| | |
|---|---|
| **Titre:** Title: | Interpretable Predictive Modeling for Patient Readmission in Intensive Care Units |
| **Auteur:** Author: | Waleed Sayed Ahmed Fathy Gharib |
| **Date:** | 2025 |
| **Type:** | Mémoire ou thèse / Dissertation or Thesis |
| **Référence:** Citation: | Gharib, W. S. A. F. (2025). Interpretable Predictive Modeling for Patient Readmission in Intensive Care Units [Thèse de doctorat, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/70097/ |

**POLYTECHNIQUE MONTRÉAL**
affiliée à l'Université de Montréal

# Interpretable predictive modeling for patient readmission in intensive care units

**WALEED SAYED AHMED FATHY GHARIB**
Département de génie informatique et génie logiciel

Thèse présentée en vue de l'obtention du diplôme de *Philosophiæ Doctor*
Génie informatique

Octobre 2025

**POLYTECHNIQUE MONTRÉAL**
affiliée à l'Université de Montréal

Cette thèse intitulée :

**Interpretable predictive modeling for patient readmission in intensive care units**

présentée par **Waleed Sayed Ahmed Fathy GHARIB**
en vue de l'obtention du diplôme de *Philosophiæ Doctor*
a été dûment acceptée par le jury d'examen constitué de :

**Amal ZOUAQ**, présidente
**Farida CHERIET**, membre et directrice de recherche
**Guillaume EMERIAUD**, membre et codirecteur de recherche
**Michael SAUTHIER**, membre
**Jean Marc LINA**, membre externe

# DEDICATION

*To my beloved family . . .*

*I dedicate this thesis to the loving memory of my dear parents, who are no longer with us. I know how much they would have loved to witnessing this special moment. This achievement is also theirs.*

# ACKNOWLEDGEMENTS

All praise is due to Allah, the Most High in status, the Most Merciful, the Most Compassionate—who has bestowed upon me from His bounty and vast generosity, and who supported me in completing this journey and overcoming all the hardships and challenges I faced. Were it not for His divine guidance, I would not have been able to overcome the obstacles that nearly caused me to collapse and lose hope. But by His great grace, He helped me rise every time I was about to fall. So, all praise is due to Allah, abundant and blessed, in a manner befitting His majestic countenance and His mighty authority.

I would like to express my deepest gratitude to my supervisor, Prof. Farida Cheriet, and my co-supervisor, Dr. Guillaume Emerieud, for their unwavering support, invaluable guidance, and continuous encouragement throughout my PhD journey.

To my beloved wife, thank you for your boundless love, patience, and care. Your presence has been my greatest source of strength and comfort.

To my brothers, your unwavering support was the pillar beneath my steps, Your belief in me lit the path when mine began to fade. When the nights were long and the weight was heavy, You stood beside me—not with words alone, But with hearts full of care and hands that never let go. You were my quiet strength, my constant refuge, The ones who saw in me what I sometimes could not. This journey is mine in name—but yours in spirit. Thank you—for everything, and more than I can ever say.

To my dear friends and lab colleagues, your encouragement and kindness have been a source of motivation along the way.

I would also like to extend my sincere thanks to Philippe Debanne, the lab assistant, for his constant willingness to help. His prompt support and readiness to solve any problem I encountered made a significant difference throughout my research journey.

Finally, I would like to thank Sally Al-Omar for her guidance during the data extraction from the Sainte-Justine Hospital database. Her expertise helped me understand the database thoroughly and substantially simplified the process.

This work is financially supported by the **Opsidian Create program** of the **Natural Sciences and Engineering Research Council of Canada (NSERC)** and the **Fonds FSISSS of Medteq**. I sincerely appreciate this funding, which made this research possible.

# RÉSUMÉ

La prédiction des réadmissions en unité de soins intensifs (USI, *ICU*) est un défi majeur, car les retours non planifiés signalent souvent une récupération incomplète, une fragilité particulière, et un risque de délais dans la prise en charge, qui résultent en un risque accru de mortalité et des coûts de santé plus élevés. La réadmission en USI pédiatrique (USIP, *PICU*) est particulièrement peu étudiée en raison de la rareté de grandes bases de données publiques et de la variabilité physiologique propre aux différents groupes d'âge pédiatriques.

La plupart des études sur la réadmission en soins intensifs portent sur des adultes, alors que cette thèse met en lumière les défis propres au contexte pédiatrique, nécessitant la prise en compte de caractéristiques spécifiques au domaine et des stratégies de modélisation adaptées. Cette thèse examine s'il est possible de développer un modèle à la fois performant et interprétable pour prédire la réadmission en USIP à partir d'ensembles de données cliniques de taille limitée et de distribution non équilibrée collectées sur l'ensemble du séjour du patient. Nous proposons une approche en deux étapes. Dans un premier temps, nous concevons une procédure prédictive complète sur la base publique adulte MIMIC-III (ICU), permettant d'évaluer systématiquement les stratégies de prétraitement, la modélisation des caractéristiques multimodales (statistiques, provenant de connaissances à priori du domaine et de techniques de traitement de signal), la sélection de caractéristiques discriminantes et l'implémentation de différentes familles de modèles. Dans un second temps, nous étendons et optimisons cette procédure pour le jeu de données pédiatrique du CHU Sainte-Justine (CathyDB) via des étapes adaptées à la pédiatrie : extraction d'informations additionnelles (p. ex. scores pédiatriques), imputations spécifiques selon la variable, filtrage des signes vitaux en deux étapes à l'aide du filtre de Kalman et une transformation par ondelettes, représentations catégorielles dépendantes de l'âge (normal/bas/élevé avec écart aux normes pédiatriques), extraction de caractéristiques basées sur des connaissances à priori du domaine plus puissantes, et sélection des caractéristiques les plus discriminantes par information mutuelle et coefficients LASSO, complétée par une recherche à l'aide d'une « enveloppe » bidirectionnelle (arrière–avant).

Sur le banc d'essai adulte, un modèle de type Light Gradient Boosting Machine (LightGBM) a atteint une aire sous la courbe ROC (AUROC) de 78.6% pour la réadmission après 3 jours, montrant une bonne discrimination dans un contexte hétérogène et soulignant l'intérêt des caractéristiques sélectionnées. Le classement des variables a mis en évidence des descripteurs du domaine fréquentiel (p. ex. coefficients FFT, transformations angulaires) ainsi que des représentations vectorielles des diagnostics (charge de comorbidités). Les résumés des signes

vitaux (fréquence cardiaque, fréquence respiratoire), les marqueurs biologiques (taux d'urée sanguine, numération leucocytaire) et les indicateurs thérapeutiques (adrénaline) figuraient parmi les facteurs de risque cliniques les plus marquants. Cette distribution des variables les plus discriminantes confirme la nécessité d'une approche multimodale intégrant les informations diagnostiques, physiologiques, biologiques et thérapeutiques.

Pour la prédiction en USIP, un modèle de régression logistique (LR) entraîné sur un sous-ensemble multimodal sélectionné a fourni les performances les plus constantes entre jeux de données (AUROC test : 87.8%), surpassant d'autres modèles linéaires, des ensembles d'arbres et des modèles profonds. Cela montre qu'un modèle simple comme le LR peut dépasser des approches plus complexes lorsque les variables sont bien choisies. Les méthodes à base d'arbres ont obtenu de très bons scores d'entraînement mais ont sur-appris et se sont dégradées sur la détection de la classe minoritaire, faute d'exemples positifs. Malgré l'usage des seules 24 dernières heures de séries temporelles et l'exclusion de données statiques (p. ex. biologie), les modèles profonds (BiLSTM avec attention et Transformer) sont restés compétitifs (AUROC 79.2% et 78.5%), laissant entrevoir des gains potentiels avec des bases de données d'entrainement plus grandes. Notre modèle LR interprétable a nettement dépassé les prédicteurs de réadmission en USIP publiés auparavant, atteignant 87.8% d'AUROC test contre 64–70% dans les études antérieures.

Nous avons démontré la cohérence du modèle prédictif proposé avec le protocole utilisé en clinique via des méthodes d'explicabilité globales (coefficients LR, SHAP) et des études d'ablations. Les analyses d'interprétabilité montrent que les caractéristiques basées sur les connaissances à priori du domaine et celles issues du traitement des signaux physiologiques sont les contributions majeures dans la prédiction—rendant leur extraction explicite essentielle. Sur le plan clinique, elles incluent les anomalies des signes vitaux, certains paramètres biologiques, l'équilibre hydrique, les scores de sédation et l'usage de médicaments spécifiques comme indicateurs clés de risque de réadmission. Les études d'ablations confirment le caractère indispensable de plusieurs variables (p. ex. acide acétylsalicylique, digoxine, anomalies du débit urinaire), suggérant des pistes de prise de décision pour les cliniciens.

Dans l'ensemble, nos résultats montrent qu'une prédiction précise et interprétable de la réadmission en USIP est réalisable avec des ensembles de données de taille limitée et de distribution non équilibrée, en s'appuyant sur un prétraitement rigoureux, des représentations spécifiques à la pédiatrie et une sélection de modèles fondée sur des métriques spécifiques. De surcroît, en combinant des techniques avancées de traitement de signal avec des variables basées sur des connaissances à priori du domaine, cette étude contribue à réduire l'écart entre la performance prédictive et l'applicabilité clinique.

D'un point de vue global, cette étude contribue à l'avancement des modèles prédictifs en soins intensifs pédiatriques en conciliant précision et interprétabilité. Contrairement aux travaux antérieurs reposant sur des approches de type « boîte noire » ou sur des données adultes, notre approche montre comment une modélisation et une sélection rigoureuse de caractéristiques permet de prendre en compte l'hétérogénéité pédiatrique. Les résultats prometteurs laissent entrevoir, après une validation plus exhaustive sur des ensembles de données provenant d'autres institutions , l'évolution vers un système d'aide à la décision clinique en temps réel, capable d'identifier les patients à haut risque de réadmission et d'expliquer les facteurs sous-jacents. Une telle interprétabilité offre aux cliniciens un appui pour orienter leur décision de prescrire la sortie du patient de l'USI, de prolonger le séjour si nécessaire ou assurer une surveillance renforcée après la sortie de l'USI, contribuant ainsi à réduire les réadmissions évitables et à améliorer la qualité des soins critiques pédiatriques.

# ABSTRACT

Intensive Care Unit (ICU) readmission prediction is a critical challenge in critical care medicine, as unplanned returns often indicate incomplete recovery, particular frailty, and a risk of delays in care, resulting in increased mortality risk, and higher healthcare costs. Pediatric Intensive Care Unit (PICU) readmission prediction is particularly underexplored due to the scarcity of large, publicly available datasets and the unique physiological variability across pediatric age groups.

While most ICU readmission studies focus on adult populations, our thesis highlights the unique challenges in PICU readmission prediction, where pediatric heterogeneity requires tailored feature engineering and modeling strategies. This thesis investigates whether a high-performing and interpretable model can predict pediatric PICU readmission using limited, imbalanced data collected over the full length of stay. We propose a two-stage approach. First, we design a complete predictive pipeline on the public adult ICU database MIMIC-III, enabling systematic evaluation of preprocessing, multimodal feature engineering (statistical, medical knowledge-based, and signal-processing descriptors), feature selection, and model families. Second, we extend and optimize this pipeline to the CHU Sainte-Justine PICU dataset (CathyDB) with pediatric-aware steps: Extracting more data such as pediatric score, using category-aware imputation, a two-stage Kalman–wavelet denoising of vital signs, age-aware categorical state representations (normal/low/high with deviation from pediatric norms), Extracting more powerful clinical meaningful features and feature selection via mutual information and LASSO coefficients plus bidirectional (backward–forward) wrapper search.

On the adult benchmark, a Light Gradient Boosting Machine (LightGBM) achieved an Area Under the Receiver Operating Characteristic (AUROC) curve of 78.6% for 3-day readmission, showing strong discrimination in a heterogeneous setting and highlighting the value of medical knowledge-based and frequency-domain features. Feature ranking highlighted medical knowledge-based features and frequency-domain descriptors (e.g., FFT coefficients, angle-based transforms) as the most predictive, alongside diagnosis embeddings (comorbidity burden). Vital-sign summaries (heart and respiratory rates), laboratory markers (blood urea nitrogen difference, white blood cell count), and medications (epinephrine) were the most significant clinical risk factors. This distribution of top-ranked features underscores the need for a multimodal approach that integrates diagnostic, physiologic, laboratory, and therapeutic information.

For PICU prediction, a Logistic Regression (LR) model trained on a selected multimodal subset delivered the most consistent performance (test AUROC 87.8%), outperforming other linear baselines, tree ensembles, and deep learning models. This shows that even a simple model like LR can outperform more complex models when features are well-selected. Tree-based methods attained high training scores but overfit and degraded on minority-class detection due to the lack of training positive samples. Despite using only last-24h time series and excluding some important static data such as lab results, deep learning models (Bidirectionl Long Short Time Memory (BiLSTM) with attention and Transformer) remained competitive (AUROC 79.2% and 78.5%), suggesting headroom with richer inputs. Our interpretable LR model significantly outperformed the previously published PICU readmission predictors, achieving a test AUROC of 87.8%, compared to 64–70% in earlier studies.

We aligned model predictions with clinical reasoning by applying global explainability methods—including LR coefficients, Shapley Additive exPlanations (SHAP), and ablation analysis—at both the feature-engineering and clinical-variable levels. To complement this, we employed Local Interpretable Model-agnostic Explanations (LIME) to provide case-specific insights into individual predictions. Feature engineering interpretability analyses showed that medical knowledge-based and signal-processing based features are the strongest contributors—making their explicit extraction essential. Clinically, it identifies vital-sign abnormalities, lab values, fluid balance measures, sedation scores, and specific medication usage as key indicators of readmission risk. Feature ablation confirmed the indispensability of several variables (e.g., Acide acetylsalicylique, Digoxine administration, urine output abnormalities), suggesting actionable insights for clinicians. In addition, we applied local interpretability methods to identify the key risk factors driving the model's prediction for each individual patient. Overall, the results show that accurate, interpretable PICU readmission prediction is feasible with limited, imbalanced data when supported by rigorous preprocessing, pediatric-specific representations, and principled model selection. In addition, by integrating advanced signal-processing with clinically interpretable features, this study bridges the gap between predictive performance and clinical usability.

From a broader perspective, this study advances the development of predictive models in pediatric critical care by combining accuracy with interpretability. Unlike prior works that often rely on black-box methods or adult ICU data, our framework demonstrates how targeted feature engineering can address the unique heterogeneity of pediatric patients. The promising findings suggest that, with further extension and refinement, this model could evolve into a real-time clinical decision support system capable of flagging high-risk patients and explaining the underlying factors driving readmission risk. Such interpretability empowers clinicians to make more informed discharge decisions, prolonging stays when needed, or pri-

oritize monitoring of key risks after discharge. Ultimately, this could help reduce preventable readmissions and improve the overall quality of pediatric critical care.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF SYMBOLS AND ACRONYMS

| | |
|---|---|
| ACA | Affordable Care Act |
| AdaBoost | Adaptive Boost |
| AHRQ | Agency for Healthcare Research and Quality |
| AFS-DT | DT based on Axiomatic Fuzzy Set theory |
| AI | Artificial Intelligence |
| ALP | ALkaline Phosphatase |
| ALT | ALanine Transaminase |
| ANIT-UKM | Anesthesiology, Intensive Care and Pain Medicine at the University Hospital Münster |
| APACHE | Acute Physiology and Chronic Health Evaluation |
| APS | Acute Physiology Score |
| ARC | Absolute Reticulocyte Count |
| ATE | Average Treatment Effect |
| AUROC | Area Under the Receiver Operating Characteristic curve |
| AUPRC | Area Under Precision-Recall Curve |
| BERT | Bidirectional Encoder Representations from Transformers |
| BFSS | Binary Fish School Search |
| BIDMC | Beth Israel Deaconess Medical Center |
| BOW | Bag-Of-Words |
| BUN | Blood Urea Nitrogen |
| CAE | Convolutional AutoEncoders |
| CathyDb | CHU Sainte-Justine PICU database |
| CC | Conceptual-Contextual |
| CCC | Complex Chronic Conditions |
| CCS | Clinical Classification Software |
| CCSR | Clinical Classifications Software Refined |
| CDF | Cumulative Distribution Function |
| CHOA | Children's Healthcare of Atlanta |
| CHU | Center Hospital University |
| CNN | Convolutional Neural Network |
| COPD | Chronic Obstructive Pulmonary Disease |
| CRF | Conditional Random Fields |
| CTCL | CodeText cross-modal Contrastive Learning |

| | |
|---|---|
| CUH | Cambridge University Hospital |
| CUI | Concept Unique Identifier |
| CFS | Correlation-based Feature Selection |
| DAG | Directed Acyclic Graph |
| DBI | Davies-Bouldin Index |
| DBP | Diastolic Blood Pressure |
| DL | Deep Learning |
| DLMs | Dynamic Linear Models |
| DT | Decision Trees |
| DFT | Discrete Fourier Transform |
| DWT | Discrete Wavelet Transform |
| EFB | Exclusive Feature Bundling |
| EBM | Explainable Boosting Machine |
| ECMO | Extracorporeal Membrane Oxygenation |
| EHR | Electronic Health Record |
| eICU | eICU Collaborative Research Database |
| EM | Expectation-Maximization |
| EFB | Exclusive Feature Bundling |
| Extra-Tree | Extremely Randomized Tree |
| ET | Extra Tree |
| FCM | Fuzzy C-Means |
| FFP | Fuzzy FingerPrint |
| FFT | Fast Fourier Transform |
| FN | False Negatives |
| FP | False Positives |
| GAT | Graph Attention Network |
| GAE | Graph Auto-Encoder |
| GBM | Gradient Boosting Machine |
| GCN | Graph Convolutional Network |
| GCS | Glasgow Coma Scale |
| GCT | Graph Convolutional Transformer |
| GK | Gustafson-Kessel |
| GLU | Glucose |
| GNN | Graph Neural Network |
| GOSS | Gradient-based One-Side Sampling |
| GPT-2 | Generative Pre-trained Transformer-2 |

| | |
|---|---|
| GROM | Graph Attention and RNN-based Neural ODE Model |
| GRU | Gated Recurrent Unit |
| GUI | Graphical User Interface |
| HAT | Hypergraph Attention Network |
| HCL | Hypergraph Contrastive Learning |
| Hgb | Hemoglobin |
| HCT | HematoCriT |
| HR | Heart Rate |
| HRRP | Hospital Readmissions Reduction Program |
| ICD | International Classification of Diseases |
| ICU | Intensive Care Unit |
| IDSS | Intelligent Decision Support System |
| IDWT | Inverse Discrete Wavelet Transform |
| IICUPM | Intelligent ICU Patient Monitoring |
| IG | Information Gain |
| INESSS | Institute National d'Excellence en Santé et en Services Sociaux |
| KNN | k-Nearest Neighbors |
| KG | Knowledge Graph |
| LASSO | L1-penalized logistic regression |
| LightGBM | Light Gradient Boosting Machine |
| LLM | Large Language Model |
| LIME | Local Interpretable Model-agnostic Explanations |
| LOCF | Last Observed Carried Forward |
| LODs | Logistic Organ Dysfunction score |
| LOS | Length Of Stay |
| LR | Logistic Regression |
| LR-Test | Likelihood Ratio-Test |
| LSTM | Long Short-Term Memory |
| Bi-LSTM | Bidirectional LSTM |
| MAD | Median Absolute Deviation |
| MBP | Mean Blood Pressure |
| MEWS | Modified Early Warning Score |
| MI | Mutual Information |
| MICE | Multiple Imputation by Chained Equations |
| MFC | Mixed Fuzzy Clustering |
| MGCN | Multi-view Graph Convolution Network |

| | |
|---|---|
| MIMIC | Multiparameter Intelligent Monitoring in Intensive Care |
| ML | Machine Learning |
| MLODS | Modified Logistic Organ Dysfunction score |
| MRI | Magnetic Resonance Imaging |
| MP-ROM | Multitask learning with Pearson and RNN-based neural ODEs Model |
| NIBIB | National Institute of Biomedical Imaging and Bioinformatics |
| NB | Naive Bayes |
| NCIT | National Cancer Institute Thesaurus |
| NLP | Natural Language Processing |
| NMF | Negative Matrix Factorization |
| NEWS | National Early Warning Score |
| NN | Neural Network |
| NRL | Noise Reduction Learning |
| OASIS | Oxford Acute Severity of Illness Score |
| ODE | Neural Ordinary Differential Equation |
| P | Phosphor |
| PCA | Principal Component Analysis |
| PELOD | Pediatric Logistic Organ Dysfunction |
| PEWS | Pediatric Early Warning Systems |
| PFS | Probabilistic Fuzzy Systems |
| PICU | Pediatric Intensive Care Unit |
| PLT | Platelets |
| PPM | Predictive Process Monitoring |
| PSD | Power Spectral Density |
| RBF | Radial Basis Function |
| RF | Random Forest |
| ReLU | Rectified Linear Unit |
| RNN | Recurrent Neural Network |
| RR | Respiratory Rate |
| S_K_GT | Statistics and Knowledge-based Graph Transformer |
| SANMF | Subgraph Augmented Non-negative Matrix Factorization |
| SAPS | Simplified Acute Physiology Score |
| SampEn | Sample Entropy |
| SBP | Systolic Blood Pressure |
| SHAP | SHapley Additive exPlanations |
| SMOTE | Synthetic Minority Over-sampling Technique |

| | |
|---|---|
| SNOMED CT | Systematized Nomenclature of Medicine-Clinical Terms |
| SOFA | Sequential Organ Failure Assessment |
| SPO2 | Oxygen saturation |
| STARR | Stanford Medicine Research Data Repository |
| SU | Symmetrical Uncertainty |
| SVM | Support Vector Machine |
| SFS | Sequential Forward Selection |
| SWIFT | Stability and Workload Index for Transfer |
| TN | True Negatives |
| TP | True Positives |
| TPE | Tree-structured Parzen Estimator |
| TSFM | Takagi-Sugeno Fuzzy Model |
| TSH | Thyroid-Stimulating Hormone |
| UMLS | Unified Medical Language System |
| VGNN | Variationally Regularized Graph Neural Network |
| WPT | Wavelet Packet Transform |
| XGBoost | Extreme Gradient Boosting |

# LIST OF APPENDICES

# CHAPTER 1 INTRODUCTION

## 1.1 Background and Motivation

Ensuring patient safety and reducing clinical risks are fundamental objectives in modern healthcare systems. Hospital administrators and healthcare providers continuously strive to enhance the quality of care, optimize resource utilization, and minimize adverse events. Among the most pressing challenges is the reduction of clinical errors, particularly in high-stakes environments such as Intensive Care Units (ICUs), where critically ill patients require close monitoring and rapid intervention. Due to the severity of cases managed within ICUs, any lapse in care can result in significant morbidity or mortality. Consequently, intensive care settings are focal points for quality improvement initiatives, clinical risk management, and operational efficiency [1–3].

The Pediatric Intensive Care Unit (PICU) represents a specialized branch of critical care, delivering life-saving treatment to children with severe or life-threatening conditions. Given the complexity of pediatric physiology and the vulnerability of the population, patient safety in the PICU is of paramount importance. PICU readmission, defined as the unplanned return of a pediatric patient to the ICU within a short interval following discharge, is a key indicator of premature discharge and unmet clinical needs. Challenges such as limited ICU bed availability and high treatment costs often necessitate difficult clinical decisions regarding discharge timing. In practice, physicians may be compelled to discharge patients early to free up beds for incoming critical cases. This practice, while addressing system-level constraints, can lead to premature discharges, exposing patients to significant risks if they are not yet clinically stable. In such cases, deterioration may occur following transfer to general wards, leading to unplanned readmission to the PICU [4–7].

Studies have reported readmission rates between 4% and 10%, underscoring its prevalence and potential preventability [8]. Importantly, readmitted patients are typically in a more fragile state and may experience delays in receiving critical care due to limited bed availability. As a result, they are significantly more likely—two to ten times—to experience mortality compared to non-readmitted patients. These consequences not only impact patient outcomes but also place a substantial strain on already scarce critical care resources, contributing to prolonged hospitalization, increased resource use, elevated treatment costs and reduced system efficiency [9–11].

Recognizing these risks, various health authorities have emphasized the importance of addres-

sing ICU readmissions. The Institute National d'Excellence en Santé et en Services Sociaux (INESSS) in 2018 formally recommended the implementation of quality improvement frameworks targeting ICU performance. Similarly, in the United States, the Affordable Care Act (ACA) introduced the Hospital Readmissions Reduction Program (HRRP), which penalizes hospitals financially for high readmission rates, thereby elevating readmissions to a metric of national importance for healthcare quality and accountability [12, 13].

With the growth of Electronic Health Records (EHRs), healthcare systems now have access to a wealth of structured and time-series data [14, 15]. EHR databases are comprehensive digital repositories that systematically collect and store patient health information, including demographics, diagnoses, treatments, laboratory results, and clinical notes. They enable large-scale analysis of real-world healthcare data and support the development of predictive models and quality improvement initiatives. However, the complex analysis of vital signs and other patient data collected in the ICU makes clinical prediction challenging. A huge amount of information is available, and it's hard for human brain to analyze all of them at the same time.

Artificial Intelligence (AI) algorithms, including Machine Learning (ML) and Deep Learning (DL), have shown great success in medical diagnosis and decision-making, utilizing numerous features that human analysis cannot handle. As doctors rely on intuition and clinical judgments, algorithms have the capacity to rapidly analyze more information with greater precision and accuracy [15–20]. Using these EHR data with these data-driven approaches makes it possible to tackle unplanned readmissions and aid doctors in decision-making. Despite this, the development of robust predictive tools in this domain remains underexplored. Challenges such as data heterogeneity, high rates of missing values, patient age-related variability, and severe class imbalance in readmission data hinder the effectiveness of conventional statistical methods [21].

Relatively few efforts have focused specifically on ICU or PICU readmissions, despite their distinct clinical and operational implications. Among existing research, many studies have examined ICU or PICU readmission as a variable correlated with patient outcomes or have investigated risk factors through retrospective statistical analyses [11, 22]. However, these studies often stop short of developing predictive models capable of forecasting readmissions in real time or guiding clinical decision-making.

## 1.2 Problem Statement

Unplanned readmissions to the ICU/PICU are associated with increased morbidity, healthcare costs, and caregiver burden [11]. They may also indicate suboptimal discharge timing, unresolved clinical issues, or gaps in care continuity. Early and accurate identification of patients at high risk for PICU readmission can support more informed discharge planning, proactive monitoring, and better allocation of healthcare resources. Despite this clinical importance, reliable and interpretable tools for predicting PICU readmission remain limited.

Predicting ICU/PICU readmissions is a challenging task for several reasons. The first challenge lies in the availability and complexity of public databases. Currently, only two widely used ICU databases are publicly available. Extracting relevant data from these sources is difficult due to their non-standardized structure, the dispersion of information across multiple tables, and the use of different identifiers. This often requires manual effort to trace and verify patient-level data such as ICU admission times and stay IDs. Additionally, missing or invalid values in crucial fields like patient identifiers and timestamps further complicate preprocessing. In some cases, high-performance computing resources are required to efficiently handle data extraction and processing, posing an accessibility barrier for some researchers.

The second challenge involves the poor quality and complexity of ICU data itself. Clinical variables frequently suffer from high rates of missingness, inconsistent measurement units, and outliers [23]. High dimensionality and the asynchronous nature of time-series measurements make data integration and alignment particularly difficult, often leading to biased predictions and reduced model accuracy [21].

The third challenge is the development of robust, generalizable prediction models in a highly heterogeneous and imbalanced clinical environment. ICU patients present with a wide range of evolving conditions and comorbidities, and confounding factors such as age, gender, and socioeconomic status influence readmission risk. The typical readmission rate is low (around 4%–10%) [8], making the prediction task inherently imbalanced. Furthermore, discharge assessments tend to show normalized values for most patients, limiting the discriminative power of standard clinical metrics. In particular, models often rely heavily on statistical features, and do not fully exploit the rich temporal and spectral information available in physiological signals. This makes it difficult to detect subtle physiological or clinical signals that may precede deterioration or readmission.

The fourth challenge is interpretability, a crucial requirement in medical decision-making systems. Clinicians must understand and trust model predictions to ensure patient safety and justify treatment decisions [24]. Yet, interpretability is difficult to achieve : complex models

capture subtle nonlinear patterns but are opaque, while simpler models are transparent yet often less accurate. Designing explanations that are both faithful to the model and clinically meaningful further complicates this issue, leaving many state-of-the-art models insufficiently interpretable for clinical practice [25].

The fifth challenge, pediatric-specific variables differ substantially from adult ICU settings, requiring tailored modeling approaches.

While data-driven methods such as ML and DL have achieved success in various clinical tasks, their application in PICU readmission prediction is constrained by these unresolved challenges. To address these limitations, this thesis proposes a transparent and clinically meaningful framework for predicting PICU readmission during the first three days after discharge. The approach combines structured clinical data with statistical, temporal, and spectral features extracted from vital signs, and integrates robust preprocessing, age-specific normalization, and interpretable machine learning models. The goal is to support physicians with a practical decision support tool that identifies high-risk patients early, improves discharge timing, and ultimately reduces preventable PICU readmissions.

### 1.3 Structure of The Thesis

This thesis is structured as follows :

— Chapter 2 : This chapter provides a structured review of the literature on ICU and PICU readmission. It begins by examining studies that have identified clinical risk factors, followed by researches that have proposed predictive models using statistical, ML, and DL approaches. In addition, it discusses feature representation and the methods employed for interpretability, highlights issues of model calibration and generalization, and assesses the extent to which practical prediction models meet key clinical requirements. The discussion then turns to the main gaps and limitations in existing works, with particular emphasis on the challenges unique to pediatric populations. Building on these insights, the chapter introduces the research questions and objectives that guide this thesis, and highlights the contribution of the proposed modeling strategy. This chapter is largely based on our published review paper, “A comprehensive review of ICU readmission prediction models : From statistical methods to deep learning approaches”, in Artificial Intelligence in Medicine (Vol. 165, 2025) [21].

— Chapter 3 : In this chapter, we present the Methodology, detailing the proposed framework for predicting ICU and PICU readmission. We begin by describing the databases used in this study, followed by the preprocessing techniques, including data cleaning

and imputation strategies. In particular, we propose variable-specific imputation methods guided by clinical considerations to better reflect the nature of each variable and to avoid implausible continuous values. We then introduce a two-stage filtering process for noise reduction to remove outliers while preserving salient signal characteristics. Next, we present the different feature extraction approaches, encompassing statistical descriptors, signal-processing–based features, and novel medical knowledge–based features. Subsequently, we outline the proposed feature selection method to identify the most relevant features, and discuss approaches for addressing data imbalance. The chapter then highlights the predictive models employed and the rationale for their selection, before introducing the interpretability techniques used to identify the most influential features for each classifier. Finally, we present the evaluation metrics adopted to assess model performance, including a custom loss-based evaluation score (class-wise loss score) introduced to mitigate bias toward the majority class in our highly imbalanced prediction problem.

— Chapter 4 : In this chapter, we present the ICU readmission prediction framework based on an interpretable Light Gradient Boosting Machine (LightGBM) model, together with its results and limitations, as described in our published paper "Prediction of ICU Readmission Using LightGBM Classifier" (2023 IEEE 20th International Symposium on Biomedical Imaging, ISBI) [26]. We then explain how this model is extended and optimized to overcome its limitations and to address pediatric heterogeneity, thereby better capturing the unique dynamics of pediatric cases. Subsequently, we discuss the results of the proposed framework when evaluated with different ML and DL models, including the impact of incorporating medical knowledge–based features and signal-processing–based features. Finally, we present the results of interpretability analyses, examining both the engineered features that most influenced model performance and the clinical variables most strongly associated with readmission.

— Chapter 5 : In this chapter, we discuss how the proposed ICU and PICU readmission prediction frameworks address the research objectives. We outline the strengths and limitations of each framework, and examine how the identified gaps in the literature are tackled, assessing whether our proposed methods effectively bridge these gaps. Furthermore, we present the results of interpretability analyses to evaluate the added value of the proposed feature extraction methods compared with standard statistical features. Finally, we highlight the key risk factors revealed by these analyses that physicians should consider when making clinical decisions related to readmission.

— Chapter 6 : In this chapter, we conclude the thesis by summarizing the key findings and reflecting on their implications for pediatric critical care. We also outline future

directions aimed at addressing the current limitations of our work and extending the proposed frameworks toward real-time prediction, with the ultimate goal of enabling practical implementation in PICUs.

# CHAPTER 2 LITERATURE REVIEW

Unplanned readmission to the ICU, particularly the PICU, is a high-risk event that reflects both clinical deterioration and potential inadequacies in discharge decision-making. As such, ICU and PICU readmissions have emerged as critical metrics of hospital performance and patient safety. In recent years, the proliferation of EHR data has enabled researchers to study both risk factors and predictive modeling techniques to identify patients at risk of readmission. However, the body of work in this domain varies significantly in terms of clinical population, modeling approach, and predictive accuracy.

This chapter is primarily adapted from our published review article, "A comprehensive review of ICU readmission prediction models : From statistical methods to deep learning approaches", in Artificial Intelligence in Medicine (Vol. 165, 2025) [21]. To systematically review the current state of the art and identify key research gaps, the literature in this chapter is categorized into six main areas :

1. Studies that have identified clinical risk factors associated with ICU/PICU readmission, primarily through retrospective analysis.
2. Studies that have developed predictive models for ICU/PICU readmission, leveraging statistical and machine learning approaches.
3. Identified gaps in the current literature and limitations of existing approaches.
4. Challenges specific to pediatric ICU readmission prediction.
5. Research questions and objectives derived from observed gaps.
6. The contribution of the proposed modeling strategy.

## 2.1 Risk Factors Associated With ICU/PICU Readmission

A broad spectrum of studies has examined the determinants and predictors associated with unplanned ICU readmissions, encompassing patient-level, clinical, physiological, and organizational domains. This section synthesizes evidence from existing literature, categorizing risk factors thematically to enhance clarity.

Severity of illness during the initial ICU stay is consistently reported as one of the strongest predictors of readmission. Multiple scoring systems—including Acute Physiology and Chronic Health Evaluation (APACHE) [27], Sequential Organ Failure Assessment (SOFA) [28], Simplified Acute Physiology Score (SAPS) II/III [29], and the Acute Physiology Score (APS) [30] have been shown to correlate with increased risk of readmission [9, 10, 31–35]. Patients with

high scores at ICU discharge are often physiologically unstable, and their early transfer may lead to deterioration in the general ward. Additionally, specialized discharge scores such as the Stability and Workload Index for Transfer (SWIFT) [36], National Early Warning Score (NEWS) [37], and Modified Early Warning Score (MEWS) [38] have been validated as strong predictors of ICU readmission [22, 39–41]. These scores capture subclinical signs of instability that may not be evident during routine assessment. Patients discharged with unresolved medical issues or requiring extensive care (reflected in high Nursing Activity Scores) face elevated risk of readmission [42]. These findings support the hypothesis that some ICU discharges occur prematurely—before the patient's condition has sufficiently stabilized.

The presence of chronic comorbid conditions consistently correlates with increased ICU readmission risk. Conditions frequently associated with higher risk include congestive heart failure, chronic kidney disease, Chronic Obstructive Pulmonary Disease (COPD), cancer, and liver cirrhosis [9, 10, 32, 35, 43, 44]. Patients with complex health histories require specialized post-ICU care, and their vulnerability to complications increases readmission likelihood.

Neurological deficits stemming from intracerebral hemorrhage, sepsis, or dysphagia independently predict readmission risk, particularly among neurological or surgical ICU cohorts [45–47]. ICU treatments indicating critical instability, such as vasopressor therapy, mechanical ventilation, renal replacement therapy, or surgical interventions, significantly elevate readmission risk [10, 43, 44, 47, 48]. Persistent organ dysfunction post-intervention highlights ongoing vulnerability despite clinical improvements at discharge.

Nighttime or after-hours discharge from the ICU is one of the most frequently cited non-clinical risk factors for readmission. Studies attribute this to reduced staffing, less experienced personnel, and inadequate handovers during these periods [48–54]. Several multicenter studies have reported significantly increased mortality and readmission rates associated with nighttime discharge. Although some studies reported that nighttime discharge was not statistically significant after multivariate adjustment, the trend remains consistent across cohorts [47, 52].

Further, unplanned or rushed discharges, often influenced by bed availability or ICU overcrowding, have been shown to increase readmission and early mortality rates [34, 55, 56]. Institutional discharge practices—including availability of step-down units, structured communication protocols, and formal ICU transition programs—yield mixed evidence regarding their effectiveness in reducing readmissions. Some studies advocate structured handovers and checklists, while others find minimal or no impact [54, 57–60]. This variability highlights the complexity of institutional influences on readmission.

Although the risk factors for unplanned readmission to PICUs overlap with those identified in adult ICUs, they also reflect the unique vulnerabilities of pediatric populations. Based on

a synthesis of findings from numerous studies, risk factors can be categorized as follows :

The most consistent and strongest predictor of PICU readmission is the presence of Complex Chronic Conditions (CCCs). Patients with multiple CCCs, medical devices or nutritional interventions dependence (e.g., tracheostomy, VP shunt, tube feeding), or severe baseline disability face significantly elevated risk [61–64]. Elevated discharge Pediatric Early Warning Systems (PEWS) [65] and its variants strongly predict readmission by capturing vital signs, responsiveness, and oxygen needs [63, 66, 67].

Abnormal clinical parameters at discharge—such as elevated Respiratory Rate (RR), Heart Rate (HR), and low Glasgow Coma Scale (GCS)—are key modifiable risk factors. Oxygen requirement at discharge is another strong predictor of clinical instability. Collectively, these signs often indicate premature discharge, and addressing them could prevent avoidable readmissions [63, 66, 68]. Low body weight and young age are consistently reported as non-modifiable risk factors. These indicators reflect overall physiological vulnerability [67, 69–71].

Longer PICU stays and admissions from general wards or emergency departments correlate with higher readmission risk due to increased severity and potential delayed ICU transfers [62, 68, 72–75]. Evening or weekend discharges yield mixed results, with some pediatric studies reporting insignificant associations after adjustments [62, 71].

Trauma or surgical admissions, direct home-to-PICU admissions, operative interventions and Extracorporeal Membrane Oxygenation (ECMO) appear to reduce readmission risk, potentially reflecting definitive treatments or enhanced care transitions [61, 62, 75].

Across ICU and PICU literature, primary factors consistently associated with readmission include :

1. Severity of illness and clinical instability at discharge.
2. Presence of chronic or complex medical conditions.
3. ICU interventions reflecting critical instability.
4. Unresolved clinical issues or premature discharge practices.
5. Operational and institutional practices (staffing, discharge timing, structured transitions).

Identifying these modifiable and non-modifiable risk factors provides valuable insights for targeted interventions aimed at reducing unplanned ICU and PICU readmissions.

## 2.2 Predictive Modeling for ICU and PICU Readmission

Recent advances in ML and DL offer promising avenues for developing predictive models capable of analyzing high-dimensional clinical data offered by the EHR databases. These models can process diverse variables such as vital signs, laboratory results, and clinical histories to uncover complex patterns associated with adverse outcomes.

### 2.2.1 Database

#### Overview of used databases

Ensuring replicable methodology is essential in data science and typically requires data sharing. However, the medical and clinical fields face ethical limitations due to the sensitivity and confidentiality of patient data. Balancing these ethical concerns with the need for reproducibility highlights the importance of open-access datasets in medical and clinical research.

The outcomes analysis of the literature review reveals a predominant reliance on publicly available databases, notably Multiparameter Intelligent Monitoring in Intensive Care (MIMIC). The MIMIC databases and the eICU Collaborative Research Database (eICU) are key public datasets for critical care research. These datasets serve as invaluable resources for advancing critical care research [76, 77]. MIMIC provides comprehensive clinical data, including physiological waveforms, demographics, and laboratory results. Three versions of the MIMIC series have been used : MIMIC-II (2001-2008) [78–81], MIMIC-III (2001-2012) [82–85], and MIMIC-IV (up to 2019) [86–88]. These versions progressively improve data quality and expand critical care research insights. The eICU database, sourced from various institutions, offers detailed ICU patient information, spanning demographics, vital signs, laboratory results, and outcomes, ensuring a diverse representation. The eICU database has been used in several studies [89–92].

In addition, various private databases have been utilized in studies, including Scottish ICU database [93], Seoul National University Bundang Hospita [94], Liverpool hospital in Australia [95], French Outcomerea network [96], Hospital da Luz in Portugal [97], Centro Hospitalar do Porto [98, 99], Anhui hospital [100], University of Chicago Medical Center [101], Stanford Medicine Research Data Repository (STARR) [86], Children's Healthcare of Atlanta (CHOA) [81], Brazilian university hospital [102], Cambridge University Hospital (CUH) [103], Amsterdam UMC [104, 105], Leiden university medical centre [105], a tertiary care hospital in the UK [103], an academic hospital in the United States [101], a Brazilian university hospital [102], Anesthesiology, Intensive Care and Pain Medicine at the University Hospital Münster (ANIT-UKM) [87], and the University of Florida Institutional [106].

**Used data types**

EHR databases store a wide array of patient information, including demographic details, vital signs, laboratory results, interventions, medical history, diagnosis, medications, procedures, treatment, imaging data, and clinical notes. These data types can be classified into five main groups :

1. Demographic Data : Include information like age, gender, and comorbidity, providing baseline characteristics for patients.
2. Temporal data : Comprise vital signs, interventions, laboratory tests, and time series data of treatments, offering dynamic health tracking over time.
3. Medical Codes : Represent diagnosis, procedures, and medication codes, providing insight into the medical history and treatment of patients.
4. Text Data : Include clinical notes, particularly discharge summaries, and other textual information, providing qualitative information on patient care.
5. Images : Comprise diagnostic images such as X-rays and Magnetic Resonance Imaging (MRI), providing visual information for diagnostics.

Each data type plays a crucial role in understanding the health status and history of a patient. Traditionally, ICU readmission prediction predominantly relied on demographics, temporal data, and comorbidities. Clinical notes are increasingly used to supplement these sources.

Current trends favor a comprehensive approach, utilizing various data modalities, including medical codes, for a holistic view of patient health and readmission risks. Interestingly, medical images remain largely unexplored for ICU readmission prediction, suggesting a potential area for future research. Effective representation of these data is crucial for accurate prediction models in healthcare.

The researchers selected variables for their models based on various approaches. Some chose variables according to previous medical knowledge, focusing on factors known to be relevant in predicting ICU readmission or following previous studies or guidelines, incorporating variables that are significant in similar research [86,95,104,107]. Additionally, some studies included all allowed variables with low missing rates and presence for the majority of patients, aiming to capture a comprehensive range of factors that could impact patient outcomes [87,93,108,109].

### 2.2.2 Feature representation techniques

Effective feature representation is paramount in predicting ICU readmission due to the rich information in EHR data. Most statistical and ML models for predicting ICU readmission

used primarily demographic data and temporal data, with some studies also exploring the use of medical codes and clinical notes. To use temporal data, statistical features were extracted using techniques such as descriptive statistics and entropy [80, 86, 97, 102, 104, 110–114]. These features fail to capture detailed temporal trends. Techniques like frequent subgraph mining were used to capture temporal trends [78, 101] but it overlooks critical patterns in understanding physiological measurement dynamics over time. To overcome this, some studies converted temporal data into qualitative variables [26, 96, 108, 115–117]. However, these methods simplify complex temporal patterns into discrete categories, neglecting time embedding in a vector space. Medical codes, especially diagnoses, were simplified in several studies using scores such as APACHE and SOFA scores [93, 95, 96, 102] but lacked the rich context and specificity required for accurate prediction. Other studies focused on using pretrained embeddings based on International Classification of Diseases (ICD) codes such as Choi [118] and the Clinical Classification Software (CCS) [119] or clustering similar diagnoses [26, 80, 89, 109, 113, 114, 120]. Treatments were represented by several features such as usage duration [81, 86, 87, 104]. Text data were represented in several ways based on Bag-Of-Words (BOW) and sting matching [100, 107, 121]. Despite some effectiveness, these methods struggled with the complexity of textual data.

These techniques do not capture the potential of these data. DL models offer automated feature learning, integrating these data to capture complex patterns and dependencies. Learning representations using Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), and Neural Ordinary Differential Equations (ODEs) was used in several studies [113, 117, 120, 122]. However, these representations face challenges due to sparse data, different modalities, variable lengths of hospital stay, and irregular time intervals. To address this, others proposed embedding time-related information to code embeddings using time-aware attention or exponential time-decay functions [106, 117].

Natural Language Processing (NLP) techniques are crucial for enhancing information extracted from clinical notes. Word embedding techniques such as Word2Vec [123] and BioWordVec [124] were used to capture semantic meanings in a continuous space, grouping similar words closely [125]. With the introduction of transformer-based Large Language Models (LLMs), especially Bidirectional Encoder Representations from Transformers (BERT) [126], the representation of textual data has undergone significant improvement as it can capture bidirectional contextual information. BioBERT [127] and ClinicalBERT [128] are both specialized versions of BERT, designed for processing biomedical and clinical text, respectively [109, 116]. However, using these embedding techniques is ineffective due to the lengthy and noisy content of clinical notes. In addition, they overlook the graphical structure mirroring the physician's decision and lack interpretability, potentially hindering performance.

Recent approaches focus on effectively extracting entities from clinical notes and medical codes, and embedding them with meaningful medical representations. One approach embedded ICD-9 codes into a hyperbolic space [129] using Poincaré embeddings to represent hierarchical structures [130]. Other approaches integrated external knowledge from medical ontologies or domain-specific databases such as the Unified Medical Language System (UMLS) [131] and the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) [132], or created a graph representation of patient encounters, or a combination of both approaches [82,115,133]. Enriching EHR data with external knowledge and generating graph embeddings improves DL performance and helps ML models better utilize clinical data.

Graphs are a flexible framework for incorporating data from multiple modalities. Most graph-based models have focused on representing either medical codes [91, 134–137] or clinical notes [138]. However, exploring optimal fusion strategies for different modalities has emerged as a promising research direction in patient representation learning [84, 85, 92, 139]. Variational regularization was proposed to handle irregular time intervals between patient encounters [137]. Graphs are limited by fuzzy patient data relevance and struggle for higher performance due to noise from diverse disease types. Leveraging external medical knowledge to integrate domain-specific information enhances node representation and facilitates a better understanding of the relationships among different entities [85, 92, 136, 138, 139]. Using contrastive learning techniques [140] to refine graph representations by generating pairs of nodes to ensure consistency, effectively bringing embeddings of patients with the same label closer in the embedding space [91,92]. These methods enhance the ability to model intricate relationships within EHR data, leading to improved performance and clinical insights.

### 2.2.3 Model types

This section reviews statistical, machine learning, and deep learning models used to predict ICU readmission.

#### Statistical approaches

Table 2.1 identifies 21 studies, that used various statistical approaches, including Logistic Regression (LR), fuzzy clustering, Dynamic Linear Models (DLMs), and Conditional Random Fields(CRF), have been employed to develop ICU readmission prediction models.

LR is a widely used statistical model for binary classification problems. In this context, LR estimates the probability that a given observation belongs to the positive class (e.g., readmission) as a function of input features. Formally, given an input vector $\mathbf{x} \in \mathbb{R}^d$, the

model computes the probability $p(y = 1|\mathbf{x})$ using the logistic sigmoid function applied to a linear combination of the predictors :

$$p(y = 1|\mathbf{x}) = \sigma(z) = \frac{1}{1 + e^{-z}} \tag{2.1}$$

where

$$z = \mathbf{w}^\top \mathbf{x} + b. \tag{2.2}$$

Here, $\mathbf{w}$ is the weight vector and $b$ is the bias term. The model is trained by maximizing the likelihood of the observed labels under the Bernoulli distribution, which is equivalent to minimizing the binary cross-entropy loss :

$$\mathcal{L}(\mathbf{w}, b) = -\frac{1}{N} \sum_{i=1}^{N} \left[ y^{(i)} \log p^{(i)} + \left(1 - y^{(i)}\right) \log\left(1 - p^{(i)}\right) \right], \tag{2.3}$$

where $y^{(i)} \in \{0, 1\}$ is the true label for the $i$-th example, and $p^{(i)} = \sigma(\mathbf{w}^\top \mathbf{x}^{(i)} + b)$. To prevent overfitting, especially when the number of features is large, a regularization term can be added. In the case of L2 regularization (Ridge), the loss becomes :

$$\mathcal{L}_{\text{reg}} = \mathcal{L}(\mathbf{w}, b) + \lambda \|\mathbf{w}\|_2^2, \tag{2.4}$$

where

$$\|\mathbf{w}\|_2^2 = \sum_j w_j^2.$$

and $\lambda$ controls the strength of regularization. LR is popular in clinical prediction tasks due to its simplicity, interpretability, and ability to estimate well-calibrated probabilities.

Ridge classifier is another statistical model that fits a linear decision function $s(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ after mapping labels to $y_i \in \{-1, +1\}$, and predicts with $\text{sign}\big(s(\mathbf{x})\big)$. It minimizes the ridge–penalized squared-error objective

$$\mathcal{J}_{\text{Ridge}}(\mathbf{w}, b) = \frac{1}{n} \sum_{i=1}^{n} \big(y_i - (\mathbf{w}^\top \mathbf{x}_i + b)\big)^2 \; + \; \lambda \|\mathbf{w}\|_2^2. \tag{2.5}$$

This quadratic objective makes it typically faster to train than the LR models that class probability via the sigmoid as shown in equation (2.1), with $y_i \in \{0, 1\}$, and minimizes the L2–regularized binary log-loss as shown in equation (2.3). But LR loss is better aligned with classification and yields calibrated probabilities, which often translates into better classifica-

tion performance than the Ridge classifier.

Campbell, *et al.* [93] initially achieved moderate Area Under the Receiver Operating Characteristic (AUROC) curve of 62% and 67% for predicting readmissions using LR. In [141], they proposed a scoring tool by normalizing LR coefficients achieving an AUROC of 76%. Subsequent improvements included feature selection techniques, with Jo, *et al.* [94] achieving an AUROC of 76% through the Likelihood Ratio-Test (LR-Test), and Frost, *et al.* [95] reaching 66% with backward-deletion. To further enhance performance, different feature extraction techniques were employed. Badawi, *et al.* [89] grouped diagnoses into 26 categories, extracted statistical features, used Multiple Imputation by Chained Equations (MICE) [142] for missing data, and selected significant features with LR-test and stepwise LR, achieving an AUROC of 71%. Ouanes, *et al.* [96] transformed temporal variables into qualitative ones and identified significant predictors through multivariate analysis, resulting in an improved AUROC of 74%. Xue, *et al.* [78] achieved an AUROC of 66% using the adapted Subgraph Augmented Non-negative Matrix Factorization (SANMF) but with limited impact on overall trend correlation. Utilizing clinical notes, in [82], they reached an AUROC of 75% using UMLS to identify medical concepts and assigning unique Concept Unique Identifiers (CUIs), generating a Bag-of-CUIs, while in [83], they attained 76% with BOW embedding. The performance of LR-based models was moderate, with AUROC values ranging from 62% to 76%, and an average AUROC of 70%.

Fuzzy clustering found application in multiple studies. It allows data points to belong to multiple clusters simultaneously. The Takagi-Sugeno Fuzzy Model (TSFM) is a fuzzy inference system that uses linear functions associated with fuzzy sets to represent rules, activated based on input values, producing a weighted average of outputs to model complex systems efficiently [143]. Techniques such as Fuzzy C-Means (FCM) [144], Mixed Fuzzy Clustering (MFC) [145], and Probabilistic Fuzzy Systems (PFS) [146] were used to determine the antecedent fuzzy sets and the number of rules of TSFM. FCM optimizes the sum of squared differences for numerical data, MFC enhances FCM for spatiotemporal data, while PFS merges fuzzy logic with probabilistic reasoning, integrating linguistic system descriptions with statistical data properties.

