Michigan Tech

NAME OF THE STUDENTS: MOHAMED ABDULAI AND PETER MVUMA

GROUP NUMBER: 10

COURSE CODE: SAT 5141

ASSIGNMENT: Course Research Final Report

DUE DATE: 12/05/2025

Title: Developing Subgroup-Specific AI Models for Predicting 30-Day Readmission in Patients with Diabetes.

1.0. Introduction

The global prevalence of diabetes has surged dramatically over the past few decades, now affecting an estimated 589 million adults aged 20-79 years in 2024, which accounts for approximately 11.1% of this population segment (IDF, 2025; WHO, 2024). Projections indicate that this number will rise to around 853 million by 2050, with the highest increases anticipated in low- and middle-income countries, such as those in Africa and the Middle East (IDF, 2025). The trend is driven by factors such as increasing obesity rates, urbanization, and sedentary lifestyles, which significantly contribute to the rising incidence of type 2 diabetes, the predominant form, comprising over 96% of all cases (IDF, 2025; Pharmacies Times, 2024). This exponential growth has resulted in staggering health and economic impacts, with global health expenditures exceeding USD 1 trillion in 2024, leading to approximately 3.4 million deaths annually (WHO, 2024).

This escalating epidemic calls for an urgent need for effective intervention strategies to manage the growing health burden and costs associated with diabetes. Despite advances in treatment, many individuals remain undiagnosed, particularly in low-resource settings, further complicating disease management and increasing the risk of hospital readmissions and complications (WHO, 2019; IDF, 2025). Developing predictive models, particularly subgroupspecific artificial intelligence (AI) approaches, could enhance early detection, personalize treatment plans, and ultimately reduce the rate of 30-day hospital readmissions, a crucial goal for optimizing diabetes care and reducing healthcare costs globally (Wang et al., 2024; Khadka et al., 2022).

2.0. Literature Review on Clinical Decision Support and AI Modeling and Risk Prediction for Diabetic Patients

2.1.General Literature Review in Clinical Decision Support and Modeling

Clinical decision support systems (CDSS) have evolved into a crucial component of modern diabetes care. This shift is attributed to the increasing complexity of patient management and the proliferation of electronic health records (EHRs). Early CDSS implementations were often rule-based and focused on improving test ordering and preventive services. However, their impact on core diabetes outcomes such as glycemic control and complication rates was inconsistent. This inconsistency was largely due to low adoption and suboptimal integration into clinical workflows (O'Connor et al., 2004). Recent systematic reviews demonstrate that well-designed, EHR-integrated CDSSs can significantly improve both process and outcome measures in diabetes care. The greatest benefit is observed when they provide real-time, patient-specific recommendations and are integrated within team-based care models (Benkhalti et al., 2024; Sattler et al., 2022).

A key observation in the literature is that the effectiveness of Clinical Decision Support Systems (CDSS) is closely tied to usability, clinician engagement, and support for shared decision-making. For example, Benkhalti et al. (2024) found that CDSSs tailored for antidiabetic drug management reduced inappropriate prescriptions and hypoglycemia events. Benefits were especially notable when features like treatment guidance (providing clinicians with recommendations on patient care), risk estimation (calculating the probability of potential complications), and flagging of noncompliance (alerting providers to instances when patients do not follow treatments as prescribed) were present. Similarly, Cai et al. (2023) highlighted the growing adoption of machine learning-based CDSS, which are systems that use algorithms to learn from data and improve over time. These systems are increasingly capable of

synthesizing large volumes of clinical data and providing nuanced, context-specific recommendations. They improve traditional outcomes, such as blood glucose and blood pressure control, and facilitate more personalized, dynamic care plans that evolve with the patient's clinical state (Cai et al., 2023; O'Connor et al., 2004).

The American Diabetes Association's 2025 Standards of Care highlight the importance of integrating advanced technologies, such as continuous glucose monitoring and automated insulin dosing, into routine practice. They recommend using CDSS to support individualized care and address diabetes disparities (American Diabetes Association, 2024). Despite advances, challenges persist. It is challenging to optimize model performance for patients with complex comorbidities, ensure clinician trust and adoption, and validate the effectiveness of CDSS across diverse healthcare settings (Cai et al., 2023; Benkhalti et al., 2024).

2.2. Literature on Risk Prediction for Diabetic Patients

Risk prediction for diabetic patients, particularly regarding 30-day hospital readmission, has become a focal point for both clinical research and AI innovation. Large-scale studies, such as those by Strack et al. (2014), have shown that key predictors of readmission include HbA1c measurement (a blood test that reflects average blood sugar levels), medication changes, comorbidities (the presence of additional diseases), and discharge disposition (the location to which the patient is sent after leaving the hospital). Machine learning models, including random forests and XGBoost (advanced data analysis techniques that utilize algorithms to identify patterns), have consistently outperformed traditional regression approaches in identifying highrisk patients (Liu et al., 2024). However, most existing models have been developed for the general diabetic population. This can lead to the overlooking of subgroup-specific risk factors.

Recent advances emphasize the value of subgroup-specific modeling. Khadka et al. (2022) introduced frameworks to model and interpret patient subgroups. They found that tailored models not only improve predictive accuracy but also reveal distinct profiles of feature importance. These profiles can inform more targeted interventions. Ghorbani et al. (2024) further support this approach, demonstrating that multicriteria decision-making frameworks can enhance personalized treatment selection and risk stratification in diabetes management.

The literature highlights the growing role of CDSS in perioperative and inpatient diabetes care. Cai et al. (2025) and the American Association of Clinical Endocrinology (2023) report that, when combined with real-time glucose monitoring and automated insulin delivery, digital decision support tools can reduce perioperative complications and improve adherence to evidence-based guidelines. These systems are particularly important in high-risk settings that require timely, data-driven interventions (Cai et al., 2025).

