Predicting 30-Day Readmissions in Diabetic Patients Using Ensemble Learning with AutoGluon

Baijiang (Noke) Yuan
Institute of Medical Science
University of Toronto
baijiang.yuan@mail.utoronto.ca

Abstract

Diabetes is a globally prevalent chronic disease associated with high rates of hospital readmissions, posing serious challenges to healthcare systems. Diabetic patients are particularly prone to complications that often lead to rehospitalization within 30 days of discharge due to poor disease control. Prior studies have shown that such readmissions can be predicted using machine learning (ML) and deep learning (DL) techniques. In this work, I propose an ensemble-based approach utilizing AutoGluon to predict 30-day readmissions in diabetic patients. I conduct a thorough comparative analysis of ensemble methods against traditional ML algorithms, deep neural networks, and a state-of-the-art Transformer-based foundation model for tabular classification tasks. Additionally, I explore how different feature sets and preprocessing strategies influence model performance and and identify the most important features associated with readmission. These findings show that the AutoGluon ensemble consistently achieves top performance across different settings, outperforming other models, with CatBoost and LightGBM as close contenders. DL and the foundation model demonstrate competitive performance but do not outperform the ensemble model. This study contributes by validating the AutoGluon ensemble model for 30-day readmission prediction in diabetic patients, benchmarking it against diverse models, and highlighting risk factors to support early intervention.

1 Introduction

1.1 Background

Diabetes is a chronic metabolic disorder characterized by the body's inability to properly regulate blood glucose levels, either due to insufficient insulin production by the pancreas or the body's reduced sensitivity to insulin (Ojo et al., 2023). As an increasingly prevalent condition, it is associated with heightened risks of acute complications and long-term health deterioration, often resulting in repeated hospitalizations. Research indicates that individuals with diabetes are nearly twice as likely to be readmitted within 30 days of discharge compared to non-diabetic patients, with reported readmission rates ranging from 14.4% to 22.7% (Rubin, 2015; Gregory et al., 2018). Hospital readmission—typically defined as a return to the hospital for unplanned care within a specific period, commonly 30 days—is considered a key indicator of healthcare quality(Kassin et al., 2012).

Financially, hospital readmissions impose significant burdens on healthcare systems, particularly for diabetic patients. In the United States, total hospital costs attributable to diabetes were approximately $124 billion in 2012, with 30-day readmissions contributing at least $20 billion(Karunakaran et al., 2018). On a per-admission basis, the average cost of a 30-day all-cause adult hospital readmission is estimated at $16,037(Kum Ghabowen et al., 2024). Specifically, for diabetic foot ulcer patients, readmissions can cost as much as the initial admission, with total care expenses averaging $79,315 for those readmitted, compared to $28,977 for non-readmitted patients (Hicks et al., 2019). These substantial costs underscore the urgent need for effective strategies to reduce readmissions among diabetic patients.

With the availability of large-scale electronic health records (EHRs), machine learning (ML) and deep learning (DL) models have emerged as promising solutions to the challenge of predicting hospital readmissions. Previous work has demonstrated that models such as random forests, gradient boosting, and neural networks can offer strong predictive capabilities (Hai et al., 2023; Liu et al., 2024; Shang et al., 2021). However, there remains a gap in comprehensive evaluations of these models, particularly when used in conjunction with ensemble learning and automated ML frameworks.

Recent literature suggests that ensembling multiple smaller models can outperform a single large model in both efficiency and accuracy(Kondratyuk et al., 2020). Building on the limitations of manual model selection and single-algorithm training, AutoGluon represents a state-of-the-art AutoML toolkit that integrates diverse base learners (e.g., LightGBM, CatBoost, neural networks) using multi-layer stacking and repeated k-fold bagging (Erickson et al., 2020). Unlike traditional AutoML frameworks that primarily focus on model or hyperparameter selection, AutoGluon emphasizes robust ensembling, scalability, and ease of deployment.

Problem Statement: This project aims to develop predictive models to determine the risk of a diabetic patient being readmitted within 30 days of discharge.

The objectives of this study are:

- To develop and evaluate an AutoML-based ensemble framework (AutoGluon) for predicting 30-day readmissions in diabetic patients.
- To compare the performance of ensemble methods with traditional ML models, the DL model, and the foundation model for tabular data.
- To analyze the impact of different feature sets and preprocessing strategies on predictive performance.
- To identify key risk factors contributing to readmission and evaluate their clinical relevance.
- To assess the generalizability of the model through sub-group analysis.

1.2 Related Works

Hai et al. (2023) employed Long Short-Term Memory (LSTM) networks to predict unplanned 30-day readmissions among over 36,000 diabetic patients using longitudinal EHR data from a large urban academic health system. Their results showed that LSTM outperformed traditional ML approaches—including Random Forest (RF), Logistic Regression (LR), and AdaBoost—achieving an AUROC of 0.79. The study also explored the impact of historical encounter data and feature engineering techniques such as Singular Value Decomposition (SVD). Notably, they demonstrated that LSTMs benefited from longitudinal input and achieved strong performance even with a reduced lab feature set.

Liu et al. (2024) conducted a comparative study using the University of California Irvine (UCI) “Diabetes 130-US Hospitals” dataset (Clore & Strack, 2014) to evaluate 11 models, including RF, XGBoost, SVMs, and LSTM. Feature selection was optimized using the Grey Wolf Optimizer (GWO). In contrast to Hai et al. (2023), this study found that traditional ensemble models—especially RF and XGBoost—outperformed DL methods, with higher AUROC scores. These findings underscore the robustness of classical ML approaches for structured tabular data and emphasize the role of feature selection and preprocessing in influencing model performance.

Shang et al. (2021) similarly examined 30-day readmissions with the same UCI diabetes dataset (Clore & Strack, 2014). They compared three ML models—RF, Naïve Bayes, and Tree Ensemble—and identified 23 key predictive features. RF achieved the highest AUC (0.661), and the study emphasized the value of structured pipelines, careful feature selection, and data balancing methods like Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002). Their findings support Liu et al. (2024)’s conclusion that readmissions for diabetic patients are predictable and that traditional ML models remain highly competitive in real-world healthcare settings.

2 Methods

2.1 Cohort Characteristics

This study utilizes the publicly available dataset from UCI Machine Learning Repository *"Diabetes 130-US hospitals for years 1999–2008"*(Clore & Strack, 2014). The dataset contains 10 years of de-identified clinical data from over 130 hospitals and integrated delivery networks across the Midwest, Northwest, South, and West regions of the United States. Each record corresponds to a single inpatient encounter of a patient diagnosed with diabetes, including details on demographics, lab results, medications, diagnoses, procedures, and hospital readmission as primary outcomes. All patients included had hospital stays of up to 14 days and received at least one medication or laboratory test.

