Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Routine to Emergency: A Machine Learning Approach for Predicting ED Visits Among Patients with T2D

Javad Mohammad Alizadeh
javad.m.alizadeh@temple.edu
Temple University
Gabriel Tajeu
University of Alabama at Birmingham
Ilene L Hollin
Temple University
Mukesh Kumar Patel
Temple University
Syeda Maryam Abidi
Temple University
Rosemary Nagy
Temple University
Huanmei Wu
Temple University

Research Article

Keywords: Type 2 diabetes, Emergency department utilization, Machine learning, Predictive modeling, Electronic health records

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11603655/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Routine to Emergency: A Machine Learning Approach for Predicting ED Visits Among Patients with T2D

Javad Mohammad Alizadeh[1*], Gabriel Tajeu[2], Ilene L Hollin[1], Mukesh K Patel[1], Syeda Maryam Abidi[1], Rosemary Nagy[1], Huanmei Wu[1*]

[1*]Health Services Administration and Policy, Temple University, Philadelphia, PA, USA.
[2]Heersink School of Medicine, University of Alabama at Birmingham, Birmingham, AL, USA.

*Corresponding authors: javad.m.alizadeh@temple.edu; huanmei.wu@temple.edu;
Contributing authors: gtajeu@uab.edu; ilene.hollin@temple.edu; mukesh.kumar.patel@temple.edu; syeda.maryam.abidi@temple.edu; rosemary.nagy@temple.edu;

Abstract

Background: Emergency department (ED) visits among patients with type 2 diabetes (T2D) are frequent and costly. A substantial proportion of ED visits arising from T2D complications are potentially preventable. Yet most high-risk patients are identified only after an acute event has already occurred. This study develops and evaluates machine learning (ML) models to predict ED utilization among patients with T2D across multiple time horizons. This may enable proactive and timely clinical intervention.

Methods: Longitudinal electronic health record data from the HealthShare Exchange (HSX) were used. A sliding-window framework was applied across three prediction periods (0–30, 30–60, and 60–90 days). A total of 185 features spanning demographics, diagnoses, vital signs, and SDoH were used. These were combined into a unified patient-level feature set. Six ML models were developed and compared. These were Logistic Regression (LR), K-Nearest Neighbors (KNN), Random Forest (RF), and Extreme Gradient Boosting (XGB). The others were

Graph Neural Network (GNN) and Feature Tokenizer-Transformer (FTT). Models were evaluated using area under the receiver operating characteristic curve (AUC ROC), accuracy, precision, recall, and F1 score.
Results: A total of 51,603 patients were included in the analytic cohort across all prediction windows. The cohort had a median age of 68 years and was 50.8% female. It was racially diverse, with the largest groups being White (48.1%) and Black or African American (37.8%). Based on AUC ROC, XGB and FTT achieved the strongest discriminative performance among the six models. XGB reached 0.847, 0.868, and 0.857 across the 0–30, 30–60, and 60–90 day windows, respectively. FTT closely matched this (0.849, 0.864, 0.858), with AUC ROC differences between the two models of 0.004 or less. RF (0.840, 0.852, 0.842), LR (0.834, 0.850, 0.844), GNN (0.836, 0.851, 0.844), and KNN (0.747, 0.756, 0.746) followed. Once evaluated at a matched decision threshold, FTT achieved the highest recall and F1 score in two windows: 0–30 and 60–90 days. XGB achieved the highest recall and F1 score in the 30–60 day window. Healthcare utilization and vital sign variables were the most consistently important predictors. Decision curve analysis confirmed that FTT provided the highest net benefit in the 30–60 and 60–90 day windows. FTT and XGB performed comparably in the 0–30 day window. Fairness evaluation revealed consistent performance across sex subgroups. Discrimination and recall were lower for Hispanic or Latino patients in the 0–30 day window. Discrimination was similar across ethnicity subgroups in the later windows, and recall varied across racial and ethnic groups. Ablation analysis identified vital signs and diagnoses as the most contributory feature groups.
Conclusions: The ML models demonstrate strong and reliable predictive performance for ED utilization among patients with T2D. These findings suggest that these models can support proactive clinical decision-making and targeted interventions to reduce ED visits.

Keywords: Type 2 diabetes, Emergency department utilization, Machine learning, Predictive modeling, Electronic health records

1 Introduction

Diabetes mellitus is one of the most prevalent and costly chronic diseases in the United States. Diagnosed cases quadrupled from 6.5 million in 1990 to 26.8 million by 2018 [1]. As of 2021, an estimated 13.6% of U.S. adults have diabetes. Global expenditures among adults with diabetes reached approximately $760 billion in 2019 [2], accounting for roughly one in every four U.S. healthcare dollars [1]. Approximately 90–95% of diagnosed cases are T2D [3], which represents the predominant driver of this burden. This growing challenge underscores an urgent need to improve T2D management and prevent acute complications.

One consequence of suboptimal diabetes control is frequent ED utilization for acute issues. These include hyperglycemia, hypoglycemia, and cardiovascular complications [4, 5]. ED visit rates among adults with diabetes climbed by approximately 55% between 2008 and 2017 [1]. In 2020, an estimated 16.8 million ED visits involved

patients with diabetes. These emergency encounters carry substantial implications for costs and outcomes [6]. Nearly 38% of these visits resulted in hospitalization [6], reflecting the severity of acute events and gaps in outpatient care [1, 7]. Patients with diabetes experience disproportionately higher rates of ED use compared to those without diabetes [7]. This highlights the importance of early identification of high-risk patients to enable timely preventive intervention.

There is growing interest in leveraging ML to predict ED utilization in the diabetes population [7, 8]. Traditional risk stratification models have had only moderate success. Established risk factors such as older age, multiple comorbidities, and prior utilization yield limited predictive accuracy for individual patients [8–10]. ML techniques offer the ability to analyze high-dimensional electronic health record (EHR) data and capture complex nonlinear interactions to improve prediction [11]. This could enable early warning systems that support targeted interventions before acute events occur.

Emerging studies have begun applying ML specifically to ED prediction in diabetes-related populations. Park et al. developed a gradient-boosting model to predict ED visits among older adults with high comorbidity burden. This achieved a validation AUC ROC of approximately 0.69 [12]. Tan et al. trained multiple ML algorithms on a multi-institutional diabetes registry to predict high healthcare utilization. Prior ED visits, chronic kidney disease, and HbA1c were among the strongest predictors, with AUC ROC values exceeding 0.80 [8]. These studies demonstrate the feasibility of ML-based ED risk prediction but leave important gaps unaddressed.

Several limitations persist in the existing literature [13]. First, most models focus on general populations or treat diabetes as one of many comorbidities. Few study ED risk within a diabetes-specific cohort [12, 14]. Second, existing work typically evaluates a limited subset of algorithms. This constrains understanding of how different ML paradigms perform for this prediction task [15]. Third, few studies explicitly anchor prediction to a clinically meaningful encounter context, such as routine care, which has fundamental implications for deployability. Fourth, most models rely on a fixed 30-day outcome horizon, providing limited guidance for distinguishing imminent from medium-term risk [13]. Fifth, feature selection strategies often present a tradeoff between under-specification and over-complexity. Some models use too few variables, while others incorporate large numbers of diagnosis codes that increase sparsity and overfitting risk [7, 8, 16, 17]. Finally, few studies integrate heterogeneous data sources including SDoH, longitudinal utilization, vital signs, and diagnoses into a unified, interpretable framework [18, 19].

To our knowledge, no published study has developed a patient-level ML model to predict binary ED visit occurrence from routine outpatient encounters. No study has done so in a T2D cohort across multiple prospective time horizons. We address this gap by developing and evaluating six ML models: LR, KNN, RF, XGB, GNN, and FTT. These models predict ED utilization among patients with T2D. We use a sliding-window framework across three clinically relevant time horizons (0–30, 30–60, and 60–90 days). This framework is anchored to routine encounters as the prediction origin.

This study extends our prior encounter-level study using the same HSX dataset [20]. That study demonstrated the feasibility of ML-based ED prediction in a

T2D cohort. However, it used a single undifferentiated outcome window, an encounter-level unit of analysis, a narrower feature set, and a restricted model comparison. It also lacked fairness, ablation, and decision curve evaluation. The current study shifts to a patient-level prediction framework and expands the cohort to 51,603 patients. It also includes patients with T2D and comorbid hypertension, who were excluded from the prior study. It enriches the feature set to 185 variables spanning demographics, healthcare utilization, vital signs, SDoH, and aggregated diagnostic categories. It also evaluates a GNN, including a dedicated no-graph baseline comparison, alongside systematic post-hoc analyses.

We also develop and evaluate FTT, a Transformer-based tabular model, alongside the other architectures. We further assess fairness across demographic subgroups and apply interpretable ML techniques to identify key predictors. We also conduct ablation analyses to quantify feature group contributions and evaluate clinical utility via decision curve analysis.

2 Methodology

2.1 Experimental Design

As shown in Figure 1, ED visit prediction was approached as a binary classification task. Data preprocessing and cohort construction were followed by model training, internal validation, and a series of post hoc analyses. These included fairness analyses stratified by demographic subgroups and feature importance analyses to interpret key predictors. We also conducted ablation studies to assess the contribution of predefined feature groups. Decision curve analysis evaluated clinical utility by comparing net benefit against treat-all and treat-none reference strategies. Model performance was assessed using AUC ROC, accuracy, precision, recall, specificity, and F1 score on an independent held-out test set.

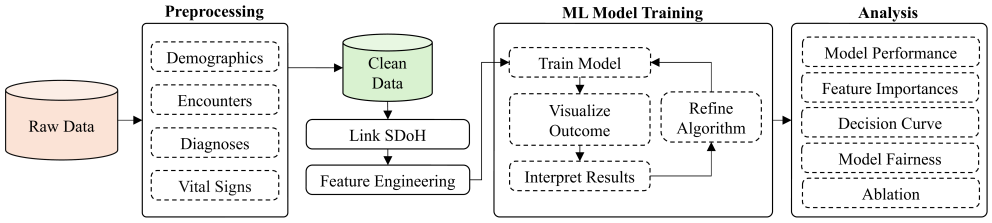


Fig. 1 System workflow overview.

2.2 Data Source and Cohort Construction

Longitudinal EHR data from the HealthShare Exchange (HSX) [21] were used to construct a patient-level analytic cohort. The HSX dataset contained electronic health records representing patients diagnosed with diabetes. These patients received care

within the Philadelphia metropolitan area and surrounding regions in Pennsylvania. The dataset captures only those who interacted with participating hospitals.

