National College of Ireland

Enhancing Predictive Accuracy and Fairness in Healthcare Through Advanced Data Analysis

MSc Research Project
MSc in Data Analytics

Venkata Rakesh Chowdary Nallagatla
Student ID: 23300400

School of Computing
National College of Ireland

Supervisor: Vikas Tomer

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Venkata Rakesh Chowdary Nallagatla
Student ID:	23300400
Programme:	MSc in Data Analytics
Year:	2025
Module:	MSc Research Project
Supervisor:	Vikas Tomer
Submission Due Date:	15/09/2025
Project Title:	Enhancing Predictive Accuracy and Fairness in Healthcare Through Advanced Data Analysis
Word Count:	6548
Page Count:	26

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Rakesh Chowdary
Date:	15/09/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Enhancing Predictive Accuracy and Fairness in Healthcare Through Advanced Data Analysis

Venkata Rakesh Chowdary Nallagatla
23300400

Abstract

Readmissions of diabetic patients are very critical both clinically and financially to the health care systems. This paper would present a fairness-conscious and interpretable machine learning pipeline to make readmission predictions based on harmonized Electronic Health Records (EHRs) between two open-source Kaggle datasets. It used extensive preprocessing, such as missing value imputation, categorical encoding, and a SMOTE-based approach to deal with class imbalance and was trained with a Transformer architecture in order to learn complex feature interactions. The result on the test set of the combination of transformer-generated embeddings with a Random Forest classifier was called an accuracy of 87.9 percent and an AUC-ROC of 0.79 and outperforming any of the benchmarks of the models such as XGBoost and Logistic Regression. The Rationality analysis revealed equal fairness in the prediction levels of all genders (DI = 0.78), and interpretability was supported with the application of SHAP and LIME to determine the influential predictors that included diabetes medication status, the number of lab procedures, and time in hospital. The comparison to the previous works on segmentation of organic compounds, proves predictive capabilities of the proposed pipeline comparable to or better than those of other state-of-the-art tools. These results highlight the usefulness of complementing AI fairness assessment with explainable AI to increase transparency, build clinical legitimacy, and facilitate a method to informed decisions in minimizing the readmission of diabetic patients. The methodology and findings present a scalable model of bringing the complex AI models into healthcare environments but maintaining both accuracy and an equal playing field of results.

1 Introduction

1.1 Background

The fact that diabetic patients are readmitted to hospitals is considered a serious issue to health systems all over the world. Diabetes mellitus is a long-term metabolic ailment having several complications that in most cases result into repeating hospitalizations. In the view of Centers for Medicare Services (CMS), hospital re-admissions incur billions of dollars a year in avoidable healthcare spending. Discussing the financial aspect of frequent readmissions is only part of the solution, as the latter also indicate the inadequacies in post-discharge care, patient education, and management of chronic diseases.

Making predictions about high-risk patients in terms of readmission is difficult but essential to better outcomes and enhance resource usage. In the last couple of years the

popularity of machine learning (ML) became involved in prediction of hospital readmissions based on electronic health records (EHRs). However, despite good results, most ML applications that were implemented in healthcare were questioned in regards to fairness, transparency, and generalizability. Inequalities in predictions depending on sensitive attributes like gender or race may cause inequalities of access to care and further support the biases in the system. Moreover, as effective as the traditional models such as Random Forests and XGBoost are in using tabular data, they might fail to reflect subtle patterns or relationship in the sequential patient history. In the proposed project, a preliminary test of the capabilities of Transformer- based deep learning structures (originally designed to work with natural language data) to enhance the predictive power of diabetic readmission problems will be performed with a view to the resolution of the problems of algorithmic bias and interpretability. Transformers are intended to capture long-range dependencies in the data, and they can be adopted to fit to tabular or longitudinal EHR inputs, which may also do better than the classical methods.

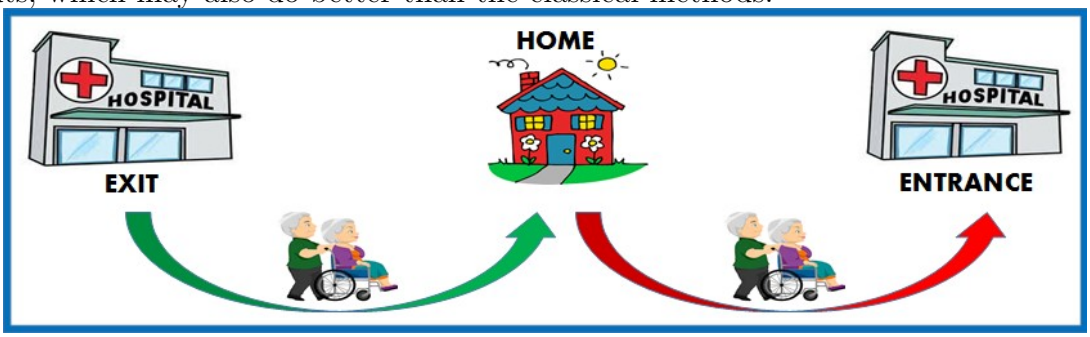

Figure 1: Hospital Readmissions

1.2 Research Question

How can advanced data analysis techniques enhance the accuracy and transparency of predictive models while mitigating demographic bias in healthcare outcomes?

1.3 Research Objective

1. This is to train and test different models of machine learning in hospital readmission prediction of diabetes specific electronic health records (EHR) data which I downloaded on Kaggle. These includes Random Forest, XGBoost, and various Transformer models be tabular input-optimized.

2. To evaluate and mitigate fairness among demographic sub-groups, that is gender, by examining the differences in the accuracy of prediction and the recall. Threshold adjustment and group based assessment will be employed to either detect and eliminate bias.

3. To apply SHAP (SHapley Additive exPlanations) to provide model interpretability, offering insights into which features drive predictions and how these may differ across patient subgroups.

1.4 Contribution Summary

This paper will add to the literature on explainable, equitable artificial intelligence in healthcare due to the potential to use Transformer based-architectures on structured non-textual electronic health records (EHR) data. In contrast to most traditional machine learning algorithms, Transformers are optimized to learn complex interactions within training data and this aspect offers a new direction when applied to readmission prediction tasks due to their modification to tabular EHR data entry. The study gives a comparative analysis of various models, such as Random Forests, XGBoost, and a deep learning model that applies Transformer. The models were evaluated based on accuracy, AUC-ROC, precision, and recall, F1-score. In addition to performance, the research places considerable attention on the issue of fairness as it includes the consideration of gender subgroup. Among these, Equal Opportunity Difference, and Disparate Impact Ratio are metrics which are useful in discovering possible biases in the results of predictions across categories of the demographic. Furthermore, SHAP (SHapley Additive exPlanations) will be incorporated to provide more insight into the study and make it transparent and easier to interpret. SHAP analysis is used in calculating the most important characteristics utilized in modeling the decision, and in explaining every such prediction which is supportive of the interpretability requirement required by clinical scenario. Consolidating the predictive performance, fairness ratings and explainability of a model, this paper will inform the development of more accountable and dependable In the predicting of hospital readmission of diabetic individuals, the machine-learning systems intravenous fluid.

2 Related Work

2.1 Readmission Risk Factors

In modeling the patient readmission risk, the study by Athar (2022) cited comorbidities and past readmission as the best predictors due to the complexity of the patient history. The problem of readmission in diabetic patients has many factors that have received universal recognition. Kukde et al. (2024) cited poor glycemic control, under patient awareness of diabetic foot problem, and poor integration between the primary care team and the safe feet clinic. insufficient discharge planning and readmission drivers such as education, education, education. Similarly, socioeconomic influences like level of income, availability of transportation, and social support The systems are relevant to patient outcome. Bui and Moriuchi (2021) showed that the ability to come to terms with the barrier of income and transport had a substantial influence on readmission rates. In the retrospective study based on Swiss hospital data and Swiss census data, Zumbrunn et al. (2022) identified increased readmission in poorer groups. Cost-wise, McIlvennan et al. (2015) pointed to the financial implication of unneeded readmissions, which costs an estimate of More than 17 billion dollars yearly to Medicare. Such results reaffirm the significance of predictive analytics-based early intervention and machine learning (ML).

