

Predicting 30-Day Hospital Readmission in Diabetic Patients Using Machine Learning

By Yinuo Isabella Chen

AUTHOR BIO

Isabella Chen is a senior at Shanghai American School Pudong Campus with a strong interest in healthcare and medical careers. She is passionate about biology, public health, and chemistry.

ABSTRACT

Hospital readmissions are a significant problem in healthcare, particularly in diabetic patients. The research question guiding this study was: Can 30-day hospital readmission for diabetic patients be predicted using common machine learning models applied to hospital record data? This study aims to predict this using three machine learning models: Logistic Regression, Random Forest, and XGBoost. The analysis was conducted on the Diabetes 130-US Hospitals dataset from the UCI Machine Learning Repository, which contains over 100,000 patient records. After data cleaning, 98,053 complete records with 20 features were used for model training and evaluation. Random Forest achieved the highest accuracy at 88.73%, Logistic Regression at 88.67%, and XGBoost at 88.59%. Moreover, the XGBoost model's Area Under the Curve (AUC) was only 0.653, indicating performance slightly better than random guessing. The number of lab procedures were found to be the strongest predictor of readmission, followed by different diagnoses and number of medications. The models struggled to identify readmitted patients due to dataset imbalance, with only 11.3% of patients readmitted. These findings suggest that while machine learning can predict readmission status with reasonable overall accuracy, current models are not reliable enough for clinical use without additional features and improved data quality.

Keywords: Hospital readmission, diabetes, machine learning, Random Forest, XGBoost, Logistic Regression, prediction, electronic health records, healthcare analytics

INTRODUCTION

Hospital readmissions are a major problem in the healthcare system. When patients return to the hospital shortly after being discharged, it greatly costs the healthcare system billions of dollars, but also causes severe stress and difficulty for patients and their families (Jencks et al., 2009). Among patients with chronic and common conditions of diabetes, readmission rates are particularly high. Studies have shown that nearly one in five diabetic patients return to the hospital within 30 days of their discharge (Rubin & Shah, 2021).

Previously, researchers have tried many different ways to predict readmission. Older methods used basic statistics, but now machine learning can find patterns in large amounts of data (Sharda et al., 2025). Salim and Ibrahim (2026) used the same dataset as this study and found combining models gave modest improvements. Wang et al. (2026) focused on older diabetic patients with heart problems and got better results with blood test data. However, many studies create complex models that are difficult for doctors to actually use. This study uses three widely used machine learning models that are easier to understand. They are Logistic Regression (Hosmer et al., 2013), Random Forest (Breiman, 2001), and XGBoost (Chen & Guestrin, 2016). The dataset analyzed in this research comes from the UCI Machine Learning Repository and includes over 100,000 diabetic patient records.

Given the dataset's imbalance (with only 11% of total patients readmitted), I hypothesized that the models would experience trouble identifying the smaller group of patients who did return. With this, this study explores the research question: Is it possible and to what extent can we predict 30-day hospital readmission for diabetic patients using their hospital record data and common machine learning models?

METHODS

This study used the Diabetes 130-US Hospitals dataset from the UCI Machine Learning Repository (Strack et al., 2014; Kahn, 1999), which contains records from over 100,000 diabetic patients. The dataset includes patient demographics, hospital admission details, different diagnoses, medications, readmission status, etc.

The data was first cleaned by removing all columns with more than 50% missing data. Any remaining rows with missing values were also removed. The target variable was defined as readmission within 30 days, with patients readmitted within that time frame labeled as "Yes" (1) and all others as "No" (0).

Next, three machine learning models (Logistic Regression, Random Forest, and XGBoost) were trained and compared. The dataset was split into 80% training use and 20% testing use. Each model was then evaluated on the accuracy, confusion matrix, and ROC curve with AUC score to ultimately determine how well they performed. All analyses were performed using Python in Google Colab with the Scikit-learn and XGBoost libraries.

I used Claude AI (Anthropic, 2026) during the week of August 17, 2026, to assist with writing and editing my paper. I used it for assistance on coding, improving the clarity of explanations, and refining the language. Prompts included “Help me fix this error in my code”, “Help me organize my results section”, and “Suggest how to improve my flow and transitions”. All analysis, interpretation, and conclusions remain my own work.