Eligible patients were adults ($\geq$ 18 years) with at least one documented T2D diagnosis who received care at a hospital participating in HSX. Patients with Type 1 diabetes, pediatric patients (age $<$ 18 years), and those with no recorded encounters were excluded. We also excluded patients recorded as 110 years of age or older, those with invalid ZIP codes, and those residing outside Pennsylvania. Additionally, patients whose admission history consisted entirely of non-emergency and non-routine encounters, single-encounter patients, and ED-only patients were excluded. The complete patient selection and filtering process is illustrated in Figure 2. After applying all exclusion criteria, 148,738 patients were removed, and the final analytic cohort comprised 51,603 patients. This provided sufficient longitudinal depth and clinical coverage to support sliding-window prediction of ED utilization across 0–30, 30–60, and 60–90 day windows.

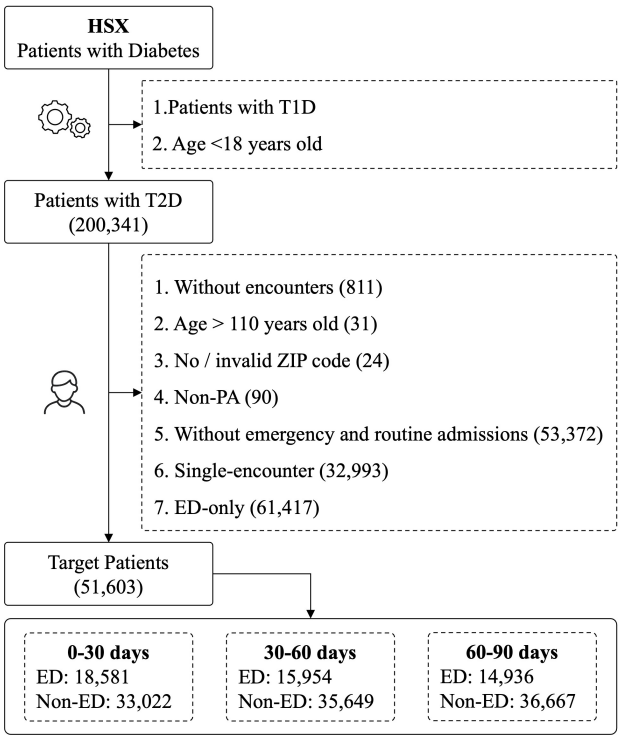


Fig. 2 Flow chart of cohort creation.

At the encounter level, duplicate records, same-day encounters, and records with missing duration or invalid geographic information were removed. Encounters were restricted to routine and emergency admission types. Routine encounters served as index prediction events, while emergency encounters were retained as outcome events.

A total of 581,002 encounters were retained after applying all encounter-level inclusion and exclusion criteria to the original 58,190,761 records. A total of 71 SDoH features were extracted from the American Community Survey (ACS) and incorporated into the dataset [22].

2.3 Data preprocessing and feature engineering

Data preprocessing and feature engineering were conducted using THReD (**T**ransforming E**H**Rs into AI-**Re**ady **D**atasets) [23]. All datasets were harmonized into a unified patient-level analytic table. Demographic and SDoH variables were retained as time-invariant attributes. Vital signs and utilization data were aggregated using summary statistics (minimum, maximum, mean, median). ICD-10 diagnosis codes were mapped to Clinical Classifications Software Refined (CCSR) [24] categories to reduce dimensionality and sparsity. Binary indicators were generated for clinically relevant CCSR categories. END002 and END003 (diabetes mellitus without and with complication) were excluded to avoid circularity. Related metabolic conditions such as obesity, dyslipidemia, and electrolyte disorders were retained.

Outcomes were defined as binary indicators of ED utilization within three forward-looking windows (0–30, 30–60, 60–90 days) from each routine index encounter. Separate binary flags were generated per window and aggregated to the patient level. A positive label was assigned if at least one qualifying ED visit occurred across all index encounters within the specified interval.

Baseline demographic characteristics were summarized and compared across ED utilization groups for each prediction window. Categorical variables were compared using chi-square tests, with Fisher's exact test applied when expected cell counts were less than five. Continuous variables were compared using the Mann–Whitney U test. All tests were two-sided with significance set at $p < 0.05$ [25]. Effect sizes were calculated using Cramér's V for categorical variables and rank-biserial correlation (r) [26] for continuous variables [27].

2.4 Model training and validation

Six ML models, including LR, KNN [28], RF [29], XGB [30], GNN [31], and FTT [32] were trained and evaluated across all prediction windows. All models used a common patient-level feature set. Each window dataset was split into training (70%) and testing (30%) sets using stratified sampling. Hyperparameters were tuned using three-fold stratified cross-validation within the training set, and the selected configuration was then evaluated with five-fold cross-validation [33]. For the GNN and FTT, an additional 20% of the training set was held out as a validation subset for early stopping.

Continuous features were standardized for LR, KNN, FTT, and GNN; tree-based models (RF, XGB) were trained on unscaled features. Missing values were imputed using the median or mode. Categorical variables were label-encoded for all models except the GNN. For the GNN, they were one-hot encoded so that Euclidean distance in the k-nearest neighbor graph would be meaningful.

For the GNN, a k-nearest neighbor graph was constructed from scaled feature vectors to capture relational structure between patients [34]. The graph was constructed

inductively, with edges built only among training patients. Validation and test patients were attached at inference time via directed edges to their nearest training neighbors. No edges were formed among validation or test patients themselves. This ensures no held-out patient's features or outcome could influence the learned node representations or another held-out patient's prediction.

Grid search with cross-validation was used to optimize each model independently per window, with AUC ROC as the primary objective [33]. For the GNN, the search jointly tuned graph and model hyperparameters (K, hidden dimension, and learning rate) rather than K alone. Graph density interacts with model capacity. AUC differences across the top GNN configurations were within one fold-to-fold standard error of each other. The simplest statistically equivalent configuration was therefore selected (Breiman's one-standard-error rule) rather than the single highest-scoring one. For all other models, the configuration with the highest mean cross-validated AUC ROC was selected.

To keep the model comparison consistent, class weighting was not applied in the final models. XGB used scale_pos_weight = 1, and LR, KNN, FTT, and GNN were trained with unweighted losses. RF tuned class_weight over balanced and none but selected the unweighted setting in all windows [30, 33, 35]. For comparability, threshold-matched metrics were reported for all models at a common decision threshold of 0.5. For the GNN, a threshold was additionally selected on the validation set.

The GNN incorporated early stopping based on validation AUC. Final models were trained on the full training set and evaluated on the held-out test set. For the GNN, test-set probabilities were additionally calibrated using isotonic regression fit on the validation set prior to decision curve analysis. Calibration was assessed via Brier score. AUC ROC values reported for the GNN are based on these calibrated probabilities, whereas its threshold-matched metrics are based on uncalibrated probabilities. Performance metrics included accuracy, precision, recall, F1 score, sensitivity, and specificity.

Model stability was assessed by bootstrapping. For LR, KNN, RF, XGB, and FTT, each model was refit on 100 bootstrap resamples of the training set. Each refit was evaluated on the fixed held-out test set. The 95% intervals for AUC ROC and accuracy were taken as the 2.5th and 97.5th percentiles of the resulting distribution. For the GNN, intervals were instead obtained by resampling the held-out test set with replacement (1,000 iterations), using the same percentile-based cutoffs [36].

An architecture-matched multilayer perceptron (MLP) with no message passing served as a no-graph baseline. This isolated the graph structure's contribution from model capacity alone. Graph quality was additionally assessed via edge homophily. This is the proportion of graph edges connecting patients with the same outcome label, relative to the base rate expected by chance.

FTT tokenized each feature and applied Transformer layers with self-attention [32]. A classification (CLS) token aggregated feature interactions for prediction. Hyperparameters (embedding size, block count, learning rate) were tuned via 3-fold cross-validation. The best configuration was evaluated with 5-fold cross-validation, matching other models' protocol.

2.5 Feature importance analysis

Feature importance was derived using model-specific approaches. For LR, absolute standardized coefficients were used [33]. For RF and XGB, impurity-based importance scores were extracted [29, 30]. For the GNN, permutation-based importance measured AUC ROC decrease after permuting raw test features [37]. After each permutation, the GNN's test-set attachment edges were rebuilt so that permuted values could not leak back through neighboring nodes. For KNN and FTT, permutation-based importance likewise measured the decrease in test-set AUC ROC. For all permutation-based approaches, each feature was permuted 10 times independently, and the mean and standard deviation of importance are reported.

2.6 Decision curve analysis

Net benefit was calculated for each model and prediction window across threshold probabilities ranging from 0.01 to 0.99. True positives were weighed against false positives relative to treat-all and treat-none strategies [38, 39]. This analysis identified threshold ranges where model-guided decisions offer greater clinical value than non-informative baselines.

2.7 Fairness assessment

Fairness was evaluated by stratifying model performance across sex, race, and ethnicity subgroups using the held-out test set. This was applied independently for each prediction window. Sex subgroups included male and female. Racial subgroups included American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, and White. Winsorization at the 1st and 99th percentiles also applied to the encoded race variable in all models except the GNN. As a result, American Indian or Alaska Native patients (n=42–49 per test set) were grouped with Asian patients. The Asian subgroup in the FTT fairness figures therefore includes them. Ethnicity subgroups included Hispanic or Latino and Non-Hispanic or Latino. Patients with missing demographic data were retained as an "Unknown" category.

Metrics included accuracy, precision, recall (sensitivity), F1 score, and AUC ROC, with sample size recorded per subgroup. AUC ROC was treated as undefined for subgroups in which only a single outcome class was observed in the test set. Demographic attributes were included as model features but were not used to define outcomes [40, 41].

2.8 Ablation analysis

Models were retrained after systematically removing predefined feature groups. These were demographics (4 features), diagnoses (59), vital signs (45), SDoH (71), and engineered utilization features (6). Engineered features comprised prior non-ED visit count and inter-encounter durations. AUC ROC on the held-out test set was compared against the full-feature baseline to quantify each group's predictive contribution [42]. For the GNN, feature groups were removed from the raw feature set [43]. The preprocessing pipeline was then refit, the k-nearest neighbor graph rebuilt, and the model

retrained. Removed features were thus excluded entirely rather than set to zero after standardization.

For the GNN and FTT, ΔAUC was averaged across three retraining seeds. ΔAUC was expressed as both an absolute difference and a relative percentage loss. For the GNN, 95% intervals were derived from variability across retraining seeds (mean ± 1.96 SD). For the other models, bootstrapped 95% confidence intervals for ΔAUC were derived from 1,000 test-set resamples [36]. Bonferroni correction was applied across the number of active feature groups tested per model [44].

3 Results

3.1 Baseline characteristics

Baseline demographic characteristics stratified by ED utilization status are summarized in Table 1 for the 0–30 day prediction window. Results for the 30–60 and 60–90 day prediction windows are available in the appendix. Patients who utilized the ED were older than those without ED utilization, with a median age of 69 years (IQR: 59–79). The proportion of female patients was similar between groups (50.9% vs. 50.7%). Male patients were slightly less represented among ED utilizers (43.1% vs. 44.9%). Patients identifying as Black or African American (45.3%) were more frequently represented in the ED cohorts. Ethnicity did not differ significantly between groups (Not Hispanic or Latino: 91.4% vs. 91.6%; p = 0.634).

