Machine Learning Application in Healthcare: A Case Study Using Ensemble Methods for Hospital Length of Stay Prediction

Hakima Reddad[1], Maria Zemzami[2], Norelislam El Hami[3], Nabil Hmina[4], Farouk Yalaoui[5]
Laboratory of Advanced Systems Engineering-National School of Applied Sciences (ENSA),
Ibn Tofail University, Kenitra, Morocco[1]
High National School of Arts and Crafts (ENSAM), University Mohammed V, Rabat, Morocco[2]
Science and Engineering Laboratory-National School of Applied Sciences (ENSA), Ibn Tofail University, Kenitra, Morocco[3]
Ibn Zohr University, Agadir, Morocco[4]
Computer Sciences and Digital Society Laboratory (LIST3N), University of Technology of Troyes (UTT),
Chaire Connected Innovation, Troyes, France[5]

***Abstract*—Artificial intelligence is driving digital transformation across multiple sectors, including healthcare, pharmaceuticals, industrial production, and the automotive industry. In healthcare specifically, AI-powered predictive analytics offer significant potential for optimizing operational efficiency and resource allocation. To demonstrate this potential, we present a case study focused on hospital length of stay (LOS) prediction using 2,125,280 admission records from the New York SPARCS database. We implemented and compared four machine learning algorithms: Linear Regression, Random Forest, Gradient Boosting, and XGBoost. Following hyperparameter optimization, the XGBoost model achieved superior performance with R²=0.8686, RMSE=3.24 days, and MAE=1.42 days, substantially outperforming Linear Regression (R²=0.5339, RMSE=6.10 days, MAE=2.86 days). Prediction accuracy reached 63.34% within ±1 day and 89.44% within ±3 days of actual LOS. SHAP analysis identified Total Costs, Total Charges, Hospital Service Area, APR Medical Surgical Description, and APR DRG Code as the most impactful predictors. Performance varied across LOS categories, with MAE ranging from 0.66 days for short stays (1-3 days) to 11.81 days for extended hospitalizations (>30 days). These results demonstrate that ensemble machine learning methods, particularly XGBoost, provide clinically meaningful accuracy for healthcare operational planning, though challenges remain for extended stays and complex cases requiring specialized modeling approaches.**

***Keywords*—Machine learning; XGBoost; healthcare operations; hospital resource management; ensemble methods; predictive analytics; SHAP analysis**

I. Introduction

Nowadays, digital transformation is revolutionizing multiple sectors, including healthcare, pharmaceutical, industrial, and automotive industries, fundamentally reshaping operational processes, decision-making frameworks, and service delivery models [1], [2]. Artificial intelligence (AI) stands as one of the main enablers for this transformation, encompassing diverse technological paradigms, including computer vision for visual perception tasks, large language models (LLMs) for natural language understanding and generation, machine learning (ML) for pattern recognition and prediction, and deep learning (DL) for hierarchical feature extraction from complex data [3,4]. Machine learning particularly enables critical analytical tasks spanning forecasting future trends, regression for continuous outcome prediction, classification for categorical decision-making, and clustering for pattern discovery [4-8]. These capabilities are realized through three primary learning paradigms: supervised learning, where models learn from labeled training data, unsupervised learning for discovering hidden patterns without explicit labels, and reinforcement learning for sequential decision-making through interaction with environments [4-8].

In the healthcare field, ML has demonstrated transformative potential across diverse applications, including medical image analysis, hospital readmission prediction, clinical decision support, and operational resource management [8-10]. These applications leverage ML's capacity to model complex, non-linear relationships in heterogeneous clinical data. Among ML algorithms, ensemble methods have proven particularly effective for healthcare tabular data [2,11], with XGBoost emerging as a leading approach due to its regularization mechanisms, efficient handling of missing values, computational scalability, and interpretability through SHAP-based feature importance analysis [12-16].

Hospital length of stay prediction represents a critical application where accurate forecasting enables proactive bed management, optimized discharge planning, evidence-based staffing, and financial forecasting [17,18]. While prior studies have applied various approaches to LOS prediction [11,19-22], most focus on specific clinical subpopulations or single institutional datasets, limiting understanding of model performance across diverse patient populations and healthcare settings. This study addresses this gap through a comprehensive evaluation of four machine learning algorithms on 2,125,280 admission records from New York's statewide hospital database [23], providing insights into algorithm selection, hyperparameter optimization value, prediction accuracy heterogeneity, and key drivers of length of stay. We address four primary objectives:

- Develop and validate predictive models for hospital LOS using four algorithms spanning traditional and modern machine learning approaches.

- Conduct rigorous performance comparison across these algorithms using multiple complementary evaluation metrics to identify optimal modeling approaches for operational deployment.

- Provide interpretable insights through SHAP analysis to identify key clinical, administrative, and financial drivers of length of stay, informing targeted interventions for LOS reduction and supporting clinical understanding of model predictions.

- Assess performance heterogeneity across patient subpopulations stratified by LOS duration to identify clinical scenarios where predictions are sufficiently accurate for operational use versus populations requiring alternative prediction strategies.

Our approach leverages 2,125,280 admission records from the New York Statewide Planning and Research Cooperative System (SPARCS), representing diverse patient populations, hospital types, and clinical conditions across an entire state healthcare system. This scale and diversity enable robust model development and evaluation that captures real-world complexity beyond single-institution studies. By comparing baseline and optimized XGBoost configurations, we quantify the value of hyperparameter tuning for healthcare prediction tasks. Through stratified performance analysis across LOS categories, we provide actionable guidance on appropriate model deployment strategies for different patient populations, conceding that uniform prediction accuracy across all cases is neither achievable nor necessary for operational value.

The remainder of this study is organized as follows: Section II reviews the state-of-the-art in machine learning applications for healthcare, with a focused analysis of prior work in hospital length of stay prediction. Section III describes the dataset, preprocessing procedures, feature engineering, and experimental setup, including model configurations and evaluation methodology. Section IV presents comprehensive results, including overall model performance comparison, train-test generalization analysis, feature importance rankings, stratified performance across LOS categories, clinical and operational implications of findings, and interprets feature importance patterns. Section V concludes with a summary of key findings and their implications for healthcare operations management and outlines future research directions.

II. Related Work and State-of-the-Art

In the healthcare field specifically, machine learning has emerged as a transformative technology for addressing longstanding challenges in clinical practice [8], operational management [9], and public health surveillance [10]. The ability of ML algorithms to process vast quantities of heterogeneous data, including electronic health records, medical imaging, laboratory results, genomic sequences, and administrative databases, enables the extraction of actionable insights that inform evidence-based decision-making at individual patient and population health levels [6,24]. Unlike traditional rule-based systems or simple statistical models, ML approaches can automatically learn complex, non-linear relationships between hundreds or thousands of variables without requiring explicit programming of decision rules, adapting to new patterns as additional data becomes available [6,7].

The application of ML in healthcare has expanded dramatically over the past decade, encompassing diverse domains with demonstrated clinical value and operational impact [7]. In medical imaging, convolutional neural networks (CNNs) and ensemble methods have achieved expert-level performance in cancer detection from radiology and pathology images, with applications spanning breast cancer screening from mammography [25], lung nodule detection from CT scans [26], diabetic retinopathy identification from fundus photographs [27], and skin lesion classification from dermoscopic images [28]. These systems not only match or exceed human radiologist performance in controlled settings but also demonstrate potential for reducing diagnostic delays, improving screening accessibility in resource-limited settings, and providing decision support for less experienced clinicians.

