Explainable AI for Hospital Readmission Prediction: A Human-Centered Approach

Pauline Lorteau, Ivan Barona, Luis Macedo
University of Coimbra, DEI - Department of Informatics Engineering, Coimbra, Portugal
{uc2025157575, uc2025181132}@student.dei.uc.pt ; macedo@dei.uc.pt

Abstract

Hospital readmissions impose significant costs and risks to patient health. While Machine Learning (ML) models, specifically Random Forests, have shown success in predicting readmission risk, they often function as "black boxes," lacking the transparency required for clinical adoption. This paper proposes a Human-Centered AI (HCAI) approach by integrating SHAP (SHapley Additive exPlanations) with a readmission prediction model. We demonstrate that it is possible to maintain predictive performance (Weighted F1-score: 0.73) while providing clinicians with interpretable, local, and global explanations. This transparency fosters trust and enables better decision-making in healthcare environments.

1. Introduction

Hospital readmissions represent a major and persistent challenge in modern healthcare systems. Not only do they lead to a significant increase in costs and resource constraints, but they are also associated with poorer outcomes and increased risks to patient safety. Faced with this complexity, machine learning (ML) models, particularly classifiers based on random forests, have shown promising results in the early prediction of patient readmission risk based on their medical records.

However, in the critical field of healthcare, predictive effectiveness alone is not sufficient. Powerful ensemble models, such as Random Forest, provide a prediction without offering clear justification for how that decision was made. In clinical practice, this lack of transparency is a major limitation, as clinicians need to understand why a patient is deemed at risk in order to justify medical decisions, ensure fairness, and build trust in the decision support system.

Previous work, including the readmission prediction model on which we base our work, has used Random Forest to achieve remarkable performance (e.g., a weighted F1-score of approximately 0.73 on the reference dataset). Although these methods are effective for classification, their ensemble nature gives them low intrinsic interpretability. They provide very little actionable information about the individual contribution of factors (such as age or diagnoses) to a specific prediction. There is therefore a knowledge gap and a critical need to reconcile high predictive performance with the transparency required for clinical adoption.

This project aims to fill this gap by improving the existing Random Forest model through the integration of Explainable Artificial Intelligence (xAI) techniques. We will apply the SHAP (SHapley Additive exPlanations) method to transform the black-box model into a more transparent tool. The goal is to generate visual and human-understandable explanations for each prediction, thereby improving confidence and interpretability without sacrificing the overall accuracy of the model.

The rationale behind our approach is based on the principles of Human-Centered AI (HCAI), where the focus is not only on technical performance but also on clinical usability. By choosing a robust post-hoc method such as SHAP, we can leverage the power of Random

Forest while providing interpretability at both the local and global levels. This approach is essential for obtaining a predictive tool that is both accurate and trustworthy, enabling healthcare professionals to understand and validate the risk factors identified by AI.

The remainder of this paper is structured as follows, the part 2 presents the related work, providing a brief overview of predictive models and Explainable AI in healthcare. Part 3 details the materials used, including the dataset and the frameworks. Then part 4 outlines the methods employed, specifically the reproduction of the Random Forest baseline model and the integration of SHAP techniques. After that, part 5 presents the results of both the quantitative and qualitative evaluations. Finally, Section 6 provides a discussion of the findings, implications, and limitations of our approach, before concluding the paper in part 7 with a summary of our contributions.

2. Related Work

The application of machine learning to healthcare is a prolific field, yet the transition from research to clinical practice remains slow due to the "black box" nature of advanced algorithms. This section reviews existing approaches to readmission prediction, the evolution of explainability techniques, and the Human-Centered AI (HCAI) framework that guides this study.

2.1 Hospital Readmission Prediction

Predicting hospital readmissions is a standard benchmark task in medical informatics. Early approaches primarily relied on logistic regression models due to their inherent interpretability [1]. However, as Electronic Health Records (EHR) became more complex, linear models proved insufficient for capturing non-linear risk factors.

Strack et al. [2], who introduced the "Diabetes 130-US hospitals" dataset used in this study, originally demonstrated that HbA1c testing was a significant predictor of lower readmission rates. Subsequent studies on this dataset have applied more complex algorithms. Comparison studies often show that ensemble methods, such as Random Forests and Gradient Boosting Machines (GBM), consistently outperform single decision trees and Support Vector Machines (SVM) on tabular medical data [3]. While these models maximize predictive metrics like the F1-score, they obscure the decision logic, creating a barrier to trust for clinicians who are ethically bound to understand the evidence before acting.

