Predicting Hospital Readmissions Using Patient Diabetes Data

Abstract

Hospital readmissions represent a critical metric for evaluating the quality of care provided to patients, particularly for chronic illnesses like diabetes. This study utilizes the Diabetes 130-US Hospitals for Years 1999–2008 dataset [1], publicly available through the UCI Machine Learning Repository, to predict 30-day hospital readmissions. Through comprehensive exploratory data analysis (EDA), preprocessing, and feature engineering, this research develops predictive models and compares them to relevant baselines and state-of-the-art techniques. Advanced machine learning methods, including ensemble approaches, were evaluated against standard classifiers to identify the most effective methods for this task. The study further explores challenges such as class imbalance, noise in the dataset, and issues related to overfitting, ultimately providing recommendations for improving readmission predictions to save lives and reduce healthcare costs associated with unnecessary hospitalizations.

1. Exploratory Data Analysis & Feature Selection

The Diabetes 130-US Hospitals dataset is comprised of 101,766 hospital encounter records collected from 130 institutions between 1999 and 2008. Each record provides detailed information on patient demographics, clinical observations, diagnoses, treatments, and outcomes. This dataset is particularly relevant for studying hospital readmissions due to its richness in features directly related to patient management. There are 50 different key attributes (mix of numerical and categorical datatypes) which include demographic details such as race, gender, and age; clinical features like the number of laboratory procedures performed, time spent in the hospital, and the number of diagnoses assigned; and treatment data such as the use of specific medications or combinations thereof. The primary outcome variable of interest is the readmission status, which indicates whether a patient was readmitted within 30 days ("<30"), after 30 days (">30"), or not readmitted ("NO").

This dataset has many missing values in various categorical and numerical features but is large enough to allow the development of robust machine learning models, while also being small enough to handle scalability constraints. Its breadth of features makes it well-suited for predictive analytics aimed at improving patient outcomes and healthcare efficiency.

Exploratory data analysis shown in Figure 1, revealed significant class imbalance in the target variable, with 53.91% of cases classified as "NO," 34.93% as ">30," and only 11.16% as "<30." This imbalance underscores the difficulty of predicting the "<30" readmission class, which is the focus of this study due to its clinical importance.

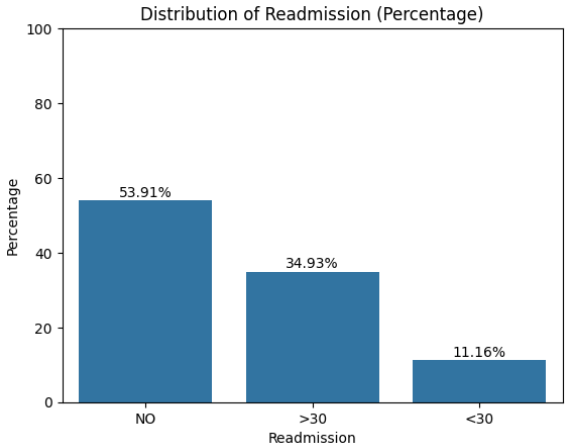

Figure 1: Distribution of Readmission results (target variable) as a percentage.

Figures 2 and 3 display how demographic variables like race and gender appear to exhibit weak correlations, if any, to readmission rates. Figure 4 shows how the type of discharge disposition also does not have an association with readmission rates. However, these weaker

predictors were retained later on in the dataset to investigate potential biases.

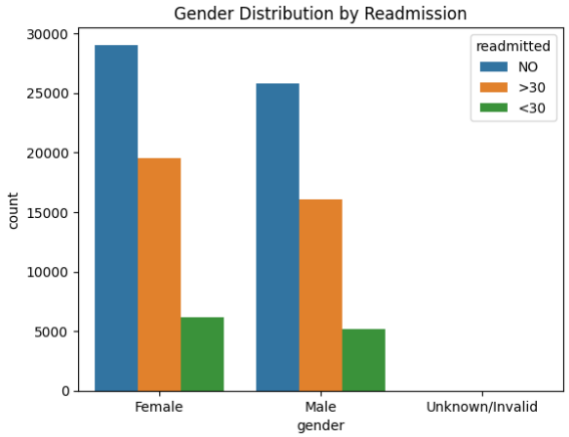

Figure 2: Gender distribution by Readmission category.