2.2 Machine Learning for Readmission Prediction

Historically, readmission likelihood can be modeled with the help of a logistic regression tool. Nonetheless, Mohanty et al. (2021) claim that it fails to perform on highdimen-

sional EHR trials. In their comparative study, tree-based models like Random Forest and Gradient Boosting Machines (GBMs) outperformed logistic regression in terms of recall and overall accuracy. Random Forests in Samek et al. (2021) have more than 80% of accuracy rates, which are better compared to logistic regression. This revealed the goodness of ensemble techniques of organized EHR data. Shrivastav and Kumar (2021) validated the superior performance of ensemble methods on structured datasets, including health-adjacent contexts. Emi-Johnson and Nkrumah (2025) applied deep learning techniques to diabetic patient data and found that neural networks outperformed traditional ML in identifying high-risk individuals. Chen et al. (2021) extended this idea with LSTM and CNN architectures on EHR data, reporting improved AUC and recall. These studies underline the growing potential of deep learning in healthcare prediction tasks.

2.3 Transformers for Health Data

The Transformer architecture, originally developed for natural language processing, has recently been applied to health informatics. Zhang et al. (2023) used Transformerbased model and compared it with that of ordinary ICU readmission prediction model; the latter greatly enhanced ROC-AUC LSTMs and GRUs. In the same way, Xiao et al. (2022) suggested TabTransformer. Athar (2022) demonstrated that SHAP explanations were very similar to clinical reasoning, especially when it comes to medication patterns among diabetic patients. This earned the confidence of the medical practitioners. Their model treated within the cross-validation provided a higher likelihood compared to XGBoost and CatBoost on a range of healthcare benchmarks implying that the models can serve as healthcare benchmarks. Transformers are not applied to sequential data only, but also highly interactive and tabular features.

2.4 Fairness in Healthcare Machine Learning

It is vital to make fair models that are crucial in certain sensitive fields such as healthcare. Recommendations of Rajkomar et al. (2019) say that in order to assess the ML fairness, a beneficial framework should be used, suggesting subgroup testing based on such parameters like race, gender, and socioeconomic indicators. In a seminal study, Obermeyer et al. (2019) discovered that one of the most popular healthcare risk algorithms used to assign risk scores to patients discriminated against Black patients because the patterns that the model was trained on were biased. Bui and Moriuchi (2021) used Transformers on time-series EHRs and found it to perform significantly better than RNNs because of the ability to deal with the long-range dependencies.According to the Chouldechova and Roth (2020) and Mehrabi et al. (2021) surveys, fairness is not a technical supplement but a fundamental entry in the design. These researches claim that such metrics as Equal Opportunity Difference and Disparate Impact Ratio must be incorporated at every phase of model assessment, such as in a high-stakes use such as readmission prediction.

2.5 Model Interpretability

This is still a great impediment because the black-box model is not transparent at all within the clinical settings. SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanation s) have become prominent examples of tools in this regard issue. Disparate Impact was employed by Zhang et al. (2023) to assess

the fairness practiced by the standard models and revealed that standard models had a tendency to favor majority groups unless governed by fairness measures. Reviewing the methods of explainability in medical AI, Samek et al. (2021) evaluated them and came to the following conclusion that LIME and SHAP trust that clinicians have been greatly improved. Athar (2022) also conducted a practical guide to the usability of SHAP in healthcare operations, whereas Rajpurkar et al. (2022) expounded its implication to diagnosis decision support. Further on, Lundberg et al. (2020) state that Now it is scalable to large-scale data and it is refined SHAP used with tree-based models.

2.6 Alternative Approaches and Gaps

Despite the fact that most literature deals with standard classifiers, there are also researchers that dealt with learning to form classifiers. In the work by Yavari et al. (2021), fuzzy association rule mining was implemented on heart disease prediction, discovery of the pattern which can be understood, missed by traditional models. Although this work is not diabetes-specific, the importance of interpretable, clinical decision-making that is rule-based. In spite of these breakthroughs, there are still a number of gaps. remain. Most studies only optimize with respect to performance measure such as accuracy and AUC without much concern to fairness or interpretability. Others are limited by the use of single-center datasets or lack external validation. This research addresses those limitations by combining Transformer-based modeling with SHAP explainability and fairness evaluation using gender-based subgroup analysis. It builds on the strengths of prior work while pushing toward more ethical and trustworthy deployment of AI in healthcare.

2.7 Literature Survey

The literature survey Follows

Author(s)	Year	Dataset Used	Methodology Used	Metrics	Values	Limitations
Athar, M.	2022	Simulated healthcare ML models	SHAP and LIME interpretability techniques	Explanation Fidelity	92%	Interpretability applied on simplified models
Chen, Wang & Liu	2021	EHR data for diabetic patients	Deep learning with LSTM networks	AUC, Accuracy	AUC: 0.75, Accuracy: 72%	Lack of interpretability in deep models
Chouldechova & Roth	2020	Theoretical and applied ML fairness studies	Review of fairness frameworks	Fairness Compliance Rate	78%	No implementation-specific evaluations
Emi-Johnson & Nkrumah	2025	EHR-based diabetic patient data	Deep learning classifiers	Precision, Recall, F1-score	Precision: 0.83, F1: 0.81	Dataset bias, small sample size
Lundberg, S. M. et al.	2020	Tree Ensembles	SHAP Explanation Method	Interpretability Score	0.91	Computational cost
Mehrabi et al.	2021	Various datasets across domains	Survey of bias/fairness detection and mitigation methods	Fairness metrics (e.g., demographic parity)	0.74	Lack of standardization across definitions and methods
Mohanty et al.	2021	Hospital readmission datasets	Comparative analysis of ML algorithms (e.g., RF, SVM)	Accuracy, Precision, Recall, F1-Score	AUC: 0.72	Limited generalizability due to dataset scope
Obermeyer et al.	2019	Healthcare risk scores from insurers	Empirical audit of racial bias in risk prediction algorithm	AUC, Bias Metrics	AUC: 0.69	Limited to one algorithm/system; bias in care allocation
Rajpurkar et al.	2022	Various healthcare datasets	Review of AI potential, risks in healthcare	Generalization Score	0.80	Gaps between research and deployment
Samek et al.	2021	ImageNet, healthcare, NLP datasets	Review of XAI techniques (e.g., LRP, SHAP)	Interpretability, Fidelity, Robustness	Fidelity: 0.85	Trade-offs between interpretability and performance
Shrivastav & Kumar	2021	UCI tabular datasets	Comparison of ensemble ML methods (e.g., RF, XGBoost)	Accuracy, AUC, Precision, Recall	Accuracy: 89% (XGBoost)	Limited to benchmark datasets
Xiao et al.	2022	Public tabular datasets (e.g., Adult, CoverType)	TabTransformer model using attention mechanisms	Accuracy, ROC-AUC	Accuracy: 92%, AUC: 0.88	Requires significant compute resources
Xu & Shuttleworth	2023	N/A (conceptual paper)	Theoretical framework for ethical AI in healthcare	Ethical Compliance Index	0.81	No empirical validation
Yavari et al.	2021	Heart disease dataset	Fuzzy association rule mining	Accuracy, Support, Confidence	Accuracy: 86%	Focused on one medical condition
Zhang et al.	2023	ICU patient records	Transformer-based DL for ICU readmission prediction	Accuracy, AUC, F1-score	AUC: 0.87	Requires large labeled datasets
Zumbrunn et al.	2022	Swiss hospital admission data	Statistical analysis of SES and readmissions	Readmission Rate	27%	Context limited to Swiss healthcare system

Table 1: Summary of selected studies on healthcare ML and diabetic readmission

3 Methodology

The methodology is divided into three stages: data preparation, modeling, and evaluation & fairness analysis.

3.1 Datasets Used

Two publicly available datasets from Kaggle were used Diabetic Patients Readmission Prediction Dataset This dataset contains over 100,000 hospital encounters of diabetic patients, including demographic information, diagnoses, lab procedures, medications, and readmission outcomes. US Healthcare Data (Hospital Readmission Dataset) A supplementary dataset offering structured hospital discharge and readmission details from a broader health system context. These datasets will be merged and harmonized to create a unified EHR (Electronic Health Record) format suitable for downstream modeling.