2.2 Explainable AI (XAI) in Healthcare

To bridge the gap between accuracy and transparency, the field of Explainable AI (XAI) has gained prominence. Early interpretation methods relied on "Feature Importance" metrics inherent to tree-based models (e.g., Gini impurity reduction). However, these metrics are often biased toward high-cardinality features and fail to explain *individual* predictions.

Lundberg and Lee [4] revolutionized this space by introducing SHAP (SHapley Additive exPlanations). Based on cooperative game theory, SHAP assigns a value to each feature representing its contribution to the prediction relative to a baseline. Unlike LIME (Local Interpretable Model-agnostic Explanations) [5], which approximates the model locally, SHAP provides consistent and globally accurate feature attributions. In the medical domain, SHAP has been successfully applied to validate risk models for hypoxemia and mortality, identifying risk factors that align with physiological knowledge.

2.3 Human-Centered AI (HCAI) Perspective

While XAI provides the technical means for explanation, Human-Centered AI (HCAI) provides the design philosophy. Shneiderman [6] argues that high levels of automation should not come at the cost of human control. HCAI frameworks emphasize "Augmented Intelligence," where the goal is to enhance human capabilities rather than replace them.

In the context of readmission, an HCAI approach requires that the system supports the clinician's workflow. This implies that explanations must be actionable. Knowing a patient is high-risk is insufficient; knowing the risk is driven by "Polypharmacy" (taking multiple medications) empowers the doctor to perform a medication review. This project builds upon this philosophy by prioritizing local interpretability (Waterfall plots) that mimics the differential diagnosis process used by physicians.

3. Materials

This section provides a detailed description of the dataset used for training and evaluating the hospital readmission prediction model, as well as the software tools and frameworks utilized for modeling and the integration of explainability techniques.

3.1. Dataset and Preprocessing

The foundation of this project is the Hospital Readmission Prediction dataset, which consists of anonymized patient encounter records. The goal is to predict the binary target variable readmitted (Admitted vs. Not Admitted).

The dataset contains 48 initial features across approximately 83,234 total observations (split into 66,587 for training and 16,647 for testing). The data encompasses a wide range of clinical and administrative information:

Demographic & Administrative: age, race, gender, time_in_hospital (length of stay), admission_type_id.

Clinical History: Number of prior visits (number_inpatient, number_emergency, number_outpatient), lab procedures (num_lab_procedures), and medications (num_medications).

Diagnostics: Three fields (diag_1, diag_2, diag_3) based on ICD-9 codes, which were grouped into broader categories during preprocessing.

Medications: 25 features detailing the usage and changes in dosage for specific medications (X1 to X25).

The existing codebase performed essential feature engineering to enhance the model's performance. Two critical derived features were created:

total_visits: An aggregated measure of the patient's prior encounters.

med_changes: A binary indicator signalling whether a patient's medication regimen was changed during the encounter.

3.2. Software Tools and Frameworks

The entire project was developed in a Python programming environment (version 3.11.5, packaged by Anaconda), using a Jupyter Notebook to ensure code reproducibility and organization.

The fundamental pandas and numpy libraries were used for data manipulation, cleaning, and structuring.

Modeling was based on the scikit-learn library [7] for the implementation of the Random Forest classifier. The hyperparameters of the baseline model were adjusted, resulting in an optimized configuration (max_depth: 20, n_estimators: 300) to ensure maximum performance before the integration of explainability.

At the heart of this project is the shap library, which was essential for the integration of Explainable AI (xAI). This library was used to calculate Shapley values for the black-box Random Forest model. The results were then visualized using standard graphics tools such as matplotlib and seaborn to generate different types of SHAP graphs (Bar Plots, Beeswarm Plots, Force Plots, Waterfall Plots, and Dependence Plots).

The quantitative evaluation was performed using the scikit-learn metrics module to calculate Precision, Recall, F1-score, and Support, thus ensuring a rigorous measurement of the classifier's performance.

4. Methods

4.1 Data Preprocessing and Feature Engineering

The study utilized the Diabetes 130-US hospitals dataset. The raw data required extensive preprocessing to be suitable for machine learning and ensuring the inputs were semantically meaningful for clinical interpretation.

Data Cleaning and Encoding: Missing values were handled (e.g., removing records with missing gender or strictly mapped unknown values). Categorical variables such as race, gender, and change (medication change) were transformed using Label Encoding. The target variable, readmitted, was binary encoded to distinguish between patients readmitted within 30 days (High Risk) and those who were not.