Variable		**Overall, n (%)**	**Readmitted = 0, n (%)**	**Readmitted = 1, n (%)**
Female		38024 (53.2%)	35372 (53.2%)	2652 (52.4%)
Medical specialty	Cardiology	4152 (5.8%)	3910 (5.9%)	242 (4.8%)
	Emergency Medicine	4450 (6.2%)	4188 (6.3%)	262 (5.2%)
	Internal Medicine	16052 (22.4%)	14789 (22.3%)	1263 (24.9%)
	Nephrology	866 (1.2%)	790 (1.2%)	76 (1.5%)
	Obstetrics & Gynecology	683 (1.0%)	668 (1.0%)	15 (0.3%)
	Orthopedics	2105 (2.9%)	1991 (3.0%)	114 (2.3%)
	Psychiatry & Psychology	720 (1.0%)	663 (1.0%)	57 (1.1%)
	Pulmonology	664 (0.9%)	604 (0.9%)	60 (1.2%)
	Radiology	852 (1.2%)	808 (1.2%)	44 (0.9%)
	Surgery	3849 (5.4%)	3602 (5.4%)	247 (4.9%)
	Other	37122 (51.9%)	34437 (51.8%)	2685 (53.0%)
Race	Caucasian	53513 (74.8%)	49622 (74.7%)	3891 (76.8%)
	AfricanAmerican	12903 (18.0%)	12036 (18.1%)	867 (17.1%)
	Other	5099 (7.1%)	4792 (7.2%)	307 (6.1%)
Admission type	Elective	13898 (19.4%)	13040 (19.6%)	858 (16.9%)
	Emergency	49840 (69.7%)	46244 (69.6%)	3596 (71.0%)
	New Born	9 (0.0%)	8 (0.0%)	1 (0.0%)
	Not Available	7747 (10.8%)	7137 (10.7%)	610 (12.0%)
	Trauma Center	21 (0.0%)	21 (0.0%)	0 (0.0%)
	Birth	3 (0.0%)	3 (0.0%)	0 (0.0%)
	Referral	22712 (31.8%)	21216 (31.9%)	1496 (29.5%)
	Transfer	5152 (7.2%)	4807 (7.2%)	345 (6.8%)
	Unknown	5168 (7.2%)	4793 (7.2%)	375 (7.4%)
	Other	38480 (53.8%)	35631 (53.6%)	2849 (56.2%)
Discharge Disposition	Expired	1273 (1.8%)	1273 (1.9%)	0 (0.0%)
	Home	51956 (72.7%)	48853 (73.5%)	3103 (61.3%)
	Hospice	561 (0.8%)	541 (0.8%)	20 (0.4%)
	Outpatient	14 (0.0%)	11 (0.0%)	3 (0.1%)
	Transfer	14181 (19.8%)	12452 (18.7%)	1729 (34.1%)
	Unknown	3150 (4.4%)	2962 (4.5%)	188 (3.7%)
	Other	380 (0.5%)	358 (0.5%)	22 (0.4%)
Primary Diagniosis	Circulatory	21599 (30.2%)	19956 (30.0%)	1643 (32.4%)
	Diabetes	5702 (8.0%)	5266 (7.9%)	436 (8.6%)
	Digestive	6665 (9.3%)	6226 (9.4%)	439 (8.7%)
	Genitourinary	3560 (5.0%)	3302 (5.0%)	258 (5.1%)
	Injury	4885 (6.8%)	4488 (6.8%)	397 (7.8%)
	Musculoskeletal	3945 (5.5%)	3696 (5.6%)	249 (4.9%)
	Neoplasms	2827 (4.0%)	2629 (4.0%)	198 (3.9%)
	Respiratory	9679 (13.5%)	9122 (13.7%)	557 (11.0%)
	Other	12653 (17.7%)	11765 (17.7%)	888 (17.5%)
Age group	0-10	154 (0.2%)	151 (0.2%)	3 (0.1%)
	10-20	528 (0.7%)	510 (0.8%)	18 (0.4%)
	20-30	1122 (1.6%)	1056 (1.6%)	66 (1.3%)
	30-40	2678 (3.7%)	2532 (3.8%)	146 (2.9%)
	40-50	6818 (9.5%)	6436 (9.7%)	382 (7.5%)
	50-60	12437 (17.4%)	11709 (17.6%)	728 (14.4%)
	60-70	15925 (22.3%)	14808 (22.3%)	1117 (22.1%)
	70-80	18185 (25.4%)	16717 (25.2%)	1468 (29.0%)
	80-90	11709 (16.4%)	10706 (16.1%)	1003 (19.8%)
	90-100	1959 (2.7%)	1825 (2.7%)	134 (2.6%)

Table 1: Distribution of patient characteristics stratified by 30-day readmission status after data preprocessing.

2.2 Data Preprocessing

The dataset underwent a comprehensive series of preprocessing steps. Demographic variables were grouped to reduce sparsity: race was collapsed into three categories—African American, Caucasian, and Other—and age intervals were mapped to their midpoint values to create numerical representations. To address potential correlation bias introduced by patients with multiple encounters, only the single encounter with the longest hospital stay (`time_in_hospital`) was retained for each patient, which reduced the dataset from 101,766 to 71,518 rows.

Categorical features such as medical specialty, admission type, admission source, and discharge disposition were consolidated into clinically meaningful groups and then one-hot encoded. Medical specialties were grouped into 11 categories: Cardiology, Emergency Medicine, Internal Medicine, Nephrology, Obstetrics & Gynecology, Orthopedics, Psychiatry & Psychology, Pulmonology, Radiology, Surgery, and Other. Admission types were grouped into 9 categories including Elective, Emergency, Newborn, Trauma Center, Birth, Referral, Transfer, Other, and Unknown. Diagnosis codes (`diag_1`, `diag_2`, and `diag_3`) originally consisted of hundreds of ICD-9 codes. To reduce dimensionality, they were grouped into 9 broader diagnostic categories: Circulatory, Diabetes, Digestive, Genitourinary, Injury, Musculoskeletal, Neoplasms, Respiratory, and Other, following the clinical mapping approach used by Strack et al. (2014). Finally, the outcome variable, `readmitted`, was binarized to readmission within 30 days (positive) and no readmission or readmission after 30 days (negative).

Features with excessive missing values, such as `weight` and `payer_code`, were removed. Constant features such as `examide`, `citoglipton`, and `glimepiride-pioglitazone` were dropped due to a lack of variability. Additionally, features with low variance (less than 0.1) were excluded to reduce noise. To address multicollinearity, highly correlated features were detected using Cramér's V for categorical variables and Pearson's correlation for numerical variables (Cramér, 1999; Pearson, 1895). For each pair of correlated features, we kept the one more strongly related to the target variable and removed the other.