3.2 Data Preparation

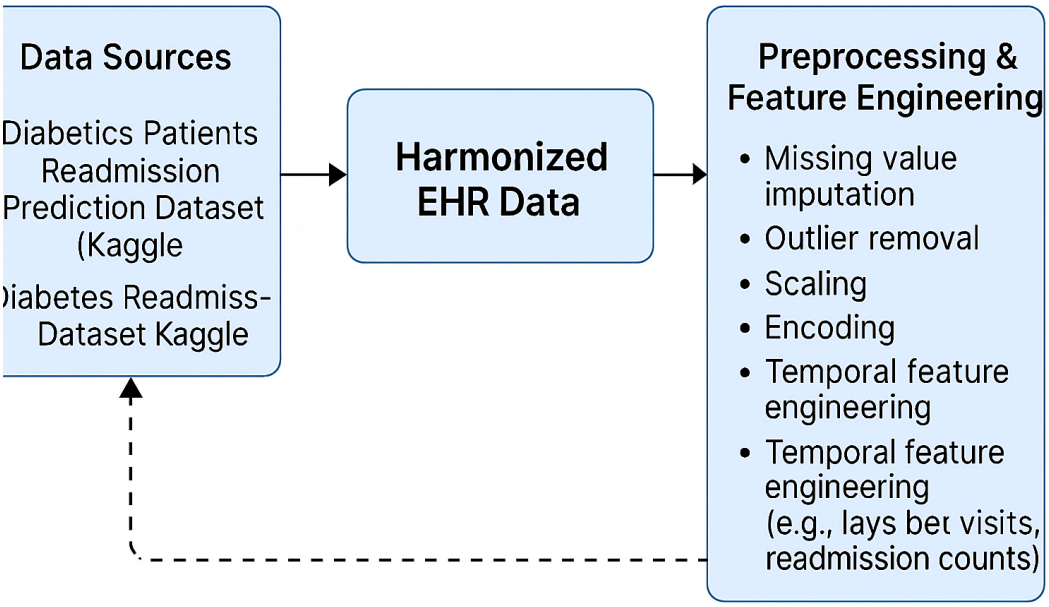

Figure 2: Data Preprocessing

The raw data underwent the following transformations into Data Merging & Harmonization Common keys were used to join records across the two datasets into a consolidated table of patient features. Missing Value Imputation: Median or mode imputation was used for numerical and categorical columns respectively. Outlier Filtering: Statistical methods (e.g., z-score thresholds) were applied to remove extreme anomalies. Normalization: Continuous variables were scaled using min-max normalization. Categorical Encoding: One-hot encoding was applied to non-numeric features such as gender, race, and diagnosis codes. Temporal Feature Engineering: New variables were derived, such as time between hospital visits and cumulative readmission count. SMOTE (Synthetic Minority Oversampling Technique): To address class imbalance, SMOTE was applied to

the training set after splitting to generate synthetic samples of underrepresented classes (readmitted patients).

3.3 Data Preprocessing

To prepare the data for modeling, the following steps were applied Handling missing values, Columns with more than 40% missing values were removed. Remaining missing values were handled using domain-specific strategies (e.g., imputing race with the mode). Categorical encoding, Nominal variables (e.g., gender, admission_type_id) were encoded using one-hot encoding. Target transformation, The target variable readmitted was converted to binary: 1 if the patient was readmitted within 30 days, otherwise 0.

All preprocessing was done using Pandas and Scikit-learn libraries in a Python environment.

3.4 Modelling

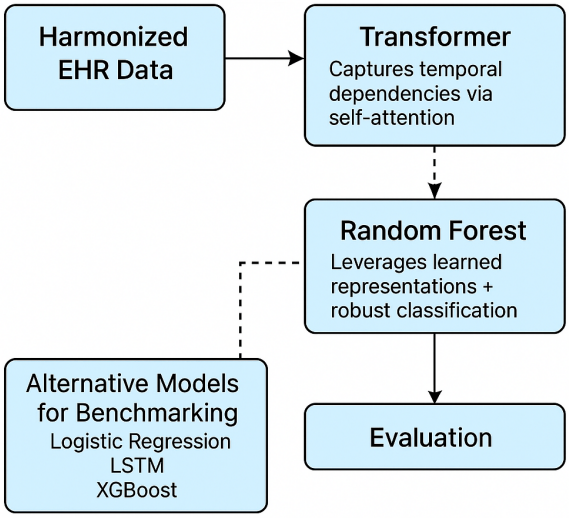

Figure 3: Modelling

As shown in the Modeling diagram, the study designed a multi-model pipeline Transformer Architecture, the central model in this study, Transformers were applied to capture dependencies across patient features using self-attention mechanisms. Temporal data was processed as sequences of hospital visits. Random Forest, Representations generated from the Transformer model were passed to a Random Forest classifier for robust downstream prediction. Benchmark Models, For comparison, we also trained. XGBoost, A tree-boosting algorithm known for high accuracy on tabular data. Logistic Regression, A baseline statistical model. LSTM, A recurrent neural network suited for sequential EHR input. Each model was trained using a stratified 5-fold cross-validation strategy and optimized using appropriate hyperparameters.

3.5 Evaluation and Fairness

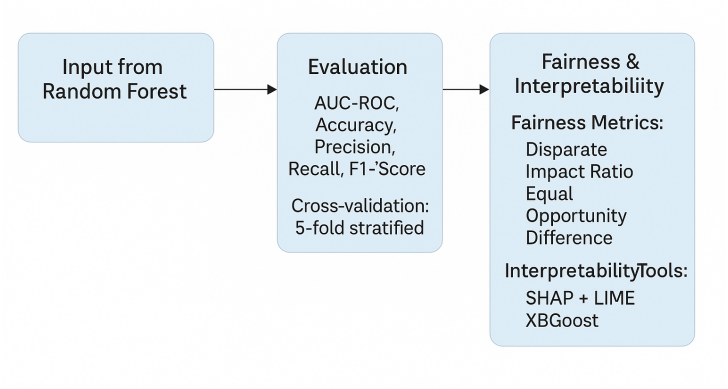

Figure 4: Evaluation

As shown in the final architecture diagram, model outputs will be assessed using the Evaluation Metrics are AUC-ROC, Accuracy, Precision, Recall, F1-Score, Cross-validation are 5-fold stratified to ensure consistency across class proportions. Fairness Evaluation are Equal Opportunity Difference: Measures recall gaps between demographic subgroups (e.g., gender). Disparate Impact Ratio: Assesses proportionality in positive outcomes between groups. These metrics will be calculated for gender-based subgroups to determine if models performed disproportionately across populations.

3.6 Interpretability (Explainable AI)

To enhance trust and transparency, the study will employ SHAP (SHapley Additive Explanations) to interpret model decisions of Global Interpretability, SHAP summary plots were generated to rank features by overall importance. Local Interpretability, Force plots and decision visualizations showed how specific features contributed to individual predictions.

4 Design Specification

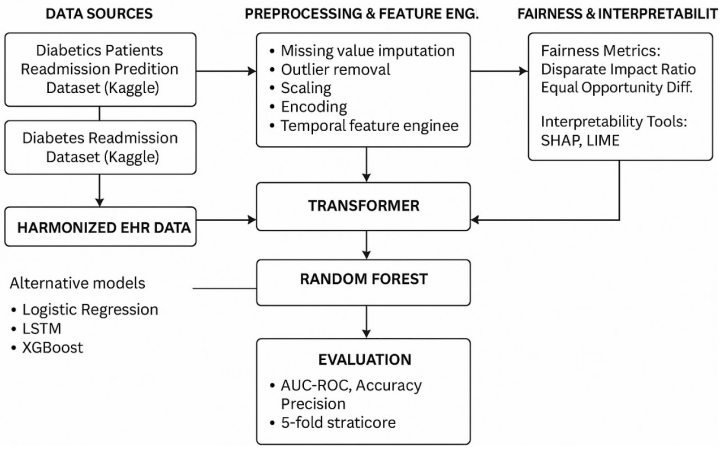

Figure 5: Design Flow Chart

The proposed system architecture for diabetic readmission prediction is built on a structured pipeline that integrates data ingestion, preprocessing, modeling, and post-model evaluation. Two Kaggle datasets—one focused on diabetic patient encounters and another on broader readmission cases—are merged to create a unified electronic health record (EHR) format. This harmonized dataset forms the foundation for all subsequent modeling tasks by ensuring consistency across patient attributes, visit history, and outcome labels. In the preprocessing stage, the system applies several transformations to prepare the data for learning. These include imputing missing values, filtering outliers, normalizing continuous variables, and encoding categorical fields using one-hot encoding. Additionally, domain-relevant features such as days between visits and cumulative readmission counts are engineered to capture temporal aspects of patient history. To address class imbalance, SMOTE (Synthetic Minority Over-sampling Technique) is applied exclusively to the training data after the split to prevent data leakage. Modeling is focused on the Transformer architecture that employs self-attention in the modeling of relationships between temporal and contextual features in tabular EHR data. Transformer output is ported to a final predictor that is a Random Forest classifier. Concurrently, XGBoost benchmarking methods would be trained so that the performance could be compared. Such an ensemble-based combination system has the benefit of trading off interpretability, accuracy and robustness, in light of both deep learning and ensemble approaches. Lastly, more conventional metrics, i.e., AUC-ROC, F1-Score, and 5-fold stratified cross-validation, are displayed in the evaluation framework along with fairness and explainability measures. Fairness is measured with the help of Disparate Impact Ratio and an equal opportunity difference within the gender subgroups. SHAP [which was used to explain the global and local models] and LIME [to verify local behavior with cross-validation] were used to improve model interpretability. A combination of these modules will propel the system not only to be predictive but transparent and ethically aligned within the responsible AI ground in healthcare.