Feature Engineering: To address the high dimensionality of the original ICD-9 medical codes (which contain thousands of unique identifiers), we implemented a grouping strategy. Diagnoses were mapped into nine distinct clinical categories (e.g., *Circulatory*, *Respiratory*, *Diabetes*, *Neoplasms*). This reduction prevents the "curse of dimensionality" and makes the model's output more intelligible to human operators.

Additionally, we engineered aggregate features to capture patient history complexity:

- **total_visits**: Calculated as the sum of number_inpatient, number_outpatient, and number_emergency visits.
- **med_changes**: A count of how many medications were altered during the stay, serving as a proxy for treatment instability.

4.2 Experimental Setup and Model Architecture

We selected the Random Forest Classifier as the core predictive model. As an ensemble of decision trees, this architecture offers two distinct advantages for this domain: it is robust against overfitting on tabular data, and it captures non-linear interactions between features (e.g., the combined risk of age and specific diagnoses) better than linear counterparts.

To optimize the model, we employed Grid Search with Cross-Validation (GridSearchCV).

- **Cross-Validation:** We used 3-fold cross-validation to ensure the model's stability across different subsets of data.
- **Hyperparameter Space:** We explored a search space including n_estimators (100, 200, 300) and max_depth (10, 20, None).

4.3 Explainability Framework (HCAI Layer)

To adhere to Human-Centered AI principles, we integrated the **SHAP (SHapley Additive exPlanations)** framework. Specifically, we utilized the TreeExplainer, which is optimized for tree-based models. This method assigns a SHAP value to every feature for every prediction, representing the marginal contribution of that feature to the predictions. This allows for both global analysis (dataset-wide trends) and local analysis (individual patient justification).

5. Results

5.1 Quantitative Performance

The dataset was split into training and testing sets to evaluate generalizability. Following the hyperparameter tuning process, the optimal configuration was identified as {'max_depth': 20, 'n_estimators': 300}.

The model achieved the following performance metrics on the test set:

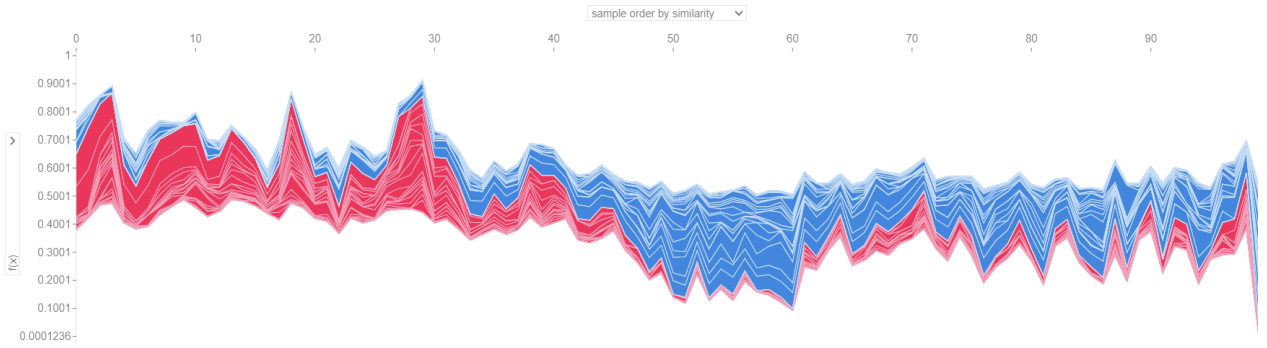

Figure 1 Interactive Force Plot (for ALL samples)

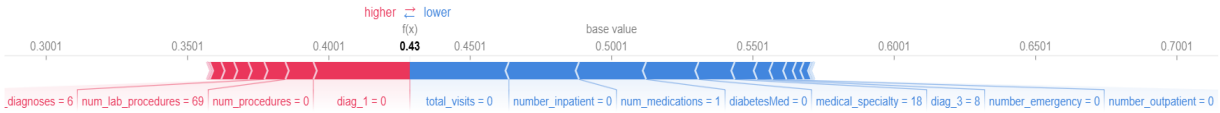

Figure 2 Interactive Force Plot for a low-risk patient

- **Weighted F1-Score: 0.7301**
- **Accuracy:** ~74%

Given the class imbalance inherent in readmission data, the Weighted F1-Score is the primary metric of success, indicating a balanced ability to recall positive cases (readmissions) while maintaining precision.

