Counterfactual Fairness under Gender Imbalance: An Empirical Study on Health Decision Models

Inês Amorim
Department of Computer Science, Faculty of Sciences, University of Porto, Portugal
(Dated: December 31, 2025)

Machine learning models are increasingly used to support clinical decision-making, yet concerns remain about their fairness across demographic groups, particularly in the presence of data imbalance. In healthcare datasets, gender imbalance and uneven disease prevalence can lead to systematic disparities in both predictive performance and model explanations. This work investigates counterfactual fairness under gender imbalance using the Cleveland Heart Disease dataset as a case study. The results reveal a consistent structural asymmetry: predictions for male patients are primarily driven by physiological features, whereas predictions for female patients frequently rely on changes to the sensitive attribute itself. While SMOTE improves group-level fairness metrics, it increases sensitivity to gender at the individual level, particularly near decision boundaries. These findings highlight the limitations of aggregate fairness metrics alone and underscore the role of counterfactual analysis in disentangling algorithmic bias from clinically grounded, sex-specific risk patterns.

Growing evidence indicates that models trained on observational medical data can inadvertently reproduce or amplify existing disparities across demographic groups, raising concerns about fairness, transparency, and trustworthiness in clinical settings. Fairness in machine learning has therefore become an active area of research, with formal criteria to quantify disparities in model behaviour across protected groups. While these metrics provide valuable aggregate-level insights, they can obscure important heterogeneity in how decisions are made for individual patients. This limitation is particularly pronounced in healthcare, where datasets often exhibit severe class imbalance and uneven subgroup representation. In such settings, models may achieve strong overall performance while relying on spurious correlations or sensitive attributes to resolve ambiguous cases. Counterfactual explanations offer a powerful framework for addressing this gap. By identifying the minimal changes required to alter a model's prediction for an individual, counterfactuals provide instance-level insight into decision logic and sensitivity to specific attributes. When applied to fairness analysis, counterfactuals can reveal whether protected attributes, such as sex, act as meaningful contextual modifiers or as unjustified shortcuts.

This work investigates counterfactual fairness under gender imbalance in health decision models through an empirical study of the Cleveland Heart Disease dataset. A Random Forest classifier is evaluated under multiple training configurations, including standard splitting, joint stratification by gender and outcome, reweighting, and SMOTE-based oversampling. Fairness is assessed using both group-level metrics and counterfactual explanations generated with DiCE.

I. THEORETICAL FRAMEWORK

Fairness in machine learning has been extensively studied, with formal definitions such as demographic parity, equalized odds, and equal opportunity providing complementary notions of group-level fairness[1]. In healthcare, prior work has shown that models trained on biased or imbalanced data can exacerbate existing disparities, leading to systematically different outcomes across demographic groups[2]. Counterfactual explanations were formalized by Wachter et al. (2017)[3] as a mechanism for providing actionable and interpretable explanations for automated decisions. Building on this framework, DiCE[4] introduced an efficient approach for generating diverse counterfactuals and has since been widely adopted in fairness analyses to assess model sensitivity to protected attributes and decision boundaries. Bias mitigation strategies are often applied at the preprocessing stage, including reweighting methods that adjust the contribution of training samples to reduce group-level disparities[5]. However, recent studies caution that such techniques may behave unpredictably under severe class imbalance, particularly in medical settings with small or noisy subgroups. Similarly, oversampling methods such as SMOTE are commonly used to address imbalance by generating synthetic minority-class examples, yet emerging evidence suggests that these synthetic instances may not faithfully represent the true data distribution, potentially increasing overfitting and misleading downstream analyses. This concern has been highlighted in recent medical machine learning literature[6], which emphasizes that synthetic samples can distort clinically meaningful feature relationships and compromise the reliability of interpretability methods.

Motivated by these findings, the present work adopts a hybrid approach. Although the predictive model is trained using SMOTE to mitigate class imbalance, counterfactual explanations are generated exclusively from real samples in the original training data. This design choice ensures that counterfactuals remain grounded in empirically observed and clinically plausible feature distributions.

II. METHODOLOGY

The Cleveland Heart Disease Dataset[7] from the UCI Machine Learning Repository is used in this study. The dataset contains clinical and demographic attributes commonly employed in the diagnosis of heart disease. It exhibits substantial imbalance both across gender and across the joint distribution of gender and disease status. Men with heart disease constitute the largest subgroup, while women with heart disease are markedly under-represented (Figure 1b). This imbalance motivates a careful evaluation of both predictive performance and fairness. Table I summarizes the 14 selected attributes used in the analysis, comprising 13 predictors and one target variable. The original target variable ranges from 0 to 4 (Figure 2), where 0 indicates the absence of heart disease and values from 1 to 4 correspond to increasing levels of disease severity. However, consistent with common practice in the literature (e.g., Reddy et al., 2021[8]), the task is reformulated as a binary classification problem by grouping all non-zero values (1–4) into a single "disease present" category and retaining 0 as "no disease." This transformation is widely adopted due to the relatively small size of the dataset and the sparse representation of higher severity classes, which makes multi-class modelling unstable and highly susceptible to imbalance. Accordingly, the target variable *num* is converted into a binary label, denoted as *HasHeartDisease*.

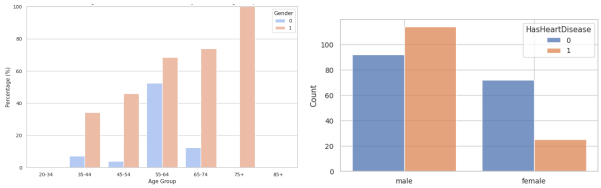
TABLE I: Summary of selected attributes in the Cleveland Heart Disease Dataset

Feature	Description	Type / Range
age	Age of the patient in years.	Continuous
sex	Biological sex (1 = male, 0 = female).	Binary (categorical)
cp	Type of chest pain experienced.	Categorical (1–4)
trestbps	Resting blood pressure (mm Hg).	Continuous
chol	Serum cholesterol in mg/dl.	Continuous
fbs	Fasting blood sugar 120 mg/dl.	Binary
restecg	Resting electrocardiographic results.	Categorical (0–2)
thalach	Maximum heart rate reached during exercise.	Continuous
exang	Exercise-induced angina.	Binary
oldpeak	ST depression induced by exercise relative to rest.	Continuous
slope	Slope of the peak exercise ST segment.	Categorical (1–3)
ca	Number of major vessels (0–3) colored by fluoroscopy.	Discrete
thal	Thalassemia test result.	Categorical (3, 6, 7)
num	Diagnosis of heart disease.	Categorical (0-4)

A. Exploratory Data Analysis

Figure 1a shows the age distribution of patients by sex. Male patients are more prevalent across most age groups, and heart disease prevalence increases with age for both sexes. This pattern is consistent with epidemiological evidence indicating that cardiovascular risk grows over the lifespan[9]. Male patients exhibit higher disease prevalence at younger and middle ages, while the gap narrows in older age groups, reflecting the well-documented increase in cardiovascular risk for women after menopause.

