Fairness Audit and Bias Mitigation in an AI-Based Heart Disease Risk Prediction Model

Aditya Bhor[1], Sujal More[2], Vedant Alhat[3], Prof. Pallavi Gholap[4]
[1,2,3,4]*Department of Science and Computer Science, MIT ACSC, Alandi, Pune*
doi.org/10.64643/ijirt.208439-459

***Abstract*—Artificial intelligence (AI) is increasingly used for healthcare prediction, but machine learning models may produce different outcomes for different demographic groups. This study presents a fairness audit and bias mitigation analysis of an AI-based heart disease risk prediction model using a dataset of 918 patient records. A Logistic Regression model was developed using one-hot encoding, feature scaling, and an 80:20 stratified train-test split. The model achieved an accuracy of 88.59%, with a precision of 87.16%, recall of 93.14%, and F1-score of 90.05%. Fairness was evaluated for Sex and AgeGroup using Demographic Parity Difference (DPD), Equalized Odds Difference (EOD), and Disparate Impact Ratio (DIR) through the Fairlearn framework. The baseline model showed noticeable demographic differences, particularly across Sex, with a DPD of 0.5144 and a DIR of 0.2637. To reduce these disparities, two bias mitigation approaches, Exponentiated Gradient and Threshold Optimizer, were applied using Sex as the sensitive feature. Exponentiated Gradient reduced the Sex DPD to 0.0512 and increased the Sex DIR to 0.9188, while Threshold Optimizer reduced the Sex DPD to 0.0865 and increased the Sex DIR to 0.8749. However, these fairness improvements were accompanied by lower overall accuracy, with Exponentiated Gradient achieving 82.61% and Threshold Optimizer achieving 81.52%. The results show that fairness improvements can involve a trade-off with predictive performance and that evaluating multiple fairness measures is important when assessing machine learning models for healthcare applications.**

***Index Terms*—Fairness in AI, Bias Mitigation, Heart Disease Prediction, Fairlearn, Logistic Regression, Demographic Parity, Equalized Odds, Disparate Impact, Responsible AI, Healthcare Analytics.**

I. INTRODUCTION

Machine learning has become an important approach for analyzing clinical data and supporting predictive tasks in healthcare. Risk prediction models can identify statistical relationships among demographic, physiological, and clinical variables and can provide quantitative estimates for binary prediction tasks. Heart disease prediction is a representative application because the underlying datasets contain a mixture of numerical and categorical clinical variables that can be processed using conventional supervised-learning methods [1], [2].

Although predictive accuracy is an important consideration, healthcare systems require more than aggregate performance. A model can achieve satisfactory overall accuracy while producing different prediction patterns across demographic groups. Research on algorithmic fairness has established that disparities can arise from the data, the learning process, the model, or the way predictions are converted into decisions [3], [4]. In healthcare, such disparities are especially important because model outputs may influence how individuals are prioritized, assessed, or recommended for further attention [5]–[7].

Fairness is not represented by a single universally applicable definition. Different fairness criteria focus on different properties of a prediction system. Demographic parity examines differences in positive prediction rates between groups, whereas equalized-odds-based criteria examine differences in error-related quantities conditional on the true outcome [8], [9]. Consequently, improving one fairness criterion does not necessarily improve all others.

Bias mitigation methods have therefore been proposed at different stages of machine-learning development. Preprocessing approaches attempt to alter data distributions [10], [11], while reduction-based methods incorporate fairness constraints during model optimization. Postprocessing methods can modify prediction thresholds to achieve a selected fairness objective. Fairlearn provides implementations for

auditing and mitigating fairness-related disparities in machine-learning systems [12], [13].

The objective of this study is to evaluate whether a conventional logistic regression heart disease risk prediction model exhibits measurable demographic disparities and to assess how two Fairlearn-based mitigation approaches alter those disparities and the predictive performance of the system. The analysis considers Sex and an age-based grouping, AgeGroup, as demographic attributes. The study does not attempt to establish clinical effectiveness or clinical deployment readiness; it focuses specifically on predictive performance and fairness characteristics within the provided dataset and experimental setup.