The dataset was randomly split into training and testing sets using an 80:20 ratio, stratified by the target label to preserve the original class distribution across both sets. Numerical features were then normalized separately using standard scaling, ensuring that the scaling parameters were derived only from the training set to prevent data leakage.

To assess the impact of different feature sets and preprocessing strategies on model performance, the pipeline includes several optional components. Feature selection can be applied using `SelectKBest`, a univariate statistical technique that retains the top k features most strongly associated with the target. Class imbalance is mitigated using the SMOTE and Adaptive Synthetic Sampling Approach (ADASYN) (He et al., 2008), which synthetically generates new samples in the training set for the minority class to improve model performance. Additionally, K-means clustering (MacQueen, 1967) can be employed to extract latent patterns within the data, with the resulting cluster labels added as new features to enhance the model's ability to capture hidden subpopulation structures. The complete pipeline is illustrated in Figure 1.

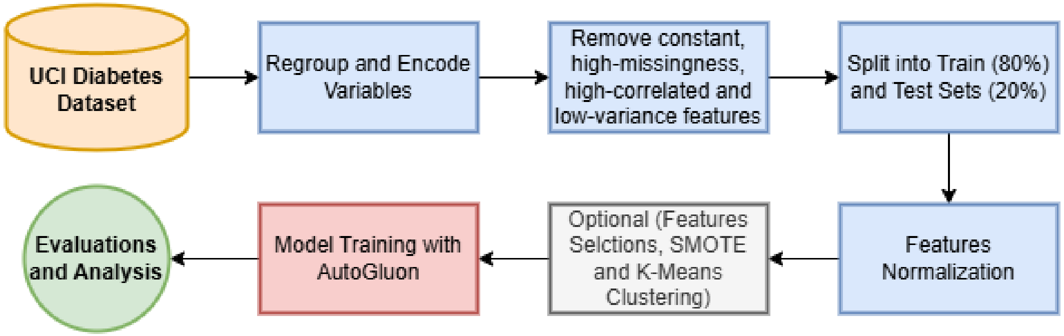

UCI Diabetes Dataset
Regroup and Encode Variables
Remove constant, high-missingness, high-correlated and low-variance features
Split into Train (80%) and Test Sets (20%)
Features Normalization
Optional (Features Selctions, SMOTE and K-Means Clustering)
Model Training with AutoGluon
Evaluations and Analysis

Figure 1: Overview of the data preprocessing and model training pipeline.

2.3 Models Training and Evaluation

In this study, I employed `AutoGluon`, an open-source framework designed to automate ML tasks with minimal manual intervention. It introduced a multi-layer ensembling strategy that combined the predictions of diverse base learners using techniques such as weighted stacking and k-fold bagging. Prior research has shown that ensemble-based approaches often outperform individually trained models by reducing variance and improving generalization performance (Erickson et al., 2020).

ML Model (with Citation)	**Description**
`WeightedEnsemble` (Erickson et al., 2020)	Meta-learner that combines the predictions of multiple base models using weighted averaging to improve predictive performance and robustness.
`LightGBM` (Ke et al., 2017)	A fast, efficient gradient boosting framework that uses histogram-based decision trees and leaf-wise growth for optimized performance on large datasets.
`XGBoost` (Chen & Guestrin, 2016)	A scalable, regularized boosting technique designed for efficiency and accuracy, with built-in support for missing data and parallel processing.
`CatBoost` (Prokhorenkova et al., 2018)	A gradient boosting algorithm with advanced handling of categorical variables and ordered boosting to reduce overfitting.
`Logistic Regression`	A linear model for binary classification that estimates the probability of a target event using the logistic function over a weighted sum of input features.
`RandomForest` (Breiman, 2001)	An ensemble of decision trees built using bootstrapped samples and random feature subsets, known for reducing overfitting and improving generalization.
`ExtraTrees` (Geurts et al., 2006)	An ensemble method that builds randomized decision trees using random splits and feature subsets, promoting model diversity and lower variance.
`Multi-layer Perceptron (MLP)`	A feedforward neural network implemented that learns complex, non-linear relationships but may require careful tuning and regularization to prevent overfitting.
`TabPFNMix` (Hollmann et al., 2022)	A transformer-based foundation model trained on synthetic data using in-context learning, enabling fast adaptation to tabular classification tasks with minimal fine-tuning.

Table 2: Descriptions of models used in this study.

I trained and evaluated 9 models within the `AutoGluon` framework. As a baseline, I used `logistic regression`, a widely adopted and interpretable algorithm for binary classification. I then benchmarked its performance against advanced gradient boosting models—`LightGBM`, `XGBoost`, and `CatBoost`—and tree-based ensemble methods such as `RandomForest` and `ExtraTrees`. Additionally, I included a `multi-layer perceptron` to capture complex nonlinear relationships, and `TabPFNMix`, a transformer-based tabular foundation model pre-trained on synthetic data. All base learners were ultimately integrated using `WeightedEnsemble`, a meta-learner that combines predictions from individual models to enhance overall accuracy. Descriptions of all models are summarized in Table 3. To evaluate the impact of different preprocessing strategies, I conducted experiments under six distinct settings: (1) training without cluster features (baseline), (2) training with cluster features added to the original variables, (3) training using only the top 10 most important features based on Random Forest importance, and (4–6) applying class imbalance correction techniques—SMOTE, ADASYN, and downsampling, respectively. During model training, I applied Bayesian optimization (Garnett, 2023) to fine-tune the hyperparameters of each base learner within AutoGluon. The tuned hyperparameter search spaces for all models are summarized in Table **??**. The rationale was to first establish a baseline without any strategies and then sequentially assess the influence of latent subgroup information (clusters), dimensionality reduction (top features), and class balancing on model performance. This approach allowed us to systematically explore the contribution of each strategy to improving performance.

