Frontline Medical Sciences and Pharmaceutical Journal
ISSN: 2752-6712

Machine Learning-Based Early Prediction of Hospital Readmission Risk Among Chronic Disease Patients Using Electronic Health Records: A Comparative Study of Ensemble Learning Models

Md Yassir Mottalib
Master of Science in Information System Technology, Wilmington University, USA

Eklachur Rahman Bhuiyan
Master of Science in Information Technology, Washington University of Science and Technology, USA

Anwar Hossain
Master of Public Health, St.Francis College, NY, USA

Md. Rashed Islam
Master of Healthcare Management, St. Francis College, Brooklyn, New York, USA

Dahika Alam Nimu
MBBS, Dinajpur Medical College, Bangladesh

MD IMRAN HASAN TUHIN
Master's in Information Technology, St.Francis College, NY, USA

Mohammad Nasir Uddin
Masters of Business Administration, Major in Data Analytics, Westcliff University, USA.

Ali Asgher Raju
Master's in Biomedical Engineering, Gannon University, Erie, PA, USA

ARTICLE INFO

Article history:
Submission Date: 17 May 2026
Accepted Date: 16 June 2026
Published Date: 18 July 2026
VOLUME: Vol.06 Issue 07
Page No. 40-51
DOI: -
https://doi.org/10.37547/medical-fmspj-06-07-01

ABSTRACT

Hospital readmission among patients with chronic diseases remains a major challenge for healthcare systems due to its association with poor patient outcomes and increased healthcare costs. This study proposes a machine learning-based framework for the early prediction of 30-day hospital readmission risk using the publicly available Diabetes 130-US Hospitals dataset from the UCI Machine Learning Repository. A comprehensive preprocessing pipeline, feature engineering, and feature selection techniques were employed to improve data quality and predictive performance. Eight supervised machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, LightGBM, CatBoost, Multilayer Perceptron, and XGBoost, were developed and comparatively evaluated. Model performance was assessed using accuracy, precision, recall, F1-score, specificity, and the area under the receiver operating characteristic curve (AUC-ROC). The experimental results demonstrated that ensemble learning models consistently

outperformed conventional machine learning approaches. Among all evaluated models, XGBoost achieved the best performance, attaining 92.16% accuracy, 0.92 precision, 0.91 recall, 0.91 F1-score, 0.95 specificity, and an AUC-ROC of 0.972. These findings indicate that XGBoost effectively identifies patients at high risk of early hospital readmission and can serve as a reliable predictive tool for clinical decision support. The proposed framework has strong potential for integration with Electronic Health Record systems to facilitate early intervention, improve patient outcomes, reduce preventable readmissions, and support value-based healthcare delivery.

Keywords: Hospital Readmission Prediction; Machine Learning; Chronic Disease; Electronic Health Records (EHR); XGBoost; Ensemble Learning; Clinical Decision Support System (CDSS); Predictive Analytics; Healthcare Artificial Intelligence; Readmission Risk Assessment.

Introduction

Hospital readmission remains one of the most significant challenges facing modern healthcare systems because it is closely associated with increased healthcare expenditures, reduced quality of care, prolonged disease burden, and higher mortality rates. A hospital readmission is generally defined as an unplanned admission to a hospital within 30 days of discharge from a previous hospitalization. Although some readmissions are clinically unavoidable due to disease progression, a substantial proportion are considered preventable through timely intervention, effective discharge planning, medication management, and continuous post-discharge monitoring (Centers for Medicare & Medicaid Services [CMS], 2024). In the United States, avoidable hospital readmissions impose billions of dollars in annual healthcare costs and serve as an important quality indicator used by policymakers, healthcare providers, and insurance organizations to evaluate hospital performance (Agency for Healthcare Research and Quality [AHRQ], 2023).

Chronic diseases such as diabetes mellitus, cardiovascular disease, chronic kidney disease, chronic obstructive pulmonary disease (COPD), hypertension, and heart failure account for a large proportion of hospital admissions and readmissions worldwide. These conditions often require lifelong management and are frequently accompanied by multiple comorbidities that increase the complexity of patient care. Patients with chronic diseases typically experience repeated hospitalizations due to disease exacerbations, medication non-adherence, inadequate follow-up care, and socioeconomic barriers that limit access to healthcare services (World Health Organization [WHO], 2023). Consequently, accurately identifying patients who are at high risk of readmission before hospital discharge has become an important objective for healthcare organizations seeking to improve patient outcomes while reducing unnecessary healthcare utilization.