The main contributions of this study are as follows. First, a clearly defined logistic regression heart disease prediction workflow is evaluated on the 918-record dataset. Second, the resulting predictions are audited across Sex and Age Group using DPD, EOD, and DIR. Third, Exponentiated Gradient with a Demographic Parity constraint and Threshold Optimizer are applied and compared against the baseline. Finally, the study quantifies the resulting trade-off between aggregate predictive performance and fairness across the evaluated demographic attributes.

II. RELATED WORK

A. Machine Learning for Heart Disease Prediction

Machine-learning methods have been widely explored for heart disease classification and risk prediction. The supplied literature includes studies using the UCI-derived heart disease data for classification and interpretability [1], as well as approaches involving machine-learning optimization for heart disease prediction [2]. These studies illustrate the value of predictive modeling for cardiovascular-risk classification but do not, by themselves, establish fairness across demographic groups.

B. Algorithmic Fairness in Machine Learning

Algorithmic fairness has been studied extensively across supervised-learning settings. Mehrabi et al. provide a broad survey of bias and fairness concepts and emphasize that fairness is multidimensional [3]. Barocas and Selbst discuss how differences in outcomes can emerge from data-driven decision systems even when sensitive variables are not explicitly used for a decision [4]. Chouldechova further demonstrates the difficulty of satisfying multiple fairness criteria simultaneously [9].

The distinction among fairness definitions is important for healthcare prediction. A model may satisfy a demographic-rate criterion while still producing unequal error-related behavior across groups. Consequently, fairness auditing should report multiple metrics rather than relying on a single score [3], [8], [9].

C. Fairness in Healthcare AI

Fairness concerns become particularly consequential in medicine because prediction systems operate on populations with heterogeneous characteristics and potentially unequal representation in training datasets. Obermeyer et al. demonstrated how a widely used healthcare algorithm could exhibit racial disparity due to the relationship between the prediction target and healthcare utilization rather than direct measurement of health need [5]. Chen et al. discuss algorithmic fairness in medicine and healthcare and emphasize the need to evaluate fairness as part of the development and assessment of medical AI systems [6]. Norori et al. similarly argues for greater attention to bias, transparency, and open scientific practices in health-related AI [7].

These studies motivate the present work's emphasis on explicit demographic auditing rather than aggregate predictive accuracy alone.

D. Bias Mitigation

Bias mitigation methods can operate before model training, during optimization, or after a predictive model has been trained. Kamiran and Calders describe preprocessing and classification strategies intended to reduce discriminatory outcomes [10], [11]. In the present study, mitigation occurs through two Fairlearn mechanisms rather than through dataset-level preprocessing intended specifically for fairness.

Exponentiated Gradient is a reduction-based approach in which a base estimator is optimized under a fairness constraint. In this study, Logistic Regression was used as the base estimator with a Demographic Parity constraint.

The Threshold Optimizer, in contrast, was applied after fitting the baseline estimator and changes the decision rule subject to a demographic-parity constraint.

E. Fairness-Aware Machine-Learning Tooling

Fairlearn was developed as a toolkit for assessing and improving fairness in AI systems [12], with subsequent work documenting its broader use and functionality [13]. The present implementation uses Fairlearn's MetricFrame and its demographic-parity, equalized-odds, and disparate-impact-related metrics for auditing, followed by Exponentiated Gradient and Threshold Optimizer for mitigation.

The research gap addressed by this study is therefore practical rather than purely theoretical: the work evaluates the same prediction workflow before and after fairness mitigation and reports how the selected fairness measures change alongside accuracy, precision, recall, and F1-score.

III. DATASET

A. Dataset Description

The study uses the Heart Failure Prediction Dataset available through Kaggle [14]. The dataset contains 918 observations and 12 original attributes: Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, Exercise Angina, Oldpeak, ST_Slope, and heart disease. The target variable, heart disease, is binary, with 0 representing the negative class and 1 representing the positive class. The positive class accounts for 55.3377% of the observations. The dataset contains 725 male and 193 female observations. An additional demographic variable, Age Group, was constructed by categorizing participants into two groups: <50 years and ≥50 years.

The demographic variables investigated in the fairness audit are therefore Sex and AgeGroup.