Hospital readmission prediction represents another critical application domain where ML has shown substantial promise for reducing healthcare costs and improving patient outcomes [29,30]. Predictive models utilizing patient demographics, comorbidities, prior utilization history, laboratory values, and medication regimens can identify high-risk patients requiring intensive post-discharge interventions such as home health visits, telephonic follow-up, or expedited outpatient appointments [31]. By enabling proactive risk stratification, these models support transition care management programs that have demonstrated reductions in 30-day readmission rates, particularly for conditions with historically high readmission burdens, including heart failure, chronic obstructive pulmonary disease, and pneumonia [31].

In pharmaceutical operations and supply chain management, ML algorithms optimize inventory levels, predict drug demand fluctuations, identify counterfeit medications through pattern analysis, and forecast medication adherence patterns to enable targeted interventions [32-34]. These applications address critical challenges in medication availability, cost containment, and therapeutic effectiveness, with particular relevance for specialty pharmaceuticals with high costs and narrow therapeutic windows. Clinical decision support systems (CDSS) powered by ML provide real-time guidance to clinicians on diagnosis, treatment selection, medication dosing, and adverse event prediction [35,36]. Modern CDSS implementations leverage ML to personalize recommendations based on individual patient characteristics rather than applying one-size-fits-all clinical guidelines, accounting for factors such as genetic variations affecting drug metabolism, comorbidity interactions modifying treatment effectiveness, and patient preferences influencing adherence likelihood [36-38].

Additional healthcare ML applications include sepsis prediction from vital sign trends and laboratory trajectories enabling early intervention before clinical deterioration; mortality risk estimation supporting goals-of-care discussions and resource allocation decisions; surgical complication forecasting informing preoperative risk stratification and patient counseling; disease progression modeling for chronic conditions enabling personalized monitoring schedules; and resource utilization prediction supporting capacity planning, staff scheduling, and financial forecasting [37,39]. Each of these applications demonstrates ML's capability to extract predictive signals from complex clinical data that traditional approaches struggle to model effectively.

Among machine learning algorithms, ensemble methods have demonstrated particular effectiveness for healthcare applications involving tabular data [2,6]. Within the ensemble learning paradigm, eXtreme Gradient Boosting (XGBoost) has emerged as a dominant algorithm for tabular data prediction, demonstrating superior performance across various healthcare applications, including disease risk prediction [40,41], operational outcome modeling [42], and cyberattack detection in the Internet of Medical Things (IoMT) for data privacy and confidentiality efficiency [43]. XGBoost extends traditional gradient boosting with several algorithmic innovations, regularization mechanisms that prevent overfitting by penalizing model complexity; a sparsity-aware algorithm for efficient handling of missing values by learning optimal default directions for tree splits; second-order gradient information in the objective function enabling more accurate optimization; parallel processing capabilities supporting scalability to large datasets; and tree pruning strategies that build trees to maximum depth then prune backward, identifying optimal tree structures more efficiently than traditional forward-stopping approaches [16].

These features make XGBoost particularly well-suited for clinical datasets characterized by heterogeneous features spanning continuous laboratory values, categorical diagnosis codes, binary indicators, and ordinal severity scores; missing observations due to selective test ordering based on clinical suspicion rather than systematic assessment; and complex non-linear relationships including feature interactions (e.g., age modifying disease severity impact) and threshold effects (e.g., laboratory values exhibiting different risk associations above versus below clinical cutoffs) [44]. The algorithm's built-in cross-validation, feature importance quantification, and interpretability through SHAP (SHapley Additive exPlanations) values further enhance its applicability to healthcare settings where model transparency and clinical interpretability are essential for adoption [16].

Hospital length of stay (LOS), the interval between admission and discharge, represents a critical outcome variable affecting both patient welfare and healthcare system efficiency [17,18]. Accurate LOS prediction enables multiple operational improvements, proactive bed management allowing anticipatory patient flow planning rather than reactive responses to capacity constraints; optimized discharge planning with early engagement of case managers and coordination of post-acute care services; evidence-based staffing decisions aligning nursing and ancillary service levels with anticipated patient volumes; improved patient and family communication through realistic expectation-setting regarding hospitalization duration; and financial forecasting supporting budget planning and resource allocation decisions [17].

Prior research on LOS prediction has employed diverse methodological approaches with varying degrees of success. Traditional statistical methods, including Bayesian belief networks (BBNs) [45], and Cox proportional hazards models treating discharge as a time-to-event outcome [46] provide interpretable coefficients, but struggle to capture non-linear relationships and complex feature interactions characteristic of clinical data.

More recent work has focused on ensemble methods and deep learning architectures. Random forest models for LOS prediction [11,21,22] have shown robust performance across multiple clinical domains, including cardiology, oncology, and general medicine, with feature importance analysis identifying key predictors such as admission diagnosis, comorbidity burden, and initial laboratory abnormalities. Gradient boosting approaches [19,20,47] have achieved competitive performance while offering computational efficiency advantages.

XGBoost specifically has been applied to LOS prediction in several clinical contexts. In a study by Chen et al. [15], the XGBoost algorithm was applied as a multi-classification model to predict ischemic stroke patients' length of hospital stay across three categories: 1–7, 8–14, or >14 days. The model utilized 28 clinical attributes and achieved a high-performance accuracy of 0.89 while identifying key predictors like NIHSS scores, coma indices, and surgical status [15]. Another study by Zeleke et al. [19], XGBoost was one of eight regression models used to predict the exact LoS for patients admitted through the emergency department. Thus, XGBoost and Ridge regressions showed the best performance by minimizing prediction error. A study by Hasan et al. [14] on ICU patients using the MIMIC-III database applied XGBoost for predicting patient length of stay, achieving an R^2 of 0.86 and an RMSE of 1.2, outperforming Random Forest and SVM models. XGBoost was applied as a classification model in a study of Chang et al. [13] to identify low-severity "triage level 3" patients predicted to have a discharge length of stay under 4 hours, using only information available at the triage stage. XGBoost demonstrated the best performance in external validation with an AUC of 0.761, suggesting its potential use for real-time patient streaming in emergency departments [13]. Furthermore, XGBoost was applied to predict postoperative hospital length of stay using 1,433 admission variables from Japan's DPC database, achieving a Mean Absolute Error of 2.82 days [12]. The model identified surgical procedure type and hospital volume as key predictors. However, comprehensive evaluation across large, diverse inpatient populations representing the full spectrum of medical and surgical conditions remains limited, with most studies focusing on specific clinical subpopulations or single institutional datasets.

III. Materials and Methods

A. Dataset Description and Preprocessing

This study utilized the New York Statewide Planning and Research Cooperative System (SPARCS) inpatient de-identified dataset for the year 2023, publicly available through the New York State Department of Health [23]. The original dataset comprised 2,125,754 hospital admission records across 33 variables, encompassing patient demographics, clinical characteristics, administrative information, and financial data from hospitals throughout New York State.

Data preprocessing involved systematic removal of records with missing values in critical clinical variables. Specifically, 474 records (0.02%) with missing values in APR (All Patient Refined) Severity of Illness Description and APR Risk of Mortality were excluded, resulting in a final analytical dataset of 2,125,280 records (99.98% retention rate).

The Length of Stay variable, originally stored as string format, was converted to integer type. Records with extreme outliers (LOS > 120 days, representing 0.01% of cases) were retained to preserve real-world distributional characteristics.

From the original 33 variables, 23 were retained based on clinical relevance, data completeness, and non-redundancy. Excluded variables included facility identifiers (Operating Certificate Number, Permanent Facility Id, Facility Name), redundant descriptive fields (CCSR Diagnosis/Procedure Descriptions, APR DRG/MDC Descriptions), and highly incomplete variables (Birth Weight, Payment Typologies 2 and 3, Zip Code).

