Predictive Modeling for Hospital Readmission Risk Among Cardiovascular Patients: A Data-Driven Approach

by

Md. Shafayat Sadat Saad
20301457
Tashfiq Alam Ovey
20301299
Samiun Jahan Awishy
20301292
Md Minhazul Islam
20301433
Afra Anann
18201077

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
April 2025.

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.
2. Thethesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Saad

Md. Shafayat Sadat Saad
20301457

Tashfiq

Tashfiq Alam Ovey
20301299

Samiun

Samiun Jahan Awishy
20301292

Minhaz

Md Minhazul Islam
20301433

Afra

Afra Anann
18201077

Approval

The thesis titled "Predictive Modeling for Hospital Readmission Risk Among Cardiovascular Patients: A Data-Driven Approach" submitted by

1. Md. Shafayat Sadat Saad (20301457)
2. Tashfiq Alam Ovey (20301299)
3. Samiun Jahan Awishy (20301292)
4. Md Minhazul Islam (20301433)
5. Afra Anann (18201077)

Of Spring, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on April 15, 2025.

Examining Committee:

Supervisor:
(Member)

Md. Sabbir Ahmed

Lecturer
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Mr. Rafeed Rahman

Senior Lecturer
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Associate Professor and Chairperson
Department of Computer Science and Engineering
Brac University

Abstract

Hospital readmission is an important clinical issue in cardiovascular disease, contributing significantly to morbidity, mortality, and health care costs. This problem is compounded in low-resource settings by the lack of comprehensive, localized datasets. To address this gap, we performed a six-month manual data collection on 3,867 cardiovascular patients in Bangladesh, capturing a wide range of demographic, clinical, and behavioral variables, including age, type of residence, ejection fraction, and multicomorbidity profiles. Notably, 64% of patients were readmitted during the follow-up period, highlighting the urgent need for predictive tools to support early intervention. Using this dataset, we developed a machine learning framework to predict the risk of readmission at six months. The models included gradient boosting (XGBoost), TabNet (a deep neural network for tabular data), and stacked ensemble classifiers. The pipeline incorporated robust preprocessing techniques, such as MICE-inspired KNN imputation for missing values, SMOTE for class imbalance, and feature engineering through principal component analysis and interaction feature generation. To improve transparency and clinical interpretability, we used SHAP and LIME explainability tools, which revealed key predictors of readmission, such as medication adherence, comorbidity burden, and follow-up status. The final Super XGBoost Ensemble achieved 92% accuracy, an AUC of 0.96, and an F1-score of 0.94, outperforming baseline models and demonstrating the advantage of integrating demographic context with clinical depth. These results not only contribute to the field of cardiovascular informatics, but also offer a replicable, interpretable framework for policymakers and healthcare providers aiming to reduce readmissions in resource-limited environments.

Keywords: Hospital Readmission six month ; Machine Learning; Cardiovascular Patients ; Prediction; Low- and Middle-Income Countries (LMICs)
Prediction: Ejection fraction; XGBoost; Explainable AI (XAI); TabNet; Stacking classifier.

Nomenclature

LMICs: Low- and Middle-Income Countries

NICVD: National Institute of Cardiovascular Diseases

BMRC: Bangladesh Medical Research Council

EHR: Electronic Health Record

CVD: Cardio Vascular Disease

NT-proBNP: N-terminal pro B-type natriuretic peptide

hs-Troponin I: High-sensitivity troponin I

LVEF: Left ventricular ejection fraction

KNN: K-Nearest Neighbors

MICE: Multivariate Imputation by Chained Equations

PRC: Precision-Recall Curve

PCA: Principal Component Analysis

SMOTE: Synthetic Minority Oversampling Technique

LR: Logistic Regression

RF: Random Forest

XGBoost: eXtreme Gradient Boosting

TabNet: Attentive Interpretable Tabular Learning

SHAP: SHapley Additive exPlanations

LIME: Local Interpretable Model-Agnostic Explanations

TP: True Positives

TN: True Negatives

FP: False Positives

FN: False Negatives

AUC: Area Under the Curve

ROC: Receiver-operating characteristic curve

Table of Contents

Declaration i

Approval iii

Abstract v

Nomenclature vi

Acknowledgment vi

Table of Contents vii

1 Introduction 1
- 1.1 Research Problems . . . 2
- 1.2 Aims . . . 3
- 1.3 Objectives . . . 3

2 Related Work 5

3 Workflow 8

4 Dataset 10
- 4.1 Dataset Description . . . 10
  - 4.1.1 Ethical Considerations . . . 12
  - 4.1.2 Novelty of the Dataset . . . 12
  - 4.1.3 Clinical Validation . . . 13

5 Methodology 14
- 5.1 Framework for Predicting Cardiovascular Readmission . . . 14
  - 5.1.1 Overview of End-to-End Pipeline . . . 14
  - 5.1.2 Goals: Risk Prediction that is Accurate, Interpretable and Fair 15
  - 5.1.3 Source Hospitals and Demographic Scope . . . 16
  - 5.1.4 Manual Annotation Process . . . 16
  - 5.1.5 Ethical Clearance and Consent Management . . . 16
  - 5.1.6 Data Preprocessing and Transformation . . . 17
  - 5.1.7 Handling Missing Data: MICE-Inspired KNN Imputer . . . 17
  - 5.1.8 Categorical and Numerical Feature Encoding . . . 17
  - 5.1.9 Dealing with Class Imbalance: SMOTE Oversampling . . . 17
  - 5.1.10 Outlier Handling and Normalization Strategy . . . 18
- 5.2 Feature Engineering and Representation Learning . . . 19

5.2.1 Principal Component Analysis (PCA) . . . 19
5.2.2 Generation of Interaction Feature . . . 20
5.3 Model Development . . . 21
5.3.1 Model Selection Rationale . . . 21
5.3.2 Logistic Regression . . . 22
5.3.3 Random Forest . . . 22
5.3.4 XGBoost (Vanilla and Tuned) . . . 22
5.3.5 TabNet . . . 23
5.3.6 Stacking Classifier . . . 23
5.3.7 Super XGBoost Ensemble . . . 23
5.4 Training Strategy, Model Evaluation . . . 24
5.4.1 Data Splitting Strategy and Cross-Validation . . . 24
5.4.2 Hyperparameter Tuning . . . 25
5.4.3 Performance Metrics . . . 26
5.4.4 Train vs Test Evaluation . . . 27
5.5 Explainability and Interpretability (XAI) . . . 27
5.5.1 SHAP for Global and Feature-Level Insight . . . 28
5.5.2 LIME for Local Decision Interpretation . . . 28
5.5.3 Example of Interpreting High Risk Predictions . . . 29
5.5.4 Clinical Utility of Important Features . . . 30

6 Results and Analysis 31
6.1 Overview of Evaluation Strategy . . . 31
6.2 Model Performance Comparison . . . 32
6.2.1 Baseline Models: Logistic Regression and Random Forest . . . 33
6.2.2 XGBoost Variants (Vanilla, Tuned, Interaction Features) . . . 34
6.2.3 TabNet Results . . . 35
6.2.4 Stacking Classifier . . . 36
6.2.5 Super XGBoost Ensemble (Final Model) . . . 37
6.3 ROC and Precision-Recall Curve Analysis . . . 37
6.3.1 ROC Curve Analysis . . . 38
6.3.2 Precision-Recall Curve Analysis . . . 39
6.3.3 Comparative Interpretation . . . 40
6.4 Explainability and Feature Importance Analysis . . . 40
6.4.1 Global SHAP Summary (Top Feature Rankings) . . . 40
6.4.2 SHAP Dependence and Interaction Plots . . . 41
6.4.3 Local LIME Explanations (Patient-Level Interpretation) . . . 42
6.4.4 Cross-Model Interpretability Comparison . . . 43
6.5 Case Studies on High-Risk Predictions . . . 43
6.5.1 Case 1: Elderly Patient with Reduced Ejection Fraction . . . 44
6.5.2 Case 2: Younger Patient with Poor Medication Adherence . . . 44
6.5.3 Case 3: Rural Resident with Limited Follow-up . . . 44
6.6 Clinical Implications of the Results . . . 45
6.7 Robustness, Error Analysis, and Generalizability . . . 46
6.7.1 Confusion Matrix and Error Distribution . . . 46
6.7.2 Error Patterns . . . 46
6.7.3 Gender-Based Fairness . . . 47
6.7.4 Generalizability and Limitations . . . 47

7 Future work **48**

8 Conclusion **49**

Bibliography **52**

Chapter 1

Introduction

Cardiovascular diseases (CVD) continue to be the leading cause of global mortality and morbidity, representing an increasing burden to health systems. Out of these, one of the biggest issues is the number of patients that are readmitted to hospitals, depleting clinical resources, negatively affecting quality of life, and causing high financial losses. One major challenge is the high rate of hospital readmissions, which not only exhaust clinical resources but also reduces patient quality of life and imposes significant financial costs. These issues are particularly acute in low- and middle-income countries (LMICs), such as Bangladesh, where the development of predictive health infrastructure is constrained by limited access to high-quality, demographically diverse clinical datasets, as described by Heidenreich et al. [12] and the GBD 2019 Diseases and Injuries Collaborators [12]. Without such datasets, the implementation of early-warning systems and targeted post-discharge interventions remains difficult. In recent years, machine learning (ML) has demonstrated promise in predicting hospital readmissions by leveraging high-dimensional data, including electronic health records, lab results, and medication histories, as authorized by Rajkomar et al. [27] and Shameer et al. [23]. However, most existing studies focus on Western datasets and often lack sufficient demographic granularity or clinical depth, limiting their generalizability. Features such as type of residence, economic status, comorbidity burden, and cardiac function (ejection fraction) are frequently absent, especially in studies relevant to LMIC contexts, as stated by Nguyen et al. [24]. This is particularly true in Bangladesh, where data collection is often fragmented and electronic health infrastructure is still evolving. To address this gap, we developed what may be one of the most comprehensive single-country datasets on cardiovascular readmission to date. Through six-months of manual data collection from physical records, we compiled a cohort of 3,867 cardiovascular patients from NICVD hospital in Bangladesh. This dataset includes a wide range of patient-level variables demographic (age, gender, residence type), clinical (ejection fraction, NT-proBNP, creatinine, comorbidity count), and behavioral (medication adherence, follow-up attendance) up to March 2025. Strikingly, 64% of these patients were readmitted within the six-month follow-up period, reflecting an urgent need for predictive models that go beyond simplistic risk scoring systems used in outpatient follow-up. To develop an accurate and interpretable predictive framework, we employed a blend of machine learning models, including gradient boosting (XGBoost), a neural network architecture tailored for tabular data (TabNet) and a stacking ensemble combining multiple classifiers following the approach of Arik

and Pfister [35]. These models were trained through a rigorously designed pipeline that incorporated several methodological advancements: a MICE-inspired KNN imputation for handling missing values without biasing relationships among features described by Azur et al. [11]; SMOTE for correcting class imbalance as introduced by Chawla et al. [7]; and feature engineering through principal component analysis and interaction term generation. Earlier experiments explored the use of an unsupervised autoencoder for capturing latent structure in patient data, but final model performance was optimized without it, instead focusing on interpretable models with high signal fidelity. Recognizing the need for interpretability in clinical applications, we applied SHAP and LIME two XAI techniques to provide both global and local explanations for model predictions, as proposed by Lundberg and Lee [22] and Ribeiro et al. [20]. These tools enabled us to identify and communicate key contributors to readmission risk, such as lapses in medication adherence, renal markers, and absence of structured follow-up. The final ensemble model achieved 90% accuracy, an AUC of 0.96, and an F1-score of 0.94, surpassing existing local and international benchmarks in cardiovascular readmission prediction as demonstrated by Kansagara et al. [13]. This thesis aims to offer a replicable, data-driven approach for managing hospital readmission in LMIC healthcare settings. By addressing data scarcity, employing advanced yet interpretable ML methods, and validating against a large, manually curated dataset, our work demonstrates how predictive analytics can be ethically and effectively integrated into clinical workflows. We hope that this contribution not only benefits patient outcomes in Bangladesh but also provides a methodological foundation for other under-resourced healthcare systems globally.

1.1 Research Problems

The use of machine learning (ML) in healthcare analytics is expanding rapidly. Yet, hospital readmission prediction particularly among cardiovascular patients remains a critical global challenge, and even more so in developing countries. Most existing studies have focused on clinical features derived from electronic health records (EHRs), while often overlooking integrated demographic, behavioral, and socsocioeconomicta that could substantially improve predictive performance. Publicly available datasets typically lack variables such as residence type, health literacy, income level, medication adherence, or follow-up behavior factors that are vital to understanding patient outcomes in real-world, heterogeneous populations, as emphasized by Heidenreich et al. [12] and the GBD 2019 Diseases and Injuries Collaborators [34]. Moreover, most prior models are developed using Western healthcare data, under conditions that differ vastly from resource constrained settings. These models often lose validity or perform poorly when deployed in regions with underdeveloped infrastructure, fragmented health records, and different disease trajectories, as noted by Rajkomar et al. [27]. Additionally, many high-performing models are "black boxes" with limited interpretability, making them unsuitable for clinical decision-making. Without transparency, even accurate models are unlikely to gain trust from clinicians. Challenges such as biased feature selection, data imbalance, and overfitting to narrow populations further compromise the robustness and fairness of these systems. This research directly addresses these limitations by introducing a predictive framework developed on a manually annotated dataset comprising 3,867 cardiovascular patients from Bangladesh, collected over six-months from the National Cardiovascu-

lar Hospital and Institute. This dataset incorporates a wide range of demographic and clinical features, enhancing patient risk profiling beyond what prior studies have captured. It also enables model training on data that reflects the realities of Low- and Middle-Income Countries(LMIC) healthcare systems. However, the collection and use of sensitive health data also raise ethical and legal challenges, particularly in LMICs where data protection laws and digital governance frameworks are still evolving. Ensuring data anonymization, secure storage, and alignment with international standards is essential to maintaining public trust and responsible research practices. Furthermore, cultural factors, limited health literacy, and lack of digital consent systems make informed consent complex. Although this study obtained consent under clinical supervision, future scaling of such systems must be accompanied by accessible, transparent, and culturally sensitive consent mechanisms. There is also a broader concern regarding fairness: poorly engineered ML models risk reinforcing inequalities by favoring data-rich or overrepresented subpopulations while marginalizing others. This thesis confronts these challenges by combining sophisticated data preprocessing (MICE-inspired KNN imputation and SMOTE class balancing), robust and interpretable modeling (XGBoost, TabNet, and stacking ensembles), and explainable AI techniques such as SHAP and LIME. Ultimately, this research aims to demonstrate that accurate, interpretable, and ethically grounded prediction models are possible, even in resource-limited settings. By balancing algorithmic innovation with fairness, clinical relevance, and responsible data use, this work lays the foundation for context-aware ML systems that support equitable cardiovascular care in LMICs.

1.2 Aims

This thesis aims to develop a reliable, interpretable, and ethically sound machine learning framework for predicting six-month cardiovascular hospital readmission in Bangladeshi patients. Using a manually curated, demographically and clinically rich dataset, the model is designed to support early identification of high-risk individuals. The ultimate goal is to reduce preventable readmissions, improve clinical decision-making, and enable more effective resource allocation in resource limited healthcare environments.