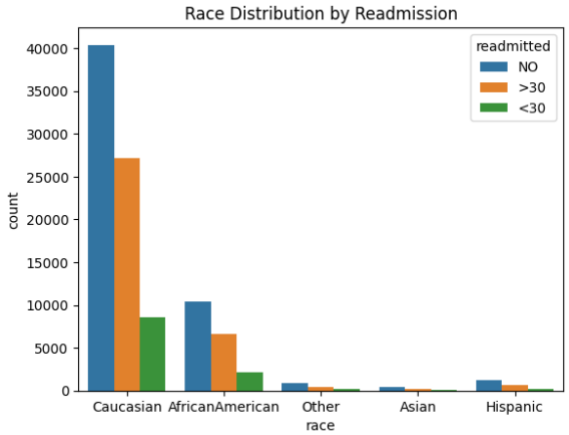

Figure 3: Race distribution by Readmission category.

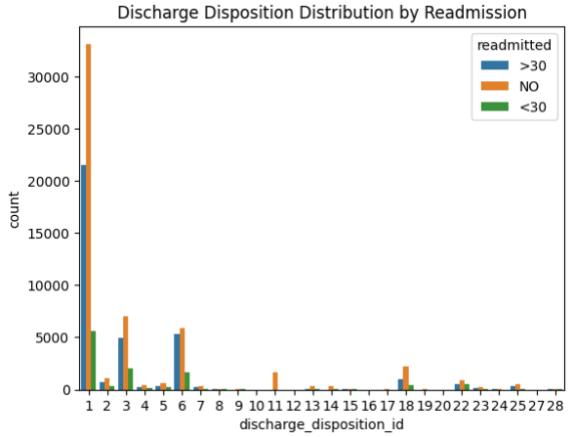

Figure 4: Discharge Disposition distribution by Readmission category.

The median age of patients observed in this study is 60-70 years old and can be seen more closely in Figure 5.

Further analysis of numerical features, such as the amount of time spent in the hospital, the number of laboratory procedures and medications, showed positive skewness in their distributions, necessitating normalization. These results are visualized in Figures 6-8.

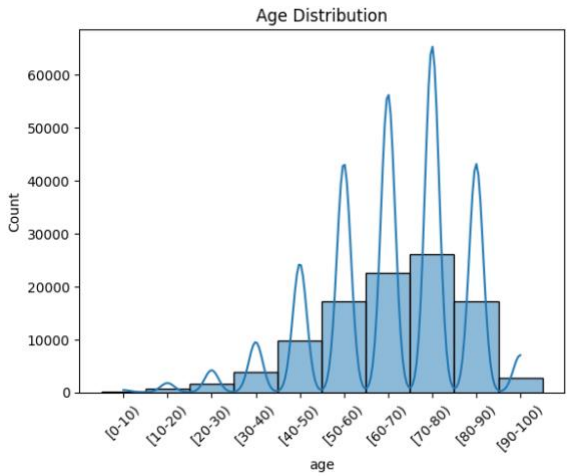

Figure 5: Age Distribution of patients.

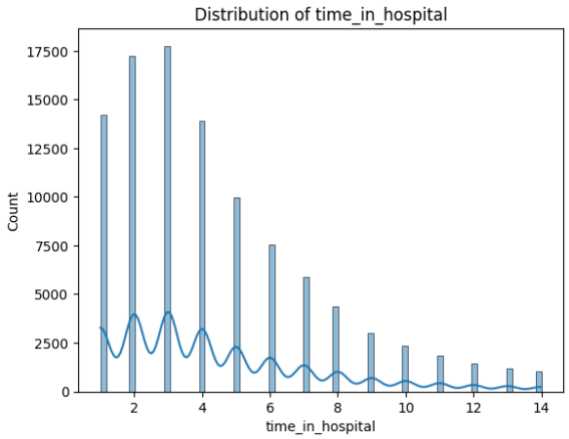

Figure 6: Distribution of number of days patients spent in the hospital.