Figure 1b presents the distribution of the target variable (`HasHeartDisease`) by sex. While the overall target variable is relatively balanced, its distribution differs substantially across gender groups. Male patients show a higher proportion of positive cases, whereas female patients are more frequently represented in the negative class. This indicates that marginal balance does not imply joint balance between sex and disease status, a factor that is particularly relevant for fairness evaluation.



(a) Age distribution of patients by sex

(b) Distribution of the target variable by sex

FIG. 1: Comparison of distributions by sex

Correlation analysis identifies several features strongly associated with heart disease. The variables `thal`, `ca`, `exang`, and `cp` show the strongest positive relationships with the target variable, reflecting well-established clinical risk indicators. In contrast, `thalach` is negatively associated with heart disease, indicating that reduced exercise capacity corresponds to higher disease risk. Among categorical variables, `thal`, `cp`, `ca`, and `exang` emerge as the most informative predictors. The feature `sex` shows a moderate association with heart disease, which is clinically plausible given known gender differences in exercise-induced and electrocardiographic diagnostic patterns[10,11].

Notice that sex plays a non-negligible role in disease prevalence and diagnostic patterns, while also constituting a sensitive attribute. These observations motivate a careful fairness-aware analysis, in which sex is retained as an input feature but explicitly examined through group-wise performance metrics and counterfactual evaluations.

B. Modelling and Explanations

The analysis began with a baseline modeling setup using a standard random train-test split. The following classifiers were trained and compared: Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machines (SVM). Among these, SVM achieved the highest overall predictive performance, while Random Forest provided a strong trade-off between accuracy and interpretability. Consequently, subsequent experiments employed Random Forest exclusively. Several variations in the training setup were explored:

- **Stratified splitting** by both gender and target to preserve subgroup distributions.
- **Weighted training** to address class imbalance.
- **Synthetic Minority Over-sampling Technique (SMOTE)** to augment underrepresented classes.

For each training configuration, fairness metrics were computed and counterfactual explanations were generated using DiCE. To ensure clinical plausibility, only continuous physiological features were allowed to vary, while one-hot encoded categorical variables were held fixed. Counterfactuals were generated separately for each combination of gender and target. Additionally, a stress test was conducted in which the only mutable feature was `sex`. Although these counterfactuals are not clinically actionable, they provide a diagnostic signal indicating the extent to which the model relies on this sensitive attribute.

III. RESULTS AND DISCUSSION

A. Splitting the data using train_test_split

1. Fairness Metrics

Despite strong aggregate performance, a disaggregated analysis by sex reveals substantial fairness concerns (Table II). The classifier achieves **perfect accuracy for female patients**, while performance for males is notably lower. This asymmetry is accompanied by a pronounced gap in selection rates, indicating that **male patients are considerably more likely to be classified as having heart disease**. The resulting demographic parity difference highlights a systematic disparity in positive predictions across genders, while differences in equalized odds are primarily driven by a **higher false positive rate among men**. These disparities are consistent with the underlying label distribution. The majority of female patients in the dataset belong to the negative class and, as a result, predicting the negative class for women yields high accuracy, while classification among men is inherently more challenging. Moreover, the smaller size of the female subgroup increases the risk of overfitting, potentially inflating minority-group performance estimates in the test set.

TABLE II: Gender-specific performance and fairness metrics for the Random Forest classifier under the train-test split setup.

Metric	**Female (0)**	**Male (1)**
Accuracy	1.00	0.88
Selection Rate	0.35	0.59
True Positive Rate (TPR)	1.00	0.95
False Positive Rate (FPR)	0.00	0.20
Demographic Parity Difference	0.235	
Equalized Odds Difference	0.20	

2. Counterfactuals

Counterfactual explanations reveal strong directional asymmetries that differ by gender. For male samples (Figure 5), flips from no disease to disease ($0 \rightarrow 1$) are primarily driven by physiological deterioration. Maximum heart rate (`thalach`) decreases substantially ($\approx -29bpm$), while ST depression (`oldpeak`) and vessel count (`ca`) increase moderately. Age, cholesterol, and resting blood pressure show smaller, more variable changes. Flips from disease to no disease ($1 \rightarrow 0$) rely more heavily on structural risk reduction, with large decreases in `ca` and moderate reductions in `oldpeak` and `trestbps`, while `thalach` changes minimally. Sex changes occur in 15.6% of male counterfactuals, exclusively for $1 \rightarrow 0$ flips, indicating that changing sex from male to female can serve as a shortcut for reversing predictions in borderline cases. For female samples (Figure 6), $0 \rightarrow 1$ flips are largely driven by reductions in `thalach` and increases in `oldpeak` and `ca`, with moderate increases in cholesterol and age. In contrast, $1 \rightarrow 0$ flips are highly rigid: `thalach` remains unchanged, while `ca`, `oldpeak`, and `trestbps` decrease uniformly. Sex changes are observed in 58.82% of the cases, exclusively in $0 \rightarrow 1$ flips. This indicates that the model often relies on changing sex to generate positive predictions for women, reflecting both data scarcity and sharp decision boundaries learned from the training set. Stress tests (Table III) further highlight sex-dependent decision thresholds. For men, flipping sex from male to female reverses the prediction for a single borderline case with conflicting clinical signals (e.g., exercise-induced angina combined with low cholesterol and no vessel obstruction). For women, flipping sex from female to male changes the prediction from positive to negative for a single older patient with isolated risk factors (e.g., high blood pressure but otherwise low ischemic risk).

TABLE III: Samples where the stress test caused prediction flips under the train-test split setup.

Age	Trestbps	Chol	Thalach	Oldpeak	CA	Gender Flip	Label FLip
44.0	120.0	169.0	144.0	2.8	0.0	1→0	1→0
59.0	174.0	249.0	143.0	0.0	0.0	0→1	1→0

Overall, these results demonstrate that male predictions are more sensitive to physiological feature changes, while female predictions often require sex flips to achieve class changes. The differing percentages of sex flips (16% for men, 63% for women) illustrate a pronounced gender asymmetry in model reliance on this sensitive attribute.

B. Stratified Split by both gender and target

1. Fariness Metrics

Table IV reports gender-specific performance and fairness metrics under the stratified train-test split. Model performance metrics are more stable and aligned with expected benchmarks for the UCI Heart Disease dataset, indicating that previous anomalies, such as perfect classification for women, were driven by structural imbalance rather than superior generalization. Although the marginal distributions of gender and heart disease prevalence are balanced, the joint distribution remains skewed, which affects fairness evaluation. While the Demographic Parity Difference is small, indicating similar positive prediction rates across genders, Equalized Odds reveals pronounced disparities. Women achieve a true positive rate of 1.00 but a **higher false positive rate compared to men**, who experience fewer false positives but a lower true positive rate. The resulting Equalized Odds Difference highlights persistent fairness violations driven by the joint distribution of sex and target, rather than marginal imbalances or sampling noise.