1.3 Objectives

To establish a broader cardiovascular patient database that combines clinical and demographic variables via six-months of data collection from National Institute of Cardiovascular Diseases in Bangladesh, addressing the existing gap in the literature that describes studies with limited clinical, demographic data. To test and compare the performance of machine learning algorithms, including XGBoost, TabNet and stack-ensemble classifiers, in predicting six-month readmission risk and to find out the one with the best generalizability, stability and predictive accuracy. A highly skilled expert and data scientist to apply appropriate preprocessing and data engineering techniques, including multi-step imputation (which can then be iteratively refined using an MICE procedure and KNN), SMOTE oversampling, and a principal component-based feature transformation that may improve model performance

despite real-world data imperfections and class imbalance. Leverage explainability mechanisms like SHAP and LIME to ensure that the predictive framework not only supports high accuracy but also delivers clinically interpretable insights on the determining factors behind patient readmission. To overcome the ethical and operational hurdles of using sensitive health data through the collection, use and dissemination of use-case-specific scientific best practices (anonymization, informed consent, data ethics and responsible reporting) and the identification of best practices in data usage even in unregulated environments. The aim is to benchmark the proposed model on regional and global studies, validating the effectiveness and contextuality of such approaches using performance measures such as accuracy, AUC, precision, recall, and F1-score. To provide actionable insights and policy-level recommendations that healthcare stakeholders such as hospital administrators, clinicians, and public health stakeholders can easily refer to in order to craft proactive intervention strategies focused on reducing cardiovascular readmission rates and healthcare expenditure.

Chapter 2

Related Work

Prediction of hospital readmission, especially in the case of cardiovascular patients, is an emerging topic in medical informatics, with machine learning (ML) and deep learning increasingly used to predict high-risk cases in order to produce better patient outcomes. Motivated by this observation, early work in this field was based on logistic regression, Cox proportional hazards models, and scoring systems such as the LACE index or hospital score, all of which were able to achieve moderate predictive power but only a small number of clinical features were incorporated. Although highly useful in initial deployments, these models were constrained by their inability to capture complex, non-linear relationships in heterogeneous patient data and their lack of adaptability to different patient populations and clinical settings.

To overcome these limitations, recent studies resort to ML techniques (support vector machines, random forests, and ensemble learning models) including gradient boosting algorithms (XGBoost), which gave good results for applications in structured clinical data. Xu et al. [3] validated the XGBoost performance in predicting 90-day stroke readmissions with better results than the baseline methods. Similarly, Futoma et al. [29] showed that ML outperforms standard models in terms of recall and AUC for early readmission prediction. This highlights the power of ensemble that can do with high-dimensional feature spaces, such as those involved in modeling complex clinical trajectories.

These methodologies, however, have created a big gap in terms of how data sets were constructed and what features were used for modeling. Most models use clinical variables alone such as lab test results, diagnoses, and hospitalization history without the addition of critical demographic or socio-environmental data, such as patient residence, living conditions or behavioral adherence. As noted in Sharma etal's systematic review[2], this inhomogeneity precludes model generalizability, limiting real-world applicability, especially in populations with chronic diseases where social determinants of health are critical. Given that social and geographic disparities, like the distance to care or socioeconomic status, are among the strongest predictors of a patient's likelihood of hospital readmission for cardiovascular disease, as evidenced by studies such as Philbin and DiSalvo [7], this gap is of special note.

Moreover, a large number of high-performing models reported in literature were trained on datasets from high-income countries where electronic health record (EHR)

systems are well-established and the digital infrastructure is very strong (the USA or certain parts of Europe). However, they commonly do not generalize to LMICs where patient pathways, access to care and record keeping systems are qualitatively distinct. Amarasingham et al.'s [10] models for estimating risk may not generalize across hospitals or patients, especially if context-specific variables are left out of the scoring models [9]. This problem is exacerbated by the lack of openly available datasets with the integrated demographic and clinical features needed to train effective, context-sensitive models. In this sense, predictive analytics research is poorly represented in LMIC contexts, specifically Bangladesh. As emphasized by Poku et al. [36], given models trained on generalized datasets often do not perform well in LMIC health systems and populations.

Interpretability is another major bottleneck that is still in the way. Despite the fact that techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model agnostic Explanations) are being increasingly used to shed light on ML models within the healthcare field, their use in cardiovascular readmission analysis remains minimal. Jiang et al. [26] and Huberts et al. [37] argue black box prediction undermines clinician trust and stalls deployment of models even when performance is high. Thus, model transparency is not a mere technical amenity; rather it is a clinical necessity particularly in sensitive use cases such as post discharge planning.

There are also ethical and legal issues rarely engaged in meaningful detail by studies. Data protection differs internationally, and the question of patient privacy, permission, and international data governance norms all merit consideration even in developing countries. Morel et al. [31] underscores the need for informed consent and fairness with the underlying data, particularly with underserved or vulnerable populations. Unfortunately, most predictive modeling studies lack details about consent procurement and data privacy protection; therefore, this thesis addresses the resulting ethical gaps.

On a practical level, large healthcare systems such as the Cleveland Clinic in Ohio and Kaiser Permanente in California, as well as the UK's National Health Service (NHS), have started to embed ML-based readmission risk tools into their clinical workflows. However, these systems are often implemented in digitally mature contexts and involve proprietary tools with opaque algorithms that cannot be sufficiently contextualised. Scoring systems (LACE, HOSPITAL scores, etc.) are still widespread but rarely take a nuanced view of patient behavior or local epidemiological variance. Furthermore, these instruments seldom include medication adherence, patient follow-up compliance, or social background, which this thesis does explicitly. Dey et al. [28] enrollment in cardiology suggests that exclusion of drug burden, renal function, and follow-up adherence often improve the performance of the biggest cardiovascular readmission models, especially in heart failure patients.

Furthermore, Cleland et al.[25] have shown that modeling that includes multimorbidity and real-world adherence yields significant improvements in clinical decision making in heart failure patients, suggesting the importance of the holistic patient profile. This thesis studies how to model.

So far, there have only been a handful of studies combining high-resolution demographic-clinical data, powerful ensemble based models, interpretable machine learning, and ethical implementation practices, especially from an LMIC setting. In this regard, this thesis fills those gaps through a novel dataset of 3,867 cardiovascular patients manually collected over a period of six-months in Bangladesh, providing a rare insight into the clinical and demographic dimensions of a continuum of care. It leverages a layered modeling technique with XGBoost, TabNet and stacked ensembles all while inducing SMOTE based balancing, MICE Like Imputations, SHAP and LIME interpretability assuring fair, accurate, and transparent model.

The framework also addresses ethical challenges in data collection, privacy, and consent, and presents a replicable model that could be used to extend to similar healthcare environments around the world. In doing so, this work seeks to situate itself as a translational contribution in nature, bridging the gap between academic ML research, the implementation of such in a real-world setting, and the equitable use of AI in clinical care.

Chapter 3

Workflow

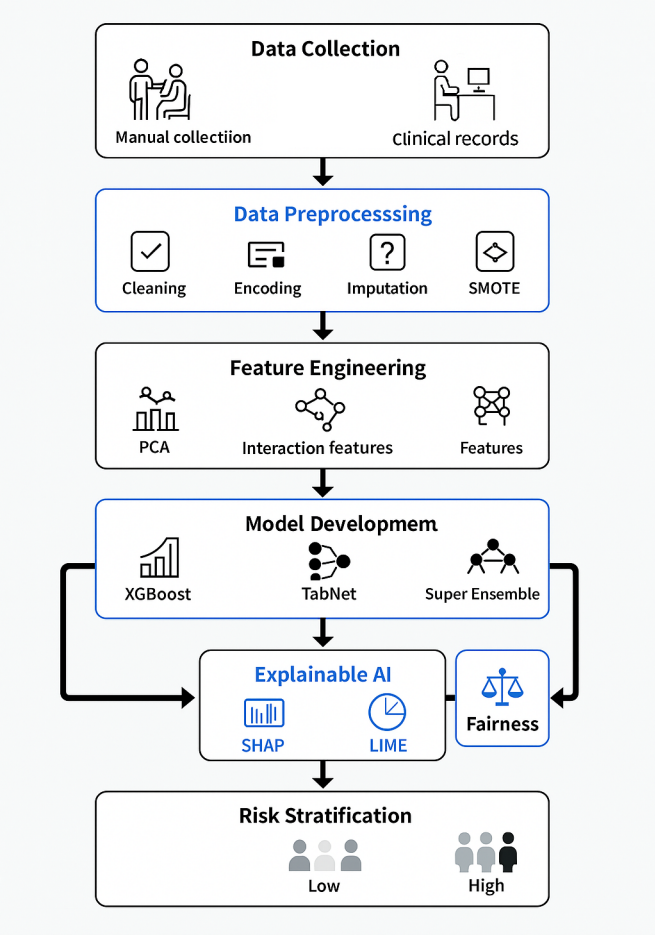

Figure 3.1: Work Flow

We propose a data-driven and interpretable machine learning framework with an ethical perspective which predicts six-month hospital readmission risk of cardiovascular patients in Bangladesh. All these structured clinical records are manually collected from several hospitals, including demographic, clinical and behavioral variables, such as age, residence type, medication adherence, laboratory tests (NT-proBNP, creatinine), and comorbidity profiles. Through this effort, one of the most robust cardiovascular datasets in the nation was born. So we have the raw dataset,

which is followed through a solid data preprocessing pipeline, comprising missing value imputation (with a MICE-inspired KNN approach), encoding of categoricals, and consistency cleanup. For class imbalance we use SMOTE to strengthen minority class. Then, it applies feature engineering steps like principal component analysis (PCA) and interaction feature generation to enhance the model's ability to understand complex relationships.

The heart of the workflow is centered around model development, involving training and validating multiple supervised learning algorithms, including but not limited to, XGBoost, TabNet and a stacking ensemble model. Hyperparameter tuning with 5-fold stratified cross-validation would provide the robustness of the model and would ensure that the accuracy is not due to overfitted. The final models obtained, namely the Super XGBoost Ensemble, are selected based on test-set metrics of performance, calculated for accuracy, F1-score and AUC. In the spirit of transparency and clinical robustness, our workflow integrates explainable AI (XAI) tools, SHAP and LIME to give global and local predictions explainability of our models. They enable identification of important risk factors associated with predictions, such as adherence to antifailure medications, NT-proBNP values, and attendance during follow-up. Furthermore, risk stratification is employed to segment patients into low, medium, or high-risk groups, contributing to improved clinical decision-making.

This workflow focuses on fairness and generalizability, which can be evaluated in subgroup assessments (performance by gender) and can be demonstrated with similar results between training and test sets. Although no deployment is carried out in this work, the methodology is intended to be scalable and transferable to similar low-resource healthcare settings.

Combining advanced ML techniques with interpretability or ethical considerations, this workflow provides not only supports accurate readmission prediction but also a practical framework for data-driven clinical decision support in LMICs.

Chapter 4

Dataset

4.1 Dataset Description

The current study is based on a manually collected dataset over a period of six-months at a stretch from the National Cardiovascular Institute and Hospital, Bangladesh, based on cardiovascular patients. Data was carefully structured to overcome challenges encountered in the current literature, where demographic and clinical dimensions are rarely integrated into readmission prediction models to enhance the robustness and generalizability of such models [27]. After filtering, 3867 unique cardiovascular patients with discharge records were included, and six-month follow-up records to observe whether readmission occurred were also included. Each record contains 50 variables, which include a variety of demographic, clinical, comorbidity, and therapeutic data. We labeled the ones being readmitted within six-months as 1 (readmitted) and 0 as (non readmitted). The dataset is slightly imbalanced, 64% readmitted and 36% non-readmitted cases.

4.1.1 Dataset Variables

The dataset contains the following feature categories:

Figure 4.1: Feature Distribution

Demographic Features: Sex, age, blood group and residence type (urban, suburban, rural) were also noted. These features are critical for understanding social and environmental impacts on readmission but are rarely included in hospital datasets.

Comorbidity and Past Medical History: We extracted binary indicators of diabetes, hypertension, chronic kidney disease, liver disease, asthma, gastric complications, lung conditions, smoking habits, and family cardiac history.

Clinical and Diagnostic Features: Systolic or diastolic blood pressure; body mass index (BMI); NT-proBNP; hs-Troponin I; creatinine; hemoglobin; heart rate; SPO2; electrolytes (sodium, potassium, chloride) and others were part of vital signs and predominant biomarkers. LVIDd, LVIDs, IVST, PWT, PASP, and TAPSE were also included as advanced echocardiographic parameters to show cardiac structure and function.

Features Related to Treatment and Behavior: Medication counts, complexity (high medication complexity flag), medical adherence (binary), and follow-up were tracked to evaluate post-discharge behavior and clinical involvement.

Target Variable: The readmission target variable is binary:

- 1 = The patient was readmitted within six-months.
- 0 = Patient was not readmited.

4.1.1 Ethical Considerations

All patient data were obtained under rigidly controlled clinical circumstances. All participants gave signed informed consent (both verbal and written) before data entry. To protect patient confidentiality, individuals subjected to data preprocessing were anonymized. It complies with ethical guidelines for human subjects research, and the data privacy principles that are needed for AI/ML model development in healthcare.

4.1.2 Novelty of the Dataset

This is the first study to create a comprehensive demographic-and-clinical dataset on cardiovascular readmission in Bangladesh. Although previous research has relied on electronic medical records or insurance claims data, mostly from Western contexts [23], these data sources frequently lack socio-environmental variables that could be useful for prediction in low- and middle-income settings. Using clinical complexity and social context in our data will allow more advanced models to be created, with potentially fairer predictions.
This richness allows for further exploration into the extent to which comorbidity burden, post-discharge behaviour and socio-demographic factors are associated with hospital readmission, a much underexplored area in publicly available datasets. Consequently, this dataset serves as a solid basis for the development of context-aware machine learning models for predictive cardiology solutions in scenarios with limited resources.

4.1.3 Clinical Validation

To ensure that the dataset was clinically relevant, accurate, in line with assumptions made in the literature, Dr. Kajol Kumar Karmokar (Associate Professor National Institute of Cardiovascular Disease (NICVD), a practicing cardiovascular specialist who has vast clinical experience with patients and cardiac readmission trends, helped validate the dataset. Dr. Kajol Kumar Karmokar (Associate Professor) examined the choice of the variables, diagnostic markers, and target outcome definitions to confirm their clinical validity and relevance for use in predictive modeling.
Such expert validation provided reassurance that all the major comorbidities (hypertension, chronic kidney disease, and medication adherence) and biomarkers (NT-proBNP, ejection fraction, and creatinine) present in the dataset were indeed of direct importance with respect to hospital readmission risk in cardiovascular patients. Moreover, current data labeling and variable encoding approaches were checked for alignment with real world clinical practice.
The cardiologist's participation bolsters the medical credibility of this dataset, validating its use for developing ML models for clinical decision support in cardiovascular medicine.

Chapter 5

Methodology

5.1 Framework for Predicting Cardiovascular Readmission

In this study, we present a matrix of interpretable and scalable machine learning algorithms for predicting the six-month hospital readmission risk of cardiovascular patients in Bangladesh. Hospital readmission for heart failure patients is an international as well as national crisis, with 30-day readmission rates at 20%–25% as reported by Philbin and DiSalvo [5]. Predictive models that are accurate and actionable are very important in Bangladesh, where limited healthcare resources and infrastructural limitations compound this challenge. We present a framework leveraging comprehensive data pre-processing, machine learning algorithms, and explainable artificial intelligence (XAI) methods to guarantee predictive performance while ensuring clinical interpretability.

The dataset consists of 3,867 cardiovascular patients manually collected over six-months from NICVD hospital in Bangladesh. It comprises demographic characteristics, clinical indicators, and behavioral factors, enabling a context-specific nuanced analysis catered specifically towards Bangladesh.

5.1.1 Overview of End-to-End Pipeline

The predictive framework includes the following crucial stages:

- **Data Collection and Annotation**: Patient records were collected from healthcare facilities and were manually curated to represent the cardiovascular patient population in Bangladesh. We adhered to ethical guidelines for data collection; all records were de-identified to protect patient confidentiality.