Fialho *et al.* [110] achieved an AUROC of 72% using a TSFM model based on FCM clustering and six variables, and an AUROC of 66% with a TSFM model based on PFS with FCM and the nearest neighbor heuristic for membership functions [97]. Vieira, e*et al.* [147] suggested improving this by merging numerical and medical text annotations, and Curto, *et al.* [107] further advanced this with a Fuzzy FingerPrint (FFP) classifier [148] and Pareto-inspired membership function [149], achieving an AUROC of 80% with text data. In [111], temporal

data were represented using Shannon entropy and weighted average, with feature selection achieved through Binary Fish School Search (BFSS) [150] combined with FCM, resulting in an AUROC of 69%. In [151], three TSFM model approaches were used : FCM, modified MFC for fixed-length multivariate time series, and MFC-FCM with feature transformation, with MFC and FCM outperforming MFC-FCM and achieving an AUROC of 58%. A follow-up [79] extended MFC to handle unequal-length multivariate time series, using FCM for transformation and MFC grouping (FCM-$U^{MFC}$), achieving an AUROC of 64%.

Fernandes, *et al.* [152] suggested using ensemble learning with FCM-based clustering to generate patient clusters for TSFM models development. They employed classifier selection techniques including a priori decisions based on cluster center distances to patient characteristics and a posteriori decisions to select outcomes with lower uncertainty. The posteriori decision outperformed the a priori approach, resulting in an AUROC of 75%. A follow-up [153] found no significant difference between aggregation techniques and classifier selection. Viegas *et al.* [112] enhanced this approach by using Sequential Forward Selection (SFS) for feature selection and Gustafson-Kessel (GK) clustering [154], combining sensitivity and specificity models with weights based on uncertainty, achieving an AUROC of 77%. Overall, fuzzy-based models performed similarly to LR-based models, with AUROCs ranging from 58% to 81% and an average of 71%.

DLMs are valuable for modeling time series data, incorporating both system dynamics and observation uncertainty. They model time series by estimating the state vector from observed data, using the Kalman filter to update the estimate with new observations [155]. Caballero, *et al.* [115] used DLMs with Bayesian time series to capture temporal dependencies and update readmission predictions, extracting features focusing on clinically named entities using UMLS and statistical topic modeling, achieving an AUROC of 93%.

CRFs are probabilistic models used for labeling sequential data. They model the conditional probability of labels given observations, capturing complex dependencies [156]. Venugopalan, *et al.* [108] used k-means and FCM for imputing temporal data, employed CRF for capturing time-varying dynamics, and applied LR and Neural Network (NN) for static data classification. CRF with FCM-based imputation achieved 90% accuracy, but models combining with k-means imputation surpassed CRF alone (80%), reaching 87% accuracy.

Table 2.1 summarizes statistical approaches, revealing AUROC scores ranging widely from 58% to 93%. Despite high AUROCs, the models' reliability is limited due to the small patient sample size.

TABLE 2.1 Statistical approaches for ICU readmission prediction (same* : same hospitalization, ** ● : used, ● : unused).

| Study | Database Name (Sample Size) | Time Period (days) Readmission Rate % | Data Type | | | | Classifier | Performance | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Demographic | Temporal | Clinical Notes | Medical Codes | | AUROC (%) | Accuracy (%) |
| Campbell 2008 [93] | Private (6,208) | same* ( 9)<br>2 (3) | ●** | ● | ● | ● | LR | 62<br>67 | - |
| Haribhakti 2021 [141] | Private (883) | same (9) | ● | ● | ● | ● | LR | 76 | - |
| Jo 2015 [94] | Private (343) | same (10) | ● | ● | ● | ● | LR | 76 | - |
| Frost 2010 [95] | Private (14,952) | same (7) | ● | ● | ● | ● | LR | 66 | - |
| Badawi 2012 [89] | eICU (704,963) | 2 (3) | ● | ● | ● | ● | LR | 71 | - |
| Ouanes 2012 [96] | Private (3,462) | 7 (2) | ● | ● | ● | ● | LR | 74 | - |
| Xue 2019 [78] | MIMIC-II (1,170) | 30 (27) | ● | ● | ● | ● | LR | 66 | - |
| Li 2019 [82] | MIMIC-III (45,305) | 30 (5) | ● | ● | ● | ● | LR | 75 | - |
| Moerschbacher 2023 [83] | MIMIC-III (4,522) | 30 (50) | ●<br>● | ●<br>● | ●<br>● | ●<br>● | LR<br>RF | 76<br>70 | 69<br>71 |
| Fialho 2012 [110] | MIMIC-II (1,028) | 3 (13) | ● | ● | ● | ● | FCM | 72 | 71 |
| Fialho 2013 [97] | Private (3,271) | 3 | ● | ● | ● | ● | PFS | 66 | 67 |
| Vieira 2013 [147] | MIMIC-II (1,028) | 3 (13) | ●<br>● | ●<br>● | ●<br>● | ●<br>● | TSFM<br>FFP | 72<br>- | -<br>- |
| Curto 2016 [107] | MIMIC-II (12,091) | 3 (6) | ●<br>● | ●<br>● | ●<br>● | ●<br>● | FFP<br>TSFM | 80<br>64 | 84<br>69 |
| Sargo 2014 [111] | MIMIC-II (726) | 3 (12) | ● | ● | ● | ● | FCM | 69 | 56 |
| Ferreira 2015 [151] | MIMIC-II (2,653) | 3 (8) | ● | ● | ● | ● | MFC | 58 | 59 |
| Salgado 2016 [79] | MIMIC-II (1,389) | 3 (10) | ● | ● | ● | ● | FCM-$U^{MFC}$ | 64 | 57 |
| Fernandes 2014 [152] | MIMIC-II (1,010) | 3 (13) | ● | ● | ● | ● | FCM posteriori | 75 | 70 |
| Salgado 2015 [153] | MIMIC-II (1,010) | 3 (13) | ● | ● | ● | ● | FCM posteriori | 81 | 72 |
| Viegas 2017 [112] | MIMIC-II (1,499) | 3 (7) | ● | ● | ● | ● | GK | 77 | 71 |
| Caballero 2015 [115] | MIMIC-II (11,648) | 30 | ● | ● | ● | ● | DLMs | 93 | - |
| Venugopalan 2017 [108] | MIMIC-II (32,331) | 30 (24) | ● | ● | ● | ● | CRF | MCC : 73 | 90 |

**Machine learning approaches**

Table 2.2 highlights 22 studies using various machine learning models, primarily supervised approaches like Decision Trees (DT), with some exploring unsupervised methods like K-means Clustering.

NN learns complex patterns through interconnected nodes using feedforward flow, activation functions, and back-propagation [157]. In [158], NN achieved AUROCs of 87% and 79% on complete and sub-sampled datasets, respectively, showing that the dataset's histogram had an improper distribution for probability models and no specific feature selection method was recommended. Junqueira, *et al.* [159] used Symmetrical Uncertainty (SU) [160] for feature selection, finding NN performed best with a 64% AUROC on a temporally split, sub-sampled dataset, showing consistent risk factors over time. In [161], four ML models performed well, with the NN model achieving the best overall performance with an F1-score of 84%. The performance of NN-based models was moderate. Naive Bayes (NB) is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions between the features [162]. In [98], NB outperformed other ML models on an oversampled database, achieving an accuracy of 99%. Oversampling by duplicating minority class data without proper separation between training and test sets leads to biased metrics and inflated specificity scores, with the model showing a significantly lower specificity of 72% without oversampling.

Support vector machines (SVM) are supervised learning models designed to find the optimal decision boundary separating two classes. In the case of a linear SVM for binary classification, the model seeks a hyperplane that maximizes the margin between positive and negative examples. Given a training dataset $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$ with $\mathbf{x}^{(i)} \in \mathbb{R}^d$ and labels $y^{(i)} \in \{-1, +1\}$, the linear decision function is defined as [163] :

$$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b, \tag{2.6}$$

where $\mathbf{w}$ is the normal vector to the hyperplane and $b$ is the bias term. The geometric margin is given by $2/\|\mathbf{w}\|$, and the optimal hyperplane is the one that maximizes this margin while correctly classifying the training samples. This leads to the following convex optimization problem :

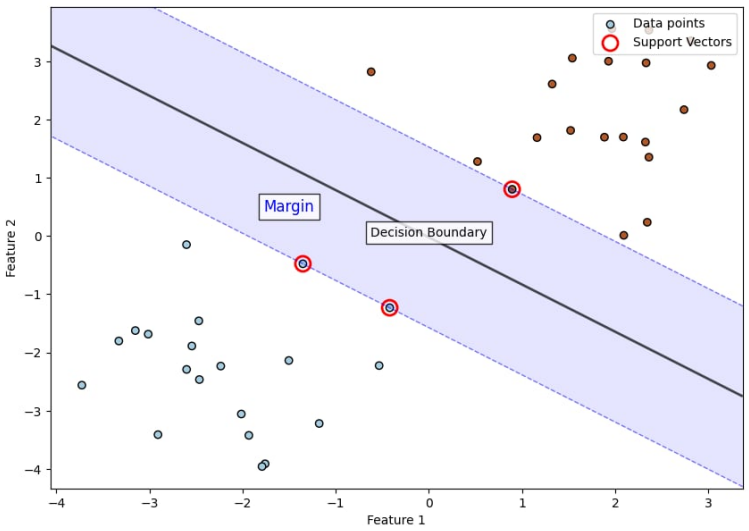



FIGURE 2.1 SVM decision boundary and margin

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|^2 \tag{2.7}$$

$$\text{subject to} \quad y^{(i)}\left(\mathbf{w}^\top \mathbf{x}^{(i)} + b\right) \geq 1, \quad \forall i. \tag{2.8}$$

In practice, many problems are not perfectly separable, so slack variables $\xi_i$ are introduced to allow margin violations. This yields the soft-margin SVM :

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \tag{2.9}$$

$$\text{subject to} \quad y^{(i)}\left(\mathbf{w}^\top \mathbf{x}^{(i)} + b\right) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i, \tag{2.10}$$

where $C > 0$ controls the trade-off between maximizing the margin and penalizing classification errors. For non-linear decision boundaries, SVMs can employ the kernel trick to implicitly map input vectors into a higher-dimensional space $\mathcal{H}$ where a linear separation becomes possible. The decision function in the kernelized SVM is given by :

$$f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i y^{(i)} K(\mathbf{x}^{(i)}, \mathbf{x}) + b, \tag{2.11}$$

where $K(\mathbf{x}^{(i)}, \mathbf{x})$ is a positive semi-definite kernel function (e.g., Radial Basis Function (RBF)),

and $\alpha_i$ are the learned Lagrange multipliers associated with each training sample. SVMs are particularly well-suited for high-dimensional spaces and often deliver strong predictive performance in settings where the number of features exceeds the number of observations.

Figure 2.1 shows the decision boundary (solid line) that maximizes the margin (shaded region) between the two classes. The highlighted support vectors (circled points) are the critical samples that determine the position and orientation of the separating hyperplane. Negar, *et al.* [121] used BOW to create a document-term matrix from clinical notes, achieving an AUROC of 71% and 74% with feature selection using SVM, outperforming other ML models on a sub-sampled database.

DT is a tree-like model that uses a flowchart-like structure of decisions and their possible consequences to make predictions or decisions [164]. Random Forest (RF) improves DT by addressing overfitting through ensemble learning. DT is interpretable, while RF excels with larger datasets. RFs are ensemble learning methods that combine the predictions of multiple decision trees to improve classification accuracy and reduce overfitting. A RF classifier consists of an ensemble of $T$ decision trees, each trained on a bootstrap sample of the training data. Formally, let $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$ denote the dataset. For each tree $t$, a bootstrap sample $\mathcal{D}_t$ is generated by randomly sampling $N$ examples with replacement. During tree construction, at each split node, only a random subset of $m$ features (with $m \leq d$) is considered to determine the optimal split. This process injects additional randomness and de-correlates individual trees [165].

Each decision tree $h_t(\mathbf{x})$ outputs a predicted class label. The final RF prediction $\hat{y}$ is determined by majority voting across all trees :

$$\hat{y} = \text{mode}\Big\{h_t(\mathbf{x})\Big\}_{t=1}^{T}. \tag{2.12}$$

Alternatively, if probabilistic predictions are desired, the estimated probability of the positive class is computed as the fraction of trees predicting the positive label :

$$p(y = 1|\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T}\mathbb{I}\Big[h_t(\mathbf{x}) = 1\Big], \tag{2.13}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function. The randomness introduced by both bootstrapping and feature subsampling helps reduce variance and improve generalization compared to a single DT. RF also provides measures of variable importance, often computed by assessing the mean decrease in impurity or by evaluating the impact of permuting each feature on prediction accuracy.

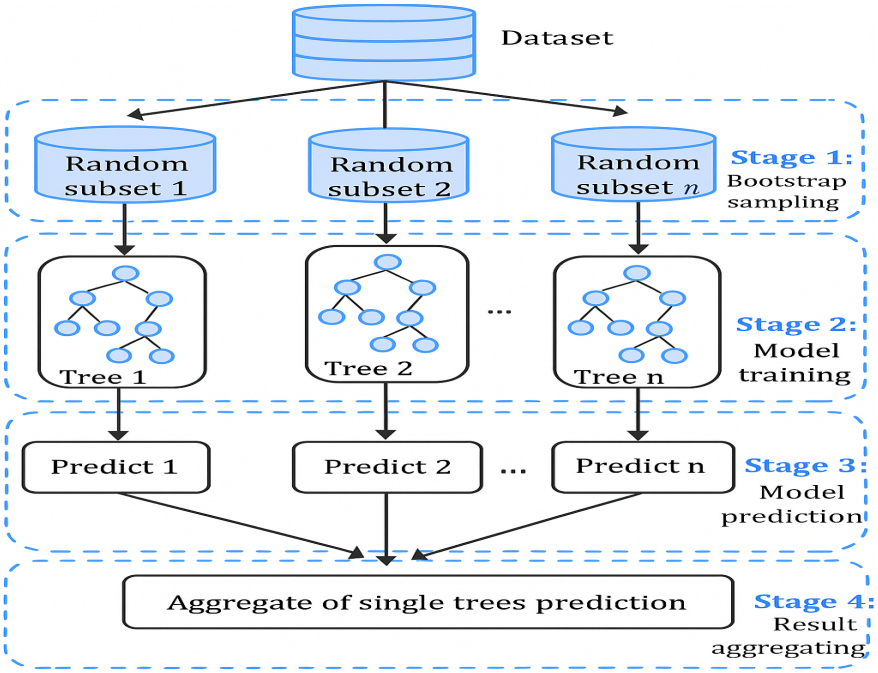



FIGURE 2.2 Random Forest training and prediction process

Figure 2.2 illustrates how RF builds multiple DTs on bootstrapped subsets of the data (Stage 1–2), generates predictions from each tree (Stage 3), and aggregates them into a final prediction (Stage 4). This ensemble approach improves accuracy and robustness compared to a single DT.

Extremely Randomized Trees (Extra-Trees) is an ensemble of DTs that injects more randomness than a standard RF. Like RFs, it builds many trees and aggregates them (majority vote for classification, mean for regression). However, There are two main differences :

1. Bootstrapping : RFs typically train each tree on a bootstrap sample (sampling with replacement), whereas Extra Trees, by default, use the entire training set for every tree (no bootstrap)—though this can be toggled. However, at each split, they sample a subset of features like RF.
2. Split selection : RFs search (nearly) optimal thresholds for each candidate feature at a node. Extra Trees instead pick random thresholds : at each split, they draw a random cut-point for each of those features (often uniformly between the feature's min and max within the node), evaluate the impurity gain for those random splits, and select the best among them. For numerical features, the range is between the minimum and

maximum values of that feature in the node. By default, it randomly generates only one split point in this range (unlike RF which considers all possible splits). For binary features, since these are already binary (0/1), there's only one possible split point. Even though it's random, there's no randomness here since it will lead to the same split.

This extreme randomization makes Extra Trees faster and typically lower-variance and higher-bias, which can help on noisy data and reduce overfitting.

In [166], a DT algorithm based on Axiomatic Fuzzy Set theory (AFS-DT) [167] achieved an AUROC of 61% using Cohen's kappa coefficient as a fitness function, with similar performance (60% AUROC) when trained on only six variables recommended in [110], indicating no significant loss of information. In [80], a proposed Noise Reduction Learning (NRL) system tackled data sparsity by under-sampling overlapping points with k-Nearest Neighbors (KNN). Ensemble learning using DT and LR models with L1 regularization achieved an AUROC of 81%. In [100], a weight decay RF model was utilized for both imputation and dataset rebalancing, leveraging features from temporal data and clinical notes to achieve an AUROC of 88%, surpassing other machine learning models. Alghatani,*et al.* [168] used ICU data from the first day and applied the percent point function to determine quantile thresholds. SVM achieved an AUROC of 59% across the entire dataset, while RF reached a best AUROC of 74% for the sub-sampled dataset.

In [90], the patient forest model was introduced, leveraging an ensemble of DTs within the gcForest framework. This approach utilized multi-grained scanning and cascade forest modules to process EHR encounters and extract features at various granularities, which were then refined through multiple levels of RFs. The model, trained with optimized convolutional filter and RF parameters via backpropagation to minimize binary cross-entropy loss, achieved an AUROC of 87%. Its superior performance was attributed to its ability to capture the heterogeneity and complexity of patient data through patient-specific DTs. The aggregation of predictions from multiple trees also reduced variance and enhanced result stability compared to models using a single global classifier.

Boosting techniques like Adaptive Boost (AdaBoost), Gradient Boosting (GBM), Extreme Gradient Boosting (XGBoost), and LightGBM enhance predictive performance by emphasizing misclassified instances sequentially. AdaBoost is effective and less prone to overfitting but sensitive to noise [169]. GBM ensures high accuracy but is computationally intensive [170]. XGBoost improves Gradient Boosting with regularization and parallel processing, enhancing computational efficiency [171]. It is an efficient, scalable implementation of gradient boosted decision trees. Unlike RF, which build trees independently and aggregate their predictions,

boosting constructs trees sequentially, where each tree attempts to correct the errors made by the ensemble so far.

XGBoost is a powerful model that improves performance and reduces overfitting. Formally, let $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$ denote the training data. The model predicts the output $\hat{y}^{(i)}$ by summing the outputs of $K$ regression trees :

$$\hat{y}^{(i)} = \sum_{k=1}^{K} f_k(\mathbf{x}^{(i)}), \quad f_k \in \mathcal{F}, \tag{2.14}$$

where $\mathcal{F}$ denotes the space of regression trees. The training objective includes both the loss function measuring model fit and a regularization term penalizing model complexity :

$$\mathcal{L} = \sum_{i=1}^{N} l\Big(y^{(i)}, \hat{y}^{(i)}\Big) + \sum_{k=1}^{K} \Omega(f_k), \tag{2.15}$$

where $l$ is a differentiable convex loss function (e.g., logistic loss), and $\Omega(f)$ is the regularization term defined as :

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2. \tag{2.16}$$

Here, $T$ denotes the number of terminal nodes (leaves) in the tree, and $w_j$ represents the weight of the $j$-th leaf. The parameters $\gamma$, which is the pruning penalty, and $\lambda$, which is the regularization coefficient, control the complexity penalty and help prevent overfitting.

At iteration $t$, the algorithm adds a new tree $f_t$ to minimize the objective function. Using a second-order Taylor expansion of the loss, the approximate objective becomes :

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{N} \Big[ g_i f_t(\mathbf{x}^{(i)}) + \frac{1}{2} h_i f_t^2(\mathbf{x}^{(i)}) \Big] + \Omega(f_t), \tag{2.17}$$

where

$$g_i = \frac{\partial l\Big(y^{(i)}, \hat{y}^{(i)}\Big)}{\partial \hat{y}^{(i)}}, \quad h_i = \frac{\partial^2 l\Big(y^{(i)}, \hat{y}^{(i)}\Big)}{\partial \hat{y}^{(i)^2}}.$$

These first and second derivatives (gradients and Hessians) are computed from the current predictions and drive the optimization of each new tree. The optimal leaf weights $w_j^*$ for each leaf $j$ can be derived in closed form :

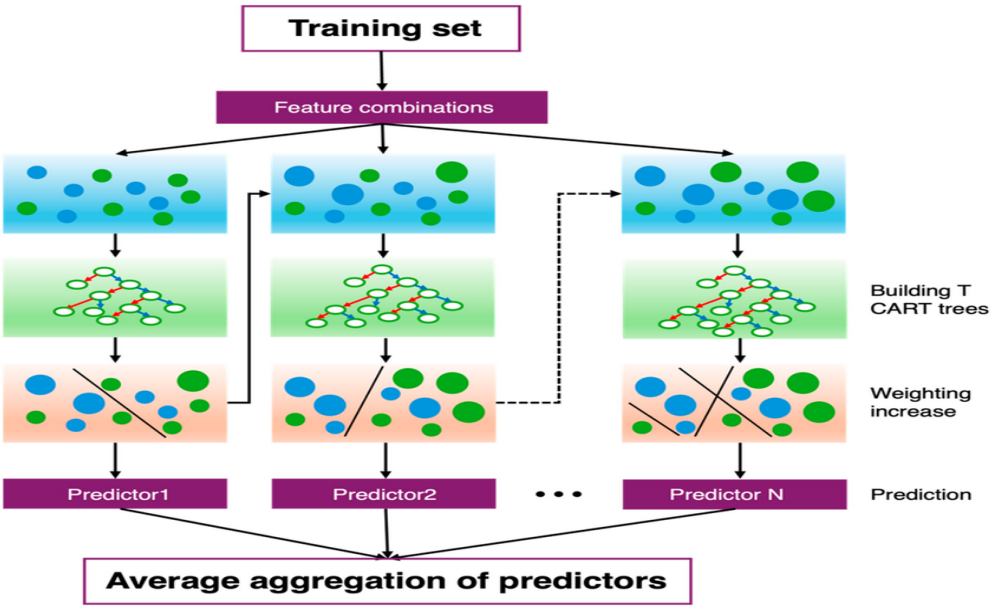



FIGURE 2.3 Flowchart of the XGBoost training process

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}, \tag{2.18}$$

where $I_j$ denotes the set of samples in leaf $j$. The corresponding optimal value of the objective reduction is :

$$\mathcal{L}_{\text{leaf}} = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T. \tag{2.19}$$

By iteratively adding trees in this manner, XGBoost produces a powerful ensemble capable of capturing complex non-linear relationships while maintaining high computational efficiency through optimized data structures and parallelization.

Figure 2.3 illustrates the XGBoost workflow, starting from iteratively reducing residuals until an optimal model is trained for prediction [172].

LightGBM is a gradient boosting framework that constructs an ensemble of DTs trained sequentially to minimize a differentiable loss function. It employs histogram-based learning for efficient processing of large datasets, although it may necessitate careful parameter tuning [173]. Like other boosting methods, it builds additive models (2.14) and minimizes the regularized objective (2.15) using the regularization term that penalizes complexity (2.16).

One key challenge in boosting algorithms is the computational cost of finding the best splitting points to grow each tree. To address this, LightGBM uses a histogram-based algorithm to discretize continuous feature values into a fixed number of bins, reducing the search space. The complexity of histogram construction is :

$$\mathcal{O}\Big(\#\text{bins} \times \#\text{features}\Big). \tag{2.20}$$

LightGBM distinguishes itself from other implementations, such as XGBoost, through two innovations that further accelerate training and improve accuracy :

— **Gradient-based One-Side Sampling (GOSS)** : During each iteration, LightGBM divides training samples based on the magnitude of their gradient statistics. Let $g_i$ denote the gradient of the loss for observation $i$. The dataset is split into :
  — A subset containing the top $X\%$ of samples with the largest gradients (those learned poorly).
  — A subset containing the remaining samples with small gradients (those learned well).

  All large-gradient samples are retained, while a random sample of size $Y$ is drawn from the small-gradient set. To correct for the sampling bias, gradients of the sampled low-gradient instances are rescaled by a factor of :

$$\frac{1-X}{Y}.$$

  This approach preserves the overall gradient distribution while focusing computational effort on poorly fitted observations.

— **Exclusive Feature Bundling (EFB)** : LightGBM also reduces the number of effective features by merging mutually exclusive (sparse) features into a single composite feature. Two features are mutually exclusive if they never take non-zero values simultaneously. By bundling these features, LightGBM reduces memory consumption and the effective dimensionality of the histogram without losing information.

Figure 2.4 shows the GOSS and the EFB optimization techniques [174]. These optimizations enable LightGBM to train faster and more efficiently than other gradient boosting frameworks, including XGBoost, particularly on high-dimensional sparse datasets common in real-world applications [173, 175].

In [101], a GBM model trained on sub-sampled data from the University of Chicago and MIMIC-III databases achieved AUROCs of 76% and 71%, respectively. In [86], GBM outper-

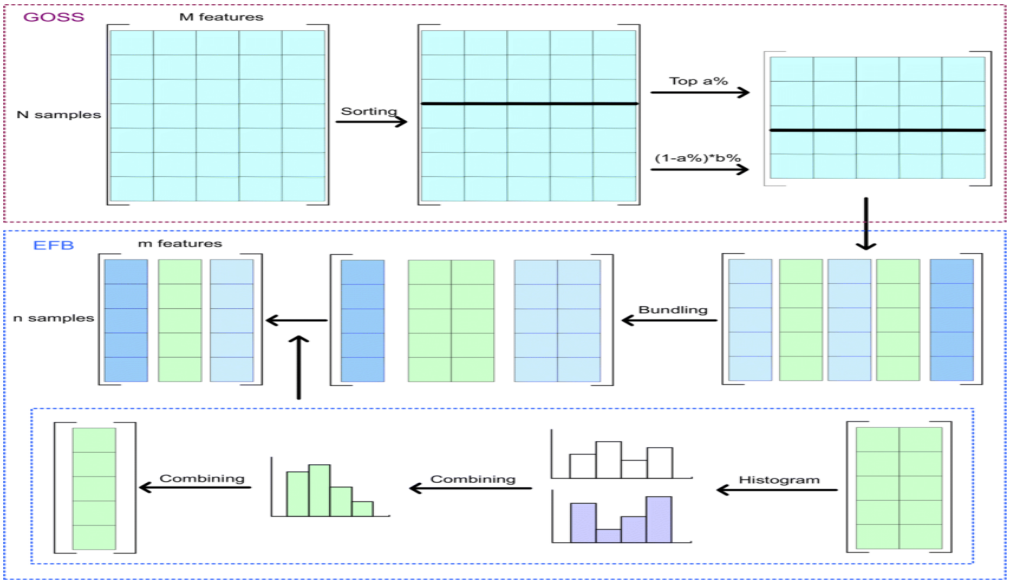



FIGURE 2.4 EFB and GOSS techniques

formed several ML models with an AUROC of 85% using the STARR database but faced a decline to 60% during external validation on the MIMIC-IV dataset, suggesting that GBM may be less prone to overfitting than LR. Zhu, *et al.* [81] applied transfer learning from MIMIC-II to CHOA databases. They applied non-Negative Matrix Factorization (NMF) and Convolutional AutoEncoders (CAE) for feature representation, and fine-tuned classifiers on CHOA data. Although transfer learning improved GBM performance with CAE-extracted features to an AUROC of 62%, a GBM classifier achieved an AUROC of 77% when trained on CHOA data.

In [102], AdaBoost achieved the best performance with an AUROC of 91% using the arrival attribute set. Cost-sensitive classification outperformed Synthetic Minority Over-sampling Technique (SMOTE) [176] for class imbalance, and the study highlighted the value of early patient characteristics over data available at ICU discharge. The study's limited private ICU sample reduces variability and generalizability, and manual data entry introduces error risk, while modeling time-varying variables as categorical data neglects their dynamic nature. Desautels, *et al.* [103] developed AutoTriage ML, which utilized transfer learning with AdaBoost, trained on MIMIC-III (source) and CUH (target) datasets. By adjusting the weighting between datasets, this approach achieved an AUROC of 71%, demonstrating that prioritizing the target dataset in transfer learning outperformed models trained solely on the source or target datasets served as an effective regularizer, preventing overfitting to the target domain and enhancing performance. In [26], the LightGBM classifier, trained with features selected for mutual information and weighted for the minority class, achieved an AUROC of 79%.

The study underscored that diagnoses are key contributors to ICU readmission.

Pakbin, *et al.* [114] used XGBoost to predict readmissions at different time windows, combining multiple feature entries to reduce missingness. The model outperformed LR, achieving an AUROC of 76% and 84% for predicting 72 hours and bounceback readmissions, respectively. Their findings highlighted distinctions in short- and long-term readmission risks, with the diagnosis as the most correlated risk factor. Thoral, *et al.* [104] used SHapley Additive exPlanations (SHAP) [177] to highlight feature importance and enhance the interpretability of an XGBoost model trained on AmsterdamUMCdb data. Feature selection with LR and L1 regularization achieved an AUROC of 77%. In a subsequent study [105], Thoral's model was assessed through a temporal validation design Leiden UMC data. External validation revealed moderate discrimination with an AUROC of 72%. Retraining the model using different time point subsets improved the AUROC to 79%, indicating no changes in data drift affecting performance over time, and highlighting the importance of retraining models on new data. In [178], XGBoost's hyperparameters were optimized using Tree-structured Parzen Estimator (TPE) [179], a Bayesian optimization technique. They achieved remarkable results with an AUROC of 92% and an Area Under Precision-Recall Curve (AUPRC) of 65%, with SHAP values identifying length of stay as the most influential feature. A low AUPRC indicates that the model struggles to correctly identify true positives while minimizing false positives, which is particularly problematic in imbalanced datasets where the positive class is rare.

Hegselmann, *et al.* [87] developed an Explainable Boosting Machine (EBM), an additive model using shape functions for interactions between variables and the logit link for dichotomous classifications. With features selected via mean absolute log-odds score from ANIT-UKM hospital data, it achieved an AUROC of 68%, outperforming LR and Recurrent neural Network (RNN) but similar to GBM. Validation on MIMIC-IV, benefiting from better data quality, improved the AUROC to 76%. The EBM's transparency, reviewed by a multidisciplinary team, highlights its advantages for healthcare applications. Tree-based models showed promising performance with AUROCs ranging from 59% to 92%, averaging 77%.

The K-means algorithm partitions patients into clusters by assigning them based on similarity to the cluster centroid and iteratively refining assignments to minimize within-cluster variance [180]. In [99], K-means outperformed k-medoids [181] and x-means [182], achieving a Davies-Bouldin Index (DBI) of 56.

Table 2.2 summarizes ML approaches. While ML models have significantly improved results, their performance tends to decline with larger sample sizes. Developing more complex models is essential for creating more reliable predictions.

TABLE 2.2 ML approaches for ICU readmission prediction ( * : External validation)

| Study | Database Name (Sample Size) | Time Period (days) Readmission Rate % | Data Type | | | | Classifier | Performance | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Demographic | Temporal | Clinical Notes | Medical Codes | | AUROC (%) | APURC (%) | Accuracy (%) |
| Inan 2018 [158] | MIMIC-III (11,000) | - | ● | ● | ● | ● | NN | 87 | 95 | 95 |
| Junqueira 2019 [159] | MIMIC-III (42,307) | 30 (11) | ● | ● | ● | ● | NN | 64 | - | 86 |
| Raza 2023 [161] | MIMIC-III (6,500) | 30 (50) | ● | ● | ● | ● | NN | F1 score : 84 | | |
| Negar 2022 [121] | MIMIC-III (10,894) | 30 (30) | ● | ● | ● | ● | SVM | 74 | - | - |
| Braga 2014 [98] | Private | 30 (1) | ● | ● | ● | ● | NB | - | - | 99 |
| Silva 2015 [166] | MIMIC-II (19,075) | same (13) | ● | ● | ● | ● | AFS-DT | 59 | - | 61 |
| He 2022 [80] | MIMIC-II (1,622) | 3 (1) | ● | ● | ● | ● | DT+ LR | 81 | - | 73 |
| Wang 2021 [100] | Private (4,697) | 2 (13) | ● | ● | ● | ● | RF | 88 | - | 87 |
| Alghatani 2022 [168] | MIMIC-III (44,626) | same (7) | ● | ● | ● | ● | RF | 74 | - | 68 |
| Khodadadi 2023 [90] | eICU (41,026) | same (17) | ● | ● | ● | ● | gcForest | 87 | 60 | - |
| Rojas 2018 [101] | Private (24,885)<br>MIMIC-III (42,303)* | same (11)<br>same (8) | ● | ● | ● | ● | GBM | 76<br>71 | -<br>- | -<br>- |
| Shi 2022 [86] | Private (3107)<br>MIMIC-IV (13,841)* | 7 (9)<br>7 (6) | ● | ● | ● | ● | GBM | 85<br>60 | 41<br>8 | 90<br>93 |
| Zhu 2022 [81] | MIMIC-II (32,331)<br>Private (5,739) | 30 (24)<br>30 | ● | ● | ● | ● | GBM | 77 | - | - |
| Loreto 2020 [102] | Private (9,926) | same (7) | ● | ● | ● | ● | RF<br>Adaboost | 91 | - | - |
| Desautels 2017 [103] | Private (2018)<br>MIMIC-III (44,741) | 2 (4)<br>2 (13) | ● | ● | ● | ● | AdaBoost | 71 | - | - |
| Fathy 2023 [26] | MIMIC-III (31,151) | 3 (3) | ● | ● | ● | ● | LightGBM | 79 | - | - |
| Pakbin 2018 [114] | MIMIC-III (3,637) | 3 (4)<br>30 (12)<br>same (7) | ● | ● | ● | ● | XGBoost | 76<br>75<br>84 | -<br>-<br>- | -<br>-<br>- |
| Thoral 2021 [104] | Private (18034) | 7 (4) | ● | ● | ● | ● | XGBoost | 77 | - | 12 |
| De Hond 2022 [105] | Private (10,052) | 7 (6) | ● | ● | ● | ● | XGBoost | 79 | - | - |
| González-Nóvoa 2023 [178] | MIMIC-III (28,557) | same (8) | ● | ● | ● | ● | XGBoost | 92 | 65 | - |
| Hegselmann 2022 [87] | Private (15,589)<br>MIMIC-IV (19,108) * | 3 (5)<br>3 (7) | ● | ● | ● | ● | EBM | 68<br>76 | 12<br>22 | -<br>- |
| Veloso 2014 [99] | Private (1043) | 30 (4) | ● | ● | ● | ● | K-means | DBI : 56 | | |

**Deep learning approaches**

Table 2.3 identifies 23 studies used to handle temporal data and clinical notes directly using models like RNN [183] or CNN [184] are often used. LSTM [185] and Gated Recurrent Unit (GRU) [186] are RNN variants that capture long-range dependencies in sequential data by retaining past information to improve future predictions. LSTM networks are a type of RNN specifically designed to model temporal dependencies and mitigate the vanishing gradient problem observed in conventional RNNs. Unlike standard RNNs, which maintain a single hidden state, LSTM introduces a memory cell that allows the network to learn which information to keep, update, or forget over long sequences.

An LSTM unit maintains two key vectors at each time step $t$ :

— the hidden state $\mathbf{h}_t$ (output of the cell), and

— the cell state $\mathbf{c}_t$ (internal memory).

Given an input vector $\mathbf{x}_t$, previous hidden state $\mathbf{h}_{t-1}$, and previous cell state $\mathbf{c}_{t-1}$, the LSTM computes the following :

**Forget gate :**

$$\mathbf{f}_t = \sigma\Big(\mathbf{W}_f\mathbf{x}_t + \mathbf{U}_f\mathbf{h}_{t-1} + \mathbf{b}_f\Big), \tag{2.21}$$

which controls how much of the previous cell state to forget.

**Input gate :**

$$\mathbf{i}_t = \sigma\Big(\mathbf{W}_i\mathbf{x}_t + \mathbf{U}_i\mathbf{h}_{t-1} + \mathbf{b}_i\Big), \tag{2.22}$$

which determines how much new information to store.

**Candidate memory :**

$$\tilde{\mathbf{c}}_t = \tanh\Big(\mathbf{W}_c\mathbf{x}_t + \mathbf{U}_c\mathbf{h}_{t-1} + \mathbf{b}_c\Big), \tag{2.23}$$

which represents new candidate values for the cell state.

**Cell state update :**

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t, \tag{2.24}$$

where $\odot$ denotes element-wise multiplication.

**Output gate :**

$$\mathbf{o}_t = \sigma\Big(\mathbf{W}_o\mathbf{x}_t + \mathbf{U}_o\mathbf{h}_{t-1} + \mathbf{b}_o\Big), \tag{2.25}$$

which controls how much of the cell state contributes to the hidden state.

**Hidden state update :**

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh\Big(\mathbf{c}_t\Big). \tag{2.26}$$

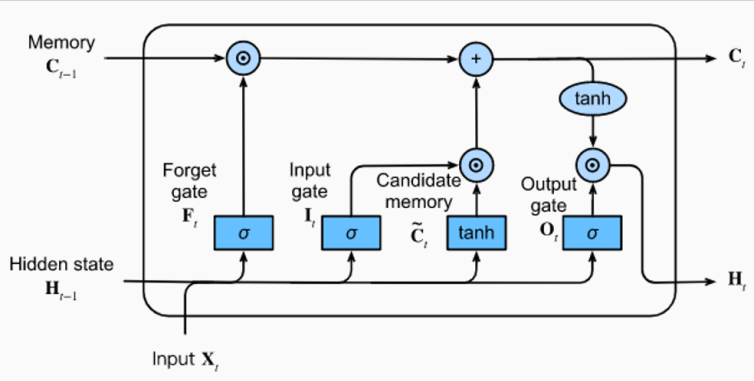



FIGURE 2.5 Architecture of the LSTM cell

Here, $\sigma(\cdot)$ denotes the sigmoid activation function, and $\tanh(\cdot)$ is the hyperbolic tangent. The gating mechanisms enable LSTMs to learn long-range dependencies by explicitly controlling the flow of information through the sequence.

Figure 2.5 illustrates the internal structure of an LSTM unit, including the input gate, forget gate, output gate, and cell state update operations [187]. The sigmoid ($\sigma$) and tanh activation functions control information flow, enabling the model to selectively retain or forget information over time.

On the other hand, CNNs learn hierarchical representations of data, which are useful for capturing patterns in sequential data with a spatial component. CNNs are a class of neural networks designed to automatically extract and learn hierarchical representations from structured data such as images or time series. Unlike fully connected layers, CNNs leverage convolutional filters to detect local patterns, enabling parameter sharing and translation invariance.

In the context of two-dimensional time series data $\mathbf{X} \in \mathbb{R}^{T \times V}$, where $T$ denotes the number of time steps and $V$ the number of variables, a 2-D convolutional layer applies learnable kernels along both the temporal and variable dimensions. This allows the model to capture joint temporal and inter-variable patterns in localized regions of the data.

Given an input matrix

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,V} \\ \vdots & \ddots & \vdots \\ x_{T,1} & \cdots & x_{T,V} \end{bmatrix},$$

and a filter $\mathbf{W} \in \mathbb{R}^{k_t \times k_v}$ of size $k_t$ (time) by $k_v$ (variables), the convolution operation produces

an output matrix $\mathbf{S}$ where each element is computed as :

$$s_{j,l} = \sum_{i=1}^{k_t} \sum_{m=1}^{k_v} W_{i,m} \, X_{j+i-1,\, l+m-1} + b,$$

where $b$ is a learnable bias term. The result is passed through a non-linear activation function such as the Rectified Linear Unit (ReLU) :

$$h_{j,l} = \max(0,\ s_{j,l}).$$

Multiple filters are applied in parallel to produce a set of feature maps, each encoding specific local patterns across time and variables.

To reduce dimensionality and improve robustness to small shifts, 2-D convolutional networks often include pooling operations. For example, with max pooling using a window of size $p_t \times p_v$, the downsampled representation is computed by taking the maximum over each pooling region :

$$h'_{j,l} = \max\Big\{ h_{\,(j-1)\,p_t+1\,:\,j\,p_t,\ (l-1)\,p_v+1\,:\,l\,p_v} \Big\},$$

where the maximum is taken over the rectangular window.

Multiple convolutional and pooling layers can be stacked to build deep architectures capable of learning increasingly abstract representations. Finally, the extracted features are flattened and passed to fully connected layers, which perform the final classification.

Figure 2.6 illustrates the main components of a CNN, including convolutional layers for feature extraction, ReLU activation functions, pooling layers for downsampling, flattening, fully connected dense layers, and a softmax output for classification [188].

The LSTM-CNN model was used in many studies. In [113], it achieved an AUROC of 79%, outperforming LSTM, CNN, and several ML models trained in statistical features. Their

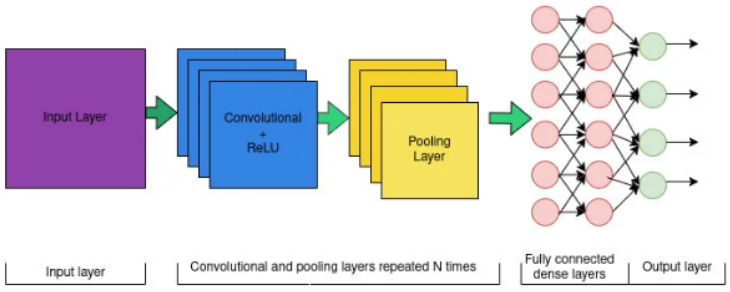



FIGURE 2.6 Schematic of the CNN architecture

study underscores the LSTM-CNN model's ability to handle time series data with high volatility and unstable conditions effectively. To enhance this work, Zebin, *et al.* [120] rebalanced the dataset using SMOTE and categorized ICD-9 codes into 17 classes, achieving an AUROC of 82%. In [130], the model was enhanced with advanced ICD-9 embeddings, including Poincaré embeddings, which achieved AUROCs of 79% and 78% for clinical notes and billing system ICD-9 codes, respectively, outperforming all graph embedding methods except TransE. Poincaré embeddings with 100 dimensions proved notably efficient, achieving an AUROC of 72% and demonstrating their effectiveness in hierarchical data representation. Chen,*et al.* [88] introduced Predictive Process Monitoring (PPM) using event logs to enhance ICU support. PPM learns from historical complete traces and makes predictions for ongoing, incomplete traces, treating each ICU stay as a process trace and utilizing rich time series information. The LSTM-CNN model improved with longer prefix lengths, achieving an AUROC of 64% with a prefix length of 21. Overall, the LSTM-CNN models performed well, with AUROC values ranging from 63% to 82%, and an average AUROC of 76%.

Although RNN and LSTM models are capable of modeling temporal dependencies, they struggle when dealing with long and irregular clinical time-series, where important information can occur far apart in time. These models process sequences sequentially, which makes training slow and can cause vanishing or diluted gradients, reducing their ability to capture long-range clinical patterns.

The attention mechanism assigns dynamic weights to input elements, helping models understand data well and focus on specific parts for better predictions [189]. To address RNN and CNN limitations, the transformer architecture processes all input tokens simultaneously, capturing long-range dependencies, handling irregularly sampled data, and improving text data analysis for identifying high-risk patients. Transformer-based attention mechanisms provide a flexible way to model dependencies between all elements of an input sequence, regardless of their distance. The core idea is the self-attention mechanism, which computes contextual representations by relating each position in the sequence to every other position.

Given an input sequence of $T$ elements, represented as vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$, self-attention projects the inputs into three spaces : queries ($Q$), keys ($K$), and values ($V$). For each element $\mathbf{x}_i$, the projections are :

$$\mathbf{q}_i = \mathbf{W}^Q \mathbf{x}_i, \tag{2.27}$$

$$\mathbf{k}_i = \mathbf{W}^K \mathbf{x}_i, \tag{2.28}$$

$$\mathbf{v}_i = \mathbf{W}^V \mathbf{x}_i, \tag{2.29}$$

where $\mathbf{W}^Q$, $\mathbf{W}^K$, and $\mathbf{W}^V$ are learned parameter matrices.

The attention score between position $i$ and position $j$ is computed as the scaled dot product of the query and key :

$$e_{ij} = \frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{d_k}}, \tag{2.30}$$

where $d_k$ is the dimensionality of the key vectors, included to prevent excessively large dot products that destabilize training.

The attention weights are obtained by applying the softmax function over the scores :

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{m=1}^{T} \exp(e_{im})}. \tag{2.31}$$

Finally, the output vector for position $i$ is the weighted sum of the value vectors :

$$\mathbf{z}_i = \sum_{j=1}^{T} \alpha_{ij}\, \mathbf{v}_j. \tag{2.32}$$

This process is referred to as single-head self-attention. In practice, Transformers use multi-head self-attention, where multiple attention mechanisms are applied in parallel and their outputs are concatenated :

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_H)\, \mathbf{W}^O, \tag{2.33}$$

where each head is computed as :

$$\text{head}_h = \text{Attention}\Big(Q\, \mathbf{W}_h^Q,\ K\, \mathbf{W}_h^K,\ V\, \mathbf{W}_h^V\Big).$$

This design allows the model to capture information from different representation subspaces jointly.

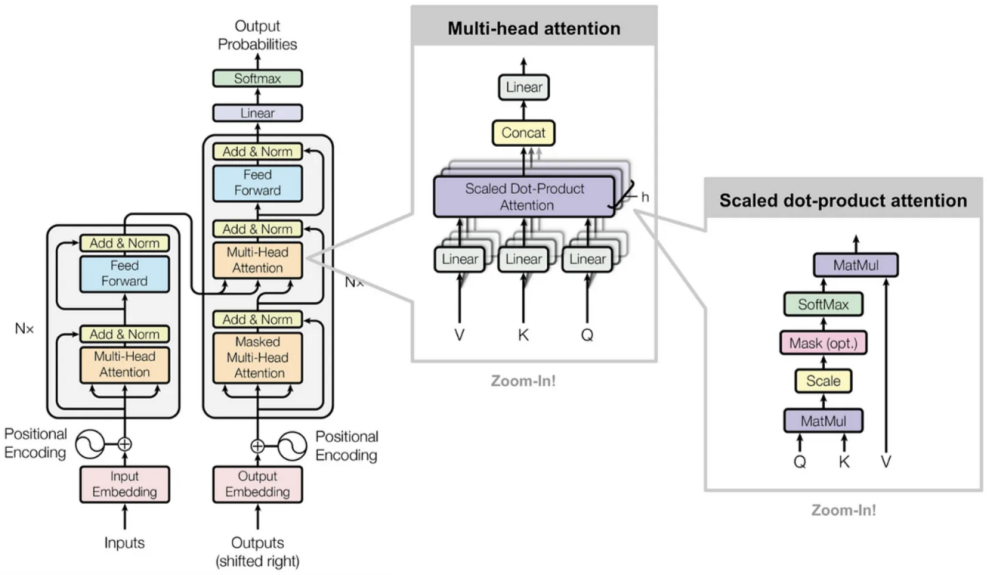



FIGURE 2.7 Transformer architecture with multi-head self-attention

Figure 2.7 shows the architecture of a Transformer model, illustrating the encoder and decoder blocks, each composed of multi-head attention layers, feed-forward networks, and layer normalization. The inset diagrams zoom into the multi-head attention mechanism and the scaled dot-product attention, highlighting how query (Q), key (K), and value (V) vectors are processed to compute attention weights and aggregate information across positions.

In [125], the patient's continuous clinical notes were concatenated and embedded using Word2Vec. An augmented CNN with a multi-headed attention mechanism was employed to extract problems, allowing for variable text spans while maintaining interpretability. The model explored various problem representations, including rolled-up ICD-9 codes and Phecodes [190], achieving AUROCs of 71% and 69% for bounceback and 30-day readmission, respectively. The model's interpretability is enhanced by using the intermediate problem list for final predictions. In [122], attention mechanisms in LSTM were examined for their correlation with feature importance and alterability. Log-odds attention showed a modest correlation and had minimal impact on predictions, raising questions about the explanatory power of attention mechanisms. Although log-odds attention offered interpretation, it differed from learned attention distributions. Despite this, incorporating attention mechanisms into LSTM consistently improved AUROC, reaching 71% with either additive or log-odds attention.

Longformer is a transformer model that handles longer sequences of text by using a sparse attention mechanism, processing longer documents more efficiently while maintaining strong

performance in various NLP tasks [191]. In [106], a model combining Longformer's global and sliding window mechanisms with BERT's special classification tokens was developed. It embedded absolute and relative temporality using event tokens. They used positional indices derived from EHR record times to create a unique and shared positional encoding. A global self-attention token was used to integrate static data. It achieved an AUROC of 84%.

Transformer-based NLP models were also investigated. In [109], the TAPER model, using BioBERT and a bidirectional GRU for text summarization, was proposed to obtain a unified text representation. An auto-encoder with GRUs further summarized sentence representations into a single patient text representation, aiming to reduce errors and capture crucial information effectively, achieving an AUROC of 67%. In [116], multimodality analysis showed that temporal abstractions of temporal data enhanced performance, particularly with gradient inclusion, but models that used ICD-9 codes outperformed them. Overall, the ClinicalBERT model trained on clinical notes achieved the best performance with an AUROC of 75%. In [192], text samples generated by the MedAug model improved the performance achieving AUROCs of 79% and 82% with ClinicalBERT and MedText classifier, respectively, though they peaked with more synthesized samples before slightly declining. Transformer-based models exhibited moderate performance, with AUROCs ranging from 67% to 84% and an average of 74%.

ODEs model system dynamics continuously over time using differential equations, treating time as a continuous variable, allowing for more efficient modeling of complex temporal dynamics [193]. Several methods were proposed in [117] to process time series sampled at irregular time intervals. RNN with time dynamics of code embeddings computed by neural ODEs, achieved the highest average AUROC of 74%. In [194], a correlation-enhanced Multitask learning with Pearson and RNN-based Neural ODEs Model (MP-ROM) was proposed, featuring a shared bottom structure and a dynamic weighting of the loss function. Task correlation enhanced the association between sub-tasks and neural ODEs enhanced feature learning to avoid local optima. The model achieved AUROCs of 74%, 74%, 74%, and 73% for predicting readmission at 5, 10, 15, and 30 days, respectively. This indicates that the 30-day prediction task benefited from multitask learning, and more improvements can be achieved by enhancing task correlations. The performance of ODE-based models was moderate, with an average AUROC of 73%.

These approaches extract features without considering their medical relevance or intermodality relationships. Recently, integrating external medical knowledge and graph methods for embedding multimodal data has shown significant performance improvements in medical applications. In [133], the Conceptual-Contextual (CC) embeddings model integrated exter-

nal knowledge into text representations. Using PubMed and MIMIC-III clinical notes with BioWordVec, it retrieved context sentences based on UMLS triplets, encoded them with a bidirectional LSTM, and modeled relationships with vector addition. They achieved an AUROC of 80%.

Graph Neural Networks (GNNs) [195], Graph Convolutional Networks (GCNs) [196], and Graph Attention Networks (GATs) [197] are key methods for graph classification. GNNs update node embeddings through message passing and aggregation, with attention weight learning enhancing neighbor importance. GCNs focus on aggregating features from neighboring nodes to capture local structures, while GATs use dynamic attention mechanisms to weigh neighbor importance. GCNs handle varying node degrees well, and GATs adapt to different edge significances, making both effective for analyzing complex graph-structured data. In [138], the MedText model represented clinical notes as document-level graphs, combining text and UMLS knowledge into a four-view graph to capture different interactions. It used GCN for encoding and an attention layer for decoding. the document is also encoded by a bidirectional LSTM, generating a second document-level representation, and the concatenated representations were classified using NN, achieving an AUROC of 83%. In [134], ME2Vec embedded medical services, doctors, and patients, emphasizing the temporal nature of EHR data. Biased random walks [198] highlighted rare services, and GAT predicted doctor specialties using a bipartite graph. An attributed multigraph was simplified using the duplication and annotation approach to derive patient embeddings. Training LR and RNN models with these embeddings achieved AUROC scores of 59% and 60%, respectively, outperforming other embedding and matrix factorization methods.

These self-attention graph models struggle to effectively learn attention parameters from scratch, often resulting in uniformly distributed attention weights among medical concepts. To address this, [135] introduced the Graph Convolutional Transformer (GCT), which uses an attention mask and KL divergence to focus on meaningful connections and an adjacency matrix based on conditional probability for visit representations. GCT outperformed the transformer, achieving an AUROC of 75% compared to 73%. Building on this, Liu, *et al.* [136] proposed the Statistics and Knowledge-based Graph Transformer (S_K_GT), which uses a knowledge attention network for optimized learning, achieving an AUROC of 76%. In [84], the CARE-30 model used a Graph Auto-Encoder (GAE) [199] to create a Directed Acyclic Graph (DAG) [200] for capturing causal relationships among variables. Latent representations from the GAE were combined with multi-modal variables encoded by transformers, achieving an AUROC of 79%. The model's interpretability was enhanced through graph weight thresholding, and robustness was improved with an Average Treatment Effect (ATE) derived loss function [201].