TABLE IV: Gender-specific performance and fairness metrics for the Random Forest classifier under the stratified train–test split setup.

Metric	Female (0)	Male (1)
Accuracy	0.789	0.857
Selection Rate	0.474	0.500
True Positive Rate (TPR)	1.000	0.826
False Positive Rate (FPR)	0.286	0.105
Demographic Parity Difference	0.0263	
Equalized Odds Difference	0.1805	

2. Counterfactuals

Counterfactual explanations under the jointly stratified train-test split reveal that the model now learns cleaner, subgroup-specific decision rules. For men (Figure 9), flips from no disease to disease (0→1) are primarily driven by decreases in `thalach` and increases in `oldpeak` and `ca`, with minor contributions from age, cholesterol, and resting blood pressure (`trestbps`). Flips from disease to no disease (1→0) rely mainly on reductions in `ca` and `trestbps`, with `thalach` playing a smaller role. Sex changes occur in only 17% of male counterfactuals, exclusively in 1→0 flips, showing similar results as the baseline. For women (Figure 10), 0→1 flips show consistent increases in `ca`, `oldpeak`, and cholesterol, while 1→0 flips are highly rigid, with zero change in `thalach` and uniform decreases in `ca` and `oldpeak`. Again, sex changes occur only in 0→1 counterfactuals (≈ 63% of samples), indicating that flipping sex is necessary in more than half the cases to predict disease for female cases. Finally, no counterfactual flips were observed under the sex-only stress test, suggesting that preserving the joint distribution of sex and target **stabilizes decision boundaries and reduces unwarranted reliance on gender**. Overall, the results demonstrate that stratification slightly improves feature-driven counterfactual explanations by mitigating spurious sex dependence in predictions.

C. Weighted Training

1. Fairness Metrics

Table V shows that reweighting the training samples exacerbates gender asymmetry under severe subgroup imbalance. While intended to improve fairness by amplifying underrepresented groups, the few women with heart disease in the dataset are insufficient to define a robust decision boundary. Consequently, the model overfits these sparse positive examples, resulting in a persistently perfect TPR for women, an increased female FPR, and almost no false positives for men. Because these effects worsen fairness and interpretability, counterfactuals were not computed for this setup. This illustrates that reweighting can increase variance and asymmetric decision boundaries when subgroups are extremely underrepresented.

TABLE V: Gender-specific performance and fairness metrics for the Random Forest classifier under the weighted training setup.

Metric	Female (0)	Male (1)
Accuracy	0.737	0.810
Selection Rate	0.526	0.405
True Positive Rate (TPR)	1.00	0.696
False Positive Rate (FPR)	0.357	0.053
Demographic Parity Difference	0.122	
Equalized Odds Difference	0.305	

D. SMOTE

1. Fairness Metrics

Table VI shows the impact of applying SMOTE to the underrepresented female-positive class. Compared to the baseline stratified split, SMOTE substantially reduces the Equalized Odds difference (0.154 vs. 0.261), indicating more balanced error rates across genders. Female TPR remains at 1.00, while male TPR increases to 0.846, narrowing the asymmetry in predictive performance. However, the Demographic Parity difference increases slightly (0.117 vs. 0.008), reflecting that overall prediction rates for positive outcomes are now less

aligned between sexes. SMOTE effectively exposes the model to female-positive patterns, smoothing subgroup-specific decision boundaries and reducing reliance on male-dominated signals. This demonstrates a fundamental trade-off between optimizing representation to improve subgroup recall and maintaining parity in overall prediction rates.

TABLE VI: Gender-specific performance and fairness metrics for the Random Forest classifier with SMOTE applied to the female positive class.

Metric	**Female (0)**	**Male (1)**
Accuracy	0.923	0.820
Selection Rate	0.423	0.540
True Positive Rate (TPR)	1.00	0.846
False Positive Rate (FPR)	0.118	0.208
Demographic Parity Difference	0.117	
Equalized Odds Difference	0.154	

2. *Counterfactuals*

Although the classifier was trained using SMOTE to mitigate class imbalance, counterfactual explanations were generated exclusively from real samples in the original training data. This ensures that all explanations reflect clinically plausible feature configurations rather than synthetic interpolations.

For men, counterfactuals exhibit reduced reliance on sex compared to earlier setups, with sex changes occurring in only 10.7% of cases and exclusively in $1 \rightarrow 0$ flips. Disease-inducing counterfactuals ($0 \rightarrow 1$) are primarily driven by substantial increases in age ($\approx +8.4$ years on average), alongside increases in the number of affected vessels (`ca`) and ST depression (`oldpeak`), and decreases in maximum heart rate (`thalach`). This represents a qualitative shift from previous experiments, where age played a negligible role, indicating that SMOTE enables age-driven pathways to disease that were previously unavailable. In contrast, reversals ($1 \rightarrow 0$) rely more strongly on reductions in `ca`, `oldpeak`, and resting blood pressure, with sex only occasionally acting as a sufficient change. Despite this overall stabilization, the sex-only stress test (Table VII) reveals two male cases where flipping sex from male to female is sufficient to reverse the prediction. These individuals present mixed clinical signals, combining high ischemic risk indicators (e.g., elevated `oldpeak` or multiple affected vessels) with strong protective features such as high exercise capacity. The persistence of these flips indicates that SMOTE alters the local decision geometry, making sex a decisive tie-breaker for borderline profiles.

For women, sex remains a more influential attribute. Approximately 47% of samples counterfactuals require changing sex from female to male, all in $0 \rightarrow 1$ flips. While this is moderately less than in previous experiments, it's still a big difference when comparing to male samples. Feature-based pathways improve relative to earlier experiments, with larger shifts in age, `thalach`, `ca`, and `oldpeak`, yet sex continues to act as a shortcut for entering the positive region of the decision space. The sex-only stress test identifies a single elderly female patient for whom flipping sex alone induces a positive prediction, despite limited ischemic evidence. This confirms that, under SMOTE, male sex systematically amplifies predicted risk for older patients near the decision boundary. These findings highlight a key tension between

TABLE VII: Samples where the stress test caused prediction flips with SMOTE applied to the female positive class

Age	Trestbps	Chol	Thalach	Oldpeak	CA	Gender Flip	Label FLip
58.0	132.0	224.0	173.0	3.2	2.0	1→0	1→0
44.0	120.0	169.0	144.0	2.8	0.0	1→0	1→0
74.0	120.0	269.0	121.0	0.2	1.0	0→1	0→1

group-level and instance-level fairness. SMOTE substantially reduces the Equalized Odds Difference by improving average recall parity across genders. However, by densifying the feature space with synthetic minority-class samples, it also places more real individuals near the decision boundary. As a result, small perturbations, such as changing sex, are more likely to flip predictions. In contrast, the jointly stratified split preserves the joint distribution $P(\text{sex}, \text{target})$, yielding fewer boundary-adjacent cases and more stable counterfactual behavior, despite slightly higher aggregate fairness disparities.