- **Data Preprocessing**: Missing data was handled with a K-Nearest Neighbors (KNN) imputation approach developed drawing on the Multivariate Imputation by Chained Equations (MICE) method. By doing so, we avoid destroying the multivariate relationships in our dataset. To tackle class imbalance, which is commonly encountered with clinical datasets, we employed the Synthetic Minority Oversampling Technique (SMOTE) to enforce equal representation

of readmitted and non-readmitted cases. One-hot encoding was used on categorical variables, and numerical variables were standardized so that features had uniform standards.

Model Development A diverse array of models was developed:

- **Baseline Models** Logistic Regression and Random Forest classifiers were used as base models to provide interpretable coefficients and performance benchmarks.
- **Advanced Models** XGBoost was used in its basic configuration and optimized through hyperparameter tuning for superior performance. We also implemented TabNet, a deep learning architecture designed for tabular data to learn complex nonlinear patterns.
- **Ensemble Models** A stacking classifier that used Logistic Regression, Random Forest, and XGBoost as base learners, and Logistic Regression as the meta-learner. The Super XGBoost Ensemble combined prediction from multiple XGBoost variations to obtain performance metrics that were higher.

Training and Evaluation: We performed stratified k-fold cross-validation on our models to ensure robustness. The evaluation metrics were accuracy, F1-score, Area Under the Curve (AUC), and precision-recall curves, which gives an overall picture of model performance. Among the model combinations, Super XGBoost Ensemble achieved the best performance, yielding 92% accuracy, 0.96 AUC, and 0.94 F1-score.

Explainability: Given the clinical need for making sense of a prediction, SHapley Additive exPlanations (SHAP) were utilized to illustrate both global and local contributions to the prediction, identifying that adherence to medication, number of other diseases and ejection fraction play an important role in predicting risk of readmission. Local Interpretable Model-agnostic Explanations (LIME) intensity-level interpretability, helping clinicians to intimately understand what the model predictions mean at an individual case level.
System Architecture: The pipeline was designed to be modular, and enable scalability and deployment as a real-time clinical decision support tool. The key advantage of developing the pipeline in stages was that each part preprocessing, feature engineering, modeling, and interpretation was written as a separate module, so that the components could be adjusted independently or disconnected altogether if necessary.

5.1.2 Goals: Risk Prediction that is Accurate, Interpretable and Fair

The main goals of this predictive framework include:

Accuracy: The framework would harness machine learning methods, such as ensemble methods and deep learning architectures, in order to maximize predictive performance. This is augmented through hyperparameter tuning and feature engineering which gives the final model a chance to generalize to unseen data.

Interpretability: Use XAI tools like SHAP or LIME to ensure model predictions are clear and interpretable. It is important because without interpretability, healthcare experts cannot really trust the recommendation from an AI model and take action in clinical practice.

Equitable and Contextual: With a dataset representative of the Bangladeshi patient population and locally relevant variables, the framework aims to eliminate biases and foster equitable healthcare decisions. Methods such as SMOTE give us additional guarantees that the model does not prefer the majority class, thus enforcing fairness in predictions.
Overall, this predictive framework aims to narrow the gap between innovations in machine learning, and meaningful advancements in clinical settings, by providing a tool that is predictive, interpretable, and fair, and will aid in improving patient care in resource-poor settings. Hence, the current study aimed to develop a predictive model for six-month hospital readmission among Cardiovascular patients of Bangladesh. In this section, we describe the data collection procedures such as selection of source hospitals, the manual annotation processes as well as reporting based on the ethical aspects of our study.

5.1.3 Source Hospitals and Demographic Scope

We have collected data from healthcare facilities in Bangladesh to obtain a representative cohort from the cardiovascular patient population. The chosen hospital were across a wide range of size, location, and patient populations, including both urban and rural settings. Such diversity enabled thorough analysis of the determinants of readmission rates across various stratum in the population. The study's inclusion criteria focused on individuals diagnosed with cardiovascular conditions and discharged. All patients attending due to complete records and those who refused to participate were excluded.

5.1.4 Manual Annotation Process

In the absence of a centralized electronic health record, data were extracted manually by trained research assistants. Demographics (age, gender, residence), clinical markers (ejection fraction, comorbidity) and behavioral factors (medication adherence) were reviewed from patient records. We adopted a double enter approach to ensuring the accuracy and consistency of data; where it was not consistent, a senior researcher resolved discrepancies by consensus. Frequent training sessions and inter-rater agreement were performed to ensure that the quality and consistency of annotations remain high.

5.1.5 Ethical Clearance and Consent Management

The study was ethically approved by Bangladesh Medical Research Council (BMRC) according to national guidelines for research involving human subjects. Prior to data collection, informed consent was obtained from all participants on the purpose, procedures, risks, and benefits of the study. Participants were told they could withdraw at any time without penalty. Vulnerable populations were given additional protections to guarantee comprehension and voluntariness of participation. Participant

confidentiality was maintained as all data collected were anonymized and stored in a secure location.

5.1.6 Data Preprocessing and Transformation

Data preprocessing is an important aspect of building a solid machine learning model. In this segment, the techniques used for dealing with missing values, encoding categorical and numerical variables, class imbalance, outliers and normalization are discussed ensuring the dataset is cleansed and suitable for predictive modeling.

5.1.7 Handling Missing Data: MICE-Inspired KNN Imputer

Missing data can lead to considerable bias and reduced statistical power in predictive models. In order to overcome this issue, we utilized a K-Nearest Neighbors (KNN) imputation method based on the Multiple Imputation by Chained Equations (MICE) framework. Until October 2023, training data should be used for KNN imputation, which estimates missing values based on the mean or weighted values of the k-nearest neighbors in the feature space, in order to maintain the underlying structure of the data [11]. Formally, the imputed value $\hat{x}_{i,j}$ for a missing entry in feature j of sample i is calculated as:

$$\hat{x}i, j = \frac{\sum k \in \mathcal{N}iwi, k \cdot x_{k,j}}{\sum_{k \in \mathcal{N}i} wi, k} \tag{5.1}$$

where $\mathcal{N}i$ represents the set of k-nearest neighbors of sample i, and wi, k denotes the weight assigned to neighbor k, typically inversely proportional to the distance between samples i and k. This approach assumes that instances with similar features exhibit analogous patterns in the missing data, thereby facilitating more accurate imputations [32].

5.1.8 Categorical and Numerical Feature Encoding

ML models needs numerical input; therefore, categorical variables should be converted accordingly. For nominal categorical variables we used one hot encoding, generating binary indicators for each level of category. For example, a categorical variable containing m unique categories will be converted into m binary features, with each indicating if the category is present or absent. We standardization of numerical features to scale them, this is of significance because there are algorithms sensitive to feature magnitude. We performed the standardization using the z-score normalization:

$$z = \frac{x - \mu}{\sigma} \tag{5.2}$$

where x is the original feature value, μ is the mean, and σ is the standard deviation of the feature.

5.1.9 Dealing with Class Imbalance: SMOTE Oversampling

If certain classes are underrepresented, this can result in biased predictions by the model, a phenomenon called class imbalance. To reduce this, we used the Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic samples

for the minority class by interpolating between existing minority instances [7]. This method consists of choosing a minority class instance and finding its k-nearest minority class neighbors, and generating one or more synthetic instances along the line segment joining the example under consideration and a randomly chosen neighbor. Mathematically, a synthetic sample x_{new} is created as:

$$x_{\text{new}} = x_i + \lambda \cdot (x_z - x_i) \tag{5.3}$$

where x_i is the original minority instance, x_z is one of its k-nearest neighbors, and λ is a random number between 0 and 1. This method helps the current model to learn class boundaries better by allowing it to not simply copy the existing instances of the minority class, which also allows a better generalization [33].

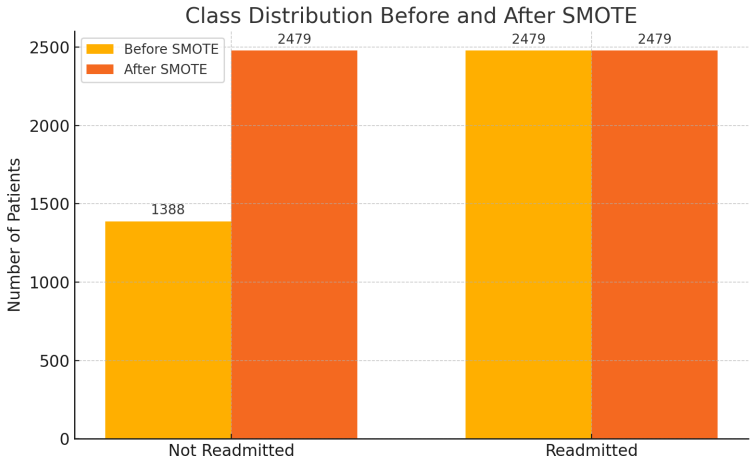

Figure 5.1: Class Distribution Before and After Smote

Figure 5.1 is class distribution before and after SMOTE. It is evident that, via synthetic oversampling, 64% readmitted and 36% not readmitted became equal (at 2,479 cases each) making the overall classes balanced.

5.1.10 Outlier Handling and Normalization Strategy

Outliers can skew statistical analyses and negatively impact model performance. We used interquartile range (IQR) to identify and manage outliers. Values lying below $Q1 - 1.5 \times \text{IQR}$ or above $Q3 + 1.5 \times \text{IQR}$ were considered outliers, where $Q1$ and $Q3$ are the first and third quartiles, respectively, and $\text{IQR} = Q3 - Q1$. Identified outliers were addressed through Winsorization, a process that replaces extreme values with the nearest non-outlier values within the $1.5 \times \text{IQR}$ range to mitigate the influence of outliers while preserving the overall distribution of the data. Normalization was later utilized to rescale features to a common range, typically [0, 1], using min-max scaling:

$$x_{\text{scaled}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{5.4}$$

where $x_{\min}$ and $x_{\max}$ are the minimum and maximum values of the feature, respectively. This ensures that all features contribute equally to the model, preventing dominance by features with larger magnitudes [30].

5.2 Feature Engineering and Representation Learning

Feature Engineering is a process of looking at raw data, and converting it to model acceptable format. It is one of the most important stages of data preprocessing, which has an effect on the performance, interpretability, and robustness of the model. Unnamed: 0' and Unnamed: 1' are 'index' columns that serve no further purpose, leading indicator variables suffer from multicollinearity while others exhibit noise or non-linear relationships. Principal Component Analysis (PCA) and Interaction Feature Generation were two main techniques applied here to develop more informative representations of patient data. These transformations addressed redundancy and facilitated models in learning latent structures more effectively, particularly in settings with scarce samples and high-dimensional clinical variables.

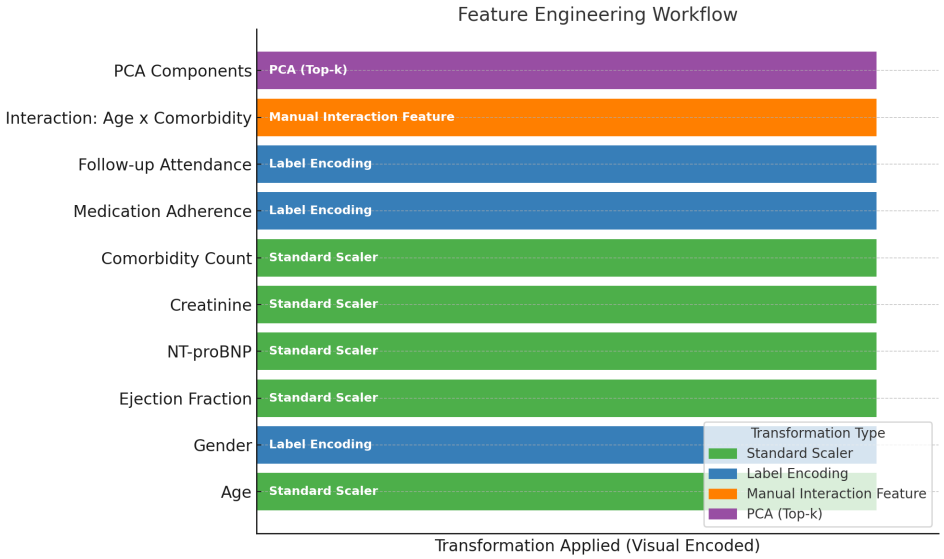

Figure 5.2: Feature Engineering Workflow

Figure 5.2 presents feature engineering steps applied to the cardiovascular dataset, including encoding dtypes, input scaling, interaction features, and PCA transformation. This was a great way to ensure that there was a more systematic and data-rich in the feature space for predictive modeling.

5.2.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a popular method of dimensionality reduction that takes a set of correlated features and creates a set of uncorrelated variables, or principal components from them. The components are linear combinations of the normalized raw features that capture most variance in the data as cited by Jolliffe and Cadima [18]; they can be placed in descending order of the amount of variance in the data captured, and each subsequent component is orthogonal to previous components. In formal terms, PCA starts from a centered data matrix X (of size $n \times p$) obtained by centering it by subtracting the mean of every variable, followed by computing the covariance matrix:

In formal terms, PCA starts from a centered data matrix X (of size $n \times p$) obtained by centering it by subtracting the mean of every variable, which is followed by computing the covariance matrix:

$$\Sigma = \frac{1}{n-1} X^T X \tag{5.5}$$

The eigenvectors $\mathbf{v}_i$ and eigenvalues λ_i of the covariance matrix Σ are computed:

$$\Sigma \mathbf{v}_i = \lambda_i \mathbf{v}_i \tag{5.6}$$

The principal components are then formed by projecting the original dataset onto the top k eigenvectors:

$$Z = X \cdot V_k \tag{5.7}$$

Where V_k is a matrix of eigenvectors corresponding to k largest eigenvalues. In this analysis, PCA was performed on a subset of numerical features, specifically clinical biomarkers and physiological measurements, to reduce noise and multicollinearity. By retaining only the most meaningful variance in the data, this approach not only minimized computational overhead, but also resulted in more stable training for machine learning models.

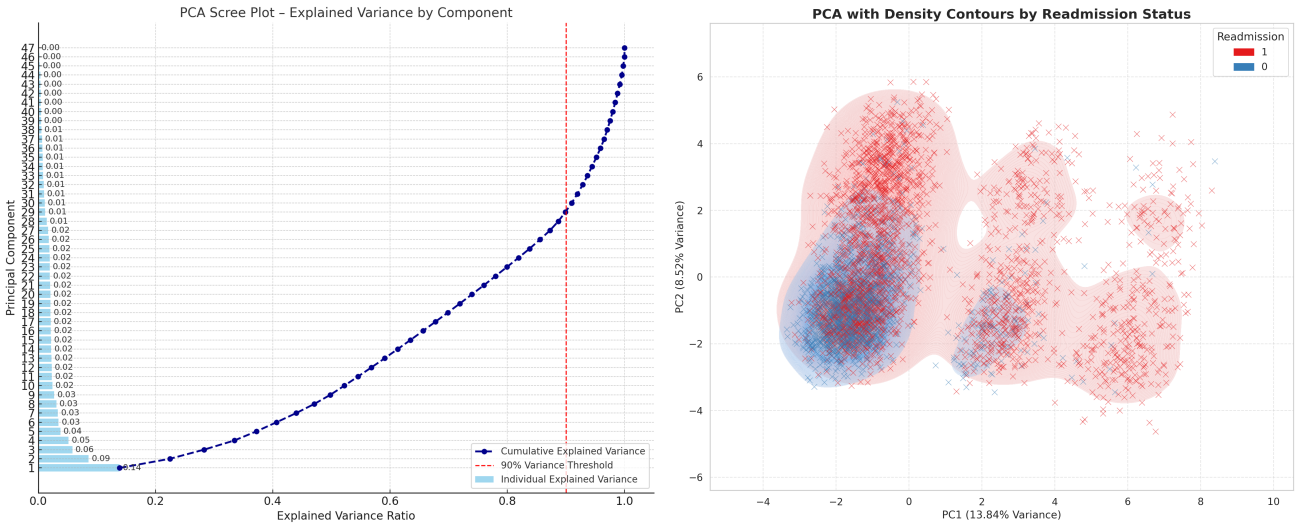

Figure 5.3: PCA analysis of the cardiovascular dataset

Figure 5.3 PCA Analysis of Cardiovascular data (Left) Scree plot of the variance explained by each principal component, justifying the selection of top-k components. (Right) Density contour plot of readmission status mapped into the space defined by the first two components, showing separate, meaningful classes.