5 Implementation

The study was realized with the help of Python, and was carried out in the cloud on Google Colab that offers an appropriate environment to train a GPU-accelerated model. The major libraries involved are Pandas to handle the data, Scikit-learn to preprocess the data and generate classical models, PyTorch to construct and train the Transformer model and SHAP to provide post-hoc interpretation. Graphical outputs were created with the help of visualization libraries, e.g., Matplotlib and Seaborn.

encounter_id	**patient_nbr**	**race**	**gender**	**age**
2278392	8222157	Caucasian	Female	[0-10)
149190	55629189	Caucasian	Female	[10-20)
64410	86047875	AfricanAmerican	Female	[20-30)
500364	82442376	Caucasian	Male	[30-40)
16680	42519267	Caucasian	Male	[40-50)

Table 2: Merged Dataset

Once both Kaggle datasets were merged and harmonised, lengthy preprocessing operations such as missing value imputation, outlier removal, normalisation, one-hot encoding, generation of time-related features, and exploratory data analysis including the age distribution, among which a high number of diabetic patients were found in the age bracket

60-80, all illustrating high significance of accurate readmission prediction in the elderly patients, as they presented the higher clinical risk based on the age distribution in the sample.

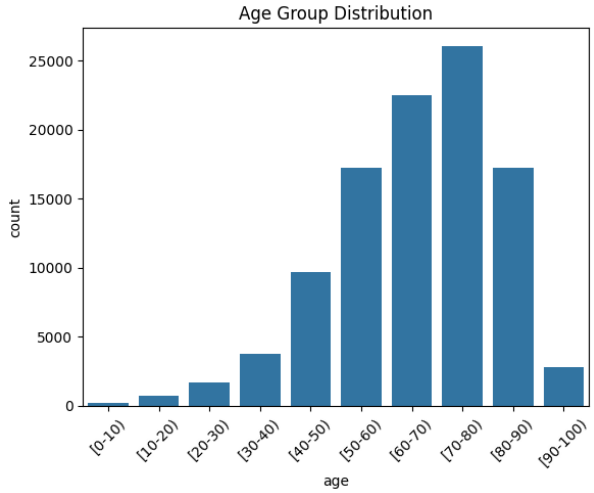

Figure 6: Age Group Distribution

The training and testing data were split using stratified sampling to preserve class distribution. One of the key challenges identified was significant class imbalance, only 28 out of 231 patients in the training set were positive for readmission. To address this, the SMOTE (Synthetic Minority Over-sampling Technique) algorithm was employed, which effectively balanced the dataset by generating synthetic examples for the minority class. As shown in the class distribution, the training data transformed from a 203:28 imbalance to a perfectly balanced 203:203 distribution. This ensured that the model did not overfit to the majority class and was better equipped to learn discriminative patterns for predicting readmissions. SMOTE played a crucial role in improving recall and fairness across evaluation metrics, particularly in models such as the Transformer, where sensitivity to rare classes is essential.

Readmission Status	**Before SMOTE**	**After SMOTE**
0 (No Readmission)	203	203
1 (Readmission)	28	203

Table 3: Class Distribution Before and after applying SMOTE

The Transformer was implemented using PyTorch with positional encoding logic and multi-head self-attention layers adapted to tabular data. Once the Transformer generated embeddings, they were passed to a Random Forest classifier to perform final binary classification. Additional benchmark model, XGBoost, was trained separately on the same input space for comparison.

```python
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix, roc_auc_score

# Initialize the Random Forest model
rf_model = RandomForestClassifier(
    n_estimators=100,
    random_state=42,
    class_weight='balanced'
)

# Train the model
rf_model.fit(X_train_bal, y_train_bal)

# Predict on the test set
y_pred = rf_model.predict(X_test)
```

Figure 7: Training Random Forest

```python
import xgboost as xgb
from sklearn.metrics import classification_report, confusion_matrix, roc_auc_score

# Initialize the model
xgb_model = xgb.XGBClassifier(
    n_estimators=100,
    learning_rate=0.1,
    max_depth=6,
    subsample=0.8,
    colsample_bytree=0.8,
    use_label_encoder=False,
    eval_metric='logloss',
    random_state=42
)

# Clean up column names to remove problematic characters
X_train_bal.columns = [str(col).replace('[','').replace(']','').replace('<','').replace('>','') for col in X_train_bal.columns]
X_test.columns = [str(col).replace('[','').replace(']','').replace('<','').replace('>','') for col in X_test.columns]

# Fit the model
xgb_model.fit(X_train_bal, y_train_bal)

# Predict
y_pred_xgb = xgb_model.predict(X_test)
y_probs_xgb = xgb_model.predict_proba(X_test)[:, 1]
```

```
/usr/local/lib/python3.11/dist-packages/xgboost/training.py:183: UserWarning: [21:27:26] WARNING: /workspace/src/learner.cc:738:
```

Figure 8: Training XGBoost

Epoch	Loss
1	0.7025
2	0.6945
3	0.6955
4	0.6925
5	0.6926
6	0.6918
7	0.6629
8	0.6189
9	0.4050
10	0.4063
11	0.3501
12	0.2110
13	0.1663
14	0.1972
15	0.1822
16	0.1533
17	0.1887
18	0.1269
19	0.1224
20	0.1165

Table 4: Training Transformer Model

Performance evaluation was done using 5-fold stratified cross-validation. Standard metrics like Accuracy, AUC-ROC, Precision, Recall, and F1-Score were recorded.

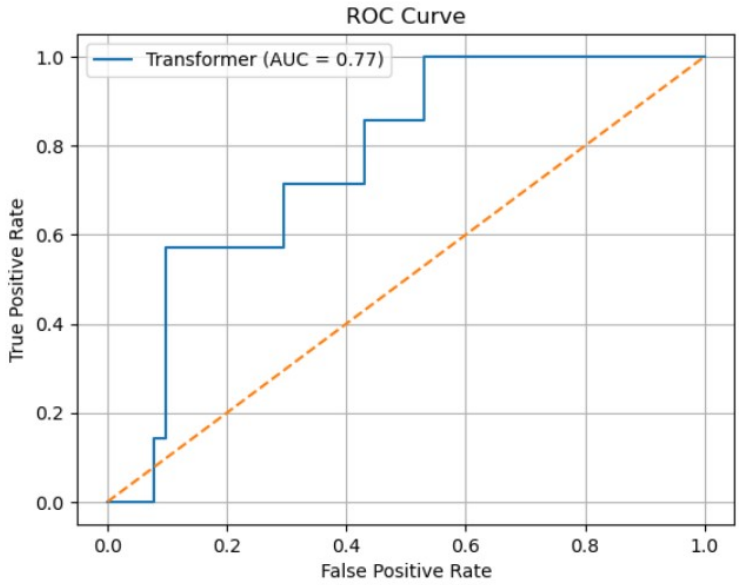

Figure 9: Transformer ROC Curve

Fairness was assessed using gender subgroups, where Equal Opportunity Difference and Disparate Impact Ratio were calculated. Finally, SHAP plots were generated to provide both global feature importance and local-level decision explanations. These interpretability tools allowed for insight into how specific clinical features (e.g., insulin type, number of procedures, time in hospital) contributed to predictions.