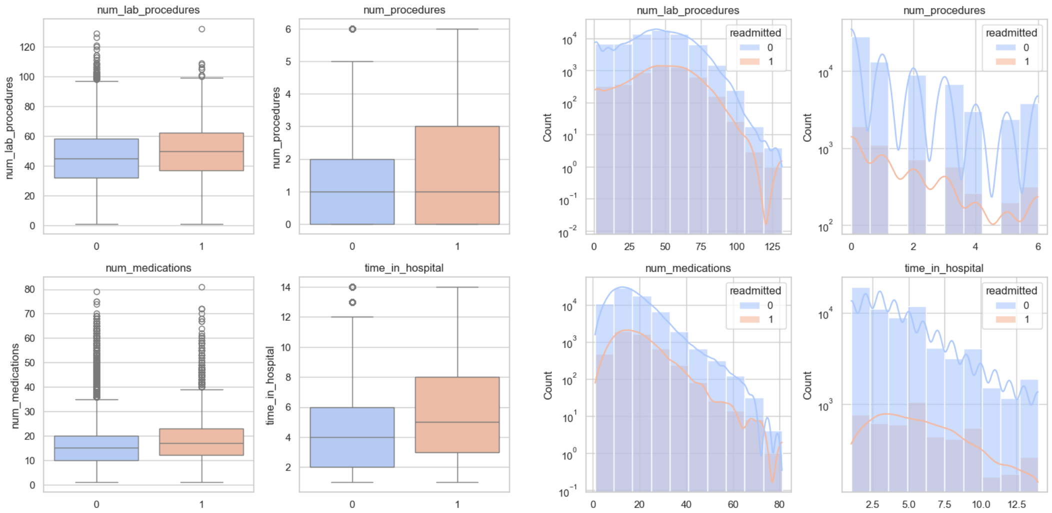
For model evaluation, I used the Area Under the Receiver Operating Characteristic Curve (AUROC) on the test set. The model achieving the highest AUROC was selected for subsequent feature importance and subgroup analyses. To investigate feature importance, I first employed the built-in feature ranking from the Random Forest model to identify the most influential predictors. I then applied SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017) values to further interpret and validate the importance of these features. SHAP provided insight into both the magnitude and direction of each feature's effect on the model's predictions, offering greater clinical interpretability. To assess generalizability, I conducted subgroup analyses across different races, age groups, primary diagnoses, and discharge dispositions. For each subgroup, model performance was evaluated using both AUROC and the Area Under the Precision-Recall Curve (AUPRC) to assess the model's generalizability across diverse patient populations.

3 Results

3.1 Exploratory Data Analysis

I first analyzed the distributions and correlations of features to better understand their relationship with the readmission outcome. Table 1 summarizes the counts of key categorical variables stratified by readmission status. Among medical specialties, Internal Medicine and Emergency Medicine were the most represented across both groups. Internal Medicine showed a higher proportion in the readmitted group (24.9%) than the non-readmitted group (22.3%), while Obstetrics & Gynecology had a much lower readmission rate (0.3%). Discharge disposition shows 34.1% of readmitted patients were transferred, compared to only 18.7% among those not readmitted. In contrast, home discharges were approximately 12% more common in the non-readmitted group.

To examine continuous features, I selected `num_lab_procedures`, `num_procedures`, `num_medications`, and `time_in_hospital` based on their clinical relevance to treatment intensity and resource use. As shown in Figure 2, boxplots (A) and histograms (B) reveal right-skewed distributions across all variables. Notably, the readmitted group generally had higher medians and wider spreads, particularly in hospital stay duration and medication counts, suggesting a potential link between increased care and readmission risk.



(a) Boxplots by readmission status

(b) Log-Scale histograms by readmission status

Figure 2: Distribution of selected numerical features by readmission status.

The correlation heatmap (Supplementary FigureS1) shows that most feature pairs demonstrate low correlations, suggesting minimal multicollinearity overall. However, some weak to moderate correlations are observed among features related to discharge disposition IDs.

3.2 Model Comparisons

Table 3 summarizes AUROC performance across all models under different feature and preprocessing configurations. For the feature sets, I evaluated: (1) models without clustering features, (2) models with clustering features, (3) models trained using only the top 10 most important features. For preprocessing, I explored the impact of applying (4) SMOTE, (5) ADASYN, and (6) downsampling to address class imbalance.

Across all configurations, the `WeightedEnsemble` achieved the highest AUROC in 3 out of 6 configurations (No Clustering (0.699), With Clustering (0.699), and Down Sampling (0.695) and was among the top two across all settings. Gradient boosting models such as LightGBM and CatBoost showed AUROC values close to the ensemble model. XGBoost had lower AUROC compared to other boosting methods in this dataset. TabPFNMix and neural networks reported lower AUROC values across all configurations.

Model	No Clustering	With Clustering	Top 10 Features	SMOTE	ADASYN	Down Sampling
WeightedEnsemble_L2	**0.699**	**0.699**	0.667	0.637	0.633	**0.695**
LightGBM	0.699	0.697	0.665	**0.657**	**0.658**	0.691
CatBoost	0.698	0.696	**0.668**	0.637	0.637	0.690
XGBoost	0.627	0.623	0.593	0.621	0.624	0.658
ExtraTrees	0.687	0.682	0.666	0.595	0.595	0.678
RandomForest	0.691	0.688	0.662	0.633	0.633	0.688
NeuralNetTorch	0.681	0.685	0.661	0.612	0.628	0.672
TabPFNMix	0.641	0.638	0.644	0.589	0.582	0.682
Logistic Regression	0.692	0.691	0.661	0.582	0.581	0.685

Table 3: Model AUROC comparisons under different settings.

Table 4 presents detailed classification metrics for the ensemble model across all configurations. The model with clustering features had the highest precision (0.1558) and F1 score (0.2273). The model trained with ADASYN achieved the highest recall (0.5726). Models using SMOTE or ADASYN had lower AUROC and precision. The configuration using only the top 10 features had the highest accuracy (0.8266), with lower recall and F1 scores.

Model Type	Accuracy	Precision	Recall	F1	AUROC
No Clustering	0.7543	0.1428	0.4936	0.2215	**0.6991**
With Clustering	0.7975	**0.1558**	0.4205	**0.2273**	0.6987
Top 10 Features	**0.8266**	0.1502	0.3110	0.2026	0.6672
SMOTE	0.6321	0.1053	0.5597	0.1773	0.6374
ADASYN	0.6180	0.1034	**0.5726**	0.1751	0.6331
Down Sampling	0.7649	0.1442	0.4699	0.2207	0.6946

Table 4: Performance metrics of the Ensemble model under different settings.

The AUROC curves for the ensemble model across different settings are shown in Figure 3. The ensemble model trained without clustering features achieved the highest AUROC and was therefore selected for downstream analysis.

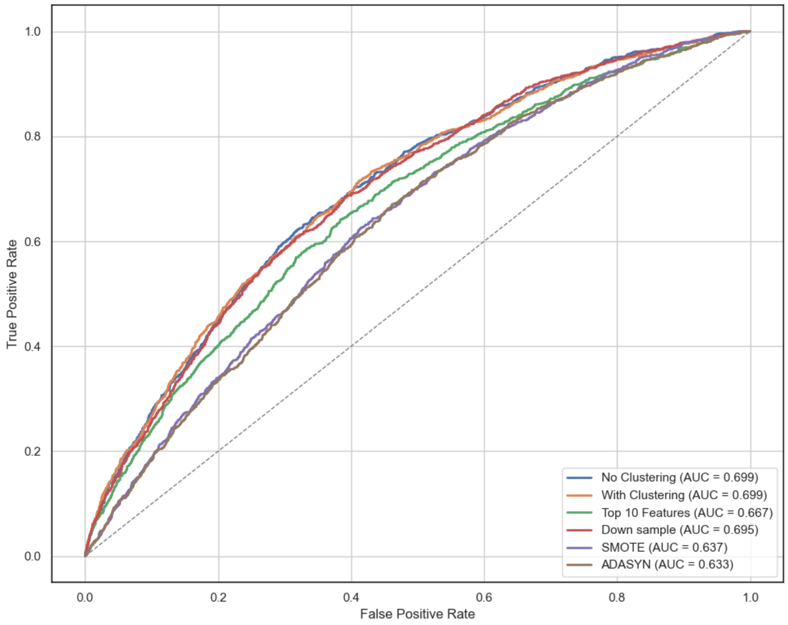


Figure 3: AUROC curves for the ensemble model under different experimental settings.