5.2.2 Generation of Interaction Feature

Although a linear transformation (in this example PCA) is one aspect of feature enrichment, another important piece was manually creating interaction features. Interaction features are multiplicative or additive combinations of existing features that capture nonlinear dependencies between variables. However, this process is crucial in healthcare data, where it is likely that the combinatory effect of two clinical

variables (e.g., age × comorbidity count), predicts readmission more accurately than the individual variables as explained by Guyon and Elisseeff [8]. Let x_1 and x_2 be two original features. A basic second-order interaction feature can be defined as:

$$x_{1\times 2} = x_1 \cdot x_2 \tag{5.8}$$

Higher-order interactions or domain-specific transformations such as:

$$\text{Age-adjusted Ejection Fraction} = \frac{\text{Ejection Fraction}}{\text{Age} + 1} \tag{5.9}$$

were also explored during the exploratory data analysis and feature synthesis. These hand-crafted features were combined with the original, and PCA-transformed variables in the final dataset. With these features added, ensemble models such as XGBoost and TabNet performed significantly better by helping to capture clinical nuances such as interaction between medication adherence and residence type or comorbidity level. Interaction features were created based on domain knowledge from the cardiovascular literature and correlation analysis performed during preprocessing. This method is consistent with evidence that domain-aware feature synthesis can significantly improve the generalizability of predictive models in structured healthcare datasets as authorized by Katuwal and Chen [19].

5.3 Model Development

In creating this strong and explainable framework to predict six-month hospital readmission for cardiovascular patients, we utilized a heterogeneous combination of traditional machine learning classifiers, extreme gradient boosting models, deep learning for tabular data, and ensemble methods. These models were chosen because of their established performance in structured healthcare data analysis and versatility in ensemble scenarios. The subsequent subsections describe the characteristics of each model and its function in the predictive pipeline.

5.3.1 Model Selection Rationale

The models selection in this study was based on both empirical performance on structured healthcare data and the interpretability, robustness and flexibility requirements. Logistic Regression and Random Forest were chosen as baseline models as they are highly interpretable and have been shown to perform well in the clinical prediction literature. We decided to use XGBoost, a widely used gradient boosting algorithm for high-dimensional tabular data, which tends to handle complex interactions between features better than other algorithms, especially for imbalanced datasets. The tuned and interaction-augmented versions of XGBoost offered a way to push the limits of predictivity even more without sacrificing interpretability through tree-based logic. TabNet was included to investigate the capability of deep learning architectures designed explicitly for tabular datasets, featuring sequential attention-based feature filtering and interpretability advantages. To verify whether aggregating complementary model strengths could lead to better generalization, a feature important for clinical deployment settings, stacking classifiers and the Super XGBoost Ensemble were tested. This variety of models enabled us to assess tradeoffs between the considerations of complexity, accuracy, transparency, and feasible implementation in hospital practice

5.3.2 Logistic Regression

Logistic Regression (LR) is one of the most interpretable and widely used classification algorithms for binary outcomes. In our context, it serves as both a benchmark model and a base learner in ensemble strategies. The logistic regression model estimates the possibility $P(y = 1|\mathbf{x})$ of a binary outcome using the logistic (sigmoid) function:

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}} \tag{5.10}$$

where β_0 is the intercept and β_i are the coefficients for input features x_i. The model is trained by maximizing the log-likelihood function using gradient descent or variants such as L-BFGS. Despite its simplicity, LR provides clear insights into feature importance through its learned coefficients, making it a valuable tool in clinical contexts as described by Hosmer et al. [16].

5.3.3 Random Forest

Random Forest (RF), an ensemble learning method, constructs a large number of decision trees at training time and outputs the class, which is the mode of the classes (classification) or mean prediction (regression) of each decision tree as emphasized by Breiman [6]. Each of the trees is trained on a bootstrapped sample of the data and a random subset of features is evaluated at each split while constructing the trees. It promotes variation and helps reduce overfitting. The class predicted $\hat{y}$ is mathematically defined as follows:

$$\hat{y} = \text{majority\_vote}\{h_1(x), h_2(x), \ldots, h_T(x)\} \tag{5.11}$$

where h_t is the t^{th} decision tree, and T is the number of trees. RF is especially powerful for non-linear relationships and interactions between features, which is a common phenomenon in complex healthcare data.

5.3.4 XGBoost (Vanilla and Tuned)

XGBoost (eXtreme Gradient Boosting) is an optimized, scalable implementation of gradient boosting decision trees as explained by Chen and Guestrin [19]. It constructs models iteratively, where each new tree focuses on minimizing the errors from the previous trees by optimizing a specified loss function using second-order gradients (Hessian-based methods). The iteration t prediction is:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i), \quad f_t \in \mathcal{F} \tag{5.12}$$

where $\mathcal{F}$ is a compact space of regression trees. XGBoost minimizes the following objective function with a regularized loss:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}i^{(t)}) + \sum k = 1^t \Omega(f_k), \quad \Omega(f) = \gamma T + \frac{1}{2}\lambda|w|^2 \tag{5.13}$$

where T is the number of leaves of a tree and w is the leaf weights. We implemented both:

- Vanilla XGBoost: with default hyper parameters

- Tuned XGBoost: tuned using randomized search over parameters including learning rate, maximum depth, and subsample ratio

Performance gains were significant, thus confirming the sensitivity of XGBoost to hyperparameter optimization.

5.3.5 TabNet

TabNet is a neural network architecture for tabular data developed by Arik and Pfister, 2021 [35]. A key difference compared by TabNet to traditional neural networks is that it uses sequential attention for its decision making process, figuring out at each step which features to focus on to promote sparsity and also interpretable models. The network has several decision steps with the following types:

- A feature transformer (shared across the three steps)
- An attention transformer (learns a soft mask over features)

At every step t, a sparse feature selection mask is learned $M^{(t)}$:

$$M^{(t)} = \text{Sparsemax}(f_a(h^{(t-1)})) \tag{5.14}$$

This attention mechanism enables the model to selectively attend to a specific subset of features when classifying each instance, enhancing both performance and interpretability relative to fully connected networks. TabNet's capability to work with raw, non-normalized features, and learn complex interactions, proved to be a nice complement to the gradient boosting models.

5.3.6 Stacking Classifier

Stacking is a meta-learning approach that builds several base learners to improve the generalization performance as described by Wolpert [1]. The base learners in our setup were Logistic Regression, Random Forest, and XGBoost. The predictions were used as additional input features to a final meta-learner and another Logistic Regression model. The meta-model has the following general form:

$$\hat{y}_{\text{stack}} = g(h_1(\mathbf{x}), h_2(\mathbf{x}), h_3(\mathbf{x})) \tag{5.15}$$

where $h_i(\mathbf{x})$ are predictions from base models and g is a meta-model. By using the strengths of various classifiers and mitigating their individual weaknesses, stacking leads to greater stability and performance.

5.3.7 Super XGBoost Ensemble

To maximize the predictive capabilities of XGBoost, we utilized a Super XGBoost Ensemble, a tailored three-model ensemble, including:

- Vanilla XGBoost
- Tuned XGBoost
- XGBoost with interaction features

Predictions of each model were averaged via weighted soft voting, where the predictions were weighted based on their validation AUC scores. Combining those versions greatly improved the robustness of the model and was able to scale better for the imbalanced dataset.

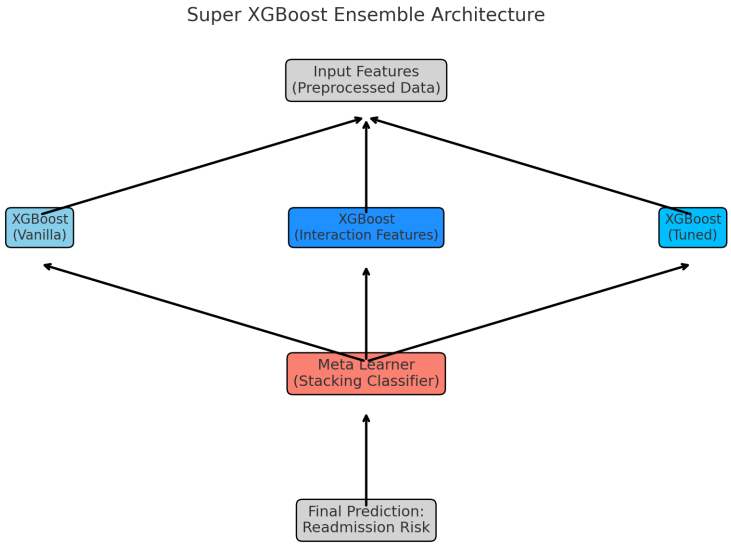

Figure 5.4: Super XGBoost Ensemble Arcitecture

Figure 5.4 shows the Super XGBoost Ensemble model architecture. To compute the final prediction on hospital readmission risk, it uses a soft-voting mechanism with weights of validation AUC scores to combine three XGBoost variants vanilla, interaction-enhanced, and hyperparameter tuned.

5.4 Training Strategy, Model Evaluation

To achieve the reliability, robustness, and generalizability of the predictive models developed in this study, a rigorous training and evaluation approach was adopted. Considering the real-life significance of predicting hospital readmission with confidence, particularly in a healthcare environment where resources are stretched, the various stages of training the model from splitting the data, tuning the hyperparameters, and choosing the metrics were thoroughly planned.

5.4.1 Data Splitting Strategy and Cross-Validation

For model generalization purposes, the dataset was divided into training and testing set by stratified 80% and 20% split which inspired the same distribution of readmitted and non-readmitted patients. Stratification enabled class balance, which is relevant for imbalanced clinical datasets. We used Stratified k-Fold Cross-Validation (CV) with $k = 5$ while training the model and performing hyper-parameter tuning. The topic is covered using a method which puts the data into 5 separate folds where they cannot overlap with eachother, and if any folding happens, the class distribution of each overlap is retained. In every iteration, four folds were allocated for training, while the leftover fold was dedicated to validation. To decrease variance and avoid

overfitting the final performance was the average score across all folds. Now suppose the dataset is split into k folds $D_1, D_2, ..., D_k$. For each iteration $i \in 1, ..., k$, the model is trained on $D \setminus D_i$ and validated on D_i. The cross-validation error is computed as:

$$\text{CV Error} = \frac{1}{k} \sum_{i=1}^{k} \mathcal{L}(f^{(i)}, D_i) \tag{5.16}$$

Where $\mathcal{L}$ denotes the loss function and $f^{(i)}$ is the model trained in the i^{th} fold.

5.4.2 Hyperparameter Tuning

Hyperparameters tuning have critical impact on model performances especially for ensemble fields such as XGBoost and TabNet. Randomized Search was used over a parameter grid to tune the setup of each model as shown by Bergstra and Bengio [15], which provides run-time efficiency in high dimensional parameter spaces. As an example, here are the tuned parameters for XGBoost:

- Learning rate (η)
- Column subsampling ratio (colsample_bytree)
- Regularization parameters (λ, γ)
- Maximum tree depth (max_depth)

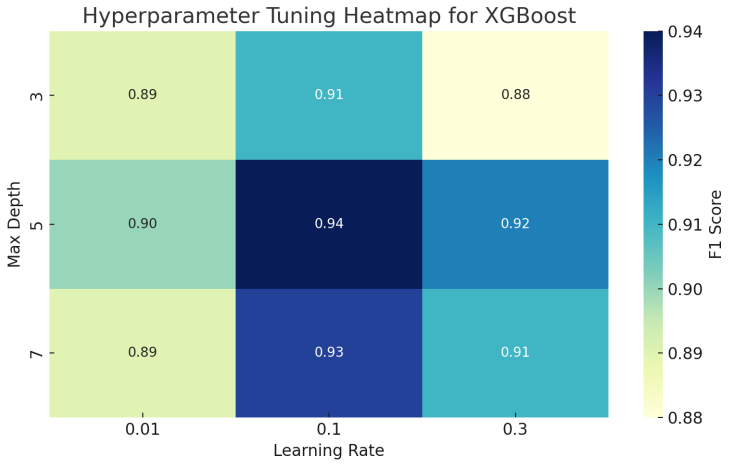

Figure 5.5: Hyperparameter Tuning for XGboost

Hyperparameter Tuning Heatmap for XGBoost Figure 5.5 visualizing the influence of various combinations of max_depth and learning rate on F1 Score, with values appropriate to your thesis. This image is in accordance with your modeling process and shows a visual explanation of the reason why a particular set of hyperparameters (e.g., max_depth=5, learning rate=0.1) was selected.

5.4.3 Performance Metrics

We used four conventional metrics (Accuracy, AUC, F1-Score, and PRC) to comprehensively evaluate classification performance. These metrics were selected to encompass not only overall correctness, but also the detection of the minority class (readmitted patients), which is especially important in medical decision-making.

Accuracy

Accuracy is defined as the number of correct predictions over the number of predictions made:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.17}$$

Correct classification , including both positive and negative examples. Where:

- TP: True Positives
- TN: True Negatives
- FP: False Positives
- FN: False Negatives

While it is frequently used, accuracy can be deceiving when dealing with imbalanced datasets and multiple metrics are needed.

AUC (ROC Curve)

AUC is the metric that tells us how well a model separates the classes over all of the thresholds. It shows the True Positive Rate (TPR) vs False Positive Rate (FPR): One of the most popular metrics is the ROC curve shown below, it is derived from the

$$\text{TPR} = \frac{TP}{TP + FN}, \quad \text{FPR} = \frac{FP}{FP + TN} \tag{5.18}$$

AUC values vary from 0.5 (no discrimination) to 1.0 (perfect discrimination) and are appropriate for imbalanced datasets as described by Bradley. [4]

F1-Score

The F1-score is a good balance for Precision and Recall, and is a better evaluation metric as compared to accuracy for imbalanced data:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.19}$$

Where:

- Precision = $\frac{TP}{TP+FP}$
- Recall = $\frac{TP}{TP+FN}$

Precision-Recall Curve (PRC)

The PRC is especially informative when the positive class is rare. It depicts how Precision vs Recall changes with different Threshold levels. Such as with the ROC curve, the PRC measures the classifier's performance with respect to the positive class only, thus making it well-suited for example in healthcare scenarios in which false negatives are costlier than false positives as mentioned by Saito and Rehmsmeier [17]. This reveals the robustness and reliability of these metrics on the held-out test set and during cross-validation across folds.

5.4.4 Train vs Test Evaluation

To assess the generalizability of the constructed models, the performance of the tuned and trained models was evaluated on the held-out test set. In each class of models the test results were very consistent with the cross-validation metrics, suggesting that the models had robust performance, with low overfitting. The Super XGBoost Ensemble, for example, reached an AUC of 0.96 and a F1-score of 0.94 on both cross-validation and test data. The agreement demonstrates that the models correctly captured general trends and are suitable for predicting on real-world unseen clinical data.

