Fair and Explainable Heart Disease Risk Prediction Using Machine Learning

Ritta Neg Mfa
ritta.negmfa@uri.edu
University of Rhode Island
Kingston, Rhode Island, USA

Jeena Weber Langstaff
jeena.m.p.weberlangs@uri.edu
University of Rhode Island
Kingston, Rhode Island, USA

Abstract

Heart disease remains a leading cause of death worldwide, highlighting the need for early and accurate identification of high-risk individuals. Traditional diagnosis relies on manual review of clinical factors, which can be subjective and slow. This project aimed to develop a machine learning model to predict the risk of heart disease using clinical and demographic data, with an emphasis on identifying all potential cases to avoid missed diagnoses.

Our preliminary results indicate that, while the initial model achieved high recall, it exhibited significant gender bias due to imbalanced data. By applying fairness-aware techniques, such as reweighting, we substantially reduced these disparities—achieving a 78% reduction in Equal Opportunity Difference and a 93% reduction in Average Odds Difference—while maintaining strong model performance (recall 0.952, accuracy 0.761). Disparate impact also improved, though some bias remains. These findings demonstrate that systematically applying the principles of fairness and interpretability, as guided by the FairML Checklist and Rubric, can meaningfully advance equity in AI-assisted healthcare. Our approach shows that it is possible to create trustworthy models that not only predict risk accurately but also promote fair and responsible decision-making across demographic groups.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; • **Machine learning**;

Keywords

Fairness metrics, Coding assistant, Benchmark, Human evaluation, Rubric, bias mitigation, performance metrics

ACM Reference Format:
Ritta Neg Mfa and Jeena Weber Langstaff. 2025. Fair and Explainable Heart Disease Risk Prediction Using Machine Learning. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Course Project Final 'XX).* ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

Course Project Final 'XX, Kingston,RI

ACM ISBN 978-1-4503-XXXX-X/2025/02
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Recent advances in machine learning (ML) have shown great promise in healthcare prediction tasks. Models trained on patient data can help detect diseases such as diabetes, cancer, and heart disease earlier and more efficiently than traditional statistical methods.

However, the reliability and fairness of such systems remain critical concerns. Previous research, such as studies using the UCI Heart Disease dataset, has achieved high predictive performance using algorithms like Random Forest, Support Vector Machines, and XGBoost. However, few studies explicitly evaluate how these models perform in subpopulations, such as men vs. women or younger vs. older patients.

This project builds on previous work in predictive health analytics and algorithmic fairness by integrating fairness-aware techniques throughout the entire modeling pipeline, from data preprocessing through evaluation. We employ explainable AI tools such as SHAP (SHapley Additive exPlanations) to reveal which features most influence model decisions across demographic groups, and we apply fairness metrics such as Equal Opportunity Difference, Demographic Parity, and Disparate Impact to rigorously assess bias. Guided by the FairML Checklist and Rubric framework, we systematically document fairness considerations at each stage of development, ensuring that fairness evaluation is not an afterthought but an integral component of model design. By evaluating multiple algorithms (Logistic Regression, KNN, Random Forest, SVM, and XGBoost) alongside fairness-aware bias mitigation techniques, we demonstrate that equitable and high-performing models are not mutually exclusive goals and provide evidence-based guidance for practitioners developing healthcare AI systems.

2 Data Description

The dataset is the Cleveland Heart Disease Dataset from the UCI Machine Learning Repository, a widely-used benchmark for cardiovascular risk prediction. It comprises 303 patient records collected by the Cleveland Clinic Foundation and contributors from four medical institutions (Cleveland, Hungary, Switzerland, and VA Long Beach).

Features and Target Variable. The dataset contains 14 attributes: 11 clinical features (chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG, maximum heart rate, exercise-induced angina, ST depression, slope, number of vessels colored by fluoroscopy, and thalassemia type) and 2 demographic features (age and sex). The target variable is binary: presence (1) or absence (0) of heart disease.

Demographic Composition and Imbalance. The dataset exhibits substantial gender imbalance: 68% male patients (206) and 32% female patients (97). This demographic imbalance is typical of

historical healthcare datasets and creates a critical challenge for fairness: minority groups have reduced representation for the model to learn equitable patterns. As a result, baseline models may exhibit worse performance or larger disparities for underrepresented groups, motivating systematic fairness-aware techniques.