3.3 Feature Importance

To identify the most influential predictors of readmission risk, I conducted feature importance analysis using two different methods: Random Forest feature importance and SHAP values.

Figure 4 shows the top 20 features ranked by Random Forest importance scores.The most influential predictor was `discharge_disposition_id_Transfer`, followed by `number_inpatient`, `time_in_hospital`, `discharge_disposition_id_Home`, and `num_lab_procedures`. These features suggest that patients who were transferred, had frequent prior inpatient visits, or experienced longer hospital stays were at higher risk of readmission. Diagnosis-related features and treatment intensity variables (e.g., `num_lab_procedures`) also ranked highly, underscoring their relevance in predicting readmission.

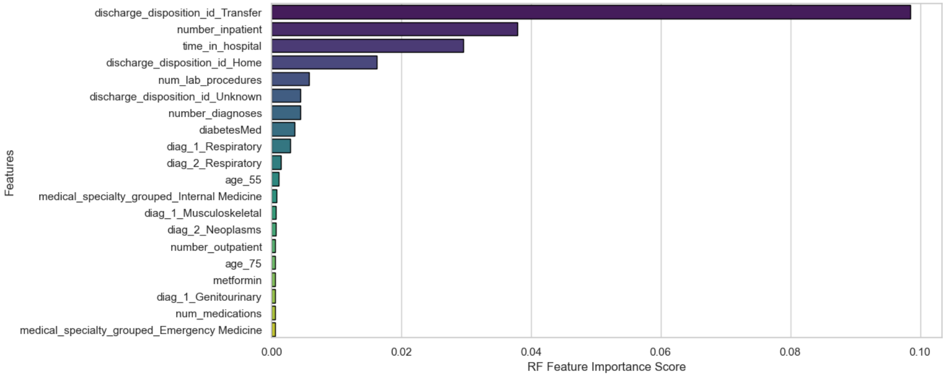


Figure 4: Top 20 features ranked by Random Forest feature importance.

Figure 5 presents SHAP summary plots, which also visualize the direction and magnitude of each feature's impact on model output. Similar with the Random Forest results, `discharge_disposition_id_Transfer`, `number_inpatient`, and `time_in_hospital` emerged as top contributors. The SHAP plot further reveals that higher values of these features (shown in red) were associated with an increased risk of readmission. Additionally, variables like `diabetesMed` and several diagnosis groups exhibited directional effects, offering clinically interpretable insights.

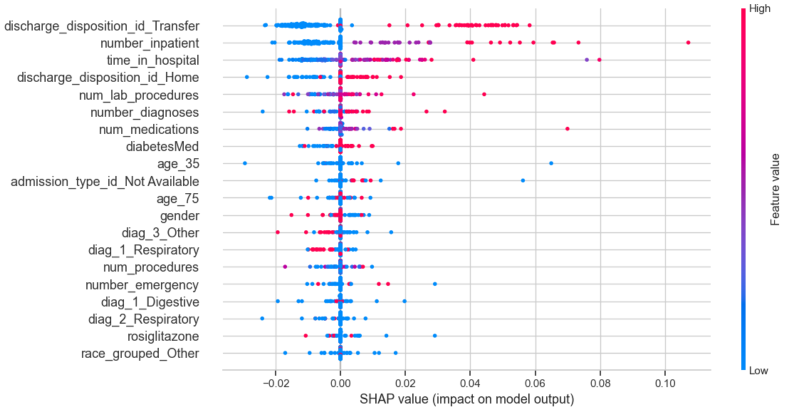


Figure 5: SHAP summary plot.

In summary, among the top 20 features, 13 were shared between both methods, and all of the top five were consistent across approaches. While RF importance provides a global measure of feature relevance through impurity reduction, SHAP adds transparency by quantifying how individual feature values affect prediction probabilities. Together, they suggested that patients with more prior inpatient visits, longer hospital stays, and discharge dispositions involving transfers are more likely to be readmitted

3.4 Subgroup Analysis

To evaluate the model's performance across diverse patient characteristics, I conducted subgroup analyses using AUROC, AUPRC, and observed readmission rates (prevalence), stratified by primary diagnosis, discharge disposition, age group, and race. The results are presented in Figure S2, with each metric shown with 95% confidence intervals.

Figure S2a shows the overall prevalence of readmission was low across all subgroups. Among discharge disposition categories, the `Transfer` group showed the highest readmission rate (0.127, 95% CI: 0.115–0.137), while the `Hospice` group had the lowest (0.038, 95% CI: 0.009–0.069). For age groups, patients aged 75 exhibited a slightly higher readmission rate (0.086, 95% CI: 0.077–0.096), whereas patients aged 15 had the lowest (0.041, 95% CI: 0.010–0.080). These patterns indicate differences in subgroup-level base rates of readmission.

Figure S2b shows AUPRC values for each subgroup. The highest AUPRC values were observed in patients discharged to `Other` (0.369, 95% CI: 0.084–0.714) and in the `age_25` group (0.336, 95% CI: 0.162–0.545). These groups also displayed wider confidence intervals. Subgroups such as the `diag_1_Injury` category (0.137, 95% CI: 0.097–0.208) and the `age_45` group (0.120, 95% CI: 0.071–0.190) had the lowest AUPRC values, though all subgroups demonstrated values above their respective prevalence rates.

Figure S2c displays AUROC across the same subgroups. Patients with `diag_1_Neoplasms` conditions had the highest AUROC (0.763, 95% CI: 0.699–0.837), followed by those with `diag_1_Musculoskeletal` (0.758, 95% CI: 0.703–0.815). Lower AUROC values were seen in discharge-related subgroups such as `Transfer` (0.625, 95% CI: 0.595–0.652) and in the `age_95` group (0.636, 95% CI: 0.539–0.727). Across most groups, AUROC values were close to the overall AUROC of 0.699.