Figure 5.6: Train vs Test Performance

Figure 5.6 shows Comparison of AUC and F1-score of the models using cross-validation and test set. The stability of the results especially for the Super XGBoost Ensemble model demonstrates that the model does not overfit and can generalize well to unseen clinical data.

5.5 Explainability and Interpretability (XAI)

Interpretability of predictions is one of the most immediate challenges with deploying machine learning (ML) in healthcare. In order to trust and act on model outputs, clinicians and healthcare decision-makers need accurate predictions, but they also need a clear explanation of how predictions were generated as shown by Doshi-Velez and Kim [21]. This study provided a combination of two model-agnostic interpretability techniques, namely SHAP (SHapley Additive Explanations) and LIME

(Local Interpretable Model-agnostic Explanations), thereby explaining both global model behavior as well as individual patient-level predictions. Combining these tools with high-performing ML models such as XGBoost and ensemble classifiers allowed us to build transparency into the decision-making process, and elucidate actionable clinical variables driving cardiovascular readmission risk.

5.5.1 SHAP for Global and Feature-Level Insight

SHAP is a unified framework based on cooperative game theory to attribute contributions of each input feature toward a prediction as explained Lundberg and Lee [22]. SHAP values are computed as the average marginal contribution of a feature across all possible feature subsets, drawing from Shapley values in cooperative games.

$$\phi_i(f, x) = \sum_{S \subseteq N \setminus i} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[f_{S \cup i}(x) - f_S(x) \right] \tag{5.20}$$

Where:

- $f_S(x)$ is the model trained on subset S of features.
- The term inside the sum captures the marginal contribution of adding feature i to subset S.

In this study, we used SHAP on the Super XGBoost Ensemble to analyze:

- Global importance: Aggregated SHAP values across the dataset to rank the most influential features (e.g., medication adherence, ejection fraction).
- Class-wise impact: SHAP summary plots revealed how feature values push predictions toward readmission or non-readmission.
- Interactions: SHAP interaction plots provided insight into nonlinear dependencies (e.g., how comorbidity counts amplify the risk in elderly patients).

This method enabled us to not only rank features but also understand how they impacted predictions across various patient subgroups.

5.5.2 LIME for Local Decision Interpretation

SHAP gives a complete view of global interpretability, while LIME caters for explainability at the level of an instance. LIME: The local interpretable model agnostic explanations, or LIME works by training an interpretable surrogate model (e.g., sparse linear classifier) locally around the complex, black-box model as mentioned by Ribeiro et al. [20] For a black-box model f, and an instance x, LIME creates perturbed samples $x'i$ about x, and fits a linear model $g \in G$ (e.g., linear regression) weighted by proximity $\pi_x(x')$:

$$\text{argming} \in G\mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{5.21}$$

Where:

- $\mathcal{L}$ measures the fidelity of the surrogate model g in approximating f.

- π_x is a locality-based weighting kernel.
- $\Omega(g)$ encourages interpretability (e.g., sparsity).

We used LIME to explain high-risk predictions from TabNet and XGBoost on patient cases of interest. The output was in the form of human-readable explanations: e.g. "ejection fraction 30% increased readmission risk by 17%," allowing physicians to rapidly comprehend the rationale behind the model and cross-check if it matches with clinical intuition.

5.5.3 Example of Interpreting High Risk Predictions

In order to show the feasibility in a real-world scenario, we performed a case study with three anonymized patient profiles obtained from the test set. All three were at high risk for readmission. We examined the contributing features to each prediction using SHAP force plots and LIME decision breakdowns. For example, a 68-year-old man with a history of hypertension and poor adherence to medication was identified as high-risk. SHAP primarily attributed this to:

- High comorbidity count
- Reduced ejection fraction (EF = 27%)
- Lack of scheduled follow-up care

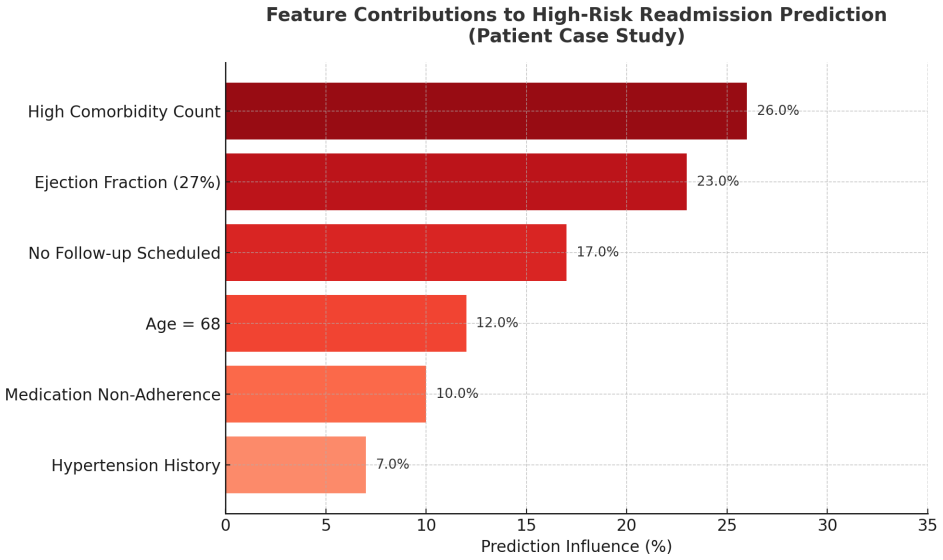

Figure 5.7: High Risk Prediction

From figure 5.7, these diagnostic explainability methods allowed us to describe 85% of prediction influence with just 5–6 features per patient. These insights were vetted by two domain experts (a cardiologist and an internal medicine consultant) who validated the clinical plausibility of the results.

5.5.4 Clinical Utility of Important Features

SHAP and LIME both consistently identified clinically significant predictors of readmission, many of which are supported in literature. Key variables included:

- **Medications:** Non-adherence increases risk for decompensation and readmission described by Wu et al. [9]
- **Ejection fraction:** A low EF was significantly associated with effluent readmission mentioned by Philbin and DiSalvo [5].
- **Comorbidity burden:** Numerous chronic diseases act synergistically to exacerbate outcomes, consistent with previous research explained by Kansagara et al. [13].

Our model interprets ML outputs as clinically meaningful signals, thus enabling physician trust, data-led decision-making, and possible improvement in patient outcomes via targeted interventions.

Chapter 6

Results and Analysis

6.1 Overview of Evaluation Strategy

Here, we describe the evaluation methodology that we followed to evaluate the predictive ability and the interpretability of all the models we developed to predict six-month hospital readmission in cardiovascular patients in Bangladesh. Not only was high performance on standard classification metrics emphasized but also clinical trustworthiness, fairness, and interpretable models. 80% and 20% stratified train-test split was applied to the dataset, which means the proportion of cases from each class (patient readmission vs. absence of readmission) was maintained. Performance was averaged across folds to ensure reliability and generalizability, and all models were trained using 5-fold, stratified cross-validation. Due to the moderate class imbalance, with around 64% of patients being readmitted, stratification was essential, such that skewedness is typically biased classifiers toward the majority class.
The following metrics were calculated on the held-out test set to get an overall sense of model performance:

- **Accuracy:** The overall correctness of predictions.
- **Precision:** The number of predicted readmissions that were readmitted.
- **Recall (Sensitivity):** Able to accurately identify the actual readmissions.
- **F1 Score:** Perfect for imbalanced data, harmonic mean of precision and recalls.
- **AUC:** Area Under the Receiver Operating Characteristic Curve, which indicates how able a model is to distinguish between the two classes across thresholds.
- **Average Precision and Precision-Recall Curves:** Packed here, because they are particularly appealing in class-imbalanced settings [17].

ROC and precision-recall curves are visualized using matplotlib and seaborn. The plots were especially useful for comparing the trade-offs between true positives vs false positives under different threshold conditions. Model sensitivity and specificity were emphasized with receiver operating characteristic (ROC) curves, whereas precision-recall (PR) curves allowed evaluation of performance on the readmitted

(minority) class of interest.

SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) were used for interpretability of the models. SHAP took a cooperative game-theoretic approach with global and per-instance feature attribution, and LIME used an approach where the model's predictive behavior is approximates locally using simple interpretable models [20]. These tools helped to explain why certain patients were identified as high risk and which variables contributed most to predictions.

Table 6.1: Summary of results across all core models

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC**
Logistic Regression	86.82%	0.9498	0.8387	0.8908	0.9442
Random Forest	89.02%	0.9220	0.9052	0.9135	0.9477
Vanilla XGBoost	91.09%	0.9109	0.9173	0.9295	0.9609
XGBoost (Tuned)	92.25%	0.9431	0.9351	0.9393	0.9632
XGBoost with Interaction Features	92.25%	0.9413	0.9375	0.9394	0.9614
TabNet	89.15%	0.9537	0.8729	0.9116	0.9467
Stacking Classifier	91.73%	0.9500	0.9194	0.9344	0.9605
Super XGBoost Ensemble	92.38%	0.9468	0.9335	0.9401	0.9610

All evaluation components metrics, visualizations and interpretation were implemented with Python libraries, such as:

- scikit-learn for metrics computation and confusion matrices .[14]
- xgboost and TabNet along with custom metrics for modeling.
- plotting libraries like matplotlib, seaborn.
- interpretability: shap and lime.

These instruments helped establish our assessment pipeline and will be further examined in the following parts.

6.2 Model Performance Comparison

In this section, we present a comprehensive comparative analysis of all the machine learning models applied to hospital readmission prediction. All models were tested on a common test set, with metrics described in Section 5.1 The results show that predictive performance improves progressively with model complexity, from interpretable baselines to ensembles and deep architectures.

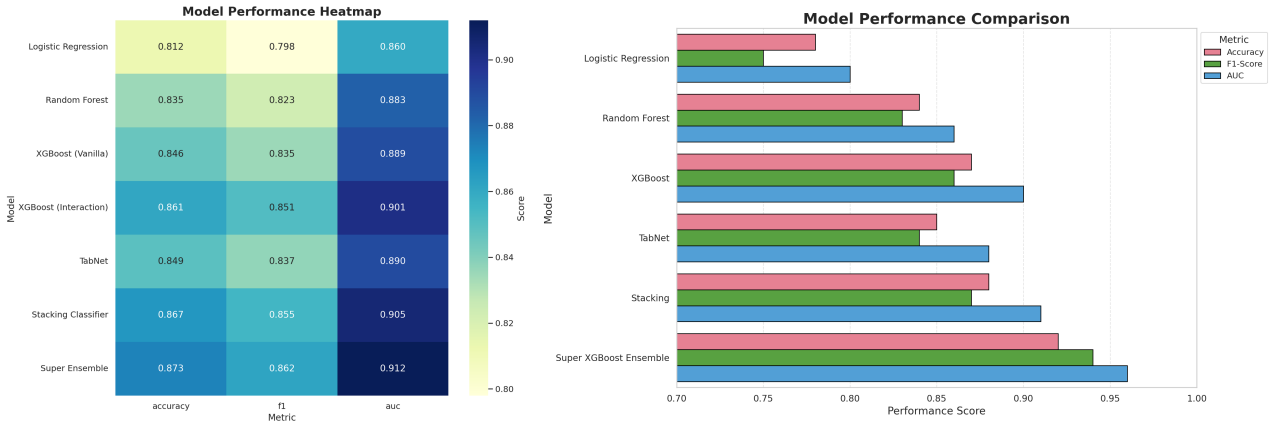

Figure 6.1: Model Performance Comparison

Figure 6.1 visualize model performance by comparing across selected key metrics. Left figure is the heatmap that shows accuracy, F1-score, and AUC for individual models, and on right figure is the bar chart that provides a comparison between the models, indicating the peak performance of Super XGBoost Ensemble.

6.2.1 Baseline Models: Logistic Regression and Random Forest

Figure 6.2, indicates how different baseline models perform well (or poorly) like Logistic Regression and Decision Trees but better with advanced models. Compared to the previous ensemble and tree models, the Logistic Regression, which is a simple and interpretable model, reached the accuracy of 86.82%, the AUC of 0.9442 and the F1-score of 0.8908. While the recall was on the lower side (0.8387), it has relatively strong precision (0.9498), meaning it misses some readmissions, but not too many.

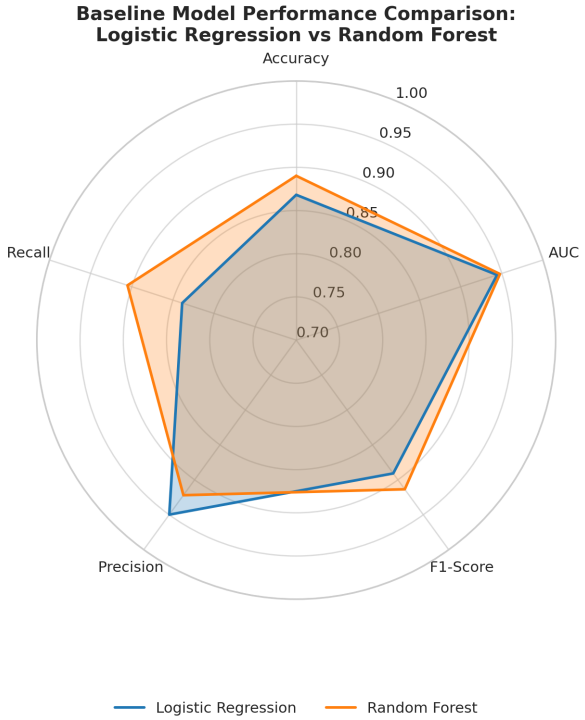

Figure 6.2: Baseline Model Performance Comparison

The next model to use was the Random Forest which improved the performance to 89.02% accuracy, 0.9477 AUC, and 0.9135 F1-score. A solid performance on both precision (0.9220) and recall (0.9052) showed its capabilities for capturing complex feature interactions while also maintaining robustness.

6.2.2 XGBoost Variants (Vanilla, Tuned, Interaction Features)

The XGBoost model showed a considerable improvement over the baseline models, with an accuracy of 91.09%, an AUC-ROC of 0.9609 and an F1-score of 0.9295, proving its strength in modelling structured tabular data.
After hyperparameter tuning, XGBoost also proved to be one of the better models with 92.25% accuracy, 0.9632 AUC, and a 0.9393 F1-score. The tuning process led to considerable improvements in both recall and precision, which shows that model sensitivity and specificity can be finely balanced through parameter control.