Graph-based models, while effective in representing complex relationships, struggle with capturing sequential dependencies and temporal dynamics present in time series due to their inherent lack of temporal processing capabilities. To address this, in [137], a Variationally Regularized Graph Neural Network (VGNN) was introduced to enhance GNN attention. The encoder processes medical embeddings to represent the graph, while the decoder provides inferences based on the graph representations. A latent layer generated latent variables, regularized by KL divergence, approximating the distributions to Gaussian where mean and standard deviation are computed from the graph by two separate feed-forward networks achieving an AUPRC of 40%. In [139], Graph Attention and RNN-based Neural ODE Model (GROM) was proposed to convert variable-length medical code sequences into fixed-size vectors. It combined a neural ODE layer for handling irregular time series data with graph attention using CCS knowledge to learn robust diagnostic code representations, reduce noise, and achieve an AUROC of 79%.

Researchers have used contrastive learning to improve graph-based models by better capturing patterns and relationships. In [91], Hypergraph Contrastive Learning (HCL) was introduced to represent complex relationships in EHR data using a Hypergraph Attention Network (HAT) [202], transformer, and GAT. HAT aggregated medical code embeddings using composer and dispatcher functions, the transformer learned code-code relationships, and GAT modeled patient-patient relationships. HCL achieved AUROC scores of 72% on the eICU database and 75% on the MIMIC-III database with supervised contrastive learning, outperforming VGNN and GCT. It also reached an AUROC of 69% with self-supervised learning on the eICU database, highlighting the effectiveness of contrastive learning and the importance of modeling diverse relationships in EHR data. In [92], the CodeText cross-modal Contrastive Learning (CTCL) framework tackled data heterogeneity and quality issues using a Multi-view Graph Convolution Network (MGCN) [203] and a cross-view contrastive learning module. BioClinicalBERT encoded clinical text, and a cross-modal encoder fused code and text representations, achieving an AUROC of 85%. In [85], semantic annotation and Knowledge Graph (KG) embeddings were used to uniformly represent multimodal data. RDF2Vec embeddings [204] from the National Cancer Institute Thesaurus (NCIT) ontology [205] selected using the BioPortal Recommender platform [206] and a RF model achieved the best results with an AUROC of 83%. The study found that multiple domain-specific ontologies did not outperform a single general-purpose ontology. Maximum performance was not achieved with discharge information alone, emphasizing the influence of data completeness, domain, and ontology appropriateness. Overall, graph and knowledge-based models showed promising performance, with AUROC values ranging from 59% to 85%, and an average AUROC of 75%.

Table 2.3 summarizes DL approaches for predicting ICU readmission, showing modest impro-

TABLE 2.3 DL approaches for ICU readmission prediction

| Study | Database Name (Sample Size) | Time Period (days) Readmission Rate % | Data Type | | | | Classifier | Performance | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Demographic | Temporal | Clinical Notes | Medical Codes | | AUROC (%) | APURC (%) | Accuracy (%) |
| Lin 2019 [113] | MIMIC-III (48,393) | 30 (14) | ● | ● | ● | ● | LSTM+CNN | 79 | - | - |
| Zebin 2019 [120] | MIMIC-III (48,393) | 30 (14) | ● | ● | ● | ● | LSTM+CNN | 82 | - | 73 |
| Lu 2019 [130] | MIMIC-III (48,411) | 30 (14) | ● | ● | ● | ● | LSTM-CNN | 79 | 48 | 75 |
| | | | ● | ● | ● | ● | | 78 | 48 | 72 |
| Chen 2022 [88] | MIMIC-IV (67,727) | 30 (14) | ● | ● | ● | ● | LSTM+CNN | 63 | - | 65 |
| Lovelace 2020 [125] | MIMIC-III (45,260) | 30 (13) | ● | ● | ● | ● | CNN+ attention | 69 | 24 | - |
| | | same (8) | | | | | | 71 | 17 | - |
| Jain 2019 [122] | MIMIC-III (34289) | 30 (22) | ● | ● | ● | ● | LSTM+ log-odds attention | 71 | 29 | - |
| Darabi 2020 [109] | MIMIC-III (38,597) | 30 | ● | ● | ● | ● | TAPER | 67 | 68 | - |
| Sheetrit 2023 [116] | MIMIC-III (15,424 ) | 30 (11) | ● | ● | ● | ● | ClinclaBERT | 75 | 30 | - |
| | | | ● | ● | ● | ● | GRU | 75 | 29 | - |
| Lu 2021 [192] | MIMIC-III (37802) | 30 (20) | ● | ● | ● | ● | MedAug+ MedText | 82 | 63 | - |
| Shickel 2022 [106] | Private (73,190) | same (6) | ● | ● | ● | ● | Longformer | 84 | - | - |
| Barbieri 2020 [117] | MIMIC-III (45,298) | 30(12) | ● | ● | ● | ● | RNN (ODE time decay) | 74 | - | - |
| Niu 2023 [194] | MIMIC-III (13,383) | 5 (20) | ● | ● | ● | ● | MP-ROM | 74 | - | - |
| | | 30 (41) | | | | | | 73 | - | - |
| Zhang 2020 [133] | MIMIC-III (48,393) | 30 (14) | ● | ● | ● | ● | CC-LSTM | 80 | 61 | 85 |
| Lu 2021 [138] | MIMIC-III (48, 393) | 30 (14) | ● | ● | ● | ● | MedText | 83 | 63 | - |
| Wu 2021 [134] | eICU (141,666) | (13) | ● | ● | ● | ● | ME2Vec+RNN | 60 | 20 | - |
| | | | | | | | ME2Vec+LR | 59 | 19 | - |
| Wang 2023 [84] | MIMIC-III (38,023) | 30 (16) | ● | ● | ● | ● | CARE-30 | 79 | 54 | 85 |
| Choi 2020 [135] | eICU (41,026) | same (17) | ● | ● | ● | ● | GCT | 75 | 52 | - |
| Liu 2021 [136] | eICU (41,026) | same (17) | ● | ● | ● | ● | S_K_GT | 76 | - | - |
| Zhu 2021 [137] | eICU (41,026) | (17) | ● | ● | ● | ● | VGNN | - | 40 | - |
| Pei 2021 [139] | MIMIC-III (45,298) | 30 (12) | ● | ● | ● | ● | GROM | 79 | - | - |
| Cai 2022 [91] | eICU (41,026) | (17) | ● | ● | ● | ● | HCL | 72 | 40 | - |
| | MIMIC-III (50,314) | (21) | | | | | | 75 | 43 | - |
| Sun 2023 [92] | eICU (15,360) | same | ● | ● | ● | ● | CTCL | 85 | 89 | |
| Carvalho 2023 [85] | MIMIC-III (48,392) | 30 (23) | ● | ● | ● | ● | KG embeddings +RF | 83 | 69 | - |

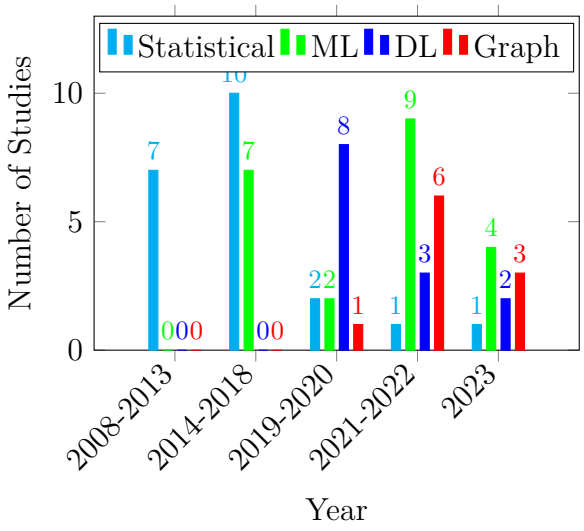



FIGURE 2.8 Number of studies per year and method type

vements over ML models. DL models fell short of expectations despite ample data availability, highlighting the problem's complexity. However, NLP significantly improved results by leveraging medical notes, and incorporating graphs and external knowledge enhanced performance. Future model advancements could potentially develop a decision-making system that meets expectations.

Figure 2.8 illustrates the evolution of study methods (Statistical, ML, DL, Graph) across five periods : 2008-2013, 2014-2018, 2019-2020, 2021-2022, and 2023. Initially, from 2008-2013, only statistical methods were used (7 studies). ML methods emerged in 2014, leading to 7 studies by 2018, due to their superior performance, causing a decline in statistical methods. DL methods appeared in 2019 with 8 studies, gaining popularity for their promising results. However, over time, there was a resurgence of interest in ML models, particularly boosting techniques, due to their interpretability compared to the black-box nature of DL methods. Interpretability is crucial in healthcare applications for clinical acceptance. The current trend shows an increase in graph-based methods, with 6 studies in 2021-2022 and 3 in 2023, as they mimic the hierarchical decisions of physicians, providing a structured approach.

### 2.2.4 Prediction models for PICU readmission

Most predictive efforts have been limited to adult ICU settings, and those that exist for PICU are scarce and often underperforming. Only two studies specifically developed predictive models for PICU readmission. Laksana *et al.* [207] employed RNN with LSTM units to predict patient readmission within the first three days following discharge, using data from

Children's Hospital Los Angeles. Despite incorporating extraneous features—synthetically generated by sampling from theoretical and empirical distributions and applying temporal masks to simulate missingness in clinical data—the model achieved a moderate predictive performance, with an AUROC of 64.4%. Arshad, *et al.* [208] compared RF, LR, and elastic net models, with RF slightly outperforming others (AUROC = 70 %). Both studies underline the role of physiological and laboratory data as strong predictors. However, these findings highlight a critical research gap, with minimal literature dedicated to predictive modeling specifically for PICU readmissions.

### 2.2.5 Interpretability, model calibration and generalization

Table 2.4 shows the summary of studies utilizing interpretation, model calibration, and generalization.

Interpretability addresses a key issue in healthcare AI : the use of "black box" models, which lack transparency in decision-making processes. Clinicians expect to see both global feature importance and patient-specific importance. Studies using LR models clarified interpretability through model weights [82,93,96], odds ratios [89], and nomograms [94,95]. In contrast, fuzzy model studies lacked interpretability due to complex rules, and non-linearity.

ML models, except tree-based ones, were black boxes due to their intricate mathematical computations. Boosting algorithms determine feature importance based on split count reducing impurity (Gini impurity or entropy) [26, 81, 87, 101, 114]. Post hoc explanation methods such as Local Interpretable Model-agnostic Explanations (LIME) [209] or SHAP were used to demonstrate local and global interpretability [104, 178]. However, post hoc methods have several shortcomings concerning robustness and adversarial attacks limiting their usefulness in health care settings [210].

Like ML models, DL and graph models are not easily interpretable due to their complex architectures, numerous parameters, and non-linear feature interactions. To address this, studies proposed using the Kolmogorov–Smirnov test [211] or dot-product attention to highlight important features [117, 125]. Ablation studies have been used to address ambiguity by systematically altering features or data modalities and observing the impact on performances [84, 85, 99, 109, 113, 116, 139, 194].

Calibration ensures that predicted probabilities align with actual outcomes, enhancing accuracy and reliability. Various calibration tests such as probability calibration curves, the Brier score [212], the Hosmer-Lemeshow goodness-of-fit test [213], the Wilcoxon signed-rank test [214], the Kruskal-Wallis test [215], post hoc sensitivity analysis [216], the Nemenyi

TABLE 2.4 Summary of studies utilizing interpretation, model calibration, and generalization

| Study | Interpretation | Model Calibration | Generalization |
|---|---|---|---|
| Campbell [93] | ● | ● | ● |
| Haribhakti [141] | ● | ● | ● |
| Jo [94] | ● | ● | ● |
| Frost [95] | ● | ● | ● |
| Badawi [89] | ● | ● | ● |
| Ouanes [96] | ● | ● | ● |
| Li [82] | ● | ● | ● |
| Vieira [147] | ● | ● | ● |
| He [80] | ● | ● | ● |
| Rojas [101] | ● | ● | ● |
| Shi [86] | ● | ● | ● |
| Zhu [81] | ● | ● | ● |
| Fathy [26] | ● | ● | ● |
| Pakbin [114] | ● | ● | ● |
| Thoral [104] | ● | ● | ● |
| De Hond [105] | ● | ● | ● |
| González-Nóvoa [178] | ● | ● | ● |
| Hegselmann [87] | ● | ● | ● |
| Veloso [99] | ● | ● | ● |
| Lin [113] | ● | ● | ● |
| Lovelace [125] | ● | ● | ● |
| Jain [122] | ● | ● | ● |
| Darabi [109] | ● | ● | ● |
| Sheetrit [116] | ● | ● | ● |
| Barbieri [117] | ● | ● | ● |
| Niu [194] | ● | ● | ● |
| Lu [138] | ● | ● | ● |
| Wang [84] | ● | ● | ● |
| Choi [135] | ● | ● | ● |
| Zhu [137] | ● | ● | ● |
| Pei [139] | ● | ● | ● |
| Cai [91] | ● | ● | ● |
| Sun [92] | ● | ● | ● |
| Carvalho [85] | ● | ● | ● |

test [217], Error bars, ANOVA test, Tukey's HSD test [218] and partial dependence plots were proposed [81, 85, 87, 89, 93, 101, 109, 114, 141, 147]. Impact analysis studies using probability-time curves and risk thresholds, case studies and generalization to low-quality data, and the value oscillation degree are also different methods used for calibration [92, 104, 113]. Various studies used ablation studies to isolate and quantify the impact of components and modifications, improving predictive accuracy and validating enhancements through comparisons with original models [84, 91, 92, 138, 194]. Some studies explored attention function behavior to understand how these mechanisms operate within models using methods such as singular value analysis and cluster compactness [122, 135, 137].

Model generalization is crucial in healthcare predictive models to ensure performance on new data. External validation tests, using independent datasets, assess robustness, realistic performance, clinician trust, and weaknesses. However, few studies report external validation performance. Some observed performance drops [86, 101], while others, like those validating on MIMIC-IV, noted improvements due to better data quality [87]. De Hond, *et al.* [105] assessed the performance of the Thoral model [104], emphasizing the need to retrain models before applying them to new data.

Implementing prediction models in production is crucial for leveraging their full potential to enhance healthcare outcomes. For a model to be practical, it must meet several key requirements. First and foremost, the model must comply with regulations regarding patient data privacy and security, ensuring that sensitive information is protected. It must consistently provide accurate and reliable predictions with high sensitivity and specificity to ensure patient safety and effective clinical outcomes. The model should be capable of handling large volumes of data and scaling with the hospital's needs, demonstrating robustness and generalization and offering timely results. Both local and global interpretability are crucial. The model should also have a user-friendly interface, regular updates, and maintenance. Lastly, the implementation and maintenance costs of the model should be proper.

Few studies implemented a practical model. INTCare is an Intelligent Decision Support System (IDSS) that published two studies [98, 99]. He, *et al.* [80] integrated their model with a secured server and a Graphical User Interface (GUI) featuring input, feedback, and maintenance layers. In [168], the Intelligent ICU Patient Monitoring (IICUPM) module was proposed. Thoral, *et al.* [104] and De Hond, *et al.* [105] proposed the Pacmed Critical module.

Table 2.5 presents an evaluation of practical prediction models against key requirements. Red flags indicate either a lack of mention or failure to meet the criteria. Models with an AUROC below 90% are deemed inadequate in terms of accuracy. Additionally, models utilizing datasets with fewer than 30,000 samples are considered unsatisfactory in terms of scalability.

TABLE 2.5 Assessment of practical prediction models against key requirements

| Study | Data Privacy and Security | Accuracy and Reliability | Scalability | Generalization | Timeliness | Interpretability | Ease of Use | Maintenance and Updates | Cost-Effectiveness |
|---|---|---|---|---|---|---|---|---|---|
| He [80] | ● | ● | ● | ● | ● | ● | ● | ● | ● |
| Alghatani [168] | ● | ● | ● | ● | ● | ● | ● | ● | ● |
| Thoral [104] | ● | ● | ● | ● | ● | ● | ● | ● | ● |

## 2.3 Gaps in Current Literature

The literature review highlighted a significant gap between current prediction models and practical applications. While many models are theoretically promising, they often fall short in performance, scalability, interpretability, generalizability, and clinical integration due to several factors. Variations in defining ICU readmission timeframes complicate comparisons. Studies have shifted from 2-day windows [89, 93, 100, 103] to 30-day windows [78, 82, 83, 108, 115], influenced by research goals, healthcare contexts, and efforts to increase instances in the minority class. Some studies inaccurately treat ICU readmissions and deaths as equivalent outcomes [85,88,89,103,104,113]. However, clinical research by Krumholz, *et al.* [219] suggests these outcomes are orthogonal, questioning the validity of modeling them jointly. Additionally, many models are unsuitable for real-time use because they rely on data from either early [93, 107, 168] or end-of-stay [78, 88, 89, 101, 113, 158] periods, or on discharge summaries and ICD-9 codes available only at the end of stay [121, 122, 125, 130]. This approach ignores ongoing patient progress, resulting in a fixed readmission probability that does not reflect treatments and health changes, thus making real-time prediction impractical. Ideally, a model should utilize all available data to understand patient changes during the ICU stay for more accurate real-time predictions. Inclusion and exclusion criteria vary across different studies, variability can lead to overlooking important factors, potentially impacting the model's performance. Table 2.6 summarizes the inclusion and exclusion criteria for patient cohorts. Additionally, some studies may use a limited number of variables based on previous knowledge or hypotheses, potentially leaving out important predictors [86, 89, 95, 104, 107, 112, 151].

Many studies do not specify whether imputation methods were employed, and models frequently encounter difficulties with incomplete data. Some imputation techniques have been utilized, with the Last Observed Carried Forward (LOCF) method being the most com-

TABLE 2.6 Summary of patient cohort inclusion and exclusion criteria

| Study | Inclusion criteria | | | | Exclusion criteria | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Age ( $\geq$ years)** | **ICU LOS** | $1st$ **ICU** | **Frequency per variable** | **died during ICU or hospital stay** | **Transferred between hospital units** | **Discharged from the hospital** | **Planned admission or discharge** | **Registration problems** |
| Campbell [93] | 16 | - | - | - | ✓ | - | - | ✓ | - |
| Haribhakti [141] | 18 | - | - | - | ✓ | - | - | ✓ | - |
| Jo [94] | 18 | - | - | - | ✓ | ✓ | - | ✓ | - |
| Forst [95] | 15 | - | - | - | ✓ | - | ✓ | ✓ | - |
| Badwi [89] | 16 | $\geq$ 4 hrs | - | - | ✓ | ✓ | ✓ | ✓ | - |
| Ouanes [96] | 18 | $\geq$24 hrs | ✓ | - | ✓ | - | ✓ | ✓ | - |
| Xue [78] | - | $\geq$ 12 hrs | ✓ | - | - | - | - | - | - |
| Li [82] | - | - | - | - | ✓ | - | - | ✓ | ✓ |
| Fialho [110] | 15 | $\geq$24 hrs | - | All | ✓ | - | ✓ | - | - |
| Fialho [97] | 15 | $\geq$ 24 hrs | - | $\geq$ 1 | ✓ | - | ✓ | - | - |
| Vieira [147] | 15 | $\geq$ 24 hrs | - | - | ✓ | - | ✓ | - | ✓ |
| Curto [107] | - | $\geq$ 24 hrs | - | $\geq$ 2 | ✓ | - | - | - | - |
| Salgado [153] | - | $\geq$ 24 hrs | - | - | - | - | - | - | - |
| Viegas [112] | 15 | $\geq$ 24 hrs | - | $\geq$ 2 | ✓ | - | - | - | - |
| Junqueira [159] | 18 | - | - | - | ✓ | - | ✓ | - | - |
| Raza [161] | 18 | - | ✓ | - | ✓ | ✓ | - | - | - |
| Negar [121] | - | - | - | - | - | - | - | ✓ | - |
| Silva [166] | 15 | - | - | - | - | - | - | - | - |
| He [80] | 15 | $\geq$ 24 hrs | - | All | ✓ | - | - | - | - |
| Wang [100] | - | $<$ 30 days | - | - | ✓ | - | - | - | - |
| Khodadadi [90] | - | $\geq$ 24 hrs | - | - | - | - | - | - | - |
| Rojas [101] | - | - | - | - | ✓ | ✓ | - | ✓ | - |
| Shi [86] | 18 | $\geq$ 24 hrs | ✓ | - | ✓ | ✓ | ✓ | ✓ | - |
| Zhu [81] | 16 | - | - | - | - | - | - | - | - |
| Desautels [103] | 16 | $\geq$ 6 hrs | - | $\geq$ 1 | - | ✓ | - | - | ✓ |
| Fathy [26] | 18 | - | ✓ | - | ✓ | - | - | - | ✓ |
| Thoral [104] | 18 | $>$ 30 days | - | - | - | - | ✓ | ✓ | - |
| De Hond [105] | 18 | $\geq$ 12 hrs | - | - | - | - | - | - | - |
| González-Nóvoa [178] | 18 | - | ✓ | - | ✓ | - | ✓ | - | - |
| Hegselmann [87] | - | - | - | - | ✓ | - | ✓ | - | ✓ |
| Lin [113] | 18 | - | - | - | ✓ | - | - | - | - |
| Zebin [120] | 18 | - | - | - | ✓ | - | - | - | - |
| Lu [130] | - | - | - | - | ✓ | - | - | - | - |
| Chen [88] | 18 | - | - | - | ✓ | - | - | - | - |
| Darabi [109] | 18 | - | - | - | - | - | - | - | - |
| Sheetrit [116] | 18 | $\geq$ 24 hrs $<$ 30 days | ✓ | $\geq$1 | ✓ | - | - | - | - |
| Shickel [106] | - | $<$ 10 days | - | - | - | - | - | - | - |
| Barbieri [117] | 18 | - | - | - | ✓ | - | - | - | - |
| Sun [92] | - | $\geq$ 24 hrs | - | - | - | - | - | - | ✓ |
| Carvalho [85] | 18 | - | - | - | - | - | - | - | - |

monly used across multiple studies [26, 79, 84, 88, 97, 103, 110, 113, 147, 151, 158]. MICE was applied in [78, 89], KNN was used in [83], while the Expectation-Maximization (EM) algorithm [220] was used in [81, 108]. Additionally, statistical techniques such as mean imputation, median imputation, interpolation, or utilizing normal variable values were employed in [86, 87, 106, 114, 116], and a weight decay term was added to the RF model in [100]. However, the impact of these imputation methods on model performance remains insufficiently studied.

Additionally, data imbalance is underexplored, with many studies using subsampling that may affect generalizability [83, 88, 101, 113, 121, 158, 159]. Solutions include using k-means clustering to resemble majority class [168], giving more weight to the minority class during classification [26, 100, 102, 117], upsampling [98], or using augmentation techniques like SMOTE [80, 102, 120]. In [151], a balanced training set was used, with the remaining data reserved for the test set. Generating synthetic temporal data remains difficult, although methods like a teacher-student framework using Generative Pre-trained Transformer-2 (GPT-2) [221] as the teacher and CNN-LSTM as the student show promise [192]. A comprehensive evaluation of the quality of these synthetic samples and their impact on model efficiency is needed.

EHR databases offer many variables, leading to a broad range of features such as statistical, temporal, and spectral. However, this results in a large number of potentially irrelevant features. Some studies employed filter feature selection methods [222], such as Information Gain (IG) and Principal Component Analysis (PCA) in [83, 102], statistical tests like chi-squared or Fisher exact test and Wilcoxon or Kruskal-Wallis test in [96, 141], NMF in [81], mutual information in [26], Correlation Coefficient, Relief, and Correlation-based Feature Selection (CFS) in [158], SU in [159], and LR-Test in [89, 94, 100]. Other studies used wrapper methods [223], such as SFS in [112], tree search feature selection [224] in [110, 147] and BFSS in [111]. Additionally, embedded methods [225] was utilized in some studies [87, 95, 121], while others applied ablation studies to define the important features [113]. Table 2.7 summarizes used preprocessing techniques. Only 39 employed at least one preprocessing technique, with only 5 studies implementing all preprocessing techniques.

Another limitation is handling inaccurate medical information, which impacts model reliability. Some studies incorrectly represented the GCS eye-opening with eight categorical values instead of the correct four [113]. Such errors arise from varied data storage techniques and a lack of medical expertise among researchers. Additionally, most studies do not address outliers or different measurement units, leading to result inconsistencies. Although some proposed pipelines handle these issues [23, 226], they have not been widely adopted. Inappropriate eva-

TABLE 2.7 Overview of preprocessing techniques utilized in reviewed studies

| Study | Feature selection | Imputation Technique | Rebalance Technique |
|---|---|---|---|
| Haribhakti [141] | ● | ● | ● |
| Jo [94] | ● | ● | ● |
| Forst [95] | ● | ● | ● |
| Badwi [89] | ● | ● | ● |
| Ouanes [96] | ● | ● | ● |
| Xue [78] | ● | ● | ● |
| Moerschbacher [83] | ● | ● | ● |
| Fialho [110] | ● | ● | ● |
| Fialho [97] | ● | ● | ● |
| Vieira [147] | ● | ● | ● |
| Sargo [111] | ● | ● | ● |
| Ferreira [151] | ● | ● | ● |
| Salgado [79] | ● | ● | ● |
| Viegas [112] | ● | ● | ● |
| Venugopalan [108] | ● | ● | ● |
| Inan [158] | ● | ● | ● |
| Junqueira [159] | ● | ● | ● |
| Negar [121] | ● | ● | ● |
| Braga [98] | ● | ● | ● |
| He [80] | ● | ● | ● |
| Wang [100] | ● | ● | ● |
| Alghatani [168] | ● | ● | ● |
| Rojas [101] | ● | ● | ● |
| Shi [86] | ● | ● | ● |
| Zhu [81] | ● | ● | ● |
| Loreto [102] | ● | ● | ● |
| Desautels [103] | ● | ● | ● |
| Fathy [26] | ● | ● | ● |
| Pakbin [114] | ● | ● | ● |
| Thoral [104] | ● | ● | ● |
| Hegselmann [87] | ● | ● | ● |
| Lin [113] | ● | ● | ● |
| Zebin [120] | ● | ● | ● |
| Chen [88] | ● | ● | ● |
| Sheetrit [116] | ● | ● | ● |
| Lu [192] | ● | ● | ● |
| Shickel [106] | ● | ● | ● |
| Barbieri [117] | ● | ● | ● |
| Wang 2023 [84] | ● | ● | ● |

luation metrics like accuracy or precision can be misleading with imbalanced data, as models may excel with the majority class but underperform with the minority [86, 98, 108, 158, 159]. Additionally, most studies lack external validation, limiting result generalizability. Many ML and DL models also suffer from poor interpretability and therefore provide limited utility in clinical contexts where transparency and explainability are critical for trust and adoption, hindering clinical use. Furthermore few models incorporate real-time data, further restricting their practical application.

## 2.4 Challenges in Pediatric Healthcare Modeling

Predicting readmission in PICU patients presents unique challenges beyond those encountered in adult ICUs that makes it more complex :

- **Physiological variability** : Pediatric vital signs vary substantially by age, requiring age-aware feature normalization.
- **Limited sample size** : PICU datasets are often smaller than adult ICU datasets, limiting model complexity.
- **Clinical heterogeneity** : Pediatric patients range from neonates to adolescents, each with distinct medical profiles.

These factors make the direct application of adult ICU prediction models to PICU cohorts unreliable and justify the need for pediatric-specific modeling approaches.

In conclusion, despite the progress in readmission prediction, several critical gaps remain :

- **Limited focus on PICU readmission** : Most models are developed for adult populations, with virtually no robust predictive models tailored to pediatric ICUs.
- **Data quality issues in EHRs** : High missing values, inconsistent identifiers, and manual entry errors.
- **Insufficient feature engineering** : Few works incorporate temporal or spectral features from physiological signals.
- **Imbalanced data** : Readmission rates typically under 4%, complicating model training and evaluation.
- **Lack of interpretability** : Many models function as black boxes, limiting their clinical trustworthiness.
- **Poor generalizability** : Most models are not tested across datasets or care settings.

.

## 2.5 Research Question and Objectives

Based on the gaps identified, this thesis poses the following research question :

> *Can we develop a high-performing and interpretable model for predicting Pediatric Intensive Care Unit (PICU) readmission using limited and imbalanced clinical data available throughout the patient's entire length of stay ?*

To answer this question, the research pursues the following specific objectives :

1. **Develop a predictive pipeline for ICU readmission using adult ICU datasets.**

   *Rationale :*

   (a) There is a scarcity of high-quality, publicly available PICU datasets.

   (b) Existing PICU readmission models are extremely rare in the literature and are typically based on private datasets, limiting reproducibility and comparison.

   (c) In contrast, adult ICU datasets (e.g., MIMIC-III) are publicly available and well-documented, with numerous published models that provide strong benchmarks.

   (d) This objective enables the systematic evaluation of a complete pipeline—including data preprocessing, feature extraction, model selection, and model architectures—to identify components that enhance predictive performance in ICU settings.

2. **Validate and Optimize the best-performing ICU pipeline to predict PICU readmission.**

   *Rationale :*

   (a) After developing the pipeline on adult ICU data, it will be further optimized to match the specific characteristics and clinical context of the PICU domain.

   (b) Deverging and integrating multimodal information (e.g., time-series vitals, lab results, treatments, and diagnoses), this phase aims to identify the most relevant features and suitable model architectures for predicting pediatric readmissions.

   (c) Addresses the specific challenges posed by limited sample sizes, class imbalance PICU dataset and the physiological heterogeneity across pediatric age groups compared to the more homogeneous adult ICU populations.

3. **Ensure model interpretability to support clinical insight and trust.**

   *Rationale :* Beyond predictive accuracy, interpretability is critical for clinical adoption.

   (a) This objective focuses on applying interpretable techniques (e.g., SHAP, ablation analysis) to explain model predictions.

(b) Emphasis is placed on aligning model explanations with feature engineering and clinical reasoning, identifying most important extracted features and the actionable risk factors to understand shared and unique readmission patterns across Adults and pediatrics, and supporting discharge decisions to reduce readmission risk.

Each objective serves a specific role in achieving the overall research goal :

— **Benchmarking on MIMIC-III Database** allows us to develop model architectures, feature extraction, feature selection, and evaluation methods using a rich, well-established adult ICU dataset.
— **Developing to Center Hospital University (CHU) Sainte-Justine PICU database (Cathy DB)** ensures the models are clinically meaningful in a pediatric setting and account for age-related variability in physiology.
— **Interpretability and Identifying risk factors** provides clinical insights and enhances trust by explaining model predictions, thereby facilitating real-world adoption.

This approach integrates traditional clinical risk analysis with advanced machine learning to address a pressing and underexplored problem in pediatric critical care.

## 2.6 Research Contributions

This thesis makes several key contributions to the field of clinical decision support systems and predictive healthcare analytics :

— Development of the first highly performing and interpretable PICU readmission prediction model using signal processing-derived features.
— Feature Innovation : Introduction of novel clinical meaningful features and extracting spectral features using signal processing techniques which enhance model discrimination between readmitted and non-readmitted patients.
— Age-aware Prediction : Implementation of models that account for the variability in normal vital sign ranges based on pediatric age groups.
— Data Quality Handling : Systematic handling of common EHR issues, including imputation strategies for missing values and techniques to address data imbalance.
— Interpretable Modeling : Development of interpretable models using algorithms like LightGBM combined with SHAP and LR combined with ablation analysis, enabling clinicians to understand the influence of individual features on prediction outcomes.
— Clinical Utility : The resulting predictive system is designed not only for accuracy but also for practical deployment in clinical settings, aiding physicians in making discharge decisions and reducing preventable PICU readmissions.

# CHAPTER 3 METHODOLOGY

## 3.1 General Framework

The general framework of predicting readmission during the first 3 days, follows a structured and reproducible pipeline designed to support robust predictive modeling. The process begins with the extraction of relevant clinical and physiological data from the designated database as shown in Figure 3.1. A set of stringent inclusion and exclusion criteria is then applied to define the target patient cohort. Specifically, we exclude patients who died during their ICU/PICU stay or within three days of discharge, those discharged directly to home, encounters with a Length Of Stay (LOS) shorter than six hours, cases entirely missing vital sign recordings or lacking admission/discharge timestamps, and patients outside the defined age brackets ($\geq$ 18 years for ICU readmission prediction and $<$18 years for PICU readmission).

In the literature, different time windows have been used to define PICU readmission, with 3-day and 30-day windows being the most common (Table 2.1, Table 2.2 and Table 2.3). A 30-day window is often selected in research contexts because it increases the number of readmission events, resulting in a less unbalanced dataset. However, this longer window does not align well with clinical priorities, as hospital quality-of-care evaluations focus primarily on early readmissions, which are more likely to be related to premature discharge, suboptimal stabilization, or complications that emerge shortly after leaving the ICU. In contrast, readmissions occurring several weeks later may reflect new or unrelated clinical events. Therefore, in this study, we selected a 3-day readmission window to reflect clinically meaningful readmission events, align with hospital priorities, and allow for comparison with prior studies that adopted the same early-readmission definition. This time window thus balances clinical relevance and methodological consistency.

Following cohort selection, the data undergo thorough preprocessing. This includes the detection and removal of outliers, correction of measurement inconsistencies, and standardization of units. Time-series variables are harmonized by aggregating values on an hourly basis using the median. Missing data are systematically addressed to ensure completeness in addition to the feature engineering phase, which encompasses both feature extraction (including time- and frequency-domain characteristics) and feature selection. Subsequently, a range of ML and DL models are developed to binary classify patients based on their risk of unplanned readmission. The final stage involves rigorous model evaluation, hyperparameter tuning, and performance assessment to ensure optimal accuracy, generalizability, and clinical relevance.

## 3.2 Databases

In this study, we utilized two critical care databases to develop and evaluate readmission prediction models. For adult ICU patients, we used the MIMIC-III database, which contains detailed clinical data from ICU stays at Beth Israel Deaconess Medical Center (BIDMC). For pediatric patients, we used the CathyDB database to predict PICU readmission. These datasets allowed us to build and compare models tailored to both adult and pediatric intensive care populations.

### 3.2.1 MIMIC-III database

We utilized the publicly available MIMIC-III database, a comprehensive and richly detailed resource that contains health data associated with nearly 60,000 ICU admissions for over 40,000 distinct patients, collected at the BIDMC in Boston, Massachusetts, between 2001 and 2012 [76].

MIMIC is the result of a longstanding collaboration beginning in the early 2000s between BIDMC (a teaching hospital affiliated with Harvard Medical School), the MIT Laboratory for Computational Physiology, and Philips Healthcare, with support from the National Institute of Biomedical Imaging and Bioinformatics (NIBIB). It was developed with the goal of democratizing access to high-quality ICU data for research and education purposes.

Unlike proprietary datasets such as the eICU database, MIMIC is freely accessible to credentialed users who complete the necessary data use agreement and training. The database includes a vast array of clinical data such as demographics, vital signs (e.g., HR, blood pres-

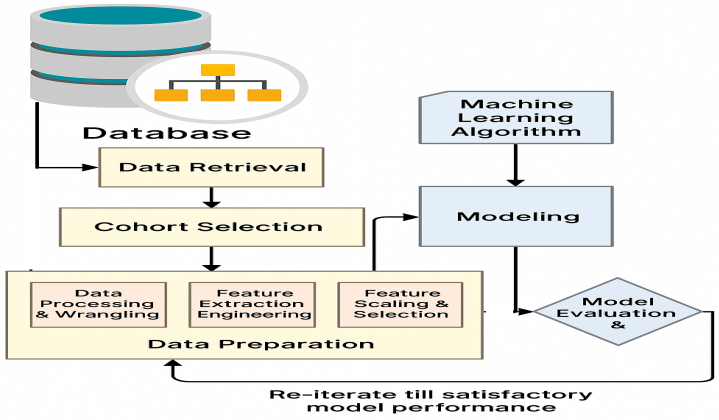



FIGURE 3.1 General framework for predicting ICU/PICU readmission

sure, RR), medications and infusions, laboratory test results, diagnosis and procedure codes (ICD-9), nursing notes and discharge summaries, ventilator settings, fluid balance, length of ICU and hospital stay and survival and readmission outcomes.

The comprehensiveness and open nature of MIMIC have led to its widespread adoption in clinical research. To date, hundreds of peer-reviewed publications have leveraged the database for tasks including mortality prediction, disease progression modeling, patient phenotyping, and ICU readmission prediction. Thousands of researchers, educators, and students across more than 30 countries use MIMIC annually, making it a cornerstone for reproducible machine learning and data science research in healthcare. Additionally, waveform data associated with MIMIC is available for a subset of patients, enabling advanced signal processing and physiological modeling.

In our study, we adopted and modified the data extraction pipeline proposed by Wang *et al.* [23], which processes MIMIC-III's raw CSV files into structured time-series data. A key adaptation in our version was the use of the median rather than the mean when aggregating vital signs within fixed time windows, providing robustness against extreme values and missingness. After applying our inclusion and exclusion criteria, the final cohort included 31,151 patients, of whom 1,033 (3.3%) were readmitted to the ICU within three days of discharge, consistent with prior studies on early ICU readmission using MIMIC data.

### 3.2.2 CathyDB database

CathyDB is a prospective high-quality and high-frequency database created in the PICU of Sainte Justine Hospital which represents a significant advancement in pediatric intensive care research. CathyDB contains many variables like the MIMIC-III database which makes it a valuable database. Designed to capture a comprehensive array of clinical data in real-time, this database collects physiological signals, respiratory and ventilator variables, pharmacotherapy details, and patient demographics at high frequencies (every 5 seconds from monitors and every 30 seconds from mechanical ventilators and infusion pumps). Diagnosis are represented using ICD-10CA codes. The study and database construction were approved by the institutional review board of Sainte Justine Hospital [227] with an opt-out option : no required consent from the parents and patients, but they are informed and can decide to be removed from the data collection to ensure that research involving human participants is conducted ethically and in accordance with established scientific and ethical principles.

The database's unique feature lies in its ability to integrate and synchronize diverse data streams, offering a detailed reconstruction of each patient's critical care journey. This robust dataset supports the development and validation of clinical decision support systems, com-

putational models of cardiorespiratory physiology, and other research initiatives aimed at enhancing patient care. While the database is currently limited to a single center, its innovative approach and rich data have the potential to transform pediatric intensive care research and pave the way for future multi-center collaborations.

The database contains information about 13,282 PICU admissions. After applying our inclusion and exclusion criteria, the final cohort included 11,288 patients, of whom 323 (2.86%) were readmitted to the PICU within three days of discharge.

## 3.3 Preprocessing

Data preprocessing is a critical step in building reliable ML models, especially in healthcare applications. It ensures data quality and consistency through cleaning, correcting errors, and standardizing formats. Handling missing data prevents bias and preserves valuable information. Feature extraction and selection help capture the most relevant patterns while reducing noise and complexity. Addressing class imbalance improves model fairness and performance, particularly when predicting rare outcomes like PICU readmission. Together, these steps enhance model accuracy, interpretability, and clinical usefulness.

### 3.3.1 Data cleaning

During data cleaning, we removed outlier values from the MIMIC-III dataset using a method proposed by Wang *et al.* [23], while outliers in the CathyDB dataset were retained due to their rarity. In addition to their minimal impact on the aggregated values, we used hourly median calculations to reduce their influence. To further mitigate potential bias, we applied a dedicated signal filtering method to enhance signal quality without distorting the underlying clinical patterns. We also corrected incorrectly registered values where some measurements were stored as strings instead of numerical values. Additionally, we standardized measurement units across all patients—for example, converting all temperature readings to Celsius—to ensure consistency in the data.

### 3.3.2 A category-aware imputation method

Imputing missing data is essential to ensure the completeness and reliability of the dataset. Instead of discarding incomplete records, which can lead to data loss or bias, we fill in missing values using appropriate techniques. This helps maintain the temporal continuity of physiological signals and ensures that ML and DL models can learn effectively from the available data.

To ensure clinically meaningful imputation for pediatric variables, we used age-based normal values whenever available. For physiological variables with established pediatric reference ranges, we imputed missing values using the mean of the normal values corresponding to the patient's age category (e.g., infants, toddlers, school-aged children). This approach accounts for developmental differences in normal physiology. For variables lacking official reference sheets or published normal ranges, we calculated the median value within each age category across the dataset and used it as a surrogate for the normal value. This allowed us to preserve age-specific variation and maintain clinical relevance in the imputed data.

Imputing missing data in a clinical setting requires careful consideration of the data type and its clinical context. We used a category-aware approach as a single imputation strategy is not suitable for all variables. Missed vital signs are often intentional, indicating that no clinically remarkable change was observed by the care team. Therefore, continuous measurements such as vital signs and fluid balance, which are expected to evolve smoothly over time, were imputed using the LOCF approach to maintain temporal continuity and avoid introducing artificial fluctuations. If the values of the variable are fully missing, we imputed the normal values of this variable. Medications and treatments, on the other hand, are administered at specific times and not continuously ; therefore, we used zero imputation to indicate that the procedure was not performed at that time. For clinical symptoms data such as signs of respiratory distress—typically recorded irregularly and showing gradual improvement—we applied interpolation to reflect symptom evolution over time. Weight values, which are crucial for pediatric care but not frequently measured, were imputed using a KNN model based on related features such as age, tidal volume and urine output.

Ventilation data introduced additional complexity due to various types of missingness. Figure 3.2 illustrates the different patterns observed. The first type is fully missing data, where an entire category of data is completely absent, ventilation type is entirely missing in Case 1, and ventilation settings are completely absent in Case 2. The second type is partially missing data, where some data exists but not for the entire time period. In Cases 3, 4 and 5, both ventilation type and settings are recorded but not cover the full duration. Additionally, in some cases, only a subset of settings is entirely missing represented by horizontal red lines in the figure. The third pattern is intermittent (time-specific) missing data, indicated by vertical red lines across cases. These indicate that data are missing at particular time points, likely due to occasional documentation errors or brief system failures during data collection.

To address these patterns, a tailored imputation strategy was applied. For fully missing ventilation types, the value was labeled as "other", and settings fully missed were imputed using median values based on age-specific norms. The same approach was used for partially

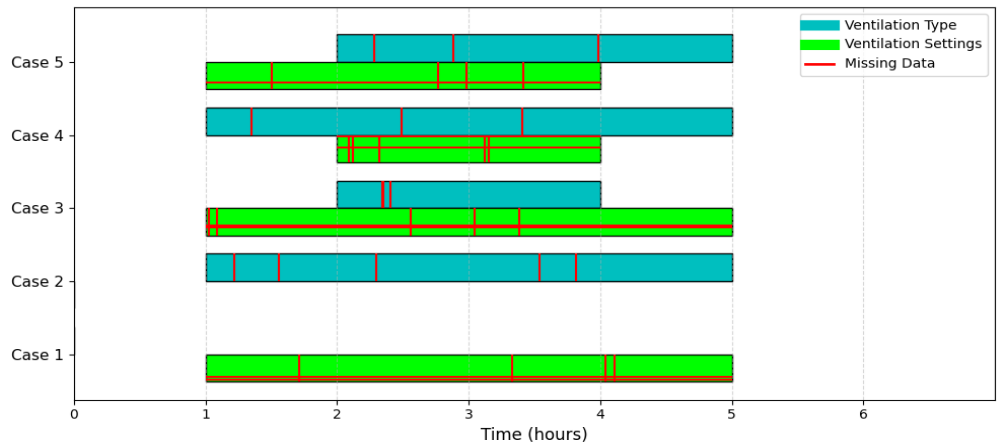



FIGURE 3.2 Visualization of missing data in ventilation type and settings

missing settings, where some data was available but not complete. For time-specific gaps, the LOCF method was used to maintain temporal continuity as ventilation settings do not change frequently. This approach preserved clinical relevance while ensuring data completeness for model development.

For variables that are not measured frequently, such as laboratory results, we chose not to impute missing values to avoid introducing misleading information, as these tests are ordered based on clinical judgment rather than at fixed intervals. In terms of demographic data, the MIMIC-III dataset had some missing values, which we imputed using the median value for each variable to preserve overall distribution without biasing the data. For contrast, the CathyDB dataset had no missing demographic information. For diagnosis, when no diagnostic information was available for a patient, we labeled it as "unknown diagnosis" to explicitly reflect the absence of clinical documentation rather than assume a specific condition. This approach ensured that imputation was applied thoughtfully, maintaining the clinical validity and interpretability of each data type.

The category-aware imputation strategy is one of the key contributions of our work, in which missing values are imputed according to the clinical and statistical nature of each variable category, rather than applying a single generic imputation method. This approach ensures that the imputed values remain realistic, clinically interpretable, and aligned with the expected behavior of each variable, thereby reducing the risk of introducing bias or artificial patterns. As a result, the final dataset better reflects actual patient physiology and clinical context, leading to more reliable model training and improved predictive performance.

### 3.3.3 Categorical states

Training models directly on age-based physiological variables may hurt performance, since raw values are weighted equally regardless of age. In addition, feature extraction from these variables may also fail to capture true abnormalities. We addressed the challenge of pediatric physiological variability and clinical heterogeneity by transforming age-based time series variables into categorical states—normal, low, or high—according to the age-specific normal range of each variable. These normal ranges were derived from the underlying data population, defined as the 5th and 90th percentiles (Q5–Q90), to ensure that the categorization reflected the real distribution of measurements across patients. In addition, we computed the absolute difference from the age-specific normal range to capture the magnitude of deviation.

This transformation is clinically important because a single numerical value (e.g., HR of 120 bpm) may indicate normality in one age group but pathology in another.For example, a given HR may be normal for an infant but abnormal for an adolescent. The example in Table 3.1 corresponds to HR measurements for a toddler aged 436 days (approximately 14 months). At this developmental stage, normal HR ranges are lower than those for infants but still higher than for older children and adults. The table illustrates how each measurement is represented by its raw numerical value (HR), categorical indicators showing whether the value is high, low, or within the age-specific normal range, and the absolute deviation (HR_diff) from that range. For this patient, most readings are flagged as high relative to the age-adjusted threshold, with deviations decreasing over time, eventually reaching a normal value at time step 5. This encoding allows the model to interpret the clinical significance of HR changes in the context of the patient's age.

By combining raw measurements with categorical normality states and age-specific deviation magnitudes, our approach will enable the model to capture the clinical meaning of numerical values rather than treating them as context-free numbers.

Transformation of age-dependent time series variables into categorical physiological states is another key contribution of our work. This transformation will allow the model to account for the natural developmental differences across pediatric age groups and will help it learn subtle physiological patterns across pediatric patients, where the same value can imply very different clinical conditions depending on age. Furthermore, by deriving normal ranges from population-based distributions rather than arbitrary cut-offs, we ensure that the model is grounded in the characteristics of the actual cohort, improving both robustness and clinical relevance in predicting PICU readmission risk.

TABLE 3.1 HR categorical indicators and deviation from age-based normal range

| Time | HR | HR_High | HR_Low | HR_Normal | HR_diff |
|---|---|---|---|---|---|
| 1 | 190.0 | 1 | 0 | 0 | 50.0 |
| 2 | 163.0 | 1 | 0 | 0 | 23.0 |
| 3 | 153.0 | 1 | 0 | 0 | 13.0 |
| 4 | 142.0 | 1 | 0 | 0 | 2.0 |
| 5 | 140.0 | 0 | 0 | 1 | 0.0 |

### 3.3.4 Feature engineering

Extracting meaningful features from raw data is crucial for capturing underlying patterns and variations that may not be immediately visible. Well-designed features enhance the model's ability to recognize clinical trends, detect abnormalities, and support accurate predictions. In this section, we describe the different feature extraction techniques used, including statistical features, signal processing-based features, and other domain-informed methods.

Capturing the essence of raw physiological signals begins with simple powerful statistical features that condense time series into informative descriptors. Basic statistics, including the mean and median for central tendency, the standard deviation for overall variability, and the minimum and maximum for observed range, quickly summarize how low or high values fluctuate. Recording the first and last measurements preserves temporal context by anchoring the signal's starting and ending states. Beyond these fundamentals, advanced statistics reveal subtler structure : skewness quantifies asymmetry, while kurtosis highlights heavy-tailed or outlier-prone distributions. Model-fit metrics deepen our understanding ; the R-squared value expresses how closely data align to a fitted linear regression $y = ax + b$, and the fitted slope $a$ and intercept $b$ capture underlying trends. Finally, the mean absolute differential, the average absolute distance between each data point and the series mean, offers a robust gauge of dispersion less sensitive to outliers than variance. Together, these features translate complex, noisy readings into concise, interpretable signals for downstream analysis and prediction.

#### Extracting time-frequency features using signal processing techniques

Vital signs are time-dependent signals that reflect dynamic changes in a patient's condition. These signals often contain important patterns, trends, and fluctuations that occur over different time scales and frequencies. Simple summary statistics may miss these subtle variations, which can be critical for early detection of clinical deterioration. Therefore, advanced signal processing techniques are essential to extract both temporal features (e.g., trends, variability) and spectral features (e.g., periodicity, frequency content), providing a deeper and

more informative representation of the data for predictive modeling.

To enhance the quality of the vital sign signals and enable more robust feature extraction, we implemented a noise reduction process. Vital sign signals acquired from bedside monitors are often contaminated by various sources of noise and artifacts, including patient movement, poor sensor contact, electrical interference, and physiological phenomena such as stress or infant crying. These noise sources can obscure clinically relevant information, introduce spurious features, and degrade the accuracy of downstream analyses.

Vital sign signals within the normal physiological range can still appear irregular due to noise. In many cases, these noise sources do not change the true state of the signal (e.g., it remains clinically "normal"), but they introduce rapid, small fluctuations that appear as ripples or jitter in the waveform. These fluctuations create the illusion of high-frequency changes as a white gaussian noise, even though no real physiological change has occurred. Therefore, removing such artifacts is essential, as they can distort downstream analyses and spectral extracted features, trigger false alarms, or mislead automated models into interpreting noise as meaningful variation.

To address these challenges and extract the true underlying physiological patterns, we employed a hybrid two-stage approach for noise reduction : first, we extracted the trend of the signal using the Kalman filter to smooth out measurement noise and short-term fluctuations ; then, we denoised the residual signal using a Discrete Wavelet Transform (DWT)–based shrinkage procedure, attenuating high-frequency noise while preserving edges and other salient features and details. This approach ensures that the resulting signals are cleaner and more representative of the patient's actual physiological state, enabling more accurate and interpretable feature extraction.

The vital signs in our dataset are recorded hourly for each patient throughout their ICU/PICU stay, which may range from as little as 6 hours to several months. This creates substantial variability in signal length, and patients with short stays yield only a small number of samples. Wavelet denoising is generally effective when the signal can be decomposed into multiple dyadic frequency bands, allowing true physiological trends to appear in coarser scales while high-frequency noise concentrates in finer scales. However, when the signal is very short, only a single-level wavelet decomposition is possible, which is insufficient to reliably separate significant information from noise, as both may occupy the same band.

To address this limitation, we adopt a two-stage filtering strategy. First, we apply a Kalman filter to extract the smooth physiological trend, leveraging the fact that vital signs follow predictable dynamic behavior : they evolve gradually, exhibit bounded variability, possess steady-state ranges under stable conditions, and their current value depends on previous

physiological states and clinical interventions. This makes them well modeled by a low-dimensional linear state-space formulation with approximately Gaussian noise, which is precisely the modeling assumption underlying Kalman filtering. After obtaining the trend component, we apply a single-level wavelet denoising step to remove residual high-frequency noise while preserving clinically meaningful edges and deviations. This two-stage Kalman–Wavelet filtering therefore ensures robust smoothing even for short-duration stays, while still retaining physiologically relevant variations necessary for accurate predictive modeling.

A Kalman filter is a recursive optimal estimator that models a system's true state through time in the presence of noise, producing a minimum-variance estimate of the underlying signal. Using a state-space representation, the filter alternates between predicting the next signal value based on a simple process model and updating this prediction with the latest noisy measurement, with the weighting determined by the Kalman gain. This balance—set by the relative uncertainties of the model and the sensor—allows the filter to dynamically adapt as the properties of the process or noise change [228].