4 Discussion

4.1 Key Findings

This study presents a comprehensive evaluation of ML approaches for predicting 30-day readmissions among diabetic patients, with a focus on automated ML using the AutoGluon framework. AutoGluon is an open-source AutoML toolkit that has shown strong performance across various domains, including tabular data, image classification, and natural language processing (Erickson et al., 2020). To the best of my knowledge, this is the first study to apply AutoGluon to the clinical problem of hospital readmission prediction in diabetic populations. I systematically compared 9 models, spanning traditional ML algorithms, deep neural networks, ensemble methods, and a state-of-the-art transformer-based tabular foundation model.

My results show that the ensemble-based `WeightedEnsemble` model achieved strong and consistent performance across different experimental configurations. It ranked first in 3 out of 6 configurations and remained within the top two across all settings. The highest AUROC (0.699) was observed both with and without clustering features, and the model also maintained high performance under the downsampling strategy (AUROC = 0.695). Among base models, gradient boosting frameworks such as LightGBM, XGBoost, and CatBoost showed strong performance. Notably, LightGBM achieved AUROC scores comparable to the ensemble model in multiple settings. These findings align with recent literature showing that boosting-based methods remain highly effective for structured tabular prediction tasks(Grinsztajn et al., 2022; Shwartz-Ziv & Armon, 2022).

4.2 Comparison with Existing Literature

When compared with previous studies on the same dataset, my results are largely consistent with the AUROC values reported in the literature. For instance, Liu et al. (2024) reported a maximum AUROC of 0.64 using XGBoost, while Shang et al. (2021) achieved an AUROC of 0.661 using Random Forest. In contrast, the ensemble-based model achieved a higher AUROC of 0.699, with boosting models such as LightGBM and CatBoost also delivering competitive performance. Notably, Liu et al.

(2024) reported F1 scores exceeding 0.84. However, such results could not be reproduced even after adopting similar preprocessing and modeling steps. Since the original code is not publicly available, it is difficult to assess the exact methodology used and whether proper validation procedures were followed.

I also reviewed top-performing solutions on Kaggle(Brandao, 2018), a public platform where users compete on standardized datasets. A recurring concern was the incorrect application of data balancing techniques, such as SMOTE to the entire dataset before the train-test split. This approach artificially inflates the target distribution in both the training and the test sets, resulting in overly optimistic F1 scores. In contrast, when class balancing techniques are applied correctly—within the training set only—performance metrics are much closer to those observed in this study. Based on the experimental design and validation, I found that predicting 30-day readmission using this dataset is challenging due to the high class imbalance. As such, relatively low to modest F1 scores are expected and likely more representative of real-world deployment.

4.3 Impact of Feature Engineering and Preprocessing

I explored how different feature engineering and preprocessing strategies influenced model performance. Adding cluster from K-means as features slightly improved F1 scores while minimally affecting AUROC, suggesting that latent subgroups offer predictive signal. Applying SMOTE and ADASYN improved recall by generating synthetic samples for the minority class, helping the model identify more readmission cases. However, this came at the cost of reduced precision and AUROC, reflecting the typical trade-off between sensitivity and specificity in imbalanced datasets.

Interestingly, training models on only the top 10 most important features still yielded relatively strong performance. This suggests the feasibility of deploying simplified models in clinical environments where computational or data resources may be limited.

4.4 Performance of Deep Learning Models

Both TabPFNMix and the multi-layer perceptron (MLP) model, which are neural network-based approaches, underperformed relative to traditional models in this experiment. TabPFNMix, a recently proposed transformer-based tabular foundation model that leverages in-context learning, did not outperform boosting models despite its theoretical potential. This may be attributed to the static and structured nature of the dataset, absence of temporal information, and the limited need for complex representation learning in this context. Furthermore, TabPFNMix was used without fine-tuning, which may have restricted its adaptability. Similarly, the MLP model showed lower performance, even falling below logistic regression in settings with reduced feature sets, such as training on only the top 10 variables. This likely reflects the dataset's modest size, which may be insufficient for deep models to learn generalizable patterns without overfitting, even with dropout regularization. Neural network models, including both MLP and TabPFNMix, are known to be sensitive to hyperparameter settings and typically require larger, more complex datasets to perform effectively.

4.5 Feature Importance and Interpretability

Feature importance analyses using RF and SHAP produced consistent results. 13 of the top 20 features were shared across both methods, with the top five features identical. The most influential predictors included `discharge_disposition_id_Transfer`, `number_inpatient`, and `time_in_hospital`—features intuitively linked to readmission risk. These findings support prior work emphasizing discharge-related factors and prior hospitalization history as strong indicators of early readmission risk. The SHAP plots further revealed that higher values of these features increased predicted readmission risk, providing clinically interpretable insights.

4.6 Subgroup Analysis

The subgroup analysis indicated that the model generalized well across clinical and demographic groups, despite the overall low prevalence of readmissions. Subgroup AUROC and AUPRC scores were generally consistent with overall performance, and AUPRC values exceeded subgroup-level prevalence in all cases. These results support the potential of this model for deployment in diverse clinical populations.

4.7 Limitations

This study has several limitations. The dataset is retrospective and cross-sectional, limiting the potential for capturing temporal patterns that may benefit time-aware models. Additionally, the population was predominantly Caucasian, with limited representation of other racial and ethnic groups, potentially affecting the model's generalizability. Some clinically important features—such as `A1Cresult`, which reflects long-term glucose control—had a high rate of missingness. Although this feature was retained during modeling to preserve clinically relevant signals, the extent of missing data may have reduced its predictive performance. Future research should incorporate longitudinal EHR data, additional social and behavioral health variables, and external validation across more heterogeneous populations.

5 Conclusions

In summary, this study demonstrates the effectiveness of AutoML-based ensemble modeling for predicting 30-day hospital readmissions in diabetic patients. By systematically evaluating a range of ML models, feature sets, and preprocessing strategies, this study offers a comprehensive assessment of model performance, interpretability, and generalizability. Among all approaches tested, AutoGluon's ensemble method consistently achieved the highest AUROC across configurations, highlighting its potential as a robust tool for readmissions risk prediction. These findings suggest that such automated, ensemble-based methods can support the early identification of high-risk patients using structured EHR data. Future work should explore the deployment of AutoGluon-based models in real-time clinical workflows, assess their utility in longitudinal datasets, and validate performance across more diverse patient populations to ensure broad applicability in real-world settings.

6 Code Availability

All code used in this study, including data preprocessing, model training, evaluation, and visualization scripts, will be made publicly available for reproducibility and further research. The repository will be hosted at:

`https://github.com/NokeYuan/Diabetes-Readmission-AutoGluon`

Please check the repository for detailed instructions on environment setup, data handling, and running the experiments.

References

Brandao, G. Diabetes 130-us hospitals for years 1999-2008. `https://www.kaggle.com/datasets/brandao/diabetes`, 2018. Accessed: 2025-03-28.