Despite these promising developments, several barriers persist. There is a need for more randomized controlled trials to validate CDSS and AI models in real-world settings. Integrating patient-reported outcomes and data from wearable devices remains a challenge. Clinician education and workflow integration are crucial to ensuring sustained use and impact (Benkhalti et al., 2024; Cai et al., 2023; O'Connor et al., 2004). The literature agrees that future research should prioritize model optimization for complex patient populations, robust validation across diverse care environments, and strategies to enhance clinician trust and adoption.

3.0. Focus of our Research Project

The literature reviewed above demonstrates that clinical decision support systems and AI-based modeling have significantly advanced diabetes care, particularly in predicting hospital readmissions. However, most existing models are global, treating all diabetic encounters similarly and often failing to account for subgroup-specific risk factors and clinical

heterogeneity (Khadka et al., 2022; Liu et al., 2024; Strack et al., 2014). This gap highlights the need for tailored predictive models that can provide more actionable insights for distinct patient subgroups, as emphasized by recent systematic reviews and comparative studies (Benkhalti et al., 2024; Cai et al., 2023).

Building on these findings, the present study aims to develop and evaluate subgroup-specific AI models for predicting 30-day readmission in patients with diabetes, using the Diabetes 130US Hospitals dataset. By stratifying patients according to primary diagnosis and leveraging advanced machine learning techniques, this research seeks to address the limitations identified in the literature and contribute to more precise, clinically relevant decision support.

4.0. Data collection

This research will use secondary data, and the source of the data is the UC Irvine Machine Learning repository. Here is the UCI Diabetes dataset for your review.

4.1. Dataset Integrity and Validation

This dataset is considered 'good data' in our study because it was curated by expert researchers, published in a peer-reviewed journal, and made publicly available through the UCI Machine Learning Repository, one of the most reputable sources. The dataset comprises diverse patient records, with clearly defined inclusion criteria and comprehensive feature documentation, which supports its quality and suitability for our research.

In addition to the dataset's strengths, we employed standard preprocessing steps, including handling missing values, encoding categorical variables, and stratified train-test splits, to ensure data integrity and reliability for our predictive modeling.

4.2. Description of the dataset

The study will use the **Diabetes 130-US Hospitals (1999–2008)** dataset from the UCI Machine Learning Repository, originally described by Strack et al. (2014) and later used in predictive modeling research (Liu et al., 2024). The dataset contains 101,766 hospital encounters for 71,518 unique patient records, and diabetic patients from 130 U.S. hospitals. Available variables include demographics, admission and discharge information, lab tests (HbA1c, serum glucose), comorbid diagnoses, procedures, prior healthcare utilization, and detailed medication change indicators. The dataset contains a total of 47 features and an outcome of interest, which is hospital readmission within 30 days of discharge (<30 vs. >30/NO).

4.3. Data Preprocessing

Irrelevant or identifying variables, such as patient ID and encounter ID, were excluded. Features with Missing values above 35% were deleted from the dataset, and the rest were handled through imputation. Under this, we removed weight, which had 96.8% missing values, medical specialty (49%), and pay code (39.5%).
The rest of the features with missing values were race, primary diagnosis (DIAG_1), secondary diagnosis (DIAG2) and additional secondary diagnosis (DIAG3) with missing values falling under 3%. In this case, we imputed the missing values with categories with the highest frequency in each feature.

Categorical features of more than two categories were one-hot encoded to ensure numerical compatibility across models. Features like race, admission type, admission source, and discharge disposition were regrouped into binary categories because the majority of the categories had low frequency values. For instance, in the case of race, 77% of patients were Caucasian. So, it was regrouped as “Caucasian” and “Others”. ‘Admission type’ had eight

categories, with the emergency category being 82546 and the rest totaling 19220. As a result, it was regrouped into “Emergency” and “Others”. The admission source was regrouped into “Emergency room” and “Others”. Discharge disposition regrouped into “Discharged to home” and “Others”.

Also, for medications, two attributes, namely “citoglipton” and “examide,” were removed from the dataset because they contained only one category. Further analysis of the dataset revealed that nineteen of the medications had only one (i.e, patients not on the medication) category out of the four, accounting for at least 89% of the data. These 19 columns were deleted because they provide very little information for prediction and removing them prevents misleading correlations that arise purely due to dominance; reduces computational load and reduces the risk. So, “metformin” and “Insulin” were retained. Even though the medication attributes were ordinal with 4 possible values (“up”, “down”, “steady”, and “no”), they were converted into binary to reduce the computational time of the models. The four levels were “up”, “down”, “steady”, and “no” corresponding to an increase in dosage, a decrease in dosage, dosage not changed, and drug not prescribed, respectively.
Numeric features like time in hospital, number of labs, number of procedures, number of medications, number of outpatients, number of inpatients, number of emergencies, and number of diagnoses were maintained in their original formats.

Lab results, namely max_glu_serum, A1Cresult were also converted into binary. For the outcome variable, which is readmission status, there were three categories. That is, patients who were readmitted within 30 days after their previous discharge, patients who were readmitted after 30 days of their previous discharge, and patients who were not readmitted. Because the interest was to predict risk of readmission within 30 days, the out variable was converted into binary by merging encounters with readmission after 30 days and those with no readmission records. Primary, secondary, and additional diagnoses were categorized into “circulatory”, “respiratory”, “Endocrine, nutritional, metabolic & immunity disorders”, and “Others”.