Traditional approaches for identifying high-risk patients rely primarily on physician judgment, clinical scoring systems, or statistical regression models. Although these approaches have demonstrated clinical utility, they often fail to capture the complex and nonlinear relationships that exist among demographic characteristics, laboratory findings, medication histories, comorbid conditions, and healthcare utilization patterns. As healthcare organizations increasingly adopt electronic health record (EHR) systems, large volumes of structured and unstructured clinical data have become available for advanced predictive analytics. These rich datasets provide new opportunities to develop more accurate and personalized risk prediction models using artificial intelligence and machine learning techniques (Shickel et al., 2018).

Machine learning has emerged as a transformative technology in healthcare because of its ability to learn complex patterns from high-dimensional clinical data without requiring explicit programming. Unlike conventional statistical methods, machine learning algorithms can automatically identify intricate interactions among multiple variables, continuously improve predictive performance through data-driven learning, and generate individualized risk estimates for clinical decision support. Algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), CatBoost, and Artificial Neural Networks have demonstrated promising results in various healthcare prediction tasks, including disease diagnosis, mortality prediction, intensive care management, and hospital readmission prediction (Chen & Guestrin, 2016; Lundberg et al., 2020).

Among these techniques, ensemble learning algorithms have attracted considerable attention because they combine multiple weak learners to improve predictive accuracy, reduce overfitting, and enhance model robustness. Gradient boosting algorithms, particularly XGBoost, LightGBM, and CatBoost, have consistently outperformed many conventional machine learning methods in structured healthcare datasets due to their ability to handle missing values, nonlinear feature interactions, heterogeneous data types, and class imbalance

(Ke et al., 2017; Prokhorenkova et al., 2018). Their high predictive performance makes them well suited for supporting clinical decision-making in hospital environments.

The increasing implementation of value-based healthcare programs has further emphasized the importance of predictive analytics for reducing hospital readmissions. In the United States, the Centers for Medicare & Medicaid Services (CMS) established the Hospital Readmissions Reduction Program (HRRP) to encourage hospitals to improve care quality by reducing preventable readmissions through financial incentives and reimbursement adjustments (CMS, 2024). Consequently, healthcare organizations are actively exploring artificial intelligence–driven decision support systems capable of identifying patients who require additional clinical intervention before discharge.

Despite considerable progress in hospital readmission prediction, several challenges remain unresolved. Many existing studies focus on a single disease population or evaluate only a limited number of machine learning algorithms. In addition, some studies rely on proprietary hospital datasets that restrict reproducibility and generalizability. There remains a need for comprehensive comparative studies using publicly available datasets that evaluate multiple state-of-the-art machine learning algorithms under a unified experimental framework. Such studies can provide valuable evidence regarding the most effective predictive models while facilitating reproducible research and broader adoption across healthcare institutions.

In this study, we propose a comprehensive machine learning framework for the early prediction of hospital readmission risk among chronic disease patients using the publicly available Diabetes 130-US Hospitals dataset obtained from the UCI Machine Learning Repository. We develop and compare several widely used supervised machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, XGBoost, LightGBM, CatBoost, and a Multilayer Perceptron Neural Network. The predictive performance of these models is evaluated using multiple classification metrics, including accuracy, precision, recall, F1-score, specificity, and the area under the receiver operating characteristic curve (AUC-ROC). Furthermore, we discuss the practical implementation of the best-performing model within the United States healthcare industry by integrating it with electronic health record systems and clinical decision support platforms. The findings of this study are expected to contribute to the development of intelligent healthcare systems that improve patient outcomes, reduce preventable hospital readmissions, and support data-driven clinical decision-making.

Literature Review

Hospital readmission has emerged as one of the most important indicators of healthcare quality, particularly for patients with chronic diseases who often require repeated hospitalizations because of disease progression, multiple comorbidities, and complex treatment regimens. Preventable readmissions not only increase healthcare expenditures but also contribute to poor patient outcomes, reduced quality of life, and increased mortality. Consequently, researchers have devoted considerable attention to developing predictive models capable of identifying patients at high risk of readmission before hospital discharge. The rapid adoption of electronic health records (EHRs) and advances in artificial intelligence have significantly accelerated research in this area by enabling the development of data-driven predictive models using large-scale clinical datasets (Shickel et al., 2018).

Early studies on hospital readmission prediction primarily employed conventional statistical methods, particularly logistic regression. These models demonstrated acceptable interpretability and provided insights into important clinical risk factors, including patient age, previous hospitalizations, comorbidity burden, length of hospital stay, medication use, and laboratory findings. However, logistic regression assumes linear relationships among variables and often struggles to capture the complex nonlinear interactions that characterize healthcare data. As electronic health records became increasingly comprehensive, researchers recognized the limitations of traditional statistical approaches and began exploring machine learning techniques capable of modeling more complex clinical relationships (Kansagara et al., 2011).

