Predicting Hospital Readmission Risk Using Explainable Machine Learning on Public Health Data

Binger Yu
School of Computing and Academic Studies, BCIT
Burnaby, BC, Canada
Student ID: A01003660
gyu42@my.bcit.ca

Savina Cai
School of Computing and Academic Studies, BCIT
Burnaby, BC, Canada
Student ID: A01493888
lcai25@my.bcit.ca

Yansong Jia
School of Computing and Academic Studies, BCIT
Burnaby, BC, Canada
Student ID: A01473470
yjial6@my.bcit.ca

Abstract

Hospital readmissions occurring within 30 days of a patient's discharge have long been a pressing issue in healthcare, and diabetes stands out as a major contributor to this problem [1]. In addition, the Hospital Readmissions Reduction Program (HRRP) penalizes hospitals exceeding expected readmission rates, yet predicting high-risk patients remains challenging [2].

This study develops machine learning classification models to predict 30-day readmission likelihood among diabetic patients using a comprehensive dataset from the UCI Machine Learning Repository [3]. Multiple algorithms are evaluated to identify patient characteristics, clinical factors, and treatment patterns associated with readmission risk.

The expected impact includes enabling targeted interventions for high-risk patients, optimizing resource allocation, and improving care transition planning [4]. This predictive capability supports operational decision-making at the hospital level and broader policy initiatives aimed at reducing preventable readmissions while enhancing patient outcomes and reducing healthcare expenditures.

Keywords

Hospital readmission prediction, healthcare analytics, diabetes dataset, data preprocessing, exploratory data analysis, machine learning models, feature engineering, explainability, SHAP, fairness analysis.

1 Introduction

1.1 Motivation and Background

The financial and clinical burden of hospital readmissions extends beyond individual institutions to affect entire healthcare systems. Although the Hospital Readmissions Reduction Program (HRRP) has created financial incentives for reducing excess readmissions [4], accurately identifying vulnerable patients remains challenging. Diabetic patients present particular complexity due to polypharmacy requirements (often 10–25 concurrent medications), frequent comorbidities, variable treatment adherence, and diverse care utilization patterns. Despite representing only 10.5% of the U.S. population, diabetic patients account for disproportionate healthcare resource utilization with costs exceeding $327 billion annually [1]. Understanding which combinations of clinical characteristics and healthcare utilization patterns best predict readmission risk remains an open question with significant practical implications.

1.2 Research Questions and Hypotheses

This study investigates whether advanced machine learning approaches can identify readmission risk patterns that simpler models miss. We address three questions: (1) Which patient factors—hospital utilization patterns, medication complexity, or laboratory testing intensity—demonstrate the strongest predictive relationships with 30-day readmission? (2) Do modern gradient boosting algorithms (XGBoost, LightGBM) offer meaningful improvements over traditional logistic regression and random forests? (3) Can engineered features, particularly composite service utilization metrics, capture risk signals beyond individual raw variables?

We hypothesize that comprehensive models will reveal nonlinear interactions between predictors. We expect paradoxical patterns—such as shorter hospital stays correlating with higher readmission rates—to emerge, potentially indicating premature discharge. Additionally, we anticipate that patients with extreme utilization patterns will constitute identifiable high-risk subgroups requiring targeted interventions.

1.3 Intended Stakeholders

This research focuses on four stakeholders. Hospital administrators and quality improvement teams require risk stratification tools to allocate case management resources and avoid HRRP penalties. Clinical teams, including discharge planners and case managers, need patient-specific risk assessments to guide discharge planning intensity and follow-up scheduling. Policy makers can leverage population-level insights to design chronic disease management programs considering systemic readmission drivers. Finally, patients themselves could benefit through improved care coordination and decreased burden of readmission.