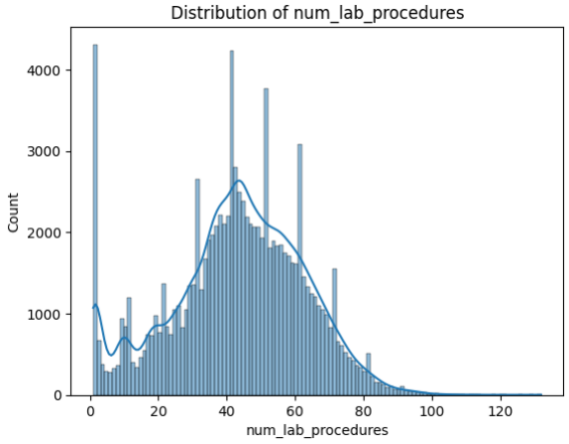

Figure 7: Distribution of number of lab procedures conducted on each patient.

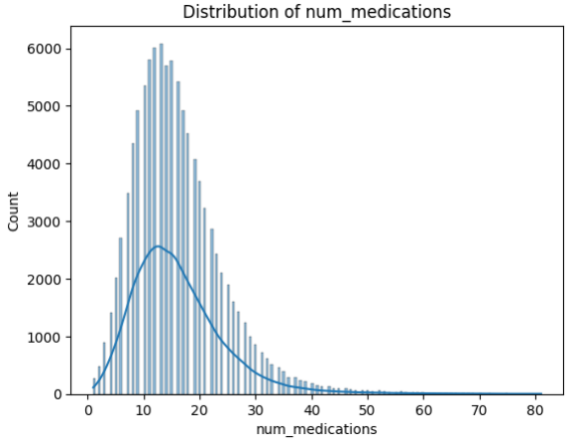

Figure 8: Distribution of number of medications prescribed to each patient.

Boxplots of these features also revealed the presence of outliers, particularly in the time spent in the hospital, which was addressed through the interquartile range (IQR) method.

The examination of categorical variables, including race, gender, and admission type, highlighted disparities in patient demographics and their association with readmission rates. For example, emergency admissions were more likely to result in readmissions, reflecting the acute nature of diabetes-related complications.

Correlation analysis in Figure 9 identified features such as the number of inpatient visits, the number of medications, and time spent in hospital as having the strongest associations with each other.

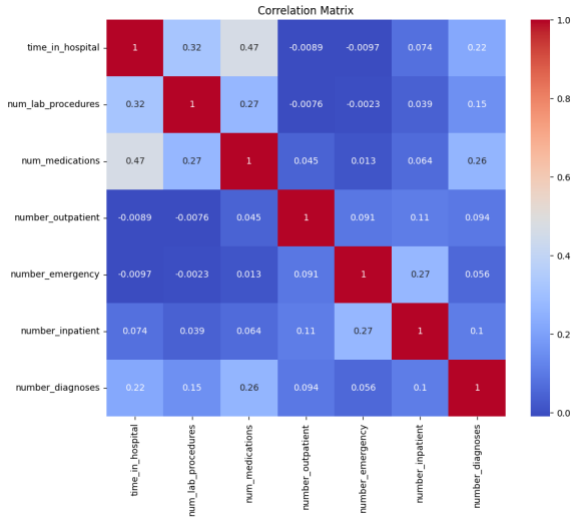

Figure 9: Correlation matrix of different numerical features.

Feature engineering played a crucial role in this study. A new variable, "hospital_interaction," was created by summing the number of inpatient, outpatient, and emergency visits to capture the overall intensity of healthcare utilization. A high score in this feature suggests a high healthcare resource utilization, a potential predictor of readmission. Another new variable called “chronic_count” was created because patients with a higher number of diagnoses are likely to have chronic conditions. This feature can serve as a proxy for patient complexity since

a high chronic count correlates with increased readmission risk, indicating patients with multiple diagnoses may require additional monitoring. These variables, along with high-correlation features like the number of medications and diagnoses, formed the basis for model training and evaluation later in the study.

2. Predictive Task Definition & Data Preprocessing

The predictive task for this study was binary classification, aimed at determining whether a patient would be readmitted within 30 days ("<30") or not ("NO" & ">30"). This task has significant implications for healthcare systems, as early identification of high-risk patients can lead to targeted interventions that reduce readmissions and improve care.

Data preprocessing involved several critical steps to tackle challenges and prepare the dataset for modeling. Noise and missing values in categorical features were imputed with the label "Unknown," while numerical features were imputed using their median to maintain the integrity of their distributions. Features like Weight and Payer Code were completely dropped from the dataset, due to high percentages of missing data.

One-hot encoding was used to map categorical features like HbA1c and Maximum Glucose Serum measurements, and Admission and Discharge Types, to binary values for inputting in to models.