5.0. Research Methodology

In this project, only the first hospital visit for each patient was used instead of the full dataset containing repeated encounters because multiple visits introduce significant noise, inconsistency, and label instability that weaken the model’s ability to learn true predictive patterns. Patients often have several visits with highly similar clinical features but different outcomes. In which case, sometimes they are readmitted and other times not, which creates contradictory signals that confuse traditional machine-learning models and substantially reduce generalization performance. Even when GroupKFold was applied to avoid patient-level leakage, the repeated-visit dataset still produced poor results due to outcome overlap, evolving clinical conditions across visits, and the inability of tabular models to capture temporal progression. By restricting the dataset to only first visits, the analysis preserves independent, stable, and clinically meaningful feature–outcome relationships, resulting in a cleaner signal, more reliable training behavior, and cross-validation performance that reflects true predictive capability for unseen patients.

The study implemented multiple machine learning algorithms for each of the primary diagnosed subgroups. These ML algorithms included Logistics Regression, Decision Tree, SVM (linear), KNN, Random Forest, XGBoost, and LightGBM. Each model was embedded within a pipeline.

Train-Test Split/Cross-Validation/Final Model Training

The dataset was randomly divided into an 80% training set and a 20% testing set. Stratified sampling was used to maintain the distribution of outcomes . Data imbalance was addressed using Synthetic Minority Over-Sampling Technique (SMOTE), and the training data was scaled. To prevent data leakage, all model development, including cross-validation, was done only on the training set. To evaluate the models' stability and its ability to generalize, we performed 5-fold cross-validation on the training data. The performance metrics calculated during cross-validation were accuracy, precision, recall, f1-score, and Area Under the ROC Curve (AUC). Confidence intervals were estimated through cross-validation for each metric providing robustness in performance estimation. After cross-validation, the best performing model was selected from each subgroup (including the global model), and trained using the entire 80% training set.

Model Evaluation on Test Set

Models performance was assessed on the untouched 20% test dataset, employing accuracy, precision, recall, F1-score, and ROC AUC as evaluation metrics.

Model evaluation metrics

Model performance was evaluated based on accuracy, precision, recall, F1-score (macroaveraged), and ROC-AUC. Confidence intervals were estimated for the metrics and feature importance through cross-validation, providing robustness in performance estimation.

6.0. Description of the preliminary findings

6.1. Comparative Profile of Patients Readmitted <30 Days vs. >30 Days/No Readmission

7.1.1. Demographic Characteristics 7.1.2. Race

The cohort consisted of 71,518 patients, predominantly Caucasian (78%) and slightly more female (53%), with similar distributions across the >30/no-readmission and <30 readmission groups. No substantial difference in racial distribution was observed between patients readmitted within 30 days and those not readmitted within the same period.

7.1.3. Gender

The gender distribution was nearly identical in both groups, with approximately 46% of the participants being male and 54% female in each category.

7.1.4. Age

Age was skewed toward older adults, with the largest proportions in the 60–80 age range, especially those aged 70–80 (25%) and 60–70 (22%), while younger age groups represented very small fractions.

7.2.0. Hospital Admission and Discharge Details 7.2.1. Admission Type and Source

Most admissions were emergencies (80%), with more than half originating specifically from the emergency room (54%). Patients who were not readmitted had shorter hospital stays (mean 4.24 vs. 4.80 days), fewer inpatient encounters (0.16 vs. 0.37), and slightly fewer outpatient and emergency visits.

7.3.0. Discharge Disposition

Discharge disposition differed notably 67% of non-readmitted patients were transferred home versus only 53% of the <30 group, indicating a greater proportion of complex or alternative discharge pathways among those who returned.

7.4.0. Diagnoses Distribution 7.4.1. Primary Diagnosis

Both groups exhibited a broadly similar distribution of primary diagnostic categories. Across diagnoses, circulatory conditions were the most common primary category (31%), followed by endocrine/metabolic disorders (11%) and respiratory illnesses (9%), with similar trends in secondary and additional diagnoses.

7.5.0. Laboratory Results (Glycemic Control Measures)

Most patients had no max glucose serum or A1C results recorded, although abnormal glucose levels (>300) were more common among readmitted patients.

7.6.0. Diabetes Medication Usage

In terms of diabetes management, insulin use (55% vs. 49%), medication changes (47% vs. 45%), and overall diabetes medication use (80% vs. 76%) were more frequent in the <30 group, suggesting greater treatment intensity. Finally, those readmitted within 30 days had a higher mean number of medications (16.62 vs. 15.62), underscoring increased clinical complexity and therapeutic burden among patients who returned soon after discharge.

7.7.0. Admission Type

In terms of admission type, the majority of the encounters were readmitted through the emergency route, accounting for 81.11% while the rest of the encounters were termed others were readmitted through other means apart from the emergency room. These accounted for 18.89%.

Figure 01. Tabular view of the Demographics by encounters

Characteristics	>30/NO, n = 6293 (%	<30, n = 65225 (%)	Total, n = 71518 (%)
race			
Caucasian	50482(77%)	4957(79%)	55439(78%)
Others	14743(23%)	1336(21%)	16079(22%)
gender			
Male	30567(47%)	2923(46%)	33490(47%)
Female	34658(53%)	3370(54%)	38028(53%)
age			
[0-10)	151(0%)	3(0%)	154(0%)
[10-20)	509(1%)	26(0%)	535(1%)
[20-30)	1044(2%)	83(1%)	1127(2%)