1.4 Scope

This analysis focuses on 30-day all-cause readmission prediction for adult diabetic patients using the UCI Diabetes 130-US Hospitals dataset containing more than 100,000 encounters with 47 raw features. **In scope**: supervised classification modeling; explainable AI techniques (SHAP, LIME); systematic algorithm comparison; fairness evaluation across age, gender, and race; high-risk patient profiling. **Out of scope**: planned readmissions; readmissions beyond 30 days; cost effectiveness analysis of specific interventions; real-time clinical deployment infrastructure; external validation; causal inference. The dataset's 1999–2008 timeframe represents a limitation, but the fundamental clinical complexity and readmission risk relationships likely persist.

2 Related Work

2.1 Studies on Diabetes Readmission Prediction

The foundational dataset for diabetes readmission research originates from Strack et al. [3], who analyzed over 70,000 encounters across 130 U.S. hospitals (1999–2008), demonstrating that inpatient HbA1c testing correlated with reduced readmission rates. This established glycemic monitoring as a quality indicator but relied on descriptive statistics and basic logistic regression without exploring complex feature interactions or modern machine learning architectures. The publicly available UCI dataset provides rich clinical detail but requires more sophisticated analytical approaches to fully leverage its predictive potential.

Duggal et al. [5] applied machine learning techniques to the same dataset, achieving modest AUC values (0.61–0.65) and identifying prior inpatient visits and medication changes as key predictors—findings aligning with our preliminary correlation analysis. However, significant gaps remain: neither study explored composite utilization metrics aggregating emergency, outpatient, and inpatient encounters; modern gradient boosting algorithms were not evaluated; demographic fairness analysis remained limited; and model interpretability received insufficient attention despite its importance for clinical adoption and building provider trust in algorithmic decision support.

2.2 Broader Readmission Prediction Literature

Methodological insights from general readmission research inform our approach. Kansagara et al. [6] systematically reviewed readmission prediction models, finding most achieve modest discriminatory power (C-statistics 0.55–0.70). They attributed performance limitations to unmeasured social determinants, behavioral factors, and post-discharge support systems rarely captured in clinical databases. Despite these constraints, even modest improvements enable valuable risk stratification. Critically, they advocated for comparison against simple baseline predictors—a principle we adopt by establishing correlation-based baselines before evaluating complex algorithms.

Artetxe et al. [7] explored natural language processing to extract features from clinical notes, achieving improved AUC values (0.70–0.75). However, their approach requires substantial data infrastructure and sacrifices interpretability—barriers to clinical adoption. Our research focuses on structured clinical data while prioritizing transparency through explainable AI techniques.

2.3 Research Gaps and Our Contribution

Current literature reveals critical gaps: (1) limited comprehensive comparison of modern gradient boosting algorithms with rigorous tuning in diabetes contexts; (2) unexplored feature engineering of composite utilization metrics despite clinical relevance; (3) insufficient demographic fairness evaluation; (4) underdeveloped actionable interpretability connecting predictions to interventions.

Our contribution addresses these through: (1) systematic algorithm benchmarking including logistic regression, random forests, XGBoost, and LightGBM with hyperparameter optimization; (2) extensive feature engineering including derived service utilization metrics; (3) explainable AI techniques (SHAP, LIME) providing feature importance rankings; (4) rigorous fairness testing across demographics; and (5) explicit connection between model outputs and clinically-defined high-risk profiles (e.g., 10.44% with excessive emergency visits identified in exploratory analysis). This yields models that are accurate, transparent, equitable, and operationally actionable for reducing preventable readmissions through evidence-based resource allocation and targeted clinical interventions.

3 Dataset Description

This study uses the *Diabetes 130-US Hospitals for Years 1999–2008* dataset from the UCI Machine Learning Repository [3]. The dataset covers hospital encounters across 130 U.S. hospitals between 1999 and 2008, with each record representing a **single inpatient encounter** (one hospital stay) for a patient diagnosed with diabetes. It includes more than 100,000 encounters and 47 attributes describing patient demographics, admission details, laboratory tests, diagnoses, medications, and hospital utilization. The dataset is publicly available under a de-identified license compliant with HIPAA regulations.

The dataset can be accessed at https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008.