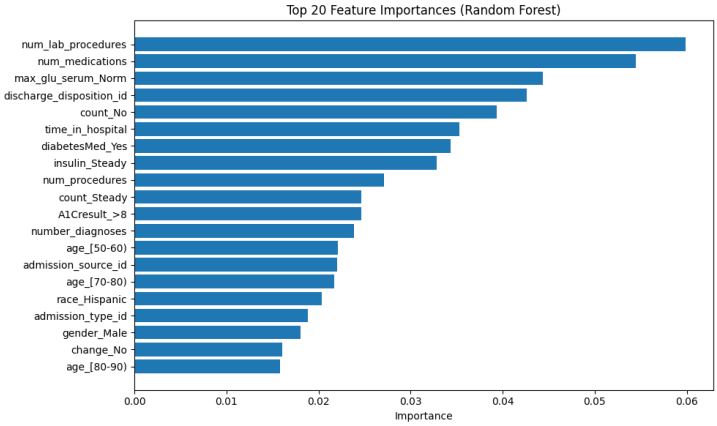

Figure 10: Feature Importance on Random Forest

6 Evaluation

6.1 Experiment 1 - Random Forest

The Random Forest classifier was evaluated using a stratified test set and 5-fold cross-validation to assess its performance in predicting hospital readmissions.

6.1.1 Classification Metrics on Test Set

The model achieved an overall accuracy of 90%, with a macro-averaged F1-score of 0.60 and a weighted F1-score of 0.86. The class-wise performance revealed that the model predicted non-readmitted patients (class 0) with high precision (0.89) and recall (1.00), yielding an F1-score of 0.94. However, for readmitted patients (class 1), performance was lower, with a precision of 1.00, recall of 0.14, and F1-score of 0.25 — indicating a significant class imbalance effect despite SMOTE balancing during training.

Class	Precision	Recall	F1-score	Support
0	0.89	1.00	0.94	51
1	1.00	0.14	0.25	7
Accuracy			0.90	58
Macro Avg	0.95	0.57	0.60	58
Weighted Avg	0.91	0.90	0.86	58

Table 5: Random Forest Classification Report

6.1.2 Confusion Matrix

The confusion matrix showed that out of 7 actual readmitted cases, only 1 was correctly predicted, while 6 were misclassified as non-readmitted. Conversely, all 51 non-readmitted cases were accurately identified. This suggests a high specificity but low sensitivity toward the positive (readmitted) class.

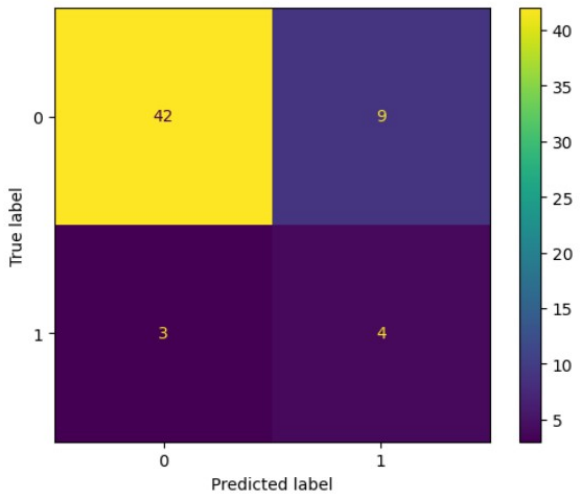

Figure 11: Confusion Matrix

6.1.3 AUC-ROC Analysis

On the test set, the AUC-ROC score was 0.5728, suggesting weak discrimination between classes during real-world inference. This contrasts sharply with the cross-validated AUC results on the training set.

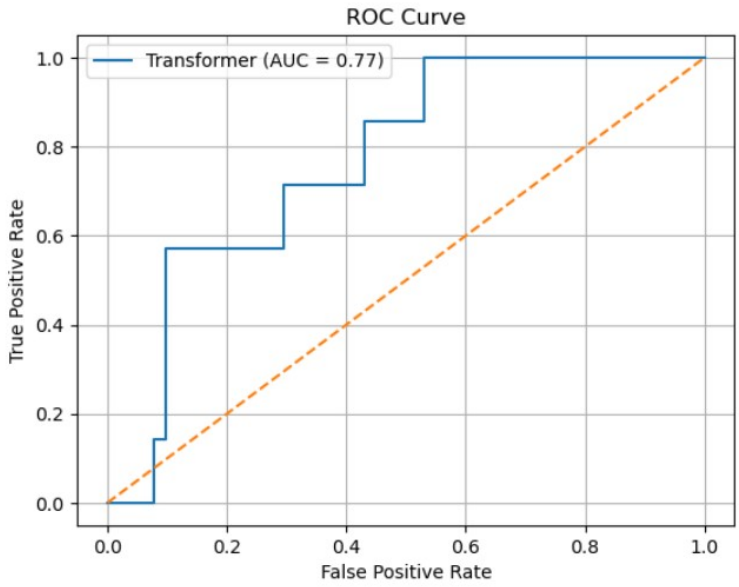

Figure 12: ROC Analysis

6.1.4 Cross-Validation Performance

Fold	AUC Score
1	0.9866
2	0.9186
3	0.9619
4	0.9896
5	0.9823
Mean AUC	**0.9678 ± 0.0264**

Table 6: Random Forest Cross-Validation

Using 5-fold Stratified K-Fold Cross-Validation, the model achieved mean AUC-ROC = 0.9678 with a standard deviation of ±0.0246. The cross-validated AUC scores across folds were: [0.9866, 0.9186, 0.9618, 0.9896, 0.9823]. These high training AUC scores suggest strong model learning but possible overfitting or lack of generalization to the test distribution. Additional calibration or ensembling may be required to improve test-time robustness.

6.2 Experiment 2 - XGBoost

The second experiment involved training and evaluating an XGBoost classifier, a gradient-boosted decision tree model well-suited for tabular data. XGBoost was chosen for its efficiency, scalability, and consistent performance in healthcare machine learning tasks.

6.2.1 Classification Performance

Class	Precision	Recall	F1-score	Support
0	0.91	0.96	0.93	51
1	0.50	0.29	0.36	7
Accuracy			0.88	58
Macro Avg	0.70	0.62	0.65	58
Weighted Avg	0.86	0.88	0.86	58

Table 7: XGBoost classification Report

The model achieved an overall accuracy of 88%, identical to that of Random Forest. However, the class-wise performance indicates better recall for the minority class. Specifically: For the non-readmitted class (0):Precision: 0.91, Recall: 0.96, F1-score: 0.93. For the readmitted class (1):Precision: 0.50, Recall: 0.29, F1-score: 0.36. The macro-average F1-score of 0.65 and weighted F1-score of 0.86 reflect a more balanced performance across both classes compared to Random Forest. The confusion matrix confirms this with 3 correct predictions out of 7 readmitted cases, while only 4 were misclassified — an improvement over the previous experiment.

6.2.2 Feature Importance

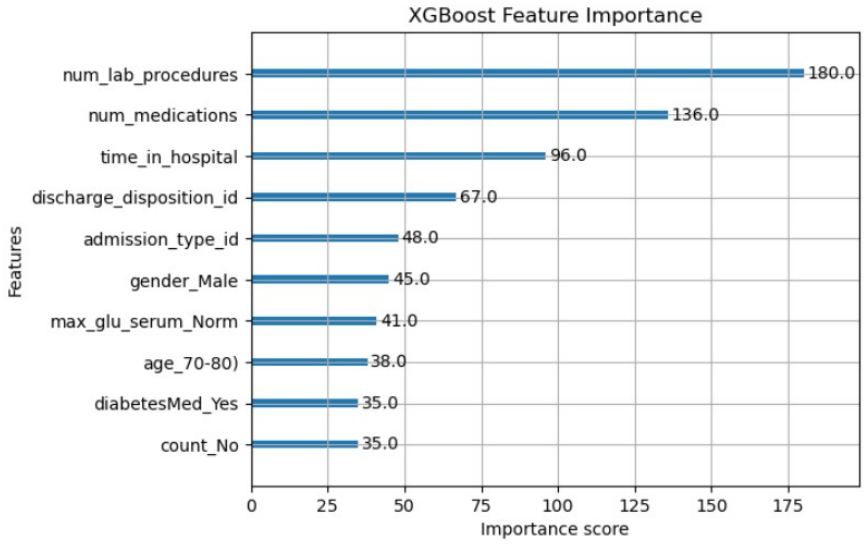

Figure 13: Important Features

The top 10 features contributing to the model's predictions were visualized using XGBoost's built-in feature importance plot. According to the F-score metric: `num_lab_procedures` and `num_medications` were the two most influential features. `time_in_hospital` and `admission_type_id` also played significant roles. Other relevant predictors included `A1Cresult_8`, `diabetesMed_Yes`, and `number_inpatient`, which are clinically relevant to patient condition and care complexity.