Numerical features were then scaled using Min-Max Normalization, shown by Equation 1, to ensure comparability across different ranges. Additionally, all outliers were removed for these features.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

To address the severe class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied, which adjusted class weights by generating synthetic samples for the minority class and enhancing the sensitivity of models to "<30" cases. This algorithm is shown in Equation 2, where λ is just a random value ranging between 0 and 1.

$$x_{new} = x_{minority} - \lambda(x_{nearest\ neighbor} - x_{minority}) \quad (2)$$

The selection of features for modeling was guided by the findings of the exploratory analysis. Features with strong correlations to the target variable, such as the number of inpatient visits and the hospital interaction variable, were prioritized. Demographic variables, despite their weaker correlations, were still included to explore their potential impact on predictions and to address fairness concerns.

A number of predictive models were explored for this task, each with their own strengths and weaknesses. Baseline models, such as a Random Guesser and Majority Class Predictor, will provide useful indications of the validity of the study's prospective models when comparing its results to that of chance.

Logistic Regression provides an interpretable and efficient approach to binary classification tasks like predicting readmissions. However, its limitations include an inability to capture non-linear relationships effectively. Support Vector Machines (SVMs) are well-suited for high-dimensional spaces and binary classification but may encounter scalability issues with large datasets, such as this one. Naïve Bayes is another simple baseline model that performs well in certain scenarios, though its assumption of feature independence often limits its performance on complex data. More advanced models like Random Forest and Extreme Gradient Boosting (XGBoost) are particularly relevant for their ability to handle non-linear relationships and interactions between features.

Random Forest provides insights into feature importance, while XGBoost's boosting mechanism reduces both bias and variance. However, both models can suffer from overfitting if not properly tuned.

To evaluate the performance of the predictive models tested, several metrics are utilized to ensure both robustness and clinical relevance. Accuracy measures the overall correctness of the model but is less informative in cases of imbalanced datasets, such as this one. Equation 3 describes Precision as the fraction of correctly predicted positive cases out of all predicted positives and becomes crucial when false positives have significant costs. Recall, on the other hand, as shown in Equation 4, assesses how well the model identifies all actual positive cases. This metric is critical when minimizing false negatives is a priority, such as in identifying high-risk patients. F1-Score serves as the harmonic mean of Precision and Recall shown, providing a balanced perspective of the model's effectiveness.

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

Additionally, AUC-ROC evaluates the model's ability to distinguish between classes across all possible thresholds by tracking both the true and false positive rates, offering insight into the model's performance. All 4 of these metrics have values ranging from 0 to 1, with 1 being the state for an ideal model. Finally, confusion matrices were also used to provide a detailed breakdown of true positives, false positives, true negatives, and false negatives, allowing for a granular understanding of the trade-offs involved in the model's predictions.

3. Model Development & Evaluation

Baseline models were first implemented to establish reference points for performance. The Random Guesser yielded a Precision of 0.80 and a Recall of 0.50, as shown in Table 1. The majority class predictor, which labeled all instances as "NO," achieved an Precision and Recall of 0.79 and 0.89, respectively. However, this approach lacked clinical relevance, as it completely failed to identify the minority class (Readmission of “<30”). Similarly, a Naïve Bayes classifier struggled with the imbalanced data, achieving an Recall score of only 0.14. These results underscored the need for more sophisticated models capable of handling class imbalance and capturing complex relationships in the data.

Advanced models, including Logistic Regression, Random Forest, and XGBoost were then implemented. Logistic Regression, when adjusted with class weights to account for imbalance, achieved a Precision of 0.83 and Recall of 0.66, but a poor F1-Score of 0.24 for the minority class. Random Forest, known for its ability to model non-linear relationships, demonstrated improved recall but moderate precision, with a promising F1-score of 0.84. XGBoost continued performing well with overall Precision and Recall, while improving with an AUC-ROC of 0.61, but still struggled significantly on the minority class with an F1-score of only 0.02.

Hyperparameter tuning was performed on the Random Forest model using techniques like Grid Search and Randomized Search, where parameters such as the number of estimators, maximum depth, and class weights were optimized to enhance performance. Regularization techniques and cross-validation were employed to mitigate overfitting and ensure generalizability. However, its performance failed to improve relative to the original Random Forest and XGBoost models.