By incorporating prior knowledge of physiological signal dynamics (such as typical HR or respiratory patterns) and expected noise characteristics, the Kalman filter can smooth vital sign measurements and suppress artifacts in real time. Its adaptive nature enables robust tracking of non-stationary signals, providing denoised estimates even as baselines drift or noise levels fluctuate [229]. For a one-dimensional time series, the standard Kalman filter equations are [228] :

**Prediction Step :**

$$\hat{x}_{k|k-1} = F\hat{x}_{k-1} \tag{3.1}$$

$$P_{k|k-1} = F\,P_{k-1}\,F^{\top} + Q \tag{3.2}$$

**Update Step :**

$$K_k = P_{k|k-1}\,H^{\top}\,(H\,P_{k|k-1}\,H^{\top} + R)^{-1} \tag{3.3}$$

$$\hat{x}_k = K_k\,z_k + \Big(1 - K_k\,H\Big)\,\hat{x}_{k|k-1} \tag{3.4}$$

$$P_k = \Big(I - K_k\,H\Big)\,P_{k|k-1} \tag{3.5}$$

Here, $\hat{x}_{k|k-1}$ denotes the predicted state (the estimated vital sign value before seeing the current measurement), while $P_{k|k-1}$ quantifies the predicted uncertainty or variance associated with this estimate—essentially reflecting our confidence in the prediction. The Kalman gain $K_k$ determines how much weight is given to the new, possibly noisy, sensor measurement $z_k$

relative to the model's prediction ; a small $K_k$ indicates greater trust in the model's prediction, while a large $K_k$ suggests more reliance on the observed data. The process noise covariance $Q$ captures our expectation of how much the true vital sign may naturally fluctuate from one time point to the next due to physiological changes, encoding our belief about the underlying signal's variability. In contrast, the measurement noise covariance $R$ models the expected error or unreliability in the sensor readings, reflecting the device's accuracy and susceptibility to artifacts. The state transition matrix $F$ and the observation matrix $H$ define the relationship between consecutive states and between the state and observations, respectively. Since we filter each vital sign individually and assume that, in the absence of noise or disturbance, the best prediction for the next value is the current value itself, both $F$ and $H$ are set to one (i.e., scalars). This choice means the state evolves according to a simple random walk, and the observation is a direct measurement of the state. Finally, the identity matrix $I$ appears in the update equation for covariance. By design, the output $\hat{x}_k$ from the Kalman filter should closely track the true, denoised value of the vital sign, and the covariance $P_k$ provides a quantitative measure of the remaining uncertainty in this estimate.

After applying a Kalman filter to a discrete-time observation signal $z[n]$, the objective is to estimate the underlying smooth process dynamics, denoted by $\hat{x}[n]$. Once this trend estimate is computed, the residual signal is obtained as :

$$r[n] = z[n] - \hat{x}[n]. \tag{3.6}$$

Kalman filtering is designed for trend estimation, so it tends to suppress high-frequency features, even important ones, like : short spikes, transient dips, abrupt real physiological changes. Wavelet-domain denoising procedure is ideal for analyzing the residual because it helps separate short bursts of signal from random noise and it can capture signal components missed by Kalman filter and suppress noise while preserving true transient events. In addition, many noise components can be removed with minimal loss.

The denoising process begins by projecting $r[n]$ onto an orthonormal wavelet basis using the DWT. The DWT is a multi-resolution analysis technique that iteratively decomposes the signal into multiple levels of approximation and detail coefficients by applying a series of filtering and downsampling operations. At each level, the signal is passed through a low-pass filter to extract the approximation coefficients, which capture the low-frequency (trend) components, and through a high-pass filter to extract the detail coefficients, representing the high-frequency (transient) features. Concretely, at each decomposition level $l$, the signal is

convolved with a pair of quadrature mirror filters :

$$A^{(l)}[k] = \sum_n h[n-2k]\, A^{(l-1)}[n], \tag{3.7}$$

$$D^{(l)}[k] = \sum_n g[n-2k]\, A^{(l-1)}[n]. \tag{3.8}$$

where :

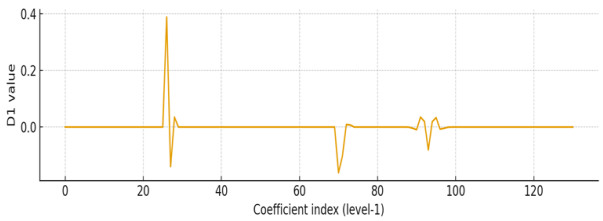
— $h[n]$ is the low-pass (scaling) filter,
— $g[n]$ is the high-pass (wavelet) filter,
— $A^{(0)}[n] = r[n]$.

The low-pass output $A^{(l)}$ retains the coarse-scale approximation, while the high-pass output $D^{(l)}$ captures the fine-scale details. Each filtering operation is followed by decimation by 2 to achieve dyadic scale separation [230].

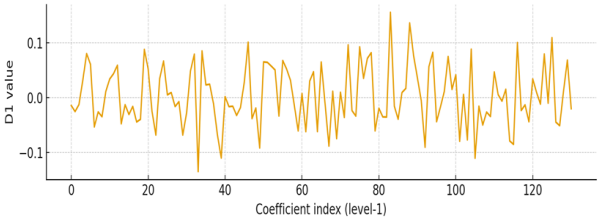
In this context, we selected the Daubechies-4 wavelet (denoted as "db4") with one decomposition level ($L = 1$). The db4 wavelet belongs to the Daubechies family of orthogonal wavelets characterized by maximal number of vanishing moments for a given filter length [231]. Specifically, db4 has four vanishing moments, allowing it to effectively represent smooth polynomial trends up to degree 3 and to capture localized discontinuities. This property makes db4 particularly well-suited for processing physiological signals such as HR or respiration, which exhibit smooth baseline trends punctuated by transient events. Using level 1 decomposition means that only the highest-frequency band is extracted, which is sufficient for isolating measurement noise from the residual.

The residual signal after trend extraction contains both the true abrupt physiological changes and the additive white Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. By applying a one-level wavelet decomposition to this residual, the detail coefficients $D^{(1)}[k]$ capture the high-frequency components at each transition index $k$. The amplitude of these coefficients indicates the strength of local rapid variations : true physiological transitions appear as large, isolated peaks in $|D^{(1)}[k]|$, while white Gaussian noise manifests as small-magnitude coefficients spread uniformly across the signal, as illustrated in Figure 3.3.

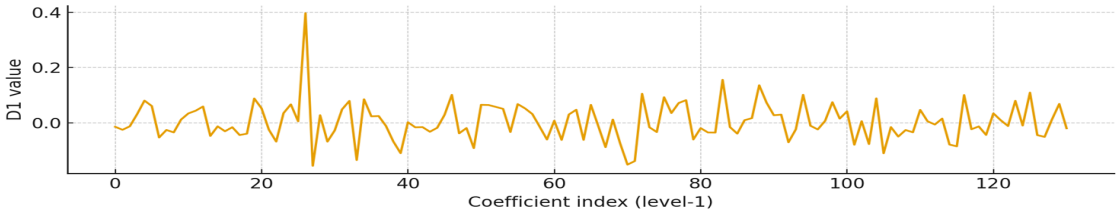
To distinguish meaningful transitions from noisy fluctuations, we apply soft thresholding using the universal threshold, which provides a statistically grounded estimate of the maximum amplitude expected from Gaussian noise, particularly as the number of samples increases (i.e., for longer patient stays).




(a) $D^{(1)}[k]$ coefficients highlighting abrupt changes in the signal.




(b) $D^{(1)}[k]$ coefficients of Gaussian noise only




(c) $D^{(1)}[k]$ coefficients of the residual signal

FIGURE 3.3 Wavelet detail coefficients $D^{(1)}[k]$ illustrating the presence of abrupt changes vs. noise.

The universal threshold is computed based on minimax risk principles (VisuShrink) [232] :

$$T = \sigma\sqrt{2\ln N}, \tag{3.9}$$

with $N$ the length of the original signal and $\sigma$ the standard deviation of the additive white noise. This threshold suppresses small coefficients, assumed to be noise, while retaining larger coefficients representing meaningful signal structures.

Under the assumption that noise predominantly manifests in the high-frequency bands, the standard deviation of the additive noise can be robustly estimated from the detail coefficients at the first scale via the Median Absolute Deviation (MAD) estimator as it will not be affected by outliers. To relate the standard deviation $\sigma$ to the Median Absolute Deviation (MAD), let $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and denote $m = \text{MAD}(|\varepsilon|) = \text{median}(|\varepsilon|)$. By definition,

$$\Pr(|\varepsilon| \leq m) = 0.5.$$

Introducing the standardized variable $Z = \varepsilon/\sigma$, we have $Z \sim \mathcal{N}(0, 1)$ and therefore

$$\Pr(|\varepsilon| \leq m) = \Pr(|Z| \leq m/\sigma) = 0.5.$$

Since the standard normal distribution is symmetric,

$$\Pr(-m/\sigma \leq Z \leq m/\sigma) = \Phi(m/\sigma) - \Phi(-m/\sigma) = 0.5,$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal. Using the identity $\Phi(-a) = 1 - \Phi(a)$, we obtain

$$2\Phi(m/\sigma) - 1 = 0.5 \quad \Rightarrow \quad \Phi(m/\sigma) = 0.75.$$

Taking the inverse CDF of both sides yields

$$\frac{m}{\sigma} = \Phi^{-1}(0.75) = 0.6745,$$

and therefore the standard deviation can be expressed in terms of the MAD as

$$\sigma = \frac{m}{0.6745} = \frac{\text{MAD}(|\varepsilon|)}{0.6745}.$$

This relationship ensures that the MAD is a robust and unbiased estimator of $\sigma$ under Gaussian assumptions [233].

Therefore the standard deviation of the $D^1[k]$ is given as :

$$\sigma = \frac{\text{MAD}\Big(|D^{(1)}[k]|\Big)}{0.6745}. \tag{3.10}$$

where the MAD of the first-level detail coefficients is defined as

$$\text{MAD}(D^{(1)}[k]) = \text{median}\Big( |D^{(1)}[k] - \text{median}(D^{(1)}[k])| \Big). \tag{3.11}$$

Since we assume that the noise component is approximately zero-mean Gaussian and the detail coefficients $D^{(1)}[k]$ are dominated by these noise coefficients (which form the majority), the median of $D^{(1)}[k]$ is effectively zero. In addition, the median operator is robust to outliers, so the few large coefficients corresponding to real abrupt physiological changes do not influence it significantly. Therefore,

$$\text{median}(D^{(1)}[k]) \approx 0, \tag{3.12}$$

and the MAD simplifies to

$$\mathrm{MAD}(D^{(1)}[k]) \approx \mathrm{median}\Big(|D^{(1)}[k]|\Big). \tag{3.13}$$

Therefore,

$$\sigma = \frac{\mathrm{median}\Big(|D^{(1)}[k]|\Big)}{0.6745}. \tag{3.14}$$

Then, soft thresholding is applied to the detail coefficients as follows :

$$\tilde{D}^{(1)}[k] = \begin{cases} \mathrm{sign}\Big(D^{(1)}[k]\Big)\,\Big(\,|\,D^{(1)}[k]|\,-\,T\Big), & \text{if } |\,D^{(1)}[k]| > T, \\ 0, & \text{otherwise.} \end{cases} \tag{3.15}$$

This operation suppresses insignificant components while preserving the continuity of significant coefficients and avoids the abrupt truncation of coefficients that occurs in hard thresholding due to discontinuity, thereby reducing Gibbs artifacts upon reconstruction.

The denoised residual is then reconstructed via the Inverse Discrete Wavelet Transform (IDWT) :

$$\hat{r}_{\mathrm{denoised}}[n] = \mathrm{IDWT}\left(A^{(1)},\, \tilde{D}^{(1)}\right). \tag{3.16}$$

Finally, to recover a refined estimate of the original process signal with both smooth trend and localized variations, the denoised residual is superimposed on the Kalman filter trend estimate :

$$\hat{x}_{\mathrm{final}}[n] = \hat{x}[n] + \hat{r}_{\mathrm{denoised}}[n]. \tag{3.17}$$

This two-stage Kalman–wavelet denoising framework is one of the key contributions of our work. It leverages the complementary strengths of both estimators : the Kalman filter optimally tracks the global temporal dynamics in the presence of stochastic measurement noise, while wavelet-domain thresholding efficiently attenuates residual, while preserving edges and other salient features and details. The resulting signal reconstruction exhibits improved smoothness and fidelity, making this approach well-suited for denoising biomedical signals and other time series with complex noise characteristics.

After preprocessing the physiological time series signals using Kalman filtering and wavelet denoising, multiple spectral features are extracted to comprehensively describe the frequency-domain characteristics, temporal complexity, and scale-dependent energy distribution of each signal. These features are important because they can capture aspects of signal behavior that are not apparent in the time domain alone, such as dominant periodicities, frequency spread,

and entropy, which are known to reflect physiological variability and regulatory mechanisms [234, 235].

To capture the frequency-domain characteristics of physiological signals, spectral features were extracted from the Discrete Fourier Transform (DFT) of the denoised signal, implemented efficiently via the Fast Fourier Transform (FFT). FFT decomposes the time series into its sinusoidal components, allowing for the analysis of periodicity and energy distribution across frequency bands. The FFT of a length-$N$ discrete signal $x[n]$ is given by :

$$X[k] = \sum_{n=0}^{N-1} x[n]\, e^{-j\,\frac{2\pi kn}{N}}, \tag{3.18}$$

where $k = 0, 1, \ldots, N-1$ denotes the frequency bins. From the Power Spectral Density (PSD) of this spectrum, several quantitative descriptors are derived. The PSD is computed as :

$$P[k] = \frac{2}{N} \Big|\, X[k] \,\Big|^2, \tag{3.19}$$

where the factor of 2 accounts for the symmetric nature of the real-valued time series (retaining only the positive frequencies). PSD shows how the signal average power is distributed over frequency components From the PSD, a set of informative features was derived, including the dominant frequency, average PSD, total spectral power, mean frequency, and spectral entropy. These features provide valuable insights into the rhythmic and oscillatory behavior of vital signs, such as HR variability or respiratory cycles. For example, dominant frequency can reflect underlying physiological rhythms, while spectral entropy quantifies the irregularity and complexity of the signal in the frequency domain. Normalizing total power by LOS enables comparisons across patients with different monitoring durations. By incorporating these spectral descriptors, models gain the ability to detect alterations in signal periodicity and complexity that may signal physiological deterioration, thus enhancing predictive performance in clinical outcome modeling.

— *Dominant Frequency* :

$$f_{\text{dom}} = f\Big(\arg\max_k P[k]\Big), \tag{3.20}$$

which identifies the most prominent oscillation in the signal. For example, in HR or respiration, it can correspond to the intrinsic cycle frequency or external pacing.

— *Mean Frequency* :

$$\bar{f} = \frac{\sum_k f[k]\, P[k]}{\sum_k P[k]}, \tag{3.21}$$

where $f[k]$ denotes the frequency corresponding to bin $k$, given by

$$f[k] = \frac{kF_s}{N} \tag{3.22}$$

where $F_s = 1$ as the data are hourly sampled. This feature represents the spectral centroid, indicating whether the signal's energy is concentrated at lower or higher frequencies.

— *Total Power* :

$$P_{\text{tot}} = \sum_k P[k], \tag{3.23}$$

representing the overall energy in the frequency domain, which can relate to signal variability.

— *Average PSD* :

$$\text{PSD}_{\text{avg}} = \frac{1}{N/2} \sum_k P[k]. \tag{3.24}$$

the mean power per frequency bin, useful for normalization and comparison across signals of different lengths.

— *Spectral Entropy* :

$$H = -\sum_k p[k] \log_2\Big(p[k]\Big), \quad p[k] = \frac{P[k]}{\sum_j P[j]}, \tag{3.25}$$

which quantifies the complexity or unpredictability of the frequency content. A flat spectrum (white noise) has high entropy, while a narrow band oscillation has low entropy. Spectral entropy has been shown to reflect autonomic regulation and disease severity [236].

These features collectively provide a succinct summary of signal periodicity, variability, and complexity in the frequency domain.

While the FFT effectively captures the frequency content of a signal, it lacks temporal resolution—it provides a global summary of which frequencies are present but not when they occur. This limitation is critical when analyzing physiological signals, which are inherently non-stationary and often contain transient events or shifts in frequency over time. For instance, a sudden change in RR, or an acute desaturation event may only occur over a short window, and such events are blurred or lost in the global frequency representation provided by FFT. Therefore, capturing the joint time-frequency behavior of the signal is essential for detecting clinically meaningful dynamics.

To address this, DWT Features were extracted. DWT plays a crucial role in analyzing phy-

siological time series, particularly vital signs, by enabling the decomposition of signals into both time and frequency domains. This dual representation is essential for capturing the inherently non-stationary and transient nature of clinical signals, enabling models to identify when specific frequency components appear, evolve, or vanish. Extracting features from wavelet channels—such as the approximation (low-frequency) and detail (high-frequency) coefficients—allows for a comprehensive characterization of both long-term trends and short-term variability. Features like energy, entropy, maximum amplitude, and transient counts provide insights into the signal's stability, complexity, and sudden fluctuations, which are often indicative of physiological deterioration or instability. By leveraging these multiscale, interpretable features, predictive models can detect subtle and clinically meaningful patterns that traditional statistical measures may overlook, ultimately enhancing the ability to forecast adverse events such as ICU or PICU readmission. We applied a one-level DWT to separate the signal into its low-frequency (approximation) and high-frequency (detail) components. This initial decomposition provides a big picture and a coarse view of the signal's spectral. From each set of coefficients $c[n]$, the following features are computed :

— *Mean and Standard Deviation* : summarizing the dispersion of wavelet coefficients.

— *Energy* :

$$E = \frac{\sum_n \Big(c[n]\Big)^2}{N}, \tag{3.26}$$

quantifying the contribution of each subband to the overall signal power.

— *Maximum Absolute Coefficient and its Time Index* ($k$) : this feature captures the most dominant variation within the signal. For the detail coefficients, the magnitude of the maximum absolute value indicates the strength of the most prominent abrupt change, while its index $k$ specifies the time at which this change occurs. For the approximation coefficients, the peak magnitude reflects the extent of the underlying trend shift (i.e., a sustained elevation or depression of the baseline), and the corresponding index $k$ identifies when this trend is most pronounced.

— *Shannon entropy of the coefficient magnitudes* :

$$H_c = -\sum_i p_i \, \log_2\Big(p_i\Big), \tag{3.27}$$

where

$$p_i = \frac{|\, c[i] \,|}{\sum_j |\, c[j] \,|}. \tag{3.28}$$

assessing the complexity of the coefficient distribution.

— *Transient Count* from detail coefficients : the proportion of coefficients exceeding twice

their standard deviation, reflecting abrupt events, such as physiologic bursts.

These features are computed separately for the approximation (low-frequency trend) and detail (high-frequency) components, enabling discrimination between baseline oscillations and transients.

To obtain a more refined representation of the signal's spectral content, we performed a two-level Wavelet Packet Transform (WPT), allowing us to analyze finer-scale frequency bands. This step effectively zooms in on different oscillatory modes and captures subtle variations that are not perceptible in a single-level DWT. Unlike the standard DWT, which decomposes only the approximation (low-frequency) coefficients at each level, WPT recursively decomposes both approximation and detail coefficients. Formally, WPT constructs a complete binary tree of subbands :

$$W_{j,k} = \text{WPT Node}(j, k), \tag{3.29}$$

where each node corresponds to a specific frequency band and its temporal localization, resulting in a richer and more balanced tiling of the time-frequency plane [237].

For each node in the wavelet packet tree, a set of statistical descriptors—including mean, standard deviation, energy, maximum absolute coefficient, and entropy—is computed. These features quantify both the intensity and variability of the signal across a finely partitioned range of frequency bands. By extracting such descriptors from each node, the model gains access to a rich multiresolution profile of the signal, capturing localized dynamics that may reflect physiologically important events—such as autonomic instability, respiratory dysregulation, or hemodynamic shifts. This more granular decomposition enables the detection of subtle spectral changes and transient anomalies that may be distributed across different frequency components, providing valuable insights into the temporal evolution of vital signs. Such detailed characterization is particularly beneficial in critical care, where early signs of deterioration are often embedded in complex and non-stationary physiological patterns.

When combined with standard DWT features—which are highly effective at capturing hierarchical trends and sudden transients—WPT features provide a complementary view by offering finer frequency resolution and better sensitivity to localized changes in physiological rhythms. Therefore, extracting features from both wavelet and wavelet packet transforms is essential for building robust models that can recognize diverse and nuanced patterns in critical care settings, ultimately improving the accuracy of predicting clinical deterioration or PICU readmission.

Finally, the temporal complexity of the signal is quantified by Sample Entropy (SampEn),

defined as :

$$\text{SampEn}(m, r) = -\log\left(\frac{A}{B}\right), \tag{3.30}$$

where $A$ is the number of matching vector pairs of length $m + 1$ within a tolerance $r$, and $B$ is the number for length $m$. Lower SampEn indicates more predictable, regular dynamics ; higher SampEn indicates higher irregularity. This measure is especially informative in HR variability studies [238].

Each feature captures a complementary aspect of the signal :

— Dominant and peak frequencies reveal core rhythmicity.
— Mean frequency and total power summarize energy distribution.
— Spectral entropy and wavelet entropy reflect complexity and disorder.
— Wavelet energies capture localized changes.
— Sample entropy measures temporal unpredictability.

Collectively, these features provide a rich representation of both periodic and nonstationary properties of biomedical signals.

We focused on extracting temporal and spectral features from the vital sign time series, as these signals inherently capture the dynamic physiological fluctuations of the patient. In contrast, other clinical time series such as medications, ventilation settings, and treatment interventions do not exhibit continuous dynamic variation and therefore do not yield meaningful spectral patterns. For this reason, we did not apply spectral feature extraction to those variables, as it would not contribute useful information to the model.

### Medical knowledge-based features

To enhance the interpretability and clinical relevance of the predictive models, we extracted a set of medical knowledge-based features that closely mimic the way physicians evaluate patient data over time. These features aim to capture not only statistical variability but also temporal and contextual patterns that often guide clinical decision-making. Extracting features that mimic physician interpretation of patient data improves model performance. Specifically, general features that capture how the measured variable changes over the entire ICU/PICU stay—such as *Is it abnormal ? (high or low ?)*—offer a direct indication of whether any abnormal values were observed. These features act as important markers of physiological instability throughout the patient's admission. Complementing these, we computed ratios and counts of abnormal values (e.g., *high_count*, *low_count*, *abnormal_ratio*) to quantify the prevalence and severity of deviations from normal ranges.

Admission-related features were derived from the first three hours of monitoring to cha-

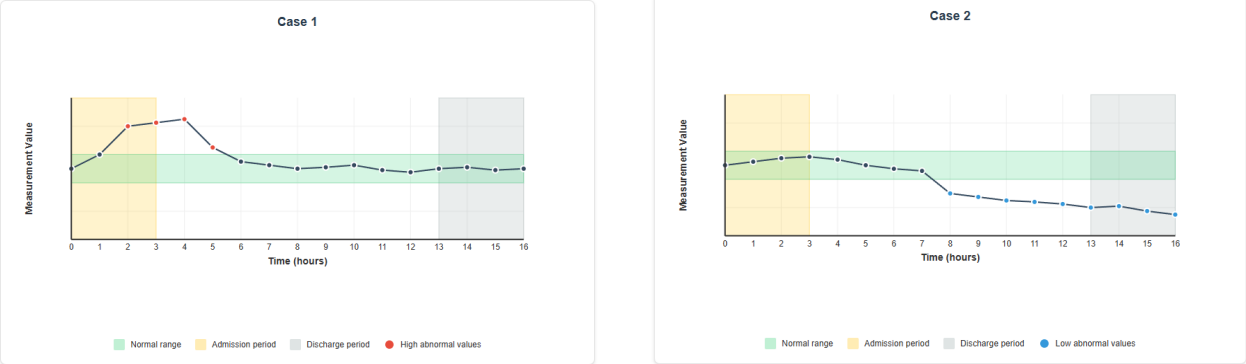
racterize the initial physiological state and potentially identify factors contributing to the admission itself. For example, *is_First_abnormal* and *is_First_high* record whether the initial measurements were already outside normal limits. Similarly, discharge-related features extracted from the last three hours (e.g., *is_last_abnormal*, *is_last_low*) are critical for identifying residual abnormalities that could increase the risk of readmission.

Time-based features further enrich this representation by explicitly modeling the temporal dynamics of recovery and deterioration. Recording the latency from admission to the first abnormal value helps distinguish whether an abnormality was likely a precipitating cause of admission or emerged as a complication. Measuring the interval between the first abnormal measurement and the first subsequent normal measurement, reflecting the patient's response to treatment. Capturing the duration between the last abnormal value and discharge, offering insight into whether the patient was discharged before fully stabilizing.

Additional features such as *max_high_diff* and *mean_low_diff* quantify the magnitude of deviation from normal reference ranges, while *num_changes* and *rate_of_change* capture temporal fluctuations and instability. For medications, we computed whether each drug or medication category (ex : Cardiovascular Medications and Antivirals) was administered during the admission window or discharge window, as well as times which indicate when the medication was initiated and when it was last administered relative to the hospitalization timeline. These time stamps are especially informative for understanding therapeutic interventions—early administration can signal initial severity or urgency, while recent use prior to discharge may indicate unresolved issues requiring ongoing management. Finally, for clinical scoring systems and categorical assessments (such as the GCS motor response), we derived severity indicators (e.g., Normal, Mild, Critical) at both admission and discharge.

Figure 3.4 illustrates two representative cases of time series measurements in relation to admission and discharge periods. In Case 1, abnormal high values occur primarily at the beginning of the timeline, suggesting these measurements may have contributed to the initial admission but were corrected before discharge. In contrast, Case 2 shows abnormal low values emerging during the discharge period, indicating potential unresolved clinical issues that could increase the risk of readmission. This comparison highlights the importance of considering the timing of abnormalities when assessing patient stability and discharge readiness.

In addition to this we extracted important medical scores such as APACHE, Logistic Organ Dysfunction score (LODs) [239], Modified Logistic Organ Dysfunction score (MLODS) [239], APS III, and Oxford Acute Severity of Illness Score (OASIS) [240] which are used to assess disease severity, predict patient outcomes, and guide clinical decision-making. APACHE evaluates patients' acute physiology, age, and chronic health conditions to predict morta-




(a) Case 1 (b) Case 2

FIGURE 3.4 Examples of measurement time series showing different patterns of abnormal values relative to admission and discharge periods.

lity risk. LODs assess organ dysfunction across six systems using physiological parameters, while MLODS refines this by adjusting variables and weights for better accuracy. APS III, part of APACHE III, considers more variables for improved prediction of ICU outcomes. OASIS is a simplified ICU severity scoring system that uses 10 easily available clinical variables—without requiring laboratory tests—to estimate the risk of in-hospital mortality. Developed to be quick and practical, OASIS offers predictive performance comparable to more complex scores like APACHE while reducing data collection burden. Each scoring system has its unique focus and calculation method, but they all aim to help clinicians better understand patients' conditions and make informed treatment decisions.

MIMIC-III contains rich clinical data in which diagnoses are coded using ICD-9-CM, while CathyDB uses the more recent ICD-10-CA codes to represent patient diagnoses. Because these coding systems contain thousands of highly granular codes, it is challenging to analyze them directly or to interpret models trained on such fine-grained inputs. To address this, we mapped all diagnosis codes to clinical classifications software categories. CCS is a classification scheme developed by the Agency for Healthcare Research and Quality (AHRQ) that groups individual ICD codes into a smaller set of clinically meaningful categories. For example, instead of dealing with hundreds of separate codes for different forms of heart failure, CCS aggregates them into a single "Heart failure" group. This mapping helps standardize the representation of diagnoses across datasets, simplifies analysis, reduces dimensionality, and improves the interpretability of predictive models by expressing each patient's diagnostic profile in terms of well-defined clinical categories rather than raw codes.

Extracting medical knowledge-based features is another key contribution of our work. This

enables the model to leverage structured expert assessments of patient status at key transitions of care. Collectively, these medical knowledge-based features complement statistical and spectral descriptors by encoding information in a form that is directly interpretable and actionable to healthcare professionals.

### 3.3.5 Feature selection

Given the large and heterogeneous pool of extracted features across all clinical domains ; including vital signs, laboratory measurements, medications, ventilations, treatments, and scores ; it was essential to apply a systematic feature selection strategy to identify the most informative and clinically relevant predictors for modeling readmission risk. The final dataset comprised thousands of candidate features, many of which were redundant, weakly informative, or noisy. To address this, we adopted a multi-step approach combining Mutual Information (MI) and L1-penalized logistic regression (LASSO).

First, for each variable category, features were pre-filtered using MI to select the top-$k$ features exhibiting the strongest dependency with the binary outcome variable (3-day readmission). MI is a powerful criterion because it quantifies any form of statistical dependence, including non-linear relationships, which are common in clinical data. By retaining only the most informative features at this stage, the dimensionality was substantially reduced while preserving relevant variability.

Second, we applied LASSO regression on the subset of top-$k$ features. LASSO imposes an L1 penalty on the magnitude of the coefficients, which has the effect of shrinking less important coefficients exactly to zero. This yields a sparse model that selects variables with the highest predictive contribution while effectively handling collinearity among correlated features. In this step, we retained features with non-zero coefficients, ensuring that the model remains interpretable and avoids overfitting.

Finally, a second MI ranking was applied to further refine the selection down to the final top-$n$ predictors with $MI > 0.001$. This step provided an additional safeguard against including features with spurious or unstable contributions.

Overall, this combination of MI and LASSO leverages both univariate and multivariate perspectives : MI ensures that non-linear and non-monotonic relationships are considered, while LASSO identifies the most robust multivariate predictors and promotes sparsity. This strategy balances predictive performance, parsimony, and interpretability. Additionally, applying this selection procedure separately within each variable category (e.g., labs, vitals, treatments) ensures that diverse types of information are retained, supporting models that reflect

the complexity of clinical decision-making.

### 3.3.6 Handling class imbalance

In this study, the prediction of unplanned ICU/PICU readmission constitutes a highly imbalanced classification problem, as the minority class (readmission within 3 days) represents only approximately 4% of the total samples. This severe imbalance can adversely affect model performance by causing predictions to be biased toward the majority (non-readmitted) class and diminishing the model's ability to correctly identify positive cases. To mitigate this challenge, we compared two strategies to improve minority class recognition. First, we trained models using cost sensitive learning with class weighting, assigning a higher penalty to misclassifying positive cases. This approach modifies the loss function to give more importance to the minority class during optimization, without explicitly generating synthetic samples. Second, we applied the SMOTE technique, which generates new synthetic samples of the minority class by interpolating between existing positive instances in feature space. This increases the representation of the minority class and helps the classifier better learn its characteristics. Figure 3.5 illustrates the concept of SMOTE, showing how synthetic examples are created between neighboring minority samples to augment the training data.

## 3.4 Models

This section focuses on the practical rationale for selecting and applying each model type within the context of PICU readmission prediction. To comprehensively assess predictive performance, we compared a diverse set of ML and DL models, spanning both classical algorithms and neural network architectures. We employed both linear and tree-based machine learning models to evaluate different types of predictive behaviors and feature interactions. Among the ML models, we evaluated LR, SVM, RF, XGBoost, and LightGBM. These me-

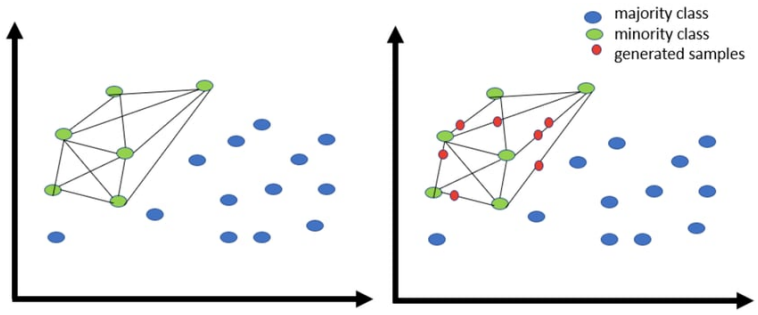



Figure 3.5 Illustration of the SMOTE algorithm

thods provide complementary strengths : for example, LR offers interpretability, SVM can capture complex decision boundaries. LR and SVM were selected as baseline linear classifiers due to their simplicity, interpretability, and effectiveness in high-dimensional spaces. Their performance provides insight into whether the relationship between input features and readmission risk is linearly separable, and they serve as a benchmark for more complex models.

To capture non-linear patterns and feature interactions, we also used ensemble tree-based models : RF, XGBoost, and LightGBM. These models are well-suited for handling structured clinical data, managing missing values, and ranking features based on importance. Their interpretability (via feature importance) also supports the clinical usability of our framework.

To model the temporal dynamics inherent in physiological time series, we used two types of deep learning architectures : LSTM , and Transformer. LSTM, which are well-suited to modeling temporal dependencies in sequential data and Transformer-Based Attention, which allows a model to weigh the importance of different input elements when encoding or generating sequences.

LSTM networks were chosen for their ability to learn long-range dependencies in sequential data. In our setup, LSTM processes sequences of hourly vital signs, fluid balance, ventilation settings, treatments and medications to capture the progression of patient status over time and the temporal trends indicative of patient deterioration and risk of readmission.

Transformers, known for their self-attention mechanism, were applied to model both short- and long-term dependencies without relying on recurrence. In our case, the Transformer architecture allows the model to dynamically attend to different parts of the patient's time series, enabling the model to flexibly relate each measurement to all others in the sequence and making it well-suited for detecting subtle yet important variations across different time windows.

To ensure robustness and comparability, all models were evaluated using the same training-validation-test split. In addition, ML models were subjected to cross-validation and ablation analysis. This comprehensive benchmarking allowed us to evaluate the relative effectiveness of traditional ML classifiers versus modern DL architectures for predicting unplanned PICU readmission.

### 3.4.1 Interpretability

Interpretability is critical in clinical predictive modeling because physicians must be able to understand, trust, and validate the factors influencing model decisions. Transparent models improve acceptance in practice and support shared decision-making by providing insight into

why a prediction was made. To enhance interpretability, we employed several complementary approaches for both local and global interpretability. For global interpretability, first, for tree based models, we used SHAP explanations, which quantify the marginal contribution of each feature to individual predictions based on cooperative game theory. Second, for LR models, we examined the learned coefficients, which directly indicate the direction and magnitude of each feature's association with the predicted risk. Positive coefficients increase the predicted probability of readmission, while negative coefficients decrease it. Finally, we assessed the stability and impact of each feature by iteratively removing features and observing the change in training and validation performance. This ablation analysis helped identify features that were essential to predictive accuracy and distinguish them from redundant or weak predictors. Together, these methods provided a comprehensive interpretability framework supporting the clinical relevance and credibility of the developed models.

For local interpretability, we used LIME which is a post-hoc technique that explains individual predictions of complex models by approximating their local decision boundary with a simple, interpretable surrogate model. For each patient in our dataset, LIME generates perturbed samples around the original feature vector, queries the black-box model for predictions, and fits a weighted linear model to highlight the most influential factors driving that specific decision. In our study, we extended LIME from the feature level to the clinical variable level, aggregating related features (e.g., multiple HR descriptors) into a single variable. This allowed us to visualize, for each patient, the clinical variables that most strongly supported the predicted outcome (readmission or non-readmission), thereby enhancing local interpretability in a clinically meaningful way.

SHAP and LIME methods highlight the most influential features driving the model's predictions. However, the extracted features (e.g., statistical or spectral descriptors) are often easier to interpret from an engineering perspective than from a clinical one.

To provide clinically meaningful insights, we aggregated the contributions of all extracted features back to their originating clinical variables. In other words, instead of reporting the importance of dozens of engineered features, we mapped their SHAP/LIME contributions to the underlying clinical measurements (e.g., heart rate, $SpO_2$, ventilation parameters). This aggregation enabled us to identify which clinical variables were most informative in driving the model's decisions.

This clinical-level interpretation is a key contribution of our work, as it bridges the gap between ML feature importance and clinical understanding. It allows clinicians to directly interpret and validate the model's reasoning using familiar clinical concepts, thereby enhancing transparency, trust, and potential clinical adoption.

### 3.4.2 Evaluation metrics

To comprehensively assess model performance, we computed a range of evaluation metrics capturing discrimination, calibration, and error trade-offs. We computed the AUROC and AUPRC metrics, which are widely used to evaluate classification performance in imbalanced clinical prediction tasks. The AUROC measures the ability of the classifier to rank positive instances higher than negative ones. Formally, AUROC is defined as the probability that a randomly chosen positive instance has a higher predicted probability than a randomly chosen negative instance. The AUPRC focuses on performance in the positive class and is particularly informative in imbalanced settings. Given the counts of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN), we computed Sensitivity (Recall) as :

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{3.31}$$

and Specificity as :

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{3.32}$$

Precision quantifies the proportion of predicted positives that are true positives :

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{3.33}$$

The F1 Score provides a harmonic mean of Precision and Recall :

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3.34}$$

Finally, Accuracy measures the proportion of correct predictions among all samples :

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{3.35}$$

In addition to conventional metrics, we defined a custom loss-based evaluation score (class-wise loss score $\mathcal{L}_{\text{score}}$) designed to reward models that achieve both low error and balanced performance across positive and negative classes. Specifically, the metric decomposes the log loss into separate contributions for positive and negative samples. Given true labels $y_{\text{true}}$ and predicted probabilities $\hat{y}$, the log loss for the positive class is computed as :

$$\text{PosLoss} = -\frac{1}{N_+} \sum_{i\,:\,y_i=1} y_i \log \hat{y}_i, \tag{3.36}$$

where $N_+$ is the number of positive samples. Similarly, the log loss for the negative class is :

$$\text{NegLoss} = -\frac{1}{N_-} \sum_{i\,:\,y_i=0} (1 - y_i) \log\Big(1 - \hat{y}_i\Big), \tag{3.37}$$

where $N_-$ denotes the number of negative samples. To convert these losses into bounded scores, each loss term is transformed as follows :

$$\text{PosScore} = \frac{1}{\Big(\text{PosLoss} + 0.1\Big) \times 10}, \tag{3.38}$$

$$\text{NegScore} = \frac{1}{\Big(\text{NegLoss} + 0.1\Big) \times 10}. \tag{3.39}$$

This scaling ensures that lower losses yield higher scores. Finally, the overall evaluation score combines these two components, averaging them and applying a penalty proportional to their imbalance :

$$\mathcal{L}_{\text{score}} = \frac{\text{PosScore} + \text{NegScore}}{2} \; - \; 0.2 \times \Big|\text{PosScore} \; - \; \text{NegScore}\Big|. \tag{3.40}$$

This formulation rewards models that simultaneously achieve low log loss for both classes while penalizing excessive discrepancy between positive and negative performance. The subtraction term encourages balanced calibration, reducing the risk of models that overfit to the majority class or systematically underperform on the minority class. The proposed class-wise $\mathcal{L}_{\text{score}}$ is one of the central contributions of our method, allowing the model to better handle imbalanced classes.

Using this diverse set of metrics allowed us to evaluate models not only in terms of overall correctness but also their ability to identify positive cases, which is essential in a clinical context with substantial class imbalance.

## 3.5 From Decision Scores to Probabilities

Many classifiers do not produce calibrated probabilities as part of training. Yet AUROC and AUPRC require continuous decision scores, and our custom class-wise loss score ($\mathcal{L}_{\text{score}}$) for linear models also expects a (probability-like) score as shown in equations (3.36) and (3.37).

Throughout, we use each model's decision function $s(x)$ as the primary continuous output at evaluation. AUROC and AUPRC depend only on the ranking of scores, so any monotone transform of $s(x)$ (e.g., a sigmoid) leaves them unchanged. When a probability is needed, we optionally map $s(x)$ to probability $\hat{p}(x)$ via a sigmoid function.

$$\hat{p}(x) = \sigma(s(x)) \tag{3.41}$$

Where :

$$\sigma(z) = 1/(1 + e^{-z}) \tag{3.42}$$

# CHAPTER 4 RESULTS

In this chapter, we address three objectives that together answer our research question. First, we establish a rigorous benchmark on the MIMIC-III database to develop a model architecture, feature extraction/selection strategies, and evaluation protocols. Second, we enhance and optimize this pipeline to the CHU Sainte-Justine PICU dataset (CathyDB), ensuring pediatric relevance by accounting for age-dependent physiology and clinical context. Third, we analyze both local and global interpretability and identify risk factors by explaining model predictions, thereby generating clinically meaningful insights and supporting real-world adoption.

## 4.1 Prediction of ICU Readmission Using LightGBM Classifier

We initially developed and evaluated ML model to predict ICU readmission using the publicly available MIMIC-III database. Using ML techniques reduce the model complexity compared to DL techniques. ML techniques are less time consuming and require less hardware building. In addition, ML facilitates the interpretation of the results obtained and thus the understanding of the key factors of the model [241]. Despite these advantages, the performance of these ML models is still insufficient compared to the requirements of ICU readmission decision-making systems. Therefore, we investigated new strategies to further improve this performance.

In this work, we study the efficiency of LightGBM classifier to predict 3 days ICU readmission using several features extracted from the clinical time series data. LightGBM is a classifier that was recently able to prove its high ability in the classification process, providing better performance compared to other ML techniques such as SVM, GBM, XGBoost [242].

### 4.1.1 Proposed approach

The End-to-end flow chart of the predicting ICU readmission pipeline is shown in Figure 4.1. We used the publicly available MIMIC-III database, as it was also used in prior works about ICU readmission, allowing us to compare our results with other methods. The required data were extracted using the proposed pipeline by Wang *et al.* [23] that address some challenges such as unifying measurement units and removing outlier values. A comprehensive set of variables encompassing demographic information, vital signs, categorical clinical variables (such as capillary refill rate, GCS eye opening, etc), laboratory test results, fluid balance

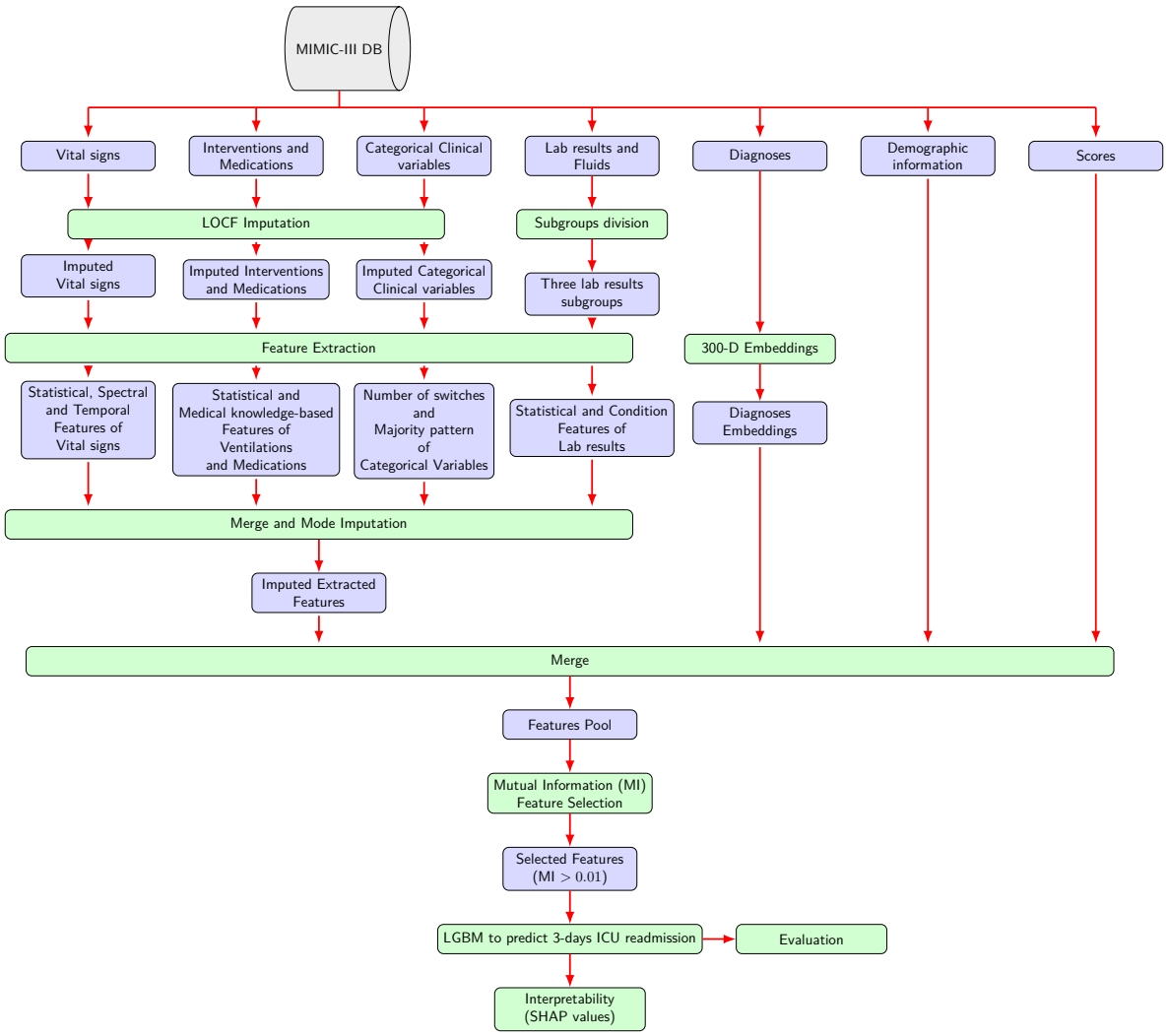



FIGURE 4.1 End-to-end flow chart of the predicting ICU readmission pipeline

measurements, ICU severity scores, mechanical ventilation parameters, and diagnostic codes based on the ICD-9 classification, were extracted. After applying exclusion criteria, we included the data of 31,151 patients, of whom 1,033 (3.3%) were readmitted to the ICU during the first 3 days of their discharge from the ICU.

EHR time series data such as vital signs, ventilations and categorical data have missing data. We handled the missing values using the following steps : If the values of the variable are fully missing, we imputed the normal values of this variable. If the variable contains some data, we replaced missing values at the beginning of the time series by the first recorded value, and for the other missing values, LOCF was used.

Then, various feature extraction methods were used to represent the information from the time series data. From the vital signs variables, we extracted the proposed statistical features.

To enrich temporal signal, we derived signal processing based descriptors using two different libraries : Time Series Feature Extraction Library (TSFEL) [243] and Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh) [244]. Using these libraries, we extracted temporal features such as energy and entropy as well as spectral features such as max frequency, in addition to the max correlation value between each two variables.

Then we represented the therapeutic interventions (such as ventilation, etc) by medical knowledge-based features such as the number of interventions used, the duration of use of each intervention, and the time between the last use of the intervention and the discharge time. The categorical data (such as capillary refill rate, etc) had been represented by two medical knowledge-based features ; the number of switches and the majority pattern presented by Lin, *et al.* [113]. Because diagnoses are key risk factors, we represented them with the pretrained 300 dimension embedding presented by Choi, *et al.* [118] to capture relationships among codes ; this could be effective for prediction, but less interpretable than code-level groupings.

For variables with high missing rates (notably laboratory tests are not measured frequently), we favored extracting simple statistics from observed values rather than aggressive imputation. We therefore divided them into three subgroups : the first subgroup contained laboratory results that were recorded at least three times for 70% of patients (such as hemoglobin, etc), the second subgroup contained the laboratory results that were recorded at least twice for 70% of the patients (such as calcium,etc), and the third subgroup contained the laboratory results that were not registered for most patients but were referred to in previous research as associated with ICU readmission (such as bilirubin, etc). In the third subgroup of the lab results, each variable was represented by the last value only, due to the few recorded values for those variables. The variables of the first and the second subgroups were represented by some simple statistical features and the first and the last value. In addition, the first and last values were converted to categorical states showing if that value was normal, high or low based on its normal range [245] as shown in Table 4.1, with an explanation of whether there was an improvement in the values of that variable or not as shown in Table 4.2. Finally, the missing values were replaced by the mode value.

TABLE 4.1 The categorical states for sodium (Normal range is 135-145)

| **Sodium Value** | **Low** | **Normal** | **High** |
|---|---|---|---|
| **130** | 1 | 0 | 0 |
| **140** | 0 | 1 | 0 |
| **150** | 0 | 0 | 1 |

TABLE 4.2 Measuring the extent to which the patient's variable values have improved

| First Value | Last value | Condition |
|---|---|---|
| **Normal** | **Normal** | Stable |
| **Abnormal** | **Normal** | Improved |
| **Abnormal** | **Abnormal** | Unstable |
| **Normal** | **Abnormal** | Deteriorate |

Extracting features from each variable led to a huge poll of features. To reduce the number of features, constant and quasi constant features were dropped, in addition to the duplicated and the high correlated features. Then we selected the features that has MI more than 0.01. The remaining features were combined together with the demographic and diagnosis data and used as the input to the LightGBM classifier.

### 4.1.2 Performance and results

We divided the cohort of 31,151 patients (1,033 readmitted, 3.3%) into two subsets : 80% for training (24,921 patients) and 20% for testing (6,230 patients). Stratified splitting ensures that the prevalence of readmission is preserved across sets, resulting in approximately 822 readmitted patients in the training set and 211 in the test set. Normalization was used to set the training values into range (0,1) and applied this later for the test split. For LightGBM, the following settings were used : learning rate of 0.04, boosting type was set to GOSS and to handle the imbalanced dataset, class weight parameter was used to give more attention to the samples of the minority class. Incorporating class-sensitive training (class weights) improved minority-class detection, thereby enhancing the model's ability to detect ICU readmissions. For the sake of comparison with previous works AUROC was used to measure performance. The proposed approach was able to distinguish between readmitted and not readmitted patients with AUROC of 78.6% showing a good performance compared with other ML methods that were used for predicting ICU readmission. To check the performance of our approach we compared it with the performance of other ML methods that were used in previous researches such as SVM, XGBoost and RF. In [113], they checked the performance of several ML models such as LR with L1 and L2 regularization, NB, SVM and RF. The best result was obtained using SVM with AUROC of 77.9%. In [168], they also compared the performance of different ML techniques such as KNN, RF and SVM. This time RF gives the best performance with AUROC of 73.8%. XGBoost was used in [114] given AUROC of 76%. Table 4.3, shows the comparison between the different ML models and it is obvious that our proposed algorithm do better than other ML algorithms. These results highlight the strength of LightGBM on

TABLE 4.3 Performance comparison

| Method | AUROC % |
|---|---|
| **LR-L2 [113]** | 77.3 |
| **LR-L1 [113]** | 77.7 |
| **NB [113]** | 70.9 |
| **SVM [113]** | 77.9 |
| **RF [168]** | 73.8 |
| **KNN [168]** | 59.3 |
| **XGBoost [114]** | 76.0 |
| **CNN+RNN [113]** | 79.1 |
| **Proposed Approach** | **78.6** |

tabular clinical data and the importance of carefully engineered features that expose trends, fluctuations, and recovery dynamics distinguishing readmitted from non-readmitted patients. Notably, LightGBM delivered performance comparable to more complex state-of-the-art DL model in [113], underscoring both the value of our feature set and the effectiveness of the selection strategy. This work was described in our published paper "Prediction of ICU Readmission Using LightGBM Classifier" (2023 IEEE 20th International Symposium on Biomedical Imaging, ISBI) [26].

### 4.1.3 Interpretability

We evaluated feature importance using SHAP values with the LightGBM model to better understand the factors influencing ICU readmission predictions. As shown in Figure 4.2, the most impactful features include several embeddings of the ICD9 diagnoses codes, which reflect specific comorbidities or clinical conditions associated with higher readmission risk. In addition, advanced signal processing features derived from DBP—such as FFT coefficients—were found to be highly informative, highlighting the value of frequency-domain analysis in capturing subtle physiological patterns. Other important predictors include statistical features from vital signs (e.g., HR, RR), lab results (e.g., blood urea nitrogen difference, white blood cell count), and medication usage (e.g., epinephrine administration). The distribution of top-ranked features illustrates the diverse nature of relevant inputs, spanning diagnostic codes, dynamic physiological trends, and treatment indicators—underscoring the need for a multimodal approach to ICU readmission prediction. However, while diagnosis embeddings enhanced discrimination, they offered limited transparency at the individual code level, complicating direct attribution to specific diagnoses or diagnostic categories.