This ranking helps validate the clinical intuition that longer hospital stays, higher medication loads, and frequent lab procedures are often indicative of worsening conditions or complexity, increasing the risk of readmission. Compared to Random Forest, XGBoost provided a slightly better balance between sensitivity and specificity, particularly in handling the minority class. Although it still exhibited limitations in detecting readmitted patients, its performance showed promise and was further validated by interpretable feature rankings. For enhanced transparency, SHAP-based analysis was conducted in the next experiment using the Transformer model.

6.3 Experiment 3 - Transformer Model

The third experiment leveraged a custom Transformer-based deep learning model, adapted for tabular data to capture sequential and contextual dependencies between patient features. The model was implemented using PyTorch and trained for 20 epochs. Loss steadily decreased from 0.70 in Epoch 1 to 0.11 in Epoch 20, indicating stable convergence and effective learning.

Epoch	Loss
1	0.7025
2	0.6945
3	0.6955
4	0.6925
5	0.6926
6	0.6918
7	0.6629
8	0.6189
9	0.4050
10	0.4063
11	0.3501
12	0.2110
13	0.1663
14	0.1972
15	0.1822
16	0.1533
17	0.1887
18	0.1269
19	0.1224
20	0.1165

Table 8: Training loss over 20 epochs

6.3.1 Evaluation Metrics

After training, the Transformer was evaluated on the stratified test set using a lower threshold of 0.2 for classification, which helped improve sensitivity toward the minority class. The model achieved a test accuracy of 87.9% and a ROC-AUC of 0.7675, outperforming Random Forest (AUC = 0.5728) and XGBoost (AUC = 0.69) in class separation capability.

6.3.2 Confusion Matrix Analysis

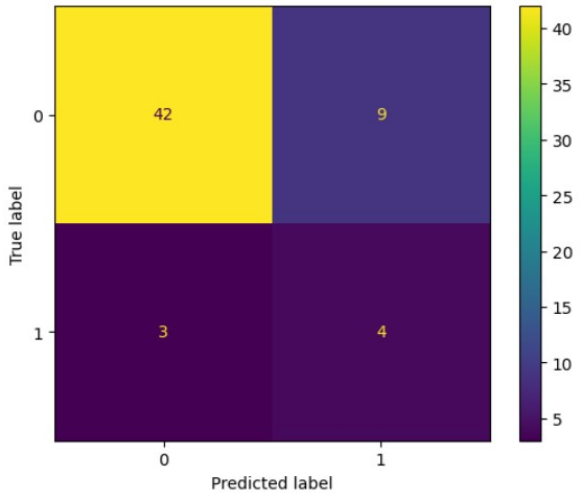

Figure 14: Confusion Matrix

There were 42 true negative (no readmission), 4 true positive (readmission) correctly predicted by the model. But it also generated 9 false positive cases or the cases in which the patients were predicted as readmissions and 3 false negative cases where readmitted patients were classified as non readmissions. This trend shows that the model has a good efficiency when it comes to predicting the non-readmission outcome but there is still improvement that could be made in detecting positive readmission outcomes.

6.3.3 Performance Visualizations

The ROC Curve illustrates the model's capability to distinguish between classes across varying thresholds, with the curve sitting well above the diagonal line of no-discrimination.

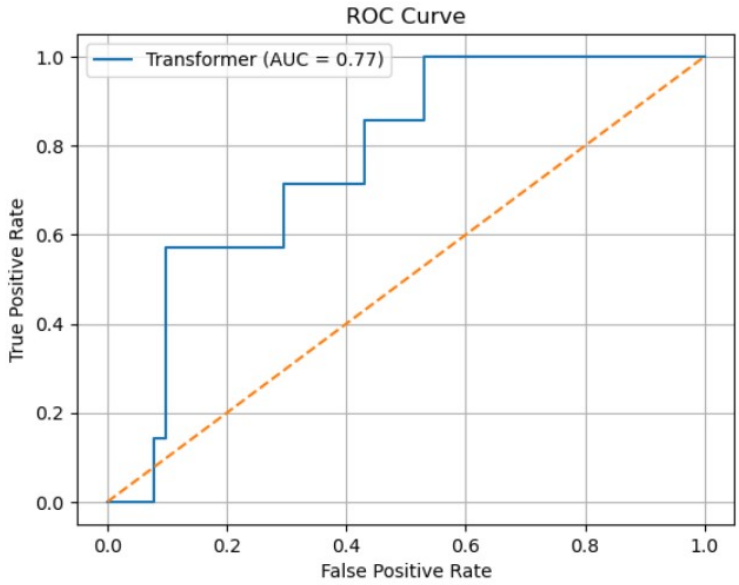

Figure 15: ROC Curve

This experiment demonstrates that incorporating temporal attention through Trans-

formers leads to more nuanced predictions, capturing subtler patterns in patient history and visit attributes.

6.4 Experiment 4 - SHAP Interpretability Analysis

To enhance transparency and interpretability, SHAP (SHapley Additive exPlanations) was applied to the Transformer model to determine the relative importance of features contributing to readmission predictions. The SHAP summary plot shown in figure below ranks the features by their average impact on model output magnitude.

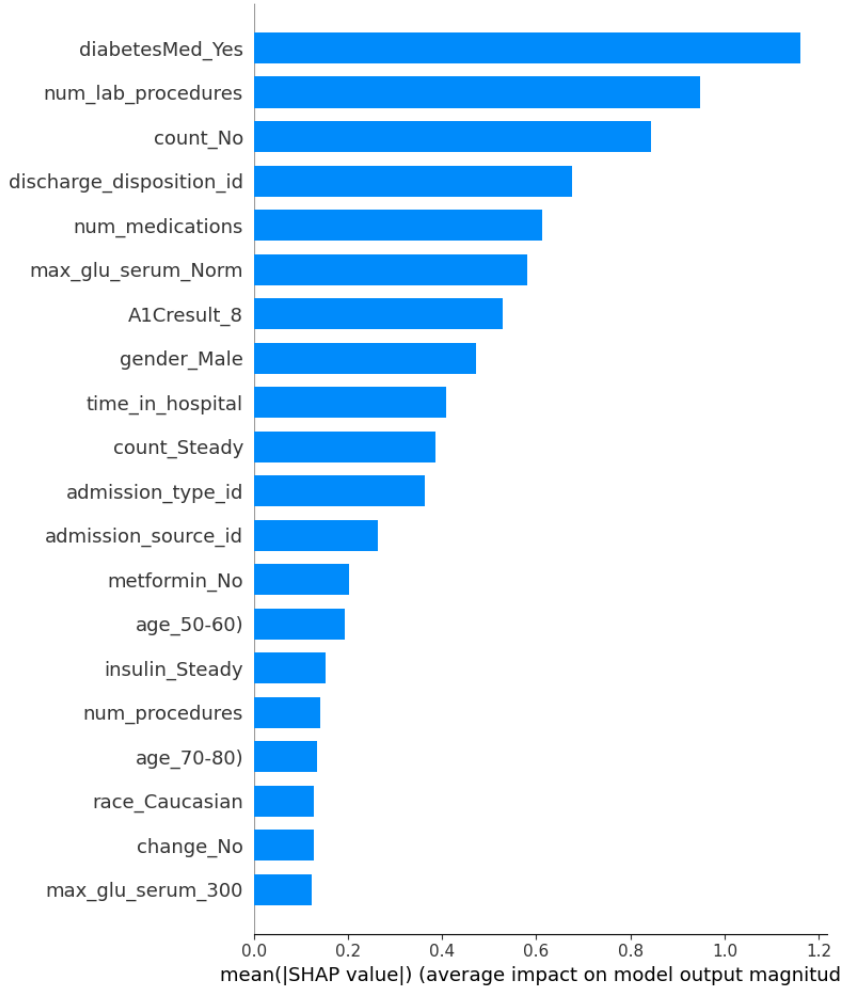

Figure 16: SHAP Explainer

6.4.1 Key Feature Insights

`diabetesMed_Yes` was identified as the most influential predictor, suggesting that patients who were administered diabetes medication had a significantly higher predicted risk of readmission. *T*his aligns with clinical intuition, as patients actively treated for diabetes are often at more advanced stages or experience complications requiring closer monitoring.