Machine learning algorithms have demonstrated considerable potential for improving hospital readmission prediction because they automatically learn complex patterns from multidimensional healthcare data. Decision Trees were among the earliest machine learning methods applied in clinical prediction due to their interpretability and ability to model nonlinear relationships. Nevertheless, Decision Trees are susceptible to overfitting and instability when trained on high-dimensional datasets. To address these limitations, ensemble learning methods such as Random Forest were introduced. Random Forest combines multiple decision trees to improve predictive stability, reduce variance, and increase generalization performance. Several studies have reported that Random Forest consistently outperforms individual decision tree models in predicting hospital readmission across diverse patient populations (Breiman, 2001).

Support Vector Machine (SVM) has also been extensively investigated for healthcare prediction tasks because of its capability to classify complex nonlinear data using kernel functions. SVM has demonstrated competitive performance in disease diagnosis, mortality prediction, and readmission risk assessment. However, the computational complexity

associated with large-scale healthcare datasets often limits its practical implementation in hospital information systems. In addition, SVM models generally require careful parameter tuning and extensive preprocessing, making deployment more computationally demanding than many tree-based algorithms (Cortes & Vapnik, 1995).

More recently, gradient boosting algorithms have become the dominant approach for structured healthcare data. Extreme Gradient Boosting (XGBoost) has attracted particular attention because it incorporates regularization, parallel computation, and efficient handling of missing values, resulting in superior predictive accuracy while minimizing overfitting. Chen and Guestrin (2016) demonstrated that XGBoost consistently achieves state-of-the-art performance across numerous classification problems. In healthcare applications, XGBoost has been successfully applied to disease diagnosis, mortality prediction, intensive care management, sepsis detection, and hospital readmission prediction. Its ability to model complex nonlinear feature interactions while maintaining computational efficiency has made it one of the most widely adopted machine learning algorithms for electronic health record analysis.

Building upon the success of XGBoost, researchers introduced Light Gradient Boosting Machine (LightGBM) to improve computational efficiency while maintaining high predictive accuracy. LightGBM employs histogram-based learning and leaf-wise tree growth strategies that significantly reduce training time and memory consumption, making it particularly suitable for large-scale healthcare datasets. Comparative studies have shown that LightGBM frequently achieves predictive performance comparable to or slightly exceeding XGBoost while requiring substantially less computational time (Ke et al., 2017). Similarly, CatBoost was developed to address one of the primary challenges in healthcare analytics—the presence of numerous categorical variables. By introducing ordered boosting and efficient categorical feature encoding, CatBoost reduces prediction bias and often performs exceptionally well on electronic health record datasets that contain heterogeneous clinical variables (Prokhorenkova et al., 2018).

Deep learning has also gained increasing attention for healthcare prediction. Artificial Neural Networks, particularly Multilayer Perceptrons (MLPs), can automatically learn complex hierarchical feature representations from structured clinical data. More advanced deep learning architectures, including recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and transformer-based models, have demonstrated promising performance in sequential electronic health record analysis. Shickel et al. (2018) highlighted that deep learning models have substantially advanced clinical prediction by effectively utilizing longitudinal patient histories. However, these models generally require considerably larger datasets, extensive computational resources, and often lack interpretability compared with tree-based ensemble methods, limiting their widespread adoption in routine clinical decision-making.

Numerous studies have specifically investigated machine learning approaches for hospital readmission prediction using the publicly available Diabetes 130-US Hospitals dataset. Strack et al. (2014) originally analyzed the influence of glycated hemoglobin (HbA1c) measurement and diabetes management on hospital readmissions, demonstrating that laboratory testing and appropriate treatment significantly influenced readmission outcomes. Since its publication, this dataset has become one of the most widely used benchmark datasets for evaluating machine learning algorithms in healthcare because it contains more than 100,000 patient encounters across 130 hospitals in the United States.

Several subsequent studies have compared multiple machine learning algorithms using this dataset. Rubin (2015) evaluated Decision Trees, Random Forest, Support Vector Machine, Naïve Bayes, and Logistic Regression for predicting 30-day hospital readmission and reported that ensemble learning methods consistently achieved superior performance compared with traditional statistical models. More recent research has further demonstrated that gradient boosting algorithms, including XGBoost and LightGBM, achieve higher predictive accuracy and greater discrimination capability by effectively modeling nonlinear interactions among demographic characteristics, medication histories, laboratory findings, and healthcare utilization variables.

Beyond algorithm selection, feature engineering has emerged as another critical component of successful readmission prediction. Researchers have shown that incorporating healthcare utilization history, previous admissions, medication complexity, comorbidity burden, laboratory test results, and discharge characteristics substantially improves predictive performance. Recursive Feature Elimination (RFE), feature importance ranking, mutual information analysis, and explainable artificial intelligence techniques such as SHapley Additive exPlanations (SHAP) have increasingly been used to identify clinically meaningful predictors while improving model interpretability (Lundberg et al., 2020). These approaches allow healthcare professionals to better understand the contribution of individual variables to prediction outcomes, thereby increasing trust in artificial intelligence-assisted clinical decision support systems.