Overall, SMOTE improves fairness in expectation, but at the cost of increased sensitivity to the protected attribute at the individual level, underscoring the importance of combining aggregate fairness metrics with counterfactual analyses when evaluating model behaviour.

IV. CONSLUSIONS AND FUTURE WORK

Across all experimental settings, a consistent structural asymmetry emerges in the model's decision logic. The classifier learns a feature-driven disease boundary for men, whereas predictions for women rely more heavily on sex as a mediating factor. For male patients, counterfactual explanations typically reach a prediction flip through changes in physiological features alone. For female patients, particularly under imbalance, prediction changes frequently require altering sex, indicating that the model lacks sufficient positive examples to learn nuanced, feature-based disease patterns for women.

Stratified splitting stabilizes performance and counterfactual behaviour by preserving the joint distribution of sex and outcome, but it does not fully eliminate this asymmetry. SMOTE partially mitigates the issue by introducing new feature-based pathways, most notably

through age, yet reliance on sex remains non-negligible. Overall, the stronger and more rigid decision boundary observed for women reflects data scarcity rather than superior modelling.

From a clinical perspective, flipping sex in a counterfactual can be justified only when sex acts as a proxy for biological mechanisms not fully captured by the available features and when the change interacts plausibly with other variables. In many instances, sex flips occur jointly with increases in age, cholesterol, or `oldpeak`, and decreases in `thalach`, which aligns with known epidemiological patterns, such as earlier male risk and age-dependent risk increases for women. In these cases, sex functions as a contextual risk modifier rather than a shortcut. However, sex flips become problematic when changing sex alone is sufficient to alter the prediction, with minimal or no changes in physiological variables. Such behaviour lacks clinical meaning and indicates shortcut learning driven by representation gaps. This pattern appears disproportionately for women, with sex flips occurring far more frequently than for men across all non-jointly stratified settings. Even after SMOTE, while the frequency of sex flips decreases, they remain more prevalent for women, suggesting that imbalance is a primary driver but not the sole explanation. Historical under-diagnosis and label bias may also contribute. The sex-only stress tests further clarify this behaviour. Prediction changes occur exclusively for individuals with borderline or conflicting clinical profiles, where strong risk and protective factors coexist. In these cases, sex acts as a threshold-shifting feature rather than a dominant causal determinant. Notably, no sex-only counterfactual flips are observed under the jointly stratified split, indicating that preserving the joint distribution of sex and outcome stabilizes the decision boundary and reduces unwarranted sensitivity to sex. In contrast, SMOTE increases the number of such flips, highlighting that synthetic oversampling can unintentionally amplify reliance on sensitive attributes near the decision boundary. Overall, the counterfactual analysis indicates that observed gender disparities arise primarily from data distribution effects and boundary instability rather than inherent clinical necessity. This underscores the importance of careful data splitting strategies and fairness-aware evaluation when deploying predictive models in sensitive medical contexts.

Several avenues remain to further investigate counterfactual fairness under gender imbalance in clinical prediction models. First, alternative methods for addressing data imbalance could be explored. These include variants of SMOTE (e.g., Borderline-SMOTE, ADASYN), generative approaches such as variational autoencoders or GAN-based oversampling, and more sophisticated reweighting schemes that account for joint feature-target distributions. Such methods may provide richer minority-class representations while reducing reliance on sensitive attributes as shortcuts. Second, the counterfactual analysis itself could be extended. While this study employed DiCE with constrained feature perturbations, future work could compare additional counterfactual generation techniques, including causal or model-agnostic approaches, to assess robustness across methods and identify any sensitivity to algorithmic assumptions. Third, incorporating causal reasoning or domain knowledge into both model training and counterfactual generation may help distinguish clinically meaningful pathways from spurious shortcuts. For instance, explicitly modelling sex-specific physiological mechanisms could reduce the model's dependence on sex flips and improve the interpretability and fairness of predictions. Finally, scaling the analysis to larger or multi-center datasets would allow evaluation of whether the observed asymmetries persist in more diverse populations and under different feature distributions.

ACKNOWLEDGEMENTS

This document was reviewed and refined with the assistance of Large Language Models, which helped check grammar, correct typos, increase code productivity, and enhance clarity. The overall content and ideas remain solely the responsibility of the author.

[1] M. Hardt, E. Price, and N. Srebro, CoRR **abs/1610.02413** (2016), 1610.02413.
[2] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, Science **366**, 447–453 (2019).
[3] S. Wachter, B. D. Mittelstadt, and C. Russell, CoRR **abs/1711.00399** (2017), 1711.00399.
[4] R. K. Mothilal, A. Sharma, and C. Tan, CoRR **abs/1905.07697** (2019), 1905.07697.
[5] F. Kamiran and T. Calders, Knowledge and Information Systems **33**, 1–33 (2011).
[6] I. M. Alkhawaldeh, I. Albalkhi, and A. J. Naswhan, World Journal of Methodology **13**, 373–378 (2023).
[7] S. W. P. M. Janosi, Andras and R. Detrano, "Heart Disease," UCI Machine Learning Repository (1989), DOI: https://doi.org/10.24432/C52P4X.
[8] K. V. V. Reddy, I. Elamvazuthi, A. A. Aziz, S. Paramasivam, H. N. Chua, and S. Pranavanand, Applied Sciences **11** (2021), 10.3390/app11188352.
[9] World Health Organization, "Global health estimates 2021: Disease burden by cause, age, sex, by country and by region, 2000–2019," Geneva: WHO (2021).
[10] P. M. Okin and P. Kligfield, Circulation **92**, 1209–1216 (1995).
[11] M. Gulati, P. Pratap, P. Kansal, J. E. Calvin, and R. C. Hendel, The American Journal of Cardiology **94**, 997–1002 (2004).

Appendix A: Exploratory Data Analysis Additional Figures

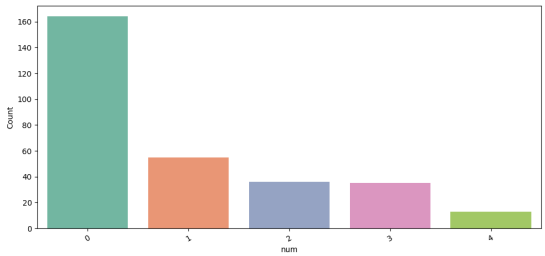


FIG. 2: Distribution of the original target variable (`num`) before conversion to a binary label.