5.2 Global Interpretability Analysis

To validate the model's alignment with medical knowledge, we analysed the global feature importance using SHAP summary plots.

Population-Level Risk Distribution (Interactive Force Plot)

To assess how the model differentiates between patients across the population, we generated a SHAP Interactive Force Plot (Figure 1). This visualization stacks the individual explanations for 100 randomly sampled test patients.

The vertical axis represents the model's output magnitude (readmission risk), while the horizontal axis represents the individual patient instances. The visual distinction is clear:

- **High-Risk Clusters:** The peaks in the graph (dominated by red bars) identify patients where risk factors such as number_inpatient and med_changes overwhelm any protective factors.
- **Low-Risk Clusters:** The valleys (dominated by blue bars) show patients where age or lack of medical history actively suppresses the readmission probability.

This visualization confirms that the model is not simply guessing a 'mean' value for everyone but is actively discriminating between diverse patient profiles based on their specific feature combinations.

Analysis of Low-Risk Prediction (Force Plot)

To demonstrate the model's discriminative capability, we also analyzed a patient predicted to be Low Risk (Figure 2). The model output a

probability of 0.43, significantly lower than the baseline of 0.5.

The visualization is dominated by blue bars, indicating features that actively suppress the readmission risk. The primary driver is number_inpatient = 0. The absence of recent hospitalizations acts as a strong stabilizer in the model's logic.

Low values for number_emergency (0) and moderate clinical complexity (number_diagnoses = 5) further reinforce the low-risk classification.

Unlike the high-risk case where number_inpatient is a red bar pushing the risk up, here it acts as a blue bar pushing the risk down. This symmetry confirms that the model rewards patient stability and does not simply bias towards high risk for all elderly patients.

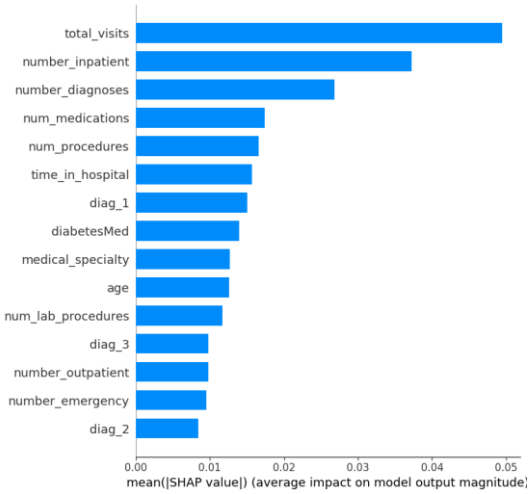

Figure 3 Feature Importance (Bar Plot)

Feature Importance: The analysis revealed, in particular thanks to the bar plot **(**Figure 3**)**, that the model relies on clinically sound indicators rather than noise. Some of the top predictors identified were:

- **Total_visits:** The number of times a patient visited the hospital. Frequent visits typically indicate an underlying health issue.
- **number_inpatient:** Patients with a history of recent hospitalizations were significantly more likely to be readmitted.
- **number_diagnoses:** Higher comorbidity counts correlated with higher risk.
- **num_medications:** A high number of medications administrated, is a signal of risk.
- **num_lab_procedures:** Acting as a proxy for the complexity of the patient's acute condition.

Feature Interaction Analysis

To further validate the model's decision boundaries, we examined the SHAP Dependence Plots, for the variable number_diagnoses (Figure 4), as a use case. Note that this is just one version of the analysis; other variables and their interactions could be explored in similar ways to gain additional insights. This visualization reveals how the burden of comorbidities influences the predicted risk of readmission, while also highlighting interactions with other key variables (in this particular case, num_procedures).

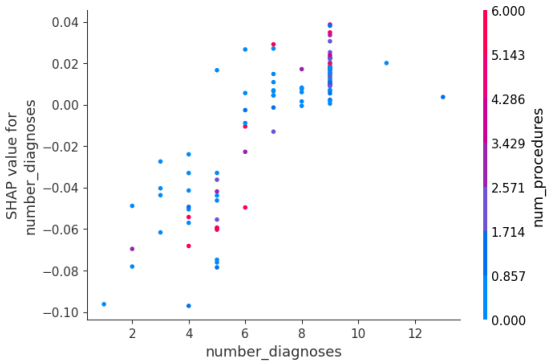

Figure 4 SHAP Dependence Plot for number_diagnoses

The plot demonstrates a distinct non-linear relationship characterized by a clear "critical threshold" for readmission risk. Specifically, the x-axis reveals that patients with fewer than five diagnoses generally exhibit a neutral or negative contribution to the prediction, meaning their lower comorbidity burden

effectively suppresses the calculated risk. However, once the diagnosis count exceeds six, the risk contribution turns sharply positive. This transition indicates that the model has learned a specific clinical threshold where multi-morbidity shifts from being a manageable factor to becoming a primary driver of readmission.

