Università Ca' Foscari di Venezia

Department of *Economics*

Master's Degree Programme in

Data Analytics for Business and Society

Explainability Tools applied to Machine Learning Models for Predicting Hospital Readmission

---

Relatore:
Prof. Mattia Stival

Graduating:
Leonardo Zanettin
Matriculation: 886433

Academic year 2025/2026

Dichiarazione del ricorso a strumenti di Intelligenza Artificiale (IA)

Il/La sottoscritto/a **Leonardo Zanettin**, matricola numero **886433**, iscritto/a presso Ca' Foscari al corso di Laurea Magistrale in **Data Analytics for Business and Society**, autore della tesi di laurea intitolata **"Explainability tools applied to machine learning models for predicting hospital readmission"**, relatore **Mattia Stival**, consapevole delle sanzioni previste dall'art. 76 del Testo unico, D.P.R. 28/12/2000 n.445, e della decadenza dei benefici prevista dall'art. 75 del medesimo Testo unico in caso di dichiarazioni false o mendaci, sotto la propria personale responsabilità ai sensi degli artt. 46 e 47 del D.P.R. 28/12/2000 n. 445,

DICHIARA

di aver fatto ricorso a tecnologie di intelligenza artificiale generativa per supportare vari aspetti della ricerca e della scrittura del suo elaborato finale, in particolare:

- **Strumenti utilizzati e versioni:**
  - *Gemini 3.1 Pro* (modello di linguaggio sviluppato da Google).
  - *Claude 4.6 Sonnet* (modello di linguaggio sviluppato da Anthropic).
- **Informazioni su come sono stati utilizzati gli strumenti:**
  - Revisione e ottimizzazione dei testi per migliorare chiarezza e coerenza.
  - Formattazione e impaginazione in linguaggio LaTeX.
  - generazione di grafici e diagrammi per la presentazione dei dati.
- **Sezioni e capitoli interessati:** L'utilizzo ha riguardato principalmente i capitoli metodologici e sperimentali (2 e 3).

Dichiara inoltre che tutto il materiale generato dall'IA è stato attentamente verificato, rivisto e rielaborato dall'autore, che si assume la piena responsabilità per la qualità, l'accuratezza e l'integrità scientifica del lavoro.

Contents

Chapter 1

Introduction

Hospital readmissions within 30 days of discharge are a frequent, costly, and clinically consequential phenomenon [1]. A substantial share of them is considered potentially preventable [2], often traceable to post-discharge adverse events, fragmented follow-up, or incomplete transfer of information between hospital-based and community care [3]. Transitional-care interventions — structured discharge planning, medication reconciliation, timely follow-up contact — have been shown to reduce early rehospitalizations [4, 5]. Yet these interventions, however effective, are resource-intensive: they demand personnel, time, and coordination that no hospital can extend to every discharged patient.

This scarcity turns readmission reduction into a targeting problem. If the patients at highest risk of early readmission could be identified before they leave the hospital, transitional-care resources could be concentrated where they are most likely to make a difference [6]. Predictive models trained on data routinely collected during the inpatient stay offer a natural instrument for this task: by estimating an individual's readmission risk from information already available at discharge, they can support clinicians at precisely the moment when intervention is still possible. The premise of this thesis is that such models, built on clinical, physiological, and patient-level data, can contribute to safer and more efficient discharge decisions — provided their predictions can be both explained and understood.

Two limitations recur in the existing literature. The first concerns scope. Although readmission has been studied with a broad range of predictors — comorbidity indices, laboratory results, sociodemographic characteristics, and social determinants of health — a systematic review of these models found that no single study had incorporated all of these categories simultaneously [6]; each captured only part of the patient's profile. The subsequent emergence of large, publicly accessible electronic health record databases has changed what is feasible, making it possible to assemble these previously separate

dimensions within a single cohort. The second limitation is more fundamental in a clinical setting: the most accurate models are typically the most opaque, and an opaque model can predict well without its reasoning being open to clinical scrutiny. Interpretability is therefore not a cosmetic addition but a precondition for clinical acceptance — and recovering it from high-performing but opaque models is a central concern of this thesis.

This thesis addresses these gaps through three connected questions. First, which model class — among architectures spanning the spectrum from transparent linear baselines to high-capacity ensembles and neural networks — most reliably stratifies 30-day readmission risk using only information available before discharge? Second, does combining demographic, social, admission-related, laboratory, medication, and comorbidity information improve risk stratification, and which features and feature categories drive the resulting predictions? Third, can a layered set of explainability tools recover a clinically coherent account of how the best-performing model reasons, including the places where that reasoning should not be trusted?

To answer these questions, the study draws on MIMIC-IV [7, 8, 9, 10], a large, deidentified electronic health record database from a single academic medical center. The analysis is deliberately built on the general hospital population rather than the intensive-care subset, so that the resulting models remain applicable to the broad cohort of discharged patients. Five complementary categories of predictors — demographic and social, admission-related, laboratory, medication-based, and comorbidity — are assembled exclusively from information recorded before discharge, ensuring that no predictor encodes knowledge of the outcome it is meant to anticipate. Readmission is defined here as any new admission within 30 days of discharge; planned and unplanned returns are not distinguished. Six models of increasing complexity are trained and compared under a temporal validation scheme that evaluates each on a chronologically later, previously unseen period, approximating prospective deployment rather than rewarding fit to historical data. Because readmission is a minority outcome, evaluation centers on the area under the precision-recall curve, which remains sensitive to performance on the high-risk class. The best model is then examined through a layered interpretability analysis — permutation feature importance, accumulated local effects, SHAP, and a regional LASSO surrogate study — that moves progressively from a global ranking of feature relevance toward a granular, context-dependent account of individual predictions.

The principal contribution of this work lies less in the absolute level of predictive accuracy it attains — which, as the results will show, is moderate and bounded by factors the in-hospital record cannot capture — than in the framework through which that accuracy is obtained and scrutinized. On the modeling side, the thesis brings demographic, social, clinical, laboratory, and healthcare-utilization variables together within a single

cohort and a single comparison, categories that earlier work has more often addressed in isolation, and shows that risk stratification learned only from pre-discharge data remains stable across a temporally distinct, previously unseen period. On the interpretive side, it pairs model-agnostic and model-specific explanation tools to progress from asking which variables matter toward asking how the best model actually uses them, surfacing localized threshold effects and feature interactions — and, with them, patterns that are predictively useful yet clinically ambiguous. In doing so, the work reframes the role of interpretability: less a means of justifying the model than a way of mapping where its reasoning can, and cannot, be relied upon.

The remainder of the thesis is organized as follows. Chapter 2 introduces the MIMIC-IV database, details the selection and preprocessing of the predictor set, and presents an exploratory analysis of the cohort. Chapter 3 formalizes the prediction problem, develops the bias–variance perspective that motivates the choice of models, introduces the models themselves, and sets out the interpretability and explainability tools used throughout. Chapter 4 reports the comparative performance of the six models under temporal validation and then turns to the interpretability analysis of the best-performing model, from global explanations down to a localized, subgroup-level study. Chapter 5 discusses the findings, situates their limitations, and outlines directions for future work.

Chapter 2

Data & Exploratory data analysis

2.1 Introduction to the dataset

The dataset used in this thesis is MIMIC-IV v3.1 ([7], [8], [9], [10]), a publicly available critical care database developed by the MIT Laboratory for Computational Physiology and hosted on PhysioNet. It contains de-identified electronic health record (EHR) data from the Beth Israel Deaconess Medical Center (BIDMC) in Boston. MIMIC-IV is fully de-identified in accordance with HIPAA standards. Patient privacy is protected through multiple mechanisms, including the replacement of personal identifiers with unique random IDs and the systematic shifting of dates. Although exact calendar dates are altered, the time elapsed between medical events remains accurate for each patient. Researchers can infer the approximate real year of care using the 3-year intervals (e.g., '2008 - 2010') provided in the `anchor_year_group` attribute. In addition, certain elements are recoded for anonymity—for example, patients aged 89 years or older are grouped into a single category with an assigned age of 91. Access to the dataset requires compliance with the governance procedures established by PhysioNet. To obtain access, I completed the required CITI Program training ("Data or Specimens Only Research"), signed the Data Use Agreement (DUA), and requested credentialed access through the PhysioNet platform.

MIMIC-IV is plausibly well-suited for predicting readmission risk using machine learning due to its large scale and comprehensive data, comparable with the kind and size of databases used in similar studies. In the systematic review by Kansagara et al. [6], which evaluated several readmission prediction models developed up to that time, the authors found that while the variables included across studies covered a wide range of categories—including comorbidity indicators, laboratory results, sociodemographic factors, and social determinants—no single study incorporated all of these categories simultaneously. Since that review, the landscape of clinical data availability has transformed

substantially.

The emergence of large, publicly accessible electronic health record databases like MIMIC-IV, which integrates detailed clinical data with administrative records for hundreds of thousands of hospitalizations, makes it now feasible to develop models that incorporate the full spectrum of potentially predictive variables.

MIMIC-IV is grouped into two modules: `hosp` and `icu`. Organization of the data into these modules reflects their provenance: data in the `hosp` module is sourced from the hospital wide EHR (Electronic Health Record), while data in the `icu` module is sourced from the in-ICU clinical information system (MetaVision). A total of 364,627 unique individuals are in MIMIC-IV v3.1, each represented by a unique subject_id. These individuals had 546,028 hospitalizations (uniquely identified by a hadm_id) and 94,458 unique ICU stays.

The hosp module includes patient demographics, admission and transfer information, laboratory results, microbiology tests, medication prescriptions and administrations, procedures, diagnoses (ICD codes), and billing-related data. In short, it provides a comprehensive overview of a patient's journey at the hospital level. The `icu` module, in contrast, contains detailed, high-resolution data specifically recorded during ICU stays. It includes continuously charted vital signs, fluid inputs and outputs, medications administered in the ICU, procedures, and time-stamped clinical observations. ICU data is organized around individual ICU stays (stay_id), allowing precise tracking of patient status during critical care periods. While patients admitted to the intensive care unit represent a clinically important subset with rich physiological data, they constitute only 15% of all hospital admissions. Restricting the analysis to this subgroup would limit the generalizability of the findings to the broader hospital population. As this study focuses on general readmission risk across all hospitalized patients, a model developed on the full cohort provides more clinically applicable insights. Hence, the focus was on obtaining useful predictors from the hosp module.

The structure of the tables used to construct the analytical dataset is illustrated in Figure 2.1. The diagram provides a simplified representation of the relevant tables from the `hosp` and `icu` module and highlights the key identifiers used to integrate patient-level information across them.

The table admissions was sort of the frame on which to attach the other info; it contains all the admissions in the dataset and many of the sociodemographic and administrative variables included in the models. Similar goes for the patients table which contained other social variables like `age` and `gender`. The variables related to clinical severity were directly or indirectly taken from the prescriptions, transfers, and services tables.

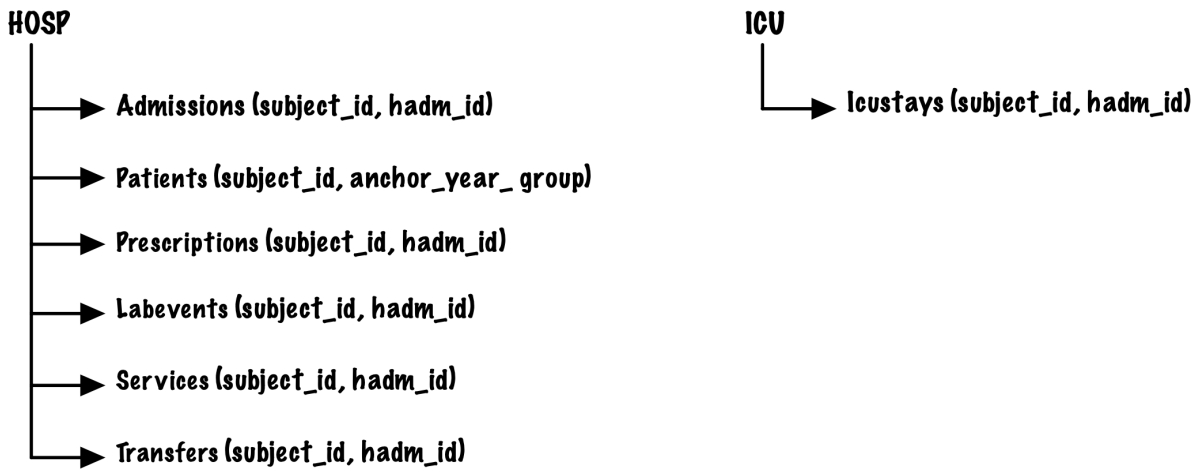


Figure 2.1: Diagram of MIMIC-IV v3.1

The remaining variables are laboratory measurements and they were retrieved from the labevents table.

The variables included in the analysis were selected to represent different dimensions of the hospitalization process and the patient's clinical profile. In particular, predictors were chosen from several broad groups that are commonly considered in studies on hospital readmissions, as they capture complementary aspects of patient risk and clinical complexity. These groups reflect baseline demographic and social characteristics, the context of the hospital admission, indicators of physiological status derived from laboratory tests, treatment intensity through medication information, and the overall burden of chronic disease. Accordingly, the variables used in this study can be organized into the following categories:

1. Demographic & Social
2. Admission-related
3. Laboratory measurements
4. Medication-based
5. Comorbidity Index

Hospital readmission prediction can be defined with respect to a generic time horizon t, representing the number of days after discharge within which a new hospital admission occurs. Different choices of t capture different clinical and operational objectives. In the healthcare literature, commonly used horizons include 7, 15, and 30 days after discharge, with the 30-day timeframe being the most widely adopted benchmark for evaluating hospital readmission risk. In this study, readmission outcomes are therefore considered

Readmission target	Event rate
7 days	9.6%
15 days	15.2%
30 days	21.8%

Table 2.1: Event rate for different readmission targets in the training set

Time Period	Usage	Instances	30-days Readmissions
2008–2013	Model training	333,677	21.8%
2014–2016	Validation	87,901	19.3%
2017–2019	Final evaluation	66,974	18.9%

Table 2.2: Temporal data splitting strategy

for multiple horizons (7, 15, and 30 days), with particular attention devoted to the 30-day interval. The choice of these intervals represents a fundamental methodological trade-off: broader intervals (such as 30 days) capture a higher number of positive readmission events, providing more observations that are crucial for training robust predictive models. However, they also introduce higher 'noise,' as patient outcomes over a longer period are increasingly influenced by external, post-discharge factors. Conversely, shorter intervals offer a clearer signal of hospital-driven acute events but yield fewer positive cases. Table 2.1 reports the observed event rates in the training set corresponding to the readmission horizons defined above.

A 'readmission' is defined as any instance where a patient is assigned a new admission ID (hadm_id) within the specified timeframe following their discharge. Notably, we do not distinguish between planned and unplanned readmissions. While MIMIC-IV provides admission types and locations, using these to definitively impute planned versus unplanned status is unreliable, so the distinction was omitted.

To ensure a clear separation between exploratory analysis and subsequent modeling stages, the dataset was partitioned chronologically into three distinct subsets: training, validation, and test. The exploratory data analysis presented in this section was conducted exclusively on the training set to prevent data leakage and ensure that feature distributions are evaluated independently of downstream validation sets.

Table 2.2 summarizes the resulting temporal partition of the dataset across the study's timeline.

2.2 Feature selection

2.2.1 Demographic and social variables

Demographic and social variables describe non-clinical patient characteristics and social context, together with some general biological information. The variables of this kind that I included in the dataset are:

1. Gender
2. Age
3. Insurance
4. Race
5. Marital status

Although no explicit documentation confirms this, the variable `gender` in MIMIC-IV appears to represent biological sex, as it is restricted to the categories "M" and "F". `insurance` is a categorical variable that contains the name of the insurance type the patient currently has. The categories in the insurance variable are:

- **Medicaid:** Public insurance for low-income individuals, often associated with socioeconomic vulnerability.
- **Medicare:** Public insurance mainly covering individuals aged 65+ and certain younger patients with disabilities or chronic conditions.
- **No Charge:** Hospital services provided without billing, typically due to institutional or charity care programs.
- **Private:** Insurance obtained through employers or purchased individually, generally associated with working-age individuals.
- **Other:** A heterogeneous group including less common or unspecified insurance types.

Similarly goes for `marital_status`, which includes "Divorced", "Single", "Widowed", "Married" and "Unknown".

All of these variables are available in MIMIC-IV without any manipulation. The inclusion of demographic and social variables such as age, gender, race, insurance and marital

status is supported by the Social Determinants of Health (SDOH) framework. According to the National Academies of Sciences, Engineering, and Medicine [11], health outcomes are shaped not only by biological characteristics but also by the social, economic, and structural conditions in which individuals live. Within this framework, age and sex may capture biological vulnerability, while variables such as insurance status, race, and marital status can reflect access to healthcare, socioeconomic resources, exposure to structural inequities, and availability of social support. These factors are recognized contributors to health disparities and may therefore influence differences in clinical outcomes, including hospital readmission risk.

2.2.2 Admission-related variables

Admission-related variables describe characteristics of the hospitalization episode and the patient's clinical trajectory during the hospital stay. These variables capture both administrative information available at admission and indicators of care complexity accrued throughout the hospitalization. The following admission-related variables were included:

1. Admission type
2. Admission location
3. Discharge location
4. Length of stay (days)
5. Intensive care unit (ICU) admission (yes/no)
6. Surgical service involvement (yes/no)
7. Number of intra-hospital transfers
8. Number of service changes during hospitalization

Admission type reflects the clinical context under which the patient was admitted (e.g., Urgent, EW emer., Observation admit, EU Observation, Elective, Direct Observation, Direct emer., Ambulatory observation, Surgical same day admit), and provides indirect information about the acuity of the presenting condition. Admission location describes the source of admission (e.g. Clinic referral, Transfer from skilled nursing facility, Transfer from hospital, Emergency room, Ambulatory surgery transfer, Walk-in/Self referral, Internal transfer to or from psych, Information not available, Procedure site, Physician referral, Post-Anesthesia Care Unit). Discharge location indicates the disposition at discharge (Psych facility, Against advice, Acute hospital, Chronic/long term acute care,

Home health care, Healthcare facility, Skilled nursing facility, Assisted living, Other facility, Unknown, Rehab, Home). These three variables were directly extracted from the `admissions` table in the `hosp` module and did not require further preprocessing beyond categorical encoding. Several additional admission-related variables were not directly available in the database and were therefore derived from other tables. Length of stay was calculated as the difference between discharge time and admission time recorded in the `admissions` table. ICU admission was defined as a binary indicator reflecting whether the patient had at least one ICU stay during the hospitalization, based on records in the `icustays` table from the ICU module. Surgical service involvement was derived from the `services` table, which records the clinical service responsible for the patient at different time points. This approach was preferred over relying solely on the `transfers` table, as patients may undergo surgical procedures without a physical ward transfer, which would otherwise be missed. The number of transfers and the number of service changes were computed by counting, respectively, the number of entries associated with each hospital admission in the `transfers` and `services` tables. All these variables were calculated using only information recorded during the index hospitalization (identified by hadm_id), prior to discharge. No information from subsequent admissions was included, ensuring that all predictors reflect data available before the occurrence of the readmission event. Together, these variables were included to characterize the intensity, complexity, and instability of care during hospitalization, complementing static admission attributes with dynamic indicators derived from the patient's in-hospital trajectory.