Four derived features were created to enhance predictive capability. First, Cost_Per_Day was calculated as Total Costs divided by Length of Stay, representing daily resource intensity. Second, Charge_Cost_Ratio was computed as Total Charges divided by Total Costs, reflecting hospital pricing practices. Third, Is_Emergency as a Binary indicator derived from Emergency Department Indicator (Y=1, N=0). And, Is_Surgical is another Binary indicator extracted from the APR Medical Surgical Description (Surgical=1, Medical=0).

The Cost_Per_Day was subsequently excluded from modeling as it is outcome-derived and requires knowledge of LOS, leaving 21 final predictive features.

Categorical variables were encoded using scikit-learn's LabelEncoder to convert string categories into numerical representations. The following 15 categorical features underwent label encoding: Hospital Service Area, Hospital County, Age Group, Gender, Race, Ethnicity, Type of Admission, Patient Disposition, CCSR Diagnosis Code, APR DRG Code, APR MDC Code, APR Risk of Mortality, APR Medical Surgical Description, Payment Typology 1, and Emergency Department Indicator. This encoding strategy preserved ordinal relationships where present (e.g., Age Groups, Severity levels) while enabling tree-based algorithms to efficiently handle categorical splits.

No feature scaling was applied to the full feature set, as tree-based ensemble methods are scale-invariant. However, for Linear Regression, features were standardized using StandardScaler to prevent scale-dependent coefficient bias.

The final analytical dataset consisted of 2,125,280 records × 23 variables, with 21 predictive features and 1 target variable (Length of Stay). All features were numerical with zero missing values. The target variable exhibited right-skewed distribution with mean=5.78 days, median=3.00 days, standard deviation=8.78 days, and range=1-120 days. A summary of the preprocessing process on the dataset is provided in Table I.

TABLE I. DATASET CHARACTERISTICS AND PREPROCESSING SUMMARY

State	Characteristic	Value
Original Dataset	**Total records**	2,125,754
	Total variables	33
	Study period	2023
	Data source	NY SPARCS Inpatient Database
After Data Cleaning	**Final records**	2,125,280 (99.98% retention)
	Final variables	23
	Records excluded	474 (0.02%)
	Missing values	0
Target Variable (Length of Stay)	**Mean ± SD**	5.78 ± 8.78 days
	Median (IQR)	3.00 (2.00-6.00) days
	Range	1-120 days
	Mode	2 days (21.5%)
Feature Types	**Categorical features**	15
	Numerical features	6
	Engineered features	4
	Total predictive features	21

B. Experimental Setup

To balance computational efficiency with statistical robustness, a stratified sampling approach was employed. From the full dataset of 2,125,280 records, 500,000 records (23.5%) were randomly sampled while maintaining the original distribution of the target variable. This subset was partitioned into training (70%, n=350,000), validation (15%, n=75,000), and test (15%, n=75,000) sets using scikit-learn's train_test_split function. The validation set was used for hyperparameter optimization in the XGBoost (Optimized) model, while the test set remained completely held out for final performance evaluation. All splits used random_state=42 for reproducibility.

Model performance was assessed using multiple complementary metrics to capture different aspects of prediction quality, such as coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and prediction accuracy bands. Overfitting was assessed by comparing training and test performance across R^2 and RMSE metrics, with train-test gaps quantifying model generalization.

C. Regression Models

1) Linear Regression: Ridge regression with L2 regularization served as the baseline model, implemented using scikit-learn with default regularization parameter (α=1.0). Features were standardized prior to training using StandardScaler to ensure equal contribution scales. The model was fit using ordinary least squares with a ridge penalty to prevent overfitting in the presence of multicollinearity. Its parameter configuration is listed in Table II. The regression model for Length of Stay prediction can be expressed as:

$$\hat{Y} = \beta_0 + \sum_{i=1}^{21} \beta_i \cdot X_i + \varepsilon \qquad i = 1, \ldots, 21 \qquad (1)$$

where, $\hat{Y}$ is the predicted length of stay (in days), β_0 is the intercept term, β_i is coefficient for feature I, X_i is the standardized value of feature i, and ε is the error term.

2) Random Forest regressor: Random Forest, an ensemble of decision trees trained on bootstrap samples with random feature subsets, was implemented using scikit-learn. Its parameter configuration is listed in Table II. The model aggregates predictions from multiple trees to reduce variance while maintaining low bias through the ensemble averaging mechanism.

3) Gradient Boosting regressor: Gradient Boosting builds an ensemble by sequentially adding trees that correct residual errors of preceding trees, implemented using scikit-learn with a squared error loss function. Each new tree is fit to the negative gradient of the loss function. Its parameter configuration is listed in Table II.

TABLE II. MODEL HYPERPARAMETERS CONFIGURATION

Hyperparameter	LR[a]	RF[b]	GB[c]	XGB[d]	XGB_opt[e]
n_estimators	N/A	100	100	100	300
max_depth	N/A	15	6	6	10
learning_rate	N/A	N/A	0.1	0.1	0.05
alpha (L2)	1.0	N/A	N/A	N/A	N/A
reg_alpha (L1)	N/A	N/A	N/A	0 (default)	0.1
reg_lambda (L2)	N/A	N/A	N/A	1.0 (default)	1.0
gamma	N/A	N/A	N/A	0 (default)	0.1
min_samples_split	N/A	10	10	N/A	N/A
min_samples_leaf	N/A	5	N/A	N/A	N/A
min_child_weight	N/A	N/A	N/A	1 (default)	5
subsample	N/A	N/A	N/A	1.0 (default)	0.8
colsample_bytree	N/A	N/A	N/A	1.0 (default)	0.8

[a] Linear Regression

[b] Random Forest

[c] Gradient Boosting

[d] XGBoost with default configuration

[e] XGBoost with optimized configuration

4) XGBoost: XGBoost (eXtreme Gradient Boosting) implements an optimized distributed gradient boosting framework with advanced regularization, parallel processing, and efficient handling of sparse data. The default configuration used standard hyperparameters (Table III), while the optimized configuration (Table II) used parameter selection focusing on balancing model complexity (tree depth, number of estimators), learning dynamics (learning rate), sampling strategies (subsample ratios), and regularization (gamma, L1/L2 penalties).

All models were trained on the 350,000 record training set with full convergence to specified iteration limits. Training time varied substantially across algorithms, Linear Regression (Ridge) with 0.07 seconds, default XGBoost with 2.25 seconds, optimized XGBoost with 11.31 seconds, Random Forest with 59.49 seconds, and Gradient Boosting with 220.17 seconds.

Model predictions were generated for both training (n=350,000) and test (n=75,000) sets to assess overfitting through train-test performance comparison. All models were evaluated using identical test data to ensure fair comparison.

Feature importance for the best-performing model (XGBoost Optimized) was quantified using SHAP (SHapley Additive exPlanations) values, a unified framework for interpreting predictions based on cooperative game theory. SHAP values represent each feature's contribution to the prediction for individual instances, with positive values pushing predictions higher and negative values lower.

The analysis was conducted on a random sample of 5,000 test set instances (6.7% of the test set) to balance computational feasibility with statistical stability. Mean absolute SHAP values across all samples provided global feature importance rankings, identifying the primary drivers of length of stay predictions. This approach offers advantages over traditional feature importance metrics by accounting for feature interactions and providing locally faithful explanations.