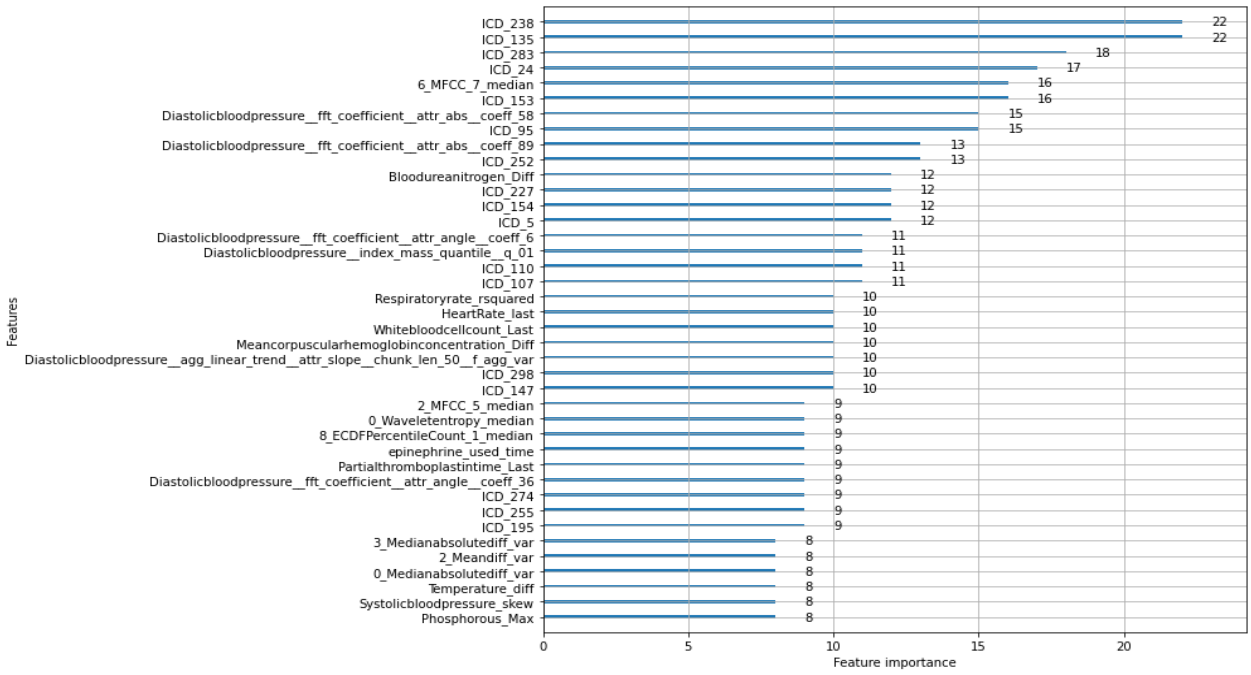



FIGURE 4.2 Top predictive features for ICU readmission

## 4.2 Interpretable Predictive Model for 3-days PICU Readmission

Encouraged by the promising ICU readmission results, we extended our approach to the PICU setting. Compared with MIMIC-III, CathyDB is smaller and contains fewer readmissions, which makes the task more challenging but also highlights the importance of problem-specific preprocessing and modeling. In this section, we present our approach for predicting PICU readmission using a combination of statistical, medical knowledge-based features and signal processing-based features extracted from patient time series data. To better capture the unique dynamics of pediatric cases and improve model performance, we introduced several modifications to the methodology used to predict ICU readmission, including enhanced preprocessing and feature engineering strategies.

### 4.2.1 Proposed Approach

We extracted patient data from the CathyDB database covering the period from January 1, 2014, to February 15, 2025. After applying the defined inclusion and exclusion criteria, a final cohort of 11,288 unique PICU admissions was identified. Among these, 323 patients (2.9%) were readmitted to the PICU within three days of discharge, which served as the primary target outcome for prediction.

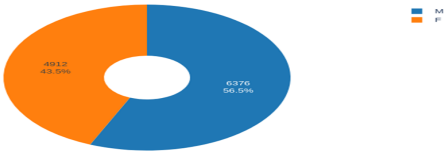



FIGURE 4.3 Gender distribution of the PICU cohort

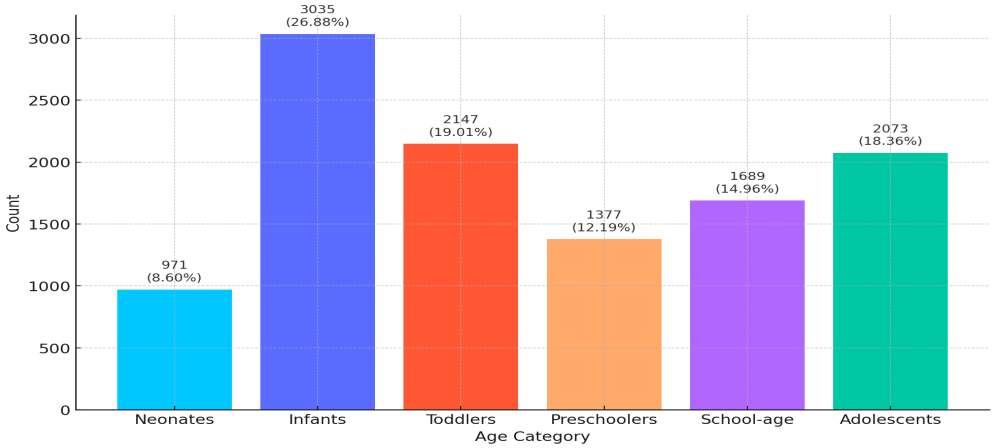



FIGURE 4.4 Age category distribution. Neonates : 0–28 days, Infants : 1–12 months, Toddlers : 1–3 years, Preschoolers : 3–5 years, School-age : 6–12 years, Adolescents : 13–18 years.

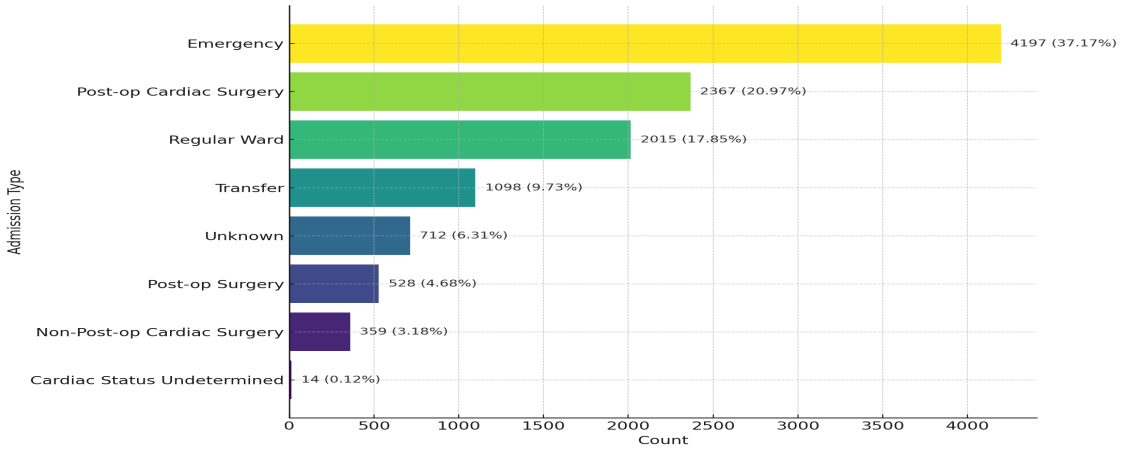



FIGURE 4.5 Admission type distribution

The cohort's demographic and clinical characteristics are summarized. As shown in Figure 4.3, the gender distribution is slightly skewed toward males, with 56.5% males (6,376 patients) and 43.5% females (4,912 patients). Figure 4.4 illustrates the age category distribution, where infants represent the largest group (27%), followed by toddlers (19%) and adolescents (18%). Figure 4.5 shows the distribution of admission types, with the majority being emergency admissions (37%) and Postoperative (Post-op)– cardiac surgery (21%).

To ensure fair evaluation and prevent data leakage, the dataset was temporally split into three splits. This time-based split simulates a real-world scenario where models are trained on past data and evaluated on future cases. The training split contains 7,758 patients, of whom 244 (3.1%) were readmitted. The validation split includes 1,693 patients with 46 readmissions (2.7%), and the test split comprises 1,694 patients with 30 readmissions (1.8%).

The end-to-end flow chart of the predicting PICU readmission pipeline is shown in Figure 4.6. From CathyDB we assembled a comprehensive, PICU-focused variable set that mirrors the major variable families in MIMIC-III while adding pediatric-specific detail. Physiologic time series include HR, RR, temperature, and other vital signs ; fluid balance records inputs and outputs (e.g., urine, blood products) and derived input/output ratios. The laboratory panel spans 51 tests (e.g., liver enzymes—ALkaline Phosphatase [ALP], ALanine Transaminase [ALT] ; hematology—Hemoglobin [Hgb], HematoCriT [HCT] ; blood gases). Diagnoses are encoded using 1,511 ICD-10-CA codes, providing a rich representation of patients' clinical conditions. Ventilatory support is captured more richly than in MIMIC-III, with explicit ventilation types (invasive, non-invasive) and 17 settings (e.g., $FiO_2$, tidal volume, PEEP). From the MIMIC-III database, we extracted only a limited set of medications that are commonly used with ventilation. The medication coverage extracted from CathyDB is broader than in MIMIC-III, 129 unique agents and administration routes (e.g., vasopressin, vancomycin via IV and PO, acetaminophen) were extracted. We also extracted treatment indicators and observed symptoms that are unavailable in MIMIC-III pipeline [23]. For clinical scoring, MIMIC-III contributed some adult-oriented scores, while CathyDB includes pediatric-specific sedation and neurologic scales. Sedation and neurologic status are recorded using eight pediatric scoring systems (e.g., CAPD, COMFORT-B, RASS). Demographics (age, gender, admission type) are included to complete the patient profile. These differences in pediatric coverage and granularity motivate tailored preprocessing and modeling for PICU readmission prediction.

We followed the preprocessing strategy outlined in the Methodology chapter to ensure consistency and reliability of the time-series data. First, we unified the time axis across all variables for each patient by aligning the start and end times based on the HR charttime. All extrac-

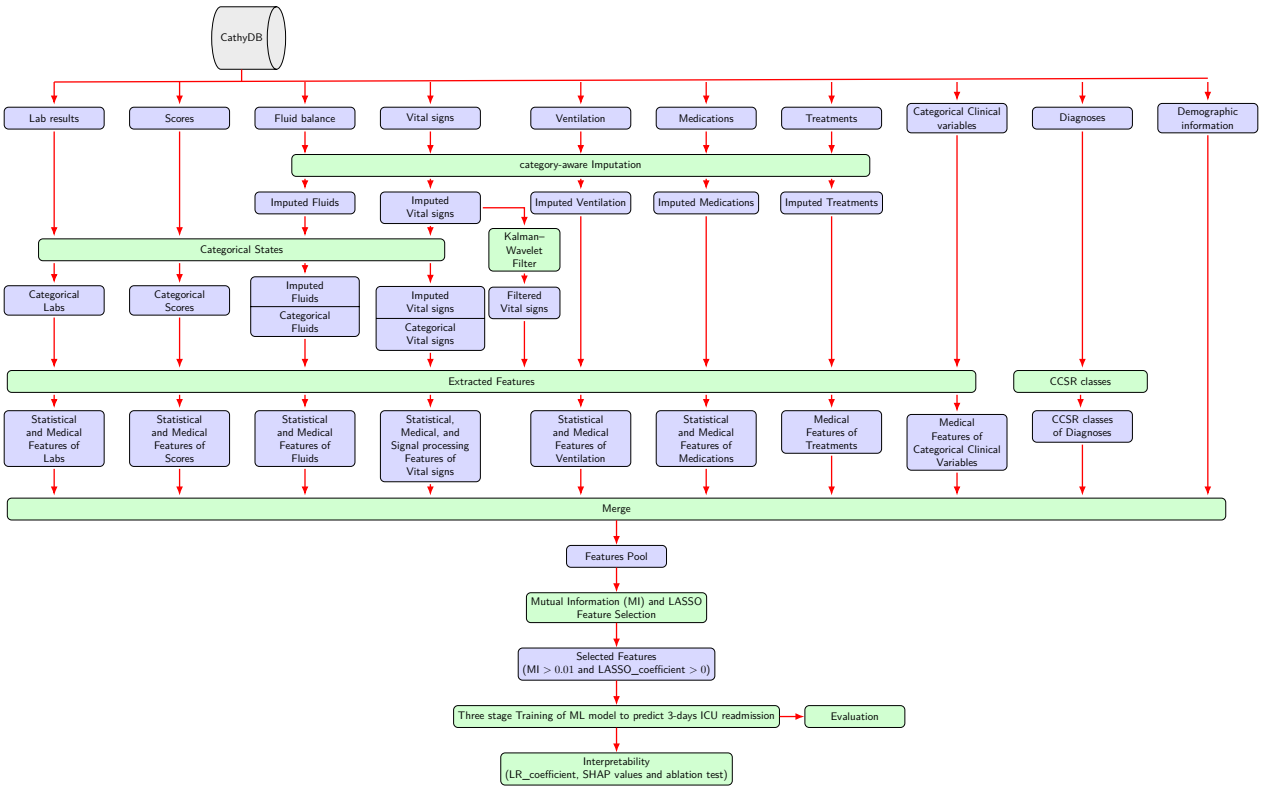



FIGURE 4.6 End-to-end flow chart of the predicting PICU readmission pipeline

ted time-series variables were then aggregated at an hourly resolution using the median value to reduce noise and variability. Pediatric laboratory reference intervals were taken from the Sainte-Justine Hospital laboratory reference sheet. For missing normal reference values, we computed and used the median value specific to each age category as a surrogate. Subsequently, we applied the imputation method proposed in the Methodology to handle remaining missing values effectively. In our work on ICU readmission, LOCF was used for all variable categories, whereas here the proposed, category-aware approach was used, selecting imputation methods appropriate to each variable clinical nature. After imputing missing values, we transformed the age-based time series variables into categorical states (normal, low, high) and computed the absolute difference from the age-specific normal range. This step is essential to address the variability in clinical interpretation of these variables across different pediatric age groups. Once the data was cleaned and standardized, we proceeded with the extraction of statistical and medical knowledge-based features from the original numerical values and the categorized states to serve as inputs to our prediction models.

In MIMIC-III database, predefined physiological ranges are available for each variable, which facilitates the removal of outliers during preprocessing. However, such ranges are not available

in the CathyDB, and the acceptable values may vary significantly across different pediatric age categories. To address this limitation and enhance the quality of the vital sign signals, we applied the two-stage filtering pipeline combining Kalman smoothing and wavelet denoising.

The effectiveness of this filtering method is demonstrated in Figure 4.7, which presents HR, RR, and SBP signals for two representative patients. In each case, the observed signal (green) exhibits high variability and noise, while the Kalman-filtered trend (blue, dotted) provides a smoothed approximation. The final denoised signal (red) captures the dominant pattern and removes sharp fluctuations likely to be artifacts. This preprocessing step was critical to ensure that the downstream feature extraction produced reliable statistical and signal-based descriptors. From the filtered vital sign signals, we extracted the proposed temporal and spectral signal processing features as detailed in the Methodology chapter. In our work on ICU readmission, we used automated libraries such as TSFEL and tsfresh to extract temporal and spectral features. However, for PICU readmission prediction, we chose not to rely on these libraries to extract features from the filtered and denoised signals, as they tend to produce an excessively large number of features—many of which are redundant or lack clinical interpretability. Instead, we extracted a focused set of specific features, as detailed in the Methodology chapter, selected specifically for their ability to capture meaningful temporal and spectral signal characteristics relevant to pediatric patients. These features enriched our overall feature pool with valuable information capturing the dynamic and frequency-based characteristics of patient physiology.

While incorporating the 300-dimensional diagnosis embedding significantly improved model performance for ICU readmission prediction, it offered limited interpretability. Although the model recognized diagnoses as key contributors, it could not clearly identify which specific diagnoses or diagnostic categories were most influential, limiting its clinical utility. To address this limitation in the PICU readmission task, we avoided using high-dimensional embeddings and instead represented diagnoses using 297 interpretable Clinical Classifications Software Refined (CCSR) categories. This approach groups ICD10-CA codes into interpretable diagnostic classes such as RSP009 : Asthma, CIR017 : Cardiac dysrhythmias, MAL007 : Respiratory congenital malformations, and RSP010 : Aspiration pneumonitis. By using these aggregated diagnostic classes, we enhanced the interpretability of the model while still preserving critical diagnostic information relevant to pediatric readmission risk.

After that, the extracted features were ranked using two complementary methods—MI and Lasso regression—as described in the Methodology chapter. This ranking strategy is more efficient than the MI approach used in our prior ICU readmission study. This ranking strategy was performed separately for each variable category (e.g., vital signs, lab results), allowing

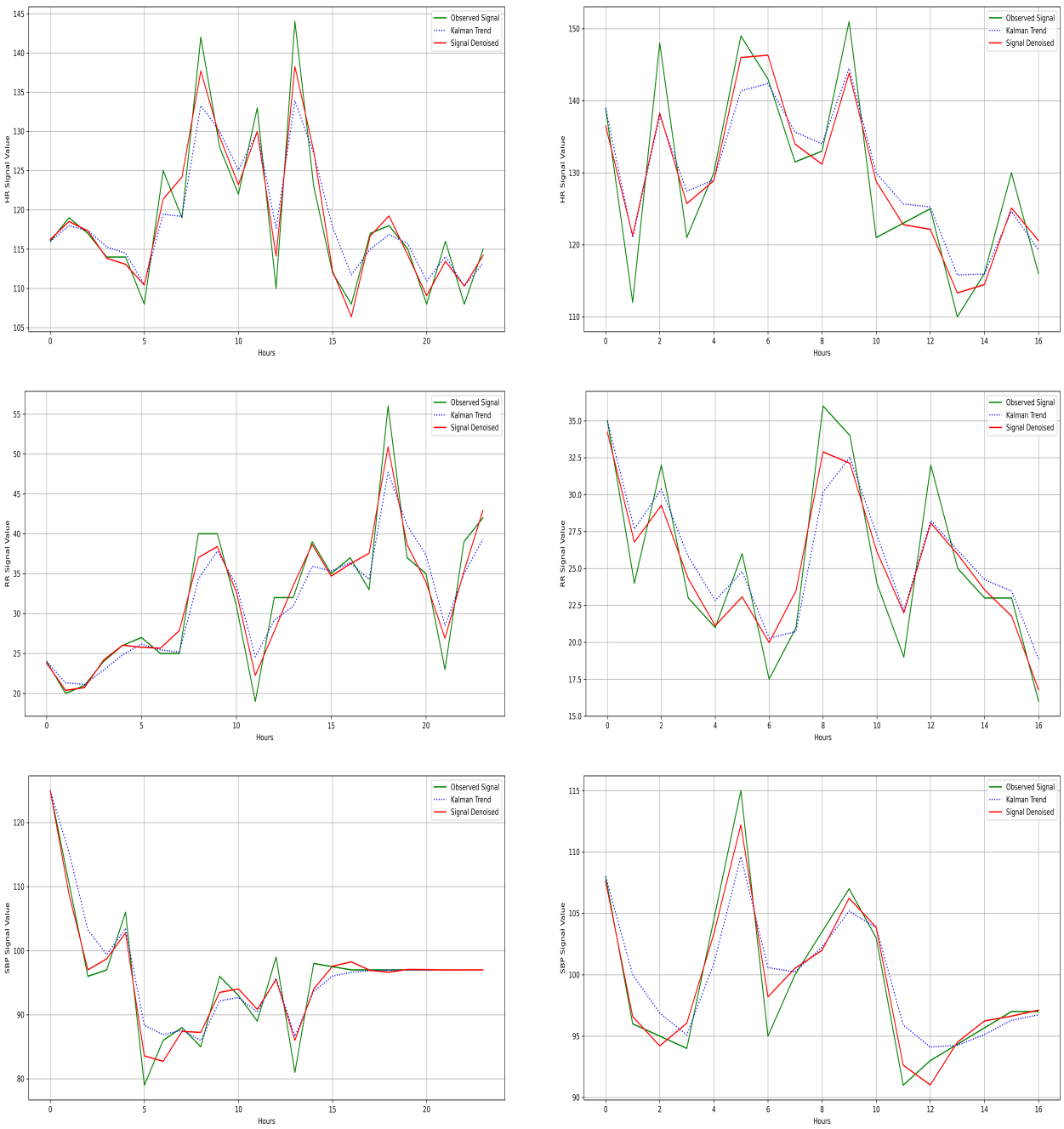



FIGURE 4.7 Filtered HR, RR, and SBP signals for patients 1 (left) and 2 (right). green : observed signal, blue : Kalman trend, red : final denoised signal.

us to later train models using features from all categories or individual categories if needed. Even after applying these ranking techniques, a substantial pool of top-ranked features $\mathcal{N}_{\text{pool}}$ are remained.

To train our model effectively, we adopted a three-stage protocol to couple hyperparameter optimization with wrapper-based feature selection and cross-validated refinement using the top ranked selected features only, aiming to maximize predictive utility while controlling variance and overfitting in a class-imbalanced PICU readmission setting.

To initialize model optimization, we formed a small seed set $S_0$ by taking the top-ranked features across major clinical categories (vital signs, medications, ventilation, fluids, and metadata), ensuring category diversity to reduce early selection bias. Using $S_0$, we tuned each model's hyperparameters $\theta$ with an objective aligned to the model family. For linear/large-margin models (LR, SVM, Ridge), the objective was our proposed custom class-wise loss score $\mathcal{O} = \mathcal{L}_{\text{score}}$ (higher is better) ; for tree-based and other ML models, the objective was the F1 score. Optimizing on a compact feature set stabilizes the search over $\theta$ by decoupling capacity control from the confounding effects of high-dimensional feature interactions. This procedure enabled us to train the model using systematically optimized parameters, thereby avoiding reliance on arbitrary or randomly chosen values.

With the best parameters $\theta^\star$ yielded from Stage 1, we expanded the candidate space by sampling randomly a larger, balanced pool $S_1$ from the top-ranked features while preserving representation from each category. We then performed a backward–forward (bidirectional) wrapper search : starting from $S_1$, we iteratively proposed removal (backward) and addition (forward) moves using the remaining features in our pool $\mathcal{N}_{\text{pool}}$ and scored each move on both the training and validation splits. The same objectives were used to guide the identification of the optimal subset of features that maximized performance on both the training and validation splits. A move was *accepted* only if it improved the *worst-split* objective by at least a tolerance $\varepsilon$, i.e.,

$$\min\{\mathcal{O}^{(\text{tr})}, \mathcal{O}^{(\text{val})}\} \text{ increased by } \geq \varepsilon.$$

The search terminated when no candidate move met this criterion. This mini-max acceptance rule ensures that features are added (or removed) only when they improve the *minimum* performance across the two splits, thereby discouraging overfitting to a single split. As a wrapper method, it evaluates conditional contributions given the current subset, addressing the limitations of univariate filters that ignore interactions and redundancy. This yields a subset $S_2$ that improved the performance across the two splits.

Feature selection performed on a single train/validation split can overfit the unique characteristics of that split, inflating apparent importance. This can lead to an overestimation of feature importance, as the selected features might perform well only on those particular splits and do not generalize. To mitigate selection bias and promote generalization, we merged the training and validation data and repeated the bidirectional search under stratified 5-fold cross-validation, keeping $\theta^\star$ fixed. Cross-validation allows for a more robust evaluation by training and validating the model across multiple data partitions, thereby reducing the potential bias introduced by fixed train-validation splits. Within this cross-validated refinement, at each add/drop proposal using the remaining features in our pool $\mathcal{N}_{\text{pool}}$, we evaluated the move on all folds using AUROC and accepted it only if it improved cross-validated performance while stabilizing it across folds. Concretely, a move was accepted when the mean AUROC increased by at least $\varepsilon$ and the across-fold standard deviation did not increase. The procedure terminated when no candidate move met these criteria, yielding the final subset $S_3$, which improved 5-fold AUROC (and maintained the Stage-2 objectives). The model was then retrained on the train split using $S_3$ and $\theta^\star$, and evaluated once on the held-out test split.

As a final important step, we conducted an ablation analysis and applied model-specific importance diagnostics (e.g., coefficient magnitudes for linear models ; SHAP for tree models). This corroborates the contribution of the selected features and identifies the most influential features contributing to the model's predictive performance. This three-stage design : seeded hyperparameter tuning, conditional (wrapper) selection, and cross-validated refinement, offers greater stability than a MI only pipeline. Unlike MI ranking, which ignores redundancy and feature interactions, the wrapper stage evaluates conditional contributions, while cross-validation ensures generalizability. This reduces variance, mitigates overfitting, and preserves strong predictive performance.

### 4.2.2 Performance and results

To evaluate model performance, we focused primarily on two metrics : Recall and AUPRC. These metrics were selected for the following reasons :

— The dataset is highly imbalanced, AUROC can be misleading as it may remain high even when the model performs poorly on the minority class, since it evaluates overall ranking performance without considering class distribution. In contrast, AUPRC directly reflects how well the model identifies positive cases by focusing on precision and recall, both of which are sensitive to the proportion of true positive predictions. Therefore, AUPRC provides a more informative and clinically meaningful evaluation

of our model's ability to detect readmissions under severe class imbalance, and is thus chosen as the primary performance metric.

— Identifying true positive cases (i.e., patients at risk of readmission) is clinically critical to minimize adverse outcomes.

In addition, to validate that our models perform meaningfully and are not random, we compared their AUPRCs with the baseline AUPRC. The baseline AUPRC represents the expected performance of a random classifier, which is equivalent to the prevalence of the positive class in the dataset. This comparison helps ensure that the model's performance is significantly better than what would be expected by chance alone. The baseline AUPRC is calculated as follows :

$$\text{Baseline AUPRC} = \frac{P}{P+N} \tag{4.1}$$

where :

— $P$ : number of positive samples,

— $N$ : number of negative samples,

— $P + N$ : total number of samples.

The baseline AUPRC for the three splits are :

— Training split baseline AUPRC : $\frac{244}{244+7638} \approx 3.1\%$

— Validation split baseline AUPRC : $\frac{46}{46+1647} \approx 2.7\%$

— Test split baseline AUPRC : $\frac{30}{30+1664} \approx 1.8\%$

In our experiments, we first investigated the impact of the proposed extracted features, which include both medical knowledge-based features and signal processing-based features, on model performance. We then compared the performance of our proposed model, which uses LR with the extracted features, against other linear and standard ML models to assess its relative effectiveness. Additionally, we evaluated our model's performance against tree-based models to examine whether more complex algorithms could offer improvements. Finally, we compared the results with deep learning models to determine whether these advanced techniques could outperform traditional machine learning approaches in this context.

**Effect of medical knowledge-based features and signal processing-based features**

To assess the contribution of the proposed extracted feature, we performed a stepwise evaluation using LR across three feature configurations :

1. Statistical features only (ST) : Standard descriptive statistics derived from time-series

data.

2. Statistical + Medical knowledge-based features (ST+MED) : Adding extracted medical knowledge-based features that mimic physician decision.
3. All features (ST+MED+SP) : Including hand-crafted signal processing features such as wavelet and FFT features.

All models were compared at a consistent cross-validation AUROC (CV_AUROC) performance level, specifically when their average CV_AUROC reached approximately 80%.

The results, summarized in Table 4.4, demonstrate a clear and consistent improvement in model performance as more informative features were added. In addition, in all scenarios, the test AUPRC scores exceeded the baseline AUPRCs computed from the class distribution.

When using only statistical features, the model achieved moderate predictive performance, with clear evidence of overfitting. As shown in the corresponding Table 4.4 and illustrated in Figure 4.8, AUPRC, Recall and AUROC scores dropped noticeably from the training split to the test split. This decline highlights limited generalizability and suggests that statistical features alone may not capture the complexity of the patient trajectories needed to robustly predict PICU readmission. The performance gap across data splits indicates the need for more informative and diverse features to enhance stability and accuracy. The maximum CV_AUROC achievable using only statistical features was 79.4%.

The addition of the medical knowledge-based features to the statistical baseline led to a notable improvement in model performance, as shown in Figure 4.9. AUPRC, Recall and AUROC scores increased across all data splits, indicating that medical knowledge-based features provides complementary predictive information. However, despite this improvement, a performance drop from training to test splits remained especially in Recall, suggesting that

Table 4.4 Performance comparison across feature configurations using LR

| **Metric** | **ST** | | | **ST + MED** | | | **ALL** | | |
|---|---|---|---|---|---|---|---|---|---|
| | Train | Val | Test | Train | Val | Test | Train | Val | Test |
| AUPRC (%) | 17.6 | 13.6 | 3.9 | 18.3 | 18.5 | 7.3 | 23.4 | 22 | 13.8 |
| Recall (%) | 74.9 | 58.7 | 53.3 | 79.4 | 73.9 | 60 | 86.2 | 71.7 | 83.3 |
| AUROC (%) | 84.1 | 78.2 | 63.5 | 85.3 | 86.4 | 79.3 | 89.1 | 88.2 | 87.8 |
| Accuracy (%) | 79.1 | 82.1 | 79.3 | 77 | 81.7 | 78.8 | 79.2 | 82.5 | 81.8 |
| F1 score (%) | 18.3 | 15.1 | 8.4 | 17.7 | 18 | 9.1 | 20.6 | 18.2 | 13.9 |
| Specificity (%) | 79.3 | 82.8 | 79.8 | 76.9 | 81.9 | 79.1 | 79 | 82.8 | 81.7 |
| Precision (%) | 10.4 | 8.7 | 4.5 | 10 | 10.2 | 4.9 | 11.7 | 10.4 | 7.6 |
| CV_AUROC (%) | 79.4 | | | 79.7 | | | 80 | | |

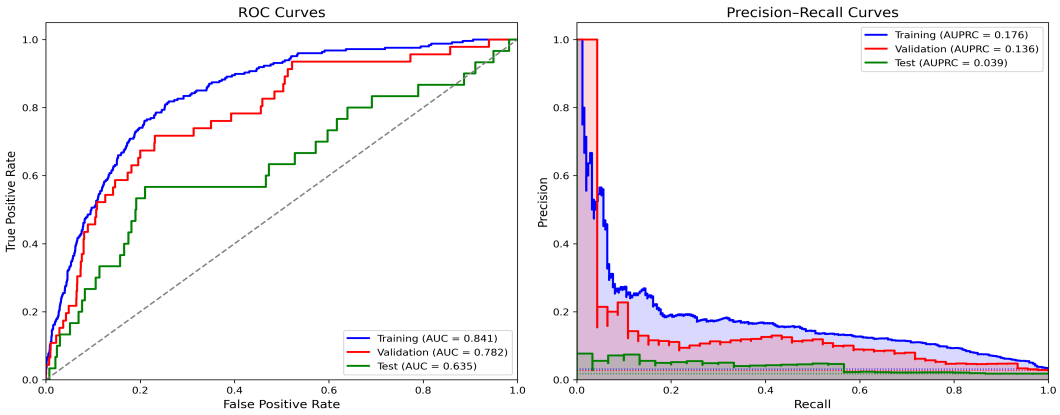



FIGURE 4.8 AUPRC and AUROC performance of the LR model using only statistical features

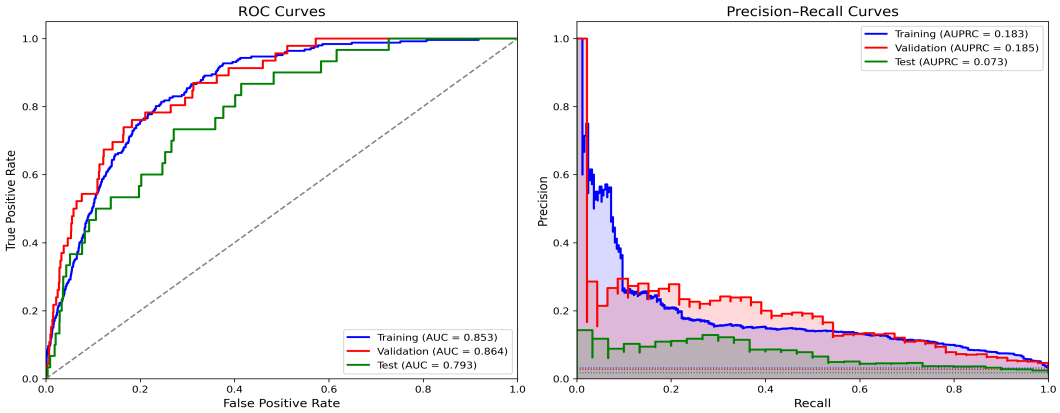



FIGURE 4.9 AUPRC and AUROC performance of the LR model using statistical and medical knowledge-based features

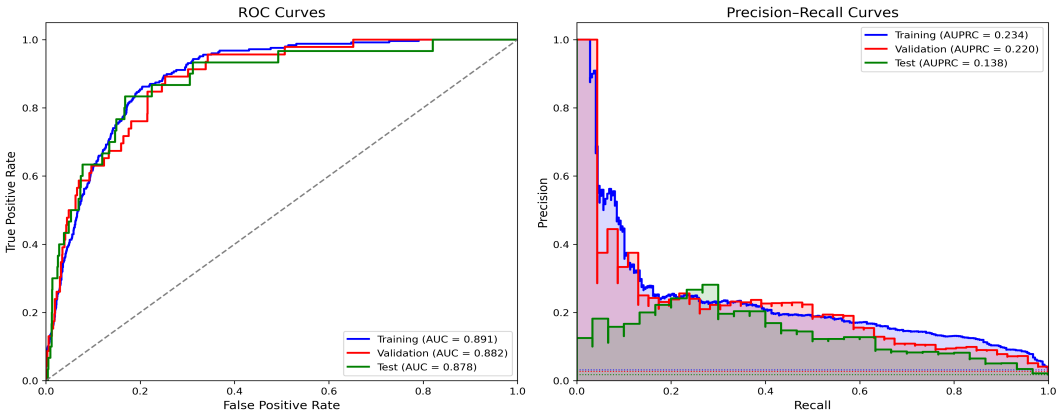



FIGURE 4.10 AUPRC and AUROC performance of the LR model using all features

the model still faces challenges in generalizing to unseen data. This implies that while medical knowledge-based features enhance model capacity, further feature enrichment is necessary to improve robustness and mitigate overfitting.

The integration of signal processing based features—such as wavelet, and entropy-based measures—resulted in a substantial improvement in model performance, in terms of AUPRC, Recall and AUROC. As shown in Figure 4.10, the LR model achieved significantly higher predictive scores across all data splits, with a remarkable increase on the test split compared to earlier feature splits. Moreover, the performance became more stable and consistent across training, validation, and test splits, indicating better generalizability and reduced overfitting. These results highlight the critical value of temporal dynamics and physiological signal patterns and the robustness of our LR model in effectively identifying the minority class, making it better suited for the imbalanced nature of our dataset.

**Performance comparison of the proposed model with alternative linear models**

We conduct a comparative analysis of the performance of our proposed LR model against other widely used linear models, specifically SVM and Ridge classifier. The optimized parameters $\theta^\star$ for each model are detailed in Table 4.5, and the corresponding performance metrics are provided in Table 4.6. This comparison serves to critically evaluate the efficacy of our proposed LR model, highlighting its advantages and potential limitations relative to other linear approaches in the context of predicting 3-days PICU readmission.

The parameters presented were optimized to strike a balance between model complexity and class imbalance. The C parameter (in LR and SVM) and alpha (in Ridge) control the strength of regularization, with smaller values enhancing regularization to prevent overfitting and improve generalization. The penalty (l2) in LR applies Ridge regularization, penalizing large coefficients and promoting model simplicity. The kernel (linear) in SVM defines a linear decision boundary, appropriate for linearly separable data. Furthermore, the class_weight parameter was adjusted to address the class imbalance, assigning greater importance to the minority class (readmitted patients) to improve model sensitivity to this critical group. These parameter selections were essential in optimizing model performance, ensuring both robustness and the ability to handle imbalanced data effectively.

SVM learns a maximum–margin separating hyperplane. With a linear kernel (no feature mapping), the learned boundary is a linear hyperplane and the decision function is $s(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ (signed distances to the hyperplane, up to a scale), and the classifier predicts $\hat{y} = \text{sign}\Big(s(\mathbf{x})\Big)$. For evaluation, we used the raw decision scores $s(\mathbf{x})$ and applied a monotonic logistic mapping to obtain pseudo–probabilities, $\hat{p} = \sigma\Big(s(\mathbf{x})\Big)$. This post–hoc transform does

TABLE 4.5 Models' optimized parameters

| Model | Parameters |
|---|---|
| **LR** | C : 0.005, penalty : l2 regularization<br>solver : liblinear, class_weight : {0 : 1, 1 : 30} |
| **SVM** | C : 0.005, kernel : linear<br>decision_threshold : {0 : 1, 1 : 30} |
| **Ridge classifier** | alpha : 0.1, class_weight : {0 : 1, 1 : 30} |
| **XGB** | n_estimators : 100, learning_rate : 0.05<br>max_depth : 3, min_child_weight : 100<br>subsample : 0.5, colsample_bytree : 0.5<br>scale_pos_weight : 30 |
| **LightGBM** | boosting_type : gbdt, n_estimators : 200<br>learning_rate : 0.01, num_leaves : 4<br>max_depth : 3, min_child_samples : 100<br>subsample : 0.5, colsample_bytree : 0.5<br>class_weight : {0 : 1, 1 : 30} |
| **RF** | n_estimators : 100, max_depth : 2<br>max_features : log2, criterion : gini<br>min_samples_split : 30, min_samples_leaf : 10<br>class_weight : {0 : 1, 1 : 30} |
| **ET** | n_estimators : 400, max_depth : 3<br>min_samples_split : 50, min_samples_leaf : 10<br>max_features : None, class_weight : {0 : 1, 1 : 30} |
| **CatBoost** | iterations : 300, learning_rate : 0.001<br>depth : 7, l2_leaf_reg : 2<br>subsample : 1, class_weights : [1, 30] |
| **KNN** | n_neighbors : 101, weights : distance<br>metric : minkowski, class_weight : {0 : 1, 1 : 30} |
| **NB** | var_smoothing : 1e-9, class_weight : {0 : 1, 1 :1} |
| **BILSTM** | Loss : BinaryCrossentropy, lr : 0.001<br>dropout : 0.3, bilstm_hiddden : 32<br>static_Dense_hidden : 32, FC_hiddens : [64, 32, 16]<br>batch_size : 64, epochs : 50, patience : 10 |
| **Transformer** | Loss : BinaryCrossentropy, lr : 0.001<br>dropout : 0.3, embedding_dim : 64<br>static_Dense_hidden : 64, FC_hiddens : [16, 8]<br>num_heads : 1, intermediate_dim : 128<br>batch_size : 64, epochs : 50, patience : 10 |

TABLE 4.6 Performance comparison of linear models

| **Metric** | **LR** | | | **Ridge** | | | **SVM** | | |
|---|---|---|---|---|---|---|---|---|---|
| | Train | Val | Test | Train | Val | Test | Train | Val | Test |
| AUPRC (%) | 23.4 | 22 | 13.8 | 14.5 | 15 | 6.5 | 17 | 16.2 | 5.3 |
| Recall (%) | 86.2 | 71.7 | 83.3 | 77.7 | 52.2 | 56.7 | 83 | 50 | 43.3 |
| AUROC (%) | 89.1 | 88.2 | 87.8 | 86.8 | 80.4 | 80.3 | 89 | 83.3 | 75.4 |
| Accuracy (%) | 79.2 | 82.5 | 81.8 | 78.3 | 81.9 | 80.8 | 87.9 | 89.3 | 87.1 |
| F1 score (%) | 20.6 | 18.2 | 13.9 | 18.3 | 13.5 | 9.4 | 30 | 16.9 | 10.4 |
| Specificity (%) | 79 | 82.8 | 81.7 | 78.4 | 82.7 | 81.2 | 88.2 | 89.6 | 87.5 |
| Precision (%) | 11.7 | 10.4 | 7.6 | 10.4 | 7.8 | 5.2 | 18.3 | 10.5 | 5.9 |
| CV_AUROC (%) | 80 | | | 80 | | | 79.9 | | |

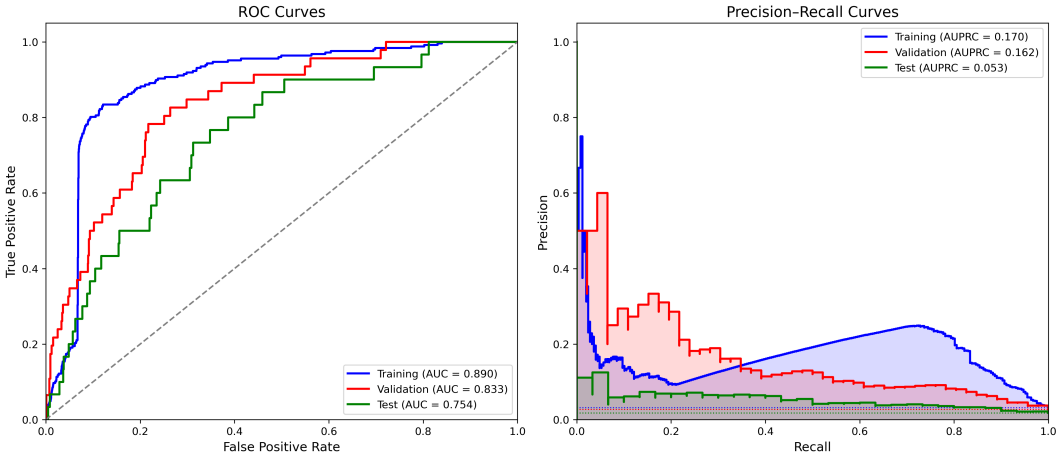



FIGURE 4.11 AUPRC and AUROC performance of the SVM model

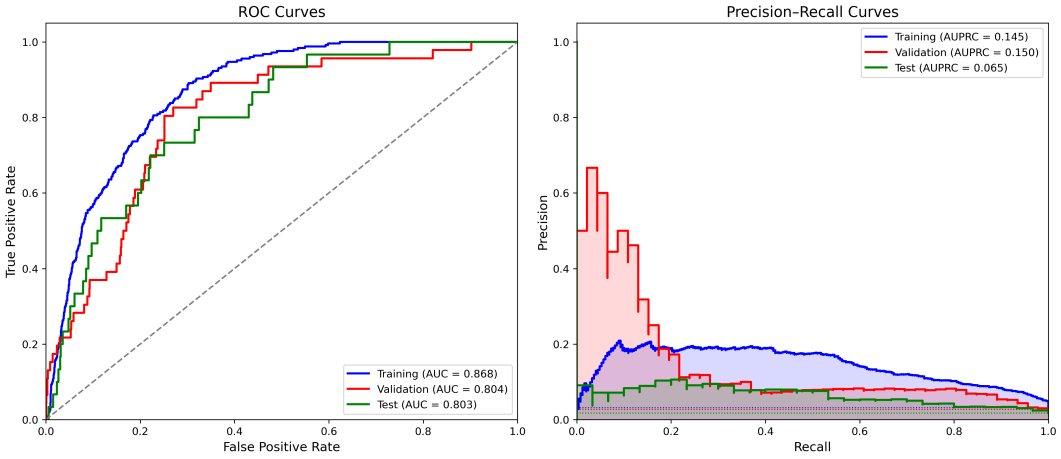



FIGURE 4.12 AUPRC and AUROC performance of the Ridge classifier

not affect training ; it simply facilitates probability–based metrics (e.g., AUROC, AUPRC).

The SVM shows strong apparent discrimination on the training and validation splits, with high AUPRC and AUROC as shown in Figure 4.11. Training AUROC/accuracy are high (AUROC = 89.0%, Acc = 87.9%) with recall of 83.0%, yet the validation split already reveals a sharp recall drop to 50.0% (AUROC = 83.3%, AUPRC = 16.2%), and the test split deteriorates further (Acc = 87.1%, recall = 43.3%, AUROC = 75.4%, AUPRC = 5.3%). The large train–test AUROC gap (−13.6%) and the collapse in AUPRC indicate overfitting and sensitivity to class imbalance ; high accuracy mainly reflects the majority class rather than true positive recovery. The test AUROC is lower than the CV_AUROC (79.9%), reinforcing limited generalization. Overall, while SVM achieves high accuracy, it struggles with class imbalance and generalization, particularly in minority-class prediction, compared with LR.

To calculate probabilities for Ridge classifer, at prediction time, we used the raw linear score $s(x)$ directly instead of $\hat{y} = \text{sign}(s(x))$. Then probability is measured by applying sigmoid function to the score and used as the hard label $\hat{y} = \sigma(s(x))$. This does not affect the training process of the Ridge classifier.

The Ridge classifier is more stable than the SVM but not the strongest on the minority class. Its AUROC declines only modestly from train to validation to test (86.8% → 80.4% → 80.3%) as shown in Figure 4.12, and the cross-validated AUROC (∼80.0%) aligns, indicating good generalization and low variance. Despite these improvements, Ridge's AUPRC (6.5%) on the test split still reflected a weaker ability to identify the minority class compared to LR. Accuracy around 81% is largely driven by the majority class and is less informative under imbalance. Compared with SVM, Ridge generalizes better and attains higher test recall and AUPRC (56.7% and 6.5 vs. 43.3% and 5.3), but overall it still underperforms LR on minority-class capture.

LR emerged as the most robust model, with superior recall and AUPRC for the imbalanced task. SVM achieved competitive accuracy but weaker recall, AUPRC, and cross-validation AUROC, indicating poor generalization. Ridge was more stable but consistently underperformed LR across all metrics.

### Performance comparison of the proposed model with other ML models

We compared our proposed LR model with other standard ML models such as KNN and NB. For these models, the optimized parameters $\theta^\star$ are as shown in Table 4.5. When comparing their performance with our proposed model, several important observations can be made regarding their performance across the three splits as shown in Table 4.7.

TABLE 4.7 Performance comparison of LR, KNN, and NB models

| **Metric** | **LR** | | | **KNN** | | | **NB** | | |
|---|---|---|---|---|---|---|---|---|---|
| | Train | Val | Test | Train | Val | Test | Train | Val | Test |
| AUPRC (%) | 23.4 | 22 | 13.8 | 15.9 | 15.7 | 3.9 | 10.6 | 12.8 | 4 |
| Recall (%) | 86.2 | 71.7 | 83.3 | 44.9 | 37 | 26.7 | 80.2 | 80.4 | 53.3 |
| AUROC (%) | 89.1 | 88.2 | 87.8 | 81.5 | 74.8 | 65.1 | 80.4 | 85.2 | 69.7 |
| Accuracy (%) | 79.2 | 82.5 | 81.8 | 90 | 91 | 89.9 | 69 | 73.3 | 73.3 |
| F1 score (%) | 20.6 | 18.2 | 13.9 | 21.9 | 18.2 | 8.5 | 13.9 | 14.1 | 6.6 |
| Specificity (%) | 79 | 82.8 | 81.7 | 91.4 | 92.5 | 91 | 68.6 | 73.1 | 73.2 |
| Precision (%) | 11.7 | 10.4 | 7.6 | 14.5 | 12.1 | 5.1 | 7.6 | 7.7 | 3.5 |
| CV_AUROC (%) | 80 | | | 72.7 | | | 80.3 | | |

KNN is a non-parametric, instance-based classifier : it learns no global model, and predicts a label for a query $\mathbf{x}$ from its local neighborhood under a chosen distance ( feature scaling is therefore important). the class of a sample is determined by the majority class of its nearest neighbors. Specifically, the number of neighbors belonging to each class is computed as follows :

$$\text{num\_0} = \sum(\text{neighbor\_labels} == 0) \tag{4.2}$$

$$\text{num\_1} = \sum(\text{neighbor\_labels} == 1) \tag{4.3}$$

If num_0 > num_1, the sample is classified as class 0 ; otherwise, it is classified as class 1. However, in the context of imbalanced datasets, the number of neighbors from the majority class (num_0) is often higher than that of the minority class (num_1). This results in a bias towards the majority class, leading to poor classification performance, especially for the minority class.

To address this limitation, we propose a weight-based adjustment to the decision rule by incorporating a ratio-based decision rule. Instead of directly comparing the raw number of neighbors from each class, we compute the ratio of the majority class to the minority class as follows :

$$\text{ratio} = \frac{\text{num\_0}}{\max(\text{num\_1}, 1)} \tag{4.4}$$

To further refine the decision boundary, we introduce a decision threshold based on the class weights, $w_0$ and $w_1$, as follows :

$$\text{decision\_threshold} = \frac{w_1}{w_0} \tag{4.5}$$

Here, $w_0$ and $w_1$ are the weights assigned to the majority and minority classes, respectively. This threshold allows the model to place more importance on the minority class, adjusting the decision rule accordingly.

Next, we simulate the probability of the sample belonging to the minority class (class 1) by modifying the standard KNN decision rule. The probability prob_class_1 is calculated using the following equation :

$$\text{prob\_class\_1} = \frac{1}{1 + \frac{\text{ratio}}{\text{decision\_threshold}}} \tag{4.6}$$

This approach results in the following decision mechanism : If the ratio of neighbors from class 0 is greater than the decision threshold, the probability of the sample belonging to class 1 will be less than 0.5, and the sample will be classified as class 0. Otherwise, the probability of the sample belonging to class 1 will exceed 0.5, and the sample will be classified as class 1.

This modification allows the KNN model to better handle imbalanced datasets by adjusting the decision boundary based on the relative importance of each class. By incorporating class weights into the probability estimation, the model becomes more sensitive to the minority class, improving its ability to correctly classify the minority samples in imbalanced scenarios and yields a usable probability for AUROC and AUPRC computation.

The KNN model exhibits high accuracy across all splits, with an accuracy of 90% on the training split, 91% on the validation split, and 89.9% on the test split. However, its recall is much lower, especially on the validation (37%) and test splits (26.7%), indicating that it struggles to correctly identify the minority class (readmitted patients). This suggests that KNN is heavily influenced by the majority class, resulting in low sensitivity to the positive samples. Additionally, both recall and AUROC exhibit a marked decline from the training to validation and test splits, as illustrated in Figure 4.13. This trend aligns with the model's low CV_AUROC of 72.7%, further underscoring its limited ability to generalize effectively to unseen data. While KNN performs relatively well in terms of AUPRC on the training (15.9%) and validation splits (15.7%), its performance on the test split (3.9%) is moderate, suggesting that the model's ability to distinguish between classes decreases on unseen data.