Table 1 Baseline characteristics of the patients with T2D across 0–30 day prediction window.

Characteristics	**Total**	**ED visit, no**	**ED visit, yes**	**P-value**
0–30 days	**n=51,603**	**n=33,022**	**n=18,581**	
Age	68 [59, 78]	68 [58, 77]	69 [59, 79]	<0.001
Sex, n (%)				<0.001
Female	26,201 (50.8%)	16,738 (50.7%)	9,463 (50.9%)	
Male	22,847 (44.3%)	14,833 (44.9%)	8,014 (43.1%)	
Unknown	2,555 (5.0%)	1,451 (4.4%)	1,104 (5.9%)	
Race, n (%)				<0.001
American Indian or Alaska Native	154 (0.3%)	104 (0.3%)	50 (0.3%)	
Asian	1,604 (3.1%)	1,208 (3.7%)	396 (2.1%)	
Black or African American	19,507 (37.8%)	11,081 (33.6%)	8,426 (45.3%)	
Native Hawaiian or Other Pacific Islander	47 (0.1%)	26 (0.1%)	21 (0.1%)	
White	24,827 (48.1%)	16,953 (51.3%)	7,874 (42.4%)	
Unknown	5,464 (10.6%)	3,650 (11.1%)	1,814 (9.8%)	
Ethnicity, n (%)				0.634
Not Hispanic or Latino	47,232 (91.5%)	30,252 (91.6%)	16,980 (91.4%)	
Hispanic or Latino	1,261 (2.4%)	804 (2.4%)	457 (2.5%)	
Unknown	3,110 (6.0%)	1,966 (6.0%)	1,144 (6.2%)	

Across all three prediction windows, age, sex, and race differed significantly between groups, but effect sizes were negligible to small. Effect sizes were |r| = 0.026–0.052 for age and Cramér's V = 0.036–0.039 for sex. Ethnicity did not differ significantly (p = 0.415–0.634; Cramér's V = 0.004–0.006). Race was the only variable

with a consistent small effect (Cramér's V = 0.120–0.146). Black or African American patients were disproportionately represented among ED utilizers (45.3–48.7%) compared to non-utilizers (33.4–33.6%). This disparity widened with longer prediction windows.

3.2 Model performance

As presented in Table 2, FTT led on accuracy (0.783) in the 0–30 day window, narrowly ahead of XGB (0.781). FTT achieved the highest recall (0.668) and F1 score (0.689) among the threshold-matched models; FTT's specificity was 0.847. KNN achieved the highest precision (0.769) and specificity (0.956). It had the lowest recall (0.261) and F1 score (0.390). LR, RF, and GNN occupied a middle ground with more balanced sensitivity-specificity tradeoffs. This pattern of KNN underperforming on recall while maintaining high specificity persisted across the 30–60 and 60–90 day windows. Recall was 0.214 and 0.187, and F1 was 0.334 and 0.297, respectively.

Table 2 Performance of the models on the held-out test set (30% of the dataset) following hyperparameter tuning via cross-validation.

0–30 days	**Accuracy**	**Precision**	**Recall**	**Specificity**	**F1 score**
XGB	0.7810	0.7441	0.5972	0.8845	0.6626
GNN	0.7779	0.7404	0.5901	0.8836	0.6567
FTT	**0.7828**	0.7112	**0.6679**	0.8474	**0.6889**
LR	0.7751	0.7526	0.5594	0.8965	0.6418
RF	0.7774	0.7357	0.5958	0.8796	0.6584
KNN	0.7058	**0.7689**	0.2614	**0.9558**	0.3901
30–60 days					
XGB	**0.8125**	0.7415	**0.6041**	0.9058	**0.6657**
GNN	0.8014	0.7375	0.5554	0.9115	0.6336
FTT	0.8091	**0.7572**	0.5631	0.9192	0.6459
LR	0.8020	0.7474	0.5428	0.9179	0.6289
RF	0.7968	0.7412	0.5265	0.9177	0.6157
KNN	0.7354	0.7533	0.2144	**0.9686**	0.3338
60–90 days					
XGB	0.8109	0.7231	0.5619	0.9124	0.6324
GNN	0.8043	0.7108	0.5459	0.9095	0.6175
FTT	**0.8117**	0.7247	**0.5635**	0.9128	**0.6340**
LR	0.8035	**0.7301**	0.5095	0.9233	0.6002
RF	0.7984	0.7225	0.4927	0.9229	0.5859
KNN	0.7428	0.7116	0.1872	**0.9691**	0.2965

The ROC curves for the 0–30 day prediction window are presented in Figure 3. Results for the 30–60 and 60–90 day windows followed comparable patterns and are available in the appendix. FTT and XGB had the highest AUCs in all three windows.

Their AUCs were 0.849 and 0.847 (0–30 days), 0.864 and 0.868 (30–60 days), and 0.858 and 0.857 (60–90 days). These differences were small and were not formally tested. RF, LR, and GNN formed a closely clustered group behind FTT and XGB. Their curves overlapped substantially, tracking well above the diagonal. Overall model performance improved in the 30–60 day window. XGB achieved its highest AUC there, and the separation from the remaining models was most pronounced. KNN consistently diverged from all other models across all three horizons (AUCs: 0.747, 0.756, 0.746). Its curve tracked markedly closer to the random chance line, confirming limited discriminative ability regardless of prediction horizon.

To isolate the specific contribution of the graph structure, GNN performance was compared against an architecture-matched MLP baseline with no message passing. The GNN did not outperform this no-graph baseline in any window. AUC differences were −0.0025, −0.0028, and −0.0033 for the 0–30, 30–60, and 60–90 day windows, respectively. This indicates that the graph topology contributed little beyond what an equivalent-capacity MLP achieved on the same features.

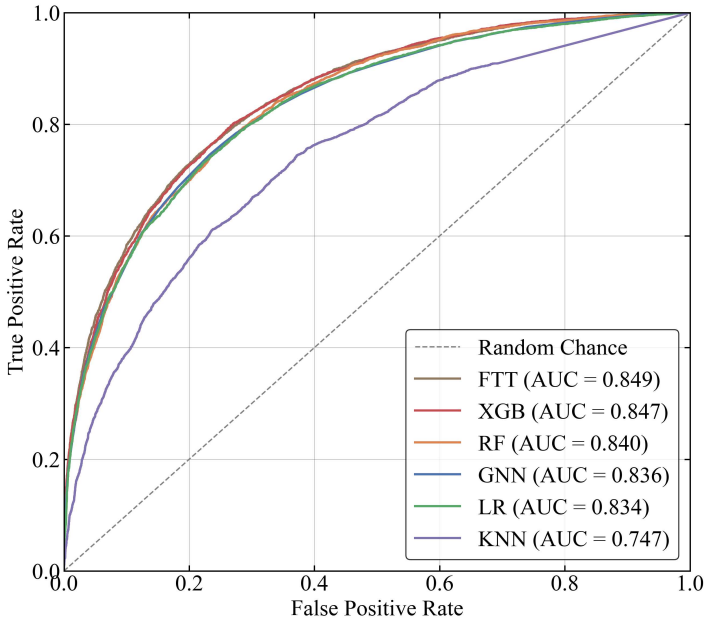


Fig. 3 AUC ROC curves of all six models for the 0–30 day window.

Bootstrap stability analysis of training resamples confirmed consistent performance of LR, KNN, RF, XGB, and FTT across all prediction windows (Table 3). XGB maintained the highest mean bootstrap AUC ROC across all three windows. Its narrow bootstrap ranges indicated stable generalization rather than sensitivity to specific training samples. FTT ranked second across all three windows, with similarly narrow bootstrap ranges (0.843, 0.858, 0.852). KNN showed slightly wider intervals. Except for

Table 3 Bootstrap AUC ROC with 95% intervals for each model. For LR, KNN, RF, XGB, and FTT, values are means across models refit on 100 bootstrap resamples of the training set and can differ from the test-set AUC ROC values reported in the text. For the GNN, values are means of the calibrated test-set AUC ROC across 1,000 test-set resamples, with 95% intervals from the same resamples.

AUC ROC	XGB	GNN	FTT	LR	RF	KNN
0–30	**0.8447** **[0.843–0.847]**	0.8357 [0.829–0.842]	0.8427 [0.840–0.845]	0.8322 [0.831–0.833]	0.8359 [0.834–0.837]	0.7179 [0.713–0.723]
30–60	**0.8628** **[0.860–0.865]**	0.8509 [0.845–0.857]	0.8575 [0.855–0.860]	0.8475 [0.846–0.849]	0.8484 [0.847–0.850]	0.7247 [0.720–0.732]
60–90	**0.8545** **[0.853–0.856]**	0.8441 [0.837–0.851]	0.8523 [0.849–0.855]	0.8416 [0.840–0.843]	0.8400 [0.838–0.842]	0.7156 [0.711–0.720]

the GNN, these bootstrap intervals reflect model stability across training resamples. They are not classical confidence intervals for the point estimate. The GNN interval was derived from test-set resampling, so its width is not directly comparable with those of the other models.

3.3 Feature importances

The cumulative normalized importance scores across all six models for the 0–30 day prediction window are presented in Figure 4. Results for the 30–60 and 60–90 day windows followed comparable patterns and are available in the appendix. Non_ED_count, representing prior routine care engagement, was the most consistently important feature across models. It ranked among the top 10 in five of the six models (all except LR) in the 0–30 and 60–90 day windows. It ranked there in all six models in the 30–60 day window. It accumulated the highest cumulative score by a substantial margin. Weight_max and Weight_min ranked prominently in RF, GNN, and FTT, and Height_min in GNN and FTT. Blood pressure features (DBP_max and SBP_min) ranked prominently in XGB and RF.

Several clinical coded variables contributed consistently in XGB and KNN across all windows. These included END011 (fluid and electrolyte disorders), CIR019 (heart failure), and BLD003 (aplastic anemia). SYM010 (nervous system signs and symptoms) ranked highly in XGB in every window. It ranked highly in KNN and FTT only in the 60–90 day window. Inter-encounter duration features were elevated in XGB, RF, and FTT. Socioeconomic variables were prominent in LR alongside anthropometric measures across all windows. GEN002 (acute and unspecified renal failure) emerged as a cross-model consensus feature in the shorter prediction horizons.