2.2.3 Laboratory measurements

Laboratory measurements provide objective clinical information on patients' physiological status and were used to complement demographic, admission-related, and comorbidity variables in the predictive models. In the MIMIC-IV database, laboratory data are recorded in the `labevents` table, where each record corresponds to a single laboratory test performed during a hospital admission. In this study, only laboratory measurements collected during the index hospitalization were used as predictors, ensuring that all laboratory information precedes the potential readmission event. Each laboratory measurement is described by three main attributes: the label, identifying the biological indicator being measured (e.g., sodium, creatinine); the fluid, indicating the biological specimen from which the sample was collected (e.g., blood, urine); and the category, grouping laboratory tests into broad clinical domains such as chemistry or hematology. These measurements are typically recorded multiple times during a hospitalization and can serve as proxies for acute illness severity, organ dysfunction, and metabolic imbalance, factors that may be associated with an increased risk of hospital readmission. Laboratory measurements

were selected primarily based on their availability across a large proportion of hospital admissions, rather than on task-specific clinical relevance (fluid and category were used for interpretability rather than as a selection criteria). This choice was motivated by the objective of constructing models applicable to the general hospital population, rather than restricted to ICU patients only. Preference was also given to laboratory tests commonly used as components of established clinical severity or comorbidity scores (e.g., platelet count and creatinine in the SOFA score), even when the full aggregated scores were unavailable outside the ICU setting. This strategy allowed the inclusion of objective physiological markers that partially capture patient complexity and acute illness severity using variables consistently recorded in the hosp module. Laboratory tests with highly sparse coverage or measurements limited to specific clinical contexts were excluded to reduce missingness and improve model robustness. The laboratory tests included in the final dataset were:

1. Sodium (Blood, Chemistry), reflecting fluid balance and electrolyte homeostasis
2. Platelet count (Blood, Hematology), providing information on coagulation status and bone marrow function
3. Creatinine (Blood, Chemistry), commonly used as an indicator of renal function
4. Hemoglobin (Blood, Hematology), reflecting oxygen-carrying capacity and anemia status
5. White blood cell count (Blood, Hematology), representing immune system activity and inflammatory response
6. Potassium (Blood, Chemistry), a key electrolyte involved in cardiac and neuromuscular function
7. Blood urea nitrogen (Blood, Chemistry), reflecting protein metabolism and renal excretory function
8. Glucose (Blood, Chemistry), indicating metabolic status and glycemic control
9. Red cell distribution width (RDW) (Blood, Hematology), quantifying variability in red blood cell size

To summarize laboratory measurements over the course of each hospitalization, three summary statistics were computed for each laboratory test. These summary measures were chosen to be clinically interpretable while capturing both extreme values and temporal patterns during the hospital stay. Specifically, the following statistics were extracted:

- Maximum value, capturing potential peaks associated with acute complications or severe physiological derangements
- Range (maximum–minimum), reflecting variability in laboratory values during hospitalization
- Mean of the last two measurements prior to discharge, providing a more stable estimate of the patient's physiological status close to discharge compared to using a single final measurement
- Binary measurement indicator, equal to 1 if the laboratory test was measured at least once during the admission, and 0 otherwise. This variable was included to explicitly capture whether the test was ordered for a patient. In cases where a laboratory test was not performed during the hospitalization, the corresponding summary statistics (maximum, range, and mean of the last two values) were inherently undefined and initially retained as missing values prior to subsequent imputation. This distinction reflects the fact that absence of a measurement in EHR data often indicates that the test was not clinically deemed necessary, rather than a data recording error.

2.2.4 Medication-based variables

Medication-based variables were designed to summarize medication exposure during each hospital admission and to capture aspects of hospitalization complexity and clinical severity. Medication data were extracted from the `prescriptions` table, which records all medications ordered for a patient during an admission. This table reflects medications prescribed by clinicians, but not necessarily those actually administered, as downstream administration data (e.g., from `pharmacy` or `eMAR` tables) were incomplete or sparsely available across the general hospital population. Medication information was used as a proxy for clinical complexity, under the assumption that more severe or unstable admissions are typically associated with a higher number of prescribed medications, reflecting multimorbidity, treatment escalation, or the management of complications. The `prescriptions` table was selected because it provides the most complete and consistent representation of medication orders across hospital admissions, regardless of the ordering system used. Several other MIMIC-IV tables potentially related to medications were considered but not included. Administration-level tables (e.g., `eMAR` and `eMAR_detail`), which record medications actually given to patients, were excluded due to limited coverage and high sparsity across admissions. Provider Order Entry tables (`poe` and `poe_detail`) were also excluded, as they include heterogeneous order types beyond medications and do

not improve medication-specific coverage compared to the prescriptions table. In addition, coding and billing-related tables (e.g., `diagnoses_icd`, `procedures_icd`, `drgcodes`, and `hcpcs_events`) were not used, as these variables are typically assigned after discharge and therefore do not represent information available during the hospitalization. Based on the selected data source, the following summary measures were constructed:

1. Number of unique medications, defined as the count of distinct medications prescribed during the admission. Medications appearing multiple times due to dose adjustments or route changes were counted only once.

2. Binary medication indicator, equal to 1 if at least one medication was prescribed during the admission and 0 otherwise. This variable also serves as an explicit indicator of medication data availability and is used as part of the missingness handling strategy described later.

2.2.5 Comorbidity index

Comorbidity indices are composite measures designed to summarize a patient's underlying disease burden by aggregating information on the presence of multiple chronic conditions. Rather than focusing on acute clinical measurements, these indices provide a structured representation of the patient's long-term health status and baseline clinical complexity, without substantially increasing model dimensionality. The MIMIC-IV database contains several comorbidity and severity scores; however, most of them are specific to intensive care unit (ICU) admissions and are therefore unavailable for the majority of hospital stays. For this reason, the analysis was restricted to the Charlson Comorbidity Index (CCI), which is available for a large proportion of admissions in the `hosp` module and can be consistently applied across the general hospital population. The Charlson Comorbidity Index is a widely used clinical scoring system that quantifies comorbidity burden based on the presence of predefined chronic diseases, such as diabetes, chronic kidney disease, malignancy, and cardiovascular conditions. Each condition is assigned a weight reflecting its association with mortality risk, and the index is computed as the sum of these weighted conditions. Higher Charlson scores indicate a greater burden of comorbid disease and worse baseline health status. In this study, the Charlson index was included as a single numeric predictor to capture baseline patient complexity and chronic disease burden.

2.2.6 Final considerations and preprocessing

The original dataset comprised 546,028 hospital admissions. Admissions resulting in in-hospital death, discharge to hospice, or occurring during the 2020–2022 period were

excluded from the analysis. The final cohort therefore consisted of 488,552 admissions corresponding to 189,640 unique patients. Prior to one-hot encoding, the modeling dataset included 52 independent variables, of which 33 were quantitative and 19 were categorical predictors.

In addition to cohort refinement and variable specification, careful consideration was given to patterns of missingness inherent to electronic health record data. Missingness in electronic health record (EHR) data can arise from different sources and may carry different meanings depending on the type of variable. Given the structured and potentially informative nature of missingness in EHR data, observations were retained and missing values were addressed through imputation, instead of removing incomplete records.

For administrative or demographic variables, missing values are typically attributable to data entry or system-related issues. In contrast, for clinical variables such as laboratory tests or medications, missingness often reflects clinical decision-making, as these measurements are ordered selectively based on the patient's condition. For this reason, different missing data handling strategies were adopted depending on the variable category.

Demographic variables exhibited near-complete availability. All demographic variables were available for approximately 100% of admissions, with the exception of marital status (95%) and insurance type (96%). Missing insurance values were imputed as "Other", a category already present in the data. For marital status, a separate "Unknown" category was created to preserve missingness information.

Admission-related variables exhibited missing values only in "discharge location" (70%), for which a new "unknown" category was substituted for null entries. All other variables of this kind were available for essentially all hospitalizations, with admission location missing in a single case. The missing value was imputed as "Unknown".

The Charlson index variable was fully available for all included admissions and therefore did not require imputation.

Medication-based variables exhibited approximately 84% availability. Missingness in this category was treated as potentially informative. In particular, the absence of medication orders may reflect a clinically stable hospitalization not requiring pharmacological treatment. For this reason, missing values in the derived variable number of unique medications were imputed as zero. In addition, a binary indicator was created to distinguish between admissions with at least one prescribed medication and admissions where the value was imputed, allowing the model to explicitly capture this distinction.

Laboratory measurements showed the highest degree of missingness, with an average availability of approximately 75%, varying by test. In this context, missing laboratory values generally indicate that a test was not ordered by clinicians, rather than a failure in

data collection. The absence of a laboratory test may therefore reflect clinical judgment that the measurement was not necessary, often suggesting a more stable physiological condition. For laboratory variables, missing values were imputed using the median value calculated from the whole MIMIC dataset. While this does not imply that the true value is necessarily normal, it provides a conservative estimate in the absence of a measurement. To preserve the information conveyed by test ordering behavior, a binary indicator was included for each laboratory variable, distinguishing between observed and imputed values.

The complete set of variables included in the analysis after feature selection and preprocessing is summarized in Table 2.3, where each variable is described in terms of its type, interpretation, and data coverage.

Table 2.3: Summary of Variables Included in the Study

Type	Variable Name	Quant./Cat.	Description	Coverage
DS	Gender	Categorical	Patient biological sex	100%
DS	Age	Quantitative	Age at hospital admission (years)	100%
DS	Race	Categorical	Self-reported race/ethnicity	100%
DS	Marital Status	Categorical	Marital status at admission	95%
DS	Insurance	Categorical	Primary insurance type	96%
AR	Admission Type	Categorical	Type of hospital admission (e.g., emergency, elective)	100%
AR	Admission Location	Categorical	Location from which the patient was admitted	100%
AR	Discharge Location	Categorical	Location to which the patient was discharged	70%
AR	Length of Stay	Quantitative	Total hospital stay duration (days)	100%
AR	Was in ICU	Categorical	Indicator of ICU admission during stay	100%
AR	Was in Surgery	Categorical	Indicator of surgical service involvement	100%
AR	Number of Transfers	Quantitative	Total intra-hospital transfers	100%
AR	Number of Service Changes	Quantitative	Number of service changes during stay	100%
MB	Number of Unique Medications	Quantitative	Count of distinct medications administered	84%
MB	Medications Present	Categorical	Indicator of any medication administered	100%
LM	Sodium (mean, max, range)	Quantitative	Summary statistics of sodium levels during stay	75%
LM	Sodium Measured	Categorical	Indicator whether sodium was measured	100%
LM	Creatinine (mean, max, range)	Quantitative	Summary statistics of creatinine levels	76%
LM	Creatinine Measured	Categorical	Indicator whether creatinine was measured	100%
LM	Hemoglobin (mean, max, range)	Quantitative	Summary statistics of hemoglobin levels	77%
LM	Hemoglobin Measured	Categorical	Indicator whether hemoglobin was measured	100%
LM	Platelet Count (mean, max, range)	Quantitative	Summary statistics of platelet count	78%
LM	Platelet Measured	Categorical	Indicator whether platelet count was measured	100%
LM	WBC (mean, max, range)	Quantitative	Summary statistics of white blood cell count	77%
LM	WBC Measured	Categorical	Indicator whether WBC was measured	100%
LM	Potassium (mean, max, range)	Quantitative	Summary statistics of potassium levels	75%
LM	Potassium Measured	Categorical	Indicator whether potassium was measured	100%
LM	BUN (mean, max, range)	Quantitative	Summary statistics of blood urea nitrogen levels	75%
LM	BUN Measured	Categorical	Indicator whether BUN was measured	100%
LM	Glucose (mean, max, range)	Quantitative	Summary statistics of glucose levels	74%
LM	Glucose Measured	Categorical	Indicator whether glucose was measured	100%
LM	RDW (mean, max, range)	Quantitative	Summary statistics of red cell distribution width	77%
LM	RDW Measured	Categorical	Indicator whether RDW was measured	100%
CI	Charlson Index	Quantitative	Charlson comorbidity index score	100%

Table 2.4: Association measures used in the exploratory analysis

Variable type association	**Association measure**	**Definition**
Quantitative – Quantitative	Pearson correlation	Measures the strength and direction of the linear relationship between two continuous variables.
Categorical – Categorical	Cramér's V	Measure derived from the chi-square statistic that quantifies the strength of association between categorical variables.
Categorical – Quantitative	Correlation ratio (η)	Measures the proportion of variance in a quantitative variable explained by the grouping defined by a categorical variable.

2.3 Exploratory Data Analysis

After completing the data preprocessing phase, the analysis proceeds with an exploratory data analysis (EDA) on the training set. The aim is to investigate the relationship between the explanatory variables (X) and the readmission outcome (y), providing insight into which features may carry predictive information and guiding the construction of the analysis. In this context, examining correlations among the explanatory variables is also important, as highly correlated predictors can introduce instability, inflate variances, and obscure the individual contribution of features. Anticipating these issues during the exploratory phase helps ensure a more robust and interpretable modeling strategy.

A key challenge in this analysis arises from the diverse nature of the variables included in the dataset. The dataset contains a heterogeneous set of variables, including both quantitative features (e.g., age, length of stay, and laboratory measurements) and categorical variables (e.g., gender, admission type, insurance status, and clinical indicators). Because these variables are measured on different scales, the strength of association between features cannot be assessed using a single statistical measure.

To account for this heterogeneity, different association metrics were used depending on the types of variables being compared. Specifically, Pearson's correlation coefficient was used for pairs of quantitative variables, Cramér's V for associations between categorical variables, and the correlation ratio for relationships between categorical and quantitative variables. Table 2.4 summarizes the selected association measures and provides a brief definition of each metric.

Given the high dimensionality of the dataset, exhaustively reporting all pairwise associations would yield an unwieldy matrix, offering limited additional insight and reducing

overall interpretability. To effectively visualize these complex relationships, isolate systemic redundancies, and map the topology of the predictor space, we constructed a feature interconnectivity network computed on the full set of variables, of which a representative subset is displayed (Figure 2.2).

In this network, each node represents a variable, while edges connect all pairs of variables. For the laboratory quantitative summaries, only the maximum is displayed as a representative node; the mean-of-last-two and range summaries are omitted from the visualization for clarity, as they are strongly correlated with the maximum (see Appendix 1, e.g. `bun_max` with `bun_last2_mean` and `bun_range`). The binary "measured" indicators are retained. Rather than applying a strict inclusion threshold, edge intensity encodes the magnitude of the association: weak relationships ($|\text{score}| < 0.3$) are represented by very light-grey edges, moderate associations ($0.3 \leq |\text{score}| < 0.65$) by medium-gray edges, and strong associations ($|\text{score}| \geq 0.65$) by dark-grey edges. Node colors correspond to the overarching variable categories defined previously (e.g., demographic, admission, laboratory, medication, comorbidity, and target variables).

Given the heterogeneity of the data types (both categorical and quantitative), multiple statistical measures were required to estimate these associations. To integrate these heterogeneous metrics into a unified spatial representation, pairwise distances were defined as $1 - |\text{score}|$. The spatial layout was then generated using multidimensional scaling (MDS), which takes the distance matrix over the displayed variables as input and embeds them in a two-dimensional space such that Euclidean distances approximate the original correlation-based dissimilarities. Consequently, strongly associated variables (with scores approaching 1) are positioned close together, while weakly associated variables are placed farther apart. Unlike force-directed layouts, MDS explicitly optimizes the global preservation of pairwise distances, providing a more faithful representation of the overall dependency structure.

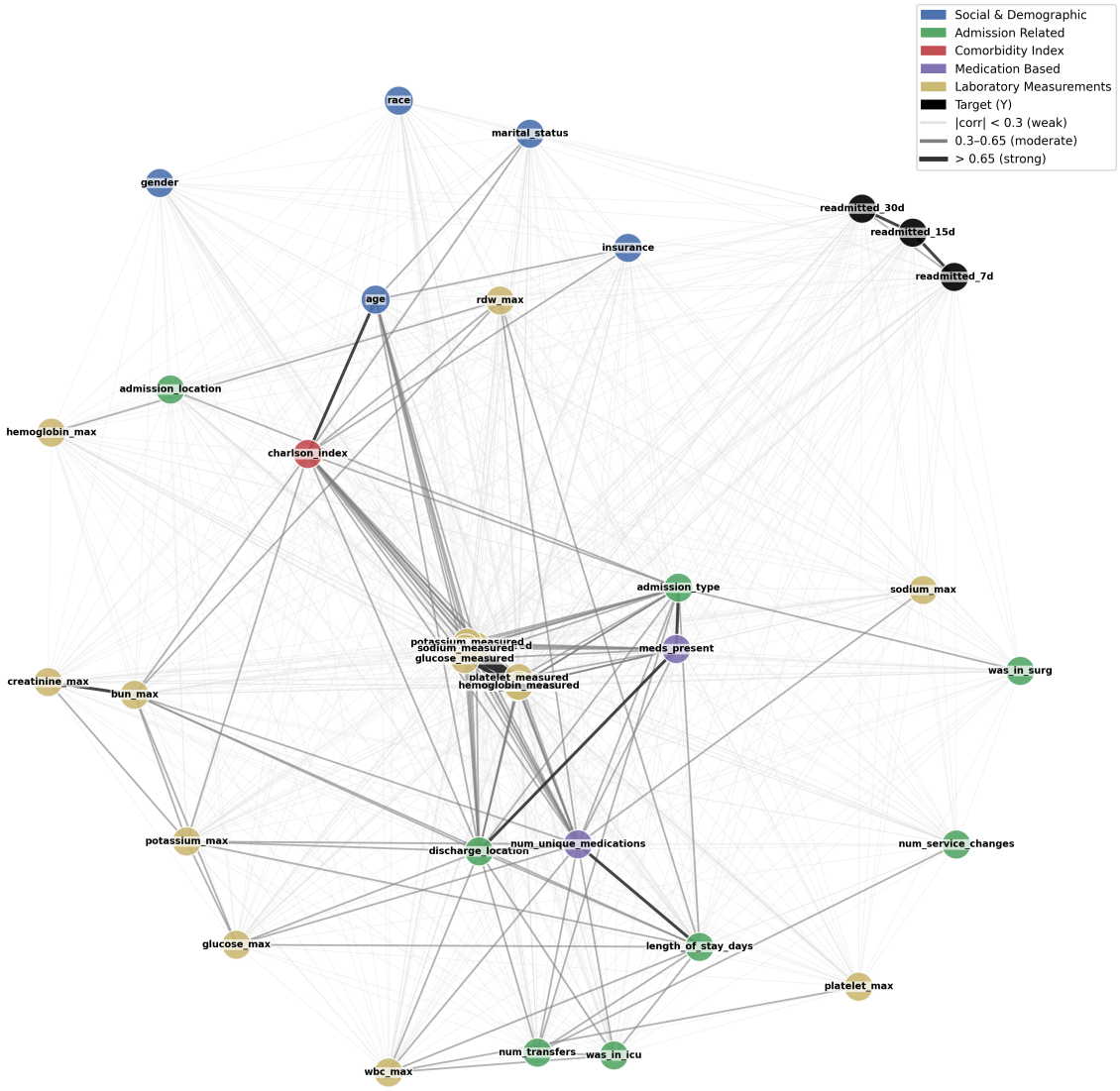


Figure 2.2: Feature interconnectivity network based on pairwise association scores. Node colors indicate variable categories, while edge intensity encodes the magnitude of the association (|score|). Distances are defined as $1 - |\text{score}|$ and embedded in two dimensions via multidimensional scaling. For the laboratory quantitative summaries, only the per-test maximum is shown; the binary measurement indicators are retained. Due to dimensionality reduction, spatial proximity should be interpreted jointly with edge intensity.