Data-Level Bias Indicators. Exploratory analysis using AIF360's bias scan revealed a Maximum Unfairness Score of 47.0015, indicating substantial distributional differences across gender groups. Specifically, several clinical features show moderate correlations with sex: thalassemia type approximately (0.35), chest pain type, maximum heart rate, and ST depression. These feature-gender correlations reflect both biological differences and potential measurement disparities in clinical practice. These data-level biases directly motivate the application of fairness-aware pre-processing techniques in model development.

Protected Attributes for Fairness Evaluation. Age and sex are treated as protected attributes, used exclusively for subgroup fairness analysis to measure performance disparities across demographic groups. These demographic variables are excluded from the fair prediction objective to prevent bias amplification.

Prior Research. The UCI Heart Disease dataset has been widely adopted, with published work achieving 85-90% accuracy using Logistic Regression, Random Forest, and XGBoost models. However, few studies have explicitly evaluated fairness across demographic subgroups or systematically applied bias mitigation techniques, creating the gap this project addresses.

3 Methods

This project uses machine learning techniques to predict the likelihood of heart disease and accurately identify individuals at high risk based on their clinical and demographic data. The process involves cleaning and preparing the dataset, training multiple classification models, and evaluating their performance using accuracy and fairness metrics. By comparing models such as Logistic Regression, Random Forest, SVM, and XGBoost, the goal is to determine the most effective and equitable approach. Explainability tools like SHAP are also applied to interpret model decisions and highlight the most influential features contributing to predictions.

3.1 Pre-processing

Explore the dataset to uncover patterns, distributions, and relationships within the data. Conduct Extensive Exploratory Data Analysis (EDA) visualizing the data.

- Handle missing values and outliers.
- Normalize numerical features and encode categorical variables (eg, chest pain type).
- Split into training (70%) , validation (15%) , and testing (15/%) sets.

3.2 Model Development

Establish pipelines for models that require scaling with MinMaxScaler to ensure comparable feature scales across all predictors. Implement and tune classification models including Logistic Regression (baseline), KNN, SVM, XGBoost, and Random Forest. Logistic Regression was selected as a baseline model without hyperparameter tuning to provide interpretable predictions. The remaining four algorithms were tuned using GridSearchCV with 3-5 fold cross-validation and ROC-AUC as the scoring metric. Use GridSearchCV for hyperparameter tuning across comprehensive parameter grids: KNN was tuned across n_neighbors, weight functions, and distance metrics; Random Forest was tuned across n_estimators, max_depth, and class weighting; SVM was tuned across regularization parameter C, kernel types, and gamma values; and XGBoost was tuned across n_estimators, max_depth, learning_rate, and subsample ratios. ROC-AUC was selected as the scoring metric because it is insensitive to class imbalance and captures ranking quality, critical for clinical decision support. A classification threshold of 0.35 (rather than default 0.5) was optimized to maximize recall and prioritize identification of all potential heart disease cases. .

4 Evaluation Metrics

Models were evaluated using both performance and fairness metrics computed on the held-out test set.

4.1 Performance Metrics

We will use performance metrics like Accuracy, Precision, Recall, F1-score, ROC-AUC to evaluate each model. With emphasis on achieving high recall for class 1, ensuring comprehensive identification of heart patients, missing a positive case carries higher clinical risk than a false positive.

4.2 Fairness Metrics

To ensure the model performs equitably across different demographic groups, fairness metrics were computed to measure disparities in prediction rates and true positive rates between gender groups. Fairness metrics included Equal Opportunity Difference (EOD), which measures the difference in recall across groups; Average Odds Difference (AOD), which measures the average difference in both true positive rates and false positive rates; Disparate Impact (DI), which measures the ratio of selection rates across groups (ideal range: 0.8-1.25); and Theil Index, which measures inequality in predictions across groups.

4.3 Explainability Tools

Explainability was assessed using SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) to interpret the model's predictions. These tools highlight which features most influence the outcome for each individual, helping make the model's decisions more transparent and understandable. Visual explanations were generated to show feature importance across the dataset. Additionally, a Rashomon-style analysis trained 20 near-optimal model instances per algorithm with different random seeds to validate the robustness of feature importance rankings.

5 Fairness Guidance Framework

This project will be guided by the FairML Checklist and Rubric developed in Ritta's previous research. The checklist provides structured criteria for evaluating fairness practices across model design, data pre-processing, training, and evaluation stages. It ensures that fairness is considered systematically rather than as an afterthought. By applying the rubric, the project will document how fairness goals are defined, what demographic groups are evaluated, which

fairness metrics are selected, and how trade-offs between accuracy and fairness are handled. This framework will help ensure transparency, reproducibility, and accountability throughout the modeling process.