IV. RESULTS AND DISCUSSION

A. ML Models' Performance

The comparative analysis of five machine learning algorithms applied to 2,125,280 hospital admission records from the New York SPARCS dataset revealed substantial differences in predictive performance for length of stay (LOS) estimation. As illustrated in Fig. 1 and summarized in Fig. 2 and Table III, the optimized version of the XGBoost model achieved superior performance across all evaluation metrics, with an R^2 of 0.8686, RMSE of 3.24 days, and MAE of 1.42 days. This represents a remarkable improvement over traditional linear regression (R^2=0.5339, RMSE=6.10 days, MAE=2.86 days), demonstrating that the complex, non-linear relationships inherent in hospital length of stay are better captured by ensemble tree-based methods.

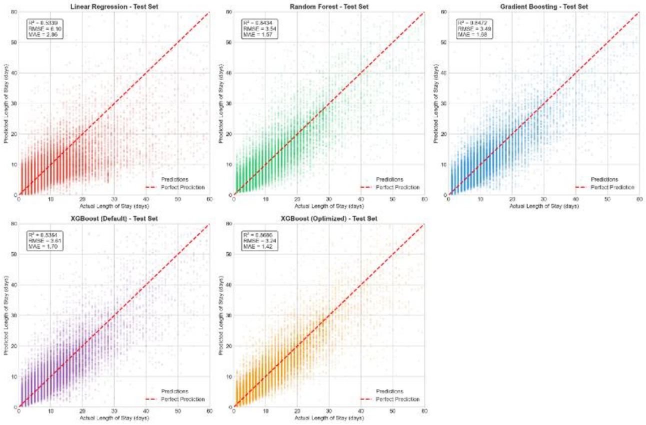

Fig. 1. Actual vs. Predicted length of stay of selected models on test set.

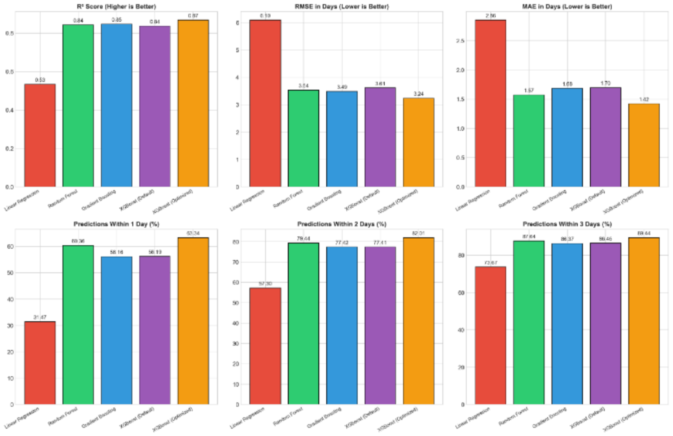

Fig. 2. Comprehensive model comparison on the test set.

TABLE III. MODEL PERFORMANCE ON TRAINING VS. TEST SETS

Model	Train R^2	Test R^2	ΔR^2	Train RMSE	Test RMSE	ΔRMSE
Linear Regression	0.5472	0.5339	0.0133	5.98	6.10	0.12
Gradient Boosting	0.8754	0.8472	0.0282	3.14	3.49	0.35
XGBoost (Default)	0.8649	0.8364	0.0285	3.27	3.61	0.34
Random Forest	0.9078	0.8434	0.0644	2.70	3.54	0.84
XGBoost (Optimized)	0.9503	0.8686	0.0817	1.98	3.24	1.26

The clinical significance of these results becomes apparent when examining the prediction accuracy bands shown in Fig. 2. The optimized version of the XGBoost model achieved 63.34% of predictions within ±1 day of actual LOS, 82.01% within ±2 days, and 89.44% within ±3 days. In contrast, linear regression achieved only 31.47% accuracy within ±1 day and 57.30% within ±2 days. For a healthcare system with an average LOS of 5.78 days, the ability to predict within 1.42 days (MAE) represents approximately 24.6% relative error, which is operationally valuable for bed management, discharge planning, and resource allocation. This level of accuracy enables hospital administrators to make data-driven decisions with reasonable confidence intervals, potentially reducing bottlenecks in patient flow and improving overall throughput efficiency.

The three ensemble methods (Random Forest, Gradient Boosting, and XGBoost) clustered closely in performance, all achieving R^2 values above 0.84, while substantially outperforming linear regression. Random Forest achieved R^2=0.8434 (RMSE=3.54 days, MAE=1.57 days), Gradient Boosting reached R^2=0.8472 (RMSE=3.49 days, MAE=1.68 days), and the optimized version of XGBoost surpassed at R^2=0.8686 (RMSE=3.24 days, MAE=1.42 days). The scatter plots in Fig. 1 reveal that all ensemble methods produce predictions that cluster tightly around the perfect prediction line (red dashed diagonal), whereas linear regression shows considerably more scatter, particularly for longer stays exceeding 20 days.

Interestingly, the default XGBoost configuration (R^2=0.8364, RMSE=3.61 days, MAE=1.70 days) performed slightly below both Random Forest and Gradient Boosting, highlighting the critical importance of hyperparameter optimization in achieving optimal model performance. The improvement from default to optimized XGBoost (ΔR^2=0.0322, ΔRMSE=0.37 days, ΔMAE=0.28 days) demonstrates that careful tuning of learning rate, max depth, min child weight, and regularization parameters can yield clinically meaningful improvements. This 0.28-day reduction in MAE, while seemingly modest, translates to approximately 20% improvement in prediction error and could impact thousands of patient flow decisions across a large hospital system annually.

The residual distributions presented in Fig. 3 provide critical insights into model behavior and potential biases. All ensemble methods exhibited nearly unbiased predictions with mean residuals close to zero (Random Forest: -0.000, Gradient Boosting: 0.003, XGBoost Optimized: 0.001, XGBoost Default: 0.006), while linear regression showed a slight positive bias (mean=0.021). The standard deviations of residuals followed the same performance hierarchy as RMSE values, with the optimized version of XGBoost showing the tightest distribution (std=3.238), followed by Gradient Boosting (std=3.492), Random Forest (std=3.535), XGBoost Default (std=3.613), and Linear Regression (std=6.099).

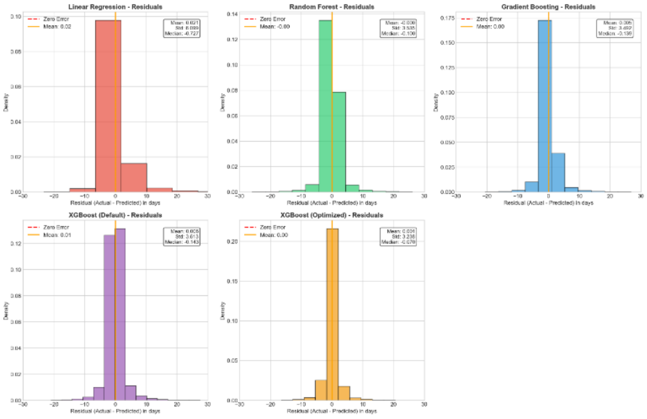

Fig. 3. Histograms of prediction residuals (actual LOS - predicted LOS, in days) for the studied models on the test set.

The nearly symmetric, approximately normal distribution of residuals for ensemble methods indicates that these models do not systematically over- or under-predict across the LOS spectrum. This is particularly important in healthcare applications where both types of errors carry consequences:

- Underpredicting LOS may result in premature discharge planning and resource shortages,
- While overpredicting unnecessarily occupies beds and resources.

The median residuals close to zero (XGBoost Optimized: -0.070, Gradient Boosting: 0.139, Random Forest: -0.100) further confirm the lack of systematic bias. Linear regression, however, shows a broader residual distribution with longer tails, suggesting inconsistent prediction quality across different patient populations and clinical scenarios.