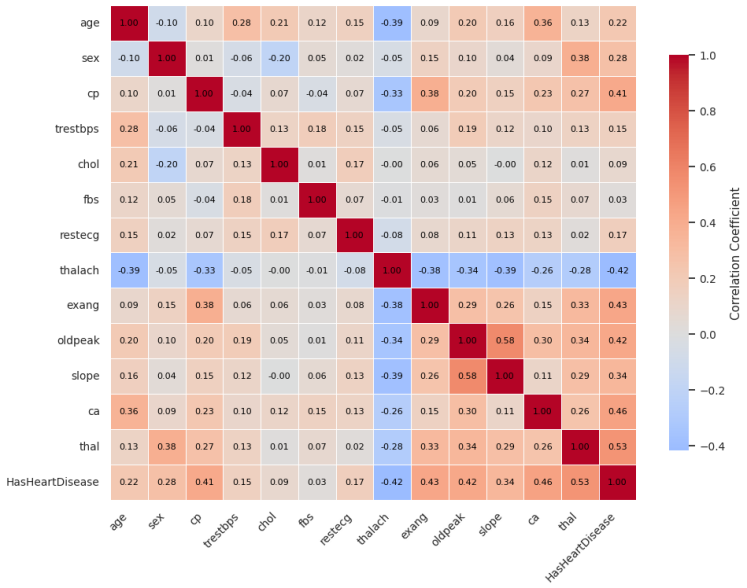


FIG. 3: Pearson correlation matrix for all features in the Cleveland Heart Disease dataset.

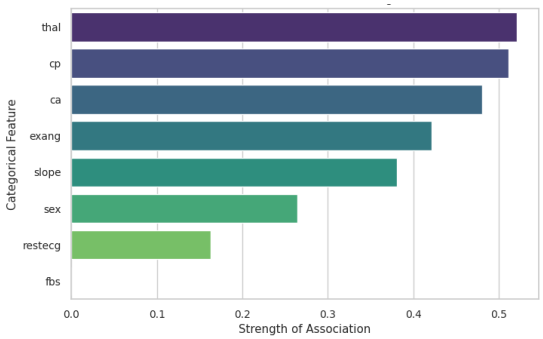
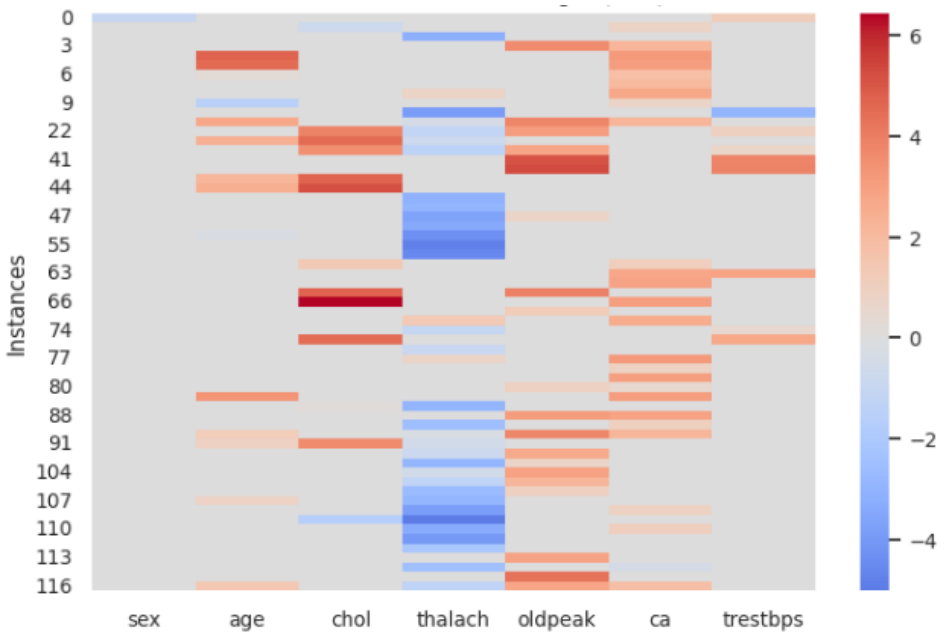
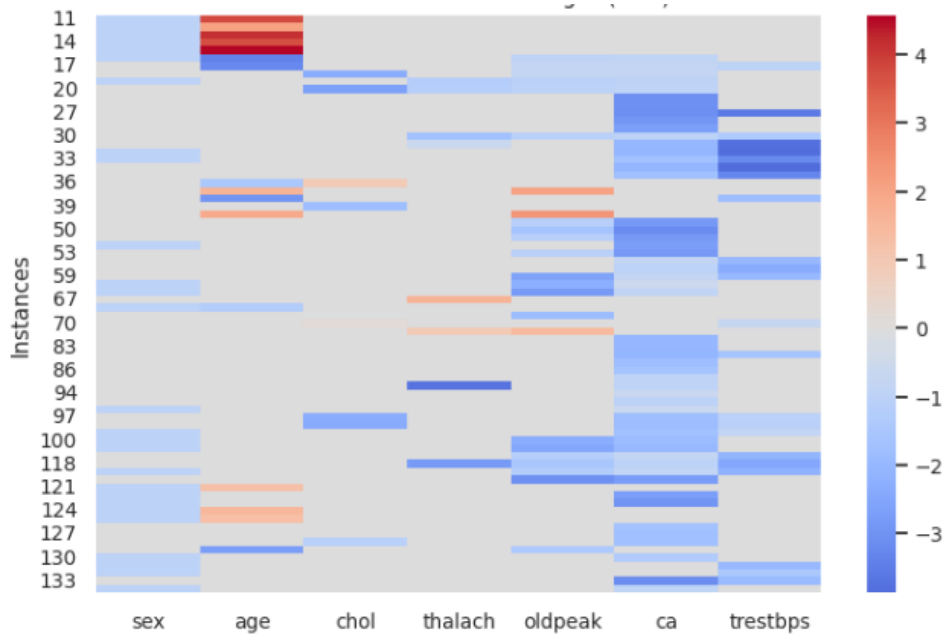


FIG. 4: Cramér's V correlation between categorical features and the target variable in the Cleveland Heart Disease dataset.

Appendix B: Counterfactuals Heatmaps

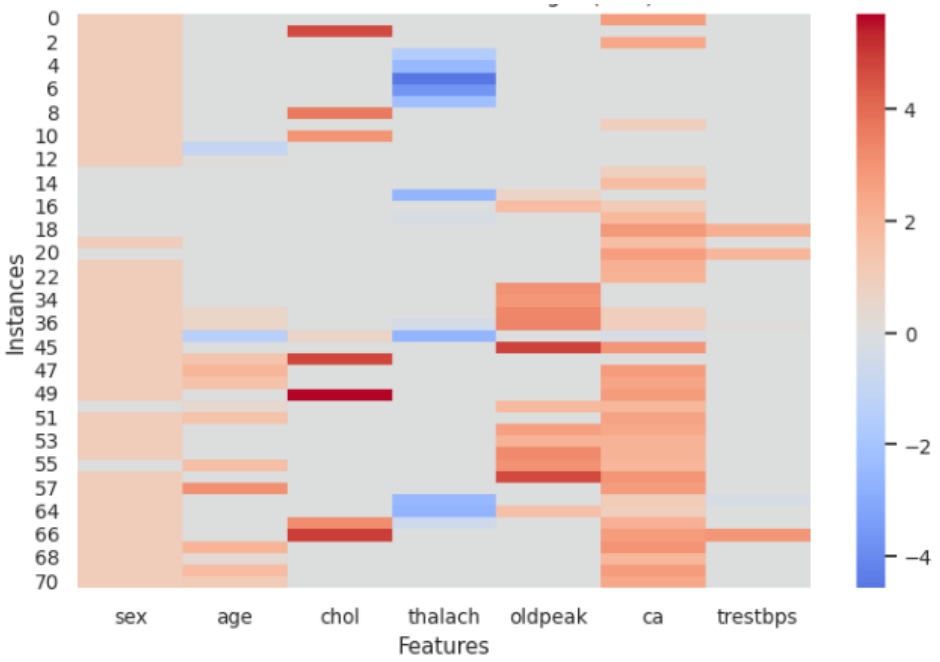


(a) Counterfactual flips: 0 → 1 (no disease → disease)