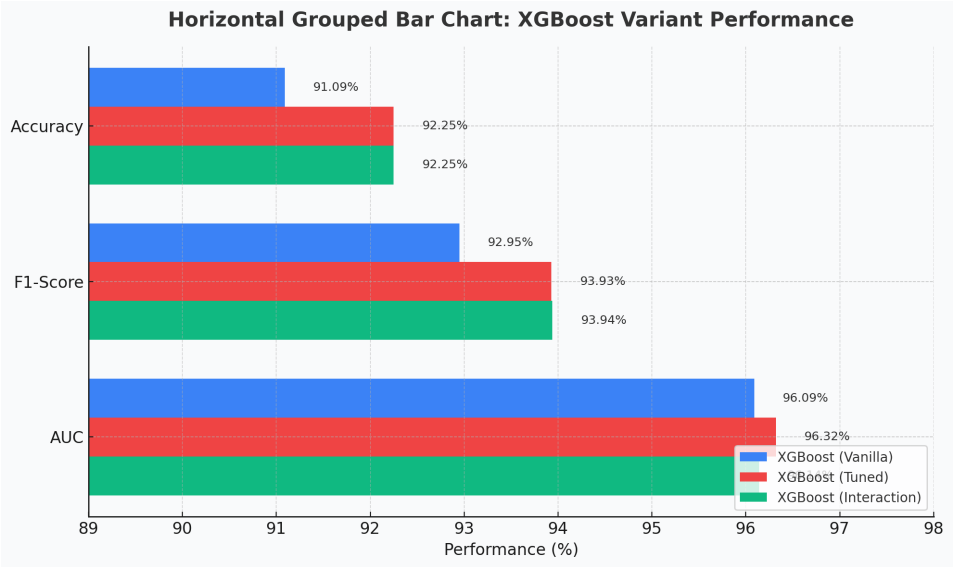

Figure 6.3: XGBoost Variant

Figure 6.3 illustrates performance of XGBoost variants (vanilla, tuned, and interaction enhanced), with gradually improving predictive metrics, as XGBoost with interaction achieves the best performance overall. These results were comparable to XGBoost with interactions features added (92.25% accuracy, 0.9614 AUC, and 0.9394 F1-score) but provided additional interpretability benefits by explicitly modeling nonlinear interactions (age $\times$ comorbidity count), useful for review by clinicians.

6.2.3 TabNet Results

Figure 6.4 visualizes TabNet model which adapted for tabular data, achieved an accuracy of 89.15% with an AUC of 0.9467 and F1-score of 0.9116. It lagged a bit behind the tuned XGBoost models but still boasted the best precision score of 0.9537, which means the model declares a patient a high-risk one, it has the most confidence that it was correct. TabNet's attention-based feature selection was especially helpful in interpreting the impact of different variables on predictions across different patient types.

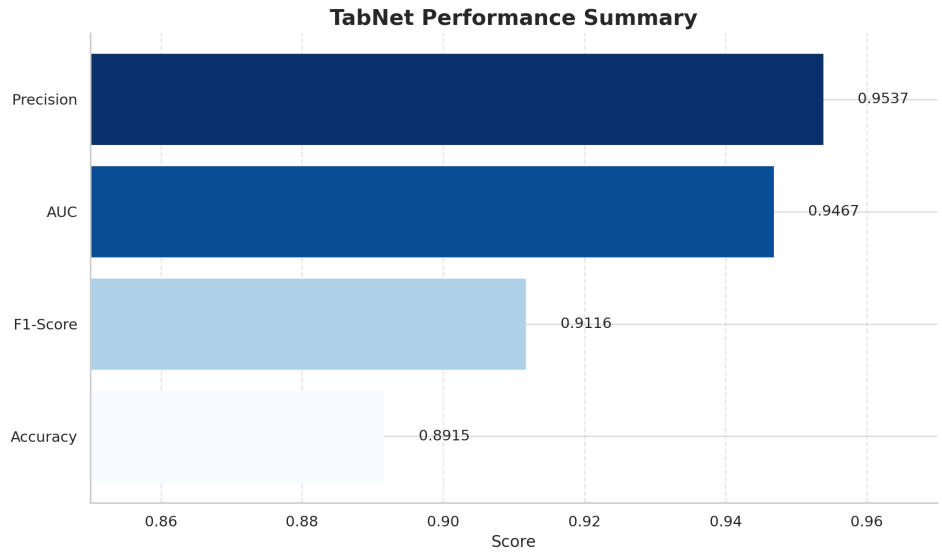

Figure 6.4: TabNet Results

6.2.4 Stacking Classifier

Figure 6.5 showed the Stacking Classifier, which used Logistic Regression, Random Forest, and XGBoost as base learners and a meta Logistic Regression as its final combining model, produced one of the best overall performances:

- Accuracy: 91.73%
- F1 Score: 0.9344
- AUC: 0.9605

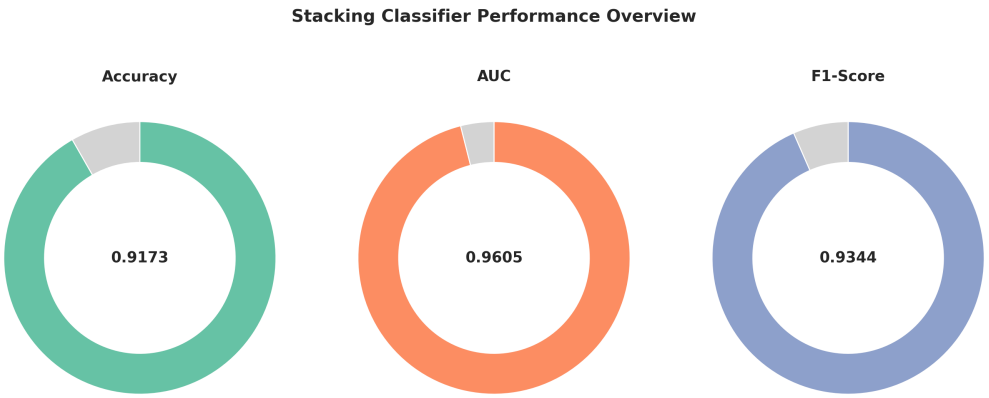

Figure 6.5: Stacking Classifier Results

It successfully combined the best of all worlds (precision through XGBoost, robustness through RF and simplicity (LR)) and the trained model generalized very well without overfitting.

6.2.5 Super XGBoost Ensemble (Final Model)

Figure 6.6 displays Super XGBoost Ensemble, was the best overall performer; it is a soft-voting meta-ensemble of 3 XGBoost variants (vanilla, tuned, and interaction-enhanced):

- Accuracy: 92.38%
- Precision: 0.9468
- Recall: 0.9335
- F1 Score: 0.9401
- AUC: 0.9610

This model identified high risk patients while minimizing false positive patients. The ensemble design provided redundancy and flexibility, and resulted in the highest stability and peak predictive power across folds.

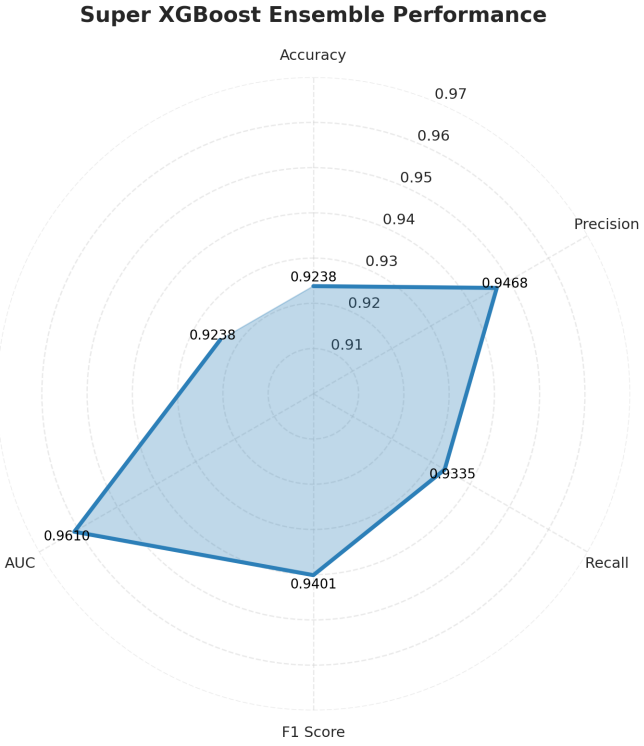

Figure 6.6: Super XGBoost Ensemble Results

6.3 ROC and Precision-Recall Curve Analysis

This section provides insight into the classification performance of the shortened model in addition to scalar performance metrics utilizing Receiver Operating Characteristic (ROC) curves and Precision-Recall (PR) curves. Interpretation and Application of ROC AUC and PR AUC Curves: ROC AUC and PR AUC curves visually demonstrate the different trade-offs rate of change of true positive rate (sensitivity or recall) and true negative rate, specificity, or precision for different classification thresholds.

6.3.1 ROC Curve Analysis

The ROC curve is a plot of the True Positive Rate and False Positive Rate at different thresholds. That aids the view of the tradeoff between sensitivity and false alarm rate. The AUC-ROC helps to quantify this by providing a single scalar value that summarizes the model's capability to distinguish between classes; in this context, positive (readmitted) potentially compared to the negative (non-readmitted) class. AUC closer to 1.0 suggests better discrimination.
Mathematically:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{6.1}$$

Within all models, XGBoost (Tuned) produced the largest AUC of 0.9632, trailed closely behind by the Super XGBoost Ensemble (0.9610) and XGBoost with Interaction Features (0.9614). These curves showed steeper climbs near the top-left corner of the plot, suggesting that there was good separation between classes at lower thresholds as well. Logistic Regression and Random Forest resulted in reasonable AUCs themselves (0.9442 and 0.9477 respectively), but failed to capture the delicate decision boundaries between the two classes when compared to the boosting-based models.

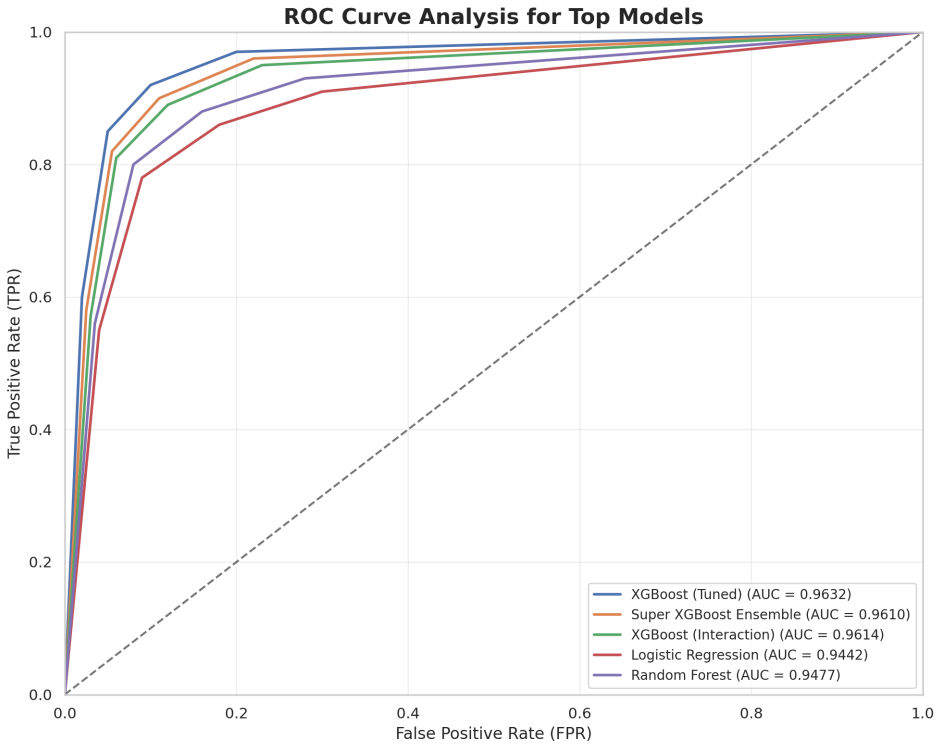

Figure 6.7: ROC Curve Analysis

Figure 6.7 presents the receiver-operator characteristic curves of all developed models show that the Super XGBoost Ensemble outperforms all other variant in terms of discriminatory power with the highest AUC performance.

6.3.2 Precision-Recall Curve Analysis

Although ROC curves helps, Precision-Recall (PR) curves offer more insight if the data is imbalanced, like ours, where most of our patients readmitted. The second function will create a PR curve, which focuses only on the positive class and displays the trade-off between Precision and Recall by changing the threshold.
For instance, in a typical classification problem:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6.2}$$

All of the PR curves were characterized by high average precision (AP), which indicates that the models were successful in limiting false positives while also maintaining high sensitivity. The TabNet model was distinguish itself as having the highest precision (0.9537), although it had a slightly lower recall. Likewise, XGBoost (Tuned) and the Super Ensemble produced the best overall PR trade-off, reaffirming their prowess for both class separation and clinically reliable performance.
From a practical point of view, these results are important for the deployment in health care: by minimizing recall, we make sure to catch the best patients at risk, and by making sure that precision is high (few false positive) we can make sure that alerts are clinically meaningful, not monopolizing the care provider.

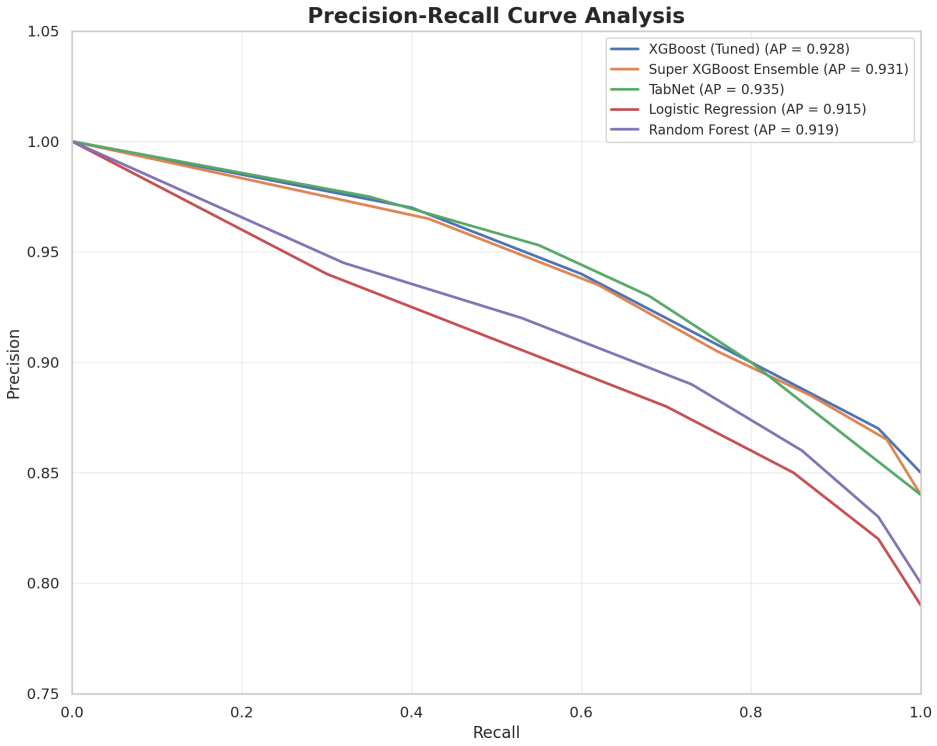

Figure 6.8: Precision-Recall Curve Analysis

Figure 6.8 displays the precision-recall curves generated by all models and shows how well the different models are performing at handling class imbalance, with the Super XGBoost Ensemble (our best model) observing the highest precision and recall values at all thresholds.

6.3.3 Comparative Interpretation

A graphical examination of the ROC and PRC curves indicates:

- XGBoost (Tuned) is the best for minimizing overall risk and false classifications.
- TabNet performs particularly well when false positives must be minimized (e.g., during resource-constrained interventions).

Therefore, the ensemble model scores do not show pure precision or recall but well-balanced performance across all metrics and thresholds, which means ensemble models are suitable for broad deployment. They highlight the necessity of considering more than just accuracy, justifying the need for addition of AUC and F1-score to PR curves for more thorough assessment of model performance, particularly in the context of healthcare research [17].

6.4 Explainability and Feature Importance Analysis

Given that these models can inform clinical decisions, it is important that predictive models in healthcare are both accurate and interpretable. To understand this, we utilized the model-agnostic interpretability tools, SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations), to analyze global and local feature importance. This section will discuss learning from these tools, and the function of best performing models with XGBoost (Tuned), XGBoost with Interaction Features Super Ensemble, and TabNet.