3.4 Decision curve analysis

The decision curve analysis for the 0–30 day prediction window is presented in Figure 5. Results for the 30–60 and 60–90 day windows followed comparable patterns and are available in the appendix. In the 30–60 and 60–90 day windows, FTT provided the highest net benefit at nearly all evaluated threshold probabilities (0.01–0.99). In the 0–30 day window, FTT and XGB performed comparably, together providing the highest

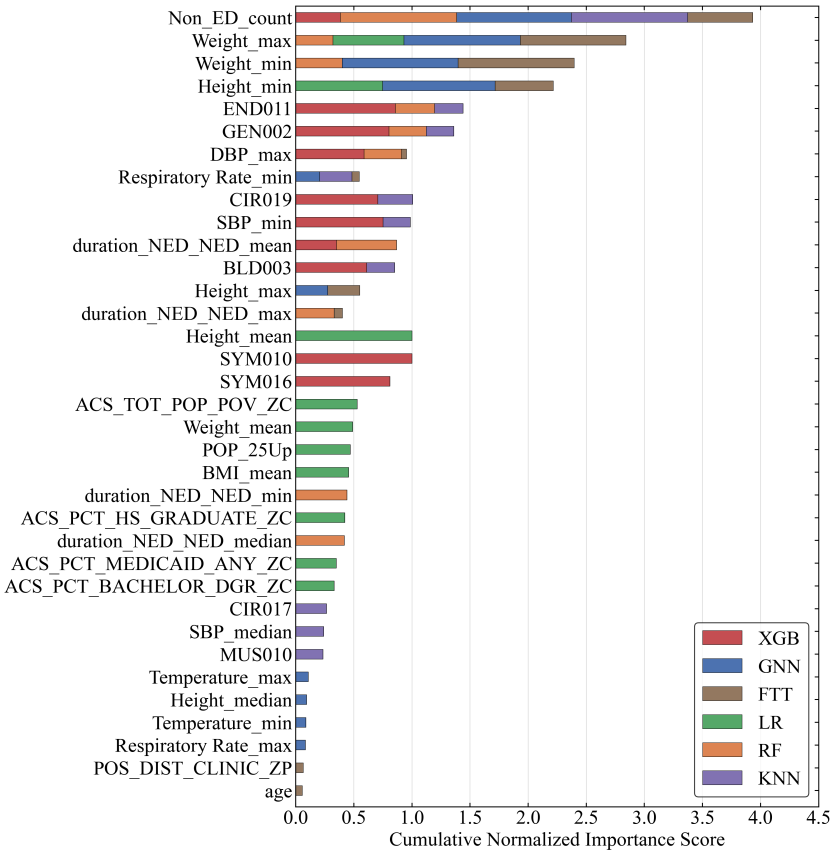


Fig. 4 Cumulative normalized feature importance scores across all six models for the 0–30 day window.

net benefit across most thresholds. FTT outperformed both the treat-all and treat-none strategies. GNN, LR, and RF tracked closely behind FTT across most thresholds, with LR representing a viable alternative where model interpretability is prioritized.

In the 0–30 day window, RF fell below the other non-KNN models at higher thresholds. XGB, despite its high AUC ROC, provided the lowest net benefit among the five non-KNN models at 30–60 and 60–90 days. It declined more steeply at higher thresholds, crossing below the treat-none line beyond a threshold of ∼0.75 in the 60–90 day window. KNN consistently underperformed across all windows, falling below the other models throughout most of the threshold range. It provided negligible net benefit at higher thresholds, corroborating its poor discriminative performance.

3.5 Model fairness evaluation

The FTT fairness evaluation results across sex, race, and ethnicity for the 0–30 day prediction window are presented in Figure 6. Results for the 30–60 and 60–90 day

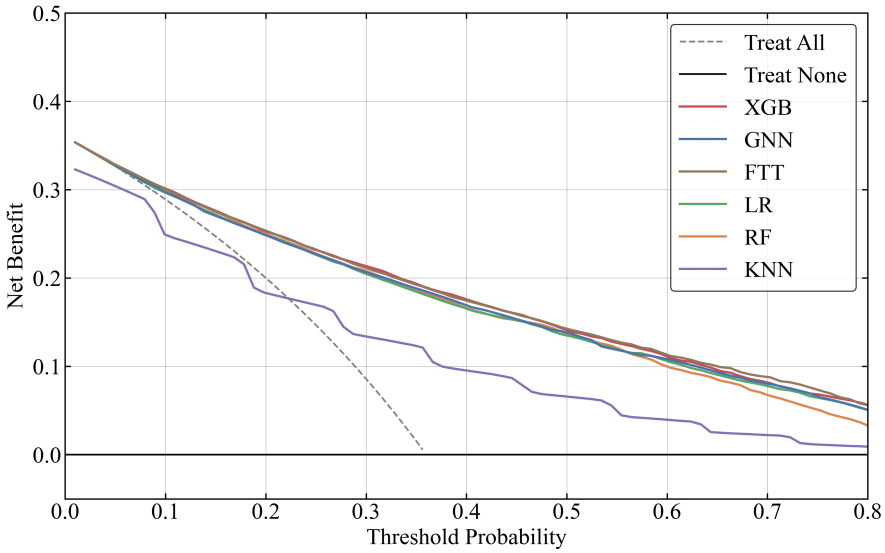


Fig. 5 Decision curve analyses for all six models across the 0–30 day window.

windows are provided in the appendix. Fairness evaluations for the remaining models (XGB, RF, LR, GNN, and KNN) are available upon request.

In the 0–30 day window, FTT performed consistently between female and male patients across all metrics, indicating no meaningful sex-based bias. Across racial groups, discrimination (AUC ROC) was similar between the two largest groups. However, recall was ∼14 points higher for Black or African American than White patients. Across ethnicity, discrimination (AUC ROC 0.818 vs. 0.851) was lower for Hispanic or Latino than for Not Hispanic or Latino patients. Recall was also lower (0.559 vs. 0.674). However, the Hispanic subgroup's smaller sample size (n=383) widens its metric variability (AUC ROC 95% CI: 0.775–0.857). This interval overlaps that of Not Hispanic or Latino patients.

This pattern was broadly consistent across the other five models. They showed similar performance between sexes and lower discrimination and recall for Hispanic or Latino patients in the 0–30 day window. They also showed higher recall for Black or African American than White patients. KNN, however, showed uniformly lower recall across all demographic subgroups. This reflects its overall tendency to underpredict the positive class regardless of group membership.

3.6 Ablation analysis

The ablation analysis results for the 0–30 day prediction window are presented in Figure 7. Results for the 30–60 and 60–90 day windows followed comparable patterns and are available in the appendix. A positive ΔAUC indicates that removing the feature group reduced model performance, reflecting its contribution to predictive ability.

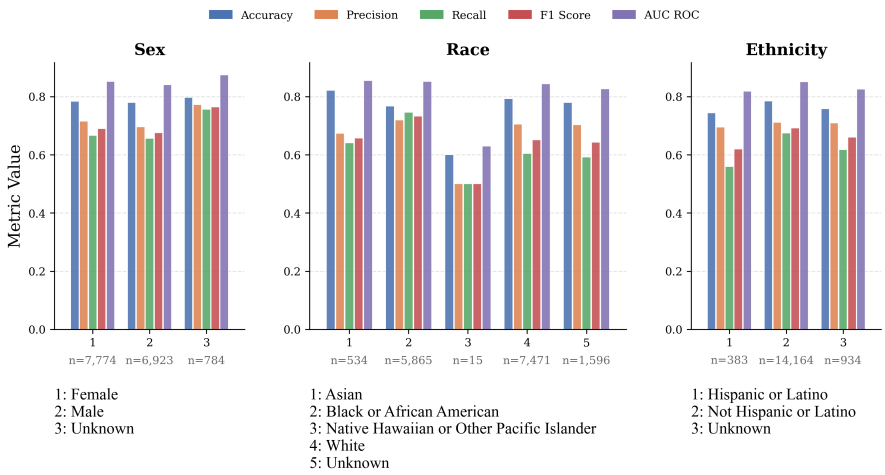


Fig. 6 Fairness evaluation of the FTT across sex, race, and ethnicity subgroups for the 0–30 day window.

Vital signs were the most important feature category in 15 of the 18 model–window combinations. They produced the largest and most consistent ΔAUC values upon removal. The exceptions were LR and KNN in the 30–60 day window, where diagnoses contributed more. KNN in the 60–90 day window was also an exception, where engineered features contributed more. RF showed the greatest sensitivity to this group in the 0–30 day window and XGB in the 30–60 and 60–90 day windows. GNN's ΔAUC declined slightly from 0–30 to 30–60 and held steady into 60–90.

Diagnoses contributed meaningfully across nearly all models and windows (except KNN in the 60–90 day window). LR showed the highest ΔAUC for this category in every window. Confidence intervals remained relatively tight, suggesting reliable and consistent contribution. FTT's Diagnoses ΔAUC was 0.019, with a tight confidence interval. Engineered features showed moderate positive contributions across all models and windows. The tree-based models (XGB and RF) benefited most in every window. This reflects their capacity to leverage temporal inter-encounter signals.

SDoH showed near-zero or slightly negative ΔAUC across most models and all windows, with RF consistently falling just below zero. This indicates that removing SDoH marginally improved its performance. KNN showed a notably negative ΔAUC for SDoH across all windows. This suggests that these features introduced noise rather than signal for this model.

Demographics contributed minimally across most models. XGB, however, showed a small but consistent positive ΔAUC for this category across all three windows, while KNN showed near-zero, non-significant values. FTT's Demographics ΔAUC was negligible and non-significant (0.0003, p=0.36).

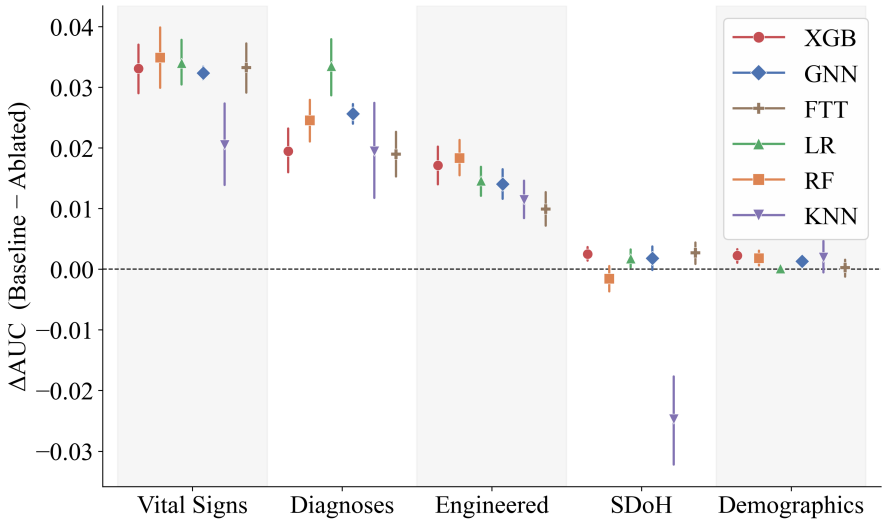


Fig. 7 Ablation analysis results showing ΔAUC for each feature group across all six models for the 0–30 day window.