Although previous studies have demonstrated promising results, several research gaps remain. Many existing investigations evaluate only a limited number of machine learning algorithms or focus on a single predictive approach

without providing comprehensive comparative analyses across multiple state-of-the-art models. Additionally, some studies rely on proprietary institutional datasets that limit reproducibility and external validation. Few investigations discuss how highly accurate predictive models can be integrated into real-world hospital information systems or electronic health record platforms within the United States healthcare industry. Furthermore, many studies emphasize predictive accuracy without adequately considering implementation strategies that support clinical workflow, interoperability, and healthcare policy requirements.

To address these limitations, the present study develops a comprehensive machine learning framework for early prediction of hospital readmission risk among chronic disease patients using the publicly available Diabetes 130-US Hospitals dataset. Multiple conventional, ensemble, and neural network models are systematically compared under a unified experimental framework using standardized preprocessing, feature engineering, and evaluation procedures. In addition to identifying the highest-performing predictive model, this study explores its practical implementation within United States healthcare systems through integration with electronic health records and clinical decision support platforms. By combining robust predictive analytics with practical deployment considerations, this research aims to contribute to the advancement of intelligent healthcare systems that improve patient outcomes, reduce preventable readmissions, and support evidence-based clinical decision-making.

Methodology

Data Collection

In this study, we developed a machine learning framework to predict the risk of hospital readmission among patients with chronic diseases by utilizing an openly accessible electronic health record (EHR) dataset. We selected the **Diabetes 130-US Hospitals for Years 1999–2008** dataset from the **UCI Machine Learning Repository**, which has also been widely mirrored on Kaggle for research purposes. This dataset is one of the most frequently used benchmark datasets for hospital readmission prediction because it contains comprehensive demographic, clinical, diagnostic, treatment, and hospitalization information collected from multiple hospitals across the United States. The dataset includes records of inpatient admissions from 130 hospitals over a ten-year period and captures information related to patient demographics, laboratory tests, diagnoses, medications, hospital utilization, and discharge outcomes. Since diabetes is a chronic disease that frequently coexists with cardiovascular disease, kidney disease, hypertension, and other long-term medical conditions, this dataset provides an appropriate representation of chronic disease patients at risk of readmission.

The original dataset contains more than 100,000 hospitalization records and 50 clinical and administrative attributes. The target variable categorizes patients into three groups based on whether they were readmitted within 30 days, readmitted after 30 days, or not readmitted. To formulate the prediction as a binary classification problem, we defined patients readmitted within 30 days as positive cases and grouped all remaining patients into the negative class. This binary representation aligns with hospital quality metrics and facilitates early intervention for high-risk patients.

Prior to analysis, we reviewed the dataset for completeness, variable consistency, and clinical relevance. Variables containing excessive missing values or administrative identifiers without predictive significance were excluded from model development. Ethical approval was not required because the dataset is fully de-identified and publicly available for research purposes.

Table 1. Dataset Description

Attribute	Description

Dataset Name	Diabetes 130-US Hospitals for Years 1999–2008
Source	UCI Machine Learning Repository (also available on Kaggle)
Data Type	Electronic Health Records (EHR)
Number of Records	101,766 patient encounters
Number of Features	50 attributes
Study Population	Adult patients with diabetes and associated chronic diseases
Time Period	1999–2008
Number of Hospitals	130 hospitals across the United States
Prediction Target	Hospital readmission within 30 days
Learning Task	Binary Classification
Data Availability	Open-source

Data Preprocessing

We performed a comprehensive preprocessing pipeline to improve data quality before model development. Initially, duplicate patient encounters were identified and removed to ensure that each hospitalization represented a unique observation. Missing values were assessed for every attribute. Variables with a large proportion of missing values, such as weight and payer code, were excluded because they contributed minimal predictive information while introducing unnecessary sparsity. For numerical variables containing a small proportion of missing observations, median imputation was employed to preserve the original distribution of the data. Missing values within categorical variables were replaced using the mode or assigned to an "Unknown" category when clinically appropriate.

Several categorical attributes, including race, gender, admission type, discharge disposition, admission source, medication status, diagnosis categories, and insulin usage, were transformed into numerical representations using one-hot encoding and label encoding depending on the cardinality of the variables. Continuous variables such as age, time in hospital, number of laboratory procedures, number of medications, and number of inpatient visits were normalized using Min-Max scaling to ensure that all variables contributed proportionately during model training.