3.1 Data Source and Access

The dataset was collected between 1999 and 2008 and contains 47 attributes describing patient demographics, admission details, laboratory results, and prescribed medications. Each record represents one inpatient encounter. It is distributed under the UCI Machine Learning Repository license and complies with HIPAA de-identification standards. This dataset has been cited in multiple predictive healthcare studies focusing on chronic disease management and hospital utilization.

3.2 Data Quality and Preprocessing

Several preprocessing steps were applied to ensure data consistency and quality:

- **Missing values:** The “?” symbol was replaced with `NaN`. Columns with more than 40% missing entries (`weight`, `payer_code`, `medical_specialty`) were dropped.
- **Identifiers:** `encounter_id` and `patient_nbr` were removed to prevent data leakage.
- **Invalid entries:** Three rows with `gender = “Unknown/Invalid”` were excluded.
- **Target encoding:** The `readmitted` variable was converted into a binary label (1 = readmitted within 30 days; 0 = no or >30 days).
- **Feature transformation:** Age ranges were converted to numeric midpoints (e.g., “50–60” → 55). A derived feature, `service_utilization`, was created as the sum of `number_outpatient`, `number_emergency`, and `number_inpatient`.

Columns with excessive missingness (e.g., `weight`, `payer_code`, `medical_specialty`) were removed based on exploratory profiling, ensuring stable model training and fair feature representation.

3.3 Dataset Summary

After preprocessing, the cleaned dataset contains 101,763 records and 46 features. Key predictors include `time_in_hospital`, `num_lab_procedures`, `num_medications`, and `number_inpatient`, which reflect patient complexity and hospital resource utilization. Table 1 summarizes the retained feature categories across demographics, diagnoses, medications, and utilization metrics.

Table 1: Summary of Feature Categories in the Cleaned Dataset

Category	Example Features
Demographic	age, gender, race
Admission Information	admission_type_id, discharge_disposition_id, admission_source_id
Hospital Stay	time_in_hospital, number_inpatient, number_emergency, number_outpatient
Lab & Procedures	num_lab_procedures, num_procedures, num_medications
Test Results	A1Cresult, max_glu_serum
Diagnosis Codes	diag_1, diag_2, diag_3, number_diagnoses
Medications	metformin, insulin, glipizide, glyburide, pioglitazone, rosiglitazone, acarbose, etc.
Medication Flags	change, diabetesMed
Utilization Metrics	service_utilization (derived feature)
Target Variable	readmitted_binary
Total Features	**46**

Figure 1 shows the number of features contained in each category after preprocessing. The dataset is dominated by medication-related features, reflecting the complexity of diabetes treatment.

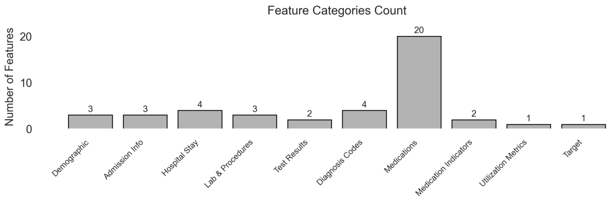

Figure 1: Count of features in each category after preprocessing.

3.4 Ethical and Privacy Considerations

The dataset is fully de-identified and contains no protected health information (PHI). All data handling processes comply to HIPAA and FIPPA/PIPEDA standards. This analysis is conducted solely for educational and research purposes, with no intent of clinical decision-making or patient profiling.

4 Exploratory Data Analysis

4.1 Distribution Analysis

The distribution of key numeric variables reveals several consistent patterns across the dataset. Hospital stay duration shows a right-skewed shape, with most patients hospitalized for fewer than seven days and a clear peak between two and five days. The number of lab procedures follows a slightly right-skewed unimodal distribution centered around roughly 40 procedures, with most values falling between 20 and 70. Medication usage displays a broader right-skewed pattern, as the majority of patients receive between 15 and 35 medications, while a small subset requires more than 80. Visit counts for outpatient, emergency, and inpatient encounters are heavily concentrated at zero, indicating that most patients had minimal recent healthcare utilization prior to admission, as illustrated in Figure 4.