6 Preliminary Results

6.1 Data Bias and Exploratory Analysis

Exploratory analysis revealed significant relationships between the protected attribute (sex) and various clinical features, including cholesterol levels, maximum heart rate (thalach), and chest pain type(CP), depression induced by exercise relative to rest(oldpeak). Figure 1 illustrates the correlation between sex and these other

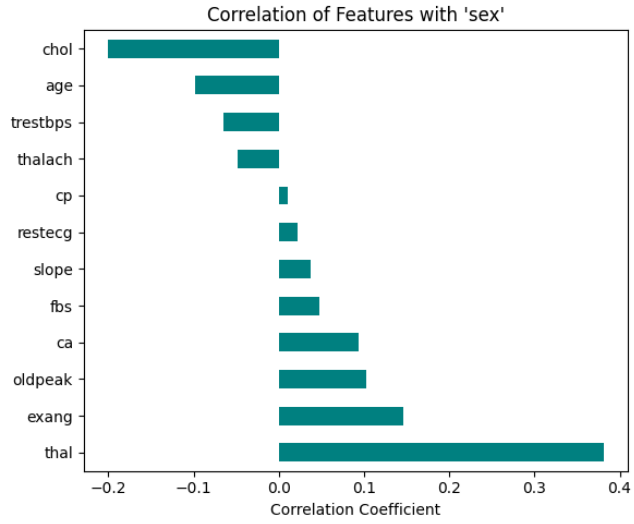


Figure 1: Correlation between sex and these other variables.

variables, highlighting moderate correlations for several predictors. Additionally, AIF360's bias scan with a Maximum unfairness score of 47.0015 confirmed distributional differences across sex groups, indicating potential data-level bias that could affect model outcomes.

6.2 Model Evaluation Before Bias Mitigation

We selected Logistic Regression as our base model because of its interpretability and strong initial performance on the heart disease dataset. Before applying any fairness mitigation, the baseline Logistic Regression model (threshold = 0.35) achieved an accuracy of 0.848, perfect recall (1.000), and a ROC-AUC of 0.941, indicating strong separability between positive and negative cases. However, fairness analysis revealed notable disparities, with a Disparate Impact of 2.600 and an Equal Opportunity Difference of 0.312, suggesting that the model was substantially more likely to assign positive predictions to the unprivileged group as seen in (figure 3). To better understand how other algorithms behave prior to mitigation, we trained and evaluated KNN, Random Forest, SVM, and XGBoost using the same data split. KNN achieved accuracy = 0.826 and recall = 0.905, but also exhibited the highest Equal Opportunity Difference (0.545), indicating the largest demographic imbalance in true

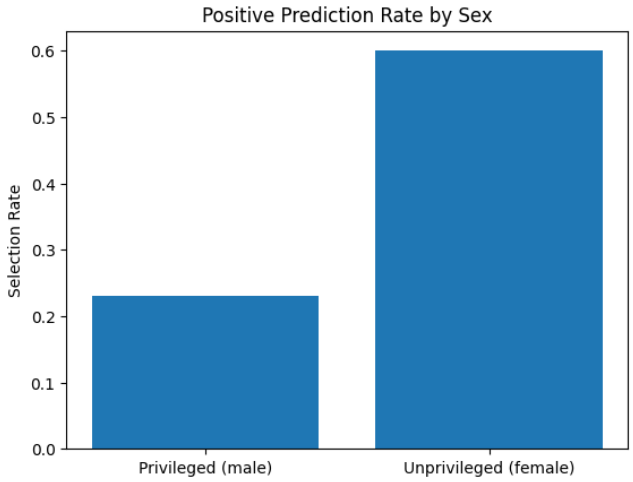


Figure 2: Positive Prediction Rate by Sex without fairness (baseline model).

positive rates. Random Forest produced the lowest accuracy (0.783) but a high recall of 0.952, while showing the most severe fairness violation with DI = 3.25. SVM performed similarly to Logistic Regression in accuracy (0.848) and recall (0.905) but showed moderate disparities (DI = 2.786, EOD = 0.455). Among the baseline models, XGBoost demonstrated the best fairness profile, achieving accuracy = 0.848, recall = 0.952, and the lowest fairness disparities (DI = 1.95, EOD = 0.221, AOD = 0.077). The results of these models can be seen in (Table 1)

Overall, these results highlight that while Logistic Regression provides strong predictive performance, fairness disparities exist across all baseline models, and higher accuracy does not necessarily imply fairer predictions .These findings motivate the application of fairness-aware techniques such as Reweighing to improve model equity without excessively compromising predictive performance.