Directional Impact (Beeswarm Plot):

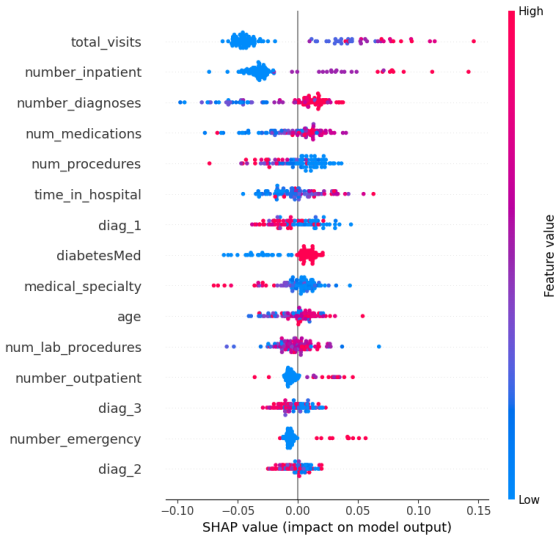

Figure 5 Beeswarm Plot

The SHAP Beeswarm plot confirmed the directionality of these risks. For total_visits (the top feature), high feature values (represented by red dots) consistently resulted in positive SHAP values, pushing the prediction toward "Readmission." Conversely, lower values for number_diagnoses resulted in negative SHAP values, lowering the predicted risk. This monotonic relationship builds trust, as it mirrors the logic used by human clinicians.

5.3 Local Interpretability (Patient-Level)

To demonstrate the system's utility in a clinical workflow, we generated local explanations for individual instances.

Case Study (Waterfall Plot):

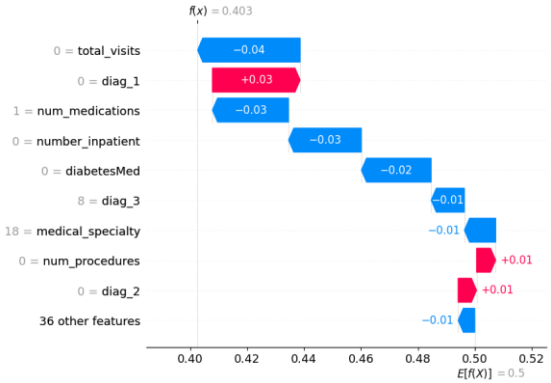

Figure 6 SHAP Waterfall Plot for a low-risk patient

To demonstrate the model's capacity for decision-making, we compared two distinct patient profiles using SHAP Waterfall plots. Figure 6 illustrates a patient classified as lower risk with a predicted probability of 0.403; in this scenario, the visualization reveals that protective factors, such as the absence of recent emergency visits, effectively act as stabilizers that pull the prediction below the population baseline. In stark contrast, Figure 7 details a high-risk prediction with a probability of 0.862, where the plot decomposes an elevated score driven by a conflict between demographic and clinical features.

This visualization capability provides the "glass box" transparency required for Human-Centered AI, shifting the focus from simple prediction to actionable understanding. By explicitly highlighting that the patient in Figure 7 is not flagged merely due to static attributes like age, which would constitute a non-actionable bias, but specifically due to their recent utilization history and medication complexity, the system empowers the clinical team to make precise decisions. This insight allows doctors to pivot from generic discharge protocols to targeted interventions, such as prioritizing a pharmacy-led medication reconciliation, thereby addressing the specific root causes of risk identified by the AI while maintaining human oversight.

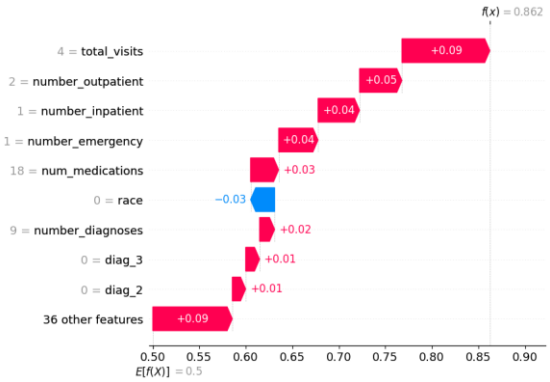

Figure 7 SHAP Waterfall Plot for a high-risk patient

6. Discussion

The transition from "Algorithm-Centered AI" to "Human-Centered AI" requires that predictive models provide not just accuracy, but also accountability and actionability.