Despite the improved global representation provided by MDS, projecting a high dimensional dependency structure onto a two-dimensional plane inevitably entails information loss. As a result, not all pairwise distances can be perfectly preserved: some variables may appear spatially distant despite exhibiting moderate associations, while others may appear closer due to global layout constraints. For this reason, interpretation should rely on a joint reading of spatial proximity and edge intensity, rather than either component in isolation.

Furthermore, due to the massive sample size of the training cohort ($\approx$ 333,000 observations), conventional statistical significance measures such as p-values become largely uninformative, as nearly all associations are statistically significant. A Bonferroni correction was also evaluated to account for multiple testing; however, even under this stringent adjustment, virtually all associations remained significant. Consequently, statistical significance was not used as a filtering criterion, and the visualization instead emphasizes the magnitude of association as the primary indicator of structural relevance. The matrix of significant association scores is reported in Appendix 1.

The resulting network topology reveals several important structural patterns. The most prominent feature remains a dense, highly interconnected module of laboratory "measured" indicators. The strong overlap among these variables reflects the operational structure of clinical practice: laboratory tests are typically ordered in standardized panels rather than individually, leading to highly correlated measurement patterns across patients. This cluster is clearly visible both in terms of spatial proximity and the presence of strong (dark) edges.

Beyond this laboratory-driven structure, the network does not exhibit sharply defined clusters, suggesting that most remaining features capture relatively distinct aspects of the patient state. Some regions of the plot—particularly among laboratory-derived quantities—show moderate spatial grouping, but these are characterized by a mix of weak-to-moderate associations rather than cohesive, high-strength modules.

Social and demographic variables display limited connectivity with the broader clinical feature space, confirming their relatively weak direct association with physiological measurements and hospital processes. However, localized relationships are still present: `age` and `insurance` exhibit moderate association, while `age` also shows moderate associations with several laboratory measurement binary indicators. This relationship is likely not purely physiological, but may instead reflect differences in clinical testing practices across age groups, where certain laboratory panels are more frequently administered to specific patient populations.

Medication-related features—particularly `num_unique_medications`—occupy an intermediate position in the network, moderate-to-strong associations with admission-related variables (notably length of stay and discharge location) and moderate associations with laboratory measurements. This supports their interpretation as proxies for overall patient complexity and clinical severity, as more complex cases typically involve both increased pharmacological intervention and more extensive diagnostic evaluation.

Importantly, the target variables (e.g., `readmitted_30d`) remain largely disconnected from the rest of the feature space in terms of strong associations. While they exhibit

internal consistency (i.e., correlation among different readmission definitions), their links to individual predictors are weak and visually represented by faint edges. This pattern reinforces the conclusion that readmission is not driven by single dominant features, but rather emerges from complex, multivariate interactions. The absence of strong one-to-one relationships highlights the need for flexible modeling approaches capable of capturing non-linear and higher-order dependencies across the feature space.

Having mapped the systemic redundancies and isolated these core themes, we proceed to investigate the direct relationship between strategic features and the target variable (`readm_30d`). Standard linear correlation metrics between individual predictors and the readmission outcome were universally low. Consequently, our bivariate analysis shifts to visualizing distributions and non-linear thresholds to uncover latent risk profiles.

To guide this exploration, a small set of representative variables was selected from the broader network structure identified in the previous analysis. Because many predictors belong to a large cluster of strongly correlated hospitalization-related variables, analyzing all of them individually would be largely redundant.

Following this principle, four variables were chosen to represent distinct stages of the hospitalization process. `admission_type` captures the conditions and urgency at the time of admission, reflecting the initial clinical context. `num_unique_medications` summarizes the intensity of pharmacological treatment during the hospital stay and acts as a proxy for clinical complexity. `discharge_location` represents the outcome and level of care required at discharge, providing insight into the patient's condition at the end of hospitalization. Finally, `age` was included as a baseline demographic factor that is not embedded within the main hospitalization cluster and therefore may provide complementary information about patient risk. The following plots examine the distribution of readmission rates across these variables in order to identify potential thresholds or subgroup patterns that may not be captured by simple linear association measures.

Figure 2.3 summarizes the relationship between the selected representative variables and the 30-day readmission outcome. Although the correlation analysis indicated weak linear associations with the target, several interpretable patterns emerge from the visual inspection. Age exhibits a non-linear relationship with readmission risk: the probability of readmission increases up to middle age and then gradually decreases among older patients, suggesting that age alone does not explain readmission risk in a monotonic way. Admission type reveals clearer structural differences between categories. Acute pathways, particularly "Direct Emergency" admissions, show higher readmission rates, whereas planned hospitalizations such as "Surgical Same Day Admissions" present substantially lower rates. Interestingly, the "Elective" category displays relatively elevated readmission rates despite being planned, likely reflecting the heterogeneous clinical con-

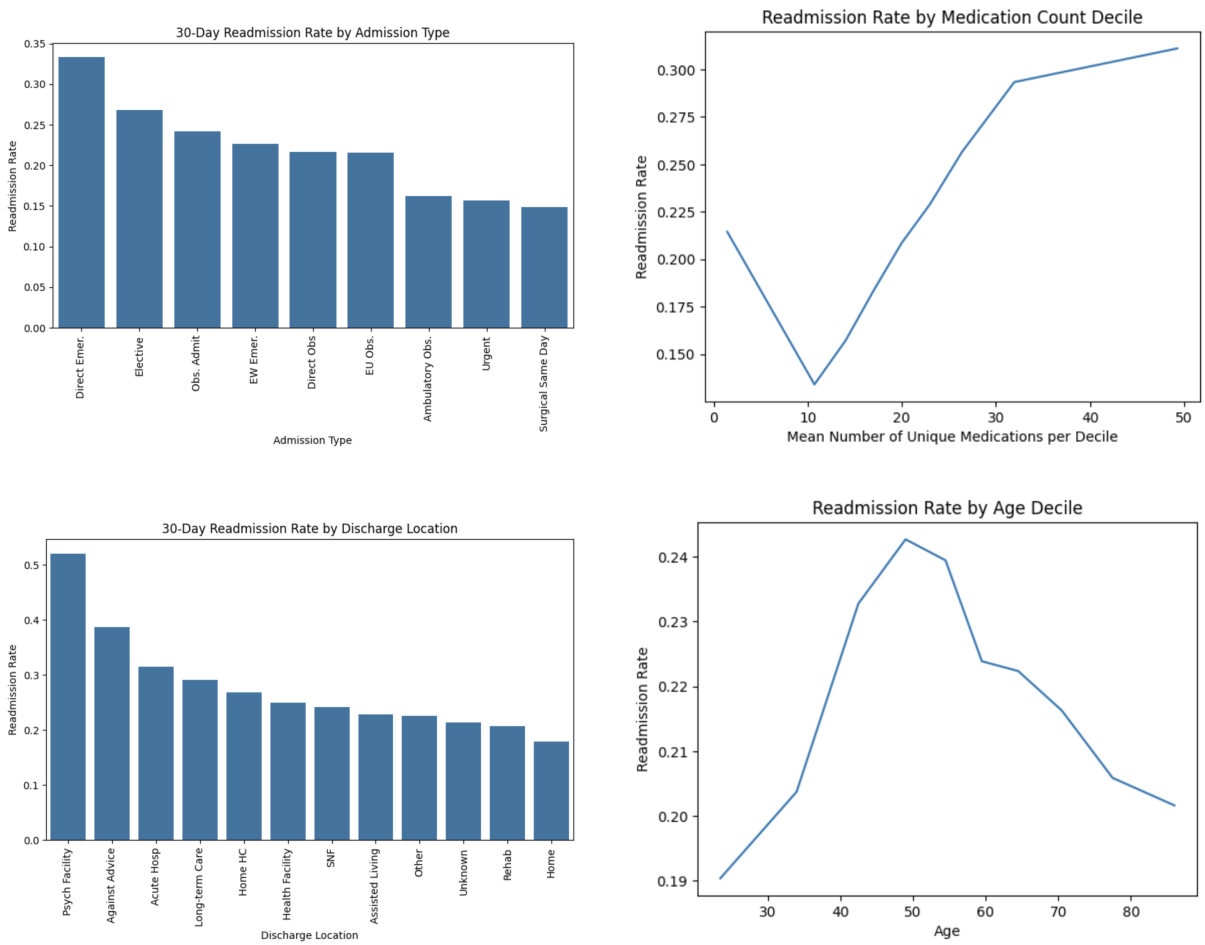


Figure 2.3: Explorative plots for the readmission target

ditions included in this group. Discharge location further highlights differences in post-hospital trajectories: categories associated with higher clinical complexity or disrupted care transitions—such as discharge to psychiatric facilities or discharge against medical advice—exhibit markedly higher readmission probabilities, while patients discharged home or with structured support show comparatively lower rates. Finally, the number of unique medications administered during the hospitalization suggests a gradual increase in readmission risk as medication burden rises. Although the difference between readmitted and non-readmitted patients remains modest, higher medication counts likely capture greater comorbidity and clinical complexity, factors that have been repeatedly associated with increased readmission risk in hospital studies. Overall, these plots indicate that while no single predictor strongly explains readmission on its own, meaningful patterns emerge when examining hospitalization stages and proxies of patient severity, supporting the need for multivariate modeling to capture these combined effects.

Having explored the main characteristics of the dataset through exploratory data analysis, we now turn to the methodological framework for predicting readmission risk. The following chapter presents the formal definition of the prediction problem, discusses the

bias–variance tradeoff that informs model selection, and introduces the set of predictive models considered. We then examine interpretability and explainability, highlighting how different models balance predictive flexibility with transparency and how explainability tools are employed to improve the usability of complex models.

Chapter 3

Methods and models

3.1 Predictive Models for Readmission Risk Classification

3.1.1 The Prediction Problem

Modeling real-world phenomena, including complex patient outcomes like hospital readmissions, is a core goal in statistical learning. These outcomes are inherently stochastic: patient trajectories are influenced by unobserved factors, measurement errors, and random variability, creating a hard limit on performance that no algorithm can overcome. As George Box famously noted [12], all models are inherently flawed approximations of reality, yet they remain highly useful tools.

This observation underpins what Leo Breiman [13] called the "two cultures" of statistical modeling. The traditional "data modeling" approach imposes simple, transparent structures on the data — such as linear equations — that generalize well precisely because they do not attempt to track every fluctuation in the training sample. The "algorithmic modeling" culture, by contrast, prioritizes predictive accuracy by letting complex algorithms learn intricate relationships directly from the data, without a fixed structural assumption. This adaptability comes with two distinct risks that are worth keeping separate.

The first is overfitting: a highly flexible model may capture patterns that are particular to the training sample and do not generalize to new observations. This is a failure of statistical generalization — the model has memorized noise rather than learned signal — and it is detectable through held-out performance evaluation.

The second risk is more subtle and does not disappear even when a model generalizes well.

Complex algorithms produce predictions through internal processes — combinations of hundreds of trees, or transformations across millions of neural network parameters — that are not directly readable by a human observer. Even a model that achieves strong predictive performance on unseen data may be doing so for the wrong reasons, exploiting spurious correlations or confounded proxies that happen to be predictive in the available data but would fail or mislead in deployment. Unlike overfitting, this failure is not detectable from performance metrics alone: it requires interrogating the model's internal logic.

To establish a rigorous framework for this interrogation, the underlying prediction problem is formalized as follows. Let $\mathbf{x} \in \mathbb{R}^p$ denote the feature vector for a given hospital admission, comprising p predictors spanning demographic, clinical, laboratory, and medication domains. The outcome of interest is a binary random variable $Y \in \{0, 1\}$, where $Y = 1$ indicates that the patient was readmitted within 30 days of discharge and $Y = 0$ otherwise. The goal is to learn a function

$$\hat{f} : \mathbb{R}^p \to [0, 1]$$

that estimates the conditional readmission probability $\hat{f}(\mathbf{x}) = \hat{P}(Y = 1 \mid \mathbf{x})$. A binary classification decision is then obtained by applying a threshold $\tau \in (0, 1)$:

$$\hat{y} = \mathbf{1}\big[\hat{f}(\mathbf{x}) \geq \tau\big].$$

Where to set τ is not a purely statistical question: it involves clinical judgement, resource availability, and institutional priorities that vary across hospitals and lie outside the scope of this study. This work therefore operates at the level of the continuous risk score $\hat{f}(\mathbf{x})$ rather than committing to a fixed decision boundary. What any choice of τ would produce, however, is a partition of predictions into four outcomes — true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) — and the clinical costs of these outcomes are not symmetric. This asymmetry is what ultimately drives the choice of evaluation metric.

A false positive ($\hat{y} = 1$, $Y = 0$) occurs when a patient is flagged as high-risk but does not actually get readmitted. In practice, this means allocating intensive transitional-care resources—personnel, time, follow-up visits—to a patient who did not need them. At scale, an excess of false positives constitutes a direct misuse of limited hospital resources and risks clinician alert fatigue, reducing the credibility and utility of the prediction system.

A false negative ($\hat{y} = 0$, $Y = 1$) occurs when a high-risk patient is not identified before discharge and consequently receives no additional support. This is the error with the most

direct human cost: a preventable readmission goes unaddressed. Both error types must therefore be managed simultaneously, and no single threshold eliminates both; reducing one necessarily increases the other.

A further structural complication arises from class imbalance. In the training cohort, the 30-day readmission rate is approximately 21.8%, meaning the negative class ($Y = 0$) is roughly four times larger than the positive class. This imbalance has important implications for model training and evaluation: a trivial classifier that predicts $\hat{y} = 0$ for every patient would achieve 78.2% accuracy while identifying exactly zero high-risk patients. Standard accuracy is therefore an uninformative metric, and evaluation must focus specifically on the model's ability to distinguish the minority positive class. The primary evaluation metric adopted in this study is the Area Under the Precision-Recall curve (PR AUC), which is particularly suited to imbalanced settings because it focuses exclusively on the positive class without being inflated by the large number of true negatives (further details on the choice of evaluation metrics are discussed in Section 4.1.2).

This pressure toward higher-capacity models — necessary to extract a reliable signal from a heavily imbalanced outcome — introduces a distinct risk of its own: as model complexity grows, the reasoning behind individual predictions becomes increasingly opaque. In high-stakes domains like healthcare, the consequences of this opacity can be severe. The literature documents both failures attributable to unexamined model reasoning and cases where interpretability tools successfully exposed them. On the failure side, Caruana et al. [14] showed that a rule-based model trained to predict pneumonia mortality learned that a history of asthma was protective — clinically backwards, but statistically accurate in the data because asthmatic patients received more aggressive care that reduced their observed mortality. The model had learned a confounded proxy rather than a causal relationship, and only an interpretable architecture made this visible. A similarly documented failure is described by Obermeyer et al. [15], who found that a widely deployed commercial algorithm for identifying high-risk patients systematically underestimated the needs of Black patients: it used healthcare costs as a proxy for health needs, embedding a structural bias that was entirely invisible from aggregate performance metrics. On the positive side, Lu et al. [16] developed an XGBoost model for sepsis-associated encephalopathy risk prediction in ICU patients and applied SHAP to extract the key drivers of the model's decisions — identifying clinical thresholds and patient-level risk profiles that were subsequently evaluated by ICU physicians and neurosurgeons, who confirmed their alignment with established clinical reasoning. The opaque model's internal logic was thus made legible enough for domain experts to scrutinize, trust, and act upon.

These cases underscore that interpretability is a clinical requirement, not merely a techni-

cal desideratum: opacity makes failures undetectable, while interpretability makes them visible and correctable. Modern practice increasingly addresses this by pairing high-capacity models with post-hoc explainability techniques, partially reconciling predictive strength with the transparency required in clinical decision-making.

3.1.2 The Bias–Variance Tradeoff

The tension between structural simplicity and representational capacity described in the previous section is not merely a conceptual heuristic; it is governed by a foundational result in statistical learning theory known as the bias-variance tradeoff [17].

Consider a model $\hat{f}$ trained on a dataset $\mathcal{D}$ drawn from some population. Because $\mathcal{D}$ is a finite random sample, $\hat{f}$ is itself a random quantity — a different training set would yield a different fitted model. For a fixed input point $\mathbf{x}$, the bias and variance of $\hat{f}$ are defined as:

$$\text{Bias}[\hat{f}(\mathbf{x})] = \mathbb{E}_{\mathcal{D}}\Big[\hat{f}(\mathbf{x})\Big] - f^*(\mathbf{x})$$

$$\text{Var}[\hat{f}(\mathbf{x})] = \mathbb{E}_{\mathcal{D}}\Big[\Big(\hat{f}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}[\hat{f}(\mathbf{x})]\Big)^2\Big]$$

where $f^*(\mathbf{x}) = P(Y = 1 \mid \mathbf{x})$ is the true conditional probability and the expectation $\mathbb{E}_{\mathcal{D}}$ is taken over all possible training sets of the same size drawn from the same population. Intuitively, bias measures how far the model's average prediction — averaged over all datasets it could have been trained on — deviates from the true underlying signal. Variance measures how much the model's prediction at a given point fluctuates depending on which particular training sample was used.