Breiman, L. Random forests. *Machine learning*, 45:5–32, 2001.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Clore, John, C. K. D. J. and Strack, B. Diabetes 130-US Hospitals for Years 1999-2008. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5230J.

Cramér, H. *Mathematical methods of statistics*, volume 9. Princeton university press, 1999.

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., and Smola, A. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*, 2020.

Garnett, R. *Bayesian optimization*. Cambridge University Press, 2023.

Geurts, P., Ernst, D., and Wehenkel, L. Extremely randomized trees. *Machine learning*, 63:3–42, 2006.

Gregory, N. S., Seley, J. J., Dargar, S. K., Galla, N., Gerber, L. M., and Lee, J. I. Strategies to prevent readmission in high-risk patients with diabetes: the importance of an interdisciplinary approach. *Current diabetes reports*, 18:1–7, 2018.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35:507–520, 2022.

Hai, A. A., Weiner, M. G., Paranjape, A., Livshits, A., Brown, J. R., Obradovic, Z., and Rubin, D. J. Deep learning vs traditional models for predicting hospital readmission among patients with diabetes. In *AMIA Annual Symposium Proceedings*, volume 2022, pp. 512, 2023.

He, H., Bai, Y., Garcia, E. A., and Li, S. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pp. 1322–1328. Ieee, 2008.

Hicks, C. W., Canner, J. K., Karagozlu, H., Mathioudakis, N., Sherman, R. L., Black III, J. H., and Abularrage, C. J. Contribution of 30-day readmissions to the increasing costs of care for the diabetic foot. *Journal of vascular surgery*, 70(4):1263–1270, 2019.

Hollmann, N., Müller, S., Eggensperger, K., and Hutter, F. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*, 2022.

Karunakaran, A., Zhao, H., and Rubin, D. J. Predischarge and postdischarge risk factors for hospital readmission among patients with diabetes. *Med Care*, 56(7):634–642, July 2018.

Kassin, M. T., Owen, R. M., Perez, S. D., Leeds, I., Cox, J. C., Schnier, K., Sadiraj, V., and Sweeney, J. F. Risk factors for 30-day hospital readmission among general surgery patients. *Journal of the American College of Surgeons*, 215(3):322–330, 2012.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

Kondratyuk, D., Tan, M., Brown, M., and Gong, B. When ensembling smaller models is more efficient than single large models. *arXiv preprint arXiv:2005.00570*, 2020.

Kum Ghabowen, I., Epane, J. P., Shen, J. J., Goodman, X., Ramamonjiarivelo, Z., and Zengul, F. D. Systematic review and meta-analysis of the financial impact of 30-day readmissions for selected medical conditions: A focus on hospital quality performance. In *Healthcare*, volume 12, pp. 750. MDPI, 2024.

Liu, V. B., Sue, L. Y., and Wu, Y. Comparison of machine learning models for predicting 30-day readmission rates for patients with diabetes. *Journal of Medical Artificial Intelligence*, 7, 2024.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

MacQueen, J. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, volume 5, pp. 281–298. University of California press, 1967.

Ojo, O. A., Ibrahim, H. S., Rotimi, D. E., Ogunlakin, A. D., and Ojo, A. B. Diabetes mellitus: From molecular mechanism to pathophysiology and pharmacology. *Medicine in Novel Technology and Devices*, 19:100247, 2023.

Pearson, K. Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 58(347-352):240–242, 1895.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018.

Rubin, D. J. Hospital readmission of patients with diabetes. *Current diabetes reports*, 15:1–9, 2015.

Shang, Y., Jiang, K., Wang, L., Zhang, Z., Zhou, S., Liu, Y., Dong, J., and Wu, H. The 30-days hospital readmission risk in diabetic patients: predictive modeling with machine learning classifiers. *BMC medical informatics and decision making*, 21:1–11, 2021.

Shwartz-Ziv, R. and Armon, A. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., and Clore, J. N. Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014(1):781670, 2014.