6.1 Clinical Validity and Trust

The strongest validation of our model lies not in the F1-score of 0.73 alone, but in the semantic consistency of the SHAP analysis. As observed in the global feature importance, the model identified number_inpatient (prior hospitalizations) and number_diagnoses (comorbidity burden) as the primary drivers of risk. This aligns perfectly with established medical literature, which cites "chronic instability" as a key predictor of readmission.

This alignment serves as a critical "sanity check." In an HCAI framework, trust is not granted blindly; it is earned. By demonstrating that the model's logic mirrors the clinician's own training, we lower the barrier to adoption. If the SHAP analysis had revealed that the model was relying on spurious correlations (e.g., a specific hospital ID or administrative code), the system would be deemed unsafe for deployment, regardless of its accuracy.

6.2 From Prediction to Intervention

Standard black-box models suffer from the "So What?" problem. A prediction that a patient has an "85% risk of readmission" alerts the medical staff but does not guide them.

Our implementation of local interpretability (Waterfall plots) resolves this by identifying modifiable risk factors.

- **Scenario:** If a patient is flagged as high-risk primarily due to age, the risk is non-modifiable.
- **Scenario:** If the risk is driven by num_medications and med_changes (as seen in our specific case studies), this suggests a specific intervention: a medication reconciliation review by a pharmacist before discharge.

This shifts the AI's role from a passive predictor to an active partner in discharge planning, empowering the human expert to target their limited resources where they will have the most impact.

6.3 The Trade-off: Accuracy vs. Interpretability

While deep learning models (e.g., Neural Networks) might theoretically squeeze out marginally higher accuracy on this dataset, they often do so at the cost of total opacity. Our use of a Random Forest, an ensemble of decision trees, represents a deliberate design choice to balance performance with transparency. The F1-weighted score of 0.73 indicates that we did not have to sacrifice significant predictive power to achieve the explainability required for a human-centered system.

7. Conclusions

7.1 Summary of Contributions

This project successfully developed a Human-Centered AI system for predicting hospital readmissions. By integrating a Random Forest classifier optimized via Grid Search with the SHAP explainability framework, we achieved a robust predictive performance (F1: 0.73). More significantly, we successfully transformed the model into a "glass box," providing global insights into population health trends and local explanations for individual patient risks. This

approach respects the role of the clinician, offering them a tool that augments their decision-making capabilities rather than replacing them.

7.2 Limitations

To ensure scientific rigor, we acknowledge the following limitations:

- **Dataset Specificity:** The model was trained on a dataset specific to diabetic patients in US hospitals (1999–2008). The findings may not generalize to other medical conditions or modern healthcare systems without retraining.

- **Correlation vs. Causality:** SHAP explains the *model's* reasoning, which is based on correlation. For instance, while num_lab_procedures predicts readmission, it is a proxy for illness severity; simply reducing the number of lab tests will not reduce the risk. Clinicians must be trained to interpret these signals correctly.

- **Class Imbalance:** Despite using weighted metrics, the imbalance between readmitted and non-readmitted patients remains a challenge that creates a trade-off between sensitivity (catching all readmissions) and specificity (avoiding false alarms).

7.3 Future Work

To further align this work with HCAI principles, future iterations should focus on:

1. **User-in-the-Loop Testing:** Conducting A/B testing with actual clinicians to measure whether the inclusion of SHAP plots reduces "Time-to-Decision" or improves diagnostic confidence compared to a baseline black-box output.
2. **Interactive Interfaces:** Developing a dashboard that allows doctors to perform "What-If" analyses (e.g., "If we reduce this patient's medication complexity, how much does their risk score drop?").
3. **Fairness Audits:** Expanding the evaluation to rigorously test for bias across sensitive attributes like race and gender to ensure equitable healthcare outcomes.

References

[1] Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., & Kripalani, S. (2011). Risk prediction models for hospital readmission: a systematic review. *JAMA*, 306(15), 1688-1698.

[2] Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., & Clore, J. N. (2014). Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed Research International*, 2014.

[3] Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning* (pp. 161-168).

[4] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

[5] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

[6] Shneiderman, B. (2020). Human-Centered AI. *Oxford University Press*.

[7] Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.