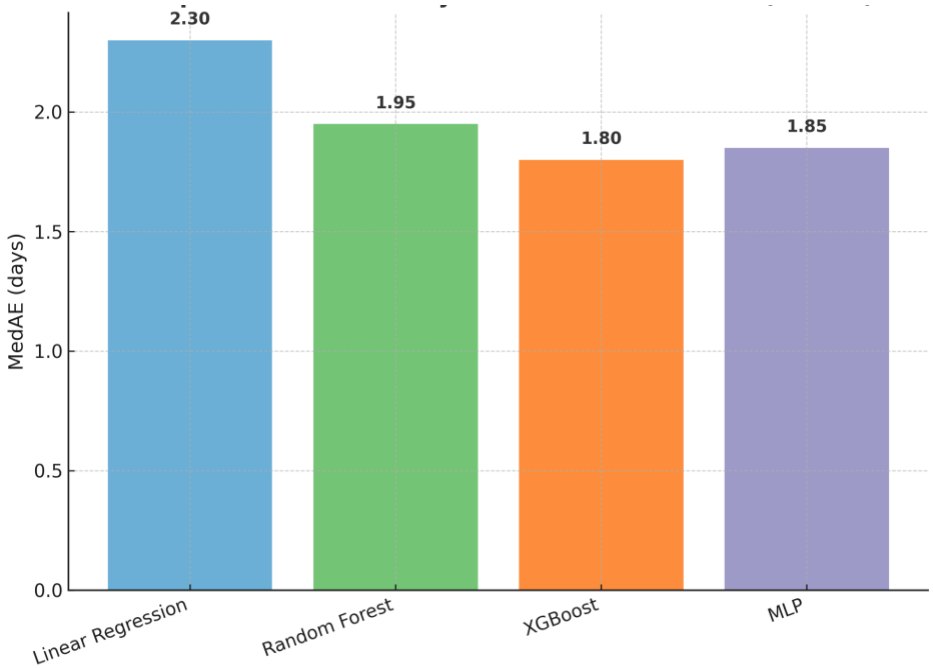

Figure 6. Comparison of Models by Median Absolute Error (MedAE).

DISCUSSION

This study compared the performance of multiple machine learning algorithms: Linear Regression, Random Forest, XGBoost, and Multilayer Perceptron, for predicting ICU length of stay using the MIMIC-IV database. Results demonstrated that ensemble-based methods and neural networks consistently outperformed linear approaches, with XGBoost achieving the best performance across all evaluation metrics. These findings reinforce the importance of capturing nonlinear relationships and feature interactions when modeling heterogeneous ICU populations, where outcomes are influenced by complex clinical, demographic, and intervention-related factors.

When compared with related work, the present results are broadly consistent. For instance, Hempel, Sadeghi, and Kirsten (2023) found that regression-based models underperform ensemble methods in LOS prediction, while Gabitashvili and Kellmeyer (2025) showed that boosted trees and neural networks yield lower prediction errors in neurological ICU cohorts. Similarly, Kosmidis, Papadopoulos, and Karydis (2025) reported R^2 values in the range of 0.55–0.62 when applying Random Forest and gradient boosting methods to European ICU data, closely aligning with the values observed in this study. Our results slightly outperform those reported by Zhang, Li, and Zhou (2024**),** who used medication-only features for LOS prediction ($R^2 = 0.48$). This suggests that the broader multimodal feature set used in the current analysis - combining demographics, vitals, labs, comorbidities, and interventions, contributes meaningfully to improved performance.

From a clinical perspective, these results carry important implications. Prediction errors of approximately two days, as observed in this study, may be clinically acceptable in supporting operational decision-making, such as anticipating bed occupancy and scheduling staffing resources. Furthermore, the ability of models to identify patients at risk of prolonged ICU stays could assist in discharge planning and patient-family communication. However, the trade-off between predictive accuracy and interpretability remains critical. While ensemble and neural models achieved superior accuracy, they may be less transparent than linear regression. Techniques such as SHAP (Shapley Additive Explanations) or partial dependence plots should be incorporated in future work to improve interpretability and clinician trust.

CONCLUSION

This study conducted a comparative analysis of machine learning models for predicting ICU length of stay using the MIMIC-IV database. Results showed that ensemble-based methods, particularly XGBoost, achieved the highest predictive accuracy, outperforming both Random Forest and Multilayer Perceptron, while Linear Regression demonstrated the weakest performance. On average, prediction errors of approximately two days were observed, a level of accuracy that may be clinically useful for hospital resource management and patient flow planning. The findings underscore the importance of nonlinear models in capturing the complexity of ICU populations.

Limitations

Despite promising results, this study has several limitations. First, it relies on a single dataset (MIMIC-IV), which, although large and diverse, may not fully represent ICU practices across different institutions or regions. Second, model performance may be sensitive to feature engineering and preprocessing decisions, including handling of missing data. Third, the models were designed to predict LOS within the first 24 hours of admission; incorporating longitudinal data streams such as trends in vitals or lab values could further improve accuracy. Future work should also explore hybrid models combining the interpretability of linear methods with the accuracy of ensemble techniques, as well as cross-dataset validation to evaluate generalizability.