4 Discussion

Patients with T2D face a high and rising burden of ED use. Yet few studies have predicted these visits within a strictly T2D cohort across multiple time horizons [1, 7, 13]. Among 51,603 patients with T2D, ML models predicted ED visits within 0–30, 30–60, and 60–90 day windows with strong discrimination. XGB and FTT performed best (AUC ROC up to 0.868). In the 0–30 day window, XGB favored precision, whereas FTT favored recall. FTT achieved the highest recall (0.668) and F1 score (0.689). This contrast was specific to that window.

Non-ED encounter count, vital signs, and diagnoses were among the most important predictors, whereas ZCTA-level SDoH features added little to discrimination. FTT performance was broadly equitable across sex. Discrimination and recall were lower for Hispanic or Latino patients in the 0–30 day window. Discrimination was similar across ethnicity in the later windows. Estimates for the smallest racial subgroups were too unstable to support conclusions about disparities. Finally, FTT provided the highest net benefit in the 30–60 and 60–90 day windows. It performed comparably to XGB in the 0–30 day window. This shows that the model with the best discrimination does not necessarily offer the greatest clinical utility.

Our preliminary study, available as a preprint, used the same HSX dataset. It showed that ML could predict ED visits in this population, with top models reaching an AUC of 0.82 [20]. However, the earlier study was conducted at the encounter level, where each of the 703,065 individual visits constituted a separate prediction instance. This design conflates patient-level risk with visit frequency. It produces performance estimates that reflect encounter-level discrimination, rather than the clinically meaningful question of which patients are at risk.

The current study is conducted at the patient level. Each of the 51,603 patients is represented as a single observation per prediction window. The current XGB model achieved AUC values of 0.847, 0.868, and 0.857 across the 0–30, 30–60, and 60–90 day windows, respectively. However, differences in unit of analysis, cohort, and feature set limit direct comparison. The main contribution of this study is therefore not the gain in AUC, but a more clinically meaningful framework. This framework identifies which patients are at risk, anchored at routine encounters, across multiple prediction windows.

Furthermore, XGB and FTT achieved the strongest discrimination. XGB reached an AUC ROC of 0.847 in the 0–30 day window and peaked at 0.868 in the 30–60 day window. FTT's AUC values (0.849, 0.864, and 0.858) were comparable to XGB's, and the differences were not formally tested. XGB's high AUC ROC reflects a consistent ability to rank patients by risk across all decision thresholds. This makes it particularly well suited for risk stratification tasks. Its capacity to capture nonlinear feature interactions contributed to its robust and stable performance across horizons, even without class reweighting [30].

RF achieved closely competitive AUC values across all three windows. This is consistent with related work where ensemble methods have frequently matched or approached gradient boosting approaches. These comparisons involved predicting ED visits and adverse outcomes in chronic disease populations [12]. Importantly, both XGB and RF were among the top performers in the prior study as well. XGB's strong performance is more clearly and consistently demonstrated here. This owes to the more rigorous evaluation framework, including held-out test sets, bootstrapped confidence intervals, and multi-horizon assessment. These results reinforce the growing consensus in the clinical ML literature that tree-based ensemble methods remain highly competitive for tabular EHR data. FTT nonetheless performed comparably to XGB in this study. This consensus holds particularly in settings with class imbalance and high feature dimensionality [8, 15].

A notable finding is the divergence between XGB's precision-oriented performance and FTT's recall-oriented profile in the 0–30 day window. FTT achieved the highest recall (0.668) and F1 score (0.689) there. This divergence was specific to that window. In the 30–60 day window, XGB had higher recall (0.604 vs. 0.563) and FTT higher precision (0.757 vs. 0.742). The two models performed nearly identically in the 60–90 day window. At the shortest horizon, FTT may therefore better suit settings where missing a high-risk patient carries greater consequence than a false positive. This is particularly relevant in T2D management. Missed high-risk patients may proceed to costly and avoidable ED visits, hospitalizations, or acute metabolic crises [6, 7]. Depending on the clinical and operational context, practitioners may therefore prefer FTT or a model ensemble that explicitly balances sensitivity and specificity [13, 15].

GNN and FTT required more compute and offered less direct interpretability than tree-based models, which may limit scalability in routine clinical deployment. Given its comparable discrimination to XGB, FTT may not justify its added complexity for some deployments.

Bootstrap stability analysis confirmed that performance estimates were consistent across training resamples for LR, KNN, RF, XGB, and FTT in all windows. Intervals

were narrow for XGB, RF, LR, and FTT, and slightly wider for KNN. The GNN intervals came from test-set resampling and are not directly comparable.

KNN exhibited high specificity (0.956) but extremely low recall (0.261) and the poorest F1 score (0.390). This reflects a severe bias toward the negative class that renders it clinically unsuitable as a standalone screening tool. The pattern is consistent with its underperformance in the prior study (AUC 0.72) and in related clinical ML literature [8, 15].

The GNN's graph-structured message passing did not outperform an architecture-matched MLP baseline with no message passing in any prediction window. This was not attributable to an uninformative graph. Edge homophily exceeded the chance level by 18–21% across all three windows. This confirms that the k-nearest-neighbor graph does connect patients with concordant outcomes more often than expected. The GNN–MLP gap therefore more likely reflects limitations of the two-layer GraphSAGE aggregation used here than an absence of graph-structured signal. This motivates future work on alternative aggregation functions or clinically informed edge definitions, rather than abandoning the graph-based approach outright.

Feature importance analyses provided interpretable and clinically grounded insights consistent across all three prediction windows. Non_ED_count emerged as the most consistently important predictor across models. It reflects variation in care engagement intensity among patients who already have some healthcare contact. It does not distinguish fully engaged from entirely disengaged patients. This is consistent with the broader literature indicating that care continuity is protective against acute utilization in chronic disease populations [7, 9]. This aligns with findings from the prior study, where visitation gap and gap difference ranked among the top five predictors. It reinforces the cross-study stability of healthcare utilization patterns as a dominant signal [20].

Vital sign features, particularly blood pressure variables (SBP_min, DBP_max), were dominant predictors in tree-based models across all windows. This underscores the clinical relevance of hemodynamic status in T2D risk stratification [4, 5]. Symptom-related CCSR categories (SYM010, SYM016) and endocrine variables (END011) were consistently prominent in XGB and, to a lesser extent, KNN. They rarely ranked highly in the other models. This reflects these models' capacity to capture disease-specific signals less salient in other model types.

SDoH variables were prominent in LR's coefficient structure but contributed minimally to AUC in the ablation analysis. This held across all models and prediction windows, with RF producing slightly negative ΔAUC values upon SDoH removal. FTT's SDoH ΔAUC was small but positive and significant in the 0–30 and 60–90 day windows (0.003 and 0.002). It was non-significant in the 30–60 day window (0.0005).

The minimal AUC contribution of SDoH features is best interpreted as a data linkage limitation rather than clinical irrelevance of social determinants [18, 45]. SDoH variables were linked at the ZCTA level using the ACS data and therefore reflect community-level rather than individual-level characteristics. All patients residing in the same ZIP code were assigned identical values for all 71 SDoH features. These variables therefore cannot distinguish between patients within the same community and capture only between-area variation. This substantially attenuates individual-level

socioeconomic variation and limits the signal available for patient-level prediction. It likely explains the near-zero or negative ΔAUC values observed upon SDoH removal. Future work should incorporate individual-level SDoH measures to more rigorously evaluate their independent predictive contribution [45, 46].

Notably, XGB showed a small but consistent positive ΔAUC for demographics across all three windows. Removing SDoH from KNN actually improved performance (ΔAUC −0.016 to −0.025), consistent with KNN's sensitivity to irrelevant dimensions in high-dimensional spaces [47]. Patients from the same ZIP code share identical values on all 71 SDoH features. These features therefore draw patients from the same area closer together in the distance calculation regardless of their clinical similarity. Together, these findings underscore the importance of model-specific rather than uniform feature interpretation [37].

The fairness evaluation found broadly equitable performance across sex groups for FTT across all three prediction windows. It also found broadly similar discrimination across ethnicity groups in the 30–60 and 60–90 day windows. This is an encouraging finding given growing concerns about algorithmic bias in clinical ML tools [40, 41]. Performance was consistent between male and female patients. Discrimination was similar between Hispanic/Latino and non-Hispanic/Latino patients in the 30–60 and 60–90 day windows. It was lower for Hispanic/Latino patients in the 0–30 day window (AUC ROC 0.818 vs. 0.851, with overlapping confidence intervals). Recall for Hispanic/Latino patients was lower in the 0–30 and 60–90 day windows. However, estimates for the smallest racial subgroups were unstable. This was particularly true for Native Hawaiian or Other Pacific Islander patients (n=15 in the 0–30 day test set). Limited sample sizes preclude reliable performance estimation and conclusions about disparities [48]. This highlights the need for intentional data collection and oversampling strategies for underrepresented groups prior to real-world deployment [41, 48].

Decision curve analysis confirmed that FTT provides the highest net benefit in the 30–60 and 60–90 day windows. This held across a clinically meaningful range of threshold probabilities. FTT performs comparably to XGB in the 0–30 day window. It consistently outperformed both treat-all and treat-none reference strategies across all windows. GNN, LR, and RF closely tracked FTT across most thresholds, with LR representing a practical alternative where interpretability is prioritized [49]. XGB, despite its high AUC ROC, provided comparatively lower net benefit than the other tree-based and deep learning models. This was particularly evident in the 30–60 and 60–90 day windows. This underscores that discrimination and clinical utility do not necessarily favor the same model. KNN's poor decision curve performance across all windows reinforces that AUC alone is insufficient for assessing clinical usefulness [39, 50]. Net benefit analysis should therefore be a standard component of clinical ML evaluation [50].

Several limitations should be acknowledged. The HSX dataset is derived from a single metropolitan region, which may limit generalizability to other geographic settings or healthcare systems. By restricting index encounters to routine admission types, this study does not model ED risk prediction from post-discharge or transitional care contexts. This restriction was intentional and methodologically motivated.

The sliding-window approach was constructed retrospectively, and performance in prospective or real-time clinical environments may differ. SDoH features were linked at the ZCTA level using ACS data, which may introduce ecological bias and underestimate individual-level socioeconomic variation. Fairness subgroup comparisons were not corrected for multiplicity. Results should therefore be interpreted with appropriate caution. Finally, key clinical variables, including medications, HbA1c trajectories, and self-reported symptoms, were not available in the EHR data. This potentially limits predictive ceiling performance.