4.2 Trend Analysis by Demographics

The demographic trends reveal clear age-related patterns in both treatment intensity and hospitalization duration. Figure 2 shows that the medication usage increases steadily with age, rising from an average of 6 prescriptions in the 0–10 age group to a peak of approximately 17 medications among patients aged 60–70. This pattern reflects the accumulation of comorbidities and the greater clinical complexity typically observed in older diabetic populations. After age 70, the number of medications declines slightly but remains substantially higher than in younger groups. Hospitalization duration follows a similar upward trend. The average length of stay increases from about 2.5 days in childhood to nearly 4.8 days in patients aged 70–80, again consistent with the higher severity and greater monitoring requirements in elderly patients. Figure 3 indicates a positive correlation between readmission rate and the length of stay. Patients with shorter hospitalizations (1–3 days) exhibit lower readmission rates (approximately 40–45%), while those with longer stays demonstrate substantially higher rates, exceeding 48% once hospitalization surpasses four days. This suggests that prolonged hospitalizations tend to correspond to more severe or unstable clinical conditions, which naturally increase the likelihood of readmission.

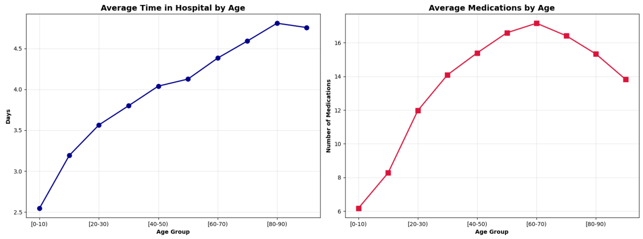

Figure 2: Average Time in Hospital and Medication by Age.

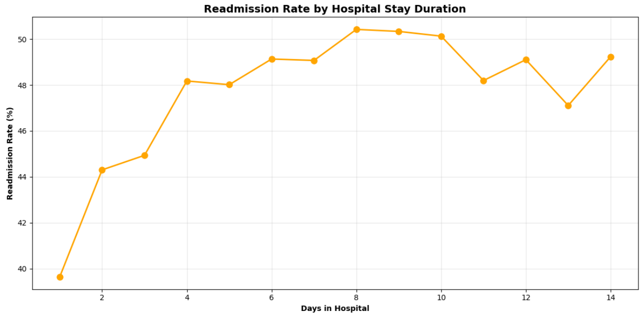

Figure 3: Readmission Rate by Hospital Stay Duration.

4.3 Correlation & Feature Relationships

The correlation patterns reveal several meaningful relationships among clinical and utilization variables. Hospitalization duration shows a moderate positive association with medication complexity (r = 0.47), indicating that longer stays are often accompanied by more intensive pharmaceutical management. Medication use is also moderately correlated with the number of clinical procedures performed (r = 0.39), suggesting that patients undergoing a higher volume of procedures tend to require broader therapeutic support. In contrast, time in hospital exhibits almost no relationship with the source of admission ($r \approx -0.01$). As shown in Figure 5, administrative variables such as admission type and discharge disposition demonstrate minimal correlation with medical complexity, underscoring that these categorical identifiers contribute little to explaining clinical variation within the dataset.

4.4 Readmission-Related Outliers

Descriptive analysis of the utilization variables shows that 10.44% of patients exhibit abnormally high numbers of emergency visits, which may indicate that their conditions are difficult to control. Additionally, 15.44% show excessive outpatient visit patterns, and 6.68% have frequent prior hospitalizations. These anomalies suggest the presence of a vulnerable patient subgroup with poorly managed chronic conditions. It is necessary to examine the key clinical indicators associated with these high-risk patients, as they are highly likely to be readmitted within 30 days of discharge.