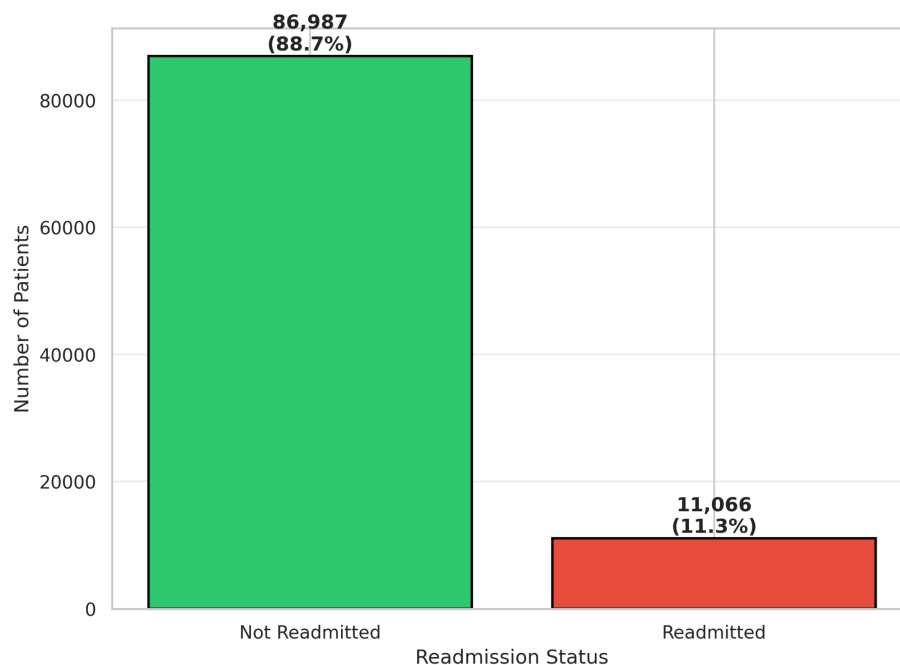
RESULTS

Data Overview

After data cleaning, the final dataset consisted of 98,053 complete patient records with 20 features. Figure 1 shows the distribution of the target variable. Of all patients, 86,987 (88.7%) were not readmitted within 30 days, while 11,066 (11.3%) were readmitted. Figure 2 shows the missing data patterns in the original dataset before cleaning. Weight has the highest missing rate (96.9%), followed by max_glu_serum (maximum glucose level) (94.7%) and A1Cresult (83.3%). All of these columns were removed during data cleaning.