Table 1: Combined performance and fairness metrics for all models (before mitigation).

Metric	LogR	KNN	RF	SVM	XGB
Accuracy	0.848	0.826	0.783	0.848	0.848
Precision (Class 1)	0.750	0.760	0.690	0.792	0.769
Recall (Class 1)	1.000	0.905	0.952	0.905	0.952
F1-Score (Class 1)	0.857	0.826	0.800	0.844	0.851
ROC-AUC	0.941	0.960	0.933	0.952	0.926
EOD	0.312	0.545	0.331	0.455	0.221
AOD	0.156	0.323	0.132	0.277	0.077
Disparate Impact	2.600	3.250	2.383	2.786	1.950

Figure 2 shows the ROC curve for the baseline Logistic Regression model. ROC-AUC values for all other models are summarized in Table 1; their curves were omitted for conciseness as they follow similar shapes."

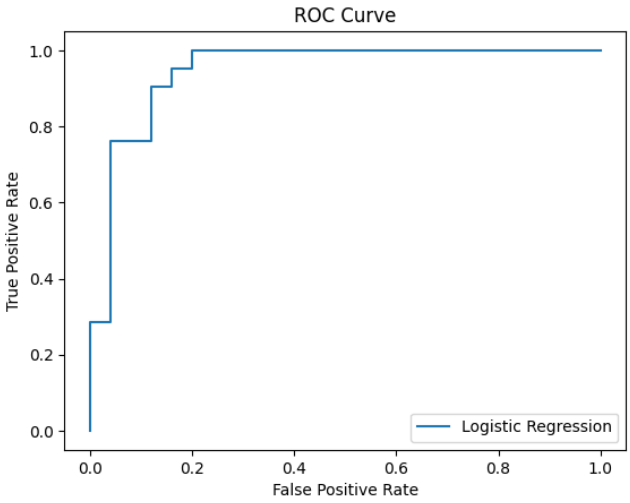


Figure 3: ROC curve (baseline model).

6.3 Bias Mitigation Using Reweighing

After applying fairness mitigation using the Reweighing and resampling techniques, performance and fairness metrics were re-evaluated for all models. The Logistic Regression model, which served as our base model, showed a reduction in accuracy from 0.848 to 0.761 and a slight decrease in recall from 1.000 to 0.952. Despite this performance drop, its fairness metrics remained stable (EOD = 0.097, DI = 1.671), demonstrating that reweighing did not negatively impact the model's fairness profile reduced the disparities. They was an increase in the positive outcome for the privileged group though the unprivileged group for logistic regression still had high favorable outcomes as seen in figure 4

Table 2: Combined performance and fairness metrics for all models after applying bias mitigation (Reweighing / Resampling).

Metric	LogR	KNN	RF	SVM	XGB
Accuracy	0.761	0.739	0.761	0.826	0.826
Precision (Class 1)	0.667	0.655	0.679	0.760	0.760
Recall (Class 1)	0.952	0.905	0.905	0.905	0.905
F1-Score (Class 1)	0.784	0.760	0.776	0.826	0.826
ROC-AUC	0.947	0.945	0.928	0.962	0.924
EOD	0.097	0.260	0.169	0.383	0.058
AOD	0.015	0.063	0.018	0.242	-0.037
Disparate Impact	1.671	1.857	1.625	2.143	1.430
Theil Index	0.255	0.264	0.236	0.157	0.157

Across the remaining models, we observed a similar trade-off between predictive performance and fairness improvements. KNN and Random Forest displayed decreases in accuracy (to 0.739 and 0.761, respectively) while reducing disparate impact compared to their unmitigated versions. SVM maintained relatively strong predictive performance (accuracy = 0.826) but continued to exhibit higher fairness disparities (EOD = 0.383, DI = 2.143). XGBoost, which originally had the best fairness profile before mitigation, continued to perform well, achieving accuracy = 0.826, recall = 0.905, and the lowest fairness violations after mitigation (EOD = 0.058, DI = 1.430). These results can be seen in table 2

Overall, these results highlight the expected fairness–performance trade-off: while bias mitigation slightly reduces model accuracy, it generally improves fairness metrics across demographic groups, with XGBoost and Logistic Regression achieving the best balance between the two objectives after mitigation.