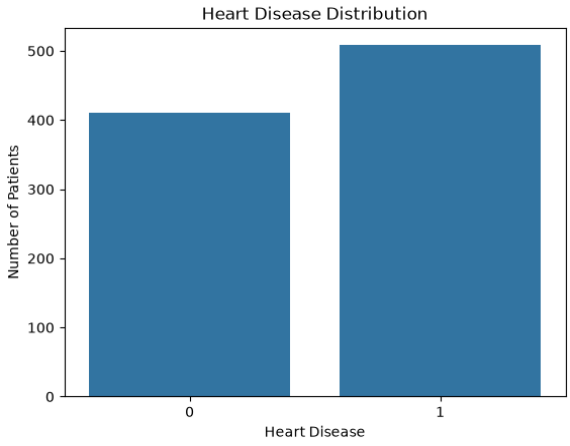

Fig. 1. Heart disease class distribution.

B. Data Quality and Preprocessing

Data-quality checks identified no missing values and no duplicate rows, leaving all 918 observations available for analysis. The categorical input variables were converted into numerical representations using one-hot encoding with drop_first=True. The target variable, HeartDisease, and the derived AgeGroup variable were excluded from the ordinary model input matrix, while Age and the encoded Sex feature were retained as predictive inputs. This preprocessing resulted in 15 model input features. Numerical features were standardized using StandardScaler. The scaler was fitted using the training data and subsequently applied to the training and test sets to avoid fitting the scaling transformation on the held-out test observations.

IV. METHODOLOGY

A. Baseline Prediction Model

Logistic Regression was used as the baseline classification model for heart disease prediction. The model was configured with random_state=42 and max_iter=1000 and used 15 input features. The model estimates the probability of the positive class using the logistic function. No model-selection or hyperparameter-search experiment was performed; therefore, Logistic Regression was used as the predefined baseline rather than being selected through comparative optimization against multiple predictive models.

B. Experimental Design

An 80:20 stratified train-test split was used with random_state=42, resulting in 734 training observations and 184 test observations. The training data were used to fit the feature-scaling transformation and the baseline Logistic Regression model. Predictions were generated on the held-out test set and evaluated using predictive performance measures. The same test set was subsequently used for fairness auditing and post-mitigation evaluation.

The overall experimental workflow is illustrated in Fig. 2. The process consists of dataset preparation, data-quality assessment, feature construction, categorical encoding, stratified train-test splitting, feature standardization, baseline prediction, predictive performance evaluation, fairness auditing, bias

mitigation, and post-mitigation evaluation. The workflow compares the baseline model with Exponentiated Gradient and Threshold Optimizer to examine the resulting fairness-performance trade-off.