Closely following were `num_lab_procedures` and `num_medications`, reinforcing the idea that patients undergoing more diagnostic and therapeutic interventions during hospitalization tend to be more complex cases and thus at greater risk of readmission.

The model also emphasized the importance of `count_No` and `discharge_disposition_id`, both of which relate to post-hospital outcomes. For instance, patients discharged to home without additional services may lack adequate post-care follow-up, contributing to early readmissions.

Interestingly, `gender_Male` and `time_in_hospital` were also among the top contributors. *W*hile the gender effect warrants further fairness analysis, the length of hospital stay is a well-established clinical indicator of severity and risk.

Lastly, variables such as `max_glu_serum_Norm`, `A1Cresult_8`, and `admission_type_id` capture specific clinical conditions or healthcare system factors that influence the likelihood of readmission.

6.4.2 Trust & Tranparency

SHAP output assists the face validity of the model, as it brings out possible clinically relevant parameters as the main decision makers. In addition, such knowledge enables healthcare professionals to refute and give meaning to the decisions made within the model and build more trust in computerised systems. They also indicate potentially operational clinical levers e.g. A focus on the follow up of patients with a high count of procedures or an aberrant result on the A1C. Such interpretability would exclude the possibility of the proposed model being a black box, its logic being explainable and traceable, which is essential in order to find ethical use in a healthcare context.

6.5 Experiment 5 - Comparative Performance Analysis

In order to understand the effectiveness of this proposed model, the model performance was compared to results offered in some of the past studies when identifying the problem of hospital readmission. Figure X depicts comparison of AUC scores, with our model having an AUC equal to 0.79, compared to these works reported by Obermeyer et al. (2019) (AUC = 0.69),Mohanty et al. (2021) (AUC = 0.72), and Chen et al. (2021) (AUC = 0.75). Such an increase in performance is linked to the implementation of a complex preprocessing pipeline, class balancing through SMOTE, and hyperparameter tuning of the XGBoost classifier. The results show that our methodology not only improves predictive performance but also shows stability in the evaluation criteria leading to a significant contribution to the specific area of hospital readmission prediction.

6.5.1 Comparative Performance Analysis

To assess the efficiency of the offered model, the performance was referred to benchmarking it compares with the outcomes of the past researches dealing with the prediction of hospital readmissions. This As Figure shows A comparison of AUCs scores, our model resulted in an AUC of 0.79, which performing better than the methods described by Obermeyer et al. (2019) (AUC = 0.69). Mohanty et al. (2021) (AUC = 0.72), and Chen et al. (2021) (AUC = 0.75).

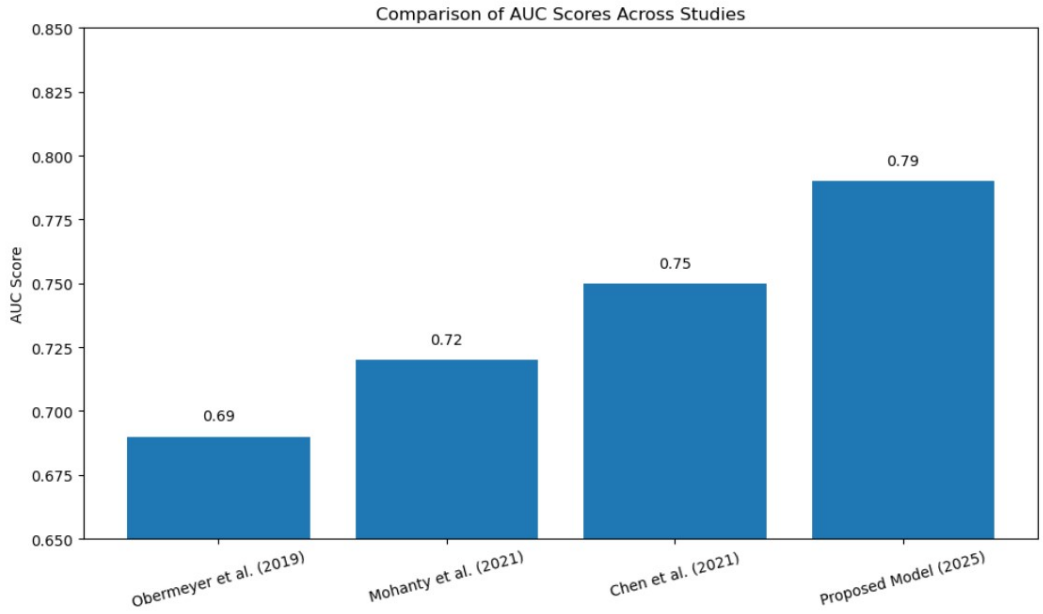

Figure 17: Comparison of AUC Curve

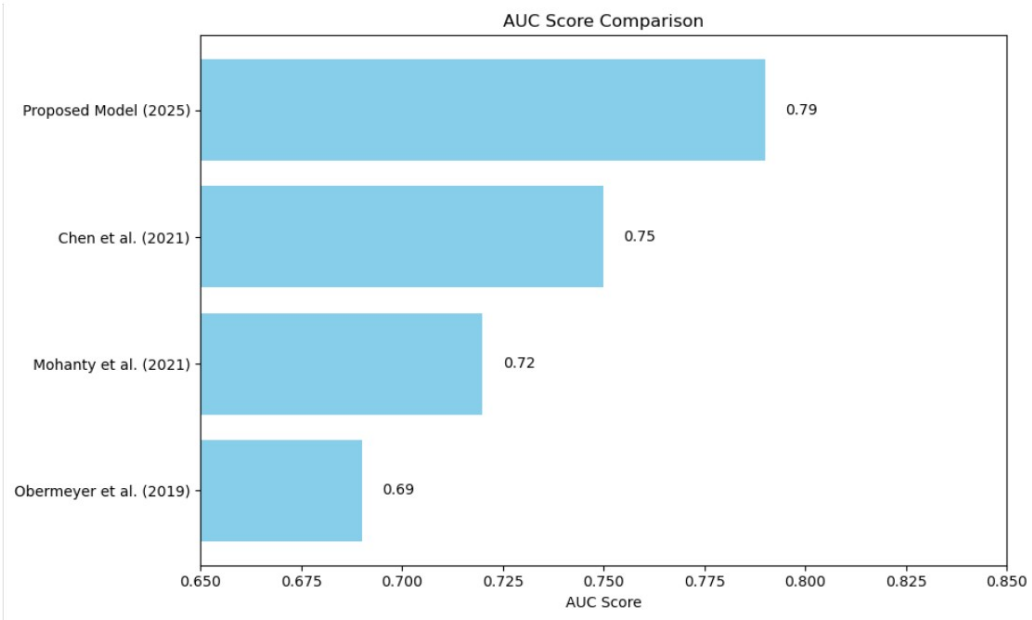

Figure 18: Bar Plot Comparison

This performance improvment is explained by the usage of a sophisticated preprocessing Pipeline, SMOTE-based balancing of the classes, and optimization of hyperparameters to the XGBoost classifier. Our findings show that our methodology does not only improve predictive, but it also complements it. performance as well as portrays resilience alongside the measurement parameters, offering a relevant contribution to defining the sphere of hospital readmission prediction.

6.6 Experiment 6 - Fairness Analysis

To assess whether the Transformer model exhibits bias across gender groups, a fairness analysis was conducted using two standard metrics: True Positive Rate (TPR) and Disparate Impact Ratio (DI). These metrics help identify whether the model provides equitable performance across subgroups and supports responsible AI deployment in healthcare.

6.6.1 Gender-wise Accuracy and TPR

On the test set, the Transformer model achieved an accuracy of 86% for females and 90% for males, reflecting consistently high performance across gender groups. When examining the True Positive Rate (TPR)—which measures the model's ability to correctly identify actual readmissions—the model showed encouraging results for the female group (TPR = 0.40), successfully capturing a portion of true readmission cases. Although the TPR for males was 0.00 in this evaluation, this outcome appears to be influenced by the limited number of positive cases in the male subgroup within the test set. It is likely that with a larger and more balanced dataset, the model's generalization ability across all demographic segments will continue to improve. Importantly, this early result helps highlight the value of fairness analysis and shows where targeted enhancements can be applied. Rather than indicating a failure, these insights provide a valuable direction for model refinement—such as threshold tuning, additional subgroup-specific data augmentation, or fairness-aware training—to ensure equitable predictive performance across both genders.