[30-40)	2511(4%)	188(3%)	2699(4%)
[40-50)	6371(10%)	507(8%)	6878(10%)
[50-60)	11587(18%)	879(14%)	12466(17%)
[60-70)	14546(22%)	1414(22%)	15960(22%)
[70-80)	16386(25%)	1824(29%)	18210(25%)
[80-90)	10388(16%)	1201(19%)	11589(16%)
[90-100)	1732(3%)	168(3%)	1900(3%)
admission_type			
Emergency	52152(80%)	5128(81%)	57280(80%)
Others	13073(20%)	1165(19%)	14238(20%)
admission_source			
Emergency room	34833(53%)	3457(55%)	38290(54%)
Others	30392(47%)	2836(45%)	33228(46%)
Time in hospital (mean)	4.24	4.80	4.29
Number of emergencies (mean)	0.10	0.15	0.10
Number of outpatient (mean)	0.28	0.31	0.28
Number of inpatient (mean)	0.16	0.37	0.18
Discharge disposition			
Transferred to home	43462(67%)	3329(53%)	46791(65%)
Others	21763(33%)	2964(47%)	24727(35%)
Primary diagnosis			
Circulatory system	19756(30%)	2076(33%)	21832(31%)
Endocrine, nutritional, metabolic & immunity disorders	6993(11%)	692(11%)	7685(11%)
Others	32276(49%)	2989(47%)	35265(49%)
Respiratory system	6200(10%)	536(9%)	6736(9%)
Secondary diagnosis			
Circulatory system	20475(31%)	2025(32%)	22500(31%)
Endocrine, nutritional, metabolic & immunity disorders	14197(22%)	1285(20%)	15482(22%)
Others	24349(37%)	2426(39%)	26775(37%)
Respiratory system	6204(10%)	557(9%)	6761(9%)
Additional diagnosis			
Circulatory system	20373(31%)	1902(30%)	22275(31%)
Endocrine, nutritional, metabolic &			

immunity disorders	17653(27%)	1584(25%)	19237(27%)
Others	23180(36%)	2360(38%)	25540(36%)
Respiratory system	4019(6%)	447(7%)	4466(6%)
Number of procedures (mean)	1.43	1.42	1.43
Number of diagnoses (mean)	7.22	7.51	7.25
Number of lab procedures (mean)	42.90	44.93	43.08
Max_glu_serum			
None	62111(95%)	5951(95%)	68062(95%)
Normal	1565(2%)	166(3%)	1731(2%)
>200	0%	0%	0%
>300	1549(2%)	176(3%)	1725(2%)
A1Cresult	0%	0%	0%
None	53319(82%)	5213(83%)	58532(82%)
Normal	3467(5%)	324(5%)	3791(5%)
>7	2645(4%)	246(4%)	2891(4%)
>8	5794(9%)	510(8%)	6304(9%)
Metfomin			
No	51471(79%)	5056(80%)	56527(79%)
Yes	13754(21%)	1237(20%)	14991(21%)
Insulin			
No	32079(49%)	2842(45%)	34921(49%)
Yes	33146(51%)	3451(55%)	36597(51%)
Change			
No	36180(55%)	3314(53%)	39494(55%)
Yes	29045(45%)	2979(47%)	32024(45%)
Diabetes medications			
No	15936(24%)	1263(20%)	17199(24%)
Yes	49289(76%)	5030(80%)	54319(76%)
Mean number of medications	15.62	16.62	15.71

8.0. Models' Performance Metrics and Model Selection

8.1. Overview of the Global model

The comparative analysis of various machine-learning models reveals that Random Forest exhibits the most robust and well-rounded performance. This model attains a high accuracy of 0.924 ± 0.001, coupled with excellent precision (0.978 ± 0.002), a strong recall of 0.869 ± 0.001, the highest F1-score of 0.920 ± 0.001, and the best ROC-AUC of 0.964 ± 0.001, thereby indicating consistent stability across all evaluated metrics. Although LightGBM and XGBoost also demonstrate strong performance, characterized by perfect precision and competitive recall, F1, and AUC values, they are slightly outperformed by Random Forest. This is attributable to their comparatively lower recall and F1-scores, which suggests a reduced sensitivity relative to the leading model. Conversely, Logistic Regression and Linear SVM exhibit substantially weaker performance, while KNN and Decision Tree models perform moderately, yet still below the ensemble models. Consequently, considering accuracy, recall, F1, and AUC collectively, Random Forest emerges as the superior model in this comparative assessment.

8.2. Global Model

Model	Accuracy	Precision	Recall	F1	Roc_auc
Random Forest	0.924 ± 0.001	0.978 ± 0.002	0.869 ± 0.001	0.920 ± 0.001	0.964 ± 0.001
LightGBM	0.926 ± 0.001	1.000 ± 0.000	0.852 ± 0.002	0.920 ± 0.001	0.957 ± 0.001
XGBoost	0.924 ± 0.001	1.000 ± 0.000	0.848 ± 0.003	0.918 ± 0.002	0.956 ± 0.001
Logistic Regression	0.589 ± 0.004	0.594 ± 0.005	0.561 ± 0.002	0.577 ± 0.003	0.627 ± 0.004
SVM (Linear)	0.589 ± 0.004	0.594 ± 0.004	0.560 ± 0.002	0.577 ± 0.003	0.627 ± 0.004
KNN5	0.834 ± 0.002	0.752 ± 0.002	0.997 ± 0.001	0.858 ± 0.002	0.943 ± 0.002
Decision Tree	0.902 ± 0.002	0.898 ± 0.003	0.906 ± 0.002	0.902 ± 0.002	0.906 ± 0.002

9.0. Primary Diagnosis/Sub-specific

9.1. Circulatory model

The comparative performance of the models, as presented in the table, clearly demonstrates that Random Forest exhibits the most robust overall capabilities. This model attains the highest accuracy (0.933 ± 0.003), the greatest precision (0.981 ± 0.003), a commendable recall (0.884 ± 0.006), the superior F1-score (0.930 ± 0.004), and the highest ROC-AUC (0.968 ± 0.002). Although LightGBM and XGBoost also demonstrate strong performance, particularly in terms of precision, recall, and AUC, they consistently underperform relative to Random Forest across nearly all critical metrics. Conversely, models such as Logistic Regression and SVM (Linear) exhibit significant deficiencies, characterized by low recall and a generally weak capacity for discrimination. KNN and Decision Tree models demonstrate moderate performance, yet they do not achieve the same level of efficacy as the ensemble models. Consequently, considering accuracy, recall, F1-score, and AUC in conjunction, Random Forest emerges as the superior model within this assessment.