7 Supplementary Materials

7.1 Correlation Matrix

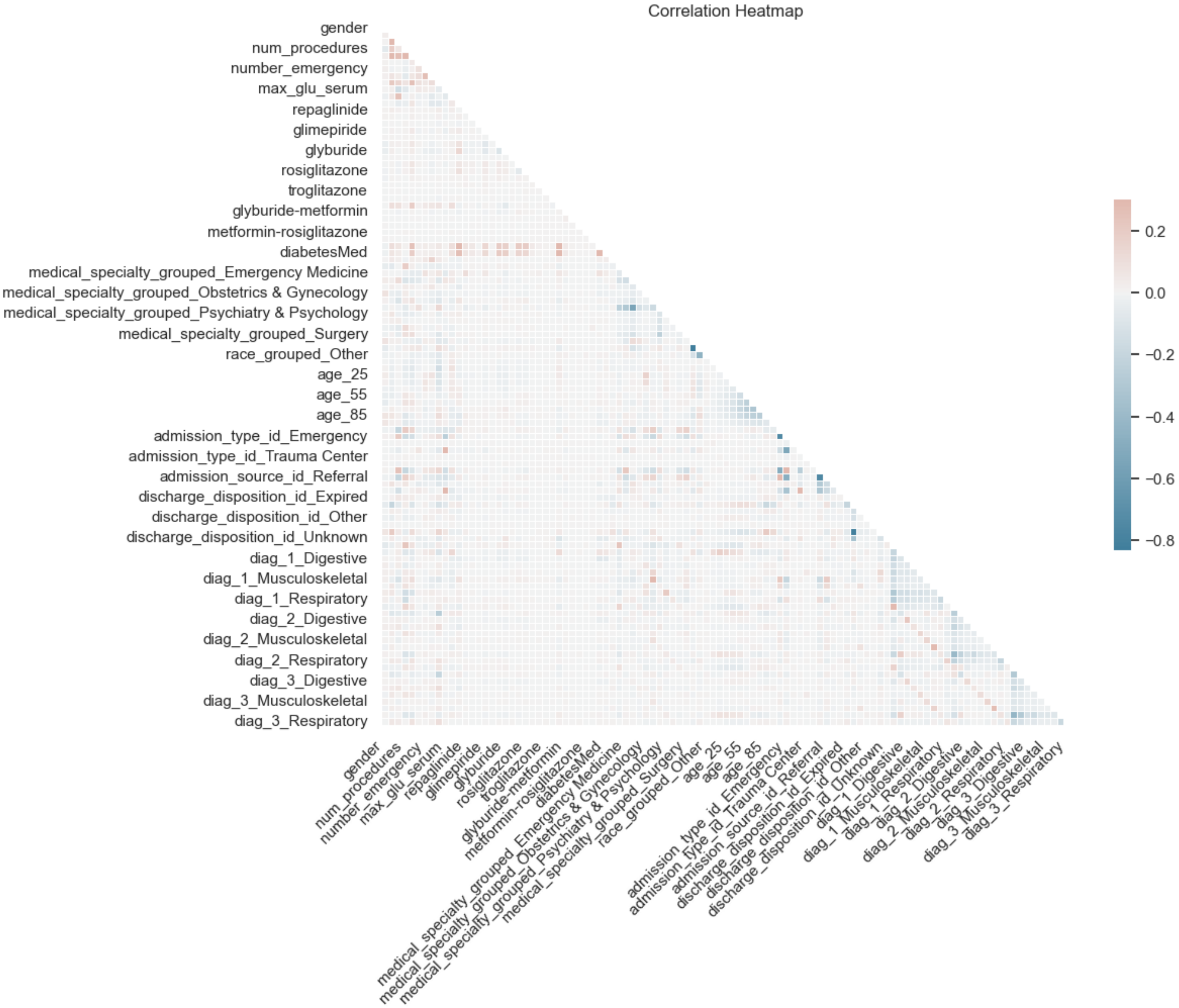
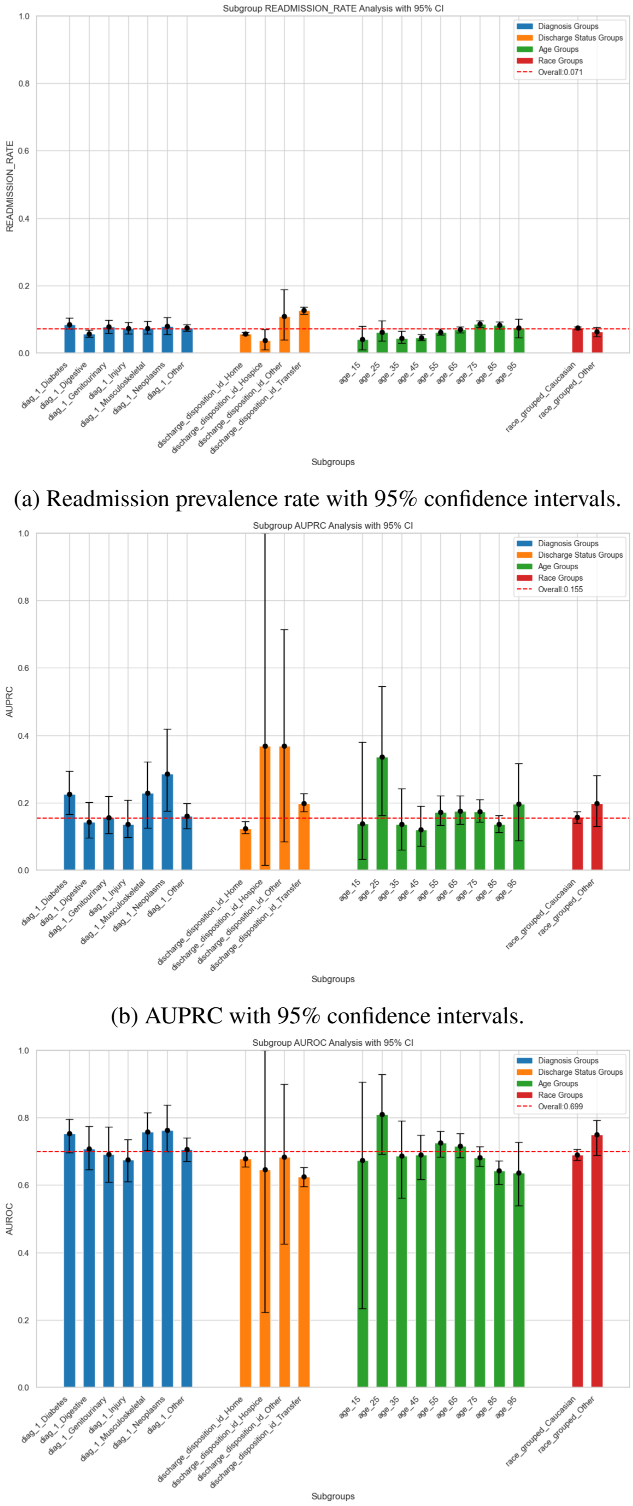


Figure S1: Correlation heatmap.

7.2 Subgroup Analysis



(a) Readmission prevalence rate with 95% confidence intervals.

(b) AUPRC with 95% confidence intervals.

(c) AUROC with 95% confidence intervals.

Figure S2: Subgroup analysis across diagnosis, discharge disposition, age, and race groups. Red dashed lines represent overall averages for each metric.

7.3 Hyperparameter Tuning Spaces for Each Model

Model	Hyperparameter Search Space
LightGBM (GBM)	learning_rate $\in$ [0.01, 0.1] num_leaves $\in$ [16, 64] min_data_in_leaf $\in$ [5, 50] feature_fraction $\in$ [0.5, 1.0]
CatBoost (CAT)	depth $\in$ [4, 10] learning_rate $\in$ [0.01, 0.15] l2_leaf_reg $\in$ [1, 10]
XGBoost (XGB)	scale_pos_weight $\in$ [1, 10] max_depth $\in$ [3, 8] learning_rate $\in$ [0.01, 0.2] min_child_weight $\in$ [1, 10]
Random Forest (RF)	n_estimators $\in$ [100, 300] max_depth $\in$ [10, 30]
Extra Trees (XT)	n_estimators $\in$ [100, 300] max_depth $\in$ [10, 30]
Logistic Regression (LR)	C $\in$ [0.01, 10.0] (log-scale) solver $\in$ {liblinear, saga}

Neural Network (NN_TORCH)	num_epochs = 5
	learning_rate $\in$ [1e-4, 1e-2] (log-scale)
	activation $\in$ {relu, softrelu, tanh}

TabPFNMix	model_path_classifier = N/A

Table S1: Hyperparameter search spaces used for tuning each model in AutoGluon.

7.4 Evaluation Metrics

To evaluate model performance, we employed a range of commonly used classification metrics, including AUROC, AUPRC, sensitivity, specificity, precision, and F1 score.

The Area Under the Receiver Operating Characteristic Curve (AUROC) measures the model's ability to distinguish between the positive and negative classes across all possible classification thresholds. A higher AUROC indicates better overall discrimination capability.

The Area Under the Precision-Recall Curve (AUPRC) is particularly informative for imbalanced datasets. It reflects the trade-off between precision and recall and is especially useful when the positive class is rare.

Sensitivity (also known as recall) is the proportion of actual positive cases that the model correctly identifies and is computed as:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Specificity is the proportion of actual negative cases that are correctly identified by the model:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Precision is the proportion of predicted positive cases that are truly positive:

$$\text{Precision} = \frac{TP}{TP + FP}$$

The F1 score is the harmonic mean of precision and recall and is particularly useful when there is an uneven class distribution. It is defined as:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$