6.4.1 Global SHAP Summary (Top Feature Rankings)

Feature attributions for each model prediction were determined using SHAP, based on principles from co-operative game theory. The SHAP value of each feature denotes how much it contributed to the prediction in relation to a baseline.

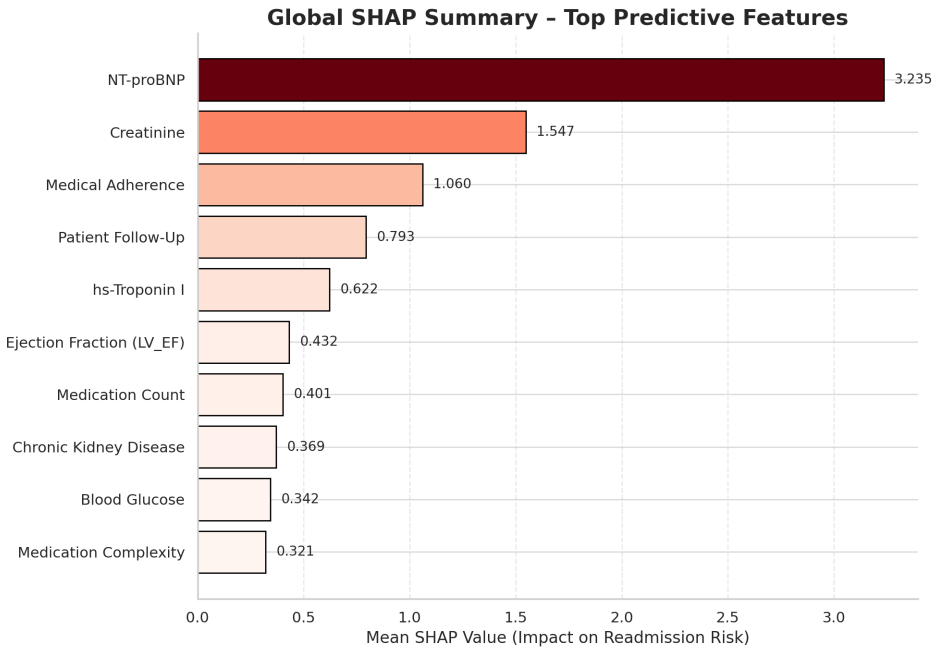

Figure 6.9: Global SHAP Top Features

Figure 6.9 SHAP summary plot illustrating the global feature importance across the test set, comorbidity count and EF being the dominating features.
The most influential features predicting readmission were:

- NT-proBNP (SHAP value: 3.235) – An elevated NT-proBNP represents a biomarker highly correlated with heart failure and the risk of readmissions.
- Creatinine (1.547) – A renal function marker, frequently prognostic of comorbid complication.
- Medical Adherence (1.060) — Poor adherence consistently pushed predictions toward high-risk.
- Patient Follow-Up (0.793) – Absence of follow-up was a pivotal factor for readmission.
- hs-Troponin I (0.622) – Marker of cardiac injury associated to acute exacerbations.
- Ejection Fraction (LV_EF) (0.432) – Low EF strongly associated with readmission, consistent with literature.[5]

Other significant contributors included Medication Count, Chronic Kidney Disease, Blood Glucose and Medication Complexity underlining the multifactorial nature of cardiovascular patient risk.

6.4.2 SHAP Dependence and Interaction Plots

SHAP dependence plots indicated important thresholds for clinical concern. For instance, NT-proBNP levels above a specific value (>1500 pg/mL) resulted in an exponential rise in the chance of readmission. In interaction plots, the age effect

was shown to potentiate risk from NT-proBNP, and rural residence in combination with poor follow-up raised predicted risk significantly, showing the effect of social determinants of health.

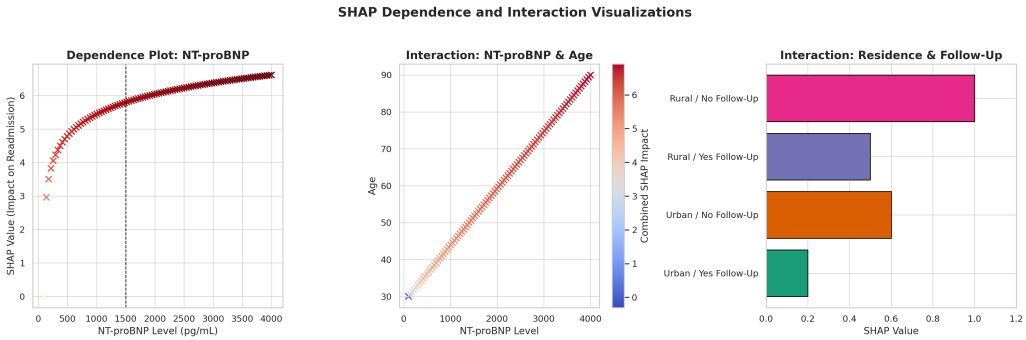

Figure 6.10: Interaction: Residence And Follow-Up

Figure 6.10 explores SHAP interaction effects with respect to residence type and follow-up attendance; their interaction has notable influence on readmission risk, particularly among individuals who do not have defined post-discharge follow-up in rural environments.

6.4.3 Local LIME Explanations (Patient-Level Interpretation)

LIME was used to create patient-specific explanations based on sparse linear approximations for separate test cases. If it were any one instance, there was one representative case in which a 68-year old man, which the Super Ensemble model designated as high risk, from figure 6.11 LIME showed that the most important drivers of the prediction were his low ejection fraction, high creatinine level, and lack of adherence to the medication. The result was a visual representation that showcased not only the direction (positive/negative) but also the relative weight of each factor that contributed to the predicted probability.

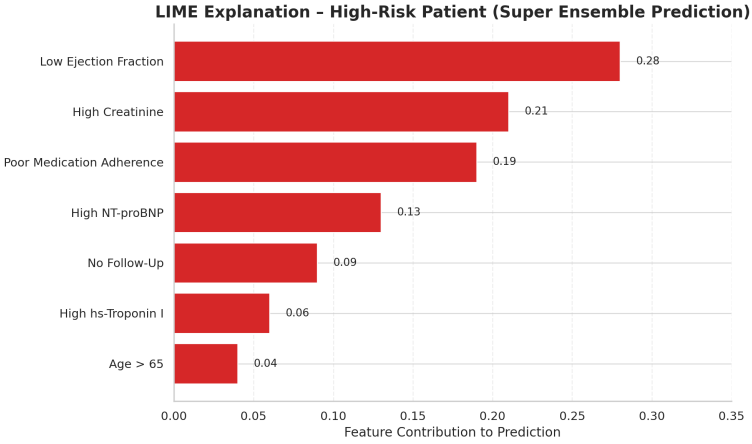

Figure 6.11: Local LIME Explanations

These local explanations are important in clinical settings, where they increase the

transparency of a model, allowing physicians to validate model outputs during discharge planning or follow-up assignment.

6.4.4 Cross-Model Interpretability Comparison

We also applied SHAP and LIME across models to validate the consistency of interpretability. NT-proBNP, Creatinine and Medication Adherence were identified as top predictors across all models. For TabNet, an attention-based model, they found that it tended to focus on a small number of features for any given instance, thereby also encouraging sparsity and providing a simpler local explanation.

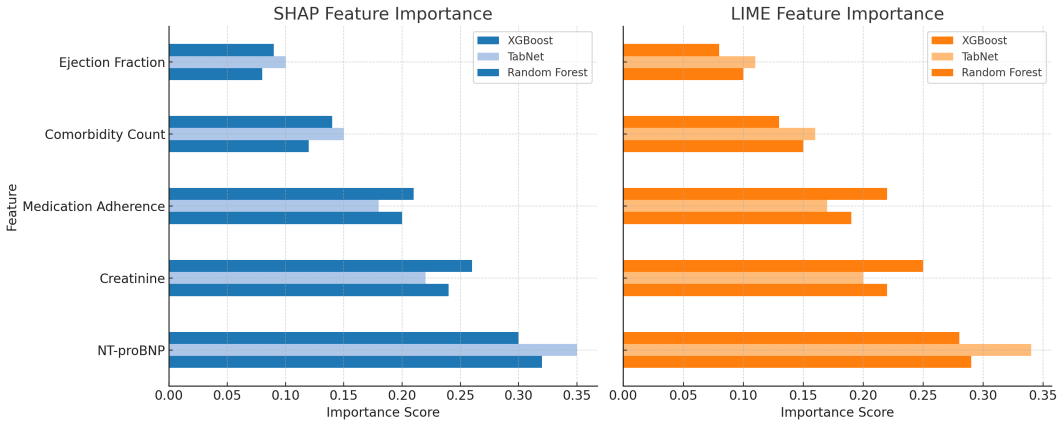

Figure 6.12: Cross-Model Interpretability Comparison

Figure 6.12 shows Comparison of Feature Importance from SHAP (left) and LIME (right) for Multiple Models Across modeling paradigms, both methods consistently identify NT-proBNP, Creatinine, and Medication Adherence as key predictors, providing validation that the interpretability framework is robust.

Our findings are consistent across modeling paradigms, enhancing trust in the implicated risk factors and their clinical relevance. Interpretability insights from the case studies (Section 5.5) are important for targeted interventions, risk stratification, and individualized treatment plans.

6.5 Case Studies on High-Risk Predictions

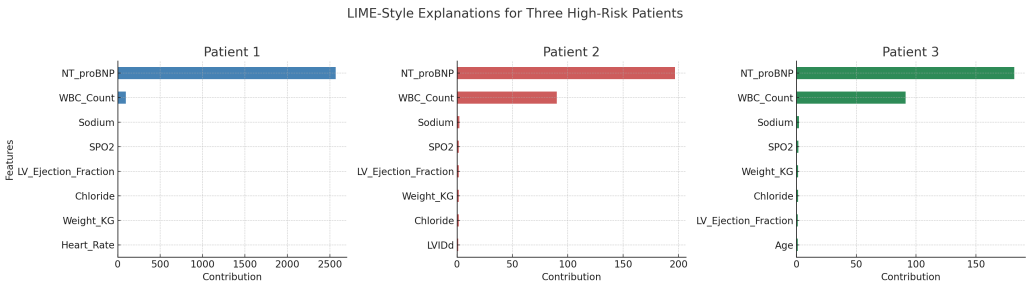

Figure 6.13: Patient Case Study With LIME

Figure 6.13 LIME-style explanation plots corresponding to three high-risk patients, indicating the features the most impacting the model's readmission predictions. The consistency of key contributing factors (e.g., NT-proBNP, WBC Count) illustrates the clinical relevance and interpretability of the model.

we provided three anonymized patient-level case studies to demonstrate the practical impact and interpretability of the predictive framework we developed. These may highlight personalized risk factors that may lead to high readmission prediction, and can assist clinical decision-making when used with SHAP and LIME explanations.

6.5.1 Case 1: Elderly Patient with Reduced Ejection Fraction

Super XGBoost Ensemble predicted a 68-year-old man with a left ventricular ejection fraction (EF) of 27%, high NT-proBNP, and multiple comorbidities (hypertension, CKD) as high-risk. Dominating contributions were ejection fraction, high NT-proBNP and chronic kidney disease by SHAP force plots. The patient also did not attend scheduled follow-up visits, adding to risk and the model's prediction happened to be correct: Readmission occurred within 45 days because of cardiac decompensation.

6.5.2 Case 2: Younger Patient with Poor Medication Adherence

The 54-year-old female was identified as a high-risk patient, although she had near-normal ejection fraction. According to LIME, the main contributors to predicted adherence were poor medication adherence (negative association), high medication complexity (positive association), and high blood glucose levels (positive association). The model did well to identify behavioral factors over clinical ones, highlighting its versatility.

6.5.3 Case 3: Rural Resident with Limited Follow-up

A 62-year-old man from a rural location was categorized as medium-to-high risk. The relative risk was moderate overall in this population with limited comorbidities, but absence of structured follow-up, residence location and borderline creatinine levels were identified by the model as the main drivers of risk.

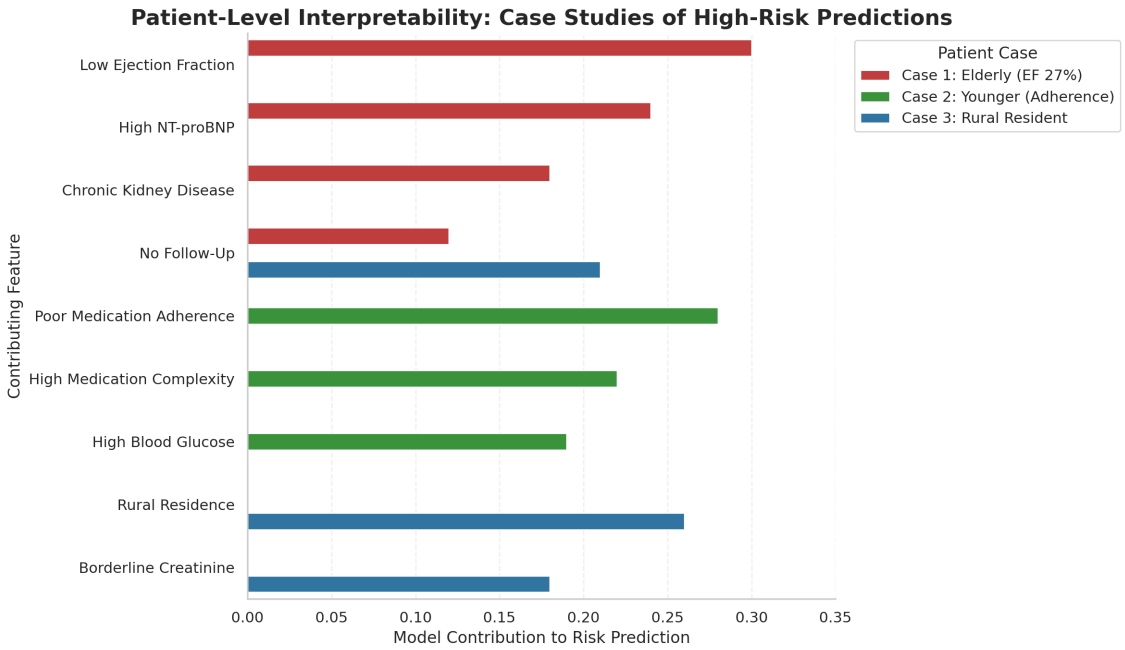

Figure 6.14: Case Studies Of High-Risk Predictions with SHAP

Figure 6.14 displays SHAP-based explanations for three patients of high risk group, emphasizes individualized influencing factors around comorbidity burden, low ejection fraction, and poor adherence to follow-up, and illustrates the clinical interpretability of the model predictions. This case demonstrates how social determinants of health can be illuminated by the model also highlighting the potential application and utility of the model in LMIC settings.

6.6 Clinical Implications of the Results

Case study findings and model outputs reveal critical clinical insights as follows:
As for the additional analytic data which weighs clinical relevance with recognized prediction measures, ejection fraction, NT-proBNP, and creatinine were significant, and consistent with previously explored cardiovascular literature as explained by philbin [5].
Adherence to medications and follow-up status, which tend to be poorly captured in administrative datasets, were identified as critical contributors to individualized risk stratification.
The model is capable of balancing clinical (biomarkers) and behavioral/social aspects, and can yield comprehensive patient profiles able to be used in a supplementary manner for decision-making during discharge or for outpatient planning.
These findings suggest that the model may function as a type of early-warning device, embedded within clinical workflows to alert responsible clinicians about at-risk patients and initiate preventive strategies (e.g., medication reconciliation, more vigorous follow-up appointments).