5 Conclusion

This study demonstrates that ML models can reliably predict ED utilization in patients with T2D using real-world EHR data. Among all models evaluated, XGB and FTT achieved the strongest discriminative performance, with results closely matched across the three prediction windows. FTT offered the highest sensitivity in the 0–30 and 60–90 day windows and XGB in the 30–60 day window. This underscores the value of aligning model selection with clinical objectives. Vital signs and prior healthcare utilization patterns were the most consistently important predictors, providing a clinically interpretable basis for risk stratification. These findings suggest that routine encounter-anchored ML prediction frameworks hold meaningful promise for supporting proactive T2D care management. They enable clinicians and health systems to identify high-risk patients before ED visits occur and direct preventive interventions accordingly.

Acknowledgements. We thank the Robert Wood Johnson Foundation (RWJF), AcademyHealth, and HealthShare Exchange (HSX) for providing the data.

Funding. This project was supported in part by the Robert Wood Johnson Foundation's Health Data for Action program, managed by AcademyHealth (Grant# 78961).

Competing interests. The authors declare no competing interests.

Human ethics and consent to participate. This research did not meet the definition of human subjects research, so Institutional Review Board approval was not required.

Clinical trial number. Not applicable.

Consent for publication. Not applicable.

Data availability. The clinical data used in this study were fully de-identified electronic health record data obtained under a data use agreement with HSX. Due to the terms of the data use agreement, the data are not publicly available but can be made available by the corresponding authors upon request.

Materials availability. Not applicable.

Code availability. The code will be made publicly available upon acceptance of this manuscript.

AI-use disclosure. During manuscript preparation, the authors used an AI writing assistance tool to improve sentence structure and clarity and to correct grammatical and typographical errors. The authors reviewed, edited, and take full responsibility for all content in this manuscript.

Appendix

Baseline demographic characteristics stratified by ED utilization status are summarized in Table 4 for 30–60 and 60–90 day prediction windows.

Table 4 Baseline characteristics of the patients with T2D across 30–60 and 60–90 day prediction windows.

Characteristics	**Total**	**ED visit, no**	**ED visit, yes**	**P-value**
30–60 days	**n=51,603**	**n=35,649**	**n=15,954**	
Age	68 [59, 78]	68 [58, 77]	69 [59, 79]	<0.001
Sex, n (%)				<0.001
Female	26,201 (50.8%)	17,895 (50.2%)	8,306 (52.1%)	
Male	22,847 (44.3%)	16,125 (45.2%)	6,722 (42.1%)	
Unknown	2,555 (5.0%)	1,629 (4.6%)	926 (5.8%)	
Race, n (%)				<0.001
American Indian or Alaska Native	154 (0.3%)	108 (0.3%)	46 (0.3%)	
Asian	1,604 (3.1%)	1,292 (3.6%)	312 (2.0%)	
Black or African American	19,507 (37.8%)	11,906 (33.4%)	7,601 (47.6%)	
Native Hawaiian or Other Pacific Islander	47 (0.1%)	31 (0.1%)	16 (0.1%)	
White	24,827 (48.1%)	18,367 (51.5%)	6,460 (40.5%)	
Unknown	5,464 (10.6%)	3,945 (11.1%)	1,519 (9.5%)	
Ethnicity, n (%)				0.618
Not Hispanic or Latino	47,232 (91.5%)	32,605 (91.5%)	14,627 (91.7%)	
Hispanic or Latino	1,261 (2.4%)	871 (2.4%)	390 (2.4%)	
Unknown	3,110 (6.0%)	2,173 (6.1%)	937 (5.9%)	
60–90 days	**n=51,603**	**n=36,667**	**n=14,936**	
Age	68 [59, 78]	68 [59, 77]	69 [58, 79]	<0.001
Sex, n (%)				<0.001
Female	26,201 (50.8%)	18,406 (50.2%)	7,795 (52.2%)	
Male	22,847 (44.3%)	16,597 (45.3%)	6,250 (41.8%)	
Unknown	2,555 (5.0%)	1,664 (4.5%)	891 (6.0%)	
Race, n (%)				<0.001
American Indian or Alaska Native	154 (0.3%)	114 (0.3%)	40 (0.3%)	
Asian	1,604 (3.1%)	1,328 (3.6%)	276 (1.8%)	
Black or African American	19,507 (37.8%)	12,239 (33.4%)	7,268 (48.7%)	
Native Hawaiian or Other Pacific Islander	47 (0.1%)	37 (0.1%)	10 (0.1%)	
White	24,827 (48.1%)	18,915 (51.6%)	5,912 (39.6%)	
Unknown	5,464 (10.6%)	4,034 (11.0%)	1,430 (9.6%)	
Ethnicity, n (%)				0.415
Not Hispanic or Latino	47,232 (91.5%)	33,555 (91.5%)	13,677 (91.6%)	
Hispanic or Latino	1,261 (2.4%)	880 (2.4%)	381 (2.6%)	
Unknown	3,110 (6.0%)	2,232 (6.1%)	878 (5.9%)	

Figure 8 and Figure 9 present the AUC ROC curves for all six models for the 30–60 and 60–90 day prediction windows.

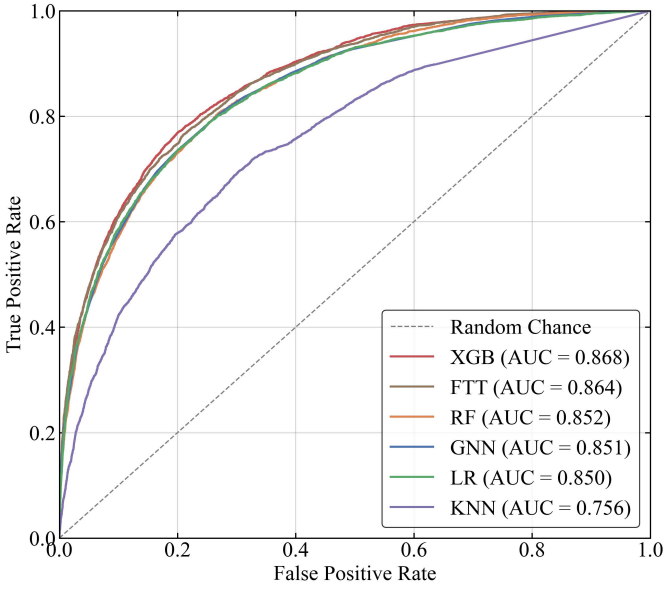


Fig. 8 AUC ROC curves of all six models for the 30–60 day window.

Figure 10 presents the cumulative normalized feature importance scores across all six models for the 30–60 day prediction window. Figure 11 presents the cumulative normalized feature importance scores across all six models for the 60–90 day prediction window.

Figure 12 presents the decision curve analyses for all six models for the 30–60 day prediction window. Figure 13 presents the decision curve analyses for all six models for the 60–90 day prediction window.

Figure 14 presents the fairness evaluation of FTT across sex, race, and ethnicity subgroups for the 30–60 day prediction window. Figure 15 presents the fairness evaluation of FTT across sex, race, and ethnicity subgroups for the 60–90 day prediction window.

Figure 16 presents the ablation analysis results showing ΔAUC for each feature group across all six models for the 30–60 day prediction window. Figure 17 presents the ablation analysis results showing ΔAUC for each feature group across all six models for the 60–90 day prediction window.

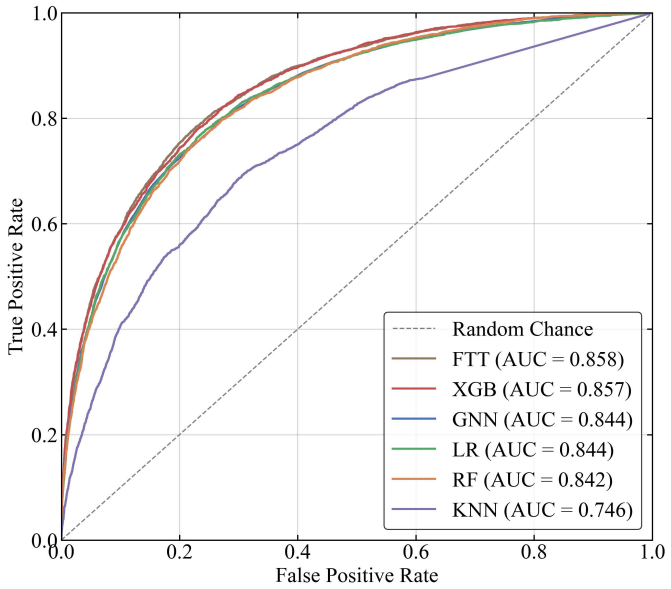


Fig. 9 AUC ROC curves of all six models for the 60–90 day window.

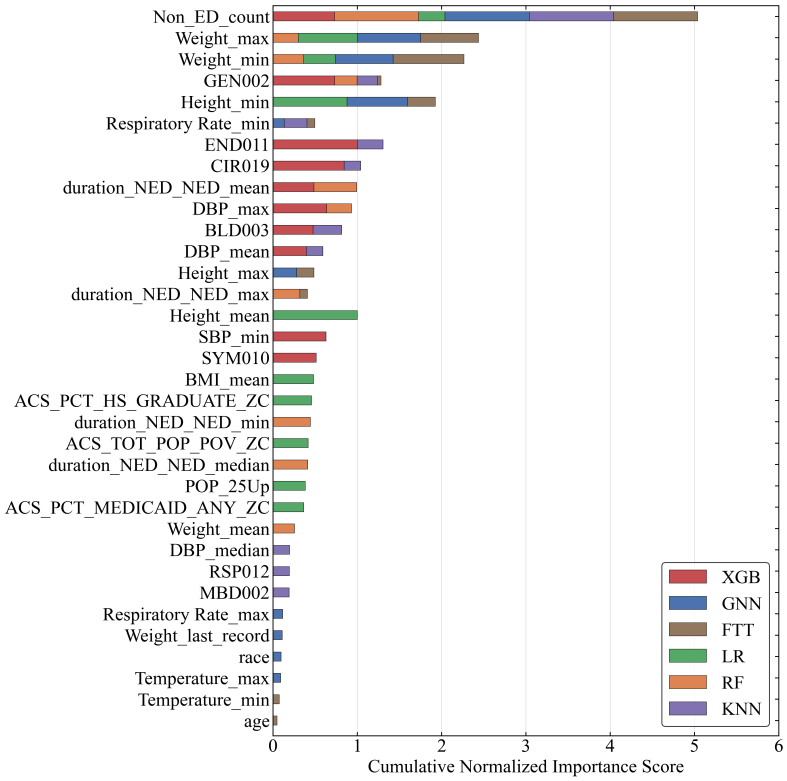


Fig. 10 Cumulative normalized feature importance scores across all six models for the 30–60 day window.