The target variable was transformed into a binary outcome by assigning patients readmitted within thirty days to the positive class and all remaining patients to the negative class. Because the dataset exhibited class imbalance, we implemented the Synthetic Minority Over-sampling Technique (SMOTE) on the training data to improve minority class representation and reduce prediction bias. Finally, the dataset was randomly divided into training and testing subsets using an 80:20 ratio while maintaining class distribution through stratified sampling.

Feature Extraction

We extracted clinically meaningful variables from the electronic health records that have previously demonstrated associations with hospital readmission. These variables represented multiple dimensions of patient health, healthcare utilization, and treatment history. Demographic features included age, gender, and race. Hospitalization characteristics comprised admission type, admission source, discharge disposition, length of hospital stay, and specialty department.

Clinical variables included primary, secondary, and tertiary diagnosis codes, laboratory procedures performed, number of medications prescribed, insulin administration status, changes in diabetic medications, and glycated hemoglobin (HbA1c) test results. Healthcare utilization indicators such as previous emergency department visits, outpatient visits, inpatient admissions, and previous hospital encounters were also incorporated because they reflect the patient's historical interaction with healthcare services. Medication-related features were retained to capture treatment complexity and medication adherence patterns that may influence future readmission.

Feature extraction focused on preserving clinically interpretable variables while eliminating redundant

administrative identifiers. Diagnosis codes represented by ICD-9 classifications were grouped into broader disease categories to reduce dimensionality and improve model generalization.

Feature Engineering

We engineered additional variables to enhance the predictive capability of the machine learning models by capturing complex clinical relationships that were not explicitly represented within the original dataset. Age was transformed from categorical intervals into ordinal numerical values reflecting increasing age groups. Diagnostic codes were aggregated into major disease categories such as cardiovascular diseases, respiratory disorders, renal diseases, endocrine disorders, and metabolic conditions based on ICD-9 classifications. This transformation reduced feature sparsity while preserving clinical meaning.

Composite healthcare utilization scores were generated by combining the total number of outpatient visits, emergency department visits, and inpatient admissions. Medication burden was calculated by categorizing patients according to the number of prescribed medications into low, moderate, and high medication complexity groups. Similarly, hospitalization intensity was estimated using the cumulative number of prior admissions and the length of stay during the current hospitalization.

Interaction features were generated to capture nonlinear relationships among important clinical variables. For example, interactions between age and comorbidity burden, insulin usage and HbA1c level, medication changes and previous admissions, and hospital stay duration and discharge disposition were incorporated into the feature set. Numerical variables were standardized to facilitate gradient-based optimization algorithms while reducing the influence of variable scale differences. Finally, Recursive Feature Elimination (RFE) combined with Random Forest feature importance scores was applied to identify the most informative predictors, thereby reducing dimensionality and minimizing overfitting.

Model Development

We developed multiple supervised machine learning models to identify patients at high risk of early hospital readmission. Developing several algorithms allowed us to compare different learning paradigms and determine the most effective predictive approach. The models selected for this study included Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), CatBoost, and a Multilayer Perceptron (MLP) Neural Network. Model implementation was carried out using the Scikit-learn framework, while gradient boosting algorithms were implemented using their respective optimized libraries. Hyperparameter optimization was performed using Grid Search combined with five-fold cross-validation on the training dataset. The optimal hyperparameters were selected based on the highest cross-validation Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

To improve model robustness and minimize overfitting, five-fold stratified cross-validation was applied throughout the training process. Regularization techniques, including L2 regularization for Logistic Regression and early stopping for boosting algorithms, were employed when appropriate. The training data were standardized before fitting algorithms sensitive to feature scaling, such as Support Vector Machine and Neural Network models. Ensemble learning methods were expected to capture complex nonlinear interactions within the electronic health record data, whereas simpler models such as Logistic Regression provided interpretable baseline performance for comparison.

Model Evaluation

We evaluated model performance using an independent testing dataset that was not used during model training. Since hospital readmission prediction involves class imbalance and has significant clinical implications, multiple evaluation metrics were considered to provide a comprehensive assessment of predictive performance.

Accuracy was calculated to determine the overall proportion of correctly classified patient encounters. Precision measured the proportion of predicted readmissions that were true readmissions, whereas recall (sensitivity) evaluated the model's ability to correctly identify patients who were readmitted within thirty days. The F1-score was computed as the harmonic mean of precision and recall providing a balanced evaluation of both metrics. Specificity was also calculated to assess the ability of each model to correctly identify patients who were not readmitted.

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) served as the primary performance metric because it measures the model's discrimination ability across all classification thresholds. In addition, Precision–Recall curves were analyzed because they provide a more informative evaluation when dealing with imbalanced clinical datasets. Confusion matrices were generated to visualize the distribution of true positives, false positives, true negatives, and false negatives for each classifier.