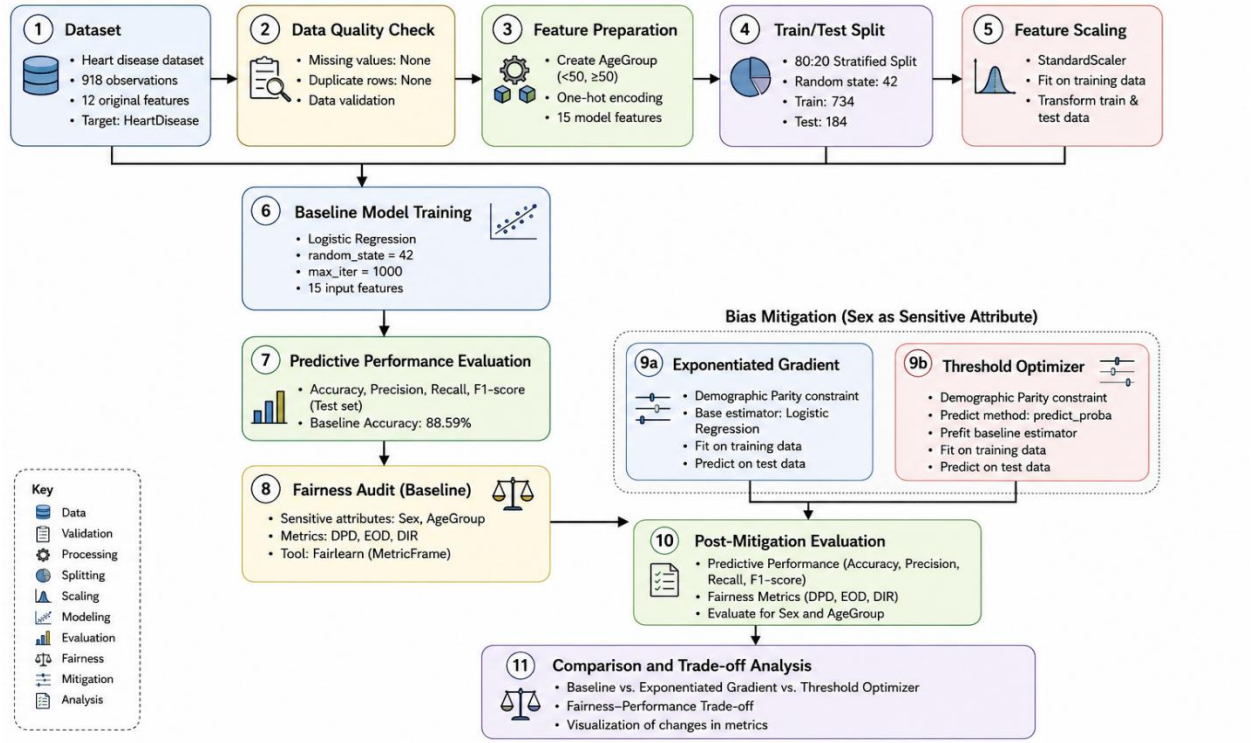

Fig. 2. Experimental workflow for heart disease prediction, fairness auditing, and bias mitigation.

No cross-validation, repeated holdout evaluation, statistical significance testing, bootstrap uncertainty intervals, or external validation was performed in this study.

C. Bias Mitigation

1) Exponentiated Gradient

Exponentiated Gradient was implemented using Logistic Regression as the base estimator with a Demographic Parity constraint. Exponentiated Gradient was implemented using Logistic Regression as the base estimator with a Demographic Parity constraint. Sex was supplied as the sensitive feature during fitting, using sex_train, and the resulting model was evaluated on the held-out test set.

2) Threshold Optimizer

Threshold Optimizer was applied to the previously fitted Logistic Regression model using a Demographic Parity constraint and probability-based prediction. The baseline estimator was treated as prefit, with Sex supplied as the sensitive feature during fitting and prediction. The optimizer was fitted using the training data and sex_train, and predictions were generated using sex_test for the held-out test set.

A critical interpretation point is that Sex is the sensitive feature used in both mitigation procedures. AgeGroup is evaluated as an additional fairness attribute after mitigation. The mitigation experiment therefore should not be interpreted as a single optimization simultaneously enforcing demographic parity for both Sex and AgeGroup.

V. EXPERIMENTAL SETUP

A. Predictive Performance Measures

Predictive performance was evaluated using accuracy, precision, recall, and F1-score.

Accuracy is

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN). \quad (3)$$

Precision is

$$\text{Precision} = TP / (TP + FP). \quad (4)$$

Recall is

$$\text{Recall} = TP / (TP + FN). \quad (5)$$

F1-score is the harmonic mean of precision and recall:

$$F1 = 2(\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}). \quad (6)$$

ROC-AUC was not calculated and is therefore not reported in this study.

B. Fairness Measures

Fairness was evaluated with respect to Sex and AgeGroup using Fairlearn's MetricFrame. Three fairness measures were considered: Demographic Parity Difference (DPD), Equalized Odds Difference (EOD), and Disparate Impact Ratio (DIR). These measures were selected to examine differences in positive prediction rates and outcome-related error rates across demographic groups.

For a sensitive attribute (A), a group (g), and predicted label ($\hat{Y}$), the demographic-parity difference used in the analysis is the range of group selection rates:

$$DPD = \max_g P(\hat{Y} = 1 \mid A = g) - \min_g P(\hat{Y} = 1 \mid A = g). \quad (7)$$

A value closer to zero indicates smaller differences in positive prediction rates.

The disparate impact ratio is represented by the ratio of the lowest group selection rate to the highest group selection rate:

$$DIR = \min_g P(\hat{Y} = 1 \mid A = g) / \max_g P(\hat{Y} = 1 \mid A = g). \quad (8)$$

A value closer to one indicates more similar selection rates.

Equalized odds evaluates whether error-related rates are similar across groups. In the form used by Fairlearn in this study, the measure can be represented as

$$EOD = \max(\max_g TPR_g - \min_g TPR_g, \max_g FPR_g - \min_g FPR_g). \quad (9)$$

where (TPR_g) and (FPR_g) are the true-positive and false-positive rates of group (g).