The train-test performance comparison, illustrated in Fig. 4, reveals important patterns regarding model generalization. Linear regression demonstrated minimal overfitting with nearly identical train and test R^2 values (0.547 vs. 0.534, gap=0.013), consistent with its fundamental simplicity and high bias. However, this stability comes at the cost of poor overall performance, making it unsuitable for practical deployment despite its generalization properties.

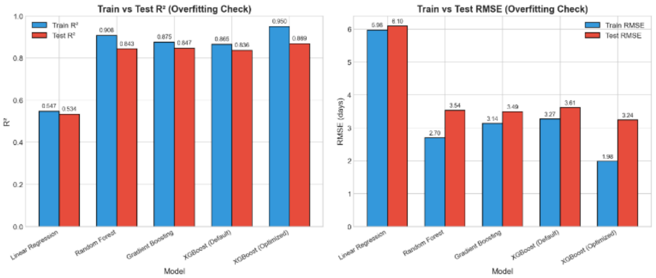

Fig. 4. Comparison of training and test set performance for the studied models using R² scores and RMSE values.

Among ensemble methods, Gradient Boosting showed the best generalization characteristics with a train-test R^2 gap of only 0.0282 (0.8754 train vs 0.8472 test) and RMSE difference of 0.35 days (3.14 train vs 3.49 test). Random Forest exhibited moderate overfitting (R^2 gap=0.0644, RMSE difference=0.84 days), while XGBoost configurations showed varying degrees. The default XGBoost demonstrated excellent generalization (R^2 gap=0.0285, RMSE difference=0.34 days), comparable to Gradient Boosting. Notably, the optimized XGBoost, despite achieving the best test performance, exhibited the largest overfitting signature (R^2 gap=0.0817, RMSE difference=1.26 days), with train R^2=0.9503 versus test R^2=0.8686.

This overfitting pattern in the optimized XGBoost model warrants careful interpretation in the healthcare context. While the train R^2 of 0.9503 suggests nearly perfect fit to training data, the test performance of R^2=0.8686 still substantially exceeds all other models, including the less-overfitted Gradient Boosting (R^2=0.8472). This indicates that the hyperparameter optimization process successfully identified genuine predictive patterns rather than merely fitting noise. However, from a clinical deployment perspective, this overfitting raises important considerations:

- The model may have learned hospital-specific patterns,
- Temporal trends specific to the training period,
- Or subtle interactions between features that may not generalize to future time periods or different healthcare settings.

Prospective validation on out-of-time test sets and external validation on data from other hospital systems would be essential before widespread deployment.

The analysis of prediction errors across different LOS groups, presented in Fig. 5, reveals critical insights about model performance heterogeneity. The optimized version of XGBoost model demonstrated strongly differential performance across stay duration categories, with MAE systematically increasing from 0.66 days for 1-3 day stays to 11.81 days for stays exceeding 30 days. This more than 17-fold increase in absolute error highlights a fundamental challenge in healthcare predictive modeling: short stays are inherently more predictable than extended hospitalizations, which often involve complications, comorbidities, and unpredictable clinical trajectories.

The box plots of residuals in Fig. 5 illustrate this heterogeneity visually. For short stays (1-3 days and 4-7 days), predictions cluster tightly around zero error with narrow interquartile ranges and few outliers, indicating consistent, reliable predictions. The median residuals remain close to zero across all groups, confirming the absence of systematic bias. However, as LOS increases, the distribution widens dramatically, with the >30 days category showing extensive scatter and numerous extreme outliers extending beyond ±50 days. This pattern reflects the clinical reality that patients with prolonged hospitalizations often experience complex, unpredictable courses involving multiple complications, ICU transfers, infections, or delayed discharge due to placement issues.

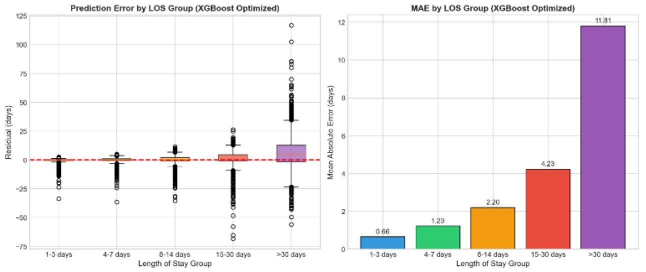

Fig. 5. Prediction error by the LOS Group of the optimized XGBoost model.

From a healthcare operations perspective, these findings suggest differentiated deployment strategies for LOS prediction models. For routine cases expected to have short stays, the model's high accuracy (MAE=0.66 days for 1-3 day group, 1.23 days for 4-7 day group) enables confident operational planning for bed turnover, same-day discharges, and routine resource allocation. For intermediate stays (8-14 days: MAE=2.20 days; 15-30 days: MAE=4.23 days), predictions provide useful guidance but should be accompanied by wider confidence intervals in planning tools. For extended stays beyond 30 days (MAE=11.81 days), the model's limited accuracy suggests it should serve as a baseline estimate rather than a precise planning tool, with frequent re-prediction as the hospitalization progresses and more clinical information becomes available. This adaptive prediction approach, updating forecasts daily as new clinical data emerges, could substantially improve accuracy for complex cases.

B. Feature Importance

Before interpreting the feature importance results, an important methodological caveat must be stated. The two highest-ranked predictors: Total Costs and Total Charges, are financial variables that accumulate during the hospital stay and are not available at the time of patient admission. Their prominence in the SHAP rankings, therefore, reflects the model's ability to exploit outcome-correlated proxies rather than purely independent clinical signals. This constitutes a form of target leakage that limits the model's applicability to prospective, admission-time prediction. Readers should interpret all subsequent importance rankings with this constraint in mind, and the clinical implications discussed below should be understood in the context of retrospective or near-discharge prediction scenarios rather than early forecasting.

The feature importance analysis, presented in Fig. 6, provides interpretable insights into the clinical and administrative factors driving LOS predictions. The dominance of financial variables, Total Costs (mean |SHAP value| ≈ 2.6) and Total Charges (≈ 1.5), as top predictors is particularly noteworthy. While these variables are highly predictive, they represent outcome proxies rather than truly independent predictors, as costs accumulate over the course of hospitalization. This finding has important implications for real-time prediction systems; at admission, these values are unknown, limiting the model's utility for early LOS forecasting unless historical cost patterns or initial procedure costs are used as proxies.

Hospital Service Area emerged as the third most important feature (mean |SHAP value| ≈ 1.0), reflecting substantial variation in LOS across different clinical departments and specialties. This aligns with clinical intuition, as surgical units, medical wards, intensive care units, and specialty services have fundamentally different patient acuity levels and treatment protocols. The importance of this feature suggests that department-specific LOS prediction models might offer further accuracy improvements by capturing unit-specific workflows and patient populations. The APR Medical Surgical Description and APR DRG (Diagnosis Related Group) Code features (mean |SHAP values| ≈ 0.7 and 0.5 respectively) capture the primary reason for hospitalization and its severity, representing core clinical determinants of resource utilization and recovery time.

Geographic factors, represented by Hospital County (mean |SHAP value| ≈ 0.5), ranked as the sixth most important predictor, potentially reflecting regional variations in practice patterns, socioeconomic factors affecting patient populations, availability of post-acute care facilities, and hospital-specific policies. The Charge-to-Cost Ratio (≈ 0.4) and Patient Disposition (≈ 0.4) provide complementary information about hospital billing practices and discharge destinations (home, skilled nursing facility, rehabilitation, etc.), both of which correlate with case complexity and LOS.