Under squared error loss, these quantities combine into the classical decomposition. Starting from the expected prediction error at a fixed point $\mathbf{x}$, we introduce the expected prediction $\bar{f}(\mathbf{x}) = \mathbb{E}_{\mathcal{D}}[\hat{f}(\mathbf{x})]$ and decompose the error as:

$$Y - \hat{f}(\mathbf{x}) = \underbrace{(Y - f^*(\mathbf{x}))}_{\text{noise}} + \underbrace{(f^*(\mathbf{x}) - \bar{f}(\mathbf{x}))}_{\text{bias}} + \underbrace{(\bar{f}(\mathbf{x}) - \hat{f}(\mathbf{x}))}_{\text{variance term}}$$

Squaring and taking expectations over both the randomness in Y and in $\mathcal{D}$, the three cross terms vanish: the noise term has zero mean by definition of f^*, the bias term is constant with respect to both expectations, and the variance term has zero mean by definition of $\bar{f}$. This leaves:

$$\mathbb{E}\Big[(Y - \hat{f}(\mathbf{x}))^2\Big] = \underbrace{\sigma^2}_{\text{noise}} + \underbrace{\Big(f^*(\mathbf{x}) - \bar{f}(\mathbf{x})\Big)^2}_{\text{Bias}^2[\hat{f}(\mathbf{x})]} + \underbrace{\mathbb{E}_{\mathcal{D}}\Big[(\bar{f}(\mathbf{x}) - \hat{f}(\mathbf{x}))^2\Big]}_{\text{Var}[\hat{f}(\mathbf{x})]}$$

where $\sigma^2 = \mathbb{E}[(Y - f^*(\mathbf{x}))^2]$ is the irreducible noise — the variance of Y around its true conditional mean, which no model can eliminate regardless of its complexity. Analogous decompositions exist for classification losses, including the 0/1 loss [18] and the log-loss relevant to probabilistic outputs; while the algebraic form differs, the qualitative structure and the tradeoff it describes carry over directly to the classification setting studied here.

Each term in this decomposition has a concrete interpretation in the context of readmission prediction. The irreducible noise reflects the fact that readmission is genuinely stochastic conditional on the available features. Factors such as a patient's social support network, medication adherence after discharge, access to outpatient follow-up, and environmental conditions all influence whether a readmission occurs — yet none of these are recorded in the in-hospital electronic health record. This irreducible component places a hard ceiling on the discriminative performance any model can achieve, regardless of its complexity, and should temper expectations when interpreting the results reported in Chapter 4.

Bias is the systematic component of error. A model with high bias makes the same kind of mistake consistently across different training samples, typically because its functional form is too rigid to capture the true relationship between features and outcome. In the present context, logistic regression exemplifies this regime: by restricting $\hat{f}$ to a linear function of the predictors on the log-odds scale, it cannot represent threshold effects or interaction terms unless these are explicitly engineered. The ALE analysis in Chapter 4 will show this concretely — the logistic regression curves are smooth and linear where the relationship is guided by non-linearities and local structures in tree ensembles.

Variance is the instability component of error. A high-variance model is highly sensitive to the specific training sample: small changes in the data produce large changes in the fitted model, and patterns particular to the training set — rather than reflective of the population — are absorbed into the predictions. The single classification tree in this study is the clearest example: despite growing to 1,584 leaves, its greedy recursive partitioning makes it susceptible to overfitting local structure, and it achieves the lowest test-set performance among all candidates.

Reducing bias and variance simultaneously is generally not possible with a fixed model class — this is the tradeoff. Ensemble methods, however, offer complementary strategies for navigating it. Random Forest addresses variance by averaging predictions across many independently trained trees: since each tree overfits to a different bootstrap sample and random feature subset, their errors partially cancel when aggregated. XGBoost, by contrast, addresses bias directly: trees are built sequentially, with each new tree fitted to the residuals of the current ensemble, progressively correcting systematic errors without substantially increasing instability. These architectural differences will manifest clearly in

the results — both in predictive performance and in the shape of the ALE curves, where XGBoost captures localized threshold effects that smoother models necessarily average out.

Balancing these two sources of error also has direct consequences for interpretability. High-bias models are structurally simple and mathematically transparent, but their rigidity limits predictive accuracy on complex outcomes. High-capacity models can capture the intricate, multivariate structure of clinical risk, but their flexibility renders their internal reasoning opaque. This tension motivates the dual focus of the study: selecting models that manage the bias-variance tradeoff effectively, while employing explainability tools to recover interpretable insights from the most complex architectures.

3.1.3 List of Models

To empirically evaluate the predictive dynamics discussed in the previous section, this study utilizes a diverse set of well-established machine learning algorithms. These models were selected to span the spectrum of mathematical complexity, ranging from traditional linear formulations to high-capacity, non-linear architectures. The specific models implemented are:

Logistic Regression (Unregularized and Regularized): Serving as the traditional baseline for binary classification, logistic regression models the conditional probability $\hat{f}(\mathbf{x}) = \hat{P}(Y = 1 \mid \mathbf{x})$ by restricting the log-odds to a linear function of the predictors:

$$\log \frac{\hat{f}(\mathbf{x})}{1 - \hat{f}(\mathbf{x})} = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}$$

which implies $\hat{f}(\mathbf{x}) = \sigma(\beta_0 + \boldsymbol{\beta}^\top \mathbf{x})$, where $\sigma(\cdot)$ is the logistic sigmoid function [19]. Parameters are estimated by maximizing the log-likelihood over the training set. To mitigate the risk of overfitting and handle potential multicollinearity, a regularized variant is also evaluated, which augments the loss function with a penalty term constraining the magnitude of $\boldsymbol{\beta}$ [17].

Decision Tree: As a foundational non-linear model, the Classification and Regression Tree (CART) algorithm [20] recursively partitions the feature space into L disjoint regions $R_1, \ldots, R_L$ based on simple threshold rules on individual features. The predicted probability for an observation falling in region R_ℓ is the proportion of positive training cases in that region:

$$\hat{f}(\mathbf{x}) = \sum_{\ell=1}^{L} \hat{p}_\ell \cdot \mathbf{1}[\mathbf{x} \in R_\ell]$$

where $\hat{p}_\ell = \frac{1}{|R_\ell|} \sum_{i:\mathbf{x}_i \in R_\ell} y_i$. This provides a highly intuitive baseline for non-parametric learning, though individual trees are prone to high variance.

Random Forest: Introduced by Breiman [21], this ensemble method addresses the variance of individual trees by training B trees on independent bootstrap samples of the data, each using a random subset of features at each split. The final risk estimate is the average across all trees:

$$\hat{f}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(\mathbf{x})$$

Since each $\hat{f}_b$ overfits to a different random sample, their idiosyncratic errors partially cancel under aggregation, substantially reducing variance without increasing bias.

Extreme Gradient Boosting (XGBoost): XGBoost [22] constructs $\hat{f}$ as an additive ensemble of T shallow trees built sequentially:

$$\hat{f}(\mathbf{x}) = \sum_{t=1}^{T} h_t(\mathbf{x})$$

where each h_t is fitted to the negative gradient of the loss with respect to the current ensemble predictions — that is, to the residual errors not yet explained by $\sum_{s<t} h_s$. Unlike Random Forest, which reduces variance through parallelism, XGBoost reduces bias through sequential correction, making it particularly effective at extracting signal from difficult-to-predict cases.

Artificial Neural Network: Representing the highest-capacity algorithm in this study, the neural network consists of interconnected layers of artificial neurons applying non-linear activation functions, optimized via backpropagation [23]. Neural networks function as universal approximators capable of mapping highly intricate, high-dimensional relationships within the data, though their performance advantage over tree-based methods on structured tabular data is not guaranteed [24].

3.2 Interpretability and Explainability in Machine Learning

3.2.1 What is an Interpretable Model?

Interpretability refers to the degree to which a human can comprehend the underlying cause of a machine learning decision or consistently predict how a model will behave. While the terms are frequently used interchangeably in practice, academic literature

often draws a distinct line between interpretability and explainability based on how understanding is achieved [25]:

- Interpretability (Passive): The model is structurally transparent. It makes sense on its own because its internal mechanics are naturally aligned with human reasoning.
- Explainability (Active): The system requires specific, external procedures to clarify its functions. This typically involves applying post-hoc methods (like SHAP or PDPs) to reverse-engineer the reasoning of an opaque black box model.

An interpretable model, often referred to as a white box or glass box, is a predictive system built with intrinsic interpretability. Rather than relying on external tools to approximate feature importance after the fact, the model's architecture is deliberately restricted in complexity so that its actual decision-making process is fully transparent to the observer. The study by Lipton et al. [26] identifies three levels at which a model can be considered transparent:

- Simulatability: The model is simple enough that a human can hold the entire system in mind and mentally calculate a prediction in a reasonable timeframe. In practice, this criterion is rarely satisfied once the number of predictors p is large — even a logistic regression with $p = 122$ features, as in this study, cannot realistically be simulated by hand.
- Decomposability: Every individual component of the model — including the inputs, parameters, and decision pathways — has a direct, intuitive meaning. A decision tree node explicitly splitting patients based on a single clinical threshold is highly decomposable; a weighted sum of 122 standardized coefficients considerably less so.
- Algorithmic Transparency: The mathematical process guiding how the model learns from the data is completely understood, allowing researchers to analyze its error surfaces and operational guarantees. This property is largely independent of dimensionality and is shared by logistic regression and decision trees, but not by ensemble methods or neural networks.

These three criteria are not binary — models satisfy them to different degrees — and they do not map uniformly across model families. Subsection 3.2.3 uses them to characterize each model in this study, examining how the mechanism by which $\hat{f}(\mathbf{x})$ is constructed determines what can and cannot be understood directly from the model itself, and what must instead be recovered through post-hoc explainability tools.

3.2.2 Explainability Tools

Explainability tools act as a bridge between complex machine learning models and human users, helping to make their decisions understandable. Rather than serving as an exhaustive review of available methods, this subsection introduces the classification framework needed to position the specific tools used in this study and clarifies the methodological choices behind their selection.

A foundational way to categorize explainability tools is by how they interact with the underlying model [25]:

- Model-agnostic tools treat the system as a black box, generating explanations solely by observing how the model's output changes in response to modified inputs. This makes them applicable to any model architecture, at the cost of not exploiting any internal structural information.
- Model-specific tools access the model's internal weights, gradients, or tree structure directly. This structural knowledge can yield more precise and computationally efficient explanations, but limits applicability to a particular class of models.

Explanations can also be characterized by their scope [27]:

- Global explainability aims to characterize the model's overall behavior across the entire input space, identifying which features drive predictions on average and how the model responds to systematic variation in each predictor.
- Local explainability focuses on a single prediction, decomposing the factors that pushed a specific individual's risk score above or below the population baseline. Local explanations are often more faithful to the model's actual behavior in a given region of the feature space.

Three tools are employed in this study, selected to cover complementary levels of analysis. Permutation Feature Importance (PFI) [21] is a model-agnostic global method that quantifies each feature's contribution by measuring the decrease in predictive performance when that feature's values are randomly shuffled, breaking its association with the outcome. It provides an interpretable ranking of feature relevance, though it is known to distribute importance unevenly across correlated features, since the model can often substitute information from a permuted variable with its correlated counterpart.

Accumulated Local Effects (ALE) [28] are a model-agnostic global tool that address this limitation by estimating the effect of a feature on the prediction within local neighborhoods of the data, rather than marginalizing over the full joint distribution as Partial

Dependence Plots do. This makes ALE plots reliable even when predictors are correlated, and they are used here to characterize the functional relationship between each key predictor and the predicted readmission probability across its range.

SHapley Additive exPlanations (SHAP) [29] provide both global and local explanations grounded in cooperative game theory. A SHAP value ϕ_j for feature j quantifies its marginal contribution to a specific prediction relative to the model's average output, satisfying desirable axiomatic properties including efficiency, symmetry, and additivity. In this study the TreeSHAP algorithm [30] is used, a model-specific implementation tailored to tree-based ensembles that computes exact Shapley values by traversing the internal tree structure and weighting branches according to the training distribution. This approach respects feature dependencies rather than generating synthetic out-of-distribution inputs, making it particularly appropriate given the correlations present in the clinical feature space.

3.2.3 Model Characteristics with Respect to Interpretability

Machine learning models inherently possess varying levels of transparency, which directly influences how their decisions can be understood and audited. Applying the three criteria introduced in Section 3.2.1 — simulatability, decomposability, and algorithmic transparency — to the models used in this study reveals a clear spectrum, from models that are natively readable to those that require post-hoc tools to recover any interpretable insight. Table 3.1 summarizes this assessment alongside the explainability tools employed in this study for each model.

Table 3.1: Model characteristics with respect to the three interpretability criteria and the explainability tools used in this study.

	Logistic Regression	**Classification Tree**	**Random Forest**	**XGBoost**	**Neural Network**
Simulatability	Low. With $p = 122$ predictors, mental simulation is not feasible.	Partial. Feasible only for shallow trees; not at 1,584 leaves.	None. Averaging across hundreds of trees is not humanly traceable.	None. The chained sum of sequential corrections is not humanly traceable.	None. The transformation across multiple hidden layers cannot be mentally simulated.
Decomposability	Partial. Individual coefficients have isolated effects, but a weighted sum of 122 parameters limits modular clinical clarity.	High. Each node applies a single, explicit clinical threshold on a single feature.	Low. Ensemble averaging obscures the contribution of any individual path.	Low. Each correction is a minor residual adjustment with no standalone clinical meaning.	Low. Individual neuron activations have no isolated or clinically meaningful interpretation.
Algorithmic Transparency	High. Maximum likelihood estimation is fully understood analytically.	High. The CART splitting criterion is fully transparent.	Partial. The bagging procedure is understood, but the aggregated result is not.	Partial. The boosting algorithm is well understood; its learned surface is not.	Partial. The backpropagation training procedure is well understood; the learned internal representations are not.
Tools Used	PFI, ALE.	PFI, ALE.	PFI, ALE.	PFI, ALE, TreeSHAP.	PFI, ALE.

Chapter 4

Results

4.1 Model setup and evaluation

4.1.1 Model Setup

Our predictive framework evaluates six models: unregularized and regularized logistic regression, classification trees, random forest, XGBoost, and a feed-forward neural network. As introduced in Section 2.1, we split the data using a temporal validation approach.

Table 2.2 shows a clear chronological shift in the target variable. The baseline 30-day readmission rate drops from 21.8% in the training period (2008–2013) to 18.9% in the final evaluation period (2017–2019). Because the patient cohort changes over time, we isolated the training partition (N = 333,677) for model fitting. Evaluating model performance on chronologically subsequent, previously unseen data therefore provides a more rigorous test than standard internal validation, capturing the impact of evolving clinical practices and patient risk profiles. In this way, the proposed framework moves beyond a static historical evaluation and instead assesses the durability of predictive performance across distinct clinical periods.

During preprocessing, categorical variables were one-hot encoded and quantitative variables were standardized using z-score normalization:

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

where μ and σ are the feature's mean and standard deviation. This step transformed our initial 52 clinical features into a final set of 122 variables.
Hyperparameter optimization was tailored to the specific complexity of each model. Unregularized logistic regression was implemented without tuning as a baseline. For regular-

ized logistic regression and classification trees, an exhaustive grid search was employed. Conversely, for high-dimensional architectures—specifically random forest, XGBoost, and the neural network—a manual search strategy was utilized, where candidate hyperparameter sets were selected based on informed heuristics and computational feasibility. To account for the class imbalance inherent in readmission data, each candidate was tested using both default and balanced class weights. The model configurations that achieved the highest performance on the validation set were selected as the winners of the tournament. Following this selection, the training and validation sets were merged, and the preprocessing pipeline was refit to the combined data to account for any changes in μ, σ, or categorical levels. The final models were then retrained on this consolidated dataset before being evaluated on the test set.

4.1.2 Model evaluation

The primary objective of the predictive framework is to identify the high-risk minority class of patients likely to face 30-day readmission while avoiding the misallocation of hospital resources toward false positives. As noted by Kansagara et al.[6], clinical interventions designed to reduce readmission are often resource-intensive; therefore, a model's utility is defined by its ability to accurately target high-risk individuals without overwhelming clinical staff with "false alarm fatigue" caused by low-precision predictions.

Traditionally, the Area Under the Receiver Operating Characteristic curve (ROC AUC) is the most utilized metric for evaluating discrimination in clinical models [31]. However, ROC AUC presents a significant "trap" in the context of class imbalance. Because the calculation of the False Positive Rate (FPR) incorporates the number of True Negatives in its denominator, the metric becomes deceptively optimistic when the negative class is dominant [32]. In the MIMIC-IV dataset, the vast majority of patients are not readmitted; this high volume of True Negatives can mask a high absolute number of False Positives, leading to an overestimation of the model's practical performance.

To address this, we employ the Area Under the Precision-Recall curve (PR AUC) as the primary metric for model selection and evaluation. The PR curve is constructed by plotting Precision ($TP/(TP+FP)$) on the y-axis against Recall ($TP/(TP+FN)$)—also referred to as the True Positive Rate or Sensitivity—on the x-axis across every possible classification threshold. Unlike the ROC curve, which incorporates True Negatives via the False Positive Rate, the PR curve focuses exclusively on these positive-class-centric ratios. By omitting True Negatives from the calculation, PR AUC ensures that the evaluation remains highly sensitive to the model's performance on the minority class of interest, accurately reflecting the impact of false alarms on predictive utility.

Beyond its mathematical sensitivity to imbalance, PR AUC addresses the limitations of threshold-dependent evaluation. While point-metrics such as the F1-score are also robust to class skew, they provide only a localized snapshot of performance at a single, often arbitrary probability cutoff. Consequently, following the logic established by Davis and Goadrich[33], we utilize PR AUC to provide a global assessment across the entire range of potential operating points. This allows us to evaluate the fundamental discriminative power of the algorithms while keeping the specific threshold discussion open for future clinical implementation.

Finally, while we monitor the Brier Score to assess probabilistic calibration—the agreement between predicted risks and observed outcomes [31]—PR AUC remains the decisive metric for the hyperparameter tournament and final model comparison. The Brier Score can be misleading in imbalanced settings because its calculation (Mean Squared Error) treats misclassifications of the majority and minority classes equally. In our context, a model could achieve a deceptively excellent Brier Score simply by being highly confident about the 80% of patients who are not readmitted, which may obscure poor performance on the 20% of patients who truly require intervention.