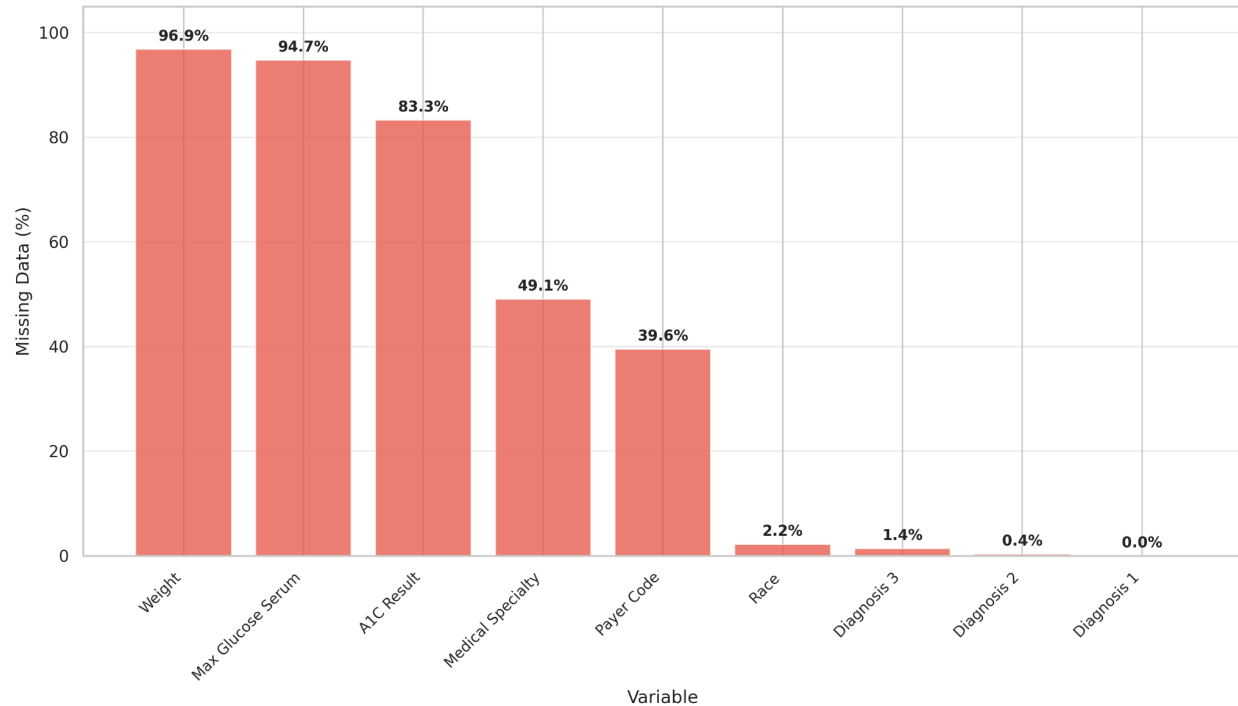
Figure 1

Distribution of 30-Day Readmission in the Dataset



Note. This bar chart shows the distribution of the readmission rate in the data set. The chart shows the number of patients who were readmitted within 30 days (11,066 patients, 11.3%) compared to those who were not readmitted (86,987 patients, 88.7%).

Figure 2
Missing Data Patterns in the Original Dataset



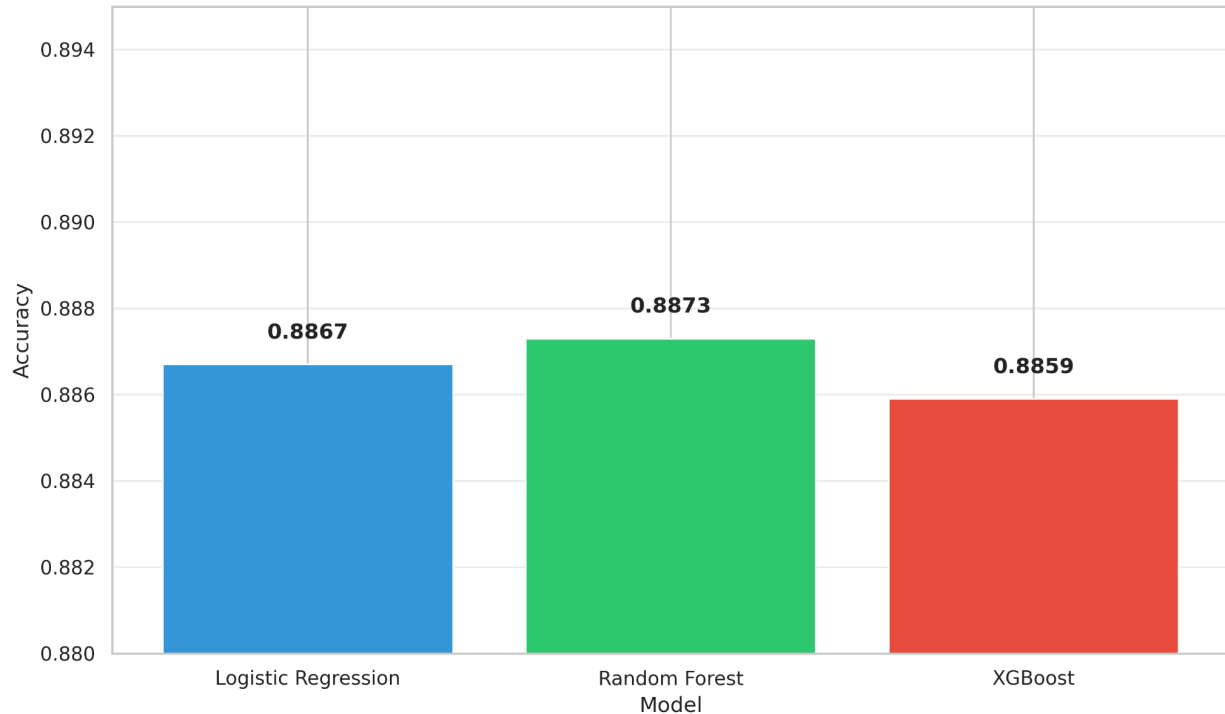
Note. This bar chart shows the percentage of missing values for each variable in the original dataset before data cleaning. Variables with missing data are shown on the x-axis, with the percentage of missing values on the y-axis. The chart includes all variables that have missing data.