Notably, demographic factors showed relatively modest importance, with Age Group ranking lowest among all features (mean |SHAP value| < 0.2). This suggests that once clinical condition, procedures, and severity are accounted for, age contributes minimally to LOS variability. Similarly, APR Risk of Mortality, Type of Admission, Emergency Department Indicator, surgical status (Is_Surgical), and APR Severity of Illness Code all showed relatively small contributions (mean |SHAP values| 0.2-0.4). This ranking challenges some clinical assumptions about LOS drivers and suggests that while these factors influence outcomes and resource intensity, their incremental predictive value beyond procedure codes and service area is limited. This insight could guide data collection priorities and inform which variables are most critical for accurate LOS prediction in resource-constrained settings.

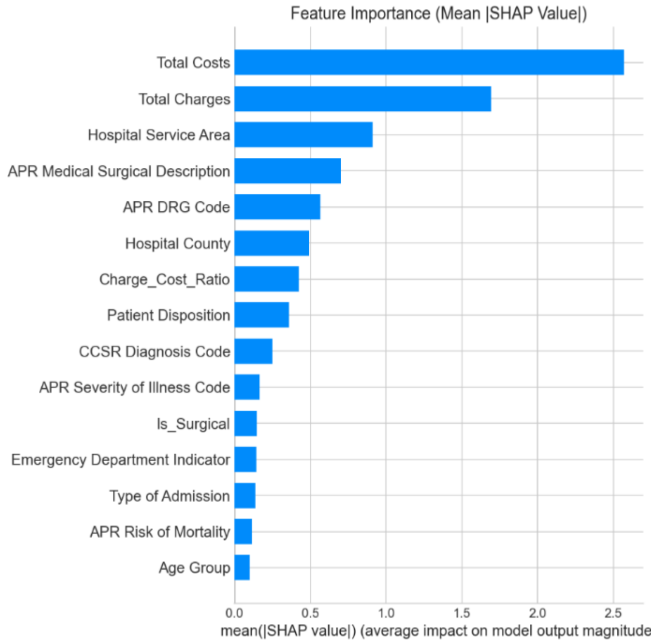

Fig. 6. Global feature importance for the optimized XGBoost model using mean absolute SHAP values on test set predictions.

C. Results Synthesis

The practical implications of these findings for healthcare delivery are multifaceted and significant. First, the high accuracy achieved by the optimized version of the XGBoost model for the majority of cases (those with LOS < 15 days, representing the bulk of admissions) enables several operational improvements. Hospital bed management systems could integrate these predictions to forecast bed availability 24-72 hours in advance with reasonable accuracy, allowing proactive management of elective admissions and emergency department patient flow. Discharge planning coordinators could receive automated alerts identifying patients likely to require post-acute care placement, enabling earlier engagement of care coordination resources when placement is most feasible.

Second, the feature importance analysis suggests actionable insights for targeted interventions. The strong predictive signal from the Hospital Service Area indicates that department-specific initiatives to reduce LOS variation could be monitored and evaluated using these models. Departments with systematically longer-than-predicted stays could be flagged for process improvement initiatives. The prominence of diagnosis and procedure codes validates current payment systems based on DRGs but also suggests opportunities for risk adjustment in benchmarking and quality reporting.

Third, the model's differential performance across LOS categories has important implications for resource allocation. Hospitals could segment their prediction systems by expected stay duration, applying the highly accurate model for routine cases to optimize daily operations, while implementing more conservative planning assumptions and clinical judgment for complex cases likely to exceed 30 days. This stratified approach acknowledges the fundamental differences in predictability across case types and avoids over-reliance on model predictions where uncertainty is high.

Fourth, the minimal contribution of age and other demographic factors to prediction accuracy, once clinical variables are included, suggests that concerns about age-based discrimination in algorithmic decision-making may be less salient for LOS prediction than for other healthcare AI applications. However, this finding should not be interpreted as evidence that demographic factors are clinically unimportant; rather, it indicates that their effects are largely mediated through clinical presentations and procedures, which the model captures through other variables.

Several methodological aspects of this study merit discussion. The large sample size (2,125,280 records after cleaning) provides robust statistical power and enables the detection of subtle patterns in LOS variation. The minimal missing data (only 0.25% missingness in Hospital Area/County, with no missing values after preprocessing) eliminates concerns about imputation bias affecting results. The 70/15/15 (train/validation/test) split follows standard practice, though the temporal ordering of data splits is not specified in the provided information. If data were split randomly without considering admission dates, the model may have inadvertently "learned" from future time periods to predict past cases, potentially inflating performance estimates. Ideally, temporal validation (training on earlier time periods, testing on later periods) would provide more rigorous evidence of prospective prediction capability.

The comparison of five distinct algorithms, ranging from simple linear regression to optimized gradient boosting, provides comprehensive evidence that ensemble tree-based methods substantially outperform traditional approaches for this application. The inclusion of both default and optimized XGBoost configurations demonstrates the value of hyperparameter tuning, while the residual analysis confirms that superior test metrics reflect genuine predictive improvement rather than distributional artifacts. The use of multiple complementary metrics (R^2, RMSE, MAE, prediction accuracy bands) provides a comprehensive assessment of model performance from different perspectives relevant to clinical decision-making.

The SHAP feature importance analysis represents a methodological strength, providing interpretable insights into model predictions rather than treating the model as a black box. This interpretability is crucial for clinical acceptance and regulatory approval of AI systems in healthcare. However, the strong dependence on Total Costs and Total Charges raises questions about whether the model is learning causal relationships or merely detecting correlations with downstream outcomes. A more rigorous causal analysis, potentially using only admission-available variables or implementing a time-aware feature selection process, might reveal different importance rankings and could improve the model's utility for admission-time prediction.

V. Conclusion

This study demonstrates that ensemble machine learning methods, particularly optimized XGBoost, substantially outperform traditional approaches for hospital length of stay prediction. Analyzing 2,125,280 admission records, our best model achieved R^2=0.8686 and MAE=1.42 days, with 63.34% of predictions within ±1 day of actual LOS. This accuracy level enables actionable operational planning, including proactive bed management, early discharge planning, and optimized resource allocation.

The performance hierarchy, with ensemble methods achieving R^2>0.84 versus 0.53 for Linear Regression, confirms that LOS prediction requires sophisticated algorithms to capture complex, non-linear relationships. Feature importance analysis revealed financial variables, Hospital Service Area, and clinical coding as top predictors, while demographic factors contributed minimally. Performance varied substantially across LOS categories (MAE: 0.66 days for 1-3 day stays versus 11.81 days for >30 days), indicating excellent accuracy for routine cases but limited predictive value for complex, extended hospitalizations.

Key limitations include geographic restriction to New York hospitals, reliance on outcome-derived financial variables limiting prospective forecasting utility, poor performance for extended stays (>30 days), and lack of external validation. Future work should prioritize external validation across diverse healthcare settings, development of specialized models for complex cases, and prospective implementation trials evaluating clinical impact. Despite these constraints, machine learning-based LOS prediction offers substantial potential for improving healthcare operational efficiency and resource management. A key constraint on the prospective deployment of the proposed model concerns the reliance on financial variables (specifically, Total Costs and Total Charges), which rank as the two most important predictors according to SHAP analysis. These variables are accumulated progressively over the course of hospitalization and are therefore unavailable at the time of admission. Consequently, the model in its current form is better suited for retrospective analysis and post-hoc resource planning than for real-time, admission-time LOS forecasting. Any operational deployment in a clinical setting would require either the exclusion of these variables or their replacement with admission-available proxies, such as estimated procedure costs derived from billing schedules or historical average costs per diagnosis-related group (DRG). Future work should evaluate a restricted model trained exclusively on admission-available features to quantify the performance trade-off and assess its practical utility for early-stage clinical decision-making.