4.1.3 Global Explanation for Models Decisions

Following the hyperparameter optimization phase—where the optimal configurations were selected based on validation set performance (as detailed in Appendix 2, 3, 4)—each "tournament winner" was evaluated on the held-out temporal test set (2017–2019). This approach ensures that the results reflect the models' ability to generalize to future clinical data, accounting for potential shifts in hospital coding or patient demographics over time. Table 4.1 summarizes the discriminative power and calibration metrics for all six model candidates.

Table 4.1: Performance Metrics on the Temporal Test Set (30-Day Readmission)

Model	**ROC AUC**	**PR AUC**	**Brier Score**
Unreg. Logistic Regression	0.6713	0.3221	0.1444
Reg. Logistic Regression[†]	0.6713	0.3220	0.1444
Classification Tree	0.6552	0.3117	0.1473
Random Forest	0.6890	0.3671	0.1416
Neural Network (MLP)	0.6889	0.3610	0.1421
XGBoost	**0.7032**	**0.3838**	**0.1388**

[†]Elastic Net regularization, α=0.2, C=10.

The experimental results demonstrate that the XGBoost ensemble outperformed all other candidates in ROC AUC and PR AUC. Specifically, its PR AUC of 0.3838 represents a

significant improvement over the linear baselines, suggesting that gradient boosting is particularly adept at handling the class imbalance inherent in readmission data.

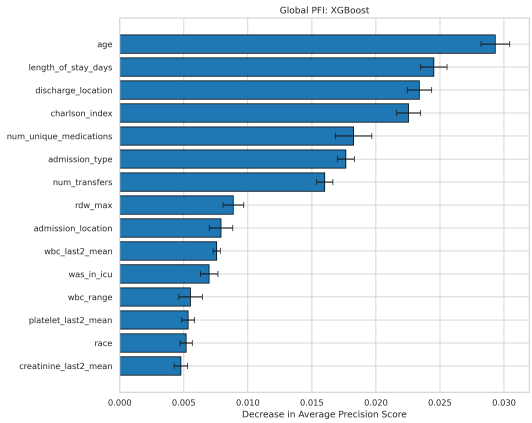
The single Classification Tree achieved the lowest performance among all candidates. Despite reaching a complexity of 1,584 leaves, its greedy recursive partitioning [20] likely failed to capture the subtle, overlapping signals present in the MIMIC-IV dataset, resulting in high variance. In contrast, the Random Forest and XGBoost models addressed this by aggregating multiple trees. The margin by which XGBoost outperformed the Random Forest can be attributed to the fundamental difference between bagging and boosting: while Random Forest builds independent trees to reduce variance, XGBoost builds them sequentially to minimize the residuals of previous trees, effectively extracting more signal out of difficult-to-predict patient profiles [22, 34].

Interestingly, the Unregularized and Regularized Logistic Regression models yielded near-identical results. The light regularization applied ($C = 10$ in the Elastic Net formulation, where C is the inverse of the penalty strength) was insufficient to materially alter the estimated coefficients relative to the unregularized baseline (the complete set of candidates with their hyperparameters and PR AUC is in Appendix 2). Consequently, the model retained all 122 features, suggesting that the clinical signal for readmission is distributed across the entire feature set rather than being concentrated in a few highly predictive variables.

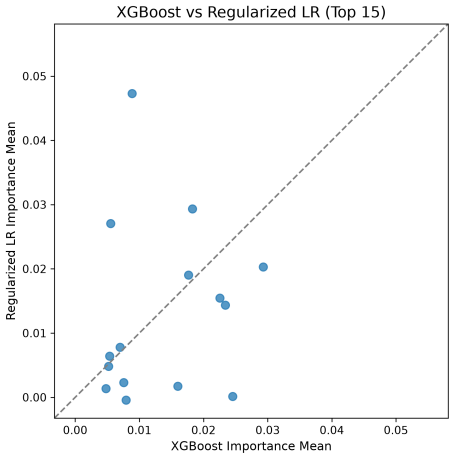
Finally, while the Neural Network (MLP) achieved competitive results (PR AUC 0.3610), it did not surpass XGBoost (all the candidates for the Neural Network model can be found in Appendix 4). This aligns with the finding by [24] that tree-based models maintain a systematic performance advantage over deep learning on tabular data, attributable to the irregular, non-smooth nature of clinical decision boundaries: neural networks carry an inductive bias toward learning continuous functions, which can be a liability when the true relationship between features and outcome is locally discontinuous and interaction-dependent [24]. Nevertheless, a more mature optimization framework—incorporating explicit non-linear feature engineering (such as interaction terms for logistic regression) and a wider hyperparameter search space—would likely unlock significant latent capacity in both baseline architectures.

With the predictive performance of the candidates established, the analysis proceeds to global interpretability. While the ensemble and neural network models achieved the highest discriminative power, their "black-box" nature necessitates the use of post-hoc explanation tools to ensure that their decision logic aligns with clinical reality. Understanding these predictive drivers is particularly critical for the top-performing models, such as XGBoost and Random Forest, as their superior performance likely arises from their ability to capture complex, non-linear interactions.

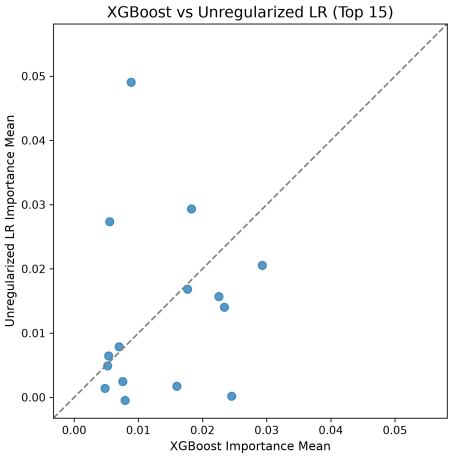
In this section, we first employ Permutation Feature Importance (PFI) to quantify the global contribution of each variable to the models' performance. By measuring the decrease in PR AUC after shuffling a feature's values, PFI provides a high-level ranking of feature relevance. However, acknowledging the limitations of PFI regarding correlated features (as detailed in Chapter 3), we further supplement this analysis with Accumulated Local Effects (ALE) plots. While PFI identifies which variables matter most in aggregate, ALE allows for a granular investigation into how specific variables—such



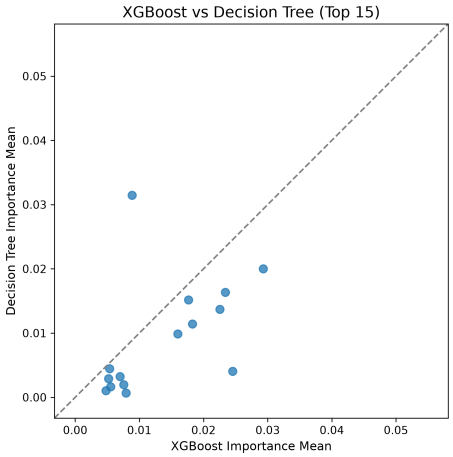
(a) XGBoost



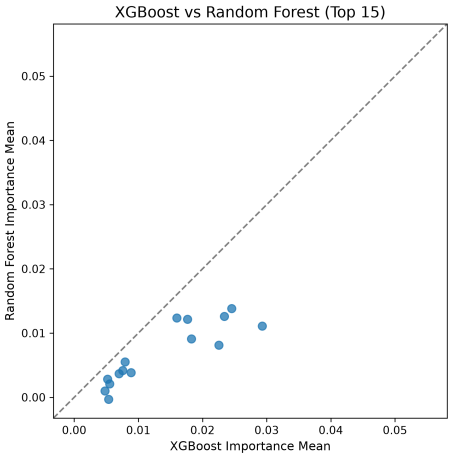
(b) XGBoost vs Logistic (Reg.)



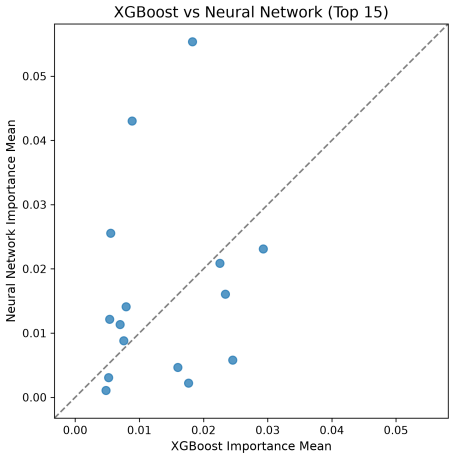
(c) XGBoost vs Logistic (Unreg.)



(d) XGBoost vs Classification Tree



(e) XGBoost vs Random Forest



(f) XGBoost vs Neural Network

Figure 4.1: PFI analysis anchored on the XGBoost top 15 features. The top-left panel shows the XGBoost importance ranking by mean drop in PR AUC. The remaining scatter plots compare those same 15 features against each of the other five models; points on the dashed identity line indicate perfect magnitude agreement, while vertical displacement reflects amplification or compression relative to XGBoost.

as Age or Length of Stay—influence the predicted probability of readmission across their functional range.

Figure 4.1 presents the PFI analysis anchored on XGBoost's top 15 features. The bar chart confirms that `age`, `length_of_stay_days`, `discharge_location`, and `charlson index` constitute the core predictive signal for this model. The scatter plots then project those same features onto the remaining five architectures, placing each model's importance estimate against XGBoost's on a common scale; a point on the identity line indicates full magnitude agreement, while vertical displacement signals amplification or compression.

The most immediate observation is that the two logistic regression variants are virtually indistinguishable: their scatter plots are near-identical in both point positions and overall spread, confirming that the regularization penalty has negligible effect on feature prioritization. Despite having several points sitting close to the identity line, both linear models conceal large disagreements at individual features: `length_of_stay_days` and `num_transfers` collapse to near-zero importance, while `rdw_max` and `wbc_range` are sharply amplified relative to XGBoost. Random Forest and the Classification Tree show stronger global alignment with XGBoost in terms of rank order, though both compress absolute importance values systematically—Random Forest to roughly half the XGBoost magnitude—placing all points uniformly below the identity line.

The Neural Network diverges most strongly from XGBoost. Its most conspicuous departure is `num_unique_medications`, which receives by far the largest importance value across any model (0.055 drop in PR AUC), nearly three times its XGBoost weight. `rdw_max` is similarly amplified, while `length_of_stay_days` and `admission_type` are heavily discounted. This suggests the Neural Network captures a materially different combination of signals, concentrating its predictive reliance on pharmacological complexity and hematological markers rather than the structural and temporal features that dominate the tree-based models.

Despite these architectural differences, a core consensus exists across all six models. Variables reflecting chronic burden and acute medical complexity—namely `age`, `charlson index`, `discharge_location`, and `num_unique_medications`—appear consistently in the upper echelons of every ranking. This cross-model agreement suggests the clinical validity of these features as robust indicators of 30-day readmission risk, independent of the mathematical assumptions of the underlying algorithm.

To overcome this limitation and move beyond merely identifying which variables are important, we must examine how the models utilize them. Therefore, Accumulated Local Effects (ALE) plots were generated for five quantitative variables. These variables are

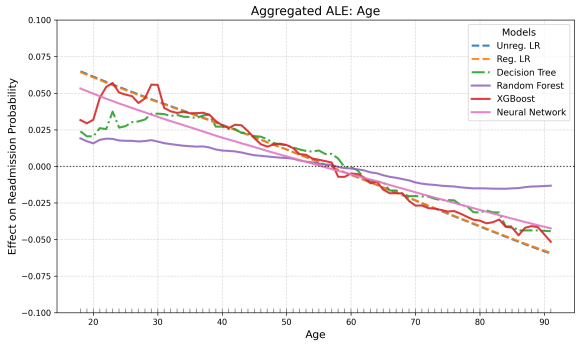
num_unique_medications, rdw_max, charlson_index, age, and length_of_stay_days and were selected because they either demonstrated striking consistency across the models or highlighted significant architectural divergences in the PFI rankings. The following paragraph details the functional relationships uncovered by the ALE analysis.

The aggregated ALE plots (Figure 4.2) provide a functional map of how the candidate models translate clinical features into readmission risk. Before interpreting the curves clinically, it is useful to recall what the y-axis represents: the ALE value at a given feature value expresses the effect of that value on the predicted probability relative to the model's average prediction, so that zero-crossings identify the tipping point at which a feature transitions from risk-increasing to risk-reducing relative to the average patient.

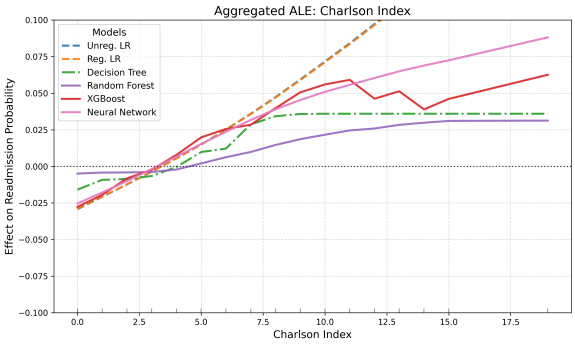
From an architectural perspective, each model exhibits a distinct signature that reflects the functional form discussed in the previous chapter. The two logistic regression curves overlap almost perfectly — a consequence of the light regularization applied — and enforce strictly linear trends throughout. XGBoost produces jagged, staircase-like curves, capturing localized threshold effects and interactions that smoother models are less prone to represent [28]. Random Forest shows a similar but considerably flatter profile, while the Neural Network, despite its capacity, tends toward smooth quasi-linear responses due to its inductive bias toward continuous functions. The single classification tree shows virtually no response outside age and charlson_index, consistent with its poor predictive performance.

Substantively, the plots confirm several clinically coherent patterns. Younger patients carry a positive risk contribution that crosses zero at approximately 55–56 years and becomes protective thereafter — a stable pattern across all architectures. The Charlson index becomes risk-increasing above a score of approximately 3, with effect sizes between +0.05 and +0.08. The medication count crosses the zero baseline at around 25 medications, with linear models imposing a steep monotone increase while tree ensembles reveal saturation and localized structure. The rdw_max effect is among the largest in magnitude but is primarily driven by elevated values, consistent with its role as a marker of physiological dysregulation. The length of stay shows modest and architecturally inconsistent effects, partly reflecting the skewed distribution of the variable rather than a genuine clinical gradient.

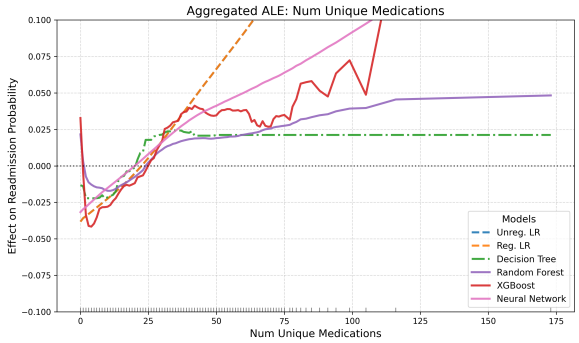
While all models broadly agree on the direction of the main predictors, they differ substantially in how faithfully they represent the underlying functional relationships. XGBoost emerges as the most reliable architecture for this task: its sequential correction mechanism captures the localized threshold effects and feature interactions that characterize clinical readmission risk, and its superiority is reflected both in predictive performance and in the granularity of its ALE curves. The logistic regression models, despite their



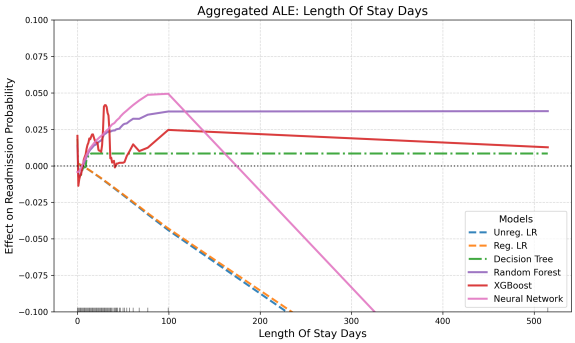
(a) Age



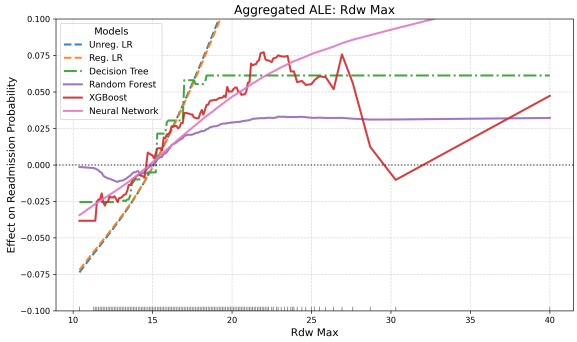
(b) Charlson Index



(c) Number of Unique Medications



(d) Length of Stay (Days)



(e) RDW Max

Figure 4.2: Aggregated Accumulated Local Effects (ALE) plots for the top five continuous predictors across all six candidate models. The y-axis represents the main effect of each feature on the predicted probability of 30-day readmission relative to the average prediction, and is fixed to a common range of $[-0.1, +0.1]$ across all plots to allow direct visual comparison; lines exceeding this window are clipped. Rug plots (black vertical tick marks) at the bottom of each graph indicate the quantile intervals used for computation.

stability and transparency, are structurally limited in this context. Their globally linear functional form cannot represent threshold effects or feature interactions without explicit engineering — spline terms, polynomial expansions, or manually constructed interaction variables — each of which would require domain-driven decisions and substantially reduce the generalizability that currently makes these models attractive. Tree-based ensembles achieve this flexibility automatically, which is precisely what the ALE curves make visible.

4.2 Interpreting the Best Performing Model

4.2.1 Global explanation

Following the identification of XGBoost as the optimal predictive architecture (the full set of candidates and their validation PR AUC is in Appendix 3), this section moves beyond performance benchmarking to examine the model's internal decision-making logic. While the previous analysis relied on Permutation Feature Importance (PFI) to ensure a model-agnostic and computationally consistent comparison across candidates, this approach has important limitations when applied to a single, complex ensemble. To address these limitations, the following analysis employs SHapley Additive exPlanations (SHAP), a framework grounded in cooperative game theory that attributes feature contributions while accounting (in the treeSHAP version) for dependencies among variables.

Specifically, we utilize the TreeSHAP algorithm, which is tailored for tree-based ensembles. Unlike model-agnostic implementations that rely on sampling background data, TreeSHAP leverages the internal tree structure to compute path-dependent conditional expectations based on the training distribution. By traversing the existing tree architecture and weighting branches according to the actual data distribution, TreeSHAP ensures that feature contributions are evaluated within realistic clinical configurations. This approach effectively mitigates issues arising from multicollinearity while remaining computationally efficient for the full test set ($N = 66,974$).