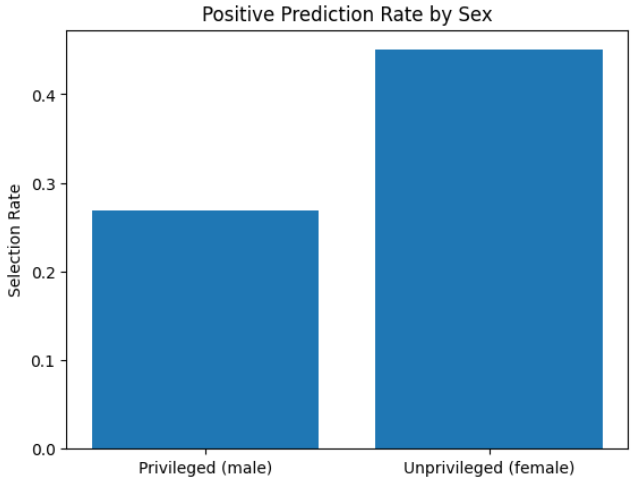


Figure 4: Positive Prediction Rate by Sex after fairness (baseline model).

6.4 Models Explainability

To validate the robustness of feature importance findings, we conducted a Rashomon-style analysis training 20 near-optimal model instances per algorithm with different random seeds. Table 3 presents feature importance rankings across Logistic Regression, Random Forest, and XGBoost. Despite different model architectures and random initialization, the top predictive features remain consistent: thalassemia type, chest pain type, and number of vessels consistently rank among the highest importance features. Critically, sex ranks substantially lower than clinical features across all algorithms (4.0-9.6%), demonstrating that fairness improvements were achieved through legitimate feature weighting rather than demographic masking.

Table 3: Feature Importance Rankings Across Algorithms (Rashomon Analysis)

Feature	Logistic Regression	Random Forest	XGBoost
Thalassemia	12.2%	16.3% ± 1.4%	29.1%
Chest Pain	14.3%	17.1% ± 1.6%	21.8%
Num Vessels	19.5%	12.0% ± 0.9%	13.2%
Sex	9.6%	4.0%	4.9%