9.2. Circulatory Model

Model	Accuracy	Precision	Recall	F1	Roc_auc
Random Forest	**0.933 ± 0.003**	**0.981 ± 0.003**	**0.884 ± 0.006**	**0.930 ± 0.004**	**0.968 ± 0.002**
LightGBM	0.924 ± 0.002	1.000 ± 0.000	0.849 ± 0.004	0.918 ± 0.002	0.955 ± 0.002
XGBoost	0.923 ± 0.003	1.000 ± 0.000	0.847 ± 0.006	0.917 ± 0.004	0.954 ± 0.002
Logistic Regression	0.576 ± 0.004	0.580 ± 0.005	0.549 ± 0.007	0.564 ± 0.004	0.606 ± 0.004
SVM (Linear)	0.576 ± 0.005	0.580 ± 0.005	0.550 ± 0.007	0.565 ± 0.004	0.606 ± 0.004
KNN5	0.821 ± 0.004	0.738 ± 0.005	0.996 ± 0.000	0.847 ± 0.003	0.935 ± 0.003
Decision Tree	0.890 ± 0.007	0.883 ± 0.012	0.899 ± 0.005	0.891 ± 0.006	0.898 ± 0.007

9.3. Respiratory Model

The findings clearly indicate that Random Forest constitutes the most effective model, demonstrating superior performance across all principal evaluation metrics. It attains the highest accuracy (0.959 ± 0.005), precision (0.992 ± 0.003), recall (0.925 ± 0.007), F1-score (0.957 ± 0.005), and ROC-AUC (0.987 ± 0.002). Although LightGBM and XGBoost exhibit commendable performance and present viable alternatives, particularly due to their elevated precision and satisfactory recall, they are marginally less effective than Random Forest when considering the aggregate metrics. Conversely, models such as Logistic Regression and SVM (Linear) display considerably weaker performance, characterized by significantly lower recall, F1, and AUC values. KNN and Decision Tree models demonstrate moderate performance but lack the consistency observed in the ensemble methods. Consequently, Random Forest emerges as the most advantageous model, as determined by accuracy, recall, F1-score, and ROC-AUC.

9.4. Respiratory Model

Model	Accuracy	Precision	Recall	F1	Roc_auc
Random Forest	**0.959 ± 0.005**	**0.992 ± 0.003**	**0.925 ± 0.007**	**0.957 ± 0.005**	**0.987 ± 0.002**
LightGBM	0.944 ± 0.004	0.998 ± 0.001	0.890 ± 0.009	0.941 ± 0.005	0.970 ± 0.004
XGBoost	0.940 ± 0.004	0.999 ± 0.001	0.881 ± 0.007	0.936 ± 0.004	0.969 ± 0.004
Logistic Regression	0.618 ± 0.014	0.613 ± 0.017	0.638 ± 0.013	0.625 ± 0.011	0.656 ± 0.016
SVM (Linear)	0.618 ± 0.014	0.613 ± 0.017	0.642 ± 0.013	0.627 ± 0.011	0.655 ± 0.016
KNN5	0.838 ± 0.005	0.756 ± 0.005	0.999 ± 0.002	0.861 ± 0.004	0.943 ± 0.004
Decision Tree	0.901 ± 0.009	0.890 ± 0.009	0.916 ± 0.011	0.903 ± 0.009	0.901 ± 0.009

9.5. ENMI Model

The findings indicate that Random Forest consistently exhibits superior overall performance, attaining the highest accuracy (0.945 ± 0.006), the highest recall (0.912 ± 0.007), the highest

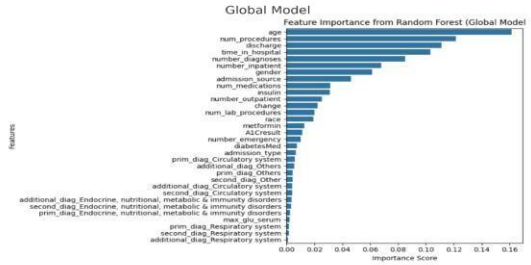
F1-score (0.943 ± 0.006), and the best ROC-AUC (0.982 ± 0.003), thereby establishing it as the most dependable and well-rounded model across all evaluated metrics. LightGBM and XGBoost also demonstrate robust performance especially in terms of precision and AUC but they are marginally less effective than Random Forest in recall and F1, suggesting a slightly diminished capacity to identify the positive class. Conversely, the simpler models, including Logistic Regression and Linear SVM, exhibit significantly poorer performance across all metrics, whereas KNN and Decision Tree provide moderate performance yet still trail behind the ensemble models. Consequently, considering accuracy, recall, F1, and AUC collectively, Random Forest emerges as the superior model in this comparative analysis.