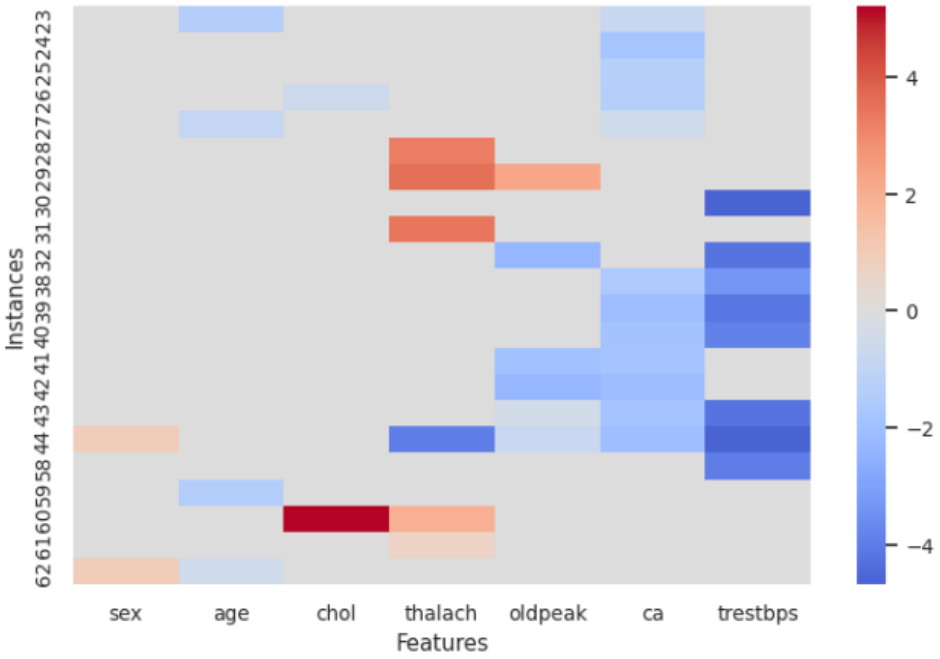


(b) Counterfactual flips: 1 → 0 (disease → no disease)

FIG. 5: Heatmaps showing feature changes in counterfactual explanations for male patients under the `train_test_split` setup. Warmer colors indicate an increase in the feature value, while cooler colors indicate a decrease.

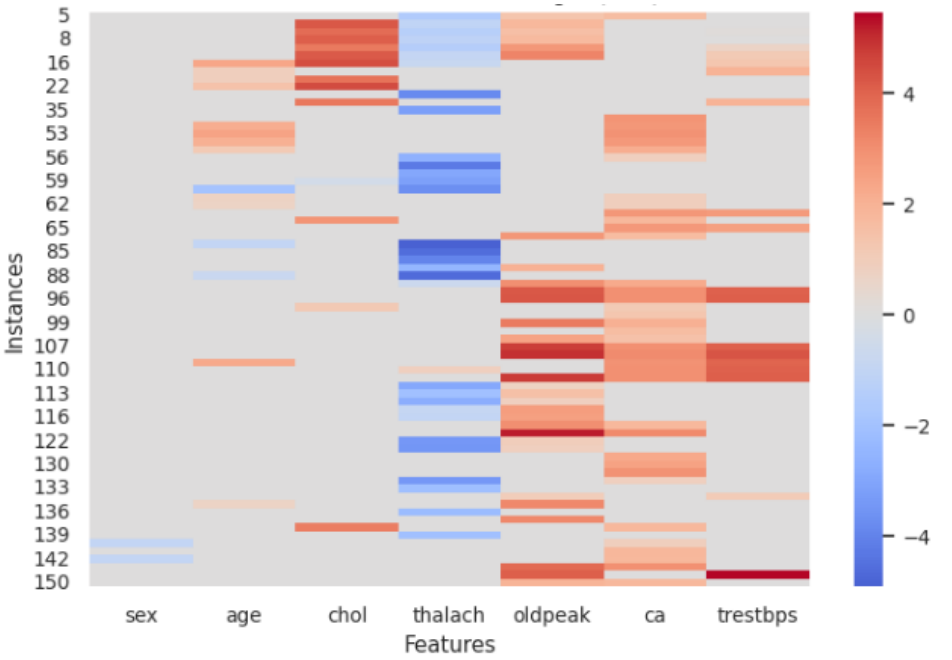


(a) Counterfactual flips: 0 → 1 (no disease → disease)

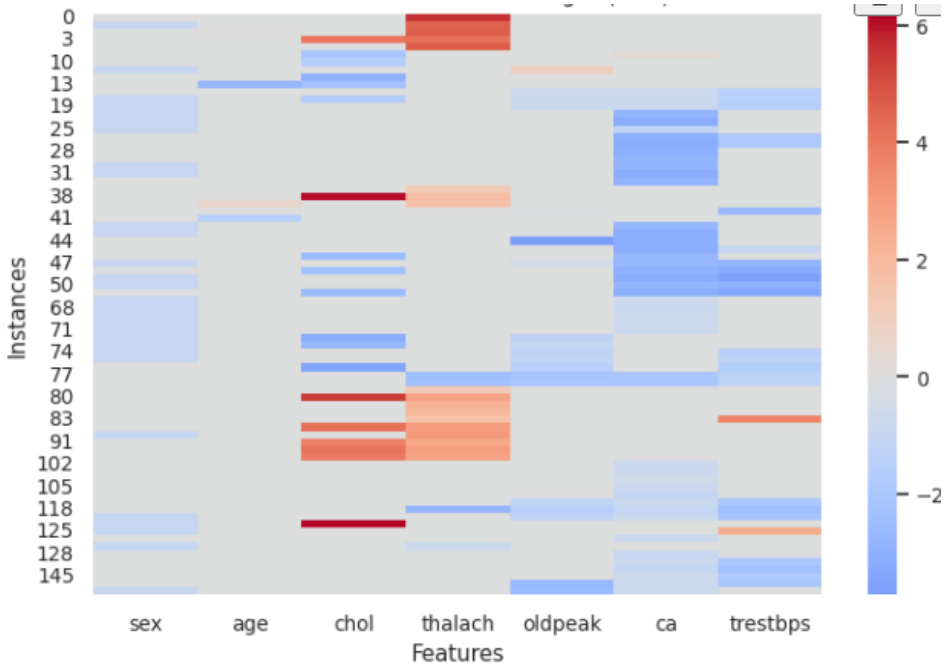


(b) Counterfactual flips: 1 → 0 (disease → no disease)

FIG. 6: Heatmaps showing feature changes in counterfactual explanations for female patients under the `train_test_split` setup.