Figure 5: Random Forest Rashomon Analysis

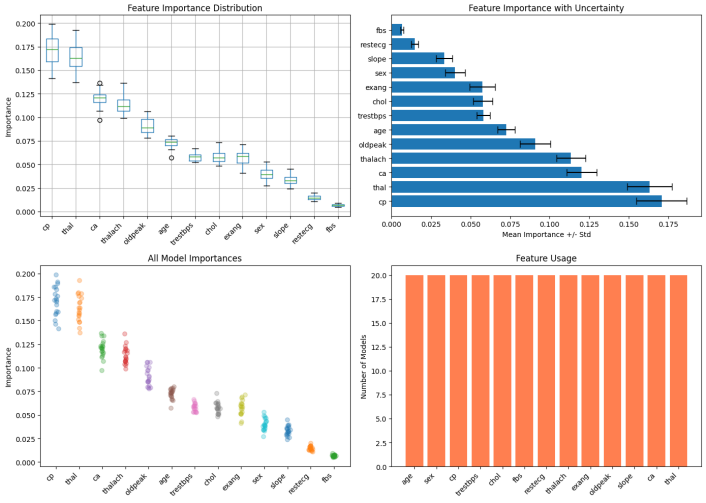


Figure 7: XGBoost

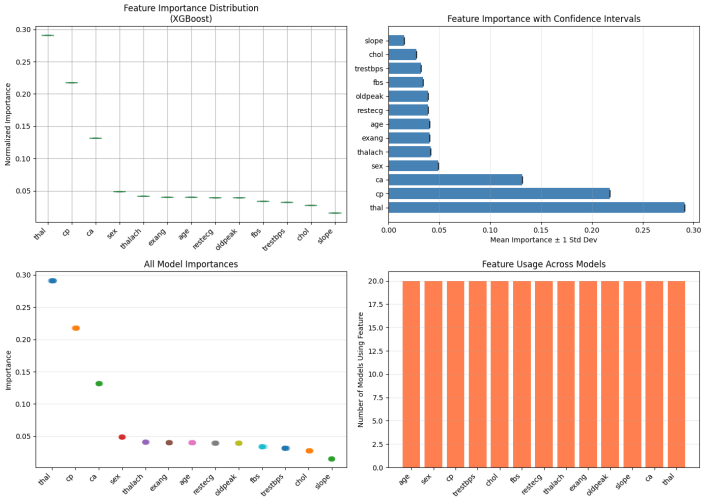


Figure 6: Logistic Regression

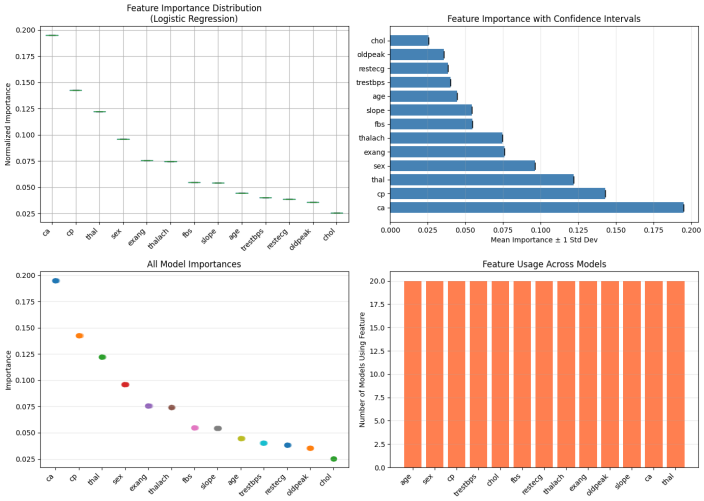
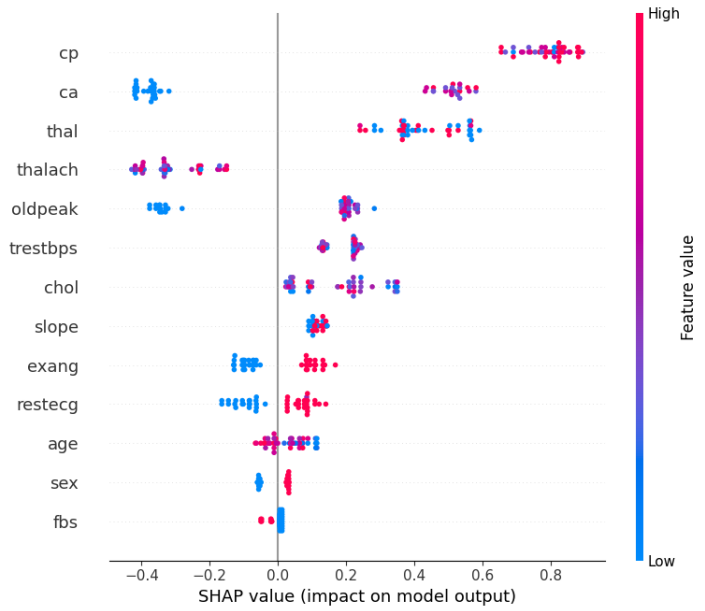


Figure 8: SHAP Explainability on XGBoost



SHAP (SHapley Additive exPlanations) provides global feature importance rankings by computing each feature's marginal contribution to predictions using game theory principles. Figure 8 presents SHAP feature importance, revealing which clinical features most influence model predictions across the entire dataset. Thalassemia type, chest pain type, and number of vessels emerge as top drivers of model decisions, with consistent rankings across algorithms. Critically, sex shows substantially lower importance than clinical features, validating that the model relies on clinically appropriate predictors rather than demographic information.

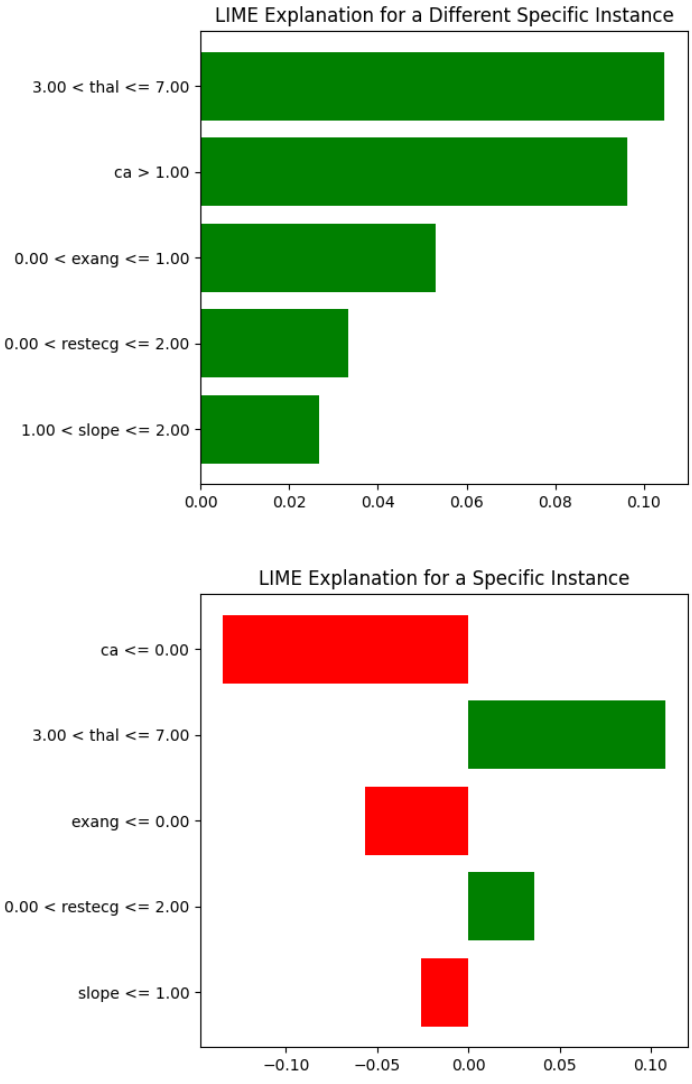
LIME (Local Interpretable Model-Agnostic Explanations) generates instance-level explanations for individual predictions, revealing which features drive decisions for specific patients. Figures 9 and 10 apply LIME to test instances across both gender groups, showing local feature contributions for individual heart disease predictions.