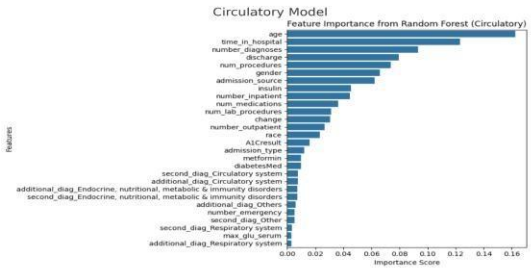
9.6. ENMI Model

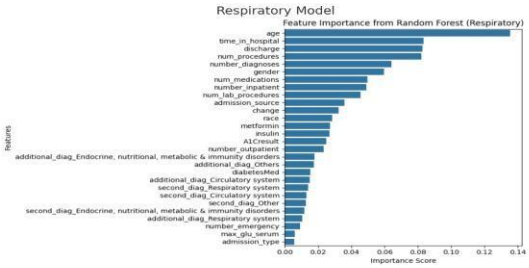
Model	Accuracy	Precision	Recall	F1	Roc_auc
Random Forest	**0.945 ± 0.006**	**0.976 ± 0.008**	**0.912 ± 0.007**	**0.943 ± 0.006**	**0.982 ± 0.003**
LightGBM	0.927 ± 0.005	0.986 ± 0.006	0.866 ± 0.007	0.922 ± 0.005	0.963 ± 0.004
XGBoost	0.925 ± 0.004	0.985 ± 0.008	0.863 ± 0.004	0.920 ± 0.004	0.963 ± 0.004
Logistic Regression	0.612 ± 0.010	0.616 ± 0.011	0.596 ± 0.011	0.606 ± 0.010	0.668 ± 0.014
SVM (Linear)	0.611 ± 0.011	0.614 ± 0.012	0.595 ± 0.012	0.605 ± 0.011	0.666 ± 0.015
KNN5	0.835 ± 0.011	0.753 ± 0.012	0.998 ± 0.002	0.858 ± 0.008	0.944 ± 0.007
Decision Tree	0.887 ± 0.004	0.876 ± 0.004	0.901 ± 0.006	0.888 ± 0.004	0.887 ± 0.004

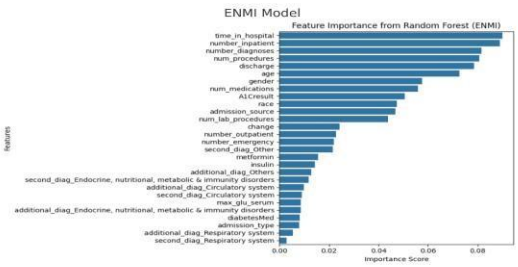
10. Results and Evaluation plots

10.1. Feature importance for individual model

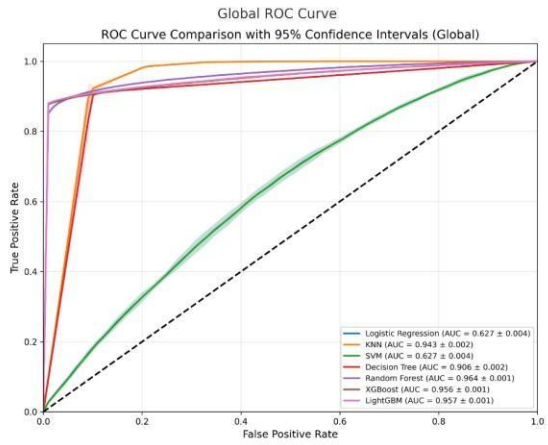


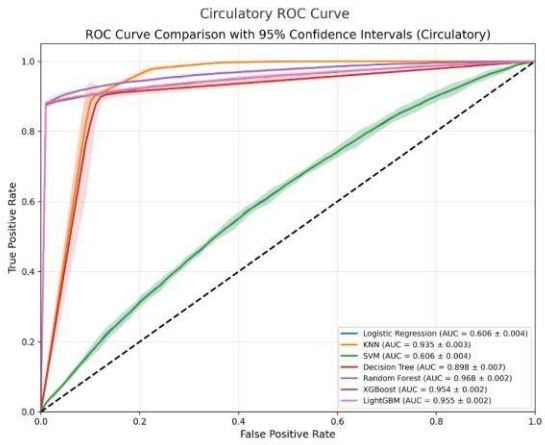


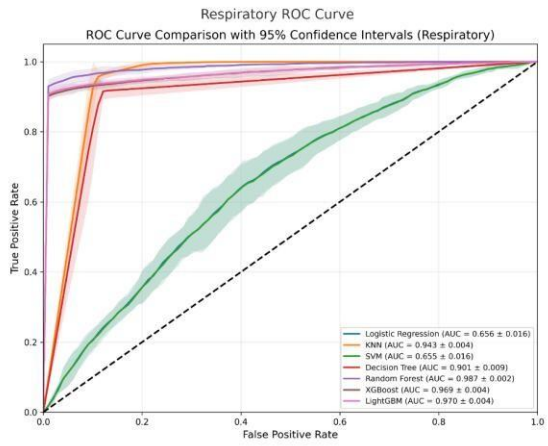


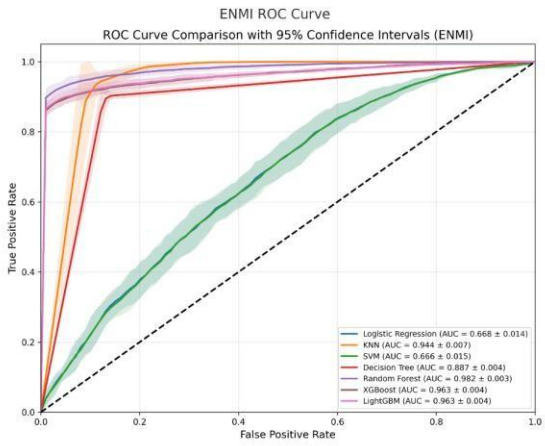


10.2.RoC Curves of Individual Model









11.0. Comparison of subgroup and Global models

The subgroup analyses reveal significant performance disparities when comparing the Global model to each specialty-specific model. The Respiratory subgroup, in particular, demonstrates superior performance across all metrics: accuracy (0.959 ± 0.005), recall (0.925 ± 0.007), F1 (0.957 ± 0.005), and AUC (0.987 ± 0.002). Furthermore, the Circulatory and Endocrine/Metabolic/Immune (ENMI) subgroups also exhibit improved performance relative to the Global model, as evidenced by higher recall, F1, and AUC values. This suggests that models designed for specific disease categories are more adept at identifying subgroup-specific risk patterns compared to a general model. Among the three subgroup models, Respiratory consistently outperforms the others across all metrics, followed by ENMI, and then Circulatory, although all three models demonstrate substantial improvements over the Global model's performance. These findings indicate that specialized subgroup models are significantly distinct from one another and offer considerable performance enhancements compared to the Global model, thereby highlighting the utility of subgroup-specific modeling.