These measures are complementary. DPD and DIR primarily characterize differences in positive prediction rates, whereas EOD incorporates group differences associated with the true outcome.

VI. RESULTS

A. Dataset Characteristics

TABLE I DATASET AND EXPERIMENTAL CHARACTERISTICS

Characteristic	Value
Total observations	918
Raw columns	12
Model input features after encoding	15
Training observations	734
Test observations	184
Male observations	725
Female observations	193
Mean age	53.5109 years
Minimum age	28 years
Maximum age	77 years
Missing values	0
Duplicate rows	0
Target positive proportion	55.3377%

The dataset contains 12 raw columns and 918 records, with no missing values and no duplicate rows.

B. Baseline Predictive Performance

TABLE II BASELINE LOGISTIC REGRESSION PERFORMANCE

Metric	Baseline
Accuracy	0.8859
Precision	0.8716
Recall	0.9314
F1-score	0.9005

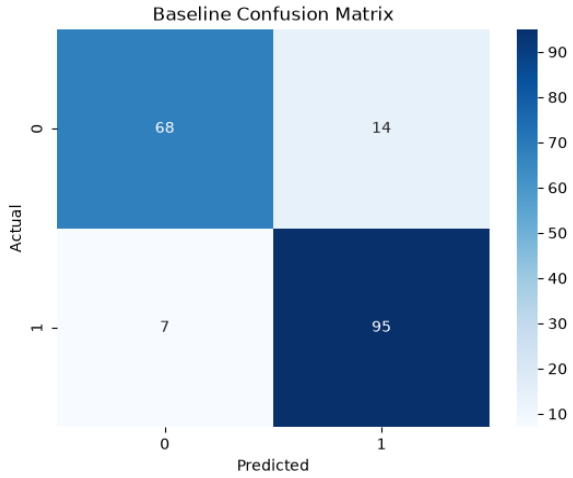

Fig. 3. Confusion matrix of the baseline logistic regression model.

The unmitigated logistic regression model achieved 88.59% accuracy, 87.16% precision, 93.14% recall, and 90.05% F1-score. The held-out test set contained 184 observations, with 82 examples in class 0 and 102 in class 1.

As shown in Fig. 3, the corresponding confusion matrix contains 68 true negatives, 14 false positives, 7 false negatives, and 95 true positives.

C. Baseline Fairness Results

TABLE III BASELINE FAIRNESS METRICS

Sensitive attribute	DPD	EOD	DIR
Sex	0.5144	0.1775	0.2637
Age Group	0.2632	0.1060	0.6184

The Sex audit shows the strongest disparity among the baseline measures. The DPD of 0.5144 indicates a substantial difference in positive prediction rates between the evaluated sex groups, while the DIR of 0.2637 indicates that the group with the lower positive prediction rate received positive predictions at a substantially lower rate than the group with the higher positive prediction rate.

For AgeGroup, the baseline DPD is lower at 0.2632, while the DIR is higher at 0.6184. Thus, the baseline model also exhibits measurable differences between the two AgeGroup groups, although the magnitude is lower than the observed disparity across Sex.

VII. FAIRNESS AUDIT AND BIAS ANALYSIS

A. Sex-Based Disparity

The baseline model exhibited the largest measured demographic disparity across Sex. The DPD of 0.5144 and DIR of 0.2637 show that positive prediction rates differed substantially between the evaluated groups.

The group-level results further indicate that the model's predictive behavior differed across the two sex groups. For the female group, the model achieved an accuracy of 0.9211, precision of 0.7143, recall of 0.8333, and F1-score of 0.7692. For the male group, the corresponding values were accuracy 0.8767, precision 0.8824, recall 0.9375, and F1-score 0.9091.

These results should be described as observed disparities in model outcomes, rather than as proof of intentional discrimination or a causal mechanism. The experiment does not isolate the source of those differences.

B. Age-Group Disparity

The baseline model also exhibits disparity across AgeGroup. AgeGroup was defined using a threshold of 50 years. The corresponding fairness statistics were DPD = 0.2632, EOD = 0.1060, and DIR = 0.6184.

At the group level, the <50 group achieved an accuracy of 0.8971, precision of 0.8276, recall of 0.9231, and F1-score of 0.8727, compared with an accuracy of 0.8793, precision of 0.8875, recall of 0.9342, and F1-score of 0.9103 for the ≥50 group.