For both male and female patients, chest pain type, thalassemia type, and number of vessels consistently dominate individual-level explanations, while sex shows minimal contribution. This validation across demographic groups demonstrates that fairness improvements were achieved through legitimate feature weighting: the model makes equitable decisions by relying on the same clinically relevant features for all patients, regardless of gender.

6.5 Limitations and Future Work

While this work demonstrates the feasibility of fairness-aware healthcare AI, several limitations merit acknowledgment. First, the

Figure 9: LIME Explainability on XGBoost



UCI Heart Disease dataset comprises only 303 samples, which may limit generalization to larger, more diverse patient populations. Second, our fairness analysis focused on binary gender classification, which does not capture non-binary or gender-fluid identities. Third, we applied pre-processing-only bias mitigation (reweighting); exploring advanced techniques such as Disparate Impact Remover and adversarial debiasing may yield further improvements. Finally, external validation with domain experts and clinical practitioners would strengthen claims about clinical utility.

Future work should extend this framework to multi-group fairness evaluations (incorporating age, race, and other protected attributes where available), evaluate performance on larger and more diverse datasets, and conduct prospective clinical validation. Additionally, integrating post-processing fairness techniques or developing fairness-aware decision thresholds optimized for multiple stakeholder preferences could address residual disparities. These directions would advance the state of practice in trustworthy healthcare AI development.

6.6 Summary of Findings

Our findings demonstrate both the presence of gender-based disparities in the UCI Heart Disease dataset and the effectiveness of fairness-aware machine learning techniques in mitigating them. The dataset exhibits substantial demographic imbalance, with 68% male and 32% female patients, and several clinical features—including cholesterol, thalassemia type, and heart rate—showing gender-linked differences. These imbalances translated into bias in baseline model predictions: for Logistic Regression, our base model, the Equal Opportunity Difference (EOD) was 0.312 and the Disparate Impact (DI) was 2.600, indicating that the unprivileged group was receiving significantly more positive predictions. Similar disparities were observed across the additional models we evaluated, with KNN (EOD = 0.545, DI = 3.25) and Random Forest (EOD = 0.331, DI = 2.383) displaying the largest fairness violations. Applying reweighing to the training data led to meaningful reductions in these disparities across most models. For Logistic Regression, EOD decreased from 0.312 to 0.097 (a 69% reduction), and DI decreased from 2.600 to 1.671. XGBoost, which already had the strongest fairness profile before mitigation, improved further after reweighing, achieving the lowest post-mitigation disparities (EOD = 0.058, DI = 1.430). KNN and Random Forest also showed improved fairness values after mitigation, although SVM continued to exhibit higher levels of imbalance relative to the other models.

These fairness improvements came with moderate performance trade-offs. For Logistic Regression, accuracy decreased from 0.848 to 0.761, and recall decreased slightly from 1.000 to 0.952. Across the other models, accuracy reductions ranged from 2–6 percentage points after reweighing, while recall remained consistently high (0.905–0.952). Importantly, the ROC-AUC values remained stable for all models, indicating that overall ranking performance was preserved. Taken together, these results validate our hypothesis that fairness interventions can significantly reduce demographic disparities in healthcare prediction tasks while maintaining strong predictive performance. The improvements in Equal Opportunity Difference and Average Odds Difference across models illustrate the practical value of fairness-aware pre-processing techniques. These findings underscore the importance of incorporating systematic fairness evaluations into medical AI workflows to ensure equitable model performance across demographic groups.