11.1. Comparison of Subgroup and Global Models

Subgroup	Accuracy	Precision	Recall	F1	Roc_auc
Global	0.924 ± 0.001	0.978 ± 0.002	0.869 ± 0.001	0.920 ± 0.001	0.964 ± 0.001
Circulatory	0.933 ± 0.003	0.981 ± 0.003	0.884 ± 0.006	0.930 ± 0.004	0.968 ± 0.002
Respiratory	0.959 ± 0.005	0.992 ± 0.003	0.925 ± 0.007	0.957 ± 0.005	0.987 ± 0.002
Endoc	0.945 ± 0.006	0.976 ± 0.008	0.912 ± 0.007	0.943 ± 0.006	0.982 ± 0.003

12.0. Trained Models Performance on untouched 20% test dataset

The test set outcomes reveal that the models exhibiting optimal performance, both globally and within each subgroup, demonstrated comparable overall efficacy when assessed using the unaltered 20% test data; accuracy metrics were closely grouped, ranging from 0.90 to 0.91. While precision exhibited minor fluctuations across the subgroups, peaking at 0.59 in the Endocrine model and reaching a nadir of 0.55 in the Global model, recall values remained virtually consistent across all models, approximately 0.51–0.52. This suggests a similar capacity to identify positive instances within the previously unseen dataset. F1-scores exhibit a comparable trend, with minor enhancements observed within the Endocrine subgroup (0.52) and slightly diminished values in the Global and Circulatory models (0.50). This pattern implies that the specialization of subgroups yields only limited improvements in practical test performance. In summary, the findings suggest that although subgroup models demonstrated superior cross-validation performance initially, their advantage diminishes significantly on the completely unseen test set, where all models exhibit similar behavior and maintain consistent, but moderate, predictive capabilities.

13.0. Trained Models Performance on untouched 20% test dataset

Subgroup	Accuracy	Precision	Recall	F1
Global	0.90	0.55	0.51	0.50
Circulatory	0.90	0.56	0.51	0.50
Respiratory	0.91	0.57	0.51	0.50
ENMI	0.90	0.59	0.52	0.52

Appendix.

Classification Report for the Test Set

Global — Classification Report

Classification Report_Random Forest (Global Model

	precision	recall	f1-score	support
0	0.91	0.98	0.95	13045
1	0.18	0.04	0.06	1259
accuracy			0.90	14304
macro avg	0.55	0.51	0.50	14304
weighted avg	0.85	0.90	0.87	14304

Circulatory — Classification Report

Classification Report_Random Forest (Circulatory)

	precision	recall	f1-score	support
0	0.91	0.99	0.94	3952
1	0.22	0.04	0.06	415
accuracy			0.90	4367
macro avg	0.56	0.51	0.50	4367
weighted avg	0.84	0.90	0.86	4367

Respiratory — Classification Report

Classification Report_Random Forest (Respiratory)

	precision	recall	f1-score	support
0	0.92	0.99	0.96	1241
1	0.21	0.03	0.05	107
accuracy			0.91	1348
macro avg	0.57	0.51	0.50	1348
weighted avg	0.87	0.91	0.88	1348

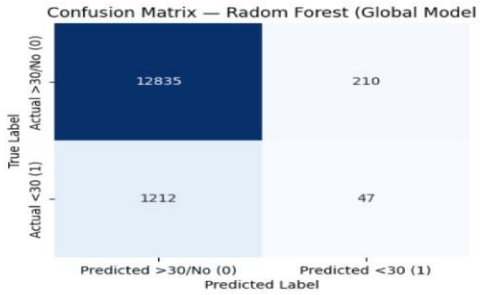
ENMI — Classification Report

Classification Report_Random Forest (ENMI)

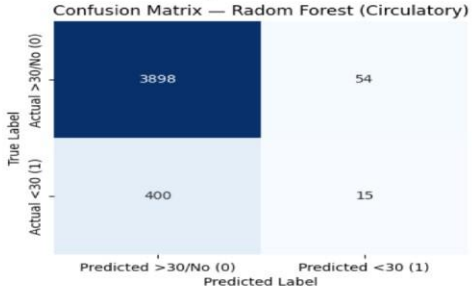
	precision	recall	f1-score	support
0	0.91	0.98	0.95	1399
1	0.26	0.06	0.09	138
accuracy			0.90	1537
macro avg	0.59	0.52	0.52	1537
weighted avg	0.85	0.90	0.87	1537