By aggregating these local values, we obtain a global characterization of the model's behavior, shifting the focus from global model error to the magnitude and direction of individual feature contributions. The resulting SHAP values, denoted as ϕ, are generated by the XGBoost `TreeExplainer` and expressed in log-odds, which is the native output space of the classifier prior to logistic transformation. To provide clinical context, these values are anchored to the 18.92% baseline readmission rate of the test set, where $\phi = 0$ corresponds to a patient with a risk exactly equal to this population average.

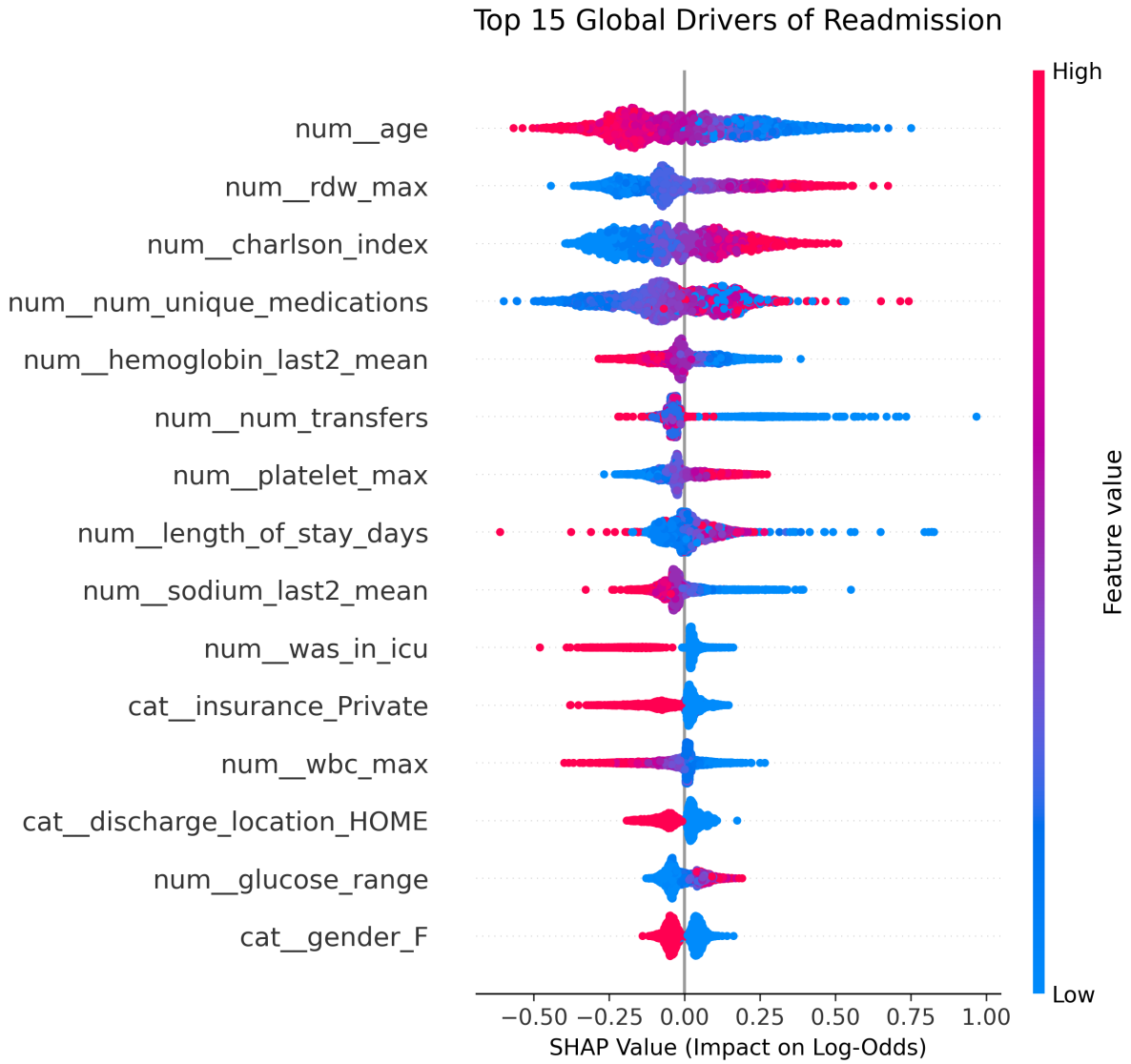


Figure 4.3: SHAP Beeswarm Plot of the top 15 features. Features are ranked by mean absolute SHAP value across the test set. Positive SHAP values indicate an increase in the log-odds of readmission, while color represents the relative value of the feature (red for high, blue for low).

Because the conversion from log-odds to probability is non-linear, the impact of a feature depends heavily on this baseline. Specifically, a SHAP contribution of $+1.0$ increases a patient's predicted probability of readmission from the baseline to approximately 38.8%, while a -1.0 contribution provides a protective effect, reducing the probability to roughly 7.9%. This framework enables a precise quantification of feature influence, demonstrating that values reaching the extremes of ± 2.0 represent massive, defining shifts in a patient's predicted clinical trajectory.

To characterize the overarching logic of the XGBoost model, Figure 4.3 presents a SHAP beeswarm plot for the top 15 features, calculated across the full test set ($N = 66,974$)

and visualized using a representative subset of $3,000$ observations. In this visualization, each point represents a single observation. The position along the x-axis indicates the feature's impact on the model's output in log-odds, while the color represents the original value of the feature, ranging from low (blue) to high (red).

A distinct horizontal spread indicates a variable with a high degree of influence on the model's variance. For instance, features where red points are concentrated on the positive side of the x-axis indicate a direct correlation between higher feature values and an increased risk of 30-day readmission. Conversely, variables where the colors are tightly clustered around the zero-line suggest a more marginalized or localized influence on the final prediction.

The SHAP beeswarm plot in Figure 4.3 reveals a feature importance structure that both confirms and refines the earlier PFI-based interpretation, while providing substantially richer insight into the model's internal logic. Age emerges as the most influential predictor, with the highest mean absolute SHAP value ($|\phi| = 0.172$) and a wide distribution of effects (from -0.64 to $+0.78$). Notably, the direction of this effect indicates that higher age values are associated with a reduction in predicted readmission risk, as evidenced by the concentration of high-value observations on the negative side of the SHAP axis. This non-monotonic and counterintuitive relationship is consistent with the pattern previously observed in the ALE analysis, reinforcing the presence of a stable, model-driven effect rather than an artifact of a specific interpretability technique. In this context, SHAP complements ALE by quantifying the magnitude and variability of this effect at the individual level, confirming that age contributes substantially to prediction variance despite its non-naïve directional behavior.

Among laboratory variables, `rdw_max` stands out as the second most important feature ($|\phi| = 0.150$), exhibiting a distinctly asymmetric impact profile. While its average contribution is high, this effect is primarily driven by extreme values: elevated RDW levels produce strong positive SHAP contributions (reaching approximately $+0.5$ to $+0.7$ in log-odds), substantially increasing predicted risk, whereas normal or low values dont exert the same influence in the opposite direction. This pattern suggests that RDW acts as a threshold-dependent predictor, where abnormal values are highly informative but normal ranges do not meaningfully reduce baseline risk, consistent with its role as a marker of physiological dysregulation rather than a continuously protective factor.

The Charlson Comorbidity Index remains a central predictor under both SHAP and PFI, with only a minor shift in ranking, indicating that its importance is robust to the choice of interpretability method and reinforcing its role as a stable measure of underlying disease burden. In contrast, a hemoglobin-based feature (`hemoglobin_last2_mean`), which was absent from the PFI ranking, emerges as a relevant predictor in the SHAP

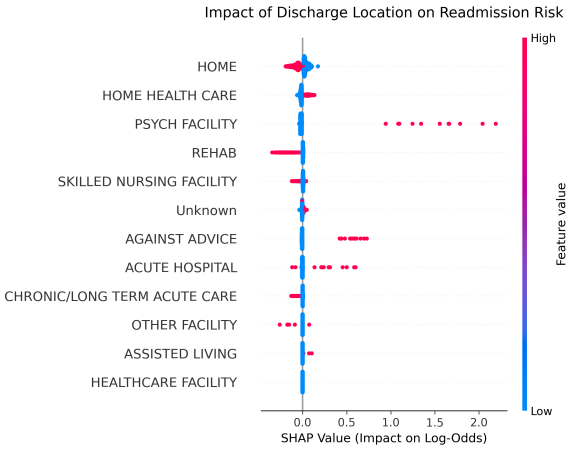
analysis. This discrepancy suggests that their contribution may be partially obscured under permutation-based approaches, likely due to correlations with other clinical variables, whereas SHAP is able to recover their marginal effect by accounting for feature dependencies.

A more nuanced divergence between PFI and SHAP emerges when examining laboratory variables. While some measurements, such as white blood cell count and platelet levels, appear in both rankings, they do so under different representations (e.g., maximum values versus recent averages), indicating that the same underlying clinical signal is captured through multiple derived features. At the same time, other variables are method-specific: sodium-related features appear among the top predictors under SHAP but are absent from the PFI ranking, whereas creatinine is identified by PFI but does not remain among the most influential features under SHAP. This pattern highlights that feature importance depends not only on the clinical variable itself, but also on how it is constructed and summarized within the dataset. When multiple correlated representations coexist, permutation-based importance may distribute their contribution unevenly or suppress it due to redundancy, whereas SHAP is better able to attribute contributions within these correlated groups, even if different derived versions of the same variable emerge as most important across methods.

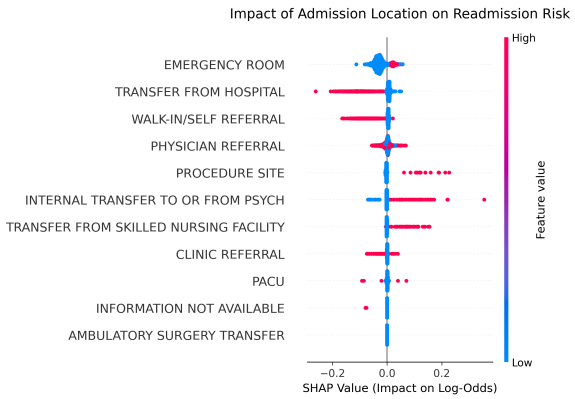
A similar contrast is observed for operational and administrative variables. ICU admission retains a comparable position across both methods, reinforcing its role as a stable predictor, while discharge-related variables exhibit a marked reduction in importance under SHAP. In particular, although discharge location is among the top features under PFI, only the "home" category appears in the SHAP ranking, and at a substantially lower position. This suggests that part of the importance attributed to such variables under PFI may stem from the independence assumption inherent in permutation, which can overstate the contribution of features evaluated in isolation. By contrast, SHAP provides a more conservative estimate by accounting for feature dependencies.

At the same time, this apparent attenuation must be interpreted with caution in the presence of one-hot encoded variables. For multi-level categorical features, the overall effect is distributed across multiple binary indicators, meaning that their global importance can appear diluted when aggregated. As a result, even when individual categories have strong and clinically meaningful effects, these may not be fully visible in the global SHAP ranking. To address this limitation, the following analysis examines feature-specific SHAP distributions for selected categorical variables, allowing the contribution of each category to be evaluated explicitly and providing a more complete understanding of their role in the model.

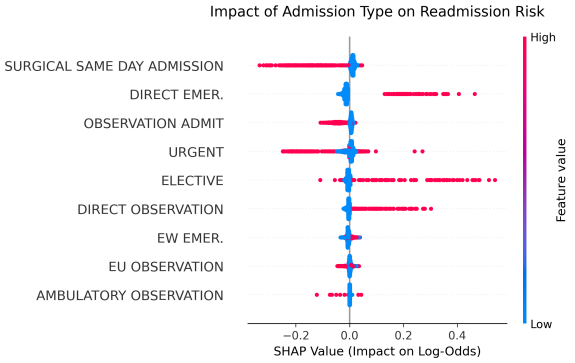
While continuous physiological variables dominate the global feature importance rank-



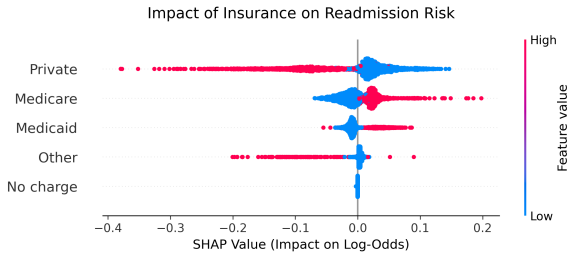
(a) Discharge Location



(b) Admission Location



(c) Admission Type



(d) Insurance Status

Figure 4.4: Categorical SHAP decomposition for the primary clinical and socioeconomic variables. Re-aggregating the one-hot encoded components reveals the isolated predictive weight of specific patient states, such as discharging to a "Psychiatric Facility" or possessing "Private" insurance, which are otherwise obscured in standard global rankings.

ings, the decomposition of categorical variables (Figure 4.4) reveals highly localized, acute risk drivers. By isolating the mean SHAP impact for patients possessing specific traits (i.e., evaluating the impact only when the categorical indicator is present), distinct clinical and socioeconomic profiles emerge.

As illustrated in Figure 4.4a, discharge location acts as the single most potent categorical indicator of readmission risk, primarily driven by behavioral health and treatment non-compliance. Patients discharged to a "Psychiatric Facility" (N=391) experienced an average local SHAP increase of +1.45. In log-odds terms, this represents a massive compounding of readmission probability, having a much larger impact than standard lab tests. Similarly, patients discharged "Against Medical Advice" (N=438) exhibited a high risk penalty (+0.49). Conversely, routine discharges to "Home" (N=24,586) or specialized "Rehabilitation" (N=2,089) act as protective factors, reducing the log-odds by -0.06 and -0.17, respectively.

The model successfully captured the socioeconomic determinants of health embedded within insurance types (Figure 4.4d). "Private" insurance acts as a global protective factor (Mean SHAP -0.09), likely serving as a proxy for better baseline health, higher socioeconomic status, and superior access to outpatient follow-up care. In contrast, "Medicaid" (+0.03) and "Medicare" (+0.02) carry positive risk penalties, reflecting populations that are structurally more vulnerable due to age, chronic disability, or economic instability.

The pathway through which a patient enters the hospital significantly alters their baseline risk (Figure 4.4c and Figure 4.4b). Planned or procedural admissions, such as "Surgical Same Day Admissions" (N=6,042, SHAP -0.14), are highly protective. These patients typically undergo elective procedures, imply a controlled clinical environment, and generally possess the physiological baseline required for proper surgical clearance. Conversely, high-acuity, unplanned entries such as "Direct Emergencies" (+0.22) or "Transfers from/to Psychiatric Units" (+0.11) strongly penalize the patient, serving as trailing indicators of systemic instability. Taken together, these categorical features confirm that the model has internalized structural determinants of readmission risk well beyond traditional clinical measurements, encoding a patient's social circumstances, care continuity, and physiological stability into its predictions. However, the global SHAP summary also reveals a more nuanced dynamic: for several continuous clinical variables, identical feature values produce markedly different attributions across patients, suggesting the model does not evaluate these markers in isolation.

4.2.2 Global-local explanation

The global SHAP summary in Figure 4.3 reveals that for variables such as `num_unique medications`, `num_transfers`, and `length_of_stay_days`, the relationship between a feature's value and its corresponding impact on the prediction is highly complex. Even when patients share identical or very similar values for a given feature, their assigned SHAP values diverge, resulting in a distinct horizontal spread of points. Because SHAP values measure how much a feature shifts an individual prediction away from the dataset's base rate, this variance generates an interesting hypothesis: the XGBoost model may not be evaluating these clinical markers in isolation, but rather conditioning its assessment on the broader physiological context of each patient.

This behavior is expected from a tree ensemble like XGBoost, which naturally creates feature interactions by splitting patients into different decision paths. To see how these interactions manifest in the predictions, we can analyze a specific subgroup where the SHAP score is highly varied.

A risk tipping analysis was conducted for the medication count to find such a subgroup.

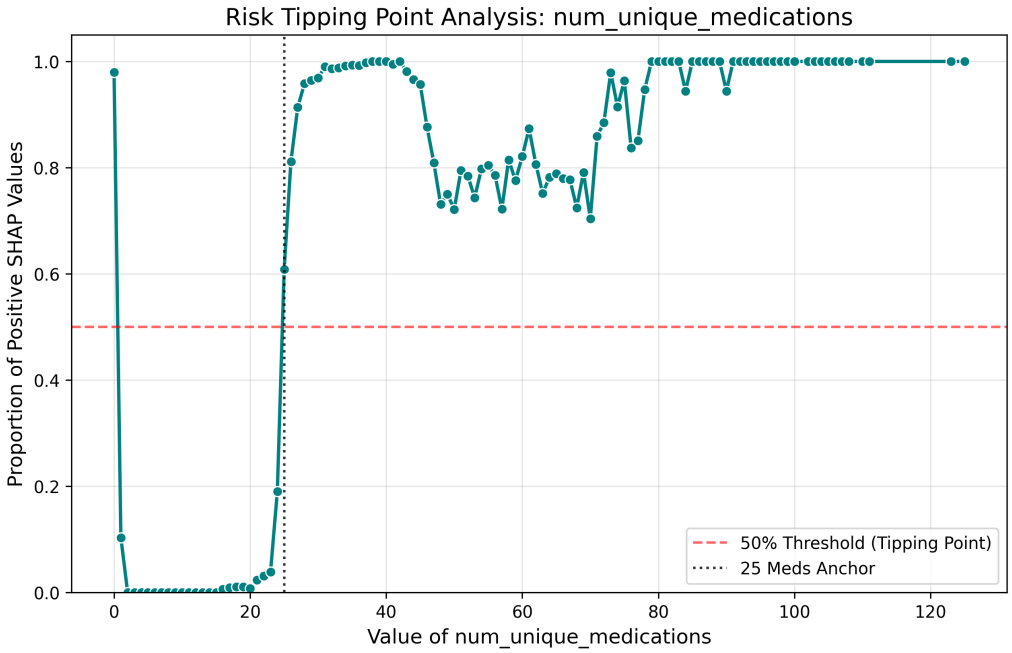


Figure 4.5: Risk Tipping Point Analysis for `num_unique_medications`. The full distribution reveals a localized region of predictive instability: while the lower and upper portions of the range exhibit near-universal positive risk attribution, the intermediate region dips to an average of 80% agreement. This fluctuation identifies a zone of maximum attribution disagreement relative to the near-100% consensus observed on either side.

Figure 4.5 tracks the proportion of patients at each discrete medication level who receive a positive (risk-increasing) SHAP value. Interestingly, while the model assigns a risk-increasing penalty to nearly all patients at lower and higher medication counts, the intermediate zone reveals a distinct drop in consensus, with positive attributions fluctuating around an average of 80%. This dip exposes a critical region of disagreement where the sheer volume of medications is no longer universally penalized, indicating that the XGBoost model is heavily conditioning its risk assessment on other interacting clinical factors.