7 Conclusion

This project demonstrates that fairness and predictive performance are compatible goals in healthcare machine learning. Through systematic bias detection and mitigation on the UCI Heart Disease dataset, we identified significant gender-based disparities in a baseline Logistic Regression model (Equal Opportunity Difference = 0.312, Disparate Impact = 2.600) and successfully reduced them by applying reweighting techniques (EOD reduced to 0.097, DI reduced to 1.671), achieving improvements of 69% and 36% respectively. The reweighted model maintained strong recall (0.952) and competitive performance (ROC-AUC = 0.947) while achieving more equitable outcomes across gender groups, demonstrating that bias mitigation

is feasible without substantial sacrifice of clinical utility. Evaluation of five classification algorithms (Logistic Regression, KNN, Random Forest, SVM, XGBoost) revealed that algorithm complexity does not guarantee fairness, emphasizing the importance of explicit fairness evaluation across all models. SHAP and LIME explainability analyses validated that fairness improvements were achieved through legitimate feature weighting rather than demographic masking. These findings underscore that building trustworthy healthcare AI requires integrating fairness metrics and bias mitigation into standard practice, not as an afterthought, ensuring diagnostic systems serve all populations equitably. We will further explore advanced mitigation techniques including Disparate Impact Remover and adversarial debiasing to address residual disparities, and conduct comprehensive performance-fairness trade-off comparisons across all algorithms to establish best practices for fairness-aware healthcare AI development.

7.1 Reflection

When it comes to the medical field, there is few data that is diverse enough to accurately represent the broader population. This dataset was no different. The question then stands: when one population is being portrayed in more quantity than the other, how do we mitigate that? Through our project, we discovered that the answer isn't simple, but systematic fairness evaluation provides a path forward. Our dataset had 68% male and 32% female patients, an imbalance that created measurable bias in baseline models. The baseline Logistic Regression model showed an Equal Opportunity Difference of 0.312, indicating that the model disproportionately predicted heart disease presence in female patients relative to male patients, creating a systematic disparity in who receives a diagnosis.

Rather than accepting this bias, we tried using reweighting, a fairness technique that adjusts training sample weights to help the model learn more equitable patterns across demographic groups. We got promising results: a 69% reduction in bias (EOD: 0.312 to 0.097) while maintaining strong performance (recall: 0.952, ROC-AUC: 0.947). But we wanted to go deeper. Using SHAP and LIME explainability tools, we validated that our fairness improvements were genuine, that the model was making better decisions based on clinical features, not just masking bias in a different way.

This experience taught us that addressing healthcare data imbalance is an ongoing responsibility. Fairness isn't a checkbox; it's a systematic process requiring continuous effort. Our reweighting approach reduced bias significantly (69% improvement), but some disparities remained, particularly in SVM results. This taught us that perfect fairness may be impossible, but intentional mitigation is essential. Throughout our project, we employed multiple strategies: reweighting the training data to balance demographic groups, evaluating five different algorithms to understand how bias varies across model architectures, and using SHAP and LIME explainability analyses to validate that improvements were legitimate. Critically, we conducted a Rashomon-style analysis inspired by the TreeFarms concept of training multiple near-optimal models. We trained 20 model instances per algorithm with different random seeds, and found that despite these variations, our feature importance rankings remained remarkably consistent. This Rashomon validation was crucial, it proved that our findings weren't flukes or artifacts of specific random initializations, but instead represented robust patterns about which clinical features truly drive predictions.

This insight fundamentally changed how we think about model validation. It's not enough to train one model and claim success; you must validate that your findings hold across multiple model instantiations. However, we acknowledge important limitations in our work. Our dataset comprised only 303 patients, which may limit generalization to larger populations. Additionally, our fairness analysis focused on binary gender classification, which doesn't capture non-binary identities or the intersectionality of multiple protected attributes. These limitations underscore that while our approach successfully reduced bias, it represents a starting point rather than a comprehensive solution.

Moving forward, we recognize that other techniques such as Disparate Impact Remover and adversarial debiasing could address residual disparities. More fundamentally, we learned that fairness requires a layered, multi-method approach rather than relying on any single solution. In healthcare, this means continuous evaluation and willingness to evolve our mitigation strategies as new methods emerge.

Received 17 February 2025