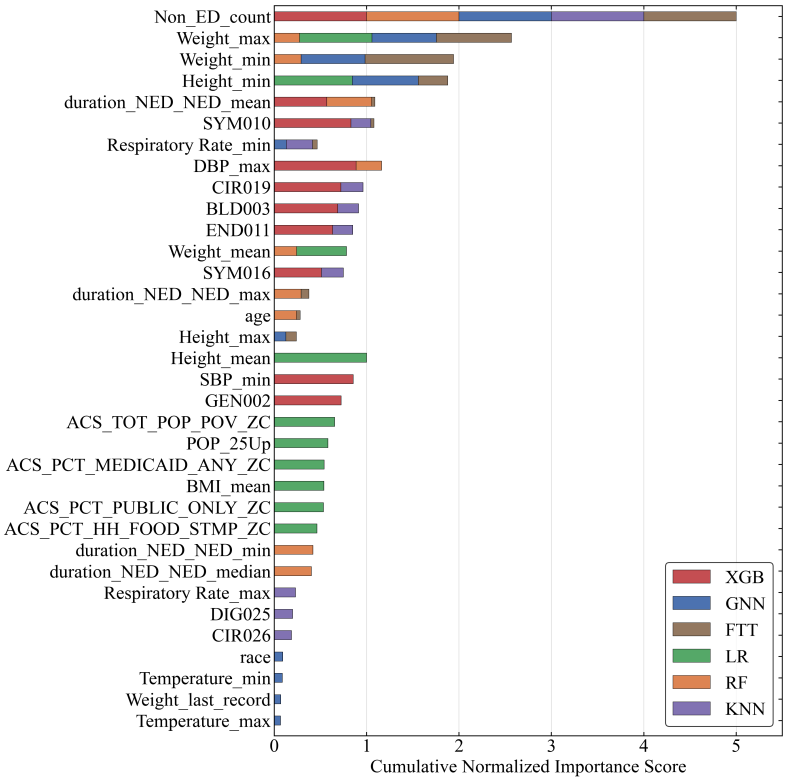


Fig. 11 Cumulative normalized feature importance scores across all six models for the 60–90 day window.

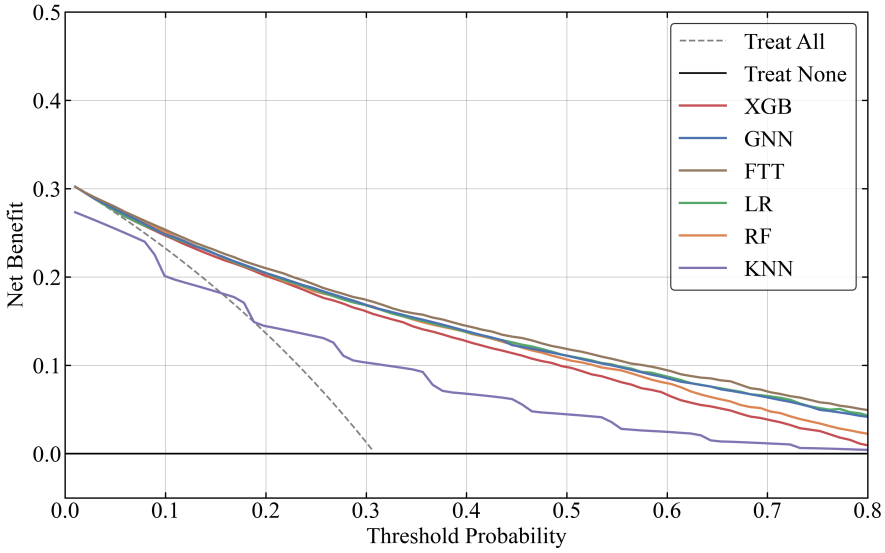


Fig. 12 Decision curve analyses for all six models across the 30–60 day window.

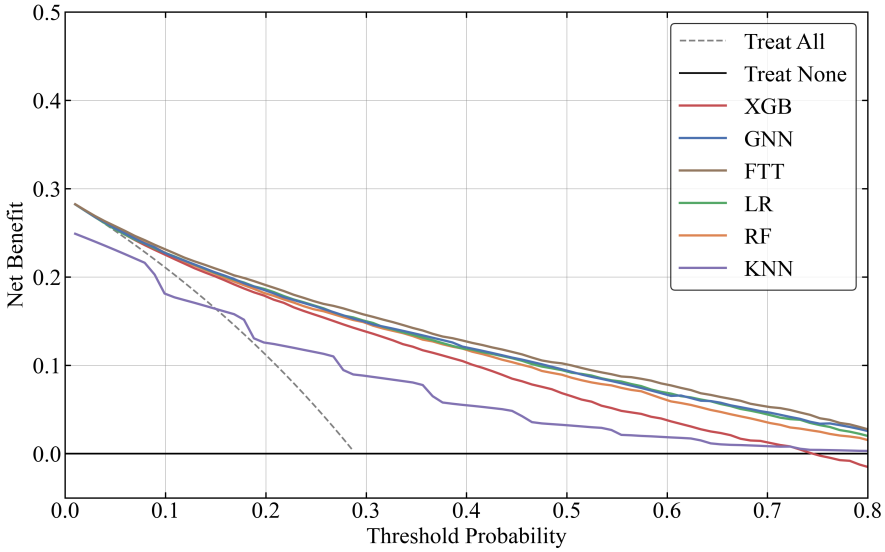


Fig. 13 Decision curve analyses for all six models across the 60–90 day window.

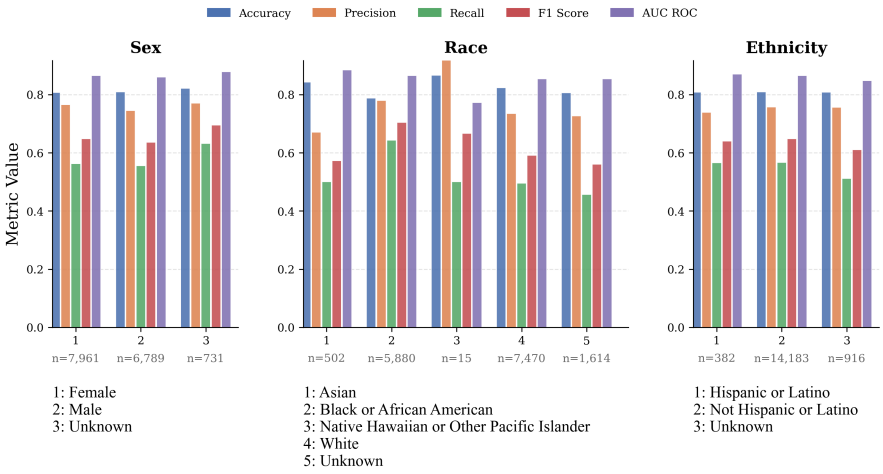


Fig. 14 Fairness evaluation of the FTT across sex, race, and ethnicity subgroups for the 30–60 day window.

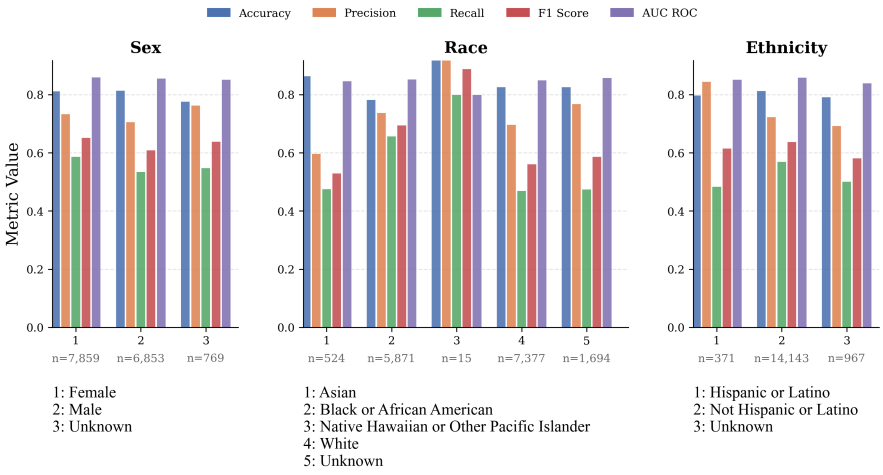


Fig. 15 Fairness evaluation of the FTT across sex, race, and ethnicity subgroups for the 60–90 day window.

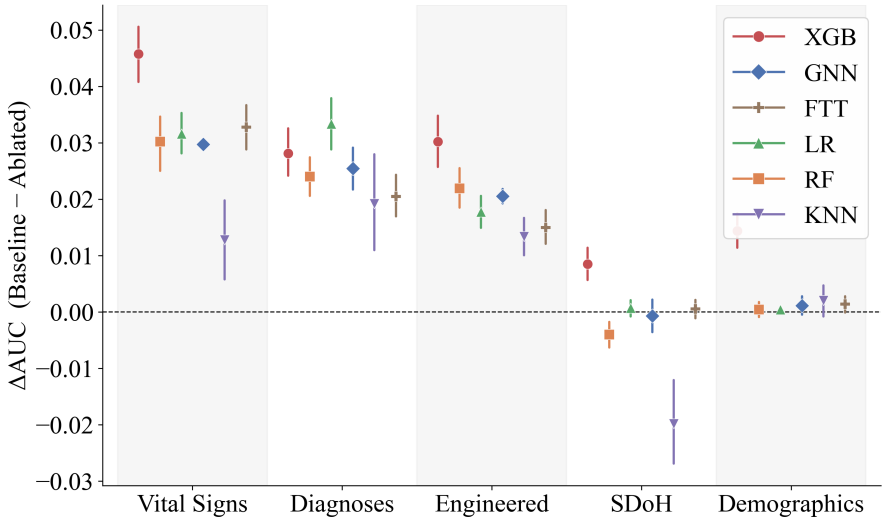


Fig. 16 Ablation analysis results showing ΔAUC for each feature group across all six models for the 30–60 day window.

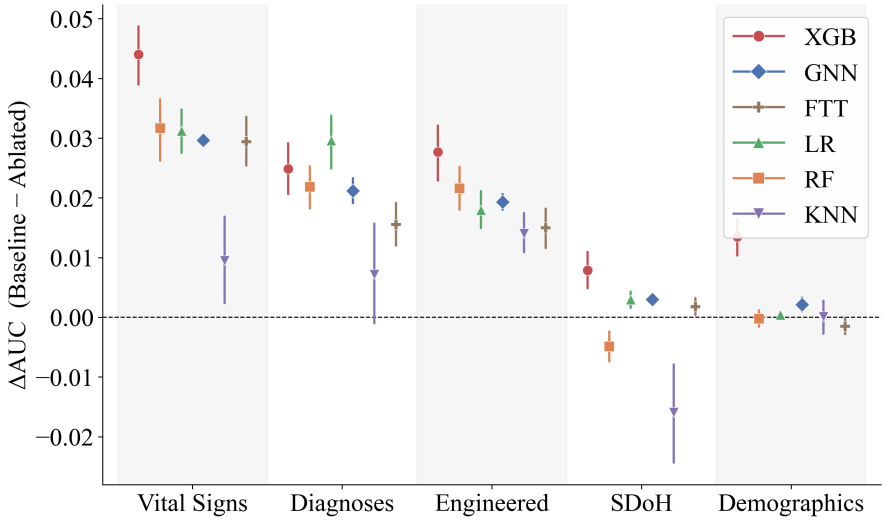


Fig. 17 Ablation analysis results showing ΔAUC for each feature group across all six models for the 60–90 day window.