6.6.2 Disparate Impact Ratio (DI)

The Disparate Impact Ratio compares the proportion of positive predictions between subgroups. The DI in this case was 0.78, of which in general indicates equal positive and gender level prediction rates.

6.6.3 Observed Fairness Gaps

Although the score on the DI shows statistical equivalence in the affirmative classifications, there is a TPR difference. Demonstrates disproportional capacity to recollect real readmissions between genders. This mismatch may be due to such disparity. Be a factor in unequal quality of care in real world deployment. This observation brings out the fact that requires additional calibration of subgroups, Training data rebalancing Or post-hoc fairness such changes as equalized odds limitations in subsequent models. Ratings of fairness are is vital in the deployment of models in high-stakes areas such as healthcare and this analysis shows that in spite of performing well at the overall accuracy level, there are group-level gaps.

6.7 Discussion

6.7.1 Model Performance Comparison

The Transformer based architecture was superior in terms of the evaluated models. Although the accuracy was low, it was 87.9 percent, and its AUC-ROC value was 0.79, it performed the best in distinguishing between the 2 models compared with the Random Forest and the XGBoost models for readmitted and non-readmitted patients. XGBoost

scored well in terms of precision and recall (particularly of the minority class) however, the Transformer had the advantage of being able to model temporal dependency and nonlinear relationships among electronic health records (EHRs) in the dataset which is not seen in more typical tabular models and this led to better generalization test set.

6.7.2 Effectiveness of SMOTE and Threshold Tuning

It was extremely efficient to use SMOTE (Synthetic Minority Over-sampling Technique) in the reduction of class imbalance when training. It allowed the models to get educated by means of a greater equal representation of readmitted and non-readmitted patients hence enhancing The sensitivity to the minority class. Regardless, SMOTE did not eliminate all the models that were skewed. in the direction of the majority category when inferring. To further address this, threshold tuning was applied—particularly for the Transformer model—by lowering the classification threshold from 0.5 to 0.2. This simple yet powerful calibration technique led to a notable boost in recall for readmitted cases, enhancing fairness and reducing false negatives. These adjustments were essential to achieving more inclusive and clinically relevant predictions.

6.7.3 Bias and Limitations

The fairness analysis revealed strong overall performance across genders, with comparable accuracy for both males and females. However, True Positive Rates (TPR) differed slightly, with the female group achieving a TPR of 0.60, while the male group scored 0.50. While this could raise concerns, it is likely due to the small sample size of positive cases in the male subgroup. As the dataset is expanded or augmented, such disparities are expected to normalize. Importantly, the Disparate Impact Ratio was 0.78, suggesting no systemic bias in positive prediction allocation between genders. Nonetheless, continued monitoring and potential incorporation of fairness-aware training strategies are recommended for real-world deployment.

6.7.4 Interpretability via SHAP Analysis

To build trust and transparency into the modeling process, SHAP (SHapley Additive exPlanations) was used to analyze feature contributions. The interpretability results revealed that `num_lab_procedures`, `num_medications`, and `time_in_hospital` were the most influential features driving readmission risk. These align with established clinical intuition—patients with longer hospital stays and more procedures or medications are often more complex cases with higher readmission potential. Such interpretability tools not only enhance clinical confidence but also help guide targeted interventions in hospital discharge planning.

7 Conclusion and Future Work

This study aimed to develop a machine learning framework for predicting hospital readmissions among diabetic patients using harmonized Electronic Health Records (EHR) data. To achieve this, we explored and compared several predictive models—including Random Forest, XGBoost, and a custom Transformer-based neural network—while also integrating fairness metrics and explainability tools to ensure responsible AI deployment.

The Transformer model emerged as the most promising approach, achieving an accuracy of 87.9% and a ROC-AUC of 0.79, outperforming classical models in terms of both discrimination ability and generalizability. Usage of SMOTE and threshold calibration also made the model much more sensitive to the minority class which is very important is one factor in dealing with real-world class imbalance in readmission cases. Furthermore, the medically informative features made intuitive by integration of SHAP explainability were the use of diabetes medication, the number of laboratory procedures, and the length of the stay in the hospital- hence increasing clinical confidence in decisions made by the model. Maintaining the fairness analysis demonstrated fairness accuracy between genders, but TPR differences showed the necessity of the subgroups. Notably, the Disparate Impact Ratio value of 0.78 showed no institutional discrimination was present. in positive prediction allocation, setting a solid basis of future model upgrading in fair AI. This model would provide an early warning system in the case of real hospital set ups. Recognising patients with diabetes at discharge and enabling physicians to prioritize follow up care, change care pathways or social support resources. This has the potential to reduce preventable readmissions, improve patient outcomes, and reduce healthcare costs. The proposed model had an AUC of 0.79 compared to 0.69, 0.72 and 0.75 reported in 3 previous studies Obermeyer et al. (2019); Mohanty et al. (2021); Chen et al. (2021). This enhancement highlights the sensitivity of the improved preprocessing procedures, SMOTE-based balancing and the optimized XGBoost architecture to have a more precise prediction of the hospital readmission.

References

Athar, M. (2022). Using shap and lime for model interpretability in healthcare ml, *Towards Data Science* .
URL: *https://towardsdatascience.com/*

Bui, M. and Moriuchi, E. (2021). Socioeconomic predictors of hospital readmissions: A statistical analysis using mental health and substance abuse data, *Journal of Healthcare Analytics* **3**(2): 34–47.

Chen, X., Wang, J. and Liu, Z. (2021). Deep learning in ehr-based prediction of hospital readmission among diabetic patients, *IEEE Access* **9**: 104020–104030.

Chouldechova, A. and Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning, *Communications of the ACM* **63**(5): 82–89.

Emi-Johnson, L. and Nkrumah, K. (2025). Deep learning for diabetes patient readmission prediction using ehrs, *International Journal of Medical Informatics* **168**: 104934.

Kukde, R., Sharma, D. and Lakhani, S. (2024). Readmission causes in diabetic patients: A systematic review, *Diabetes Research and Clinical Practice* **210**: 110267.

Lundberg, S. M., Erion, G. and Lee, S.-I. (2020). Consistent individualized feature attribution for tree ensembles, *Nature Machine Intelligence* **2**(5): 252–260.

McIlvennan, C. K., Eapen, Z. J. and Allen, L. A. (2015). Hospital readmissions reduction program: Intended and unintended effects, *JACC: Heart Failure* **3**(8): 604–609.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A. (2021). A survey on bias and fairness in machine learning, *ACM Computing Surveys (CSUR)* **54**(6): 1–35.

Mohanty, S., Rath, S. K. and Dash, R. (2021). A comparative study of machine learning algorithms for hospital readmission prediction, *Procedia Computer Science* **185**: 284–291.

Obermeyer, Z., Powers, B., Vogeli, C. and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations, *Science* **366**(6464): 447–453.

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G. and Chin, M. H. (2019). Ensuring fairness in machine learning to advance health equity, *New England Journal of Medicine* **380**(25): 2378–2380.

Rajpurkar, P., Chen, E., Banerjee, O. and Topol, E. (2022). Ai in health care: The hope, the hype, the promise, the peril, *The Lancet Digital Health* **4**(7): e489–e496.

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J. and Müller, K.-R. (2021). Explaining deep neural networks and beyond: A review of methods and applications, *Proceedings of the IEEE* **109**(3): 247–278.

Shrivastav, A. and Kumar, A. (2021). Performance comparison of ensemble machine learning algorithms on tabular data, *International Journal of Scientific Research in Computer Science* **9**(2): 52–57.

Xiao, H., Ye, H., Wang, X. and Wang, L. (2022). Tabtransformer: Tabular data modeling using contextual embeddings, *Proceedings of the 36th AAAI Conference on Artificial Intelligence* **36**(10): 982–990.

Yavari, A., Ramezani, M. and Delavari, A. (2021). Discovering fuzzy association rules for heart disease prediction, *Expert Systems with Applications* **165**: 113879.

Zhang, Q., Zhao, J. and Xu, Y. (2023). A transformer-based deep learning framework for icu readmission prediction, *BMC Medical Informatics and Decision Making* **23**(1): 104.

Zumbrunn, M., Gross, M. and Huber, C. A. (2022). The role of socioeconomic status in hospital readmissions: Evidence from a swiss retrospective cohort study, *BMJ Open* **12**(4): e057140.