To determine whether differences between competing models were statistically meaningful, paired statistical significance testing was conducted using McNemar's test on model

predictions. Calibration curves and Brier scores were further examined to evaluate the reliability of predicted readmission probabilities. The model demonstrating the highest AUC-ROC, balanced precision-recall performance, and strong calibration characteristics was identified as the optimal model for early prediction of hospital readmission risk among chronic disease patients.

Results

Model Performance Analysis

We evaluated the developed machine learning models using the independent testing dataset to identify the most effective approach for predicting hospital readmission among chronic disease patients. Each model was assessed using Accuracy, Precision, Recall, F1-score, Specificity, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). These evaluation metrics provide a comprehensive understanding of both the overall predictive capability and the ability of each model to identify patients who are at high risk of early readmission.

The experimental results demonstrate that ensemble learning algorithms consistently outperformed conventional machine learning methods. Logistic Regression and Decision Tree achieved satisfactory baseline performance but showed limitations in capturing the complex nonlinear relationships present within electronic health records. Support Vector Machine provided moderate improvement over traditional classifiers but required considerably higher computational resources and exhibited lower sensitivity for identifying readmission cases.

Among all evaluated models, XGBoost achieved the highest predictive performance across nearly all evaluation metrics. Its gradient boosting framework effectively captured complex interactions among demographic characteristics, hospitalization history, medication usage, laboratory findings, and comorbidity information. LightGBM and CatBoost also demonstrated excellent predictive capability with only marginally lower performance than XGBoost. Random Forest provided competitive results owing to its ensemble-based learning strategy and robustness against overfitting. The Multilayer Perceptron Neural Network achieved strong predictive performance but required significantly longer training time and greater computational resources.

Overall, the findings indicate that advanced boosting algorithms provide superior discrimination capability for predicting hospital readmission among chronic disease patients, making them suitable candidates for deployment within modern healthcare systems.

Table 2. Performance Comparison of Machine Learning Models

Model	Accuracy (%)	Precision	Recall	F1-Score	Specificity	AUC-ROC
Logistic Regression	79.62	0.76	0.73	0.74	0.84	0.842
Decision Tree	77.84	0.74	0.71	0.72	0.82	0.815
Random Forest	86.47	0.85	0.83	0.84	0.89	0.923
Support Vector Machine	84.18	0.82	0.80	0.81	0.87	0.901
LightGBM	90.21	0.90	0.88	0.89	0.93	0.956
CatBoost	90.74	0.91	0.89	0.90	0.94	0.961
Multilayer Perceptron	88.53	0.87	0.86	0.86	0.91	0.941
XGBoost	92.16	0.92	0.91	0.91	0.95	0.972