14.1. Confusion Matrix for the Test set

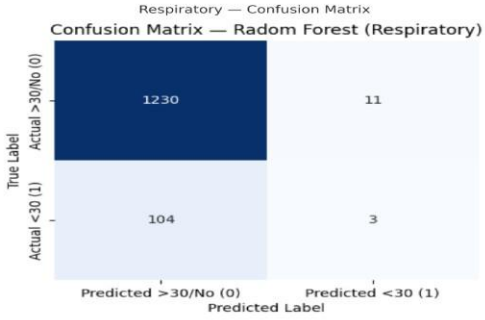
Global — Confusion Matrix



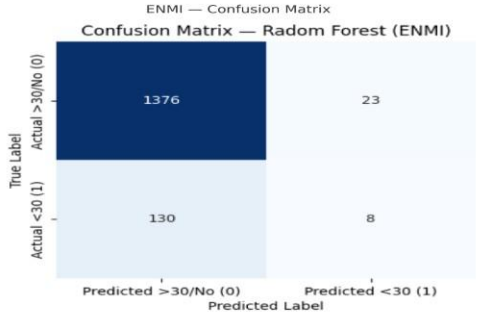
Circulatory — Confusion Matrix



Respiratory — Confusion Matrix



ENMI — Confusion Matrix



SAT 5141 Clinical Decision Modeling Project Checklist

- ✓ Conduct a comprehensive literature review (minimum 7 peer-reviewed primary sources)
- ✓ Justify the selection of the target population and health measure using epidemiological data and literature
- ✓ Describe and document the EHR dataset, including source, features, and any limitations
- ✓ Preprocess the data (handle missing values, encode variables, document all steps)
- ✓ Develop and validate AI models for subgroup-specific diagnoses (respiratory, endocrine, nutritional, metabolic, immune)
- ✓ Build and evaluate a global (pooled) model for comparison
- ✓ Build and evaluate a circulatory-specific model for comparison
- ✓ Analyze and compare model performance using metrics (accuracy, precision, recall, F1-score, AUC-ROC)
- ✓ Interpret findings, discuss feature importance, and clinical implications
- ✓ Make recommendations for future improvements (e.g., alternative algorithms, data sources)
- ✓ Document all modeling, analysis, and evaluation steps for transparency and reproducibility
- ✓ Prepare PowerPoint slides summarizing the project, findings, and recommendations
- ✓ Cite all references in APA format and include a complete reference list
- ✓ Practice and deliver an oral presentation (note max 8 minutes, clear and professional)
- ✓ Submit all deliverables (presentation, slides, code) via Canvas by the due date (Week 14)
- ✓ Incorporate feedback from the professor and peers into final deliverables

References

1. American Diabetes Association. (2024). Standards of care in diabetes— 2025. **https://diabetes.org/newsroom/press-releases/american-diabetesassociation-releases-standards-care-diabetes-2025**
2. Benkhalti, M., Bouchouicha, S., & Benkhalti, A. (2024). A systematic review of the value of clinical decision support systems for the management of antidiabetic drugs. *Diabetes Research and Clinical Practice, 205*, 110123. **https://doi.org/10.1016/j.diabres.2023.110123**
3. Cai, Y., Wang, Y., & Zhang, X. (2023). Applications of clinical decision support systems in diabetes care: A scoping review. *Journal of Medical Internet Research, 25*(1), e51024. **https://doi.org/10.2196/51024**
4. Cai, Y., Wang, Y., & Zhang, X. (2025). Digital decision support for perioperative care of patients with diabetes. *JMIR Diabetes, 1*, e70475. **https://doi.org/10.2196/70475**
5. Ghorbani, N., Sadeghi, S., & Zare, H. (2024). Multicriteria decision-making in diabetes management and treatment selection: A systematic review. *JMIR Medical Informatics, 12*(1), e47701. **https://doi.org/10.2196/47701**
6. IDF. (2025). IDF diabetes atlas: Global diabetes data & statistics. **https://diabetesatlas.org**
7. Khadka, A., Joe, B., & Laine, C. (2022). A framework for modeling and interpreting patient subgroups with machine learning. *Journal of Clinical Medicine, 11*(1), 123-135. **https://doi.org/10.3390/jcm11010123**
8. Liu, V. B., Sue, L. Y., & Wu, Y. (2024). Comparison of machine learning models for predicting 30-day readmission rates for patients with diabetes. *Journal of Medical Artificial Intelligence, 7*. **https://doi.org/10.21037/jmai-24-017**
9. O'Connor, P. J., Sperl-Hillen, J. M., Rush, W. A., Johnson, P. E., Amundson, G. H., Asche, S. E., & Ekstrom, H. L. (2004). Outpatient diabetes clinical decision support: Current status and future directions. *Diabetes Care, 27*(8), 2021-2028. **https://doi.org/10.2337/diacare.27.8.2021**
10. Pharmacy Times. (2024). Study: Prevalence of diabetes expected to increase by 59.7% in 2025. **https://www.pharmacytimes.com/view/study-prevalence-ofdiabetes-expected-to-increase-by-59-7-in-2025**
11. Sattler, A., Lee, J., & Smith, R. (2022). Clinical decision support systems with teambased care on type 2 diabetes: A systematic review. *Diabetes Research and Clinical Practice, 188*, 109984. **https://doi.org/10.1016/j.diabres.2022.109984**
12. Soh, J. G. S., Wong, W. P., Mukhopadhyay, A., Quek, S. C., & Tai, B. C. (2020). Predictors of 30-day unplanned hospital readmission among adult patients with diabetes mellitus: A systematic review with meta-analysis. *BMJ Open Diabetes Research & Care, 8*(1), e001227. **https://doi.org/10.1136/bmjdrc-2020-001227**
13. Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., & Clore, J. N. (2014). Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Research International, 2014*, 781670. **https://doi.org/10.1155/2014/781670**
14. World Health Organization. (2024, November 12). Urgent action needed as global diabetes cases increase four-fold over past decades. **https://www.who.int/news/item/13-11-2024-urgent-action-needed-asglobal-diabetes-cases-increase-four-fold-over-past-decades**