Ensemble methods were explored to leverage the strengths of individual models. The Stacking

Classifier, which used Logistic Regression as a meta-model, again maintained much of the same performance. A Weighted Voting Classifier, combining Logistic Regression, Random Forest, and XGBoost with soft voting and higher weighting on XGBoost, achieved the best overall balance of metrics, with the high Precision and Recall, and an improved AUC-ROC of 0.62. These results highlight the utility of ensemble approaches in improving predictive performance for imbalanced datasets. The Weighted Voting Classifier proved to be the most appropriate model this study was able to produce for the given predictive task.

Table 1: Evaluation results of all models (baseline, advanced, & ensemble methods) used in study.

Model	Precision	Recall	F1-Score	AUC-ROC
Random Guesser	0.80	0.50	0.59	0.51
Majority Class Predictor	0.79	0.89	0.84	N/A
Naïve Bayes	0.84	0.14	0.07	0.51
Logistic Regression	0.83	0.66	0.72	0.63
Random Forest	0.81	0.88	0.84	0.58
XGBoost	0.83	0.89	0.84	0.61
Random Forest w/ Hyperparameter Tuning	0.81	0.87	0.84	0.58
Stacking Classifier	0.82	0.87	0.84	0.58

4. Related Literature & Context

The dataset has been used in prior research, including the study titled "Impact of HbA1c Measurement on Hospital Readmission Rates," to investigate the role of HbA1c measurements in reducing hospital readmission rates [2]. This research employed multivariable logistic regression to analyze the relationship between readmissions and various features, such as HbA1c measurements, patient demographics, and primary diagnoses. Additionally, it aimed to identify potential improvements in the care of diabetic patients during hospital stays, thus providing insights into how such metrics can inform healthcare strategies.

Similar datasets from the UCI Machine Learning Repository and other healthcare repositories have been extensively used in the past. These datasets have facilitated research on predicting hospital readmissions for a range of conditions, analyzing the efficacy of diabetic management protocols, and studying the impact of patient demographics and care patterns on healthcare outcomes. For instance, datasets like the CMS Hospital Readmissions Reduction Program and MIMIC-IV have been used to evaluate healthcare policies and predict readmission risks, offering a broader understanding of factors contributing to patient outcomes.

The state-of-the-art methods for readmission prediction often include traditional statistical techniques and advanced machine learning methods. Multivariable logistic regression has been a foundational approach for binary classification tasks like readmission prediction due to its interpretability and effectiveness in analyzing relationships between features. Decision trees and ensemble methods, such as Random Forest and XGBoost, are commonly employed for their ability to handle both categorical and numerical features effectively, making them suitable for datasets with diverse feature types. Advanced approaches like neural networks and attention-based models have also been explored, especially for larger datasets, though they often require substantial preprocessing and feature engineering.

Commonly used features for these prediction tasks include patient demographics, medical history (diagnoses, lab results, etc.), and hospitalization details such as length of stay and procedures performed. Derived features, such as chronic condition counts or interaction terms between variables (e.g., race and diagnosis), are frequently utilized to enhance model performance. Insights from the literature, particularly regarding the predictive value of HbA1c measurements and their interactions with diagnoses, provide a basis for incorporating similar features into new models.

In terms of conclusions drawn from existing literature, several findings stand out. Research indicates that HbA1c measurements taken during hospital admissions are associated with reduced readmission rates, irrespective of the actual test results. This highlights the importance of systematic testing protocols for diabetic patients. Furthermore, many studies have noted the lack of structured diabetic care protocols in non-critical settings, suggesting a need for systematic improvements. Enhanced focus on diabetic management during hospitalization has been shown to significantly improve outcomes. These findings align with some of the results obtained in this study, particularly the identification of critical predictors like hospital interaction and the number of inpatient visits. However, discrepancies remain, as the low recall for the minority class in this study indicates that the current model requires further tuning and a more balanced training dataset to effectively identify high-risk groups, as emphasized in the literature.

5. Results & Discussion

The results of this study demonstrate that predicting 30-day readmissions is a feasible, yet challenging task. Baseline models, while achieving high overall accuracy, failed to identify the minority class, highlighting the limitations of simplistic approaches. Advanced models, particularly Random Forest and XGBoost, showed improved Recall and Precision but still struggled to accurately capture the minority class, reflecting the inherent difficulty in addressing class imbalance and emphasizing the shortcomings of the final model's findings.