Looking at the full distribution, the 25-medication mark immediately stands out as a major milestone. This is the exact point where the risk balance flips, and about 60% of patients start receiving a positive risk penalty from the model. It feels like the most natural place to start looking for answers.

However, there is a practical problem with isolating this specific point: even though it represents a clear tipping point in terms of consensus among patients, the actual SHAP values at 25 medications are tightly bunched around zero. This means that while the model is starting to flag these patients as higher risk, the actual size of the penalty is incredibly small. There simply isn't enough spread or variance in the risk scores at this level to clearly separate what makes one patient receive a positive or a negative SHAP

Table 4.2: Predictive and distributional diagnostics for the candidate medication slices. Except for the sample size (N), all metrics describe the distribution of local SHAP values within that specific cohort. The SHAP Range explicitly measures the spread between the 90th and 10th percentiles, highlighting the superior predictive variance available at 47 medications.

Meds Count	N	**% Pos SHAP**	**% Neg SHAP**	**Mean Pos SHAP**	**Mean Neg SHAP**	**10th Pct SHAP**	**90th Pct SHAP**	**SHAP Range (90th−10th)**
25	1,480	60.8%	39.2%	0.0269	-0.0172	-0.0240	0.0458	0.0699
47	320	80.9%	19.1%	0.1178	-0.0508	-0.0417	0.1993	0.2410

score. To really understand how the model differentiates patients, we need a slice where the SHAP scores have a much wider, more dramatic horizontal spread.

A wider slice—for instance, grouping all patients taking between 46 and 57 unique medications—would naturally provide a larger sample size to study. While this range seems narrow enough on the surface, expanding the window inevitably introduces unwanted clinical heterogeneity into the dataset. Even a modest variation in the medication count means the feature is no longer a strict control; it continues to fluctuate. This internal variance shifts the baseline risk within the cohort, making it much harder to cleanly separate the patient's underlying clinical context from the changing drug count itself.

This motivates the choice to isolate and analyze the specific slice where `num_unique medications`= 47. Although focusing on a single discrete value sacrifices sample size compared to the larger patient pool available at lower medication counts, this cohort provides significantly greater predictive variance. As a point of contrast, while the 25-medication mark represents a major transition threshold where the majority of SHAP values flip from negative to positive, its distribution remains tightly clustered near zero. The distinct statistical characteristics of these two candidate slices are compared in Table 4.2.

To unpack the risk variance observed within this high-polypharmacy sub-group, we implement a regional surrogate LASSO model. The ultimate objective of this step is to isolate a compact, high-impact subset of 15 key features. By identifying the primary drivers of model behavior within this specific slice, we can use these features to anchor a downstream K-means clustering algorithm, partitioning the cohort into distinct sub-phenotypes to explain the underlying diversity in the SHAP scores.

While global surrogate models attempt to approximate a black-box model across an entire dataset, and local surrogates focus on explaining a single individual, a regional surrogate maps the model's behavior strictly within a specific slice of the decision surface. To construct this surrogate, we first isolate the 320 test observations where `num_unique medications` = 47, effectively removing the direct influence of medication count by hold-

Table 4.3: Diagnostic statistics for the conditional slice at `num_unique_medications` = 47 and the regional LASSO surrogate.

Diagnostic Metric	**Value**
Conditional slice size (N)	320
Range of predicted readmission risk within slice	[0.0108, 0.8279]
Surrogate R^2 fidelity	0.8465
Active features (Non-zero LASSO coefficients)	43 / 122

ing it constant. We then map the bounded XGBoost readmission probabilities ($\hat{\pi}$) into unbounded logit space using $\text{logit}(\hat{\pi})$ to ensure a mathematically appropriate and unconstrained linear fit. By restricting our scope to this actual empirical distribution, we avoid generating synthetic, randomly perturbed data points that frequently destroy natural clinical correlations.

The properties of this isolated cohort and the resulting fit statistics of the cross-validated LASSO surrogate are detailed in Table 4.3.

The empirical distribution within this single slice reveals that despite sharing the exact same medication count, patients receive predicted probabilities spanning an enormous range from 1.1% to 82.8%. The surrogate achieves a high local fidelity ($R^2 = 0.8465$), indicating that a linear combination of features in logit space captures the clear majority of this predictive variance.

Crucially, LASSO applied an aggressive regularization to the feature space, pruning the pool of 122 available features down to just 43 active coefficients. This sparse behavior is highly advantageous for our methodology; it demonstrates that the local predictive signal is heavily concentrated within a compact subset of clinical drivers rather than being diffusely distributed across the entire dataset. This clean compression justifies extracting the highest-magnitude features to guide our subsequent sub-phenotyping clustering. To isolate these primary drivers, Figure 4.6 displays the top features ranked by the absolute magnitude of their LASSO coefficients.

The features driving the risk variance within this high-polypharmacy slice are primarily characterized by a mix of baseline chronic burden, laboratory measurements, and administrative or admission-related metrics. The importance of the Charlson comorbidity index (`num__charlson_index`) highlights the baseline chronic disease burden of the patient. Meanwhile, variables like acute laboratory metrics (`num__rdw_max`, `num__hemoglobin max`) and hospital tracking variables (`num__length_of_stay_days`, `num__was_in_icu`) capture the specific clinical path and physiological trajectory of the patient during their stay, completely independent of the number of medications prescribed. Finally, the prominence of missing data indicators, such as unknown marital status or race, suggests the

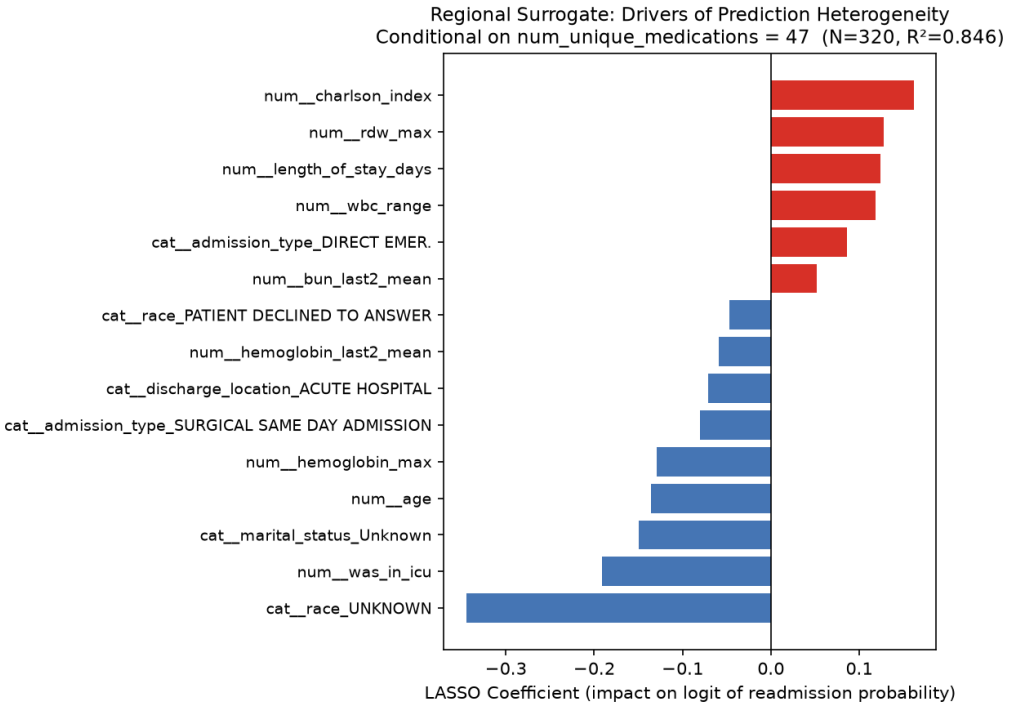


Figure 4.6: Regional surrogate coefficients identifying the top 15 drivers of prediction heterogeneity for observations with exactly 47 unique medications.

model is leveraging these structural recording patterns to capture hidden dynamics related to how this information is collected or asked of patients. Unpacking these patterns further would require a deeper understanding of the American healthcare system or the specific hospital's administrative workflows.

With the 15 highest-magnitude LASSO features identified as the primary drivers of prediction heterogeneity within the slice, a K-means algorithm with $K = 2$ was applied to partition the 320 patients into two sub-phenotypes. The choice of $K = 2$ is deliberate: the goal is not to recover the full distributional complexity of the cohort, but to obtain the sharpest possible contrast between patients who receive a stronger versus a weaker medication SHAP attribution, given an identical polypharmacy footprint. While the K-means algorithm was performed on the standardized and encoded representations to ensure a balanced distance metric, the resulting sub-phenotypes are profiled on their original clinical scales to maintain practical clinical meaning. Table 4.4 reports the resulting cluster-level summaries. Before interpreting the clinical profiles, it is worth acknowledging a structural feature of these results. Both clusters carry a positive mean medication SHAP value, meaning neither group is being actively protected by the medication signal — the model assigns a risk-increasing attribution to the majority of patients in both groups. This is consistent with the composition of the slice itself: at exactly 47 medications, approximately 80% of patients already receive a positive SHAP attribution, leaving only a small minority with negative values. The clustering therefore does not recover a

Table 4.4: K-means clustering results for the meds = 47 slice. Cluster labels are assigned post-hoc based on mean predicted readmission probability.

Metric	**Low Risk Cluster**	**High Risk Cluster**
N	187 (58.4%)	133 (41.6%)
Mean predicted risk	0.144	0.358
Mean SHAP (medications)	+0.040	+0.149
Δ Predicted Risk	0.215	
Δ SHAP (medications)	+0.109	

Table 4.5: Clinical profiles of the two clusters in the meds = 47 slice, reported on original clinical scales.

Clinical Variable	**Low Risk**	**High Risk**
Charlson index	3.52	6.17
Length of stay (days)	8.60	16.90
RDW max	14.20	17.95
BUN (last 2, mean)	16.89	27.13
WBC range	8.44	9.29
Hemoglobin max	11.96	9.99
Hemoglobin (last 2, mean)	10.04	8.53
Was in ICU	95.2%	39.9%
Admission: surgical same day	36.4%	8.3%
Admission: direct emergency	4.3%	6.0%
Age	63.5	63.4

clean separation between patients penalized and patients protected by their medication count. What it does recover is a meaningful gradient: patients in the high-risk cluster receive a medication attribution that is, on average, 0.109 logit units stronger than those in the low-risk cluster, while their mean predicted readmission probability is 0.215 higher. Rather than looking for factors that completely flip the direction of the risk signal, this framing allows us to examine which underlying clinical characteristics tend to scale the magnitude of the polypharmacy penalty up or down. The clinical profiles of the two clusters, reported in Table 4.5, reveal a coherent answer to this question. The high-risk cluster is characterized by a substantially greater burden of chronic comorbidity, with a mean Charlson index of 6.17 compared to 3.52 in the low-risk group. This is accompanied by a markedly longer inpatient stay (16.9 versus 8.6 days) and laboratory markers consistent with physiological instability approaching discharge: elevated RDW, higher BUN in the final measurement window, and slightly lower hemoglobin. The BUN finding is particularly noteworthy, as it reflects renal function close to the point of discharge rather than at admission, suggesting that the high-risk cluster contains patients whose condition had not fully stabilized by the time they left the hospital. The low-risk cluster, by contrast, is predominantly composed of patients admitted for elective or planned surgical

procedures (36.4% surgical same-day admissions versus 8.3%), which typically implies a pre-screened, clinically stable baseline. The ICU finding warrants a specific note. The low-risk cluster shows a dramatically higher ICU rate (95.2% versus 39.9%), which appears counter-intuitive at first reading. A plausible interpretation is that patients on 47 medications who were routed through intensive care received a level of clinical monitoring and pharmacological treatment that stabilized their condition before discharge, attenuating the residual risk that the model detects. An alternative explanation is a selection effect: patients who survived an ICU stay at this level of polypharmacy represent a positively selected group, with the highest-acuity cases either not surviving to discharge or being readmitted in another hospital. Distinguishing between these mechanisms is not possible within the scope of this study. Notably, age does not differentiate the two groups (63.5 versus 63.4 years), despite both clusters sitting well above the test set mean of approximately 57 years. This suggests that within the high-polypharmacy stratum, age alone carries little additional discriminating power — clinical severity, as captured by comorbidity burden and laboratory trajectory, is the dominant axis of differentiation. Taken together, these findings suggest that the XGBoost model treats medication count not as an independent risk signal, but as a severity multiplier whose weight is conditioned on the surrounding clinical profile. Patients on 47 medications who present with high comorbidity, prolonged stays, and signs of physiological instability near discharge receive a substantially amplified medication attribution. Those with a similar prescription burden but a stable, planned clinical trajectory receive a comparatively muted one. The model appears to reserve its strongest polypharmacy penalty for patients where the drugs are embedded in a broader picture of unresolved clinical complexity.

Determining whether these context-dependent shifts represent genuine physiological patterns or merely complex statistical correlations remains outside the scope of this study. Nevertheless, documenting the exact behavior of these localized variations is a necessary first step, providing a precise empirical map to guide future clinical validation.

Chapter 5

Discussion and conclusion

This thesis set out to develop and interpret predictive models for 30-day hospital readmission using routinely collected, hospital-wide data from MIMIC-IV. Starting from a cohort of 488,552 admissions across 189,640 patients, five complementary categories of predictors — demographic and social, admission-related, laboratory, medication-based, and comorbidity — were assembled and used to train six models spanning the complexity spectrum, from linear logistic regression to gradient-boosted trees and a feed-forward neural network. The models were compared under a temporal validation scheme intended to approximate prospective deployment, and the analysis then moved beyond performance benchmarking toward a layered interpretability study combining permutation feature importance, accumulated local effects, SHAP, and a regional surrogate analysis of the best-performing model.

Two broad conclusions emerge. First, gradient boosting (XGBoost) provided the most reliable discrimination, reaching a PR AUC of 0.384 on the temporally held-out test set and outperforming both the linear baselines and the more flexible neural network. The margin over the other tree ensembles was modest but consistent across all three metrics considered. Second, and more relevant to the aims of this work, the interpretability analysis indicated that this predictive advantage is tied to the model's capacity to represent localized threshold effects and feature interactions that smoother models necessarily average out. The accumulated local effects curves, the SHAP decomposition, and the regional surrogate analysis together produced a coherent picture in which variables such as age, comorbidity burden, length of stay, and medication count appear to drive risk not in isolation but in combination.

The principal contribution of this thesis is therefore not the absolute level of predictive performance, which is moderate, but the framework through which that performance is obtained and examined. On the modeling side, the work brings together demographic,

social, clinical, laboratory, and healthcare-utilization variables within a single cohort and a single comparison — categories that earlier readmission studies have more often addressed in isolation — and shows that a model trained only on information available before discharge can stratify risk in a way that remains stable across a temporally distinct, previously unseen period. On the interpretive side, the layered use of model-agnostic and model-specific tools allowed the analysis to move from a global ranking of feature relevance toward a granular account of how the best model actually reasons. The regional surrogate analysis of the high-polypharmacy slice is the clearest example: by isolating patients with an identical medication count and fitting a sparse surrogate within that slice, the analysis suggested that the model does not treat medication count as an independent risk signal but rather modulates its weight according to the surrounding clinical profile — amplifying it for patients with high comorbidity, prolonged stays, and signs of physiological instability near discharge, and muting it for those on a stable, planned trajectory. Whether this pattern reflects genuine clinical structure or a complex statistical correlation cannot be established here, but documenting it precisely is a necessary first step toward that question.

Several limitations qualify these results. The most fundamental is that predictive performance is bounded by information the in-hospital record does not contain. As discussed in the bias–variance treatment, readmission is genuinely stochastic conditional on the available features: post-discharge factors such as social support, medication adherence, and access to outpatient follow-up strongly influence outcomes yet are entirely absent from the data. The moderate PR AUC observed across all models is consistent with this irreducible component and should temper any expectation that architecture alone could close the gap. A second limitation concerns the target itself: because planned and unplanned readmissions could not be reliably distinguished in MIMIC-IV, the outcome modeled here is a broader and noisier proxy than the unplanned-readmission measure that motivates much of the clinical literature, and part of the residual error likely reflects this. Third, the analysis is single-center, drawn entirely from one institution; no external validation on a geographically or institutionally distinct cohort was performed, so the transportability of these models beyond this setting remains untested. Fourth, laboratory missingness was handled by median imputation under the assumption that an unordered test signals a clinically stable patient whose value would plausibly fall near the population norm; while the paired measurement indicators preserve the information carried by ordering behavior, this assumption is necessarily imperfect, and imputed values should not be read as measured physiology.

A further consideration concerns the interpretive status of some of the patterns recovered by the interpretability analysis. The model assigns, on average, a protective contribution to higher age, and within the high-polypharmacy slice the lower-risk sub-phenotype

was characterized by a markedly higher ICU rate. Both findings are internally consistent across methods and stable across the test set, but neither admits a straightforward causal reading. Older age is not plausibly protective against readmission in a biological sense; the observed pattern more likely reflects competing risks and care-pathway effects — older patients who do not return to the hospital may have died, been transferred to settings where a subsequent admission is recorded differently, or received more structured post-discharge support. The ICU pattern is open to a similar reading, in which the apparent protection reflects either the attenuating effect of intensive monitoring on residual instability or a survivor-selection mechanism among patients who reached discharge at all. These are structurally the same kind of confounded proxies documented by Caruana et al. [14] for asthma in pneumonia mortality and, in a different form, by Obermeyer et al. [15] for healthcare cost as a proxy for need: associations that are statistically real, predictively useful, and clinically misleading. Their presence here does not undermine the model's discriminative performance, but it does mark the limits of what predictive signal alone can be asked to support. A formal subgroup performance evaluation across demographic strata, which the present work does not include, would be a necessary component of any deployment and is left to future work.

These limitations point directly to the most valuable directions for future work. The single most promising lever is the integration of post-discharge and longitudinal information — follow-up contacts, outpatient access, social-support indicators — since this is the very information whose absence plausibly caps performance. Incorporating diagnosis codes or unstructured clinical notes, validating the models externally across institutions, and conducting an explicit fairness evaluation across demographic strata would each strengthen the case for clinical relevance. Finally, the localized, context-dependent behavior surfaced by the regional surrogate analysis invites a more targeted follow-up: the sub-phenotypes identified within the high-polypharmacy stratum, and the counterintuitive findings attached to them, are natural candidates for a dedicated causal or prospective study rather than a purely observational one.