Acknowledgment

The authors would like to acknowledge the New York State Department of Health for providing access to the Statewide Planning and Research Cooperative System (SPARCS) data through the Health Data NY open data platform. This study used the Hospital Inpatient Discharges (SPARCS De-Identified) dataset for calendar year 2023, which is publicly available at https://health.data.ny.gov

The views expressed in this publication are those of the authors and do not necessarily reflect the official policy or position of the New York State Department of Health.

REFERENCES

[1] P. C. Verhoef et al., "Digital transformation: A multidisciplinary reflection and research agenda," J. Bus. Res., vol. 122, pp. 889–901, Jan. 2021, doi: 10.1016/J.JBUSRES.2019.09.022.

[2] M. Wazid, J. Singh, A. K. Das, and J. J. P. C. Rodrigues, "An Ensemble-Based Machine Learning-Envisioned Intrusion Detection in Industry 5.0-Driven Healthcare Applications," IEEE Transactions on Consumer Electronics, vol. 70, no. 1, pp. 1903–1912, Feb. 2024, doi: 10.1109/TCE.2023.3318850.

[3] C. H. Lee, C. Wang, X. Fan, F. Li, and C. H. Chen, "Artificial intelligence-enabled digital transformation in elderly healthcare field: Scoping review," Advanced Engineering Informatics, vol. 55, p. 101874, Jan. 2023, doi: 10.1016/J.AEI.2023.101874.

[4] K. Sharifani and M. Amini, "Machine Learning and Deep Learning: A Review of Methods and Applications," 2023. Accessed: Jan. 30, 2026. [Online]. Available: https://papers.ssrn.com/abstract=4458723

[5] M. Javaid, A. Haleem, R. Pratap Singh, R. Suman, and S. Rab, "Significance of machine learning in healthcare: Features, pillars and applications," International Journal of Intelligent Networks, vol. 3, pp. 58–73, Jan. 2022, doi: 10.1016/J.IJIN.2022.05.002.

[6] A. Zhang, L. Xing, J. Zou, and J. C. Wu, "Shifting machine learning for healthcare from development to deployment and from models to data," Nature Biomedical Engineering 2022 6:12, vol. 6, no. 12, pp. 1330–1345, Jul. 2022, doi: 10.1038/s41551-022-00898-y.

[7] Q. An, S. Rahman, J. Zhou, and J. J. Kang, "A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges," Sensors 2023, Vol. 23, vol. 23, no. 9, Apr. 2023, doi: 10.3390/S23094178.

[8] A. Rajkomar, J. Dean, and I. Kohane, "Machine Learning in Medicine," New England Journal of Medicine, vol. 380, no. 14, pp. 1347–1358, Apr. 2019, doi: 10.1056/NEJMRA1814259;ISSUE:ISSUE:DOI.

[9] A. Duddalwar and P. Khobragade, "An Optimization of Healthcare Operation Management Using Machine Learning," Lecture Notes in Electrical Engineering, vol. 1281 LNEE, pp. 439–453, 2025, doi: 10.1007/978-981-97-8422-6_36.

[10] F. Ekundayo, "Using machine learning to predict disease outbreaks and enhance public health surveillance," https://wjarr.com/sites/default/files/WJARR-2024-3732.pdf, vol. 24, no. 3, pp. 794–811, Dec. 2024, doi: 10.30574/WJARR.2024.24.3.3732.

[11] H. Liu, F. Xing, J. Jiang, Z. Chen, Z. Xiang, and X. Duan, "Random forest predictive modeling of prolonged hospital length of stay in elderly hip fracture patients," Front. Med. (Lausanne)., vol. 11, p. 1362153, May 2024, doi: 10.3389/FMED.2024.1362153/BIBTEX.

[12] T. Maruyama, K. Ikezawa, H. Suzuki, T. Kurokawa, Y. Akashi, and T. Oda, "Explainable machine learning for predicting postoperative length of stay after gastrectomy: a nationwide study using XGBoost and SHAP," Front. Med. Technol., vol. 7, p. 1732580, Dec. 2025, doi: 10.3389/FMEDT.2025.1732580/BIBTEX.

[13] Y. H. Chang et al., "Machine learning–based triage to identify low-severity patients with a short discharge length of stay in emergency department," BMC Emergency Medicine 2022 22:1, vol. 22, no. 1, pp. 88-, May 2022, doi: 10.1186/S12873-022-00632-6.

[14] M. N. Hasan, S. Hamdan, S. Poudel, J. Vargas, and K. Poudel, "Prediction of Length-of-stay at Intensive Care Unit (ICU) Using Machine Learning based on MIMIC-III Database," Proceedings - 2023 IEEE Conference on Artificial Intelligence, CAI 2023, pp. 321–323, 2023, doi: 10.1109/CAI54212.2023.00142.

[15] R. Chen et al., "A study on predicting the length of hospital stay for Chinese patients with ischemic stroke based on the XGBoost algorithm," BMC Medical Informatics and Decision Making 2023 23:1, vol. 23, no. 1, pp. 49-, Mar. 2023, doi: 10.1186/S12911-023-02140-4.

[16] H. Reddad, M. Zemzami, N. El Hami, and N. Hmina, "Machine Learning and Artificial Intelligence with XGBoost Algorithm for Binary Classification," Methods and Applications of Artificial Intelligence: Dynamic Response, Learning, Random Forest, Linear Regression, Interoperability, Additive Manufacturing and Mechatronics, pp. 161–192, Dec. 2025, doi: 10.1115/1.862MAA_CH7.

[17] K. Stone, R. Zwiggelaar, P. Jones, and N. Mac Parthaláin, "A systematic review of the prediction of hospital length of stay: Towards a unified framework," PLOS Digital Health, vol. 1, no. 4, p. e0000017, Apr. 2022, doi: 10.1371/JOURNAL.PDIG.0000017.

[18] S. Gokhale et al., "Hospital length of stay prediction tools for all hospital admissions and general medicine populations: systematic review and meta-analysis," Front. Med. (Lausanne)., vol. 10, p. 1192969, Aug. 2023, doi: 10.3389/FMED.2023.1192969/BIBTEX.

[19] A. J. Zeleke, P. Palumbo, P. Tubertini, R. Miglio, and L. Chiari, "Machine learning-based prediction of hospital prolonged length of stay admission at emergency department: a Gradient Boosting algorithm analysis," Front. Artif. Intell., vol. 6, p. 1179226, Jul. 2023, doi: 10.3389/FRAI.2023.1179226/BIBTEX.

[20] M. Suchithra, K. Shashwat, and M. Shoaib Khan, "Predicting Hospital Length of Stay Using Light Gradient Boosting Machine Regression," IFIP Adv. Inf. Commun. Technol., vol. 718 IFIPAICT, pp. 487–498, 2024, doi: 10.1007/978-3-031-69986-3_37.

[21] R. Jain, M. Singh, A. R. Rao, and R. Garg, "Predicting hospital length of stay using machine learning on a large open health dataset," BMC Health Services Research 2024 24:1, vol. 24, no. 1, pp. 860-, Jul. 2024, doi: 10.1186/S12913-024-11238-Y.

[22] N. Boff Medeiros, F. S. Fogliatto, M. Karla Rocha, and G. L. Tortorella, "Predicting the length-of-stay of pediatric patients using machine learning algorithms," Int. J. Prod. Res., vol. 63, no. 2, pp. 483–496, Jan. 2025, doi: 10.1080/00207543.2023.2235029;JOURNAL:JOURNAL:TPRS20;WGROUP:STRING:PUBLICATION.

[23] New York State Department of Health, "Hospital Inpatient Discharges (SPARCS De-Identified): 2023 [Data set]. Health Data NY. https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/46xm-urtu."