In the training and validation splits, the NB model exhibited high recall of 80.2% and 80.4%, respectively, alongside relatively low accuracy of 69% and 73.3%. This indicates a tendency to favor the minority class, which influenced its modest AUPRC values of 10.6% (training)

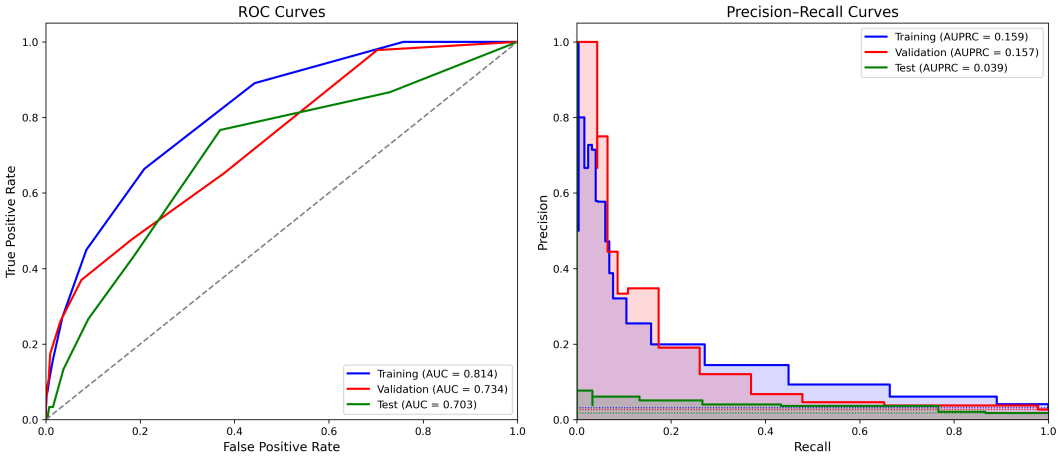



FIGURE 4.13 AUPRC and AUROC performance of the KNN classifier

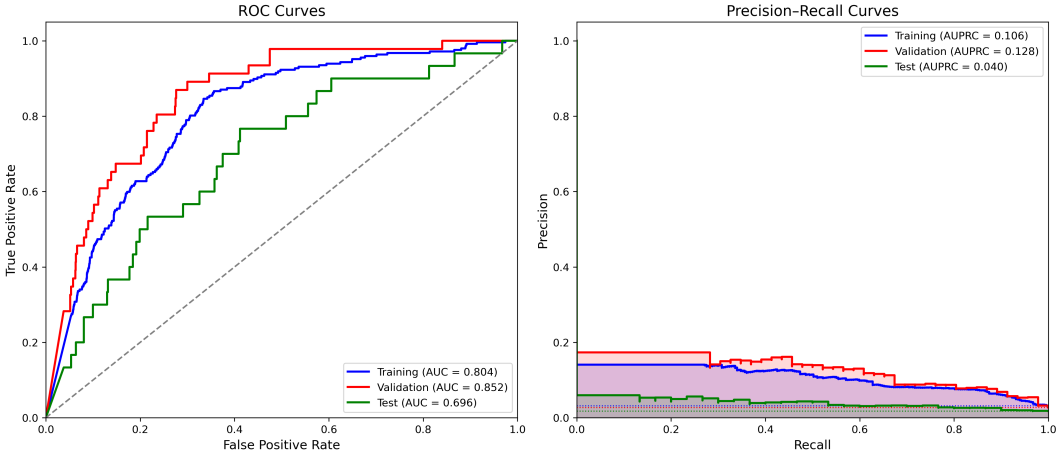



FIGURE 4.14 AUPRC and AUROC performance of the NB classifier

and 12.8% (validation) as shown in Figure 4.14. The model demonstrated high AUROC performance of 80.4% and 85.2% on the training and validation splits, consistent with its CV_AUROC of 80.3%. However, a substantial decline in AUROC to 69.7% was observed on the test split, accompanied by low accuracy (73.3%), recall (53.3%), and AUPRC (4%). This pattern suggests that NB model struggles to generalize to unseen data, leading to diminished predictive performance on the test split.

KNN achieved high accuracy but consistently failed to identify minority cases, reflecting overfitting and poor generalization. NB favored the minority class with high recall but low accuracy, and its AUROC gains on training and validation collapsed on the test set. By contrast, LR delivered balanced performance across all splits, indicating that this task benefits from models with stronger inductive biases for heterogeneous, sparse clinical time series.

**Performance comparison of the proposed model with tree-based models**

In addition to Linear and standard ML models, we evaluated several tree-based models, including XGBoost, Extra Trees (ET), RF, LightGBM, and CatBoost, to predict PICU readmission. The model optimized parameters are shown in Table 4.5. The parameters for each model were chosen to enforce strong regularization and mitigate the risk of overfitting, particularly given the class imbalance and heterogeneity in our dataset. For the boosting models (XGBoost and LightGBM), we restricted tree depth, applied subsampling of both observations and features, and used small learning rates to control model complexity, while incorporating class weights to address imbalance. For RF and ET, shallow trees and minimum sample constraints were applied to reduce variance, combined with ensemble averaging to stabilize performance. These conservative configurations ensured that the models captured clinically relevant patterns without fitting spurious noise, although at the cost of increased bias and a higher risk of underfitting. Table 4.8 shows the performance comparison between these tree-based models.

XGBoost demonstrated high accuracy across the training, validation, and test splits, reaching 81.4%, 83.6%, and 86.1%, respectively. Despite this, the model's recall was high in the training split (72%) but dropped considerably in the validation (45.7%) and test splits (46.7%), indicating a tendency to favor the majority class. This behavior also affected the AUPRC, which decreased from 20.6% in training to 9% in validation and 5.1% in testing, reflecting limited ability to capture the minority class as shown in Figure 4.15. Although the AUROC was strong in training (85.7%), it declined in the validation (72.9%) and test splits (72.4%), aligning closely with the CV_AUROC of 74.2%. Overall, XGBoost showed limited generalization and was heavily influenced by class imbalance, resulting in poor performance for minority class prediction.

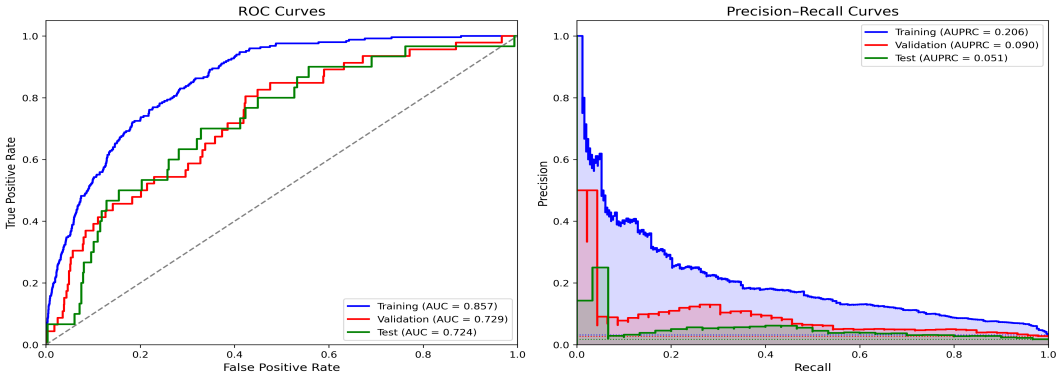



FIGURE 4.15 AUPRC and AUROC performance of the XGBoost classifier

TABLE 4.8 Performance comparison with tree-based models

| **Metric** | **XGBoost** | | | **ET** | | | **RF** | | | **LGBM** | | | **CatBoost** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Train | Val | Test | Train | Val | Test | Train | Val | Test | Train | Val | Test | Train | Val | Test |
| AUPRC (%) | 20.6 | 9.0 | 5.1 | 15.3 | 11.9 | 4.3 | 13.4 | 11.8 | 4.5 | 16.6 | 10.7 | 4.7 | 35.6 | 11.5 | 5.1 |
| Recall (%) | 72.0 | 45.7 | 46.7 | 49.4 | 47.8 | 46.7 | 53.0 | 50.0 | 36.7 | 64.4 | 50.0 | 46.7 | 64.8 | 34.8 | 23.3 |
| AUROC (%) | 85.7 | 72.9 | 72.4 | 72.1 | 73.8 | 71.4 | 74.7 | 73.7 | 68.1 | 77.5 | 71.3 | 74.2 | 83.5 | 75.4 | 67.9 |
| Accuracy (%) | 81.4 | 83.6 | 86.1 | 77.1 | 78.9 | 80.6 | 79.2 | 79.7 | 82.1 | 77.2 | 79.9 | 82.6 | 85.1 | 85.3 | 86.6 |
| F1 score (%) | 19.5 | 13.2 | 10.6 | 11.9 | 11 | 7.8 | 13.7 | 11.8 | 6.8 | 15 | 11.9 | 8.7 | 21.4 | 11.4 | 5.8 |
| Specificity (%) | 81.7 | 84.7 | 86.8 | 78 | 79.7 | 81.2 | 80 | 80.5 | 82.9 | 77.7 | 80.7 | 83.3 | 85.8 | 86.7 | 87.7 |
| Precision (%) | 11.3 | 9 | 6 | 6.8 | 6.2 | 4.3 | 7.9 | 6.7 | 3.7 | 8.5 | 6.7 | 4.8 | 12.8 | 6.8 | 3.3 |
| CV_AUROC (%) | 74.2 | | | 70.4 | | | 70.2 | | | 72.5 | | | 67.5 | | |

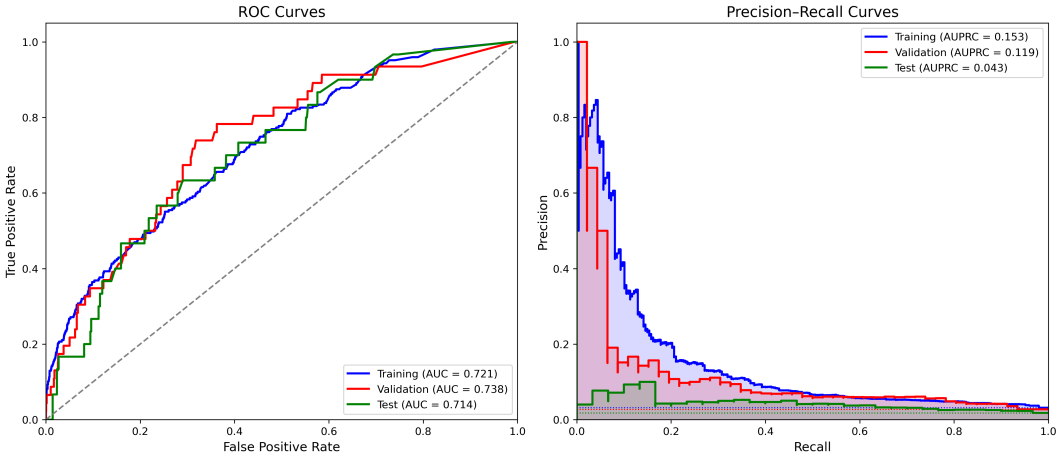



FIGURE 4.16 AUPRC and AUROC performance of the ET classifier

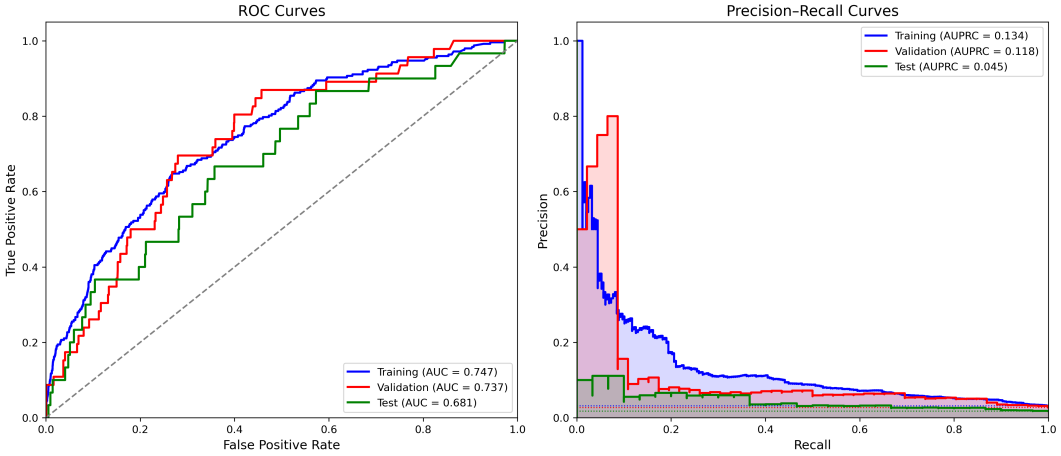



FIGURE 4.17 AUPRC and AUROC performance of the RF classifier

ET showed a moderate stabel performance across the three splits. It achieved moderate training accuracy (77.1%) and low recall (49.4%), which remained similar in validation (Accuracy : 78.9%, Recall : 47.8%) and test splits (Accuracy : 80.6%, Recall : 46.7%). The AUPRC dropped from 15.3% in training to 11.9% in validation and 4.3% in testing as shown in Figure 4.16. AUROC values were also moderate across all splits (Train : 72.1%, Val : 73.8%, Test : 71.4%), reflecting both poor discrimination for the minority class and limited generalization.

RF demonstrated slightly higher training accuracy and recall than ET with accuracy (79.2%) and recall (53%). This remained simlar in validation with recall (50%) but experiencing a considerable drop in test recall to 36.7% and AUPRC to 4.5%. AUPRC values were low across all splits (Train : 13.4%, Val : 11.8%, Test : 4.5%) but better than random. As shown

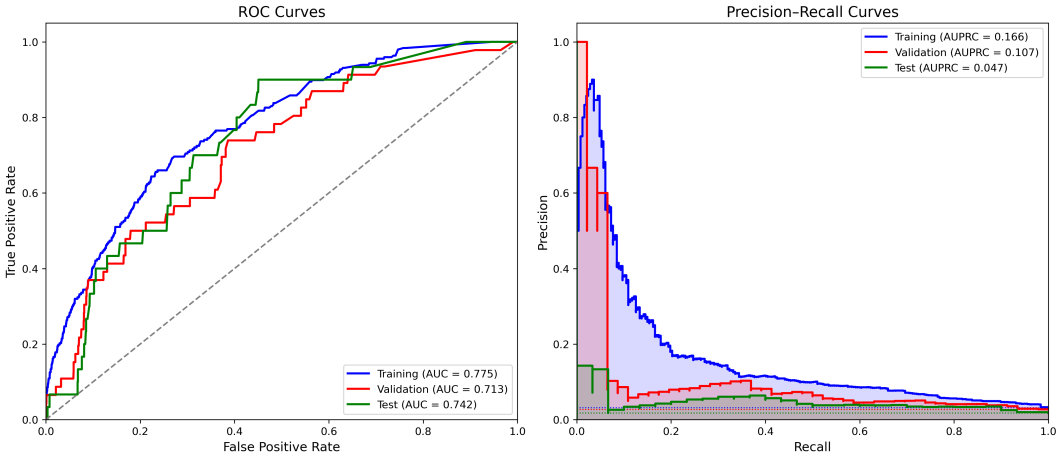



FIGURE 4.18 AUPRC and AUROC performance of the LightGBM classifier

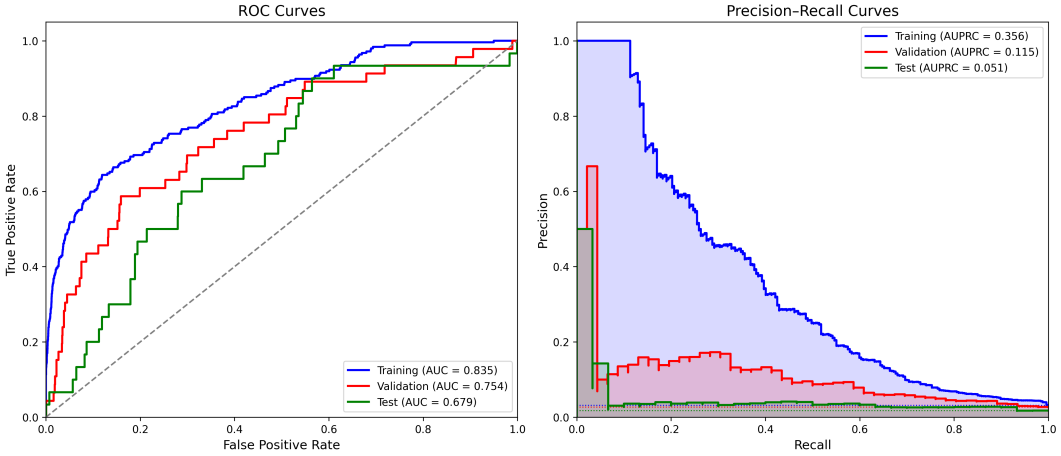



FIGURE 4.19 AUPRC and AUROC performance of the CatBoost classifier

in Figure 4.17, AUROC decreased from 74.7% in training and 73.7% in validation to 68.1% in test split matching the low CV_AUROC of 70.2%. RF exhibited similar limitations as ET, with poor minority class detection and weak generalization.

LightGBM showed somewhat better performance than RF in terms of recall, achieving 64.4% on training and 50% on validation, but still dropped to 46.7% on the test split, with AUPRC declining from 16.6% to 4.7% as shown in Figure 4.18. It has the best test AUROC across the tree based models. However, AUROC also decreased from 77.5% (train) to 71.3% (val) and 74.2% (test) like the CV_AUROC of 72.5%, reflecting some generalization issues and sensitivity to class imbalance.

CatBoost showed strong training performance as it demonstrated high training accuracy

(85.1%) and strong training AUPRC (35.6%), yet its recall and AUROC decreased drastically on validation (recall 34.8%, AUROC 75.4%) and test splits (recall 23.3%, AUROC 67.9%), indicating overfitting to the training data. Figure 4.19 shows that the AUROC dropped dramatically from 83.5% in training to 75.4% and 67.9% in validation and test splits, respectively matching the CV_AUROC of 67.5% showing poor performance.

Unexpectedly, tree-based models lagged behind linear models in the small-sample regime, a reversal of the ICU setting. Trees are inherently high-variance learners and require sufficient positive events to establish stable splits ; with few cases, they are prone to overfitting. While tree-based models often attained higher training accuracy, their recall and AUPRC consistently declined on validation and test sets, especially for the minority class, underscoring their limited generalizability under data scarcity. Among tree-based approaches, gradient-boosted ensembles such as XGBoost and LightGBM delivered the best and most consistent results, typically outperforming simpler baselines such as KNN and NB. Nevertheless, our proposed LR achieved more balanced performance across all splits.

**Performance comparison of the proposed model with DL models**

We compared the performance of our proposed model with two DL models : Bidirectional LSTM (BiLSTM) based attention and Transformer models that integrate temporal clinical measurements and static patient information to predict PICU readmission. In this study, the time-series representation was enriched by combining multiple complementary feature types for each measurement. Alongside the raw numerical values, we included corresponding categorical states and the absolute difference between each measured value and its age-adjusted normal range, enabling the model to quantify the degree of deviation in a clinically meaningful way. The model parameters are shown in Table 4.5. Given the limited size of our dataset, we deliberately designed our DL models with relatively simple architectures. Restricting network depth, number of units, and overall parameter count helped reduce the risk of overfitting while ensuring that the models remained trainable with the available data. This minimalist structure allowed the networks to capture essential temporal and clinical patterns without introducing unnecessary complexity that could compromise generalization.

Figure 4.20 shows the architecture of the dual-input BiLSTM–attention model. The time-series branch processes the most recent 24 hours of patient data through a masking layer to ignore padded values, followed by a BiLSTM network that captures dependencies and exploits both past and future contextual information within the observed sequence, an important advantage in retrospective clinical prediction where the complete sequence is available. This bidirectional processing enhances the ability to detect subtle patterns, such as deteriora-

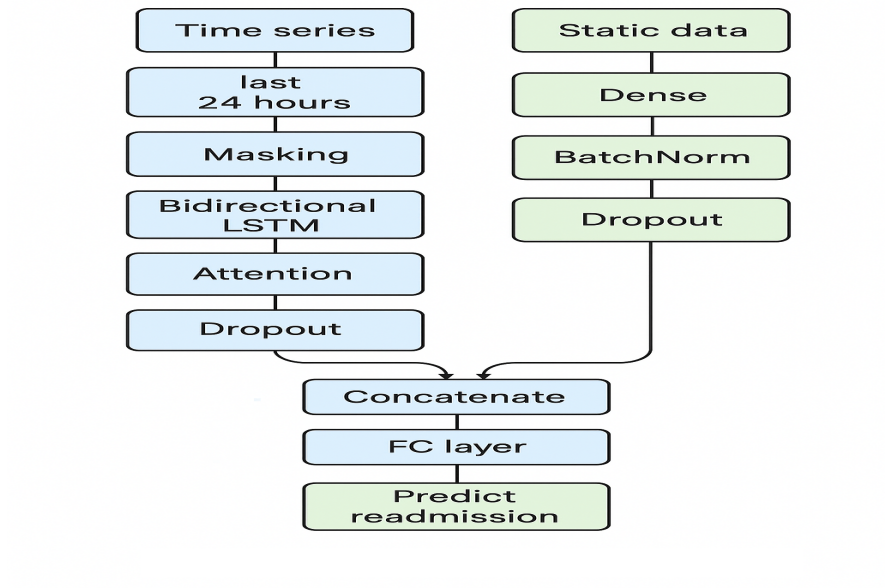



FIGURE 4.20 Architecture of the dual-input BiLSTM–attention model

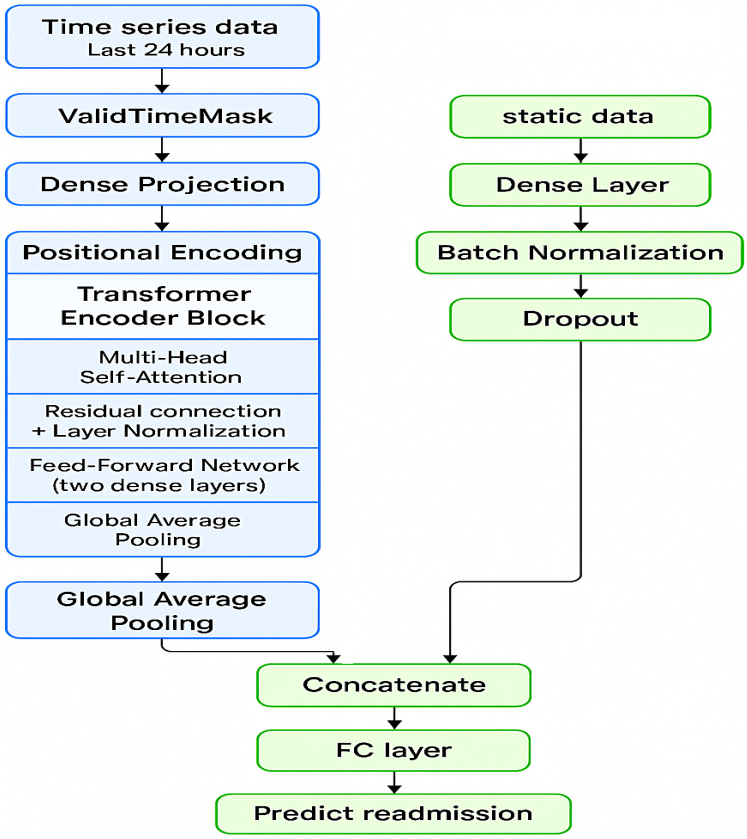



FIGURE 4.21 Architecture of the dual-input Transformer model

tion–recovery dynamics, that may influence readmission risk. An attention mechanism is then applied to selectively weight the most informative time steps and improve predictive power, enabling the model to highlight specific hours or events that are most relevant for the decision, reducing the influence of noise and allowing for clinically meaningful insight into the temporal factors driving the prediction. Batch normalization and dropout are incorporated to stabilize training and mitigate overfitting. In parallel, the static branch processes diagnosis codes and metadata through a dense layer with ReLU activation, batch normalization, and dropout, producing a compact representation of the static features. The outputs of the two branches are concatenated and passed through multiple fully connected layers to generate the final readmission prediction.

Figure 4.21 shows the architecture of the dual-input Transformer model. The time-series branch of the proposed model incorporates a preprocessing step that is crucial for handling irregular and padded clinical sequences. A Valid Time Mask is generated before any positional encoding is applied, ensuring that padded time steps—introduced to standardize sequence length—are accurately identified while they still contain only zero values. This mask is passed directly to the Transformer's self-attention mechanism, preventing padded positions from influencing attention weights. By contrast, a standard Masking Layer is also applied to mark these positions for masking-aware layers, but it operates internally during layer computations and does not explicitly control attention. Placing the Valid Time Mask before positional encoding is essential, as adding positional embeddings would otherwise obscure which time steps were originally padding.

Following masking, the model applies learnable positional encoding to embed temporal order into the sequence representation, enabling the Transformer to capture clinically relevant progression patterns that pure self-attention cannot model inherently. The Transformer encoder processes the entire sequence in parallel, using multi-head self-attention to relate events across both short and long temporal spans, while the attention mask ensures only valid time points are considered. This architecture allows the model to focus adaptively on the most informative moments in a patient's history, improving its ability to detect complex, temporally distributed risk factors for PICU readmission.

The BiLSTM and Transformer models achieved comparable performance across the three splits, as shown in Table 4.9. Both models performed well on the training set, with BiLSTM slightly ahead, but their performance declined on the validation and test sets, indicating overfitting. Their test accuracies of 85.7% and 85.1%, respectively, indicating their ability to capture overall patterns in the data. However, both models exhibited limited sensitivity to PICU readmissions, with recall values of 50.0% (BiLSTM) and 56.7% (Transformer) and

TABLE 4.9 Performance comparison with the two DL models

| **Metric** | **BiLSTM** | | | **Transformer** | | |
|---|---|---|---|---|---|---|
| | Train | Val | Test | Train | Val | Test |
| AUPRC (%) | 23.0 | 7.6 | 8.3 | 17.2 | 8.6 | 7.3 |
| Recall (%) | 70.9 | 43.5 | 50.0 | 68.0 | 50.0 | 56.7 |
| AUROC (%) | 84.7 | 74.8 | 79.2 | 83.2 | 76.6 | 78.5 |
| Accuracy (%) | 80.3 | 84.3 | 85.7 | 80.7 | 84.5 | 85.1 |
| F1 score (%) | 18.3 | 13.1 | 11 | 18.1 | 14.9 | 11.9 |
| Specificity (%) | 80.6 | 85.4 | 86.3 | 81.1 | 85.4 | 85.6 |
| Precision (%) | 10.5 | 7.7 | 6.2 | 10.4 | 8.8 | 6.6 |

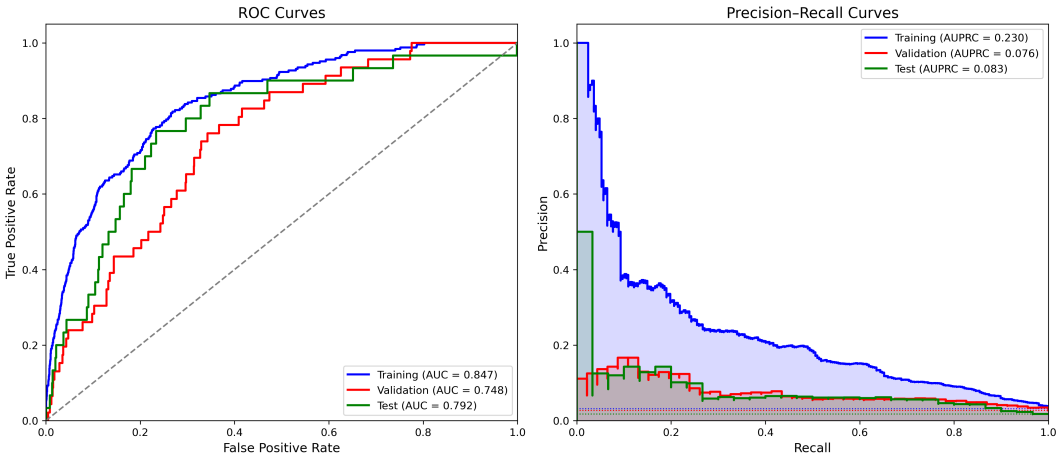



FIGURE 4.22 AUPRC and AUROC performance of the BiLSTM classifier

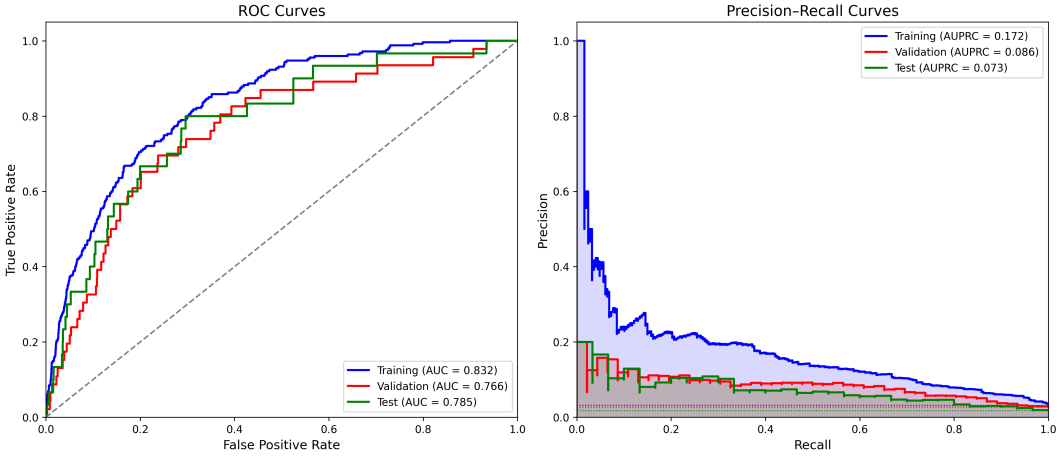



FIGURE 4.23 AUPRC and AUROC performance of the Transformer classifier

low AUPRCs of 8.3% and 7.3%, respectively as shown in Figure 4.22 and Figure 4.23. The Transformer's slightly higher recall suggests that its attention mechanism may better identify critical time steps. Despite their capacity to model complex temporal dependencies, the low recall and precision for the minority class highlight the combined effects of class imbalance and limited sample size, which constrain the models' ability to learn robust, discriminative patterns.

Despite the small cohort and deliberately simple architectures, both DL models achieved competitive test AUROCs, outperforming most classical ML baselines except Ridge and our proposed LR. These findings suggest that sequence models can capture clinically meaningful temporal patterns in multivariate time-series data, even under data scarcity. However, their performance remained constrained by severe class imbalance, limited sample size, and reliance on only the last 24 hours of data.

Although the DL models achieved slightly higher overall accuracy, they exhibited low sensitivity to readmission cases, reflecting the challenges of imbalance and limited data. Between the two architectures, the Transformer achieved marginally higher test recall than the BiLSTM, underscoring the potential of attention mechanisms to identify relevant temporal dependencies. Nevertheless, given its superior recall, interpretability, and generalization, LR remains the most clinically reliable model in this setting.

### Performance comparison with previous works

Compared to previous work, our DL and linear models demonstrate markedly higher predictive performance for PICU readmission as shown in Table 4.10. Laksana *et al.* [207] employed an LSTM-based RNN to predict patient readmission within the first three days post-

Table 4.10 Performance comparison with previous PICU readmission models

| **Study** | **AUROC (%)** |
|---|---|
| LSTM [207] | 64.4 |
| RF [208] | 70.0 |
| **Our proposed ET** | 71.4 |
| **Our proposed XGBoost** | 72.4 |
| **Our proposed LightGBM** | 74.2 |
| **Our proposed SVM** | 74.9 |
| **Our proposed Transformer** | 78.5 |
| **Our proposed BiLSTM** | 79.2 |
| **Our proposed Ridge** | 80 |
| **Our proposed LR** | 87.8 |

discharge, achieving a moderate AUROC of 64.4%, despite augmenting their input with synthetic features and temporal masking to handle missingness. Similarly, Arshad *et al.* [208] evaluated RF, LR, and ElasticNet models, reporting a maximum AUROC of 70% with the RF model. In contrast, our deep learning models deliver competitive performance, with BiLSTM and Transformer reaching test AUROCs of 79.2% and 78.5%, respectively. These results highlight the value of integrating categorical state representations with raw numeric time series to mitigate age-based variability. Moreover, when trained on our proposed engineered features, classical ML models achieve substantially higher discrimination with the interpretable LR attains an AUROC of 87.8%, underscoring the effectiveness of the feature engineering and the selection procedure used to identify the most informative predictors.

### 4.2.3 Global interpretability

For the LR model, we assessed interpretability with two complementary analyses. First as shown in Figure 4.24, we ranked features by the absolute values of their LR coefficients which indicate the direction and magnitude of each feature's contribution to the prediction. Second as shown in Figure 4.25, we performed a leave-one-feature-out ablation, measuring the drop in AUPRC of both training and validation splits when each feature was removed. This criterion highlights features that are most critical to preserving discrimination for the minority class. Together, coefficient-based ranking (global direction and strength) and ablation-based importance (conditional utility given other predictors) provide a coherent picture of which variables drive predictions and which are indispensable for maintaining minority-class sensitivity.

The results indicate that our proposed, medical knowledge-based features are highly effective, emerging as the strongest drivers of model performance. In particular, temporally informed descriptors—such as the time to first abnormal value, the recovery latency (time from an abnormal value to the first return to normal), the severe-state ratio (fraction of time spent in abnormal sever state), the discharge-state flag (low/high at discharge), and the last observed state—consistently rank among the most informative predictors. In addition, signal-processing–based features derived from wavelet and wavelet–packet representations contribute meaningfully. Collectively, these engineered features and the diagnoses CCSR categories outperform simple statistical summaries by capturing fluctuation dynamics and recovery trajectories that differentiate readmitted from non-readmitted patients.

The analyses also highlight clinically coherent contributors : abnormal vital signs (HR, RR, Systolic Blood Pressure (SBP), Mean Blood Pressure (MBP)) ; abnormal laboratory results (phosphorus, monocytes, albumin, Absolute Reticulocyte Count (ARC), glucose, and blood

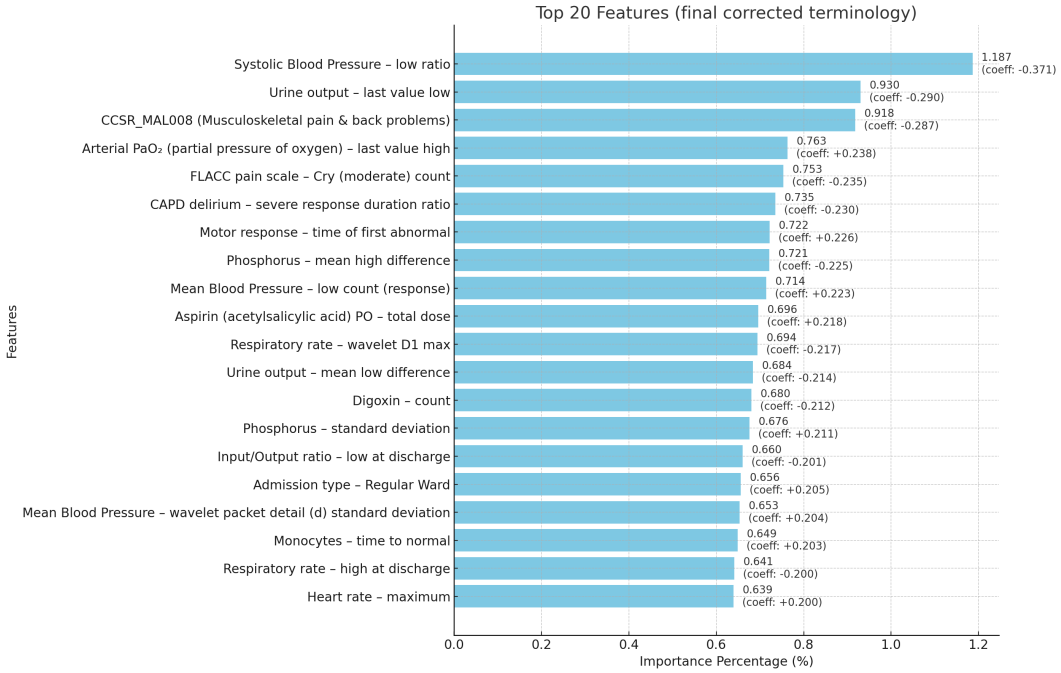



FIGURE 4.24 Feature ranking based on LR model coefficients

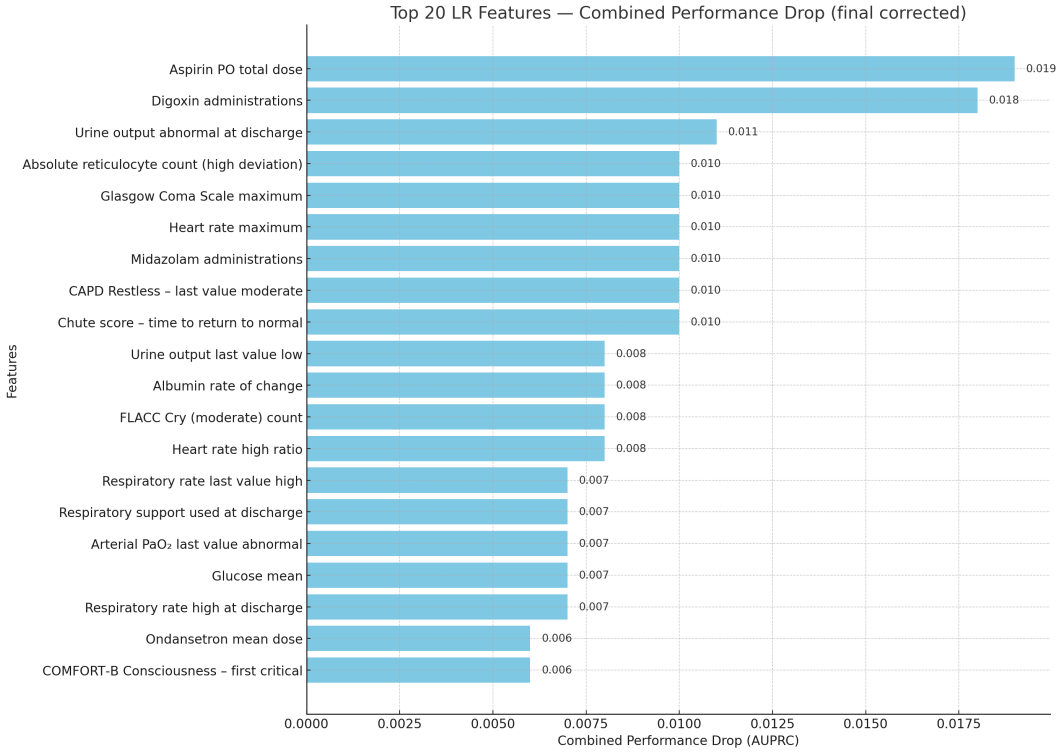



FIGURE 4.25 Feature ranking based on LR model ablation test

gases such as $PO_2$) ; abnormal fluid balance (urine output and input/output ratios) ; abnormal scores (GCS, FLACC, CAPD, Évaluation du risque de chute (Fall Risk Assessment), COMFORT-B) ; use of medications (Acide acetylsalicylique, Digoxine, Midazolam, Ondansetron) ; and the presence of musculoskeletal pain and back problems. Notably, variables such as Acide_acetylsalicylique_PO_sum and Digoxine_Count rank highly across both coefficient-based and ablation-based importance, reinforcing their predictive value and clinical relevance for PICU readmission. Taken together, these complementary results increase confidence in the model by revealing not only which features drive predictions but also which are indispensable for maintaining performance. Table A.1 provides the full description of all extracted features ranked by the absolute value of their LR coefficients.

For tree-based models, we used SHAP values to quantify feature importance. Figure 4.26 and Figure 4.27 summarize the XGBoost explainer. Consistently, medical knowledge-based features and signal-processing–based features dominate, indicating that the model relies on trajectories and state transitions rather than isolated measurements.

The most impactful pattern was variability in elevated Glucose (Glu) levels, a strong determinant of predictions. Other key contributors included urine-related indicators, oxygen saturation stability, and the timing of respiratory deterioration (Mechanical_Ventilation). Laboratory and treatment-related variables, as well as sedative usage timing, further contributed to risk stratification. Collectively, these findings show that the model integrates absolute physiological values with their temporal dynamics, emphasizing recovery speed, variability, and intervention timing as critical factors for predicting PICU readmission. The distribution of top-ranked features underscores the need for a multimodal approach that integrates diagnostic, physiologic, laboratory, and therapeutic information. .

The DL models, while demonstrating high predictive performance, are regarded as black-box approaches due to their complex internal architectures and the non-linear interactions between a large number of parameters. Unlike traditional statistical models, where feature contributions can be directly quantified through coefficients or performance drop analysis, DL models encode knowledge in multi-layered representations that are not inherently interpretable to humans. This opacity makes it challenging to understand the specific reasoning behind individual predictions or to identify which input features most strongly influence the model's decision-making process.

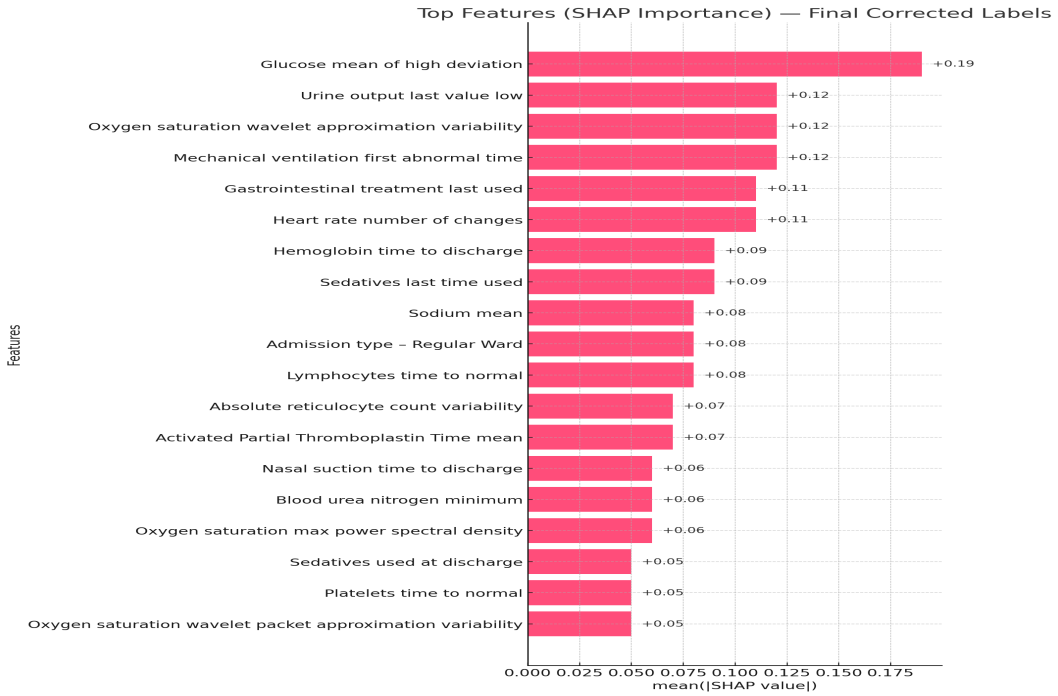



FIGURE 4.26 SHAP values summary for XGBoost model

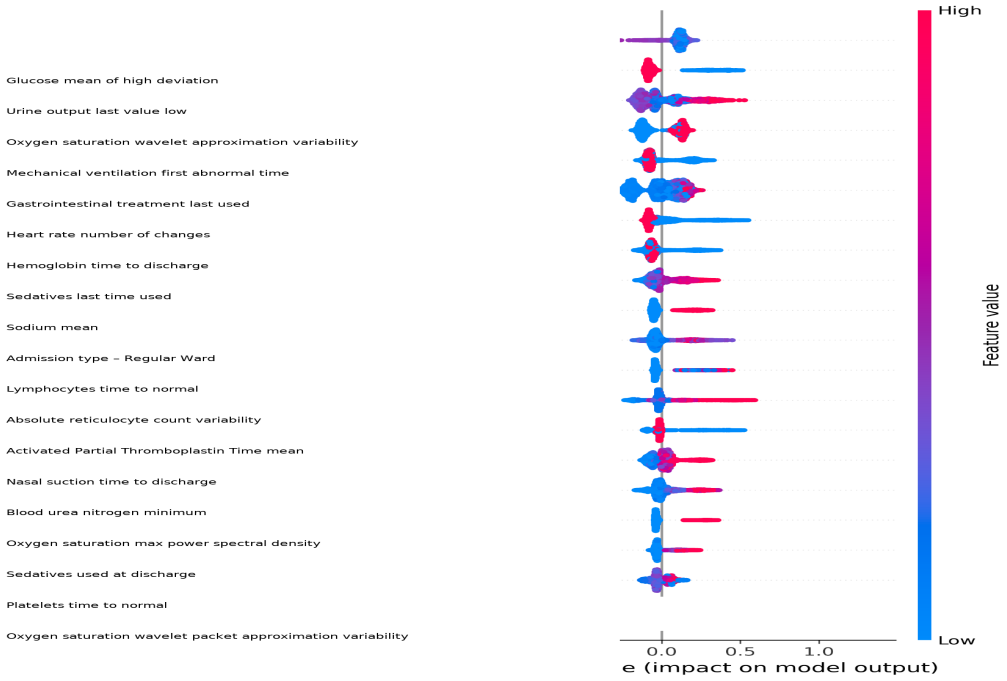



FIGURE 4.27 SHAP beeswarm of XGBoost model

### 4.2.4 Global interpretability of clinical variables

To assess global interpretability at the clinical variable level, we applied two complementary strategies. First, we aggregated the absolute LR coefficients of all features belonging to each clinical variable, allowing us to quantify their overall contribution to the prediction of PICU readmission. This approach highlights which variables consistently exert the strongest influence across the entire cohort. Second, we conducted an ablation analysis, in which all features corresponding to a given clinical variable were removed and the model retrained, with the resulting drop in AUPRC used as a measure of importance. Together, these methods provide a robust understanding of global interpretability : coefficient-based aggregation reveals how strongly variables are weighted by the model, while ablation analysis captures their unique predictive value when considered in the broader modeling context.

The global interpretability analysis based on LR coefficients (Figure 4.28) shows that Diagnoses related variables dominate the model, contributing the largest cumulative effect, followed by hemodynamic measures such as SBP, MBP, Diastolic Blood Pressure (DBP), and other vital signs such as HR, RR and Oxygen saturation (SPO2), as well as functional scores including FLACC, Comfort-B, and CAPD. In addition to Urine output and other laboatory results such as (Phosphor (P), Blood Urea Nitrogen (BUN), Glucose, Platelets (PLT) and pH). This indicates that both diagnostic information and physiological instability captured by vital signs, laboratory results and sedation scores are central to the model's predictions. Table A.2 presents the cumulative absolute effects of all clinical variables used to train the LR model for PICU readmission, together with the corresponding extracted features contributing to each variable.

In contrast, the ablation-based analysis using combined AUPRC drop (Figure 4.29) highlights Diagnoses again as the most influential variable, but also reveals that medications (e.g., Digoxine, Acide acetylsalicylique, Midazolam) play a more pronounced role than suggested by coefficient-based importance alone. The consistency of Diagnosis across both methods reinforces its strong predictive value, while the differences between the two analyses emphasize the complementary nature of these interpretability approaches : coefficient aggregation reflects the weight assigned by the model, whereas ablation analysis uncovers the unique predictive contribution of each clinical variable in the broader feature set.

### 4.2.5 Local interpretability

For local interpretability, we applied LIME to individual patients. We then aggregated feature-level contributions into clinical variables, which allowed us to identify, for each pa-

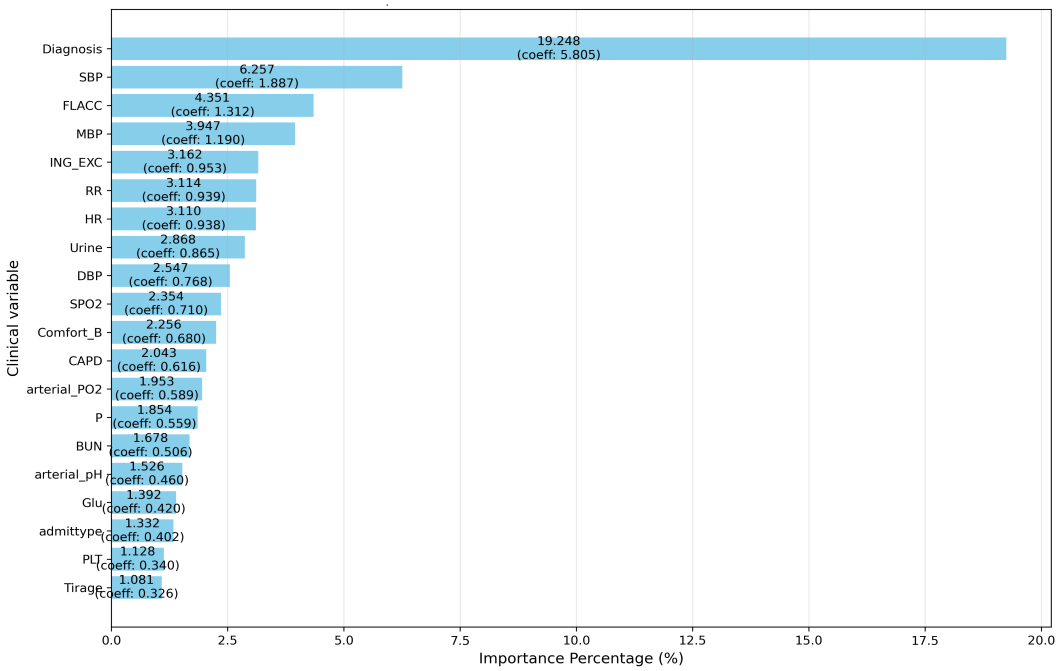



FIGURE 4.28 Top clinical variables based on LR coefficients

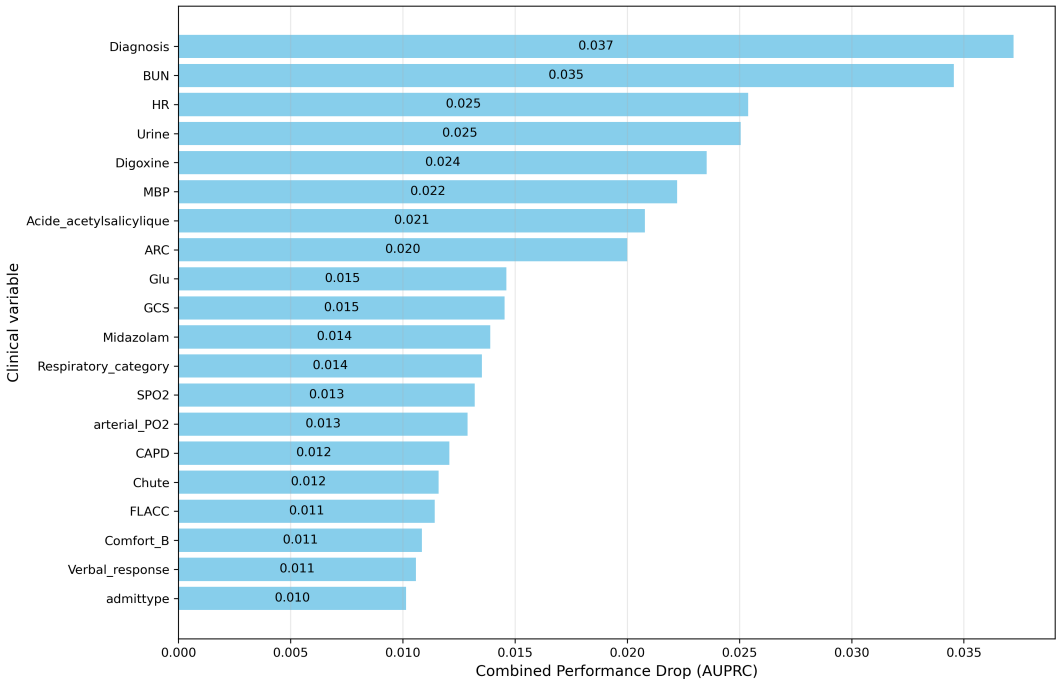



FIGURE 4.29 Top clinical variables based on LR ablation test

tient, the variables that most strongly supported the predicted outcome (readmission or non-readmission).

We next examine cases where the model correctly predicted that a patient would be readmitted, aligning with the actual outcome. Figures 4.30 and 4.31 illustrate the local interpretability results using LIME for two representative patients. For patient 1, the prediction was strongly supported by laboratory variables such as (capillary lactate, Troponin, Thyroid-Stimulating Hormone (TSH)) and medication exposures like Ciprofloxacine and Acetylcysteine, alongside diagnostic codes, all of which contributed meaningful weight to the model's decision. For patient 2, the model's high confidence in readmission was driven by a slightly different profile, with strong influence from Acetylcysteine, Ciprofloxacine, and Diagnosis, supplemented by laboratory indicators (e.g., TSH, venous HCO3, capillary lactate) and clinical scores (e.g., Comfort-B). These examples demonstrate how different clinical variables, spanning diagnoses, labs, and treatments, combine to support readmission risk predictions, highlighting the model's ability to capture patient-specific factors.