Taken together, this work supports a measured conclusion. A model trained solely on routinely collected in-hospital data can produce risk stratification that is both reasonably discriminative and, with the right tools, genuinely explainable. The path toward materially better prediction, however, appears to run less through model complexity than through information the electronic health record does not yet capture — and the role of interpretability, in the meantime, is less to justify the model than to map where its reasoning can and cannot be trusted.

Bibliography

[1] Stephen F. Jencks, Mark V. Williams, and Eric A. Coleman. "Rehospitalizations among Patients in the Medicare Fee-for-Service Program". In: *The New England Journal of Medicine* 360.14 (2009), pp. 1418–1428. DOI: 10.1056/NEJMsa0803563.

[2] Carl van Walraven et al. "Proportion of hospital readmissions deemed avoidable: a systematic review". In: *CMAJ* 183.7 (2011), E391–E402. DOI: 10.1503/cmaj.101860.

[3] Sunil Kripalani et al. "Deficits in communication and information transfer between hospital-based and primary care physicians: implications for patient safety and continuity of care". In: *Journal of Hospital Medicine* 2.5 (2007), pp. 314–323. DOI: 10.1002/jhm.224.

[4] Eric A. Coleman et al. "The Care Transitions Intervention: Results of a Randomized Controlled Trial". In: *Journal of the American Geriatrics Society* 52.11 (2004), pp. 1821–1830. DOI: 10.1111/j.1532-5415.2004.52504.x.

[5] Brian W. Jack et al. "A Reengineered Hospital Discharge Program to Decrease Rehospitalization: A Randomized Trial". In: *Annals of Internal Medicine* 150.3 (2009), pp. 178–187. DOI: 10.7326/0003-4819-150-3-200902030-00007.

[6] Devan Kansagara et al. "Risk Prediction Models for Hospital Readmission: A Systematic Review". In: *JAMA* 306.15 (2011), pp. 1688–1698. DOI: 10.1001/jama.2011.1515.

[7] Alistair E. W. Johnson et al. "MIMIC-IV, a freely accessible electronic health record dataset". In: *Scientific Data* 10.1 (2023), p. 1. DOI: 10.1038/s41597-022-01899-x.

[8] Ary L. Goldberger et al. "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals". In: *Circulation* 101.23 (2000). RRID:SCR_007345, e215–e220.

[9] Alistair Johnson et al. *MIMIC-IV (version 3.1)*. PhysioNet. RRID:SCR_007345. 2024. DOI: 10.13026/kpb9-mt58.

[10] *MIMIC-IV v3.1 on PhysioNet*. https://physionet.org/content/mimiciv/3.1/. Accessed: 2025-10-17.

[11] National Academies of Sciences, Engineering, and Medicine et al. *Communities in Action: Pathways to Health Equity*. Ed. by James N. Weinstein et al. Washington, DC: National Academies Press, 2017. ISBN: 978-0-309-45296-0. DOI: 10.17226/24624.

[12] George E. P. Box. "Science and statistics". In: *Journal of the American Statistical Association* 71.356 (1976), pp. 791–799.

[13] Leo Breiman. "Statistical modeling: The two cultures (with comments and a rejoinder by the author)". In: *Statistical science* 16.3 (2001), pp. 199–231.

[14] Rich Caruana et al. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2015, pp. 1721–1730. DOI: 10.1145/2783258.2788613.

[15] Ziad Obermeyer et al. "Dissecting racial bias in an algorithm used to manage the health of populations". In: *Science* 366.6464 (2019), pp. 447–453. DOI: 10.1126/science.aax2342.

[16] Xiao Lu et al. "Prediction and risk assessment of sepsis-associated encephalopathy in ICU based on interpretable machine learning". In: *Scientific Reports* 12.1 (2022), p. 22621. DOI: 10.1038/s41598-022-27134-6.

[17] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[18] Pedro Domingos. "A Unified Bias-Variance Decomposition and its Applications". In: *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000)*. San Francisco, CA: Morgan Kaufmann, 2000, pp. 231–238.

[19] David R. Cox. "The Regression Analysis of Binary Sequences". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 20.2 (1958), pp. 215–232.

[20] Leo Breiman et al. *Classification and regression trees*. CRC press, 1984.

[21] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

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

[23] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. "Learning representations by back-propagating errors". In: *nature* 323.6088 (1986), pp. 533–536.

[24] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. "Why do tree-based models still outperform deep learning on tabular data?" In: *Advances in Neural Information Processing Systems (NeurIPS 2022) Datasets and Benchmarks Track*. Vol. 35. 2022, pp. 507–520.

[25] Alejandro Barredo Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58 (2020), pp. 82–115.

[26] Zachary C. Lipton. "The Mythos of Model Interpretability". In: *Queue* 16.3 (2018), pp. 31–57.

[27] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2020. URL: https://christophm.github.io/interpretable-ml-book.

[28] Daniel W. Apley and Jingyu Zhu. "Visualizing the effects of predictor variables in black box supervised learning models". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82.4 (2020), pp. 1059–1086. DOI: 10.1111/rssb.12377.

[29] Scott M. Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*. 2017, pp. 4765–4774.

[30] Scott M. Lundberg et al. "From local explanations to global understanding with explainable AI for trees". In: *Nature Machine Intelligence* 2.1 (2020), pp. 56–67. DOI: 10.1038/s42256-019-0138-9.

[31] Ewout W Steyerberg, Andrew J Vickers, Peggy R Cook, et al. *Assessing the performance of prediction models: a framework for traditional and novel measures*. Vol. 21. 1. LWW, 2010, pp. 128–138.

[32] Takaya Saito and Marc Rehmsmeier. "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets". In: *PloS one* 10.3 (2015), e0118432.

[33] Jesse Davis and Mark Goadrich. "The relationship between Precision-Recall and ROC curves". In: (2006), pp. 233–240.

[34] Jerome H. Friedman. "Greedy function approximation: A gradient boosting machine". In: *Annals of Statistics* 29.5 (2001), pp. 1189–1232.

Appendix

Table 1: Complete pairwise association scores (≥ 0.6)

Var1	Var2	Type 1	Type 2	Method	Score	P-value
wbc_measured	rdw_measured	Cat.	Cat.	Cramér's V	0.998	< 0.001
hemoglobin_measured	rdw_measured	Cat.	Cat.	Cramér's V	0.996	< 0.001
hemoglobin_measured	wbc_measured	Cat.	Cat.	Cramér's V	0.995	< 0.001
sodium_measured	potassium_measured	Cat.	Cat.	Cramér's V	0.989	< 0.001
potassium_measured	bun_measured	Cat.	Cat.	Cramér's V	0.987	< 0.001
creatinine_measured	bun_measured	Cat.	Cat.	Cramér's V	0.981	< 0.001
platelet_measured	wbc_measured	Cat.	Cat.	Cramér's V	0.981	< 0.001
platelet_measured	rdw_measured	Cat.	Cat.	Cramér's V	0.981	< 0.001
sodium_measured	bun_measured	Cat.	Cat.	Cramér's V	0.980	< 0.001
rdw_last2_mean	rdw_max	Quant.	Quant.	Pearson	0.979	< 0.001
hemoglobin_measured	platelet_measured	Cat.	Cat.	Cramér's V	0.978	< 0.001
sodium_measured	glucose_measured	Cat.	Cat.	Cramér's V	0.976	< 0.001
creatinine_measured	potassium_measured	Cat.	Cat.	Cramér's V	0.974	< 0.001
sodium_measured	creatinine_measured	Cat.	Cat.	Cramér's V	0.966	< 0.001
potassium_measured	glucose_measured	Cat.	Cat.	Cramér's V	0.966	< 0.001
bun_measured	glucose_measured	Cat.	Cat.	Cramér's V	0.964	< 0.001
creatinine_measured	glucose_measured	Cat.	Cat.	Cramér's V	0.948	< 0.001
platelet_last2_mean	platelet_max	Quant.	Quant.	Pearson	0.945	< 0.001
creatinine_last2_mean	creatinine_max	Quant.	Quant.	Pearson	0.941	< 0.001
glucose_max	glucose_range	Quant.	Quant.	Pearson	0.924	< 0.001
hemoglobin_last2_mean	hemoglobin_max	Quant.	Quant.	Pearson	0.872	< 0.001
creatinine_measured	platelet_measured	Cat.	Cat.	Cramér's V	0.869	< 0.001
bun_last2_mean	bun_max	Quant.	Quant.	Pearson	0.868	< 0.001
wbc_measured	glucose_measured	Cat.	Cat.	Cramér's V	0.866	< 0.001
glucose_measured	rdw_measured	Cat.	Cat.	Cramér's V	0.866	< 0.001
hemoglobin_measured	glucose_measured	Cat.	Cat.	Cramér's V	0.865	< 0.001
platelet_measured	bun_measured	Cat.	Cat.	Cramér's V	0.864	< 0.001
creatinine_measured	hemoglobin_measured	Cat.	Cat.	Cramér's V	0.857	< 0.001
platelet_measured	potassium_measured	Cat.	Cat.	Cramér's V	0.857	< 0.001
sodium_measured	platelet_measured	Cat.	Cat.	Cramér's V	0.856	< 0.001
creatinine_measured	wbc_measured	Cat.	Cat.	Cramér's V	0.854	< 0.001
creatinine_measured	rdw_measured	Cat.	Cat.	Cramér's V	0.854	< 0.001
sodium_measured	wbc_measured	Cat.	Cat.	Cramér's V	0.853	< 0.001
platelet_measured	glucose_measured	Cat.	Cat.	Cramér's V	0.852	< 0.001
sodium_measured	rdw_measured	Cat.	Cat.	Cramér's V	0.852	< 0.001
sodium_measured	hemoglobin_measured	Cat.	Cat.	Cramér's V	0.851	< 0.001
wbc_measured	bun_measured	Cat.	Cat.	Cramér's V	0.849	< 0.001
bun_measured	rdw_measured	Cat.	Cat.	Cramér's V	0.849	< 0.001
hemoglobin_measured	bun_measured	Cat.	Cat.	Cramér's V	0.848	< 0.001
wbc_measured	potassium_measured	Cat.	Cat.	Cramér's V	0.842	< 0.001

Continued on next page

Table 1 – continued from previous page

Var1	Var2	Type 1	Type 2	Method	Score	P-value
potassium_measured	rdw_measured	Cat.	Cat.	Cramér's V	0.842	< 0.001
hemoglobin_measured	potassium_measured	Cat.	Cat.	Cramér's V	0.841	< 0.001
wbc_max	wbc_range	Quant.	Quant.	Pearson	0.806	< 0.001
creatinine_max	creatinine_range	Quant.	Quant.	Pearson	0.804	< 0.001
readmitted_30d	readmitted_15d	Cat.	Cat.	Cramér's V	0.803	< 0.001
bun_max	bun_range	Quant.	Quant.	Pearson	0.795	< 0.001
readmitted_7d	readmitted_15d	Cat.	Cat.	Cramér's V	0.771	< 0.001
wbc_last2_mean	wbc_max	Quant.	Quant.	Pearson	0.758	< 0.001
potassium_max	potassium_range	Quant.	Quant.	Pearson	0.749	< 0.001
creatinine_range	bun_range	Quant.	Quant.	Pearson	0.745	< 0.001
sodium_range	potassium_range	Quant.	Quant.	Pearson	0.738	< 0.001
admission_type	meds_present	Cat.	Cat.	Cramér's V	0.732	< 0.001
sodium_last2_mean	sodium_max	Quant.	Quant.	Pearson	0.726	< 0.001
num_unique_medications	sodium_range	Quant.	Quant.	Pearson	0.688	< 0.001
num_unique_medications	hemoglobin_range	Quant.	Quant.	Pearson	0.684	< 0.001
creatinine_max	bun_max	Quant.	Quant.	Pearson	0.682	< 0.001
age	charlson_index	Quant.	Quant.	Pearson	0.680	< 0.001
num_unique_medications	potassium_range	Quant.	Quant.	Pearson	0.666	< 0.001
sodium_range	hemoglobin_range	Quant.	Quant.	Pearson	0.664	< 0.001
discharge_location	meds_present	Cat.	Cat.	Cramér's V	0.661	< 0.001
hemoglobin_range	platelet_range	Quant.	Quant.	Pearson	0.660	< 0.001
discharge_location	num_unique_medications	Cat.	Quant.	Corr. Ratio	0.652	< 0.001
length_of_stay_days	num_unique_medications	Quant.	Quant.	Pearson	0.651	< 0.001
hemoglobin_range	potassium_range	Quant.	Quant.	Pearson	0.643	< 0.001
hemoglobin_range	wbc_range	Quant.	Quant.	Pearson	0.642	< 0.001
platelet_max	platelet_range	Quant.	Quant.	Pearson	0.641	< 0.001
meds_present	platelet_measured	Cat.	Cat.	Cramér's V	0.635	< 0.001
potassium_last2_mean	potassium_max	Quant.	Quant.	Pearson	0.635	< 0.001
meds_present	hemoglobin_measured	Cat.	Cat.	Cramér's V	0.626	< 0.001
meds_present	wbc_measured	Cat.	Cat.	Cramér's V	0.623	< 0.001
meds_present	rdw_measured	Cat.	Cat.	Cramér's V	0.623	< 0.001
length_of_stay_days	rdw_range	Quant.	Quant.	Pearson	0.621	< 0.001
hemoglobin_range	rdw_range	Quant.	Quant.	Pearson	0.620	< 0.001
readmitted_30d	readmitted_7d	Cat.	Cat.	Cramér's V	0.619	< 0.001
creatinine_range	bun_max	Quant.	Quant.	Pearson	0.619	< 0.001
sodium_range	bun_range	Quant.	Quant.	Pearson	0.618	< 0.001
platelet_range	wbc_range	Quant.	Quant.	Pearson	0.618	< 0.001
sodium_range	glucose_range	Quant.	Quant.	Pearson	0.617	< 0.001
creatinine_last2_mean	bun_last2_mean	Quant.	Quant.	Pearson	0.616	< 0.001
glucose_last2_mean	glucose_max	Quant.	Quant.	Pearson	0.616	< 0.001
num_unique_medications	platelet_range	Quant.	Quant.	Pearson	0.614	< 0.001
potassium_range	bun_range	Quant.	Quant.	Pearson	0.611	< 0.001
num_unique_medications	wbc_range	Quant.	Quant.	Pearson	0.609	< 0.001
meds_present	creatinine_measured	Cat.	Cat.	Cramér's V	0.608	< 0.001
creatinine_last2_mean	bun_max	Quant.	Quant.	Pearson	0.600	< 0.001

Table 2: Hyperparameter tuning results for logistic regression sorted by validation PR-AUC.

Penalty	C	L1 Ratio	Class Weight	Validation PR-AUC	Rank
elasticnet	10.00	0.2	—	0.3283816442	1
l2	10.00	—	—	0.3283815948	2
elasticnet	10.00	0.8	—	0.3283813448	3
l1	10.00	—	—	0.3283810910	4
elasticnet	10.00	0.5	—	0.3283809122	5
elasticnet	1.00	0.2	—	0.3283743591	6
l2	1.00	—	—	0.3283724171	7
elasticnet	1.00	0.5	—	0.3283721491	8
l1	1.00	—	—	0.3283703291	9
elasticnet	1.00	0.8	—	0.3283697085	10
l2	0.10	—	—	0.3282687827	11
elasticnet	0.10	0.2	—	0.3282661221	12
elasticnet	0.10	0.5	—	0.3282426176	13
l1	0.10	—	—	0.3282220185	14
elasticnet	0.10	0.8	—	0.3282172666	15
l2	0.01	—	—	0.3274501344	16
elasticnet	0.01	0.2	—	0.3272477532	17
l1	1.00	—	balanced	0.3268349956	18
elasticnet	1.00	0.8	balanced	0.3268330680	19
l1	10.00	—	balanced	0.3268313260	20
elasticnet	10.00	0.5	balanced	0.3268309571	21
elasticnet	10.00	0.8	balanced	0.3268308951	22
l2	10.00	—	balanced	0.3268308517	23
elasticnet	10.00	0.2	balanced	0.3268306704	24
elasticnet	1.00	0.5	balanced	0.3268296758	25
elasticnet	1.00	0.2	balanced	0.3268258885	26
l2	1.00	—	balanced	0.3268250832	27
l1	0.10	—	balanced	0.3268077245	28
elasticnet	0.10	0.8	balanced	0.3267918233	29
elasticnet	0.10	0.5	balanced	0.3267823503	30
elasticnet	0.10	0.2	balanced	0.3267758231	31
l2	0.10	—	balanced	0.3267621901	32
elasticnet	0.01	0.5	—	0.3267048351	33
elasticnet	0.01	0.8	—	0.3264610991	34
l1	0.01	—	—	0.3261625478	35
elasticnet	0.01	0.2	balanced	0.3260151347	36
l2	0.01	—	balanced	0.3259931855	37
elasticnet	0.01	0.5	balanced	0.3257844684	38
elasticnet	0.01	0.8	balanced	0.3254839493	39
l1	0.01	—	balanced	0.3253139563	40

Table 3: XGBoost hyperparameter tuning results sorted by validation PR-AUC.

Max Depth	**Learning Rate**	**Trees**	**Validation PR-AUC**	**Rank**
6	0.050	500	0.375441	1
5	0.050	500	0.375021	2
7	0.020	500	0.373983	3
4	0.200	500	0.373286	4
6	0.050	500	0.372927	5
7	0.020	500	0.372073	6
5	0.050	500	0.370666	7
4	0.200	500	0.369132	8
8	0.100	500	0.368383	9
4	0.050	500	0.366843	10
4	0.050	500	0.364994	11
8	0.100	500	0.363586	12
3	0.010	500	0.338839	13
3	0.010	500	0.337136	14

Table 4: Neural Network hyperparameter tuning tournament results, ranked by Validation PR-AUC.

Rank	**Model ID**	**Hidden Layers**	**Units per Layer**	**Dropout Rate**	**L2 Penalty**	**Learning Rate**	**Batch Norm**	**Val PR-AUC**
1	**Model 7**	**4**	**256**	**0.5**	10^{-4}	**0.0005**	**No**	**0.3804**
2	Model 9	4	128	0.3	10^{-4}	0.0010	No	0.3795
3	Model 3	2	256	0.4	10^{-4}	0.0001	No	0.3777
4	Model 6	2	64	0.1	10^{-4}	0.0010	No	0.3760
5	Model 2	2	256	0.3	0	0.0010	No	0.3735
6	Model 1	1	512	0.2	10^{-5}	0.0010	No	0.3733
7	Model 8	3	256	0.5	10^{-3}	0.0005	No	0.3732
8	Model 10	3	64	0.4	10^{-3}	0.0010	No	0.3664
9	Model 4	3	128	0.2	10^{-4}	0.0010	Yes	0.3389
10	Model 5	2	256	0.4	10^{-5}	0.0005	Yes	0.3387