[24] H. Reddad, M. Zemzami, F. Yalaoui, N. Q. Nguyen, and C. Benmhamed, "EMR-NG: New-Generation Solutions for Healthcare," 2024 International Conference on Smart-Digital-Green Technologies and Artificial Intelligence Sciences, CSDGAIS 2024, 2024, doi: 10.1109/CSDGAIS64098.2024.11064782.

[25] S. M. McKinney et al., "International evaluation of an AI system for breast cancer screening," Nature 2020 577:7788, vol. 577, no. 7788, pp. 89–94, Jan. 2020, doi: 10.1038/s41586-019-1799-6.

[26] H. Mkindu, L. Wu, and Y. Zhao, "Lung nodule detection of CT images based on combining 3D-CNN and squeeze-and-excitation networks," Multimedia Tools and Applications 2023 82:17, vol. 82, no. 17, pp. 25747–25760, Mar. 2023, doi: 10.1007/S11042-023-14581-0.

[27] E. Özbay, "An active deep learning method for diabetic retinopathy detection in segmented fundus images using artificial bee colony algorithm," Artificial Intelligence Review 2022 56:4, vol. 56, no. 4, pp. 3291–3318, Aug. 2022, doi: 10.1007/S10462-022-10231-3.

[28] B. Shetty, R. Fernandes, A. P. Rodrigues, R. Chengoden, S. Bhattacharya, and K. Lakshmanna, "Skin lesion classification of dermoscopic images using machine learning and convolutional neural network," Scientific Reports 2022 12:1, vol. 12, no. 1, pp. 18134-, Oct. 2022, doi: 10.1038/s41598-022-22644-9.

[29] S. Davis et al., "Effective hospital readmission prediction models using machine-learned features," BMC Health Services Research 2022 22:1, vol. 22, no. 1, pp. 1415-, Nov. 2022, doi: 10.1186/S12913-022-08748-Y.

[30] N. C. da Silva, M. K. Albertini, A. R. Backes, and G. das G. Pena, "Machine learning for hospital readmission prediction in pediatric population," Comput. Methods Programs Biomed., vol. 244, p. 107980, Feb. 2024, doi: 10.1016/J.CMPB.2023.107980.

[31] M. Jamei, A. Nisnevich, E. Wetchler, S. Sudat, and E. Liu, "Predicting all-cause risk of 30-day hospital readmission using artificial neural networks," PLoS One, vol. 12, no. 7, p. e0181173, Jul. 2017, doi: 10.1371/JOURNAL.PONE.0181173.

[32] A. I. Sierra Espinel and M. J. Suarez Barón, "Applying Deep Learning and Forecasting Techniques to the Pharmaceutical Supply Chain,"

Procedia Comput. Sci., vol. 253, pp. 2791–2800, Jan. 2025, doi: 10.1016/J.PROCS.2025.02.003.

[33] S. Al-Hourani and D. Weraikat, “A Systematic Review of Artificial Intelligence (AI) and Machine Learning (ML) in Pharmaceutical Supply Chain (PSC) Resilience: Current Trends and Future Directions,” Sustainability 2025, Vol. 17, vol. 17, no. 14, Jul. 2025, doi: 10.3390/SU17146591.

[34] C. Benmhamed, H. Reddad, M. Zemzami, F. Yalaoui, and N. Hmina, “An integrated nine-dimensional framework for pharmaceutical digital transformation (I9D-PDT),” Proceedings of the IEEE International Conference on Advanced Healthcare Systems (IEEE-ICAHS) (2025). IEEE. in press, 2026.

[35] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, “An overview of clinical decision support systems: benefits, risks, and strategies for success,” NPJ Digit. Med., vol. 3, no. 1, Dec. 2020, doi: 10.1038/S41746-020-0221-Y.

[36] A. P. Susanto, D. Lyell, B. Widyantoro, S. Berkovsky, and F. Magrabi, “Effects of machine learning-based clinical decision support systems on decision-making, care delivery, and patient outcomes: a scoping review,” Journal of the American Medical Informatics Association, vol. 30, no. 12, pp. 2050–2063, Nov. 2023, doi: 10.1093/JAMIA/OCAD180.

[37] N. Hong et al., “State of the Art of Machine Learning–Enabled Clinical Decision Support in Intensive Care Units: Literature Review,” JMIR Med Inform 2022;10(3):e28781 https://medinform.jmir.org/2022/3/e28781, vol. 10, no. 3, p. e28781, Mar. 2022, doi: 10.2196/28781.

[38] L. Pumplun, F. Peters, J. F. Gawlitza, and P. Buxmann, “Bringing Machine Learning Systems into Clinical Practice: A Design Science Approach to Explainable Machine Learning-Based Clinical Decision Support Systems,” J. Assoc. Inf. Syst., vol. 24, no. 4, pp. 953–979, Jan. 2023, doi: 10.17705/1jais.00820.

[39] S. H. Miao, Y. J. Liu, M. Li, and J. Yan, “Clinical subtypes identification and feature recognition of sepsis leukocyte trajectories based on machine learning,” Scientific Reports 2025 15:1, vol. 15, no. 1, pp. 12291-, Apr. 2025, doi: 10.1038/s41598-025-96718-9.

[40] S. A. Alzakari et al., “Enhanced heart disease prediction in remote healthcare monitoring using IoT-enabled cloud-based XGBoost and Bi-LSTM,” Alexandria Engineering Journal, vol. 105, pp. 280–291, Oct. 2024, doi: 10.1016/J.AEJ.2024.06.036.

[41] M. Hanoon Tuama, “International Journal of Professional Studies A Comparative Evaluation of Random Forest and XGBoost Models for Disease Detection Using Medical Indicators,” INTERNATIONAL JOURNAL OF PROFESSIONAL STUDIES, no. 19, p. 2025, doi: 10.37648/ijps.v1.

[42] B. Li et al., “Using Machine Learning (XGBoost) to Predict Outcomes After Infrainguinal Bypass for Peripheral Artery Disease,” Ann. Surg., vol. 279, no. 4, pp. 705–713, Apr. 2024, doi: 10.1097/SLA.0000000000006181.

[43] B. Guembe, S. Misra, S. Misra, and A. Azeta, “Federated Bayesian optimization XGBoost model for cyberattack detection in internet of medical things,” J. Parallel Distrib. Comput., vol. 193, p. 104964, Nov. 2024, doi: 10.1016/J.JPDC.2024.104964.

[44] R. P. Sheridan, W. M. Wang, A. Liaw, J. Ma, and E. M. Gifford, “Extreme Gradient Boosting as a Method for Quantitative Structure–Activity Relationships,” J. Chem. Inf. Model., vol. 56, no. 12, pp. 2353–2360, Dec. 2016, doi: 10.1021/ACS.JCIM.6B00591.

[45] A. H. Marshall, S. I. McClean, C. M. Shapcott, and P. H. Millard, “Modelling patient duration of stay to facilitate resource management of geriatric hospitals,” Health Care Manag. Sci., vol. 5, no. 4, pp. 313–319, 2002, doi: 10.1023/A:1020394525938.

[46] M. Roimi et al., “Development and validation of a machine learning model predicting illness trajectory and hospital utilization of COVID-19 patients: A nationwide study,” Journal of the American Medical Informatics Association, vol. 28, no. 6, pp. 1188–1196, Jun. 2021, doi: 10.1093/JAMIA/OCAB005.

[47] B. Alsinglawi et al., “An explainable machine learning framework for lung cancer hospital length of stay prediction,” Sci. Rep., vol. 12, no. 1, Dec. 2022, doi: 10.1038/S41598-021-04608-7.