We also examined cases where the model predicted that a patient would be readmitted, while in reality the patient was not readmitted. Figures 4.32 and 4.33 show two such examples of false positives. For patient 3, the model's decision was mainly driven by medication exposures (e.g., Ciprofloxacine, Acetylcysteine, Isuprel, Tobramycine), Diagnosis, and supported by clinical scores (e.g., FLACC) and laboratory results (TSH) . Similarly, in patient 4, the prediction was supported by laboratory results (capillary lactate, Glucose, TSH), meidcations (Ciprofloxacine, Acetylcysteine, Isuprel, Digoxine) and Diagnosis, along with clinical scores (e.g., FLACC). These examples illustrate how treatments, diagnoses, and laboratory markers may lead the model to infer elevated risk of readmission, even in patients who ultimately recovered without requiring readmission. Such false positives suggest that while the model is sensitive to clinically relevant risk factors, it may overestimate their combined effect in certain individuals.

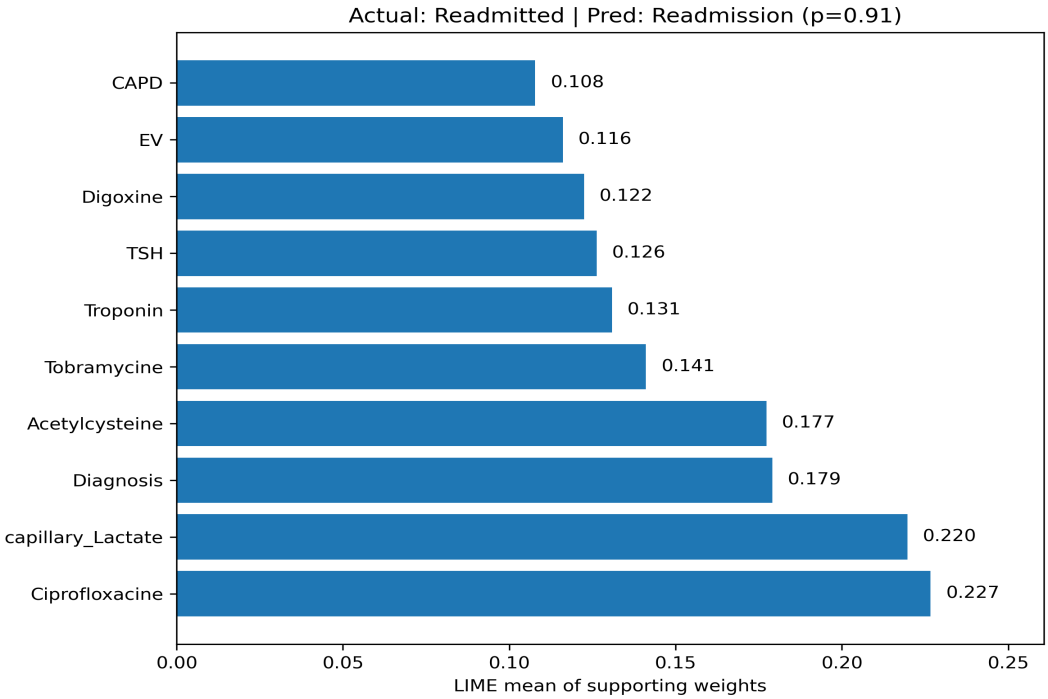



FIGURE 4.30 Local interpretability for patient 1

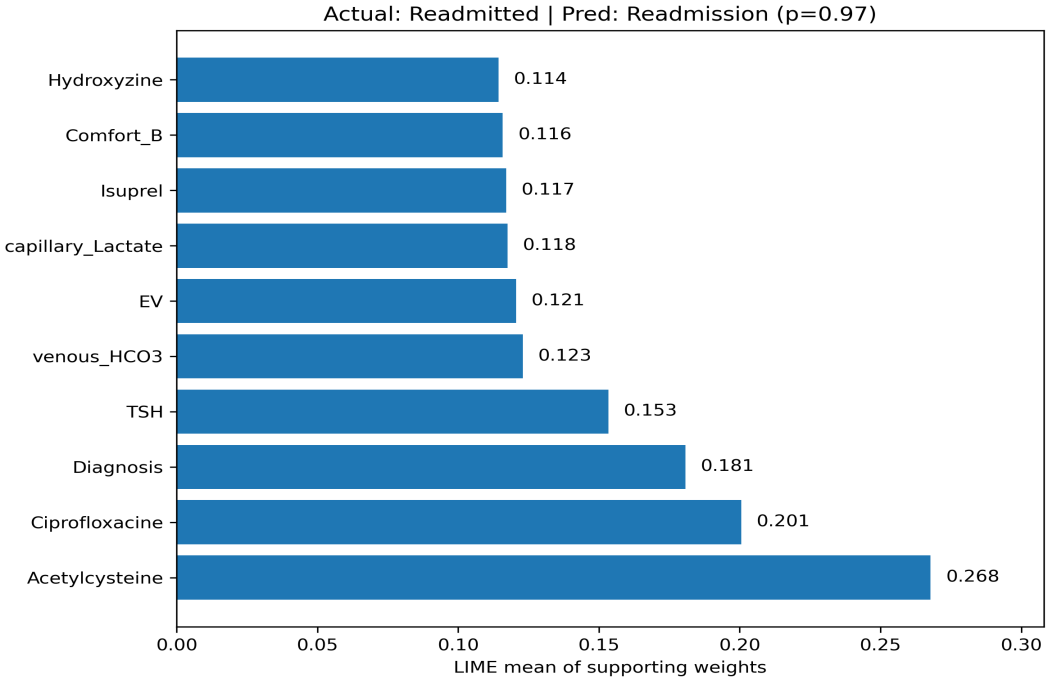



FIGURE 4.31 Local interpretability for patient 2

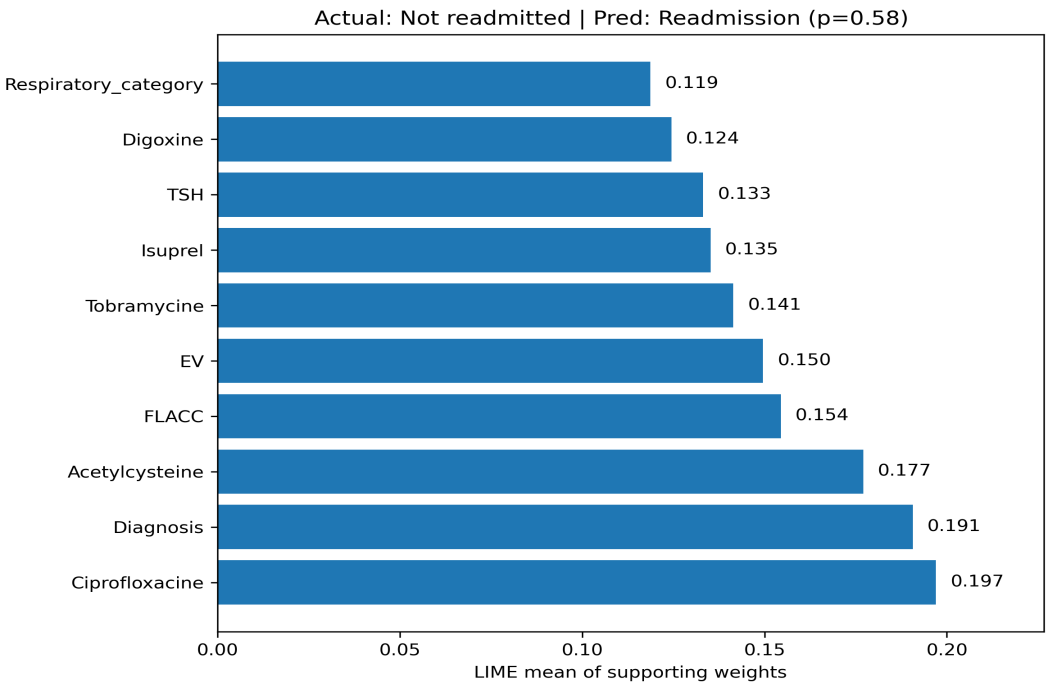



FIGURE 4.32 Local interpretability for patient 3

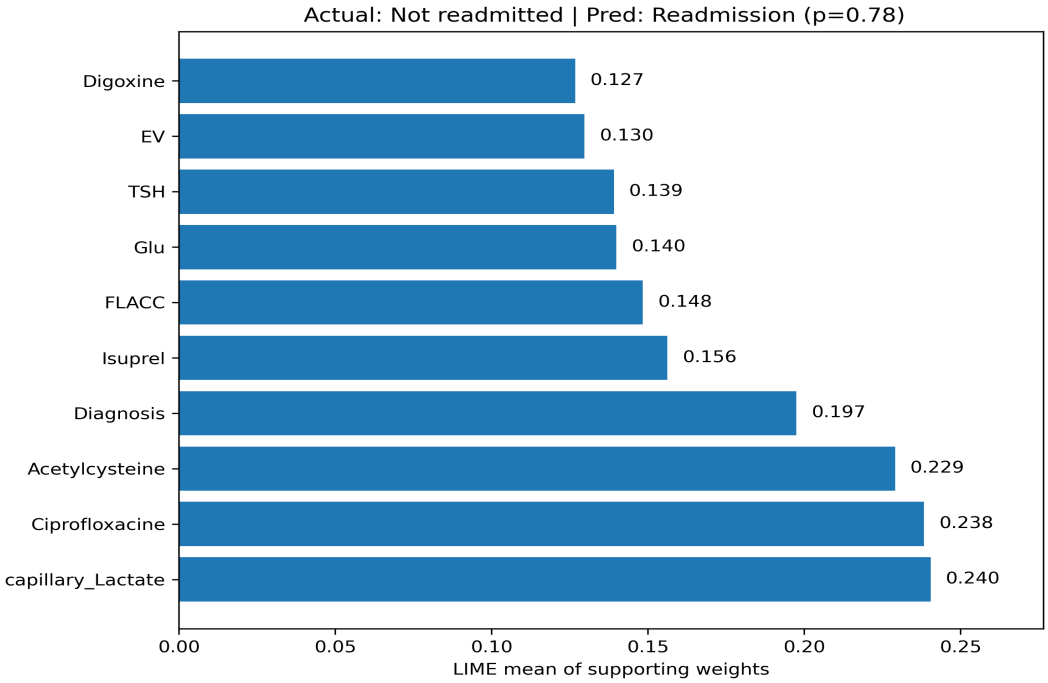



FIGURE 4.33 Local interpretability for patient 4

### 4.2.6 Impact of SMOTE oversampling on model performance

Table 4.11 summarizes the effect of applying the SMOTE technique to generate synthetic samples for the minority class. In our setting, SMOTE did not improve generalization. Across models, we observed pronounced overfitting : apparent gains on the training and validation splits failed to transfer to the test set. This likely reflects the fact that SMOTE interpolates in feature space without preserving patient-level trajectories, joint dependencies among laboratory values, vital signs, and treatments, or coding artifacts. As a result, synthetic samples distorted clinically meaningful correlations and temporal structure, producing patterns absent from real records. The problem was exacerbated by the extreme imbalance in our dataset (3% positives, 247 cases), which would have required approximately 30-fold augmentation.

Tree-based models tolerated the added noise somewhat better and benefited modestly from the larger sample size, achieving AUROC around 60% on the test set, whereas several linear baselines dropped below random performance (<50%). Overall, however, generalization remained weak, underscoring that naïve oversampling is insufficient on its own. In contrast, cost-sensitive training consistently improved minority-class detection and yielded more reliable test performance across classifiers.

TABLE 4.11 Model performance with SMOTE oversampling

| **Model** | **Train** | | | | **Validation** | | | | **Test** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUPRC % | Recall % | AUROC % | Acc % | AUPRC % | Recall % | AUROC % | Acc % | AUPRC % | Recall % | AUROC % | Acc % |
| LR | 93.7 | 90.3 | 94.3 | 87.6 | 39 | 89.1 | 94.3 | 85.4 | 2 | 20 | 50 | 78.1 |
| Ridge | 90.3 | 84.2 | 91 | 83.2 | 9.3 | 41.3 | 67.8 | 82 | 1.9 | 33.3 | 51.5 | 75.5 |
| SVM | 93.8 | 78.8 | 93.8 | 85.8 | 9.1 | 23.9 | 55.3 | 93.7 | 1.9 | 13.3 | 50.7 | 90.2 |
| KNN | 96.2 | 99.1 | 96.3 | 78.1 | 10.4 | 80.4 | 68.5 | 63.8 | 1.5 | 23.3 | 43.8 | 55.9 |
| NB | 82.7 | 38.3 | 85.6 | 65.9 | 9.5 | 39.1 | 64.1 | 95.2 | 1.9 | 6.7 | 48.4 | 91.1 |
| ET | 84 | 64.6 | 85.7 | 75.4 | 5.1 | 23.9 | 65.8 | 90.9 | 2.9 | 20 | 61.4 | 90.2 |
| RF | 86 | 69.2 | 86.2 | 76.2 | 13.6 | 39.1 | 65.3 | 86.8 | 2.5 | 16.7 | 60.6 | 83.4 |
| LightGBM | 95.4 | 85.8 | 95.4 | 88 | 9.2 | 41.3 | 67.5 | 90.5 | 2.6 | 26.7 | 59.6 | 86.2 |

# CHAPTER 5 DISCUSSION

## 5.1 Discussion

We proposed a practical strategy to predict ICU readmission within the first 72 hours of admission in the MIMIC-III database. The approach combines a LightGBM classifier with clinically guided feature engineering for time series, followed by feature selection to retain the most informative predictors and discard weak or redundant ones.

The proposed pipeline demonstrated several notable strengths. First, we carefully defined inclusion and exclusion criteria and adopted a clinically meaningful three-day readmission window, ensuring that the problem formulation aligns with hospital practices and provides actionable insights for discharge planning. Our pipeline integrated variables spanning multiple domains—including vital signs, laboratory results, fluid balance, and diagnostic codes—thereby capturing the multifaceted nature of an ICU stay. Rigorous preprocessing steps, such as harmonizing units, correcting inconsistencies, and removing physiologically implausible values, further enhanced data quality. Missing data were systematically addressed with imputation strategies, and numerical values were complemented with categorical state representations (e.g., low, normal, high), improving interpretability and supporting clinical reasoning.

Another strength lies in the diversity and depth of feature extraction. Beyond standard statistical descriptors, we introduced some clinical knowledge-based, temporal, and spectral features that enriched the data representation and contributed substantially to model performance. The use of diagnosis embeddings offered an innovative way to capture the complex interactions between comorbidities and care pathways, with these embeddings emerging among the most influential predictors. To mitigate redundancy, we employed MI for feature selection, preserving both linear and non-linear associations while reducing noise. Importantly, we addressed the severe class imbalance problem not by discarding data through sub-sampling but by applying cost-sensitive training, thereby giving greater weight to readmitted patients without distorting the underlying distribution. The resulting LightGBM model not only outperformed traditional ML approaches but also achieved performance comparable to advanced DL methods, while retaining interpretability through SHAP analysis—an essential property for clinical adoption.

Despite these advantages, several limitations temper the strength of our findings. Our imputation strategy was uniform across variables of very different clinical natures, which may not have been optimal. For example, LOCF is inappropriate for medications : it would falsely

imply continuous, hourly administration. Therefore, addressing missing data with methods that respect temporal structure and clinical constraints is crucial. Likewise, temporal and spectral feature extraction through automated libraries generated a vast number of highly correlated descriptors, many of which were redundant. The MI–based feature selection method, while effective in capturing univariate relevance, was limited in its ability to identify the most robust multivariate interactions. Furthermore, our approach to class imbalance relied solely on cost-sensitive training ; we did not evaluate alternative oversampling techniques such as SMOTE.

Additional limitations are tied to evaluation and generalizability. The random splitting of the dataset into training and test sets may have introduced optimistic bias by neglecting temporal trends, as patient characteristics and treatment practices evolve over time. We emphasized AUROC as the primary performance metric, but in the context of highly imbalanced outcomes such as ICU readmission, metrics like recall, precision and AUPRC are equally, if not more, important. Diagnosis embeddings, although effective predictors, were difficult to interpret clinically, limiting their explanatory value for physicians. Although the model demonstrates promising performance, prospective evaluation and integration into existing clinical workflows without disrupting routine practice remain essential priorities to ensure that it becomes both robust and actionable in clinical practice. Finally, because the model was developed on a single-center database, its external validity and generalizability remain untested. To support deployment as a decision aid, performance should be validated and calibrated on external databases and across care settings. Addressing these challenges will be critical for future work aiming to move beyond proof-of-concept toward safe, reliable, and clinically deployable decision support systems.

To address limitations encountered in the ICU pipeline, we extended and optimized the end-to-end workflow on *CathyDB* (Sainte-Justine Hospital). The pipeline developed for predicting readmission in the PICU builds on the methodological advances established with MIMIC-III while incorporating several innovations tailored to the unique challenges of pediatric data. We extracted variables spanning multiple clinical domains—including treatments, medications, and sedation scores—which provided a more comprehensive representation of a child's PICU stay than the adult pipeline. Rigorous preprocessing ensured data consistency : units of measurement were harmonized, documentation errors corrected, and noise managed through a proposed two-stage Kalman–wavelet filter that reduced noise while preserving key and real abrupt physiological patterns. By denoising the signals before extracting temporal and spectral features, we reduce the influence of measurement artifacts and random fluctuations. This enhances the sensitivity of the extracted features to real physiological changes, improving both model performance and interpretability. Missing data were addressed with a category-

aware imputation strategy that accounted for the clinical meaning of each variable, and for variables that were completely missing, we used age-specific median imputation based on population distributions. This design avoids implausible trajectories, preserves clinically meaningful dynamics, and improves downstream learning. Many pediatric variables are age-dependent with reference intervals that shift across pediatric age bands, making it difficult for the model to consistently distinguish normal from abnormal values. To reduce the burden on the model to infer age-specific normal ranges, we recoded numerical values into categorical states (low, normal, high) using quantile-based thresholds, thereby enhancing interpretability and robustness in the face of pediatric heterogeneity.

A major strength of this pipeline lies in its feature engineering. Beyond conventional statistical measures, we extracted clinical knowledge-based and signal processing based descriptors that captured hidden dynamics and substantially improved predictive performance. Not all variables exhibit oscillatory behavior ; therefore, we restricted signal-processing features to vital signs, which possess clear temporal and spectral structure. From the smoothed vital signals, we computed a targeted set of spectral/temporal descriptors that proved most informative in the ICU setting : FFT-based bandpower summaries, energies from wavelet approximation and detail coefficients, wavelet-packet node energies, power-spectral density statistics, and spectral entropy. Non-oscillatory variables were instead represented with clinically derived state and burden measures, yielding complementary, mechanism-aligned information. We summarized patient status over three clinically salient windows : (i) admission (first 3 hours post-PICU admission), (ii) discharge (last 3 hours pre-PICU discharge), and (iii) the entire PICU stay. These windows capture initial severity, recovery trajectory, and terminal status, enabling the model to quantify fluctuation burden and recovery dynamics and to distinguish features associated with admission versus those indicative of subsequent readmission risk. Unlike the opaque 300-dimensional diagnosis embeddings, we we represented diagnoses using 297 CCSR categories, which group related conditions into clinically coherent classes and facilitate attribution of risk to specific diagnostic domains. This design choice allowed us to preserve clinical transparency while retaining predictive ability, making the risk factors more actionable for physicians. Feature selection combined MI (to capture univariate associations) with Lasso regularization (to retain robust multivariate interactions), producing a compact but highly informative set of predictors. Together, these strategies enhanced the model's ability to capture both individual effects and higher-order patterns across the dataset.

The pipeline also addressed class imbalance through cost-sensitive learning and systematically compared it to oversampling with SMOTE, providing a balanced view of trade-offs. Importantly, we did not rely solely on AUROC, but also reported AUPRC, recall, and accuracy—metrics that are critical in the context of few but clinically serious events such as

PICU readmissions. We further proposed a custom class-wise loss function for linear models that penalized misclassification in both positive and negative classes, preventing the majority class from dominating optimization and guiding feature selection toward minority-class sensitivity. For classifiers based on decision scores, we introduced a procedure to convert scores into calibrated probabilities, and we adapted the KNN decision rule to better handle highly imbalanced data. Evaluation was conducted using a temporal split to mimic real-world deployment scenarios, ensuring that the reported performance more closely reflects prospective clinical use.

For some models, the test performance exceeded the validation performance. This can be attributed to the temporal split of the data, where the readmission rate was not identical across the validation and test sets. Moreover, since the final model was optimized using five-fold cross-validation, its final parameters were influenced by multiple training–validation splits rather than a single validation set, making the validation performance less representative of the final generalization ability.

Although the model achieved a high recall (83.3%) and a good specificity (81.7%), the precision remained low. This outcome reflects the highly imbalanced nature of the dataset, where the number of readmitted (positive) cases is much smaller than the non-readmitted (negative) ones. In such settings, even a modest number of false positives can substantially lower precision. Achieving high precision would require an extremely high specificity (approaching 99%), which is difficult without sacrificing recall. In practice, models that maximize specificity tend to become biased toward the majority negative class, thereby missing true readmissions. However, from a clinical standpoint, maintaining a high recall is more critical, as correctly identifying patients at risk of readmission is essential for timely intervention and potentially life-saving care.

The results highlight the power of thoughtful feature engineering. Even with a limited-size pediatric dataset, our approach enabled a relatively simple LR model to outperform more complex ML and DL methods, underscoring the importance of domain-specific feature design. At the same time, our DL models with attention mechanisms showed promising performance, surpassing prior studies while using relatively simple architectures. Crucially, interpretability was embedded throughout : we provided both global explanations at the feature-engineering and clinical-variable levels, as well as local explanations for individual predictions. This is essential for clinician trust and clinical decision-making. Taken together, the PICU pipeline represents a robust and interpretable approach that not only advances the state of the art but also holds tangible potential to improve discharge planning, reduce preventable readmissions, decrease length of stay, and ultimately enhance the quality of pediatric critical care.

## 5.2 Limitations

Despite the strengths of our PICU pipeline, this study has several limitations that temper its generalizability and suggest directions for future work. First, the dataset is relatively small and severely imbalanced, which constrains model capacity and increases variance in performance estimates. Another major limitation is the set of variables that were not included in our analysis. Severity of illness scores such as Pediatric Logistic Organ Dysfunction (PELOD) and specialized discharge scores were absent from our database. Similarly, we did not capture neurological deficits, surgical interventions, dependence on medical devices, or nutritional interventions, all of which are clinically relevant factors that could influence readmission risk and consistently identified in the literature as strong predictors of readmission. In addition, we limited our analysis to structured tabular data and did not exploit other valuable modalities such as free-text clinical notes or medical imaging (e.g., chest X-ray, MRI), which have shown promising results in other studies and could enhance predictive power.

Certain methodological choices also introduce constraints. Some of the variables we used, such as ICD-10 diagnostic codes, are only available at discharge, which makes them unsuitable for real-time deployment ; we included them here to better understand their contribution and to inform future database design. We defined normal ranges using quantiles instead of the common z-score method, which does not always represent true population distributions. Our imputation strategy, though category-aware and clinically informed, was not validated against alternative methods such as MICE, leaving uncertainty about its robustness. Likewise, our two-stage Kalman–wavelet filtering approach to noise reduction in vital signs was not benchmarked against direct feature extraction from raw signals or against other filtering techniques, limiting confidence in its relative advantages. Additional spectral and temporal descriptors, as well as alternative sequence representation methods such as learned embeddings, were also not explored, which could have revealed better performance. Moreover, we did not leverage external knowledge from medical ontologies or domain-specific databases such as UMLS or SNOMED ; integrating such resources could enhance feature representations and facilitate a deeper understanding of relationships among clinical variables.

From a modeling standpoint, our approach faced challenges related to dataset size and computational complexity. The limited sample size restricted our ability to train more advanced DL models, and we did not pursue synthetic temporal data generation that could have mitigated this limitation. Consequently, our evaluation of DL was limited to two relatively simple architectures, while prior studies suggest that more sophisticated models could yield stronger results. In addition, our DL experiments were limited to time-series inputs with demographic and diagnoses data and excluded laboratory results and other informative mo-

dalities, which may have capped their performance. Similarly, although boosting algorithms such as XGBoost and LightGBM are known to be highly competitive, their performance was constrained by the limited dataset size. Our training strategy also relied on computationally intensive procedures : grid search for hyperparameter optimization and a three-stage training with wrapper feature selection process based on greedy search. While effective, these approaches were time-consuming and increased the cost of experimentation. Moreover, the objective function used to select features by maximizing the min performance across splits, was not sufficiently robust, as it could improve weaker folds while degrading stronger ones.

In addition, there are some limitations concern interpretability and representation choices. Although we improved diagnostic interpretability by using CCSR categories, this representation was less informative than high-dimensional embeddings, highlighting a trade-off between transparency and richness of information. At the interpretability stage, we combined all diagnoses into a single category to enable global analysis, but this may obscure the contribution of individual diagnosis categories, which could carry distinct clinical implications. Local interpretability may be biased, as variables with many features can appear more important, while true risk factors with fewer features may be underestimated. The DL models were not paired with post-hoc explanation methods (e.g., DeepSHAP), limiting insight into which features drive their predictions. These shortcomings underscore the need for richer, more nuanced feature representation and interpretability strategies that balance clinical usability with predictive accuracy.

Finally, although our model achieved high AUROC and recall with reasonable AUPRC, overall accuracy was modest, reflecting threshold trade-offs. More importantly, the system currently provides only a single risk estimate at discharge, rather than time-updated predictions during the PICU stay, which limits its real-time clinical utility. The study was also restricted to a single-center retrospective cohort, and external validation across multiple hospitals, as well as temporal recalibration, will be essential before deployment in practice.

# CHAPTER 6 CONCLUSION

## 6.1 Summary of Works

Readmission is a critical quality and safety metric, associated with increased mortality, longer LOS, and higher healthcare costs. Early and accurate risk identification is therefore essential to reduce avoidable morbidity and mortality, limit PICU "bounce-backs," and lessen the burden on hospital resources and staff by informing discharge readiness, guiding targeted post-discharge monitoring, enabling proactive interventions, and optimizing resource allocation. While adult ICU readmission has been extensively studied, pediatric investigations remain relatively sparse and, to date, have reported only modest performance.

This thesis focused on developing high-performing and interpretable models to predict unplanned PICU readmission within 72 hours of discharge. The work used limited, imbalanced, and heterogeneous EHR data covering entire ICU and PICU stays. Guided by our research question, we pursued three objectives : (i) to design an end-to-end ICU readmission pipeline on public adult data (MIMIC-III) ; (ii) to adapt and optimize this pipeline for the pediatric context (CathyDB, CHU Sainte-Justine) ; and (iii) to ensure that model predictions are interpretable and clinically meaningful.

To achieve these goals, we built a pediatric-aware pipeline that handles noisy, irregular, and heterogeneous PICU data. The framework combined variable-aware preprocessing, age-normalized representations, medical knowledge-based feature engineering, advanced temporal and spectral descriptors, and imbalance-aware learning strategies. Results showed that interpretable models can achieve state-of-the-art performance. In MIMIC-III, our LightGBM baseline reached an AUROC of 78.6%, comparable to more complex deep learning networks. In the pediatric dataset (CathyDB), our logistic regression model achieved strong performance (AUROC 87.8%, Recall 83.3%, AUPRC 13.8%, Accuracy 81.8%) with cross-validation AUROC of 80%, outperforming both classical ML and deep learning models as well as previous published work. Importantly, interpretability analyses showed that the most influential predictors aligned with established clinical reasoning, which supports the model's credibility for clinical use.

The main contribution of this work is bridging the gap between technical rigor and clinical relevance. It lays the foundation for deploying interpretable AI systems in critical care, demonstrating that transparent models can match or surpass the performance of black-box approaches. This achievement carries significant implications for healthcare systems : suppor-

ting safer discharge decisions, providing real-time decision support to physicians, improving outcomes for vulnerable pediatric populations, strengthening post-discharge monitoring, and enabling more efficient use of scarce PICU resources.

In a broader context, this thesis illustrates how pediatric-aware AI can reshape critical care by reducing preventable complications, improving hospital efficiency, and ultimately enhancing the quality of life for children and their families. The methodological principles advanced here—robust preprocessing, transparent modeling, and clinically faithful explanations—are not limited to PICU readmission but extend to other high-stakes medical prediction tasks. Taken together, these findings affirm that interpretable and reliable AI can serve both as a decision-support tool and as a catalyst for transforming critical care into a safer, more proactive, and more equitable practice.

## 6.2 Future Research

The potential of ICU and PICU readmission prediction to reduce preventable readmissions and improve outcomes is significant, yet several critical areas require further research to fully realize this potential. At the data level, there is an urgent need for standardization of EHR databases and cohorts. Harmonized data collection, consistent inclusion and exclusion criteria, and unified definitions of readmission timeframes would allow more comparable and reliable research across institutions. Hospitals should also aim to improve data entry practices, for example by providing training for nurses and physicians to reduce documentation errors. Extraction pipelines should be enhanced to capture the full spectrum of available variables rather than only those deemed relevant to a particular research problem, ensuring that valuable predictors are not systematically overlooked.

A second priority is the expansion of variable scope and multimodal data integration. Future pipelines should include clinically validated predictors that were absent from this study, such as severity-of-illness scores (e.g., PELOD), discharge readiness scores, neurological status, surgical interventions, device dependence, and nutritional interventions. These factors have been repeatedly identified in the literature as strong predictors of readmission. In addition, future systems should leverage modalities beyond structured tabular data : free-text notes, discharge summaries, and medical images such as X-rays and MRI scans can provide essential contextual and diagnostic information. Disease representation also requires advancement : while CCSR categories improved interpretability, they lacked depth compared to embeddings. Promising avenues include graph-based and hierarchical embeddings, which can capture relationships among diseases, and the integration of external medical ontologies (e.g., UMLS, SNOMED), which can ground embeddings in clinically meaningful hierarchies.

From a methodological perspective, data preprocessing and feature representation remain areas where systematic benchmarking is needed. Our category-aware imputation and Kalman–wavelet filtering approaches were promising but require comparison against established methods such as MICE and alternative filtering strategies. Normalization of physiological values should also be revisited ; while we used quantile thresholds, z-scores are widely used in pediatrics and may more accurately capture underlying distributions. In terms of feature design, future work should expand the repertoire of temporal and spectral features and move toward direct representation learning from raw signals, using autoencoders or sequence embedding models to reduce reliance on hand-crafted features.

Data augmentation and representation learning are particularly pressing given the small size and imbalance of pediatric datasets. Traditional oversampling techniques such as SMOTE proved insufficient, highlighting the need for clinically constrained augmentation methods that preserve temporal structure and physiological plausibility. Deep generative models, variational autoencoders, and self-supervised pretraining on large EHR corpora could be used to generate realistic synthetic trajectories, reduce variance, and support the training of more complex architectures. Reinforcement learning and transfer learning represent further opportunities to overcome small-sample limitations and better exploit related datasets. As the database continues to grow, incremental updating of models with new cases, potentially through reinforcement learning or transfer learning, could mitigate data scarcity.

At the modeling level, training strategies and optimization methods should also be refined. Our feature selection relied on a greedy, three-stage pipeline and an objective function that prioritized improving weaker folds, but this approach may not yield globally optimal feature subsets. Future work should explore stability selection, multi-objective optimization, and variance-reducing alternatives. Hyperparameter optimization should move beyond exhaustive grid search toward more efficient methods such as Bayesian optimization, random search, or Hyperband. Deeper and more diverse architectures are needed, including ensemble approaches combining multiple models, as well as multitask architectures that learn shared representations across related prediction tasks, may further enhance performance and robustness.

Achieving clinical deployment and generalization requires several additional steps. Scaling beyond a single-center to multi-center pediatric cohorts, coupled with external validation and temporal recalibration, is essential for generalizability. Interpretability also requires expansion : while our LR model provided global and local explanations, deep models must be paired with post-hoc explanation methods such as DeepSHAP. Local interpretability should be strengthened to make predictions clinically actionable. Visual tools that dynamically re-

present patient trajectories and risk factors could help physicians understand and trust model outputs. Prediction systems should evolve from static discharge-only estimates to dynamic, time-updated risk predictions refreshed every few hours during a patient's stay. Models must be prospectively evaluated in real-time workflows and hardened with distribution-drift monitoring, fairness audits across subgroups, and scheduled recalibration or retraining.

Finally, leveraging cloud technologies would also facilitate model updating and retraining periodically using continuously accumulating data, support deployment across multiple hospitals, and reduce the hardware burden on individual institutions. With these advances, predictive pipelines can mature from retrospective analyses into clinically reliable decision-support tools that reduce preventable readmissions, optimize discharge decisions, and ultimately improve both patient outcomes and resource use. Cloud-based architectures offer a promising solution by enabling secure, distributed storage, rapid model inference, and integration with electronic health record systems in near real-time.

# REFERENCES


[1] S. J. Weaver *et al.*, "Promoting a culture of safety as a patient safety strategy : A systematic review," *Annals of Internal Medicine*, vol. 158, pp. 369–374, 3 2013. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/23460092/

[2] J. F. Bion and J. E. Heffner, "Challenges in the care of the acutely ill," *Lancet*, vol. 363, pp. 970–977, 3 2004.

[3] B. D. Winters *et al.*, "Rapid-response systems as a patient safety strategy : A systematic review," *Annals of Internal Medicine*, vol. 158, pp. 417–425, 3 2013. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/23460099/

[4] J. A. Town *et al.*, "Relationship between icu bed availability, icu readmission, and cardiac arrest in the general wards," *Critical Care Medicine*, vol. 42, pp. 2037–2041, 2014. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/24776607/

[5] P. Zajic *et al.*, "Intensive care unit caseload and workload and their association with outcomes in critically unwell patients : a large registry-based cohort analysis," *Critical Care*, vol. 28, 12 2024. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/39277756/

[6] T. Leary and S. Ridley, "Impact of an outreach team on re-admissions to a critical care unit," *Anaesthesia*, vol. 58, pp. 328–332, 4 2003.

[7] F. Karachi, R. Gosselink, and S. Hanekom, "Public sector physiotherapists' organisation and profile : Implications for intensive care service," *South African Journal of Physiotherapy*, vol. 79, 2023. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/37065455/

[8] J. I. Lai and *et al.*, "Readmission to the intensive care unit : A population-based approach," *Journal of the Formosan Medical Association*, vol. 111, pp. 504–509, 9 2012.

[9] A. L. Rosenberg and C. Watts, "Patients readmitted to icus : A systematic review of risk factors and outcomes," *Chest*, vol. 118, pp. 492–502, 2000. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/10936146/

[10] C. R. Ponzoni and *et al.*, "Readmission to the intensive care unit : Incidence, risk factors, resource use, and outcomes. a retrospective cohort study," *Annals of the American Thoracic Society*, vol. 14, pp. 1312–1319, 8 2017.

[11] M. Katsiari and *et al.*, "Predictors of adverse outcome early after icu discharge," *Int J Crit Care Emerg Med*, vol. 5, 6 2019.

[12] "Inesss : Publication : Modes d'organisation des services de soins intensifs : état de connaissances et indicateurs de qualité." [Online]. Available : https://www.inesss.qc.ca/publications/repertoire-des-publications/publication/modes-dorganisation-des-services-de-soins-intensifs-etat-de-connaissances-et-indicateurs-de-qualite.html

[13] C. K. McIlvennan, Z. J. Eapen, and L. A. Allen, "Hospital readmissions reduction program," *Circulation*, vol. 131, pp. 1796–1803, 2015. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/25986448/

[14] J. Montomoli, M. P. Hilty, and C. Ince, "Artificial intelligence in intensive care : moving towards clinical decision support systems," *Minerva Anestesiologica*, vol. 88, pp. 1066–1072, 12 2022. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/36287392/

[15] B. Bardak and M. Tan, "Improving clinical outcome predictions using convolution over medical entities with multimodal learning," *Artificial Intelligence in Medicine*, vol. 117, p. 102112, 7 2021.

[16] M. Herland, T. M. Khoshgoftaar, and R. Wald, "Survey of clinical data mining applications on big data in health informatics," *12th International Conference on Machine Learning and Applications*, vol. 2, pp. 465–472, 2013.

[17] Z. Rayan, M. Alfonse, and A. M. Salem, "Machine learning approaches in smart health," *Procedia Computer Science*, vol. 154, pp. 361–368, 1 2019.

[18] M. A. Dhoayan, H. Alghamdi, and Y. M. Arabi, "Machine learning applications in critical care," *Saudi Critical Care Journal*, vol. 3, p. 29, 2019.

[19] J. M. Lee and M. Hauskrecht, "Modeling multivariate clinical event time-series with recurrent temporal mechanisms," *Artificial Intelligence in Medicine*, vol. 112, p. 102021, 2 2021.

[20] M. M. Ahsan and Z. Siddique, "Machine learning-based heart disease diagnosis : A systematic literature review," *Artificial Intelligence in Medicine*, vol. 128, p. 102289, 6 2022.

[21] W. Fathy, G. Emeriaud, and F. Cheriet, "A comprehensive review of icu readmission prediction models : From statistical methods to deep learning approaches," *Artificial Intelligence in Medicine*, vol. 165, p. 103126, 7 2025. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S0933365725000612?via%3Dihub

[22] A. Mahmoodpoor *et al.*, "Prognostic value of national early warning score and modified early warning score on intensive care unit readmission and mortality : A prospective observational study," *Frontiers in Medicine*, vol. 9, 8 2022. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/35991649/

[23] S. Wang and *et al.*, "Mimic-extract : A data extraction, preprocessing, and representation pipeline for mimic-iii," in *The ACM Conference on Health, Inference, and Learning*, ser. CHIL '20. New York, NY, USA : Association for Computing Machinery, 2020, p. 222–235.

[24] S. C. Lu *et al.*, "On the importance of interpretable machine learning predictions to inform clinical decision making in oncology," *Frontiers in Oncology*, vol. 13, p. 1129380, 2023. [Online]. Available : https://pmc.ncbi.nlm.nih.gov/articles/PMC10013157/

[25] Q. Teng *et al.*, "A survey on the interpretability of deep learning in medical diagnosis," *Multimedia Systems*, vol. 28, p. 2335, 12 2022. [Online]. Available : https://pmc.ncbi.nlm.nih.gov/articles/PMC9243744/

[26] W. Fathy, G. Emeriaud, and F. Cheriet, "Prediction of icu readmission using lightgbm classifier," in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, 2023, pp. 1–4.

[27] W. A. Knaus *et al.*, "Apache-acute physiology and chronic health evaluation : a physiologically based classification system." *Critical care medicine*, vol. 9, pp. 591–597, 1981. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/7261642/

[28] J. L. Vincent *et al.*, "The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure," *Intensive Care Medicine*, vol. 22, pp. 707–710, 1996. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/8844239/

[29] J. R. L. Gall *et al.*, "A simplified acute physiology score for icu patients," *Critical Care Medicine*, vol. 12, pp. 975–977, 1984. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/6499483/

[30] D. P. Wagner and E. A. Draper, "Acute physiology and chronic health evaluation (apache ii) and medicare reimbursement," *Health Care Financing Review*, vol. 1984, p. 91, 1984. [Online]. Available : https://pmc.ncbi.nlm.nih.gov/articles/PMC4195105/

[31] V. Badilla-Morales *et al.*, "Factors associated with early readmission to intensive care units. a systematic review," *Enfermería Intensiva (English ed.)*, vol. 36, p. 100498, 4 2025. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S2529984025000102

[32] A. A. Kramer, T. L. Higgins, and J. E. Zimmerman, "Intensive care unit readmissions in u.s. hospitals : Patient characteristics, risk factors, and outcomes," *Critical Care Medicine*, vol. 40, pp. 3–10, 1 2012. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/21926603/

[33] A. L. Woldhek *et al.*, "Readmission of icu patients : A quality indicator ?" *Journal of Critical Care*, vol. 38, pp. 328–334, 4 2017. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/27939901/

[34] M. Rodríguez-Carvajal *et al.*, "Impact of the premature discharge on hospital mortality after a stay in an intensive care unit," *Medicina Intensiva*, vol. 35, pp. 143–149, 4 2011. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/21419522/

[35] A. Kaben *et al.*, "Readmission to a surgical intensive care unit : incidence, outcome and risk factors," *Critical Care*, vol. 12, p. R123, 10 2008. [Online]. Available : https://pmc.ncbi.nlm.nih.gov/articles/PMC2592757/

[36] O. Gajic *et al.*, "The stability and workload index for transfer score predicts unplanned intensive care unit patient readmission : Initial development and validation," *Critical Care Medicine*, vol. 36, pp. 676–682, 2008. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/18431260/

[37] B. Williams, "The national early warning score : from concept to nhs implementation," *Clinical Medicine*, vol. 22, p. 499, 11 2022. [Online]. Available : https://pmc.ncbi.nlm.nih.gov/articles/PMC9761416/

[38] C. Stenhouse *et al.*, "Prospective evaluation of a modified early warning score to aid earlier detection of patients developing critical illness on a general surgical ward," *British Journal of Anaesthesia*, vol. 84, p. 663P, 5 2000. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S0007091217380480

[39] M. Kastrup *et al.*, "Predictive ability of the stability and workload index for transfer score to predict unplanned readmissions after icu discharge," *Critical Care Medicine*, vol. 41, pp. 1608–1615, 7 2013. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/23660731/

[40] C. Doğu *et al.*, "Importance of the national early warning score (news) at the time of discharge from the intensive care unit," *Turkish Journal of Medical Sciences*, vol. 50, p. 1203, 2020. [Online]. Available : https://pmc.ncbi.nlm.nih.gov/articles/PMC7491295/

[41] A. N. Balshi *et al.*, "Modified early warning score as a predictor of intensive care unit readmission within 48 hours : a retrospective observational study," *Revista Brasileira de Terapia Intensiva*, vol. 32, p. 301, 6 2020. [Online]. Available : https://pmc.ncbi.nlm.nih.gov/articles/PMC7405753/

[42] J. Haruna *et al.*, "Nursing activities score at discharge from the intensive care unit is associated with unplanned readmission to the intensive care unit," *Journal of Clinical Medicine*, vol. 11, 9 2022. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/36079134/

[43] S. E. Brown *et al.*, "The epidemiology of intensive care unit readmissions in the united states," *American Journal of Respiratory and Critical Care Medicine*, vol. 185, pp. 955–964, 5 2012. [Online]. Available : https://www.atsjournals.org/doi/pdf/10.1164/rccm.201109-1720OC?download=true

[44] R. Maharaj, M. Terblanche, and S. Vlachos, "The utility of icu readmission as a quality indicator and the effect of selection," *Critical Care Medicine*, vol. 46, pp. 749–756, 5 2018. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/29394182/

[45] R. Tangonan *et al.*, "Frequency, risk factors, and outcomes of unplanned readmission to the neurological intensive care unit after spontaneous intracerebral hemorrhage," *Neurocritical Care*, vol. 37, pp. 390–398, 10 2022. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/35072926/

[46] A. L. Rosenberg *et al.*, "Who bounces back ? physiologic and other predictors of intensive care unit readmission," *Critical Care Medicine*, vol. 29, pp. 511–518, 2001. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/11373413/

[47] E. E. T. Álvarez *et al.*, "Risk factors for readmission to icu and analysis of intra-hospital mortality," *Medicina Clinica*, vol. 158, pp. 58–64, 1 2022. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/33516522/

[48] T. Hanane *et al.*, "The association between nighttime transfer from the intensive care unit and patient outcome," *Critical Care Medicine*, vol. 36, pp. 2232–2237, 2008. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/18664778/

[49] L. C. Azevedo *et al.*, "Association between nighttime discharge from the intensive care unit and hospital mortality : A multi-center retrospective cohort study," *BMC Health Services Research*, vol. 15, pp. 1–9, 9 2015. [Online]. Available : https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-015-1044-4

[50] D. Gantner *et al.*, "Mortality related to after-hours discharge from intensive care in australia and new zealand, 2005–2012," *Intensive Care Medicine*, vol. 40, pp. 1528–1535, 9 2014. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/25118868/

[51] G. J. Duke, J. V. Green, and J. H. Briedis, "Night-shift discharge from intensive care unit increases the mortality-risk of icu survivors," *Anaesthesia and Intensive Care*, vol. 32, pp. 697–701, 2004. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/15535498/

[52] S. Chatterjee, S. Sinha, and S. K. Todi, "Transfer time from the intensive care unit and patient outcome : A retrospective analysis from a tertiary care hospital in india," *Indian Journal of Critical Care Medicine*, vol. 23, pp. 115–121, 3 2019. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/31097886/

[53] F. A. Priestap and C. M. Martin, "Impact of intensive care unit discharge time on patient outcome," *Critical Care Medicine*, vol. 34, pp. 2946–2951, 12 2006. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/17075364/

[54] J. L. Nates *et al.*, "Icu admission, discharge, and triage guidelines : A framework to enhance clinical operations, development of institutional policies, and further

research," *Critical Care Medicine*, vol. 44, pp. 1553–1602, 8 2016. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/27428118/

[55] C. A. Chrusch *et al.*, "High occupancy increases the risk of early death or readmission after transfer from intensive care," *Critical Care Medicine*, vol. 37, pp. 2753–2758, 2009. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/19707139/

[56] K. M. Sauro *et al.*, "Adverse events after transition from icu to hospital ward : A multicenter cohort study," *Critical Care Medicine*, vol. 48, pp. 946–953, 7 2020. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/32317594/

[57] N. V. Sluisveld *et al.*, "Variation in rates of icu readmissions and post-icu in-hospital mortality and their association with icu discharge practices," *BMC Health Services Research*, vol. 17, p. 281, 4 2017. [Online]. Available : https://pmc.ncbi.nlm.nih.gov/articles/PMC5393034/

[58] R. L. Hoffman *et al.*, "Development and implementation of a risk identification tool to facilitate critical care transitions for high-risk surgical patients," *International Journal for Quality in Health Care*, vol. 29, pp. 412–419, 6 2017. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/28371889/

[59] H. T. Stelfox *et al.*, "Critical care transition programs and the risk of readmission or death after discharge from icu," *Intensive Care Medicine*, vol. 42, pp. 401–410, 3 2016. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/26694189/

[60] D. G. Coughlin *et al.*, "Preventing early bouncebacks to the neurointensive care unit : A retrospective analysis and quality improvement pilot," *Neurocritical Care*, vol. 28, pp. 175–183, 4 2018. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/28929392/

[61] E. A. Sharp *et al.*, "Frequency, characteristics, and outcomes of patients requiring early picu readmission," *Hospital Pediatrics*, vol. 13, pp. 678–687, 2023. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/37476936/

[62] J. D. Edwards *et al.*, "Frequency, risk factors, and outcomes of early unplanned readmissions to picus," *Critical Care Medicine*, vol. 41, pp. 2773–2783, 12 2013. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/23982030/

[63] I. M. Mandell *et al.*, "Pediatric early warning score and unplanned readmission to the pediatric intensive care unit," *Journal of Critical Care*, vol. 30, pp. 1090–1095, 10 2015. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/26235654/

[64] H. Kaur *et al.*, "Proper : Development of an early pediatric intensive care unit readmission risk prediction tool," *Journal of Intensive Care Medicine*, vol. 33, pp. 29–36, 1 2018. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/27601481/

[65] H. Duncan, J. Hutchison, and C. S. Parshuram, "The pediatric early warning system score : A severity of illness score to predict urgent medical need in hospitalized children," *Journal of Critical Care*, vol. 21, pp. 271–278, 9 2006. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/16990097/

[66] M. H. Zoham *et al.*, "Validity of pediatric early warning score in predicting unplanned pediatric intensive care unit readmission," *Journal of Pediatric Intensive Care*, vol. 12, p. 312, 12 2021. [Online]. Available : https://pmc.ncbi.nlm.nih.gov/articles/PMC10631837/

[67] A. Kotsakis *et al.*, "Description of picu unplanned readmission," *Pediatric Critical Care Medicine*, vol. 17, pp. 558–562, 6 2016. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/27261644/

[68] A. M. Bernard and A. S. Czaja, "Unplanned pediatric intensive care unit readmissions : A single-center experience," *Journal of Critical Care*, vol. 28, pp. 625–633, 10 2013. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/23602033/

[69] M. N. M. Bah *et al.*, "Unplanned picu readmission in a middle-income country : Who is at risk and what is the outcome ?," *Pediatric Critical Care Medicine*, vol. 21, pp. E959–E966, 11 2020. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/32590834/

[70] M. K. Barnwell, H. Zhou, and S. Erickson, "Prevalence and risk factors associated within 48-hour unplanned paediatric intensive care unit readmissions : An integrative review," *Australian Critical Care*, vol. 38, 1 2024. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/38724409/

[71] G. J. Prutsky *et al.*, "Is unplanned picu readmission a proper quality indicator ? a systematic review and meta-analysis," *Hospital Pediatrics*, vol. 11, pp. 167–174, 2 2021. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/33504562/

[72] M. A. Bhat, G. Soto-Campos, and M. C. Scanlon, "Relationship between pediatric intensive care unit length of stay and 24-h unplanned readmission rate," *Health Services Research*, vol. 57, pp. 598–602, 6 2022. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/35149985/

[73] A. H. Smith *et al.*, "Variation in case-mix adjusted unplanned pediatric cardiac intensive care unit readmission rates," *Critical care medicine*, vol. 46, p. e1175, 12 2018. [Online]. Available : https://pmc.ncbi.nlm.nih.gov/articles/PMC6239958/

[74] S. Linton *et al.*, "The development of a clinical markers score to predict readmission to paediatric intensive care," *Intensive and Critical Care Nursing*, vol. 25, pp. 283–293, 12 2009. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/19819698/

[75] F. O. Odetola *et al.*, "Going back for more : An evaluation of clinical outcomes and characteristics of readmissions to a pediatric intensive care unit,"

*Pediatric Critical Care Medicine*, vol. 8, pp. 343–347, 2007. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/17545926/

[76] A. E. Johnson *et al.*, "Mimic-iii, a freely accessible critical care database," *Scientific Data*, vol. 3, pp. 1–9, 5 2016. [Online]. Available : https://www.nature.com/articles/sdata201635

[77] T. J. Pollard *et al.*, "The eicu collaborative research database, a freely available multi-center database for critical care research," *Scientific Data*, vol. 5, pp. 1–13, 9 2018. [Online]. Available : https://www.nature.com/articles/sdata2018178

[78] Y. Xue, D. Klabjan, and Y. Luo, "Predicting icu readmission using grouped physiological and medication trends," *Artificial Intelligence in Medicine*, vol. 95, pp. 27–37, apr 2019.

[79] C. M. Salgado, S. M. Vieira, and J. M. Sousa, "Fuzzy modeling based on mixed fuzzy clustering for multivariate time series of unequal lengths," *Communications in Computer and Information Science*, vol. 611, pp. 741–751, 2016.

[80] L. He *et al.*, "An embedded machine learning model for early detection and intervention of high-risk intensive care unit readmission patients," *Proceedings - 2022 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2022*, pp. 1544–1549, 2022.

[81] Y. Zhu and *et al.*, "Domain adaptation using convolutional autoencoder and gradient boosting for adverse events prediction in the intensive care unit," *Frontiers in artificial intelligence*, vol. 5, 4 2022.

[82] Z. Li and *et al.*, "Early prediction of 30-day icu re-admissions using natural language processing and machine learning," *Biomedical Statistics and Informatics*, vol. 4, no. 3, p. 22, 2019.

[83] A. Moerschbacher and Z. He, "Building prediction models for 30-day readmissions among icu patients using both structured and unstructured data in electronic health records," *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 4368–4373, 2023.

[84] L. Wang and *et al.*, "Care-30 : A causally driven multi-modal model for enhanced 30-day icu readmission predictions," *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1509–1516, 12 2023.

[85] R. M. S. Carvalho, D. Oliveira, and C. Pesquita, "Knowledge graph embeddings for icu readmission prediction," *BMC Med Inform Decis Mak*, vol. 23, pp. 1–17, 12 2023.

[86] K. Shi and *et al.*, "Predicting unplanned 7-day intensive care unit readmissions with machine learning models for improved discharge risk assessment," *AMIA Annual Sym-*

*posium Proceedings*, vol. 2022, p. 446, 2022.