Model Performance

Logistic Regression, Random Forest, and XGBoost were trained and evaluated. Figure 3 represents the comparison of the accuracy of each model. The models had a relatively small variation of accuracy. Specifically, Random Forest achieved the highest accuracy at 88.73%, followed by Logistic Regression at 88.67% and XGBoost at 88.59%. All three models performed within 0.2% of each other.

Figure 3

Comparison of Machine Learning Model Performance for Prediction



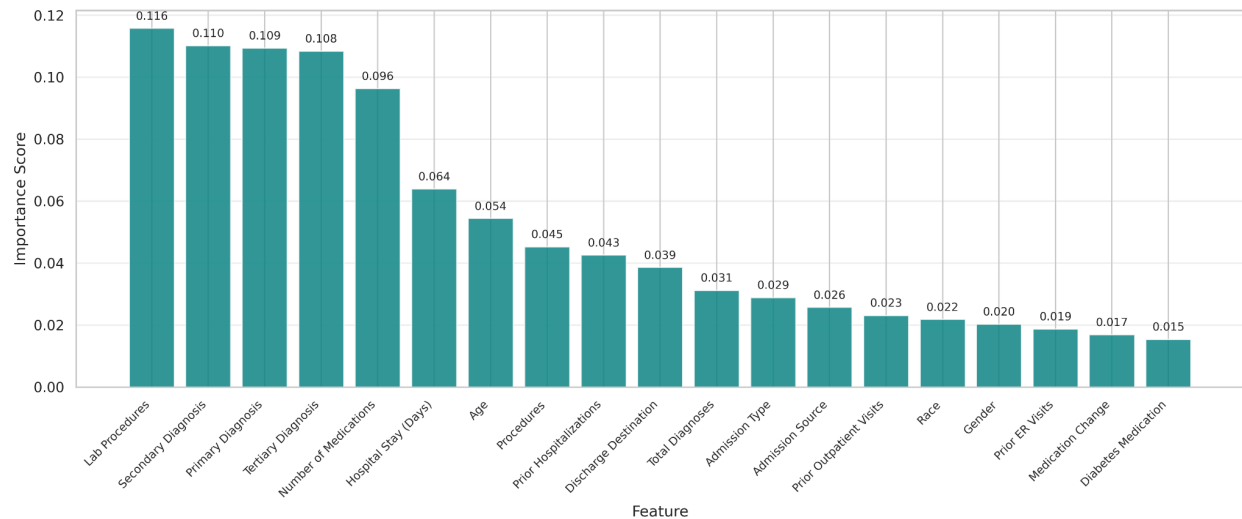
Note. This bar chart compares the accuracy of three machine learning models, Logistic Regression, Random Forest, and XGBoost, in predicting 30-day hospital readmission for diabetic patients. Accuracy represents the percentage of correct predictions made by each model.

Feature Importance

To determine which features play the most important role in determining readmission, the following figure (Figure 4) displays the results from the Random Forest model on the top 10 most important features for predicting readmission. The number of lab procedures was the most important predictor with an importance score of 0.116, followed by secondary diagnosis (0.110), primary diagnosis (0.109), and tertiary diagnosis (0.108). The number of medications (0.096) and time in hospital (0.064) also ranked among the top predictors, though less important.