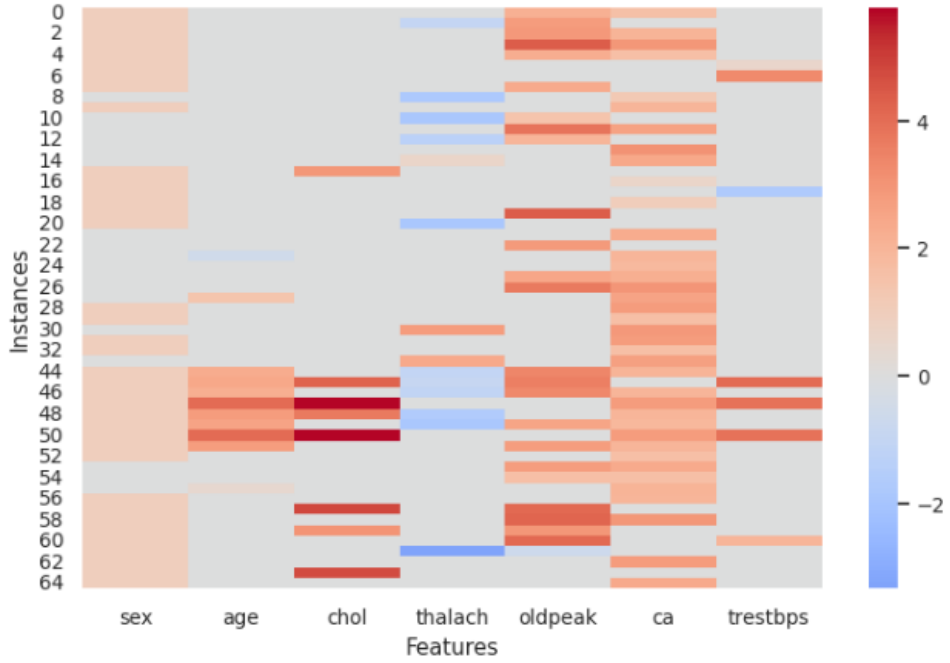


(a) Counterfactual flips: 0 → 1 (no disease → disease)

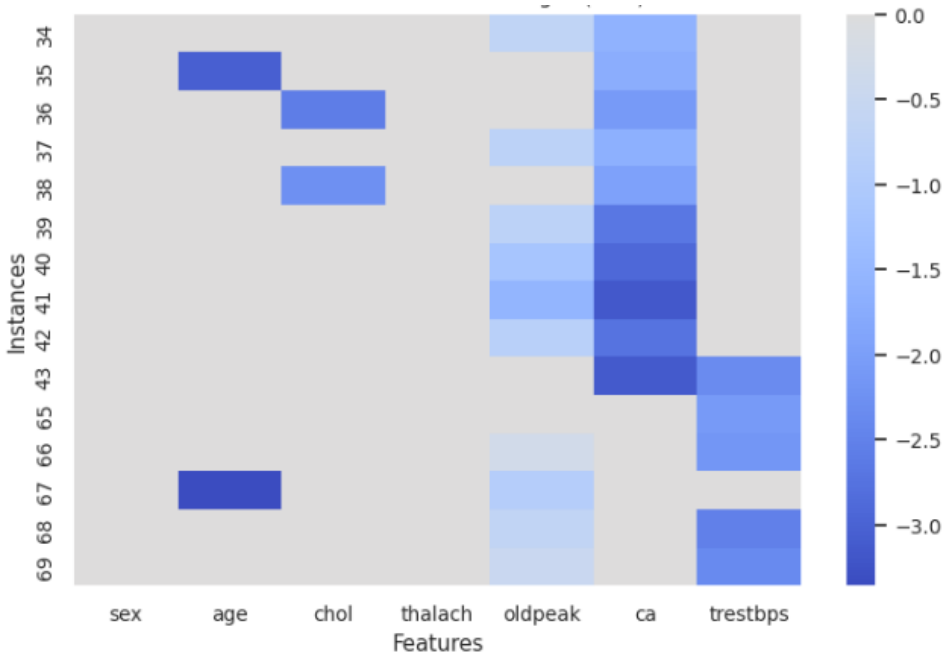


(b) Counterfactual flips: 1 → 0 (disease → no disease)

FIG. 7: Heatmaps showing feature changes in counterfactual explanations for male patients under the stratified split setup.

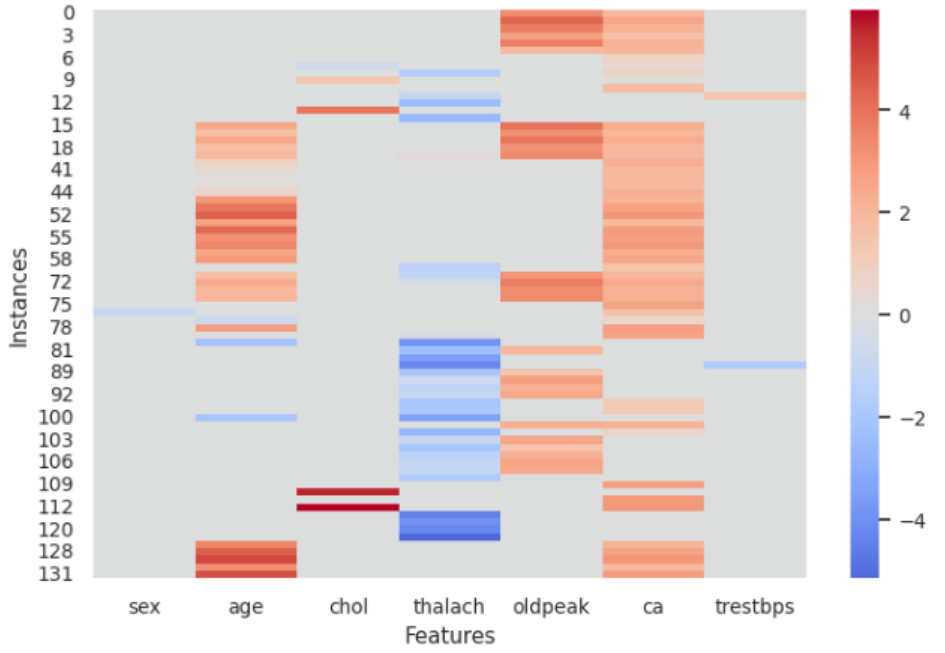


(a) Counterfactual flips: 0 → 1 (no disease → disease)