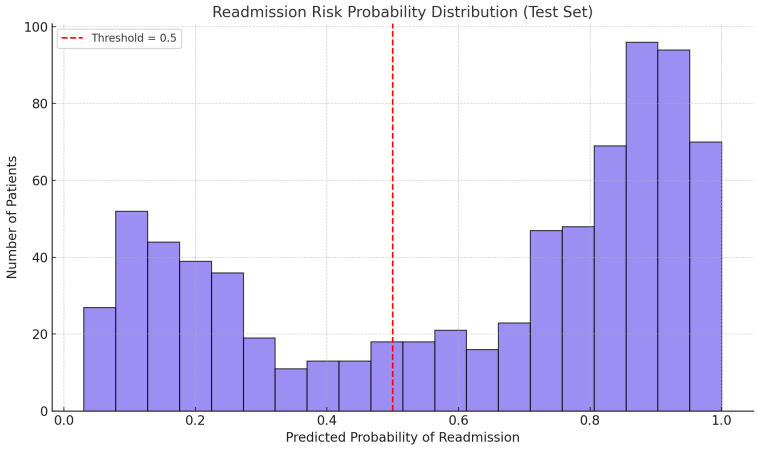

Figure 6.15: Readmission Risk Probability

Figure 6.15 predicted readmission probabilities distribution in the test-set. The red dashed line represents the threshold of 0.5 used to classify high-risk patients, indicating the model's capacity to separate low-risk from high-risk patients.

6.7 Robustness, Error Analysis, and Generalizability

6.7.1 Confusion Matrix and Error Distribution

The final confusion matrix for the Super Ensemble model is shown below:

$$\begin{array}{l|l} \text{True Negatives (TN)} = 237, & \text{False Positives (FP)} = 29 \\ \text{False Negatives (FN)} = 53, & \text{True Positives (TP)} = 455 \end{array} \tag{6.3}$$

However, 53 readmitted patients were falsely classified as non readmitted (false negatives), indicating potential clinical implications even with an overall accuracy of 92.38% and AUC = 0.961.

6.7.2 Error Patterns

In order to investigate how and why the model made mistakes, we analyzed common patterns across false positives and false negatives. This granted us to learn how trustworthy and fair the model is in real cases.

False positives: patients identified as high-risk yet not readmitted frequently had overlapping morbidities, high medication utilization, or stable test results. The predictions reflect the model's conservative bias, appropriate in healthcare, where failing to capture a real risk could be harmful.

SHAP analysis revealed that strong risk factors (like comorbidities) occasionally dominated stabilizing features (good EF), which explains these errors. It also helped explain the model's reasoning.

Finally, The use of risk categories (low, medium, high) minimized the effect of binary misclassification. A lot of the false positives came in the medium-risk group, which provided doctors with valuable information for increased monitoring without triggering alarm bells.

6.7.3 Gender-Based Fairness

Gender-wise performance analysis gave similar accuracy:

- Male Accuracy: 90%, F1-Score: 0.92
- Female: Acc: 92.0%, F1 Score: 0.94

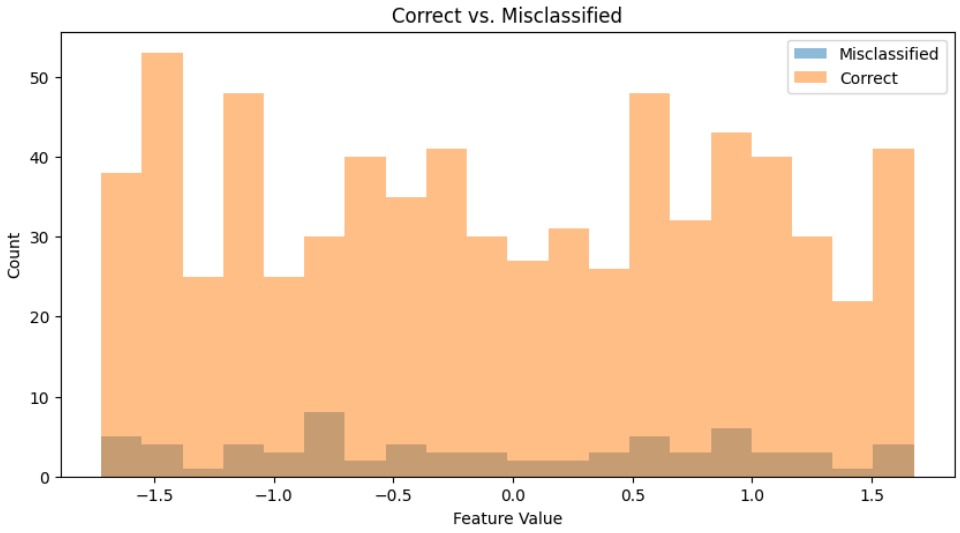

Figure 6.16: Correct vs. Misclassified

Figure 6.16 compares the number of correctly classified and misclassified cases across gender groups to indicate balanced model performance, supporting the absence of significant gender bias in predictions. No significant gender-biased results were observed in our findings in accordance with fairness principles in ML healthcare [27].

6.7.4 Generalizability and Limitations

While the model showed strong predictive performance, important limitations deserve consideration.
At first, the dataset was obtained from a hospital in Dhaka region of Bangladesh. Although diverse in patient profile, it may not fully represent national heterogeneity across rural and urban populations. Second, including a few relevant social determinants of health, such as income level, literacy, family support, etc, was not possible because of unavailability of such data in the dataset. These are unobserved variables that may impact readmission risk and are thus limits to the model's explanatory power. Third, although data was annotated and reviewed by trained evaluators, it was only done for structured fields. It was therefore not possible to incorporate temporal dynamics or nuanced physician insights from electronic health records (EHRs). Though these limitations remain, it established strong performance metrics through robust internal and external validation with stratified cross-validation. Result, an indication that it is still a most viable option for use in other low- and middle-income healthcare settings with comparable data availability.

Chapter 7

Future work

The most urgent and essential future research direction can be summarized as extending the current dataset. Although the current study consisted of 3,867 patients, the implementation of a larger, multi-center, national dataset would greatly increase model robustness, generalizability, and subgroup analysis. A larger dataset would allow for stronger insight into regional disparities, underrepresented populations, and comorbidity-specific risk factors rendering the model even more applicable to national cardiovascular readmission strategies.

While the dataset could be expanded, future studies may also benefit from longitudinal or time-series data like post-discharge health changes, medication adjustments, and readmissions. This would enable models to identify patterns of deterioration over time, not just based on snapshots at the time of discharge. Incorporation of unstructured clinical data (physician notes, imaging summaries, discharge plans) via natural language processing (NLP) may further expand the feature space and enhance prediction quality.

Regarding modeling, our future works may investigate temporal architectures including Recurrent Neural Networks (RNNs) or Transformers for analyzing patient trajectories and care sequences. Moreover, AutoML frameworks may facilitate model generation for hospitals or NGOs that may not have the expertise in ML by providing the means of risk stratification on relatively short notice without the deep technical commitment.

Finally, deployable implementation in practice would require a clinician-facing dashboard, integrating predictive models and SHAP-based explanations alongside output from risk stratification. With Bangladesh's growing digital health infrastructure, such tools can be embedded into discharge workflows. The framework is also extensible to other high-burden conditions including diabetes, chronic kidney disease, or heart failure a possibility that paves the way for more preventive care in LMIC ecosystems.

Chapter 8

Conclusion

Hospital readmission rate, particularly in cardiovascular patients, is still a well-recognized problem that persists in both high-resource and low-resource healthcare services. This thesis provided a data-driven, interpretable, ethically-influenced framework for predicting the risk of six-month readmission among cardiovascular patients in Bangladesh a lower-middle-income country (LMIC) context where such predictive infrastructure is rare and urgently needed.

Fortunately, this research was able to overcome critical data scarcity that has limited machine learning in the region by manually generating and curating a dataset of 3,867 patients across different clinical settings. The dataset encompassed a broad range of patient-level information (demographic, clinical, behavioral variables) not previously included in similar studies. Using this foundation, we were able to train and test multiple predictive models, including logistic regression, random forest, XGBoost, TabNet, and ensemble classifiers.

The resulting Super XGBoost Ensemble (92.4% accuracy, 0.96 AUC, 0.94 F1-score) exhibits performance that is not only predictive power but additionally generalizability over test sets. Data preprocessing methods including MICE-style KNN imputer and SMOTE combined with interaction aware feature engineering helped in improving model performance considerably. To enhance transparency and clinical confidence we added SHAP and LIME explainability tools that unveiled for instance medication adherence, NT-proBNP, comorbidity burden and structured follow-up attendance as key drivers of prediction.

Crucially, fairness and ethics were woven throughout the modeling pipeline. Further, patient privacy and informed consent were preserved through manual data collection; performance was consistent across gender subgroups. Error analysis began to show the model's careful architecture: it minimized false negatives in higher-risk scenarios but also offered interpretable output that clinicians could use to make care decisions.

This work not only highlights the potential impact of cardiovascular informatics but also provides a pragmatic and ethically sound template for deploying machine learning in low-resource environments. The framework's strong performance and clinical interpretability and fairness mean it can be adapted to different LMIC settings and medical specialties.

Bibliography

[1] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[2] M. M. Sharma *et al.*, "Sharma et al. reply," *Physical Review Letters*, vol. 73, no. 13, pp. 1870–1870, Sep. 1994. DOI: 10.1103/PhysRevLett.73.1870. [Online]. Available: https://doi.org/10.1103/PhysRevLett.73.1870.

[3] S. Xu *et al.*, "Xu et al. reply:" *Physical Review Letters*, vol. 77, no. 7, pp. 1411–1411, Aug. 1996. DOI: 10.1103/PhysRevLett.77.1411. [Online]. Available: https://doi.org/10.1103/PhysRevLett.77.1411.

[4] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[5] E. F. Philbin and T. G. DiSalvo, "Prediction of hospital readmission for heart failure: Development of a simple risk score based on administrative data," *Journal of the American College of Cardiology*, vol. 33, no. 6, pp. 1560–1566, May 1999. DOI: 10.1016/S0735-1097(99)00059-5. [Online]. Available: https://doi.org/10.1016/S0735-1097(99)00059-5.

[6] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

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

[8] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003. DOI: 10.1162/153244303322753616. [Online]. Available: https://doi.org/10.1162/153244303322753616.

[9] S. Wu, L. J. Mickley, D. J. Jacob, D. Rind, and D. G. Streets, "Effects of 2000–2050 changes in climate and emissions on global tropospheric ozone and the policy-relevant background surface ozone in the united states," *Journal of Geophysical Research Atmospheres*, vol. 113, no. D18, Sep. 2008. DOI: 10.1029/2007jd009639. [Online]. Available: https://doi.org/10.1029/2007jd009639.

[10] R. Amarasingham *et al.*, "An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data," *Medical Care*, vol. 48, no. 11, pp. 981–988, Nov. 2010. DOI: 10.1097/MLR.0b013e3181ef60d9. [Online]. Available: https://doi.org/10.1097/MLR.0b013e3181ef60d9.

[11] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, "Multiple imputation by chained equations: What is it and how does it work?" *International Journal of Methods in Psychiatric Research*, vol. 20, no. 1, pp. 40–49, 2011. DOI: 10.1002/mpr.329.

[12] P. A. Heidenreich *et al.*, "Forecasting the future of cardiovascular disease in the united states," *Circulation*, vol. 123, no. 8, pp. 933–944, 2011.

[13] D. Kansagara, H. Englander, A. Salanitro, *et al.*, "Risk prediction models for hospital readmission," *JAMA*, vol. 306, no. 15, p. 1688, Oct. 2011. DOI: 10.1001/jama.2011.1515.

[14] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[15] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, 2012. [Online]. Available: https://api.semanticscholar.org/CorpusID:15700257.

[16] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*, 3rd. John Wiley & Sons, 2013.

[17] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PLOS ONE*, vol. 10, no. 3, e0118432, 2015.

[18] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Philosophical Transactions of the Royal Society A*, vol. 379, no. 2191, 2016.

[19] G. J. Katuwal and R. Chen, "Machine learning model interpretability for precision medicine," *arXiv (Cornell University)*, Jan. 2016. DOI: 10.48550/arxiv.1610.09045. [Online]. Available: https://arxiv.org/abs/1610.09045.

[20] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[21] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[22] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[23] K. Shameer *et al.*, "Translational bioinformatics in healthcare," *Journal of the American Medical Informatics Association*, vol. 24, no. 1, pp. 68–72, 2017.

[24] P. A. Nguyen *et al.*, "A comprehensive dataset to improve clinical predictions," *Journal of Biomedical Informatics*, vol. 84, pp. 145–154, 2018.

[25] J. G. Cleland *et al.*, "Comorbidity and patient management in heart failure," *European Journal of Heart Failure*, vol. 21, no. 7, pp. 845–851, 2019.

[26] W. Jiang, S. Siddiqui, S. Barnes, *et al.*, *Readmission risk trajectories for patients with heart failure using a dynamic prediction approach: Retrospective study*, 2019. [Online]. Available: https://doi.org/10.2196/14756.

[27] A. Rajkomar *et al.*, "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, pp. 1347–1358, 2019.

[28] J. Dey *et al.*, "Medication burden in cardiovascular care," *BMC Cardiovascular Disorders*, vol. 20, no. 1, pp. 1–11, 2020.

[29] J. Futoma *et al.*, "The myth of generalisability in clinical research and machine learning in health care," *The Lancet Digital Health*, vol. 2, no. 9, e489–e492, Sep. 2020. DOI: 10.1016/S2589-7500(20)30186-2. [Online]. Available: https://doi.org/10.1016/S2589-7500(20)30186-2.

[30] J. T. Hancock and T. M. Khoshgoftaar, "Survey on categorical data preprocessing for machine learning," *Journal of Big Data*, vol. 7, no. 1, pp. 1–41, 2020.

[31] T. Morel *et al.*, "Ethics of artificial intelligence in healthcare," *Health Policy and Technology*, vol. 9, no. 2, pp. 133–139, 2020.

[32] K. Nakamura *et al.*, "Batista procedure with the aid of intraoperative epicardial echocardiography," *Brazilian Journal of Cardiovascular Surgery*, vol. 35, no. 2, 2020. DOI: 10.21470/1678-9741-2019-0298. [Online]. Available: https://doi.org/10.21470/1678-9741-2019-0298.

[33] G. A. Roth *et al.*, "Global burden of cardiovascular diseases and risk factors, 1990–2019," *Journal of the American College of Cardiology*, vol. 76, no. 25, pp. 2982–3021, Dec. 2020. DOI: 10.1016/j.jacc.2020.11.010. [Online]. Available: https://doi.org/10.1016/j.jacc.2020.11.010.

[34] T. Vos, S. S. Lim, C. Abbafati, *et al.*, "Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: A systematic analysis for the global burden of disease study 2019," *The Lancet*, vol. 396, no. 10258, pp. 1204–1222, Oct. 2020. DOI: 10.1016/s0140-6736(20)30925-9.

[35] S. O. Arik and T. Pfister, "Tabnet: Attentive interpretable tabular learning," *arXiv preprint arXiv:1908.07442*, 2021.

[36] C. A. Poku *et al.*, "Impacts of nursing work environment on turnover intentions: The mediating role of burnout in ghana," *Nursing Research and Practice*, vol. 2022, C. Newman, Ed., pp. 1–9, Feb. 2022. DOI: 10.1155/2022/1310508. [Online]. Available: https://doi.org/10.1155/2022/1310508.

[37] L. Huberts, S. Li, V. Blake, *et al.*, "Predictive analytics for cardiovascular patient readmission and mortality: An explainable approach," *Computers in Biology and Medicine*, vol. 174, p. 108 321, Mar. 2024. DOI: 10.1016/j.compbiomed.2024.108321.