Figure 4

Top 10 Most Important Predictors of 30-Day Hospital Readmission



Note. This bar chart displays the top 10 most important features for predicting 30-day hospital readmission, as determined by the Random Forest model. The importance score represents each feature's contribution to the model's prediction, with higher scores indicating greater influence.

Model Evaluation Metrics

To evaluate the predictions significance beyond accuracy, Figure 5 and 6 evaluates the specificity and sensitivity of a model, in this case, the XGBoost. Figure 5 represents the normalized confusion matrix for the XGBoost model's predictions. There, the model can be seen to correctly identify 99.4% of non-readmitted patients (True Negatives) but only 3.4% of the readmitted patients (True Positives).

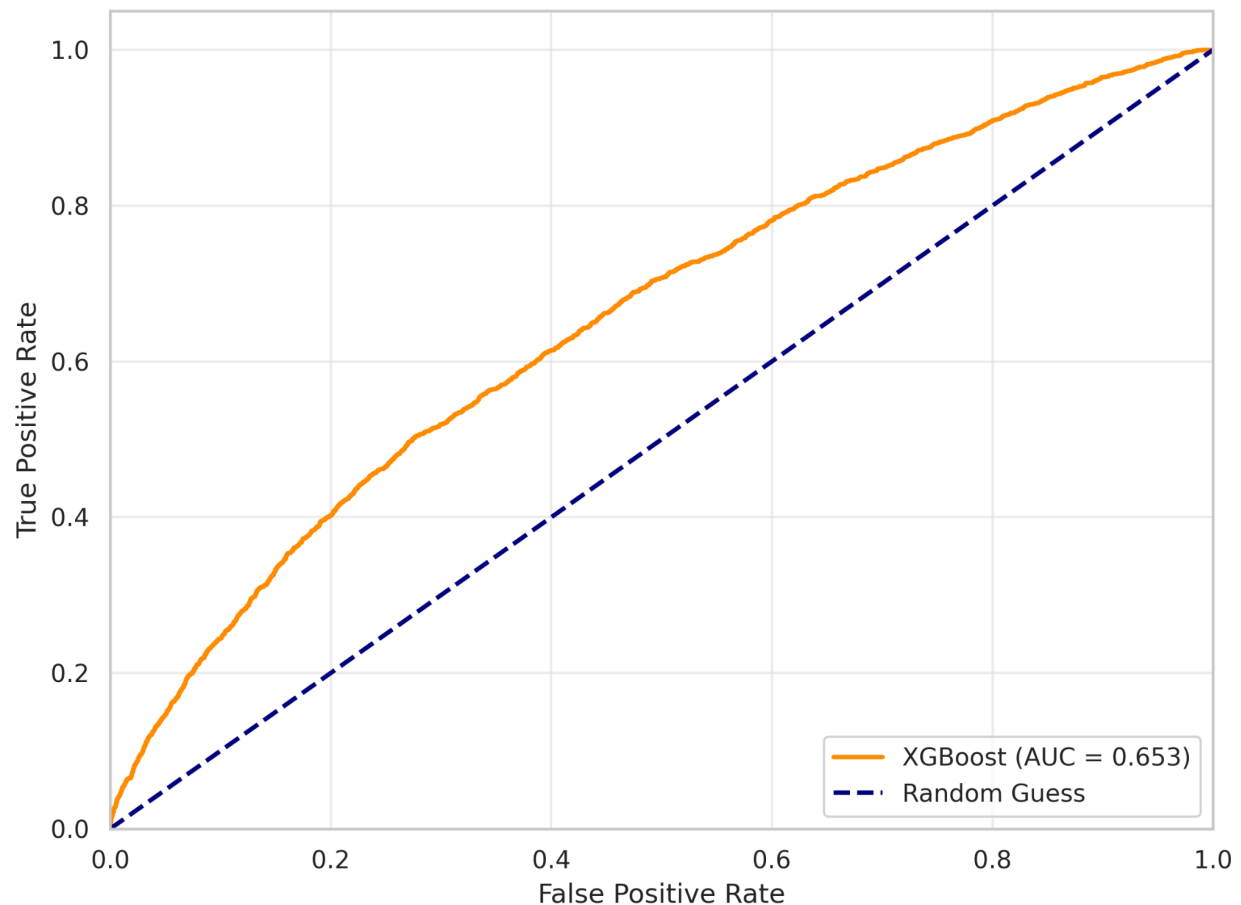
Figure 6 displays the ROC curve for the XGBoost model. The model achieved an Area Under the Curve (AUC) of 0.653. This indicates the performance is slightly better than random guessing, which would have an AUC of 0.5. This is represented in the figure through the two lines. The orange curve represents the model's performance, while the dashed diagonal line represents random guessing.