4.5 Preliminary Baseline Model

To establish a performance reference for subsequent modeling efforts, we construct a simple baseline model with the most correlated predictor. Specifically, this baseline estimates readmission risk using only the number of prior hospital visits (number_inpatient), which demonstrates the strongest correlation with the outcome variable. This naïve, correlation-based approach serves as a benchmark to evaluate whether advanced machine learning models—such as Random Forest or XGBoost—offer substantial performance improvements. The baseline provides a minimum threshold against which to measure the added value of feature engineering, algorithm sophistication, and model optimization efforts.

5 Team Plan and Timeline

Our project is organized into three main phases to help us stay coordinated and maintain steady progress.

Phase 1 — Proposal and Early Exploration (Nov 1–13). During this stage, we reviewed related research, explored the dataset, and carried out the initial EDA. We also discussed preprocessing decisions, finalized feature categories, and set up the shared GitHub and Overleaf environments.

Phase 2 — Modeling and Analysis (Nov 14–27). The focus of this phase is building and evaluating the models. Student A continues managing data quality and feature engineering updates, Student B implements and tunes both baseline and advanced models, and Student C works on model interpretability and fairness analysis. The team updates notebooks and shares results regularly through GitHub.

Phase 3 — Final Report and Presentation (Nov 28–Dec 4). In the final phase, the team brings everything together. We consolidate model results, refine visualizations, and complete the written report in Overleaf. The presentation slides are also prepared during this time.

A detailed breakdown of team roles, task assignments, and project milestones is provided in Appendix B.

GitHub Repository: GitHub Repo Link

Overleaf Report Workspace: Overleaf Link Proposal | Overleaf Link Final Paper

Google Drive Folder: Drive Link

Appendices

A Figures

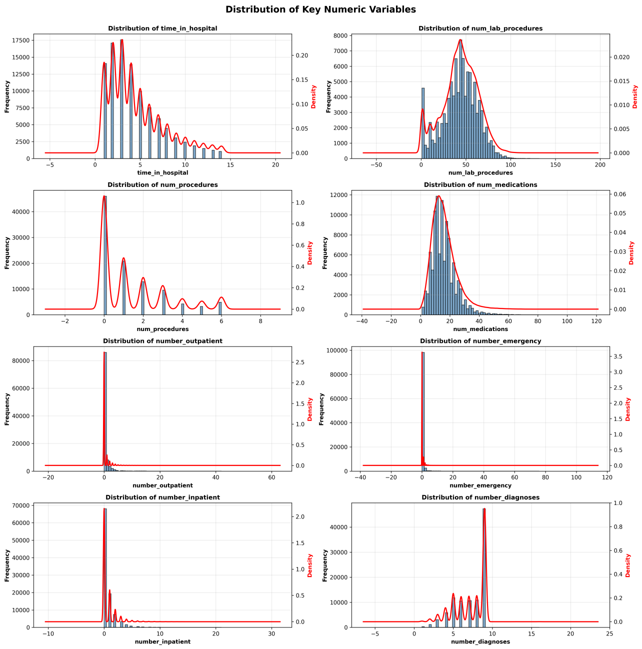

Figure 4: Distribution of Key Numeric Variables.

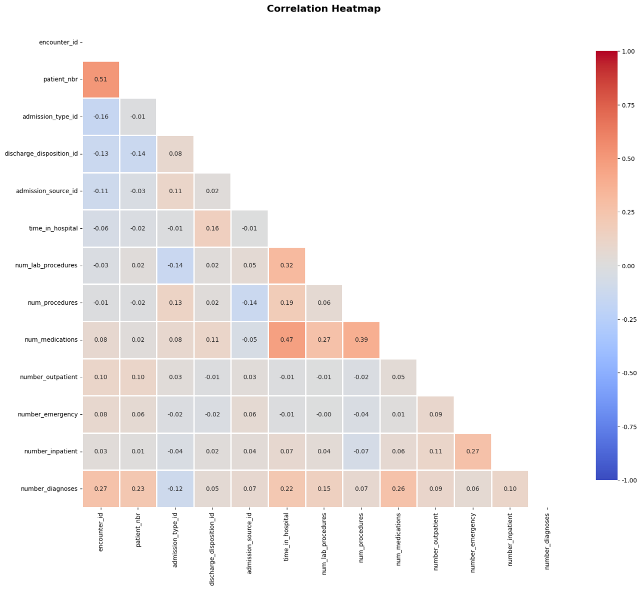

Figure 5: Correlation Heatmap.