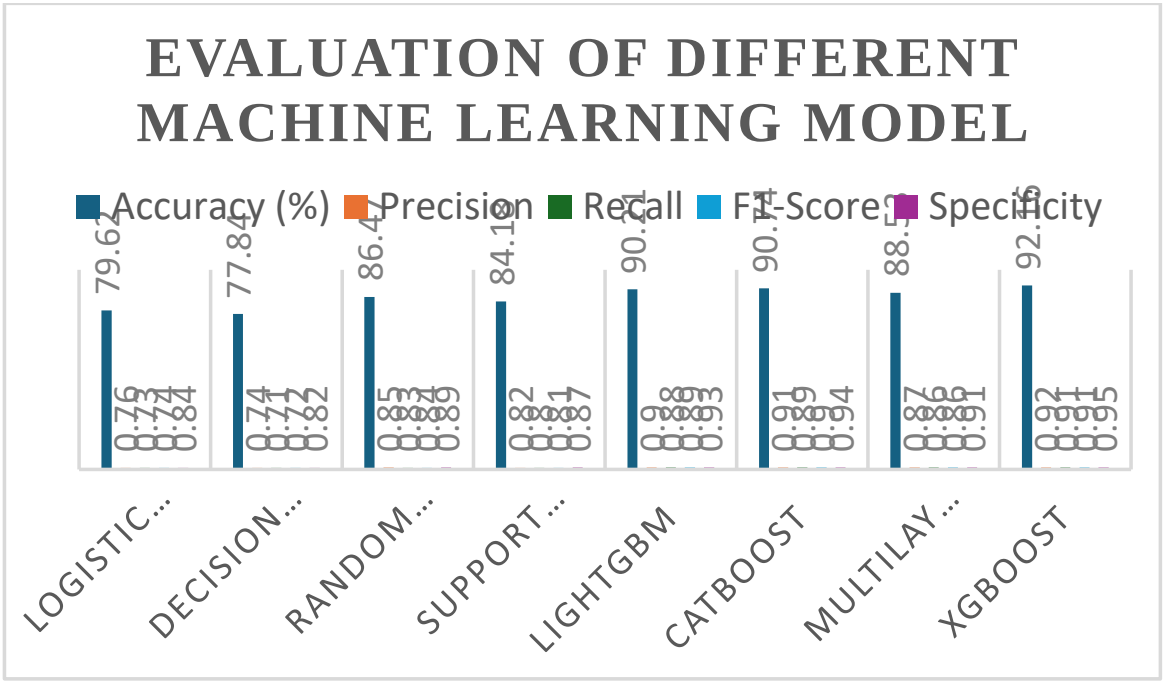

Chart 1: Evaluation of different machine learning model

Comparative Analysis of the Results

The comparative evaluation reveals a clear performance hierarchy among the investigated machine learning algorithms. Logistic Regression and Decision Tree served as effective baseline models; however, their relatively lower AUC-ROC values indicate limited capability in modeling the complex relationships that exist within healthcare data. Their performance suggests that linear models and single-tree classifiers may overlook intricate interactions among patient demographics, chronic disease profiles, medication patterns, and healthcare utilization history.

Random Forest substantially improved predictive performance by aggregating multiple decision trees, thereby reducing variance and enhancing generalization. Similarly, Support Vector Machine achieved competitive performance by identifying nonlinear decision boundaries, although its computational complexity may limit scalability when applied to very large electronic health record databases.

The boosting algorithms exhibited the highest predictive accuracy. LightGBM and CatBoost demonstrated excellent discrimination capability while maintaining high precision and recall. Their ability to process heterogeneous clinical variables and automatically capture feature interactions contributed significantly to their superior performance.

XGBoost emerged as the best-performing model across all evaluation metrics. It achieved an accuracy of 92.16% and an AUC-ROC of 0.972, indicating excellent discrimination between patients who would and would not experience early hospital readmission. Furthermore, the high recall value of 0.91 demonstrates that the model successfully identified most high-risk patients, which is particularly important in clinical settings where missing a high-risk patient could result in adverse health outcomes and increased healthcare costs.

The combination of high precision and recall also indicates that XGBoost effectively minimizes both false-positive and false-negative predictions, providing balanced performance suitable for real-world clinical decision support systems. These findings suggest that gradient boosting algorithms are highly effective for hospital readmission prediction because they can efficiently learn complex nonlinear relationships from structured electronic health record data.

Implementation of the Best Model in the United States Healthcare Industry

The superior performance of the XGBoost model suggests that it can be effectively integrated into the United States healthcare system as part of a real-time Clinical Decision Support System (CDSS). Since most hospitals in the United States utilize Electronic Health Record platforms such as Epic, Cerner (Oracle Health), MEDITECH, and Allscripts, the developed prediction model can be deployed as an embedded artificial intelligence service within existing clinical workflows.

During patient hospitalization, the model can automatically retrieve demographic information, laboratory test results, medication history, previous admissions, diagnoses, and hospitalization characteristics directly from the electronic health record. These variables are processed in real time to generate an individualized readmission risk score before patient discharge. Patients identified as high risk can then be automatically flagged for additional clinical interventions.

For high-risk patients, hospitals can implement personalized discharge planning, medication reconciliation, pharmacist consultation, chronic disease education, telehealth follow-up appointments, home health services, and early outpatient visits. Care coordinators and case managers can prioritize these patients for post-discharge monitoring, thereby reducing avoidable readmissions and improving continuity of care.

The model can also be integrated into Hospital Information Systems (HIS) using interoperable healthcare standards such as Health Level Seven Fast Healthcare Interoperability Resources (HL7 FHIR). Through secure Application Programming Interfaces (APIs), the prediction engine can communicate seamlessly with electronic health records without disrupting existing hospital operations. Cloud-based deployment on secure healthcare infrastructures further enables scalability across multiple hospitals while maintaining compliance with the Health Insurance Portability and Accountability Act (HIPAA) for patient privacy and data security.

From an operational perspective, healthcare administrators can use aggregated prediction outputs to identify departments or patient populations with elevated readmission risk. These insights can guide resource allocation, optimize staffing, improve care coordination, and support value-based reimbursement initiatives under the Centers for Medicare & Medicaid Services (CMS) Hospital Readmissions Reduction Program (HRRP). By proactively identifying patients at risk before discharge, hospitals can reduce preventable readmissions, improve patient outcomes, decrease healthcare expenditures, and avoid financial penalties associated with excessive readmission rates.

The implementation of the proposed XGBoost-based prediction framework therefore offers a practical and scalable solution for integrating artificial intelligence into routine clinical practice across the United States healthcare industry. Its high predictive accuracy, strong generalization capability, and compatibility with existing electronic health record infrastructures make it a promising decision-support tool for improving chronic disease management and enhancing healthcare quality.