[87] S. Hegselmann and *et al.*, "Development and validation of an interpretable 3 day intensive care unit readmission prediction model using explainable boosting machines," *Frontiers in Medicine*, vol. 9, 8 2022.

[88] Q. CHEN and *et al.*, "Outcome-oriented predictive process monitoring to predict unplanned icu readmission in mimic-iv database," *ECIS Research-in-Progress Papers*, jun 2022.

[89] O. Badawi and M. J. Breslow, "Readmissions and death after icu discharge : Development and validation of two predictive models," *PLOS ONE*, vol. 7, p. e48758, 11 2012.

[90] A. Khodadadi and *et al.*, "Improving diagnostics with deep forest applied to electronic health records," *Sensors*, vol. 23, 7 2023.

[91] D. Cai and *et al.*, *Hypergraph Contrastive Learning for Electronic Health Records*. Society for Industrial and Applied Mathematics Publications, 2022, pp. 127–135.

[92] M. Sun and *et al.*, "A cross-modal clinical prediction system for intensive care unit patient outcome," *Knowledge-Based Systems*, vol. 283, p. 111160, 1 2024.

[93] A. J. Campbell and *et al.*, "Predicting death and readmission after intensive care discharge," *British journal of anaesthesia*, vol. 100, pp. 656–662, 2008.

[94] Y. S. Jo and *et al.*, "Readmission to medical intensive care units : Risk factors and prediction," *Yonsei Medical Journal*, vol. 56, p. 543, 3 2015.

[95] S. A. Frost and *et al.*, "Readmission to intensive care : development of a nomogram for individualising risk," *Crit Care Resusc*, 2010.

[96] I. Ouanes and *et al.*, "A model to predict short-term death or readmission after intensive care unit discharge," *Journal of Critical Care*, vol. 27, pp. 422.e1–422.e9, 8 2012.

[97] A. S. Fialho and *et al.*, "Predicting intensive care unit readmissions using probabilistic fuzzy systems," *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2013.

[98] P. Braga and *et al.*, "Data mining models to predict patient's readmission in intensive care units," *ICAART 2014 - Proceedings of the 6th International Conference on Agents and Artificial Intelligence*, vol. 1, pp. 604–610, 2014.

[99] R. Veloso and *et al.*, "A clustering approach for predicting readmissions in intensive medicine," *Procedia Technology*, vol. 16, pp. 1307–1316, 1 2014.

[100] B. Wang and *et al.*, "Predictive classification of icu readmission using weight decay random forest," *Future Generation Computer Systems*, vol. 124, pp. 351–360, 11 2021.

[101] J. C. Rojas and *et al.*, "Predicting intensive care unit readmission with machine learning using electronic health record data," *Ann. Am. Thorac. Soc.*, vol. 15, no. 7, pp. 846–853, jun 2018.

[102] M. Loreto, T. Lisboa, and V. P. Moreira, "Early prediction of icu readmissions using classification algorithms," *Computers in Biology and Medicine*, vol. 118, p. 103636, 3 2020.

[103] T. Desautels and *et al.*, "Prediction of early unplanned intensive care unit readmission in a uk tertiary care hospital : a cross-sectional machine learning approach," *BMJ Open*, vol. 7, no. 9, p. e017199, sep 2017.

[104] P. J. Thoral and *et al.*, "Explainable machine learning on amsterdamumcdb for icu discharge decision support : Uniting intensivists and data scientists," *Critical Care Explorations*, vol. 3, p. e0529, 9 2021.

[105] A. A. D. Hond and *et al.*, "Predicting readmission or death after discharge from the icu : External validation and retraining of a machine learning model," *Critical Care Medicine*, vol. 51, pp. 291–300, 2 2023.

[106] B. Shickel and *et al.*, "Multi-dimensional patient acuity estimation with longitudinal ehr tokenization and flexible transformer networks," *Frontiers in digital health*, vol. 4, 11 2022.

[107] S. Curto and *et al.*, "Predicting icu readmissions based on bedside medical text notes," *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 2144–2151, nov 2016.

[108] J. Venugopalan and *et al.*, "Combination of static and temporal data analysis to predict mortality and readmission in the intensive care," *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, vol. 2017, p. 2570, sep 2017.

[109] S. Darabi and *et al.*, "Taper : Time-aware patient ehr representation," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 11, pp. 3268–3275, aug 2019.

[110] A. S. Fialho and *et al.*, "Data mining using clinical physiology at discharge to predict icu readmissions," *Expert Systems with Applications*, vol. 39, no. 18, pp. 13 158–13 165, dec 2012.

[111] J. A. G. Sargo and *et al.*, "Binary fish school search applied to feature selection : Application to icu readmissions," *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1366–1373, sep 2014.

[112] R. Viegas and *et al.*, "Daily prediction of icu readmissions using feature engineering and ensemble fuzzy modeling," *Expert Systems with Applications*, vol. 79, pp. 244–253,

aug 2017.

[113] Y. W. Lin and *et al.*, "Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory," *PLOS ONE*, vol. 14, no. 7, p. e0218942, jul 2019.

[114] A. Pakbin and *et al.*, "Prediction of icu readmissions using data at patient discharge," *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, vol. 2018, pp. 4932–4935, oct 2018.

[115] K. Caballero and R. Akella, "Dynamic estimation of the probability of patient readmission to the icu using electronic medical records," *AMIA Annual Symposium Proceedings*, vol. 2015, p. 1831, 2015.

[116] E. Sheetrit, M. Brief, and O. Elisha, "Predicting unplanned readmissions in the intensive care unit : a multimodality evaluation," *Scientific Reports 2023 13 :1*, vol. 13, pp. 1–9, 9 2023.

[117] S. Barbieri and *et al.*, "Benchmarking deep learning architectures for predicting readmission to the icu and describing patients-at-risk," *Sci Rep*, vol. 10, no. 1, pp. 1–10, jan 2020.

[118] Y. Choi, C. M. Y.-I. Chiu, and D. Sontag, "Learning low-dimensional representations of medical concepts," *AMIA Summits on Translational Science Proceedings*, vol. 2016, p. 41, 2016.

[119] "Clinical classifications software (ccs) for icd-9-cm." [Online]. Available : https://hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

[120] T. Zebin and T. J. Chaussalet, "Design and implementation of a deep recurrent model for prediction of readmission in urgent care using electronic health records," *EEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, jul 2019.

[121] N. Orangi-Fard, A. Akhbardeh, and H. Sagreiya, "Predictive model for icu readmission based on discharge summaries using machine learning and natural language processing," *Informatics*, vol. 9, no. 1, p. 10, jan 2022.

[122] S. Jain, R. Mohammadi, and B. C. Wallace, "An analysis of attention over clinical notes for predictive tasks," in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, jun 2019, pp. 15–21.

[123] T. Mikolov and *et al.*, "Efficient estimation of word representations in vector space," in *International Conference on Learning Representations*, 1 2013.

[124] Y. Zhang and *et al.*, "Biowordvec, improving biomedical word embeddings with subword information and mesh," *Scientific Data*, vol. 6, pp. 1–9, 5 2019.

[125] J. Lovelace and *et al.*, "Dynamically extracting outcome-specific problem lists from clinical notes with guided multi-headed attention," pp. 245–270, 9 2020.

[126] J. Devlin and *et al.*, "Bert : Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics*, vol. 1, 10 2019, pp. 4171–4186.

[127] J. Lee and *et al.*, "Biobert : a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 09 2019.

[128] K. Huang, J. Altosaar, and R. Ranganath, "Clinicalbert : Modeling clinical notes and predicting hospital readmission," *ArXiv*, vol. abs/1904.05342, 2019.

[129] M. Nickel and D. Kiela, "Poincaré embeddings for learning hierarchical representations," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[130] Q. Lu and *et al.*, "Learning electronic health records through hyperbolic embedding of medical ontologies," in *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, ser. BCB '19. New York, NY, USA : Association for Computing Machinery, 2019, p. 338–346.

[131] B. L. Humphreys and *et al.*, "The unified medical language system : An informatics research collaboration," *Journal of the American Medical Informatics Association*, vol. 5, pp. 1–11, 1 1998.

[132] "Snomed international." [Online]. Available : https://www.snomed.org/

[133] X. Zhang, D. Dou, and J. Wu, "Learning conceptual-contextual embeddings for medical text," *AAAI*, vol. 34, no. 05, pp. 9579–9586, apr 2020.

[134] T. Wu and *et al.*, "Leveraging graph-based hierarchical medical entity embedding for healthcare applications," *Sci Rep*, vol. 11, no. 1, pp. 1–13, mar 2021.

[135] E. Choi and *et al.*, "Learning the graphical structure of electronic health records with graph convolutional transformer," *AAAI Conference on Artificial Intelligence*, pp. 606–613, jun 2020.

[136] X. Liu and *et al.*, "Research on intelligent diagnosis model of electronic medical record based on graph transformer," in *6th International Conference on Computational Intelligence and Applications (ICCIA)*, 2021, pp. 73–78.

[137] W. Zhu and N. Razavian, "Variationally regularized graph-based representation learning for electronic health records," in *Proceedings of the Conference on Health, Inference, and Learning*. New York, NY, USA : Association for Computing Machinery, 2021, p. 1–13.

[138] Q. Lu, T. H. Nguyen, and D. Dou, "Predicting patient readmission risk from medical text via knowledge graph enhanced multiview graph convolution," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, dec 2021, p. 1990–1994.

[139] S. Pei and *et al.*, "Readmission prediction with knowledge graph attention and rnn-based ordinary differential equations," *Knowledge Science, Engineering and Management*, vol. 12817 LNAI, pp. 559–570, 2021.

[140] Y. L. Oord, Aaron V. and O. Vinyals, "Representation learning with contrastive predictive coding," 2018.

[141] N. Haribhakti and *et al.*, "A simple scoring tool to predict medical intensive care unit readmissions based on both patient and process factors," *Journal of General Internal Medicine*, vol. 36, pp. 901–907, 4 2021.

[142] M. J. Azur and *et al.*, "Multiple imputation by chained equations : what is it and how does it work ?" *International Journal of Methods in Psychiatric Research*, vol. 20, p. 40, 3 2011.

[143] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-15, pp. 116–132, 1985.

[144] J. C. Bezdek, "Pattern recognition with fuzzy objective function algorithms," *Advanced Applications in Pattern Recognition*, 1981.

[145] H. Izakian, W. Pedrycz, and I. Jamal, "Clustering spatiotemporal data : An augmented fuzzy c-means," *IEEE Transactions on Fuzzy Systems*, vol. 21, pp. 855–868, 2013.

[146] J. V. D. Berg, U. Kaymak, and W. M. V. D. Bergh, "Fuzzy classification using probability-based rule weighting," *IEEE International Conference on Fuzzy Systems*, vol. 2, pp. 991–996, 2002.

[147] S. M. Vieira and *et al.*, "A decision support system for icu readmissions prevention," *Joint IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS)*, pp. 251–256, 2013.

[148] N. Homem and J. P. Carvalho, "Authorship identification and author fuzzy "fingerprints"," in *Annual Meeting of the North American Fuzzy Information Processing Society*, 2011, pp. 1–6.

[149] C.-H. Chen, T.-P. Hong, and V. S. Tseng, "Finding pareto-front membership functions in fuzzy data mining," *International Journal of Computational Intelligence Systems*, vol. 5, no. 2, pp. 343–354, 2012.

[150] C. J. Filho and *et al.*, "A novel search algorithm based on fish school behavior," *IEEE International Conference on Systems, Man and Cybernetics*, pp. 2646–2651, 2008.

[151] M. C. Ferreira and *et al.*, "Fuzzy modeling based on mixed fuzzy clustering for health care applications," *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, vol. 2015-Novem, nov 2015.

[152] M. P. Fernandes and *et al.*, "Multimodeling for the prediction of patient readmissions in intensive care units," *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1837–1842, sep 2014.

[153] C. M. Salgado and *et al.*, "Ensemble fuzzy classifiers design using weighted aggregation criteria," *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, vol. 2015-Novem, nov 2015.

[154] D. E. Gustafson and W. C. Kessel, "Fuzzy clustering with a fuzzy covariance matrix," *IEEE Conference on Decision and Control*, pp. 761–766, 1978.

[155] M. West and J. Harrison, *The Dynamic Linear Model*. Springer, New York, NY, 1989, pp. 105–141.

[156] J. D. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields : Probabilistic models for segmenting and labeling sequence data," in *International Conference on Machine Learning*, 2001.

[157] Z. R. Yang and Z. Yang, "Artificial neural networks," *Comprehensive Biomedical Physics*, vol. 6, pp. 1–17, 7 2014.

[158] T. T. Inan and *et al.*, "A decision support model to predict icu readmission through data mining approach," *PACIS 2018 Proceedings*, jun 2018.

[159] A. R. Junqueira, F. Mirza, and M. M. Baig, "A machine learning model for predicting icu readmissions and key risk factors : analysis from a longitudinal health records," *Health Technol.*, vol. 9, no. 3, pp. 297–309, may 2019.

[160] L. Zhang and X. Chen, "Feature selection methods based on symmetric uncertainty coefficients and independent classification information," *IEEE Access*, vol. 9, pp. 13 845–13 856, 2021.

[161] S. Raza and S. R. Bashir, "Auditing icu readmission rates in an clinical database : An analysis of risk factors and clinical outcomes," *IEEE International Conference on Healthcare Informatics*, pp. 722–726, 2023.

[162] G. I. Webb, "Naïve bayes," *Encyclopedia of Machine Learning*, pp. 713–714, 2011.

[163] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pp. 144–152, 1992.

[164] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81–106, 3 1986.

[165] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 10 2001.

[166] C. Silva, S. M. Vieira, and J. M. Sousa, "Fuzzy decision tree to predict readmissions in intensive care unit," *Lecture Notes in Electrical Engineering*, vol. 321 LNEE, pp. 365–373, 2015.

[167] X. Liu and W. Pedrycz, "The development of fuzzy decision trees in the framework of axiomatic fuzzy set logic," *Applied Soft Computing Journal*, vol. 7, pp. 325–342, 1 2007.

[168] K. Alghatani and *et al.*, "Precision clinical medicine through machine learning : Using high and low quantile ranges of vital signs for risk stratification of icu patients," *IEEE Access*, vol. 10, pp. 52 418–52 430, may 2022.

[169] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," *Lecture Notes in Computer Science*, vol. 904, pp. 23–37, 1995.

[170] J. H. Friedman, "Greedy function approximation : A gradient boosting machine," *The Annals of Statistics*, vol. 29, pp. 1189–1232, 2001.

[171] T. Chen and C. Guestrin, "Xgboost : A scalable tree boosting system," *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-August-2016, pp. 785–794, 3 2016.

[172] X. Yao, X. Fu, and C. Zong, "Short-term load forecasting method based on feature preference strategy and lightgbm-xgboost," *IEEE Access*, vol. 10, pp. 75 257–75 268, 2022. [Online]. Available : https://ieeexplore.ieee.org/document/9832627

[173] G. Ke and *et al.*, "Lightgbm : A highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[174] L. Jiang *et al.*, "Lightcpg : a multi-view cpg sites detection on single-cell whole genome sequence data," *BMC Genomics 2019 20 :1*, vol. 20, pp. 1–17, 4 2019. [Online]. Available : https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-5654-9

[175] J. barbier and *et al.*, "Mutual information for symmetric rank-one matrix estimation : A proof of the replica formula," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[176] N. V. Chawla and *et al.*, "Smote : Synthetic minority over-sampling technique," *Artificial Intelligence Research*, vol. 16, pp. 321–357, 6 2011.

[177] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *International Conference on Neural Information Processing Systems*, vol. 2017-December, pp. 4766–4775, 5 2017.

[178] J. A. González-Nóvoa and *et al.*, "Improving intensive care unit early readmission prediction using optimized and explainable machine learning," *International Journal of Environmental Research and Public Health*, vol. 20, 2 2023.

[179] S. Watanabe, "Tree-structured parzen estimator : Understanding its algorithm components and their roles for better empirical performance," *arXiv preprint arXiv :2304.11127*, 4 2023.

[180] X. Jin and J. Han, "K-means clustering," *Encyclopedia of Machine Learning*, pp. 563–564, 2011.

[181] H. S. Park and C. H. Jun, "A simple and fast algorithm for k-medoids clustering," *Expert Systems with Applications*, vol. 36, pp. 3336–3341, 3 2009.

[182] D. Pelleg and A. W. Moore, "X-means : Extending k-means with efficient estimation of the number of clusters," in *Proceedings of the Seventeenth International Conference on Machine Learning*, ser. ICML '00. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 2000, p. 727–734.

[183] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," *Physica D : Nonlinear Phenomena*, vol. 404, p. 132306, 3 2020.

[184] J. Teuwen and N. Moriakov, "Convolutional neural networks," *Handbook of Medical Image Computing and Computer Assisted Intervention*, pp. 481–501, 1 2020.

[185] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, 11 1997.

[186] J. Chung and *et al.*, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *NIPS Workshop on Deep Learning*, 12 2014.

[187] O. Calzone, "An intuitive explanation of lstm. recurrent neural networks medium." [Online]. Available : https://medium.com/@ottaviocalzone/an-intuitive-explanation-of-lstm-a035eb6ab42c

[188] "Understandin convolutional neural network (cnn) architecture." [Online]. Available : https://www.codecademy.com/article/understanding-convolutional-neural-network-cnn-architecture

[189] A. Vaswani and *et al.*, "Attention is all you need," *Neural Information Processing Systems (NIPS)*, pp. 5999–6009, 6 2017.

[190] J. C. Denny and *et al.*, "Phewas : demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations," *Bioinformatics*, vol. 26, pp. 1205–1210, 5 2010.

[191] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer : The long-document transformer," *arxiv*, 4 2020.

[192] Q. Lu, D. Dou, and T. H. Nguyen, "Textual data augmentation for patient outcomes prediction," *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2817–2821, 2021.

[193] R. T. Q. Chen and *et al.*, "Neural ordinary differential equations," *Neural Information Processing Systems (NIPS)*, vol. 109, pp. 31–60, 6 2018.

[194] K. Niu and *et al.*, "Intensive care unit readmission prediction with correlation enhanced multi-task learning," *Computers and Electrical Engineering*, vol. 110, p. 108780, 9 2023.

[195] J. Zhou and *et al.*, "Graph neural networks : A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, 1 2020.

[196] S. Zhang and *et al.*, "Graph convolutional networks : a comprehensive review," *Computational Social Networks*, vol. 6, pp. 1–23, 12 2019.

[197] P. Veličković and *et al.*, "Graph attention networks," *International Conference on Learning Representations (ICLR)*, 10 2017.

[198] P. Leleux and *et al.*, "Design of biased random walks on a graph with application to collaborative recommendation," *Physica A : Statistical Mechanics and its Applications*, vol. 590, p. 126752, 3 2022.

[199] I. Ng and *et al.*, "A graph autoencoder approach to causal structure learning," *ArXiv*, 11 2019.

[200] R. Foraita, J. Spallek, and H. Zeeb, "Directed acyclic graphs," *Handbook of Epidemiology : Second Edition*, pp. 1481–1517, 1 2014.

[201] K. Kuang and *et al.*, "Causal inference," *Engineering*, vol. 6, pp. 253–263, 3 2020.

[202] C. Chen and *et al.*, "Hypergraph attention networks," *IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pp. 1560–1565, 12 2020.

[203] M. R. Khan and J. E. Blumenstock, "Multi-gcn : Graph convolutional networks for multi-view networks, with applications to global poverty," *The AAAI Conference on Artificial Intelligence*, vol. 33, pp. 606–613, 7 2019.

[204] P. Ristoski and *et al.*, "Rdf2vec : Rdf graph embeddings and their applications," *Semantic Web*, vol. 10, pp. 721–752, 2019.

[205] "Nci thesaurus." [Online]. Available : https://ncit.nci.nih.gov/ncitbrowser/

[206] M. Martínez-Romero and *et al.*, "Ncbo ontology recommender 2.0 : An enhanced approach for biomedical ontology recommendation," *Journal of Biomedical Semantics*, vol. 8, pp. 1–22, 6 2017.

[207] E. Laksana *et al.*, "The impact of extraneous features on the performance of recurrent neural network models in clinical tasks," *Journal of Biomedical Informatics*, vol. 102, p. 103351, 2 2020. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S1532046419302709

[208] A. Arshad *et al.*, "490 : Predicting intensive care readmission among hospitalized children," *Critical Care Medicine*, vol. 50, pp. 236–236, 1 2022. [Online]. Available : https://journals.lww.com/ccmjournal/fulltext/2022/01001/490___predicting_intensive_care_readmission_among.456.aspx

[209] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?" : Explaining the predictions of any classifier," *Conference of the North American Chapter of the Association for Computational Linguistics : Demonstrations*, pp. 97–101, 2 2016.

[210] D. Slack and *et al.*, "Fooling lime and shap : Adversarial attacks on post hoc explanation methods," *The AAAI/ACM Conference on AI, Ethics, and Society*, pp. 180–186, 11 2019.

[211] F. J. Massey, "The kolmogorov-smirnov test for goodness of fit," *The American Statistical Association*, vol. 46, p. 68, 3 1951.

[212] G. W. BRIER, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78, pp. 1–3, 1 1950.

[213] D. W. Hosmer and S. Lemeshow, "Goodness of fit tests for the multiple logistic regression model," *Communications in Statistics - Theory and Methods*, vol. 9, pp. 1043–1069, 1 1980.

[214] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research (JMLR)*, vol. 7, pp. 1–30, 2006.

[215] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *The American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[216] M. Drummond and *et al.*, *Methods for the Economic Evaluation of Health Care Programmes*. Oxford University Press, 1997.

[217] P. B. Nemenyi, "Distribution-free multiple comparisons." *ProQuest Dissertations and Theses*, p. 127, 1963.

[218] H. J. Keselman and J. C. Rogan, "The tukey multiple comparison test : 1953-1976," *Psychological Bulletin*, vol. 84, pp. 1050–1056, 9 1977.

[219] H. M. Krumholz and *et al.*, "Relationship between hospital readmission and mortality rates for patients hospitalized with acute myocardial infarction, heart failure, or pneumonia," *JAMA*, vol. 309, pp. 587–593, 2 2013.

[220] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Processing Magazine*, vol. 13, pp. 47–60, 1996.

[221] A. Radford and *et al.*, "Language models are unsupervised multitask learners," 2019.

[222] N. Sánchez-Maroño, A. Alonso-Betanzos, and M. Tombilla-Sanromán, "Filter methods for feature selection – a comparative study," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4881 LNCS, pp. 178–187, 2007.

[223] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, pp. 273–324, 12 1997.

[224] GuyonIsabelle and ElisseeffAndré, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, 3 2003.

[225] T. N. Lal *et al.*, "Embedded methods," *Studies in Fuzziness and Soft Computing*, vol. 207, pp. 137–165, 2006.

[226] S. Tang and *et al.*, "Democratizing ehr analyses with fiddle : a flexible data-driven preprocessing pipeline for structured clinical data," *The American Medical Informatics Association*, vol. 27, no. 12, pp. 1921–1934, 10 2020.

[227] D. Brossier *et al.*, "Creating a high-frequency electronic database in the PICU : The perpetual patient," *Pediatric Critical Care Medicine*, vol. 19, no. 4, pp. e189–e198, apr 2018. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/29406373/

[228] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, pp. 35–45, 3 1960. [Online]. Available : https://dx.doi.org/10.1115/1.3662552

[229] "Understanding adaptive kalman filters in digital signal processing : Applications & techniques." [Online]. Available : https://www.einfochips.com/blog/understanding-adaptive-kalman-filters-in-digital-signal-processing/

[230] S. G. Mallat, "A theory for multiresolution signal decomposition : The wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 674–693, 1989.

[231] I. Daubechies, "Ten lectures on wavelets," *Ten Lectures on Wavelets*, 1 1992.

[232] D. L. Donoho and J. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 9 1994. [Online]. Available : https://dx.doi.org/10.1093/biomet/81.3.425

[233] D. L. Donoho and L. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of American Statistical Association*, vol. 90, pp. 1200–1224, 1 1995.

[234] U. R. Acharya *et al.*, "Heart rate variability : A review," *Medical and Biological Engineering and Computing*, vol. 44, pp. 1031–1051, 12 2006. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/17111118/

[235] M. Bračič and A. Stefanovska, "Wavelet-based analysis of human blood-flow dynamics," *Bulletin of Mathematical Biology*, vol. 60, pp. 919–935, 1998. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/9739620/

[236] T. Inouye *et al.*, "Quantification of eeg irregularity by use of the entropy of the power spectrum," *Electroencephalography and Clinical Neurophysiology*, vol. 79, pp. 204–210, 9 1991. [Online]. Available : https://www.sciencedirect.com/science/article/abs/pii/001346949190138T

[237] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Transactions on Information Theory*, vol. 38, pp. 713–718, 1992.

[238] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate and sample entropy," *American Journal of Physiology - Heart and Circulatory Physiology*, vol. 278, 2000. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/10843903/

[239] J. R. L. Gall *et al.*, "The logistic organ dysfunction system : A new way to assess organ dysfunction in the intensive care unit," *JAMA*, vol. 276, pp. 802–810, 9 1996. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/8769590/

[240] A. E. Johnson, A. A. Kramer, and G. D. Clifford, "A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy," *Critical Care Medicine*, vol. 41, pp. 1711–1718, 7 2013. [Online]. Available : https://pubmed.ncbi.nlm.nih.gov/23660729/

[241] "Deep Learning vs Machine Learning: The Ultimate Battle."

[242] "LightGBM vs XGBOOST - Which algorithm is better."

[243] M. Barandas *et al.*, "Tsfel : Time series feature extraction library," *SoftwareX*, vol. 11, p. 100456, 2020.

[244] M. Christ *et al.*, "Time series feature extraction on basis of scalable hypothesis tests (tsfresh – a python package)," *Neurocomputing*, vol. 307, pp. 74–83, 2018.

[245] "Hyponatremia - Symptoms and causes - Mayo Clinic."

# APPENDIX A FEATURE AND CLINICAL VARIABLE IMPORTANCE TABLES OF LR MODEL

This appendix provides detailed tables summarizing the importance of features and clinical variables in the LR model for predicting PICU readmission.

Table A.1 lists all extracted features ranked by the absolute value of their LR coefficients, providing a complete description of each feature's relative influence on the model's predictions.

TABLE A.1 – LR features and absolute coefficients

| **Feature** | $\|\beta\|$ |
|---|---|
| SBP_low_ratio | 0.382350 |
| Urine_is_last_low | 0.290492 |
| CCSR_MAL008 | 0.287367 |
| arterial_PO2_is_last_high | 0.239641 |
| Motor_response_first_abnormal_time | 0.235616 |
| CAPD_long_time_respond_Severe_ratio | 0.230649 |
| FLACC_Cry_Moderate_Pleurs_count | 0.226806 |
| RR_wavelet_d1_max | 0.222201 |
| MBP_low_count_Response | 0.219693 |
| Comfort_B_Restlessness_Mild_ratio | 0.215927 |
| admittype_Unité de soin | 0.215681 |
| ING_EXC_has_low_Discharge | 0.215137 |
| Acide_acetylsalicylique_PO_sum | 0.212472 |
| Digoxine_Count | 0.212011 |
| P_mean_high_diff | 0.211842 |
| HR_max | 0.209735 |
| P_std | 0.206553 |
| RR_has_high_Discharge | 0.205466 |
| Mono_time_to_normal | 0.205264 |
| MBP_wp_d_std | 0.201050 |
| arterial_PO2_is_last_abnormal | 0.199305 |
| Troponin_rate_of_change | 0.197536 |
| Respiratory_category_used_Discharge | 0.196208 |

*(continued)*

| **Feature** | $\|\beta\|$ |
|---|---|
| Urine_mean_low_diff | 0.194927 |
| BUN_high_count | 0.192355 |
| DBP_has_abnormal_Discharge | 0.192221 |
| HR_high_ratio | 0.189471 |
| Insuline_sum | 0.189358 |
| num_used_medications_categories | 0.189146 |
| SBP_Discharge_state_Deteriorate | 0.188655 |
| Tot_Bil_max_high_diff | 0.188270 |
| ALT_min | 0.187620 |
| ING_EXC_abnormal_count_Discharge | 0.187197 |
| PLT_num_changes | 0.187039 |
| Rocuronium_Count | 0.186206 |
| SPO2_is_last_low | 0.183847 |
| arterial_PCO2_mean | 0.183392 |
| ING_EXC_abnormal_count | 0.182684 |
| DBP_wp_a_energy | 0.180479 |
| CCSR_END016 | 0.179315 |
| TSH_abnormal_ratio | 0.178614 |
| CCSR_INJ075 | 0.177931 |
| capillary_PCO2_high_ratio | 0.176755 |
| CAPD_Restless_is_last_Moderate | 0.175012 |
| RR_total_power | 0.174211 |
| SBP_wp_a_energy | 0.173199 |
| SBP_wavelet_approx_energy | 0.173199 |
| HR_has_high | 0.172666 |
| Glu_mean | 0.169875 |
| BUN_abnormal_count | 0.166380 |
| arterial_pH_mean_low_diff | 0.165618 |
| Urine_Discharge_state_Abnormal | 0.165204 |
| CCSR_INJ019 | 0.163233 |
| EV_is_last_Moderate_pain | 0.161468 |
| arterial_pH_has_low | 0.161464 |
| Chute_time_to_normal | 0.160008 |
| Glu_low_count | 0.159869 |

*(continued)*

| **Feature** | $\|\beta\|$ |
|---|---|
| Urine_Response_state_Normal | 0.159762 |
| Comfort_B_Facial_tension_is_first_Moderate | 0.158807 |
| MBP_wp_a_max | 0.158638 |
| CCSR_MAL003 | 0.158605 |
| CCSR_NEO048 | 0.158097 |
| RR_is_last_high | 0.157075 |
| ARC_abnormal_count | 0.153745 |
| PLT_has_abnormal | 0.153200 |
| CCSR_GEN002 | 0.152896 |
| CCSR_CIR006 | 0.152867 |
| SPO2_max_diff_Discharge | 0.152066 |
| Prednisolone_sum | 0.151047 |
| venous_PCO2_mean_low_diff | 0.150851 |
| arterial_PO2_high_ratio | 0.149930 |
| MBP_first_abnormal_time | 0.149600 |
| FLACC_Face_Mild_Sourcils_count | 0.148295 |
| FLACC_Consolability_Moderate_consoler_ratio | 0.147870 |
| BUN_max | 0.147275 |
| Comfort_B_Consciousness_is_first_Critical | 0.143835 |
| CCSR_PNL002 | 0.142811 |
| Hgb_num_changes | 0.141006 |
| RR_total_mean | 0.140841 |
| P_time_to_discharge | 0.140828 |
| NaCl_IV_sum | 0.140279 |
| venous_PCO2_rate_of_change | 0.136847 |
| SPO2_psd_freq | 0.136071 |
| Tirage_gen_first_abnormal_time | 0.134226 |
| VentNonInv_ratio | 0.133836 |
| SBP_high_count | 0.133768 |
| FLACC_Cry_Moderate_Plaintes_count | 0.133332 |
| arterial_pH_first_abnormal_time | 0.132996 |
| Ciprofloxacine_PO_mean | 0.132093 |
| Midazolam_Count | 0.131610 |
| DBP_max_high_diff | 0.131182 |

*(continued)*

| **Feature** | $\|\beta\|$ |
|---|---|
| CL_has_low | 0.130721 |
| Tobramycine_SA_mean | 0.129592 |
| FLACC_Consolability_time_to_normal | 0.129042 |
| ARC_max_high_diff | 0.125653 |
| ING_EXC_is_First_low | 0.125633 |
| Albumin_rate_of_change | 0.124733 |
| venous_HCO3_is_First_high | 0.124322 |
| MAP_min | 0.123552 |
| DBP_abnormal_count_Admission | 0.123043 |
| AST_time_to_normal | 0.122251 |
| Hgb_has_abnormal | 0.122239 |
| FLACC_Revised_Leg_time_to_normal | 0.121751 |
| admittype_Unknown | 0.121668 |
| CAPD_long_time_respond_Severe_count | 0.120316 |
| Troponin_std | 0.119662 |
| SBP_mean | 0.119129 |
| GCS_max | 0.119085 |
| MBP_low_count | 0.118860 |
| SBP_is_last_abnormal | 0.118021 |
| FLACC_Face_is_last_Moderate_Mâchoires | 0.116977 |
| CCSR_END012 | 0.115526 |
| Mg_is_First_high | 0.115331 |
| CCSR_NVS014 | 0.113939 |
| CCSR_MUS007 | 0.112960 |
| CCSR_NEO049 | 0.112856 |
| Tobramycine_IV_sum | 0.111183 |
| CCSR_DIG012 | 0.110023 |
| MCV_abnormal_count | 0.109893 |
| CCSR_SYM013 | 0.109679 |
| Cr_is_First_low | 0.109306 |
| Ondansetron_mean | 0.108580 |
| CCSR_RSP017 | 0.108334 |
| Tot_Bil_high_count | 0.107170 |
| FLACC_Revised_Leg_is_last_Moderate_augmentée | 0.105694 |

*(continued)*

| **Feature** | $\|\beta\|$ |
|---|---|
| COMFORT_B_is_first_Weaning_sedation | 0.105472 |
| SBP_is_First_low | 0.104852 |
| venous_pH_is_First_high | 0.104419 |
| Verbal_response_Mild_count | 0.103866 |
| last_weight_median | 0.103428 |
| Cal_Ion_Serum_max | 0.103197 |
| Clobazam_PO_sum | 0.103081 |
| age_category_Infants | 0.102059 |
| SBP_low_count_Response | 0.101645 |
| CCSR_DIG010 | 0.101525 |
| CCSR_NEO058 | 0.101070 |
| Gastrointestinal_category_used_Admission | 0.100386 |
| GCS_has_Low | 0.099927 |
| Acide_tranexamique_IV_Count | 0.099111 |
| CCSR_EXT018 | 0.098407 |
| CCSR_DIG024 | 0.097158 |
| CCSR_EXT015 | 0.097144 |
| CCSR_NEO070 | 0.096261 |
| Inspiratory_press_std | 0.095254 |
| MBP_wavelet_d1_max | 0.095081 |
| Naproxen_sum | 0.094935 |
| CCSR_CIR024 | 0.094914 |
| CCSR_CIR021 | 0.094879 |
| CCSR_MBD010 | 0.093427 |
| FLACC_Revised_Consolability_Mild_distraire_ratio | 0.093191 |
| age_category_Neonates | 0.092610 |
| arterial_SO2_abnormal_ratio | 0.092543 |
| TSH_is_First_high | 0.091436 |
| CCSR_INJ031 | 0.090812 |
| Lympho_max_high_diff | 0.090749 |
| CAPD_Aware_surroundings_Severe_ratio | 0.090229 |
| Glu_high_count | 0.089933 |
| CCSR_NEO064 | 0.089477 |
| NaCl_IV_Count | 0.089295 |

*(continued)*

| **Feature** | $\|\beta\|$ |
|---|---|
| FLACC_Revised_Cry_Moderate_Plaintes_count | 0.089269 |
| CCSR_EXT002 | 0.088998 |
| LNHD_usage_time | 0.088739 |
| Gabapentine_mean | 0.087343 |
| CCSR_FAC009 | 0.087246 |
| CCSR_DIG025 | 0.086852 |
| Caspofungine_IV_mean | 0.086239 |
| Ceftriaxone_mean | 0.085884 |
| SPO2_psd_max | 0.084467 |
| SPO2_has_abnormal_Discharge | 0.084411 |
| CCSR_DIG023 | 0.083970 |
| Verbal_response_Critical_count | 0.083845 |
| Acetylcysteine_PO_mean | 0.083261 |
| CCSR_CIR031 | 0.083029 |
| HR_has_low | 0.082975 |
| SBP_low_count | 0.082866 |
| CCSR_NEO015 | 0.082417 |
| CCSR_SYM001 | 0.082314 |
| CCSR_DIG001 | 0.082101 |
| CCSR_SKN002 | 0.081974 |
| venous_SO2_min | 0.081277 |
| CCSR_SYM010 | 0.080751 |
| CO2_tot_gaz_art_time_to_discharge | 0.079192 |
| CCSR_CIR017 | 0.078426 |
| Tirage_supra_claviculaire_is_last_3 | 0.078308 |
| Hydroxyzine_mean | 0.078265 |
| HR_sampen | 0.077017 |
| ING_EXC_is_last_high | 0.076618 |
| Cardiovascular_category_time_last_use | 0.076588 |
| HR_wp_d_energy | 0.076412 |
| Antihistamine_category_used_Discharge | 0.075009 |
| CCSR_MUS025 | 0.074739 |
| CCSR_DIG002 | 0.073082 |
| CCSR_CIR005 | 0.072476 |

*(continued)*

| **Feature** | $\|\beta\|$ |
|---|---|
| CCSR_FAC025 | 0.071988 |
| CCSR_MAL009 | 0.071376 |
| CCSR_INJ030 | 0.070913 |
| MBP_high_count_Admission | 0.070448 |
| Cal_Ion_Serum_has_high | 0.069739 |
| venous_PO2_abnormal_ratio | 0.069455 |
| capillary_Lactate_max_low_diff | 0.066388 |
| CCSR_DIG009 | 0.066301 |
| CCSR_MUS011 | 0.065683 |
| CCSR_MUS004 | 0.065139 |
| ING_EXC_Discharge_state_Deteriorate | 0.063348 |
| CCSR_END013 | 0.062982 |
| CCSR_MUS023 | 0.062788 |
| CCSR_FAC006 | 0.062574 |
| Nifedipine_sum | 0.062543 |
| bld_prd_count | 0.061257 |
| CCSR_END015 | 0.061207 |
| Tirage_sous_costal_first_abnormal_time | 0.060474 |
| CCSR_INJ013 | 0.060035 |
| CCSR_DIG008 | 0.059904 |
| Comfort_B_Ventilation_is_last_Mild | 0.059721 |
| CCSR_INJ014 | 0.059121 |
| SBP_is_First_high | 0.058870 |
| ALP_first_abnormal_time | 0.057484 |
| RBC_rate_of_change | 0.057382 |
| DBP_low_count_Discharge | 0.057243 |
| Comfort_B_Facial_tension_Moderate_count | 0.056674 |
| Isuprel_Count | 0.055908 |
| GCS_time_to_normal | 0.055285 |
| CCSR_CIR013 | 0.053299 |
| Tirage_supra_sternal_2_count | 0.052960 |
| Vancomycine_IV_Count | 0.052390 |
| arterial_Lactate_high_count | 0.051695 |
| DBP_wavelet_d1_transients | 0.051551 |

*(continued)*

| **Feature** | $\|\beta\|$ |
|---|---|
| HR_has_high_Discharge | 0.051013 |
| SBP_mean_low_diff | 0.049662 |
| ING_EXC_low_ratio | 0.049482 |
| HR_is_First_low | 0.049356 |
| Cal_max_high_diff | 0.049231 |
| Hydromorphone_mean | 0.048895 |
| MBP_spectral_entropy | 0.048486 |
| CCSR_NVS006 | 0.048138 |
| MBP_has_high_Admission | 0.047729 |
| CCSR_RSP010 | 0.047348 |
| CCSR_CIR011 | 0.047078 |
| admittype_SOP_chirnoncar_NP | 0.046521 |
| CCSR_CIR012 | 0.046344 |
| SBP_max_diff_Response | 0.045836 |
| Comfort_B_Consciousness_Moderate_count | 0.045404 |
| CCSR_MAL005 | 0.045224 |
| ARC_time_to_normal | 0.045071 |
| Amoxicilline_Acide_sum | 0.043794 |
| CCSR_MBD001 | 0.043301 |
| CCSR_NEO022 | 0.043019 |
| SBP_is_First_abnormal | 0.041204 |
| CCSR_INJ039 | 0.041200 |
| SBP_wp_d_mean | 0.040562 |
| SBP_wavelet_d1_mean | 0.040562 |
| MBP_wp_a_mean | 0.040412 |
| MBP_wavelet_approx_mean | 0.040412 |
| capillary_PCO2_max_low_diff | 0.036531 |
| COMFORT_B_has_Weaning_sedation | 0.036111 |
| Chute_rate_of_change | 0.035732 |
| Methylprednisolone_Count | 0.034594 |
| Motor_response_is_last_Severe | 0.033353 |
| SBP_iqr | 0.032774 |
| SPO2_wavelet_approx_entropy | 0.030386 |
| HR_wp_a_entropy | 0.029393 |

*(continued)*

| **Feature** | $\|\beta\|$ |
|---|---|
| ING_EXC_first_abnormal_time | 0.028376 |
| Clonidine_PO_sum | 0.026386 |
| ING_EXC_has_abnormal | 0.025010 |
| DBP_total_power | 0.023492 |
| RR_has_high_Response | 0.020908 |
| SPO2_abnormal_count | 0.019332 |
| SPO2_low_count | 0.019332 |
| Urine_first_abnormal_time | 0.018694 |
| RR_wp_a_entropy | 0.018383 |
| Urine_time_to_discharge | 0.018020 |
| admittype_Urgence | 0.017988 |
| Urine_has_abnormal | 0.017725 |
| age | 0.017312 |
| Fentanyl_Timbre_sum | 0.013553 |
| DBP_Response_state_Improved | 0.008781 |

Table A.2 aggregates features into their corresponding clinical variables and reports the cumulative absolute effect of each variable, along with the list of features contributing to it. This view highlights the overall role of higher-level clinical variables in shaping the model's decision-making.

TABLE A.2 – Clinical variables, their extracted features, and cumulative absolute LR effect

| **Clinical variable** | **Extracted Features** | **Effect** (sum of $\|\beta\|$) |
|---|---|---|
| Diagnosis | CCSR_* | 5.8048 |
| SBP | SBP_low_ratio, SBP_Discharge_state_Deteriorate, SBP_wp_a_energy, SBP_wavelet_approx_energy, SBP_high_count, SBP_mean, SBP_is_last_abnormal, SBP_is_First_low, SBP_low_count_Response, SBP_low_count, SBP_is_First_high, SBP_mean_low_diff, SBP_max_diff_Response, SBP_is_First_abnormal, SBP_wp_d_mean, SBP_wavelet_d1_mean, SBP_iqr | 1.8872 |
| FLACC | FLACC_Cry_Moderate_Pleurs_count, FLACC_Face_Mild_Sourcils_count, FLACC_Consolability_Moderate_consoler_ratio, FLACC_Cry_Moderate_Plaintes_count, FLACC_Consolability_time_to_normal, FLACC_Revised_Leg_time_to_normal, FLACC_Face_is_last_Moderate_Mâchoires, FLACC_Revised_Leg_is_last_Moderate_augmentée, FLACC_Revised_Consolability_Mild_distraire_ratio, FLACC_Revised_Cry_Moderate_Plaintes_count | 1.3122 |
| MBP | MBP_low_count_Response, MBP_wp_d_std, MBP_wp_a_max, MBP_first_abnormal_time, MBP_low_count, MBP_wavelet_d1_max, MBP_high_count_Admission, MBP_spectral_entropy, MBP_has_high_Admission, MBP_wp_a_mean, MBP_wavelet_approx_mean | 1.1904 |

*(continued)*

| **Clinical variable** | **Extracted Features** | **Effect** (sum of $\|\beta\|$) |
|---|---|---|
| ING_EXC | ING_EXC_has_low_Discharge, ING_EXC_abnormal_count_Discharge, ING_EXC_abnormal_count, ING_EXC_is_First_low, ING_EXC_is_last_high, ING_EXC_Discharge_state_Deteriorate, ING_EXC_low_ratio, ING_EXC_first_abnormal_time, ING_EXC_has_abnormal | 0.9535 |
| RR | RR_wavelet_d1_max, RR_has_high_Discharge, RR_total_power, RR_is_last_high, RR_total_mean, RR_has_high_Response, RR_wp_a_entropy | 0.9391 |
| HR | HR_max, HR_high_ratio, HR_has_high, HR_has_low, HR_sampen, HR_wp_d_energy, HR_has_high_Discharge, HR_is_First_low, HR_wp_a_entropy | 0.9380 |
| Urine | Urine_is_last_low, Urine_mean_low_diff, Urine_Discharge_state_Abnormal, Urine_Response_state_Normal, Urine_first_abnormal_time, Urine_time_to_discharge, Urine_has_abnormal | 0.8648 |
| DBP | DBP_has_abnormal_Discharge, DBP_wp_a_energy, DBP_max_high_diff, DBP_abnormal_count_Admission, DBP_low_count_Discharge, DBP_wavelet_d1_transients, DBP_total_power, DBP_Response_state_Improved | 0.7680 |
| SPO2 | SPO2_is_last_low, SPO2_max_diff_Discharge, SPO2_psd_freq, SPO2_psd_max, SPO2_has_abnormal_Discharge, SPO2_wavelet_approx_entropy, SPO2_abnormal_count, SPO2_low_count | 0.7099 |

*(continued)*

| **Clinical variable** | **Extracted Features** | **Effect** (sum of $\|\beta\|$) |
|---|---|---|
| Comfort_B | Comfort_B_Restlessness_Mild_ratio, Comfort_B_Facial_tension_is_first_Moderate, Comfort_B_Consciousness_is_first_Critical, Comfort_B_Ventilation_is_last_Mild, Comfort_B_Facial_tension_Moderate_count, Comfort_B_Consciousness_Moderate_count | 0.6804 |
| CAPD | CAPD_long_time_respond_Severe_ratio, CAPD_Restless_is_last_Moderate, CAPD_long_time_respond_Severe_count, CAPD_Aware_surroundings_Severe_ratio | 0.6162 |
| arterial_PO2 | arterial_PO2_is_last_high, arterial_PO2_is_last_abnormal, arterial_PO2_high_ratio | 0.5889 |
| P | P_mean_high_diff, P_std, P_time_to_discharge | 0.5592 |
| BUN | BUN_high_count, BUN_abnormal_count, BUN_max | 0.5060 |
| arterial_pH | arterial_pH_mean_low_diff, arterial_pH_has_low, arterial_pH_first_abnormal_time | 0.4601 |
| Glu | Glu_mean, Glu_low_count, Glu_high_count | 0.4197 |
| admittype | admittype_Unité de soin, admittype_Unknown, admittype_SOP_chirnoncar_NP, admittype_Urgence | 0.4019 |
| PLT | PLT_num_changes, PLT_has_abnormal | 0.3402 |
| Tirage | Tirage_gen_first_abnormal_time, Tirage_supra_claviculaire_is_last_3, Tirage_sous_costal_first_abnormal_time, Tirage_supra_sternal_2_count | 0.3260 |
| ARC | ARC_abnormal_count, ARC_max_high_diff, ARC_time_to_normal | 0.3245 |
| Troponin | Troponin_rate_of_change, Troponin_std | 0.3172 |
| Tot_Bil | Tot_Bil_max_high_diff, Tot_Bil_high_count | 0.2954 |
| venous_PCO2 | venous_PCO2_mean_low_diff, venous_PCO2_rate_of_change | 0.2877 |

*(continued)*

| **Clinical variable** | **Extracted Features** | **Effect** (sum of $\|\beta\|$) |
|---|---|---|
| GCS | GCS_max, GCS_has_Low, GCS_time_to_normal | 0.2743 |
| TSH | TSH_abnormal_ratio, TSH_is_First_high | 0.2701 |
| Motor_response | Motor_response_first_abnormal_time, Motor_response_is_last_Severe | 0.2690 |
| Hgb | Hgb_num_changes, Hgb_has_abnormal | 0.2632 |
| Tobramycine | Tobramycine_SA_mean, Tobramycine_IV_sum | 0.2408 |
| NaCl | NaCl_IV_sum, NaCl_IV_Count | 0.2296 |
| capillary_PCO2 | capillary_PCO2_high_ratio, capillary_PCO2_max_low_diff | 0.2133 |
| Acide_ acetylsalicylique | Acide_acetylsalicylique_PO_sum | 0.2125 |
| Digoxine | Digoxine_Count | 0.2120 |
| age | age_category_Infants, age_category_Neonates, age | 0.2120 |
| Mono | Mono_time_to_normal | 0.2053 |
| Respiratory_ category | Respiratory_category_used_Discharge | 0.1962 |
| Chute | Chute_time_to_normal, Chute_rate_of_change | 0.1957 |
| Insuline | Insuline_sum | 0.1894 |
| medications_ categories | num_used_medications_categories | 0.1891 |
| Verbal_response | Verbal_response_Mild_count, Verbal_response_Critical_count | 0.1877 |
| ALT | ALT_min | 0.1876 |
| Rocuronium | Rocuronium_Count | 0.1862 |
| arterial_PCO2 | arterial_PCO2_mean | 0.1834 |
| Cal_Ion_Serum | Cal_Ion_Serum_max, Cal_Ion_Serum_has_high | 0.1729 |
| EV | EV_is_last_Moderate_pain | 0.1615 |
| Prednisolone | Prednisolone_sum | 0.1510 |
| VentNonInv | VentNonInv_ratio | 0.1338 |
| Ciprofloxacine | Ciprofloxacine_PO_mean | 0.1321 |

*(continued)*

| **Clinical variable** | **Extracted Features** | **Effect** (sum of $\|\beta\|$) |
|---|---|---|
| Midazolam | Midazolam_Count | 0.1316 |
| CL | CL_has_low | 0.1307 |
| Albumin | Albumin_rate_of_change | 0.1247 |
| venous_HCO3 | venous_HCO3_is_First_high | 0.1243 |
| MAP | MAP_min | 0.1236 |
| AST | AST_time_to_normal | 0.1223 |
| Mg | Mg_is_First_high | 0.1153 |
| MCV | MCV_abnormal_count | 0.1099 |
| Cr | Cr_is_First_low | 0.1093 |
| Ondansetron | Ondansetron_mean | 0.1086 |
| venous_pH | venous_pH_is_First_high | 0.1044 |
| last_weight | last_weight_median | 0.1034 |
| Clobazam | Clobazam_PO_sum | 0.1031 |
| Gastrointestinal_ category | Gastrointestinal_category_used_Admission | 0.1004 |
| Acide_ tranexamique | Acide_tranexamique_IV_Count | 0.0991 |
| Inspiratory_press | Inspiratory_press_std | 0.0953 |
| Naproxen | Naproxen_sum | 0.0949 |
| arterial_SO2 | arterial_SO2_abnormal_ratio | 0.0925 |
| Lympho | Lympho_max_high_diff | 0.0907 |
| LNHD | LNHD_usage_time | 0.0887 |
| Gabapentine | Gabapentine_mean | 0.0873 |
| Caspofungine | Caspofungine_IV_mean | 0.0862 |
| Ceftriaxone | Ceftriaxone_mean | 0.0859 |
| Acetylcysteine | Acetylcysteine_PO_mean | 0.0833 |
| venous_SO2 | venous_SO2_min | 0.0813 |
| CO2_tot_gaz_art | CO2_tot_gaz_art_time_to_discharge | 0.0792 |
| Hydroxyzine | Hydroxyzine_mean | 0.0783 |

*(continued)*

| **Clinical variable** | **Extracted Features** | **Effect** (sum of $\|\beta\|$) |
|---|---|---|
| Cardiovascular_ category | Cardiovascular_category_time_last_use | 0.0766 |
| Antihistamine_ category | Antihistamine_category_used_Discharge | 0.0750 |
| venous_PO2 | venous_PO2_abnormal_ratio | 0.0695 |
| capillary_Lactate | capillary_Lactate_max_low_diff | 0.0664 |
| Nifedipine | Nifedipine_sum | 0.0625 |
| bld_prd | bld_prd_count | 0.0613 |
| ALP | ALP_first_abnormal_time | 0.0575 |
| RBC | RBC_rate_of_change | 0.0574 |
| Isuprel | Isuprel_Count | 0.0559 |
| Vancomycine | Vancomycine_IV_Count | 0.0524 |
| arterial_Lactate | arterial_Lactate_high_count | 0.0517 |
| Cal | Cal_max_high_diff | 0.0492 |
| Hydromorphone | Hydromorphone_mean | 0.0489 |
| Amoxicilline_Acide | Amoxicilline_Acide_sum | 0.0438 |
| Methylprednisolone | Methylprednisolone_Count | 0.0346 |
| Clonidine | Clonidine_PO_sum | 0.0264 |
| Fentanyl_Timbre | Fentanyl_Timbre_sum | 0.0136 |