B Team Plan and Timeline

Table 2: Team Members and Roles

Member	**Role**	Responsibilities
Binger Yu	**Data Engineer /Preprocessing Lead**	Acquire and preprocess the UCI Diabetes dataset; handle missing values, encoding, and normalization; perform exploratory data analysis (EDA) with descriptive statistics and visualizations.
Savina Cai	**Modeling & Evaluation Lead**	Implements baseline models (logistic regression, random forest); tune and train advanced models (XGBoost, LightGBM); evaluate models using AUC, F1, and calibration plots.
Yansong Jia	**Explainability & Reporting Lead**	Applys SHAP/LIME to interpret feature importance; conducts fairness analysis across demographics; prepares visualizations, report sections, and final presentation slides.

Table 3: Team Task Assignment for Proposal

Member	Main Writing Tasks	Supporting Tasks
Binger Yu	Write Sections Keywords, 3 (Dataset Description) and 5 (Team Plan and Timeline); provide dataset summary, feature overview and figures.	Create, compile and format the final proposal in Overleaf.
Savina Cai	Write Section 4 (Exploratory Data Analysis); review EDA findings for statistical validity; contribute to discussion of model preparation and expected results.	Review abstract and expected results for clarity and alignment with proposal objectives.
Yansong Jia	Write Sections 0–3 (Abstract, Introduction, Related Work).	Insert references and ensure consistent citation and formatting style throughout the document.

Table 4: Project Phases and Deadlines

Phase	**Dates**	Deliverables
Proposal	**Nov 1–13**	3-page IEEE-style proposal with literature search, dataset plan, and preliminary results.
Modeling	**Nov 14–27**	Baseline and advanced models, evaluation results, and performance metrics.
Final Report & Presentation	**Nov 28–Dec 4**	Complete report and 10-minute presentation summarizing methodology, explainability, and findings.

References

[1] American Diabetes Association, "Economic costs of diabetes in the u.s. in 2017," *Diabetes Care*, vol. 41, pp. 917–928, May 2018.

[2] K. E. Joynt and A. K. Jha, "Thirty-day readmissions—truth and consequences," *New England Journal of Medicine*, vol. 366, pp. 1366–1369, Apr 2012.

[3] B. Strack, J. P. DeShazo, C. Gennings, J. L. Olmo, S. Ventura, K. J. Cios, and J. N. Clore, "Impact of hba1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records," *BioMed Research International*, vol. 2014, p. 781670, 2014. Dataset source. Available: https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008.

[4] S. F. Jencks, M. V. Williams, and E. A. Coleman, "Rehospitalizations among patients in the medicare fee-for-service program," *New England Journal of Medicine*, vol. 360, pp. 1418–1428, Apr 2009.

[5] R. Duggal, S. Shukla, S. Chandra, B. Shukla, and S. K. Khatri, "Predictive risk modelling for early hospital readmission of patients with diabetes in india," *International Journal of Diabetes in Developing Countries*, vol. 36, pp. 519–528, Dec 2016.

[6] D. Kansagara, H. Englander, A. Salanitro, D. Kagen, C. Theobald, M. Freeman, and S. Kripalani, "Risk prediction models for hospital readmission: A systematic review," *JAMA*, vol. 306, pp. 1688–1698, Oct 2011.

[7] A. Artetxe, A. Beristain, and M. Graña, "Predictive models for hospital readmission risk: A systematic review of methods," *Computer Methods and Programs in Biomedicine*, vol. 164, pp. 49–64, Oct 2018.