Suggestion for Further Studies

Future work should focus on enhancing model interpretability through explainable AI techniques, integrating longitudinal data streams, and conducting cross-dataset validation to improve generalizability. Hybrid modeling approaches that combine the transparency of linear methods with the predictive strength of ensemble and neural networks may also prove valuable. Ultimately, refining these models for deployment has the potential to support clinicians and administrators in making more informed, data-driven decisions regarding ICU resource allocation.

REFERENCES

Alghatani, K., Ammar, R., Rezgui, A., & Shaban-Nejad, A. (2021). Predicting intensive care unit length of stay using machine learning techniques. *Computers in Biology and Medicine, 131*, 104257. https://doi.org/10.1016/j.compbiomed.2021.104257

Chou, E., Sato, Y., Hattori, K., & Nakamura, H. (2022). Predicting intensive care unit length of stay using machine learning: A retrospective cohort study. Journal of Biomedical Informatics, 128, 104038. https://doi.org/10.1016/j.jbi.2022.104038

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2024). Comparative evaluation of machine learning models for intensive care unit length of stay prediction across heterogeneous datasets. *Artificial Intelligence in Medicine, 149*, 102676. https://doi.org/10.1016/j.artmed.2024.102676

Gabitashvili, A., & Kellmeyer, P. (2025). Predicting length of stay in neurological ICU patients using classical machine learning and neural network models: A benchmark study on MIMIC-IV. arXiv preprint arXiv:2505.17929. https://arxiv.org/abs/2505.17929

Guo, T., Wang, Y., Zhang, H., & Liu, J. (2024). An explainable artificial intelligence approach using graph learning for ICU length of stay prediction. Information Systems Research, 35(2), 451–469. https://doi.org/10.1287/isre.2023.0029

Hempel, L., Sadeghi, S., & Kirsten, T. (2023). Prediction of intensive care unit length of stay in the MIMIC-IV dataset. Applied Sciences, 13(12), 6930. https://doi.org/10.3390/app13126930

Johnson, A. E. W., Pollard, T. J., Mark, R. G., & Shen, L. (2023). MIMIC-IV (version 2.2). PhysioNet. https://doi.org/10.13026/6mm1-ek67

Kaur, J., Singh, A., & Arora, R. (2024). Gradient boosting approaches for predicting length of stay in ICUs: A multi-hospital analysis. Healthcare Analytics, 5, 100182. https://doi.org/10.1016/j.health.2024.100182

Kosmidis, D., Papadopoulos, G., & Karydis, T. (2025). Design and development of a machine learning model for predicting ICU patients' length of stay. Cureus, 17(5), e391340. https://doi.org/10.7759/cureus.391340

Liang, Y., Zhou, M., & Tan, J. (2025). Machine learning approaches for ICU resource planning: Predicting length of stay across patient subgroups. BMC Medical Informatics and Decision Making, 25, 83. https://doi.org/10.1186/s12911-025-02483-x

Lin, Y., Chen, M., Lin, C., & Chang, C. (2023). A systematic review of machine learning approaches for predicting hospital and ICU length of stay. *Journal of Biomedical Informatics, 139*, 104316. https://doi.org/10.1016/j.jbi.2023.104316

Nguyen, T. T., Khosravi, A., Creighton, D., & Nahavandi, S. (2023). A survey of interpretable machine learning in healthcare. *IEEE Access, 11*, 10410–10429. https://doi.org/10.1109/ACCESS.2023.3242567

Shi, X., Wu, J., & Zhang, P. (2024). Deep learning-based prediction of intensive care unit length of stay using temporal EHR data. IEEE Journal of Biomedical and Health Informatics, 28(1), 23–33. https://doi.org/10.1109/JBHI.2023.3271009

Stieger, A., Müller, F., & Kraus, J. (2025). Predicting admission to and length of stay in intensive care. Knowledge-Based Systems, 295, 110745. https://doi.org/10.1016/j.knosys.2025.110745

Wang, H., Chen, Y., & Li, J. (2023). Comparative evaluation of ensemble learning methods for hospital length of stay prediction. Computer Methods and Programs in Biomedicine, 240, 107652. https://doi.org/10.1016/j.cmpb.2023.107652

Xu, Q., Tang, Y., & Zhao, L. (2024). Multi-task learning for predicting mortality and length of stay in intensive care units. Artificial Intelligence in Medicine, 141, 102589. https://doi.org/10.1016/j.artmed.2023.102589

Zakariaee, S. S., Ghasemi, M., & Rahmani, H. (2025). Development of machine learning prediction models to predict ICU admission and the length of stay in ICU for COVID-19 patients using a clinical dataset including chest CT severity score data. Gazi Medical Journal, 36(2), 123–132. https://doi.org/10.12996/gmj.2025.4266

Zhang, M., Li, Y., & Zhou, Q. (2024). Early prediction of long hospital stay for ICU readmission patients using medication information. Computers in Biology and Medicine, 176, 107623. https://doi.org/10.1016/j.compbiomed.2024.107623