The Weighted Voting Classifier emerged as the most effective approach, balancing the strengths of individual models and achieving the highest AUC-ROC score. Feature importance analysis on the Random Forest model shown in Figure 10, revealed that variables such as the number of inpatient visits, hospital interactions, number of medications, number of lab procedures, and time spent in the hospital were the most predictive indicators of readmission results, aligning with clinical intuition and the correlation analysis discussed earlier in Figure 9. Unexpectedly, the presence of an HbA1c measurement is found to be ranked only 13th in terms of predictive importance, despite the findings of previous research.

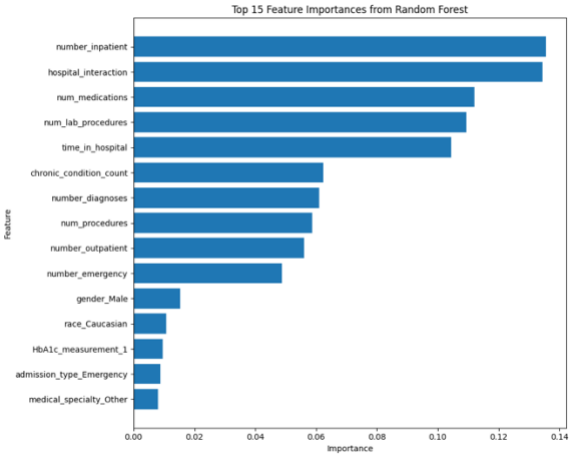

Figure 10: Distribution of top 15 most important features in dataset, when using Random Forest model.

Several challenges were encountered during the processing and model building phases. Noise in categorical variables, such as medical specialty, limited their predictive utility. The imbalanced nature of the target variable posed significant challenges, necessitating the use of techniques

like SMOTE. Overfitting was observed in certain models, particularly XGBoost, underscoring the need for regularization and robust validation strategies.

This study demonstrates the potential of machine learning to predict 30-day hospital readmissions for diabetic patients. By combining rigorous preprocessing, feature engineering, and ensemble modeling, the research achieves meaningful insights and improves predictive accuracy for a clinically significant task. The findings underscore the importance of balancing recall and precision, particularly for imbalanced datasets, and highlight the utility of ensemble methods in achieving this balance. The Precision-Recall curve shown in Figure 11, demonstrates the difficulty in balancing the 2 metrics in the case of the Random Forest model.

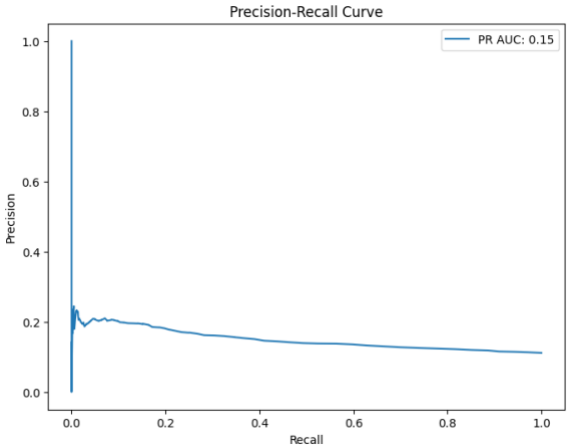

Figure 11: Precision-Recall curve for Random Forest Decision model.

In this case, Recall is significantly important as it is critical to minimize missed high-risk cases, such as life-threatening situations. However, F1-Score provides an essential balance between the two by ensuring a low ratio of False Positives, thereby limiting hospital resource waste.

Future work could explore the incorporation of temporal features, such as time-series data, to capture trends over multiple hospital visits. Deep learning models could be investigated for their ability to extract complex feature representations, though their interpretability remains a concern. Addressing biases in demographic features and expanding the dataset to include additional clinical variables and an increased ratio of 30-day readmission instances could further enhance model performance and generalizability.

These enhancements could further reduce readmissions and improve healthcare quality, translating into better patient outcomes and cost savings. By focusing on these directions, future studies can build on the findings of this research to develop more robust and clinically impactful predictive models.

References

[1] Clore, J., Cios, K., DeShazo, J., & Strack, B. (2014). Diabetes 130-US Hospitals for Years 1999-2008 [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5230J.

[2] Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records,” BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014.