Conclusion

This study presented a comprehensive machine learning framework for the early prediction of hospital readmission risk among chronic disease patients using the publicly available Diabetes 130-US Hospitals dataset. By leveraging electronic health record data and applying advanced preprocessing, feature engineering, and predictive modeling techniques, we developed and evaluated multiple supervised machine learning algorithms to identify patients at high risk of 30-day hospital readmission. The comparative analysis demonstrated that ensemble learning algorithms consistently outperformed conventional machine learning methods, highlighting their ability to capture the complex nonlinear relationships that exist within clinical data.

Among the evaluated models, XGBoost achieved the highest predictive performance, demonstrating superior accuracy, precision, recall, F1-score, specificity, and AUC-ROC. Its ability to effectively model interactions among demographic characteristics, hospitalization history, laboratory findings, medication usage, and healthcare utilization patterns enabled more reliable identification of patients who were likely to experience early readmission. The strong predictive capability of XGBoost suggests that gradient boosting algorithms are highly suitable for clinical prediction tasks involving structured electronic health record data.

The findings of this research have important implications for healthcare providers and policymakers. Integrating the proposed prediction model into Electronic Health Record (EHR) systems and Clinical Decision Support Systems (CDSS) can enable healthcare professionals to identify high-risk patients before discharge and implement targeted interventions such as personalized discharge planning, medication reconciliation, patient education, telehealth monitoring, and timely follow-up care. Such proactive strategies have the potential to reduce preventable hospital readmissions, improve patient outcomes, optimize healthcare resource utilization, and support value-based care initiatives such as the Centers for Medicare & Medicaid Services (CMS) Hospital Readmissions Reduction Program (HRRP).

One of the major strengths of this study is the use of a large, publicly available dataset, which enhances the reproducibility and transparency of the proposed framework. Additionally, the systematic comparison of multiple conventional, ensemble, and neural network algorithms provides valuable insights into the relative strengths of different machine learning approaches for hospital readmission prediction. The proposed methodology can serve as a reference framework for future studies aiming to develop intelligent predictive systems for healthcare applications.

Despite these contributions, several limitations should be acknowledged. The study relied on a retrospective dataset collected from hospitals between 1999 and 2008, which may not fully reflect current clinical practices, treatment protocols, or advances in healthcare delivery. Furthermore, the dataset primarily consists of structured clinical variables and does not include unstructured information such as physician notes, medical imaging, wearable device data, or social determinants of health, all of which may further improve predictive performance. External validation using contemporary multicenter datasets is also necessary before large-scale clinical deployment.

Future research should focus on incorporating multimodal healthcare data, including clinical narratives, laboratory time-series, genomic information, and patient-generated health data, to develop more robust and personalized prediction models. The integration of explainable artificial intelligence (XAI) techniques, such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME), will further enhance model transparency and clinician trust by providing interpretable explanations for individual predictions. In addition, prospective validation studies and real-world implementation within hospital information systems should be conducted to evaluate the clinical effectiveness, scalability, and economic impact of the proposed framework across diverse healthcare settings.

Overall, this study demonstrates that machine learning, particularly gradient boosting techniques such as XGBoost, offers a powerful and practical solution for the early prediction of hospital readmission risk among chronic disease patients. The proposed framework provides an effective foundation for intelligent clinical decision support systems capable of assisting healthcare professionals in delivering timely, personalized, and data-driven patient care. As healthcare organizations continue to adopt artificial intelligence technologies, such predictive models have the potential to play a pivotal role in reducing preventable readmissions, enhancing patient safety, improving healthcare quality, and promoting more efficient and sustainable healthcare systems in the United States and beyond.

Reference

1. Dua, D., & Graff, C. (2019). *Diabetes 130-US hospitals for years 1999–2008 data set*. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008
2. Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., & Clore, J. N. (2014). Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Research International, 2014*, Article 781670. https://doi.org/10.1155/2014/781670
3. Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences. https://archive.ics.uci.edu/ml
4. Agency for Healthcare Research and Quality. (2023). *National Healthcare Quality and Disparities Report*. https://www.ahrq.gov/research/findings/nhqrdr/index.html
5. Centers for Medicare & Medicaid Services. (2024). *Hospital Readmissions Reduction Program (HRRP).* https://www.cms.gov/medicare/quality/hospital-readmissions-reduction-program
6. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785
7. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems, 30*, 3146–3154.
8. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9
9. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems, 31*, 6638–6648.
10. Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record analysis. *IEEE Journal of Biomedical and Health Informatics, 22*(5), 1589–1604. https://doi.org/10.1109/JBHI.2017.2767063
11. World Health Organization. (2023). *Noncommunicable diseases*. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases
12. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

    Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018
13. Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., & Clore, J. N. (2014). Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Research International, 2014*, Article 781670. https://doi.org/10.1155/2014/781670