References

[1] Uppal TS, Chehal PK, Fernandes G, et al. Trends and Variations in Emergency Department Use Associated With Diabetes in the US by Sociodemographic Factors, 2008-2017. JAMA Netw Open. 2022;5:E2213867. https://doi.org/10.1001/JAMANETWORKOPEN.2022.13867.

[2] Reynolds EL, Mizokami-Stout K, Putnam NM, et al. Cost and utilization of healthcare services for persons with diabetes. Diabetes Res Clin Pract. 2023;205:110983. https://doi.org/10.1016/J.DIABRES.2023.110983.

[3] Centers for Disease Control and Prevention. National Diabetes Statistics Report - United States Diabetes Surveillance System [Internet]; n.d. Available from: https://gis.cdc.gov/grasp/diabetes/diabetesatlas-statsreport.html. Accessed 13 May 2026.

[4] McCoy RG, Lipska KJ, Van Houten HK, Shah ND. Association of Cumulative Multimorbidity, Glycemic Control, and Medication Use With Hypoglycemia-Related Emergency Department Visits and Hospitalizations Among Adults With Diabetes. JAMA Netw Open. 2020;3:e1919099–e1919099. https://doi.org/10.1001/JAMANETWORKOPEN.2019.19099.

[5] Benoit SR, Hora I, Pasquel FJ, et al. Trends in Emergency Department Visits and Inpatient Admissions for Hyperglycemic Crises in Adults With Diabetes in the U.S., 2006–2015. Diabetes Care. 2020;43:1057–1064. https://doi.org/10.2337/DC19-2449.

[6] Centers for Disease Control and Prevention. National Diabetes Statistics Report [Internet]; 2024. Available from: https://www.cdc.gov/diabetes/php/data-research/index.html. Accessed 24 Nov 2025.

[7] Rubin DJ, Shah AA. Predicting and Preventing Acute Care Re-Utilization by Patients with Diabetes. Curr Diab Rep. 2021;21:34. https://doi.org/10.1007/S11892-021-01402-7.

[8] Tan JK, Quan L, Salim NNM, et al. Machine Learning-Based Prediction for High Health Care Utilizers by Using a Multi-Institutional Diabetes Registry: Model Training and Evaluation. JMIR AI. 2024;3:e58463. https://doi.org/10.2196/58463.

[9] Selya A, Anshutz D, Griese E, et al. Predicting unplanned medical visits among patients with diabetes: translation from machine learning to clinical implementation. BMC Medical Informatics and Decision Making. 2021;21:111. https://doi.org/10.1186/S12911-021-01474-1.

[10] Yang Y, Yu J, Liu S, et al. Predicting Avoidable Emergency Department Visits Using the NHAMCS Dataset. AMIA Summits on Translational Science

Proceedings. 2022;2022:514.

[11] Alhumaidi NH, Dermawan D, Kamaruzaman HF, Alotaiq N. The Use of Machine Learning for Analyzing Real-World Data in Disease Prediction and Management: Systematic Review. JMIR Med Inform. 2025;13:e68898. https://doi.org/10.2196/68898.

[12] Park S, Lee C, Lee SB, Lee JY. Machine learning-based prediction model for emergency department visits using prescription information in community-dwelling non-cancer older adults. Scientific Reports. 2023;13:18887. https://doi.org/10.1038/s41598-023-46094-z.

[13] Olawade DB, Da'Costa A, Origbo JE, et al. Smart emergency care: a narrative review of predictive machine learning models. Ann Transl Med. 2025;13:60–60. https://doi.org/10.21037/ATM-25-83.

[14] Cicek M, Buckley J, Pearson-Stuttard J, Gregg EW. Characterizing Multimorbidity from Type 2 Diabetes: Insights from Clustering Approaches. Endocrinol Metab Clin North Am. 2021;50:531–558. https://doi.org/10.1016/J.ECL.2021.05.012.

[15] Williams EL, Huynh D, Estai M, et al. Predicting Inpatient Admissions From Emergency Department Triage Using Machine Learning: A Systematic Review. Mayo Clinic Proceedings: Digital Health. 2025;3:100197. https://doi.org/10.1016/J.MCPDIG.2025.100197.

[16] Kavakiotis I, Tsave O, Salifoglou A, et al. Machine Learning and Data Mining Methods in Diabetes Research. Comput Struct Biotechnol J. 2017;15:104–116. https://doi.org/10.1016/J.CSBJ.2016.12.005.

[17] Weng SF, Reps J, Kai J, et al. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One. 2017;12:e0174944. https://doi.org/10.1371/JOURNAL.PONE.0174944.

[18] Vest JR, Ben-Assuli O. Prediction of emergency department revisits using area-level social determinants of health measures and health information exchange information. Int J Med Inform. 2019;129:205–210. https://doi.org/10.1016/j.ijmedinf.2019.06.013.

[19] Lee SY, Hayes LW, Ozaydin B, et al. Integrating Social Determinants of Health in Machine Learning-Driven Decision Support for Diabetes Case Management: Protocol for a Sequential Mixed Methods Study. JMIR Res Protoc. 2024;13:e56049. https://doi.org/10.2196/56049.

[20] Mohammad Alizadeh J, Patel JS, Tajeu G, et al. Predicting Emergency Department Visits for Patients with Type II Diabetes. arXiv preprint. 2024;arXiv:2412.08984. https://doi.org/10.48550/arXiv.2412.08984.

[21] HealthShare Exchange (HSX). Clinical Data Repository (CDR) Overview [Internet]; 2017. Available from: https://healthshareexchange.org/wp-content/uploads/2021/08/11.20.17_cdr_overview_final.pdf.

[22] Agency for Healthcare Research and Quality. Community-Level Health Database [Internet]; 2026. Available from: https://www.ahrq.gov/data/innovations/clh-data.html#download. Accessed 31 Oct 2026.

[23] Mohammad Alizadeh J, Tajeu G, Hollin IL, et al. THReD: An agentic modular and reproducible system for transforming electronic health records into AI-ready datasets; 2026. Manuscript under review.

[24] Agency for Healthcare Research and Quality. Clinical Classifications Software Refined (CCSR) for ICD-10-CM Diagnoses [Internet]; 2026. Available from: https://hcup-us.ahrq.gov/toolssoftware/ccsr/dxccsr.jsp#overview. Accessed 5 Jan 2026.

[25] Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods. 2020;17:261–272. https://doi.org/10.1038/s41592-019-0686-2.

[26] Kerby DS. The Simple Difference Formula: An Approach to Teaching Nonparametric Correlation. Comprehensive Psychology. 2014;3:11.IT.3.1. https://doi.org/10.2466/11.IT.3.1.

[27] Sullivan GM, Feinn R. Using Effect Size—or Why the P Value Is Not Enough. J Grad Med Educ. 2012;4:279–282. https://doi.org/10.4300/JGME-D-12-00156.1.

[28] Zhang Z. Introduction to machine learning: k-nearest neighbors. Ann Transl Med. 2016;4. https://doi.org/10.21037/ATM.2016.03.37.

[29] Breiman L. Random Forests. Mach Learn. 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.

[30] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016;13-17-August-2016:785–794. https://doi.org/10.1145/2939672.2939785.

[31] Scarselli F, Gori M, Tsoi AC, et al. The graph neural network model. IEEE Trans Neural Netw. 2009;20:61–80. https://doi.org/10.1109/TNN.2008.2005605.

[32] Gorishniy Y, Rubachev I, Khrulkov V, Babenko A. Revisiting Deep Learning Models for Tabular Data. Adv Neural Inf Process Syst. 2021;34:18932–18943.

[33] Pedregosa F, Michel V, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825–2830.

[34] Kipf TN, Welling M. Semi-Supervised Classification with Graph Convolutional Networks. 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings. 2017;arXiv:1609.02907.

[35] He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21:1263–1284. https://doi.org/10.1109/TKDE.2008.239.

[36] Raschka S. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning; 2018.

[37] Fisher A, Rudin C, Dominici F. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. Journal of Machine Learning Research. 2019;20:1–81.

[38] Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ. 2016;352. https://doi.org/10.1136/BMJ.I6.

[39] Vickers AJ, Elkin EB. Decision curve analysis: A novel method for evaluating prediction models. Medical Decision Making. 2006;26:565–574. https://doi.org/10.1177/0272989X06295361.

[40] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366:447–453. https://doi.org/10.1126/SCIENCE.AAX2342.

[41] Chen IY, Pierson E, Rose S, et al. Ethical Machine Learning in Healthcare. Annu Rev Biomed Data Sci. 2021;4:123–144. https://doi.org/10.1146/annurev-biodatasci-092820-114757.

[42] Zeiler MD, Fergus R. Visualizing and Understanding Convolutional Networks. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2013;8689 LNCS:818–833. https://doi.org/10.1007/978-3-319-10590-1_53.

[43] Hamilton WL, Ying R, Leskovec J. Inductive Representation Learning on Large Graphs. Adv Neural Inf Process Syst. 2017;2017-December:1025–1035.

[44] Sedgwick P. Multiple hypothesis testing and Bonferroni's correction. BMJ. 2014;349. https://doi.org/10.1136/BMJ.G6284.

[45] Moss JL, Johnson NJ, Yu M, et al. Comparisons of individual- and area-level socioeconomic status as proxies for individual-level measures: evidence from the Mortality Disparities in American Communities study. Population Health Metrics. 2021;19:1. https://doi.org/10.1186/S12963-020-00244-X.

[46] Buajitti E, Chiodo S, Rosella LC. Agreement between area- and individual-level income measures in a population-based cohort: Implications for population health research. SSM Popul Health. 2020;10:100553. https://doi.org/10.1016/J.SSMPH.2020.100553.

[47] Mirkes EM, Allohibi J, Gorban A. Fractional Norms and Quasinorms Do Not Help to Overcome the Curse of Dimensionality. Entropy. 2020;22:1105. https://doi.org/10.3390/E22101105.

[48] Chen IY, Johansson FD, Sontag D. Why Is My Classifier Discriminatory? Adv Neural Inf Process Syst. 2018;2018-December:3539–3550.

[49] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019;1:206–215. https://doi.org/10.1038/s42256-019-0048-x.

[50] Van Calster B, Wynants L, Verbeek JFM, et al. Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators. Eur Urol. 2018;74:796–804. https://doi.org/10.1016/J.EURURO.2018.08.038.