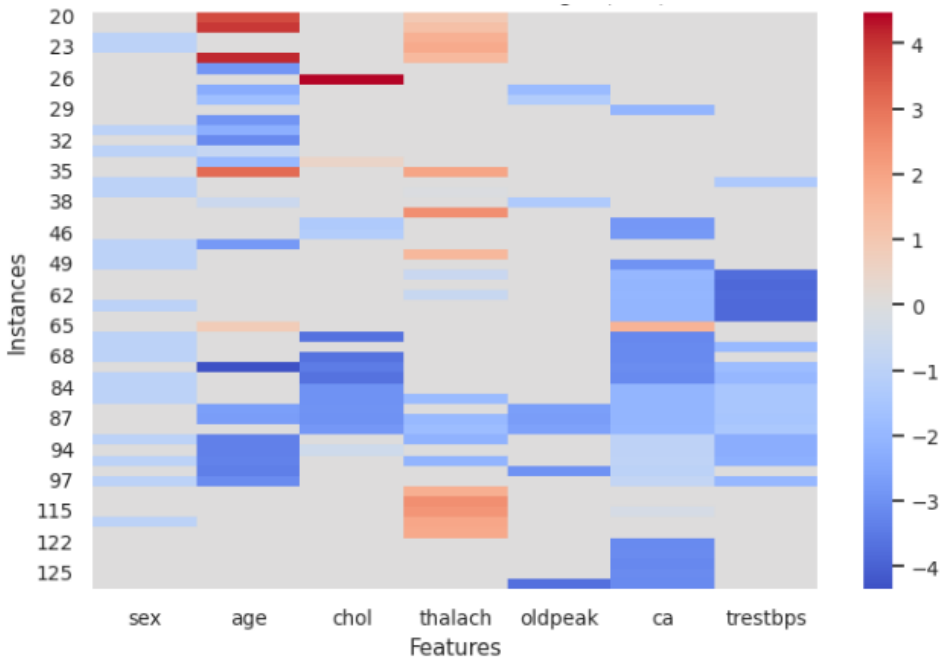


(b) Counterfactual flips: 1 → 0 (disease → no disease)

FIG. 8: Heatmaps showing feature changes in counterfactual explanations for female patients under the stratified split setup.

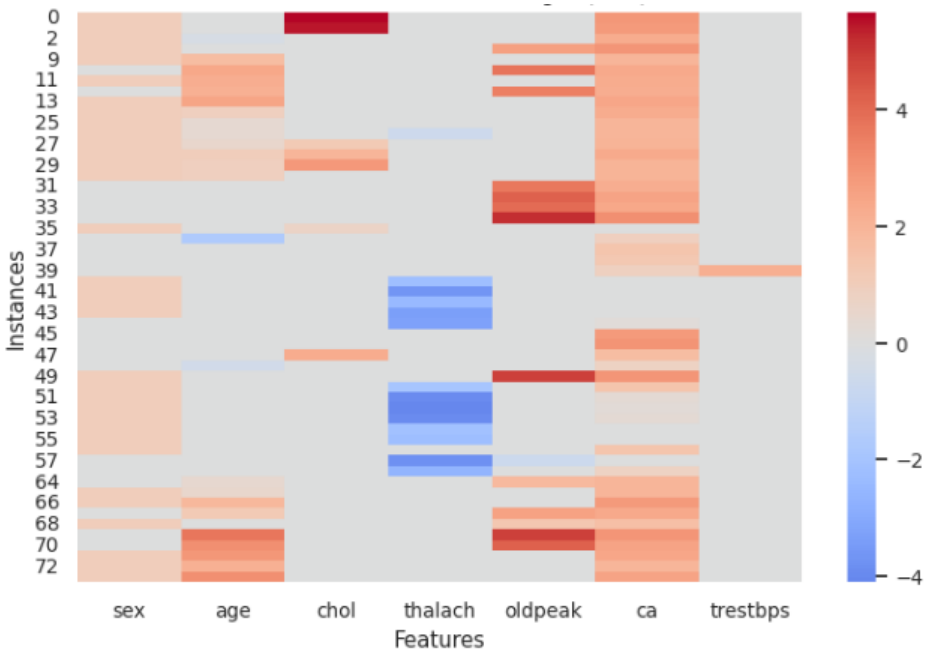


(a) Counterfactual flips: 0 → 1 (no disease → disease)

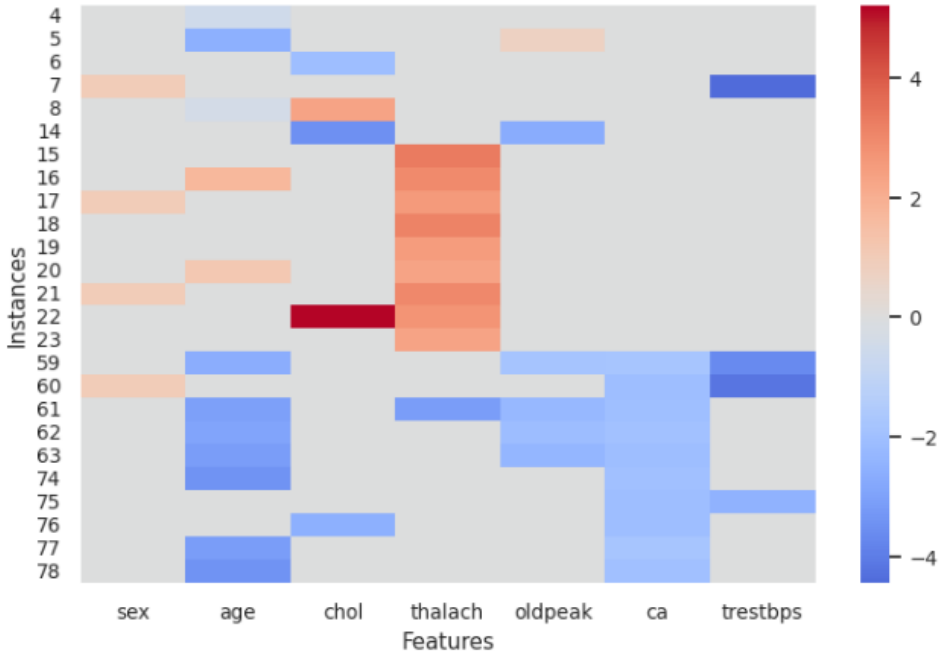


(b) Counterfactual flips: 1 → 0 (disease → no disease)

FIG. 9: Heatmaps showing feature changes in counterfactual explanations for male patients under the stratified split setup.



(a) Counterfactual flips: 0 → 1 (no disease → disease)



(b) Counterfactual flips: 1 → 0 (disease → no disease)

FIG. 10: Heatmaps showing feature changes in counterfactual explanations for female patients under the stratified split setup.