Figure 5
Confusion Matrix (Normalized)



Note. This figure shows the normalized confusion matrix for XGBoost model’s predictions. Each row represents the actual class and each column represents the predicted class. Both broken into Readmitted & Not Readmitted categories. Values are shown as proportions with each row summing to 1.0 (100%).

Figure 6
ROC Curve



Note. This figure is the Receiver Operating Characteristic (ROC) curve for the XGBoost model. This plots the True Positive Rate against the False Positive Rate across different classification thresholds. The dashed diagonal line represents random guessing (AUC = 0.5) and the orange line represents the XGBoost's performance (AUC = 0.653).

DISCUSSION

This study is aimed to predict the 30-day hospital readmission of diabetic patients using three different machine learning models applied to hospital record data. The Random Forest model achieved the highest accuracy at 88.73%, but all three models performed similarly. The number of lab procedures was identified to be the strongest predictor of readmission, followed by the different diagnoses (primary, secondary, tertiary) and the number of medications. However, as seen on the ROC curve, the AUC of the

XGBoost model is at 0.653, suggesting that the available features alone may be insufficient for highly reliable readmission prediction. This data point signals that the model can only predict slightly better than a random guess, which would yield an AUC of 0.5.

Previous studies have explored readmission predictions using similar machine learning approaches. Salim and Ibrahim (2026) used the same Diabetes 130-US Hospitals dataset to predict readmission using a stacking ensemble model. Through this, their results achieved an AUC of 0.688, slightly higher than the XGBoost model used in this study. This implies that using a group of models together only slightly improves accuracy compared to using a single model.

On the other hand, in a similar study conducted by Wang et al. (2026), their model achieved an AUC of 0.867 when predicting readmission in elderly patients with both diabetes and heart failure. This improvement in accuracy comparatively may be due to their more specific patient population and the inclusion of laboratory biomarkers such as C-reactive protein and B-type natriuretic peptide, which were not available in the dataset used in this study. This finding suggests that models trained on more specific patient populations with richer feature sets may positively impact the ability to predict performance.

A systematic review by Sharda et al. (2025) found that Random Forest and Artificial Neural Networks consistently outperformed traditional regression methods for readmission prediction across multiple clinical conditions. In this study, the finding that Random Forest achieved the highest accuracy aligns with this broader trend. However, the review also noted that subgroup modeling provided only marginal improvements in predictive accuracy, which may explain the limited gains observed in our study.

This study was successful in several ways. Firstly, the analysis has sufficient statistical power as it was conducted on a large datasets of 98,053 patient records. Moreover, three different machine learning models were trained and compared, allowing for a very comprehensive evaluation. The authors also implemented a rigorous data-cleaning process to ensure data quality and transparently reported limitations and missing-data patterns.

A big limitation seen in this study is that the dataset used was heavily imbalanced, with only 11.3% of patients being readmitted. This likely biased the models toward predicting “not readmitted”, heavily influencing the ability of the models to accurately predict results. This can be seen in the confusion matrix where only 3.4% of readmitted patients were correctly identified. Because of this imbalance, accuracy is not the most reliable metric for this problem, since a model that simply predicts ‘not readmitted’ for every patient would already achieve 88.7% accuracy without being meaningful. Future research should address this by applying resampling techniques to the training set to give the models more examples of readmitted patients to learn from. Another limitation seen is that several important clinical variables, including weight, A1C results, and maximum glucose serum, had substantial missing data and had to be excluded from the analysis. This may have resulted in important decision factors for readmission being overlooked due to the lack of resources available.

Despite the modest predictive performance, the finding that lab procedures were the strongest predictor suggests that clinical complexity drives readmission. Patients who require more lab tests during their hospital stay may benefit from additional follow-up care and discharge planning. Additionally, the high false negative rate (96.6%) reflects that the current model would miss many at-risk patients, limiting its clinical utility in its current form.

Future research and development on this topic should incorporate a broader range of features, including socioeconomic status, medication adherence, and post-discharge follow-up data. Additionally, addressing the class imbalance through techniques may improve the detection of patients at risk of readmission. Wang et al. (2026) demonstrated that including laboratory biomarkers significantly improved model performance, suggesting that richer clinical data will benefit and enhance predictive accuracy.

CONCLUSION

This study's goal is to determine to what extent common machine learning models can predict 30-day hospital readmission for diabetic patients using hospital record data. The models achieved overall accuracy of approximately 88.7%, but this number is misleading. The high accuracy mostly reflects the model's ability to correctly predict non-readmissions, which make up 88.7% of the dataset. In reality, the models still struggled to identify patients who were actually readmitted. This may be caused by the significant imbalance of the original dataset. For example, the XGBoost model only correctly identified 3.4% of readmitted patients and achieved a low AUC of 0.653, indicating performance only slightly better than random guessing. While the models can predict readmission status with reasonable overall accuracy, it is not yet reliable enough for identifying high-risk patients in a clinical setting. With potential future improvements like including more features and balancing the dataset, machine learning could become a useful tool for identifying patients who need extra support after discharge. In particular, improving the recall for readmitted patients should be a key goal for future works, since missing high-risk patients is the most clinically concerning error currently.

ACKNOWLEDGEMENTS

I would like to sincerely thank my research supervisor, Dr. Ramin Ramezani (Adjunct Associate Professor of Computer Science, UCLA) , and my research advisor, Mr. Sultan Mussakhan (Ph.D. candidate, Brock University, Ontario, Canada), for their guidance, feedback, and support throughout this research project.

REFERENCES

- Anthropic. (2026). *Claude* [Large language model]. <https://claude.ai>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>
- 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>
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley. <https://doi.org/10.1002/9781118548387>
- Jencks, S. F., Williams, M. V., & Coleman, E. A. (2009). Rehospitalizations among patients in the Medicare fee-for-service program. *New England Journal of Medicine*, 360(14), 1418-1428. <https://doi.org/10.1056/NEJMsa0803563>
- Kahn, M. (1999). Diabetes 130-US Hospitals for Years 1999-2008 [Dataset]. UCI Machine Learning Repository. <https://doi.org/10.24432/C5T59G>
- Rubin, D. J., & Shah, A. A. (2021). Predicting and preventing acute care re-utilization by patients with diabetes. *Current Diabetes Reports*, 21(9). <https://doi.org/10.1007/s11892-021-01402-7>
- Salim, S. S., & Ibrahim, A. A. (2026). A machine learning approach for predicting 30-day hospital readmission in patients with diabetes. *Healthcare*, 14(9), 1185. <https://doi.org/10.3390/healthcare14091185>
- Sharda, M., Sharma, S., Raikar, S., Verhagen, N., Wagle, J., Mathur, R., Gowda, S., Kommu, S., Prasad, R., Bhandari, S., & Jha, P. (2025). The role of machine learning in predicting hospital readmissions among general internal medicine patients: A systematic review. *Cureus*, 17(5), e84761. <https://doi.org/10.7759/cureus.84761>
- 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, 781670. <https://doi.org/10.1155/2014/781670>
- Wang, L., Wei, C., Chen, Y., Lu, D.-S., Zhang, H.-X., & Yang, M. (2026). Machine learning-based prediction model for 30-day readmission risk in elderly patients with type 2 diabetes mellitus and heart failure: A retrospective cohort study with SHAP interpretability analysis. *Frontiers in Cardiovascular Medicine*, 12, 1673159. <https://doi.org/10.3389/fcvm.2025.1673159>