The difference in group-level predictive metrics demonstrates that aggregate accuracy alone would not fully represent model behavior across the evaluated demographic categories.

C. Interpretation of the Baseline Audit

The fairness audit establishes that the baseline model is not uniformly distributed in its predictive outcomes across the analyzed demographic attributes. However, the audit alone does not establish the causal source of these disparities. They may reflect differences in the underlying dataset, feature distributions, class prevalence, model behavior, or combinations of those factors.

Accordingly, this study uses the term fairness disparity when describing measured differences and avoids asserting that the model has been proven to discriminate against a particular demographic group.

VIII. BIAS MITIGATION AND PERFORMANCE-FAIRNESS TRADE-OFF

A. Exponentiated Gradient

The Exponentiated Gradient implementation reduced the Sex-based DPD from 0.5144 to 0.0512 and increased the Sex-based DIR from 0.2637 to 0.9188. These changes indicate a substantial reduction in the disparity measured by those two demographic-parity-related metrics.

However, the Sex-based EOD increased from 0.1775 to 0.3800. Thus, the mitigation method did not improve every fairness criterion simultaneously.

Predictive performance also declined:

- Accuracy decreased from 0.8859 to 0.8261.
- Precision decreased from 0.8716 to 0.8070.
- Recall decreased from 0.9314 to 0.9020.
- F1-score decreased from 0.9005 to 0.8519.

Exponentiated Gradient was implemented with the demographic-parity constraint described above, and the resulting predictive and fairness measures are reported in Table V.

B. Threshold Optimizer

Threshold Optimizer reduced the Sex-based DPD from 0.5144 to 0.0865 and increased Sex-based DIR

from 0.2637 to 0.8749. Sex-based EOD, however, increased from 0.1775 to 0.3113.

The predictive metrics after Threshold Optimizer were:

- Accuracy: 0.8152
- Precision: 0.7742
- Recall: 0.9412
- F1-score: 0.8496

The optimizer used the demographic-parity constraint, probability-based prediction, and the prefit baseline estimator. Figs. 4–7 summarize the predictive and fairness comparisons.

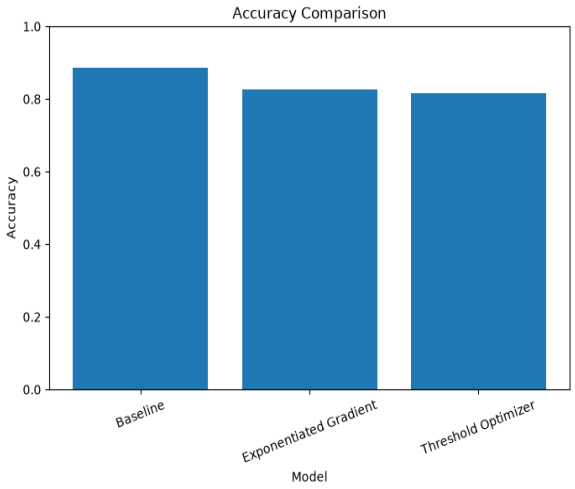

Fig. 4. Accuracy comparison of the baseline and two mitigation approaches.

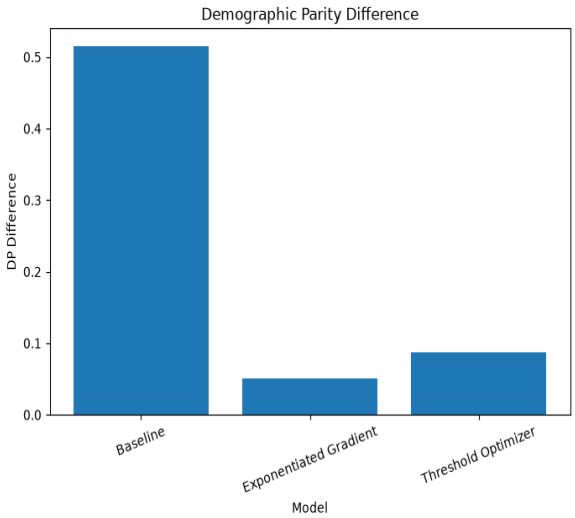

Fig. 5. Sex-based demographic parity difference across the baseline and two mitigation approaches.

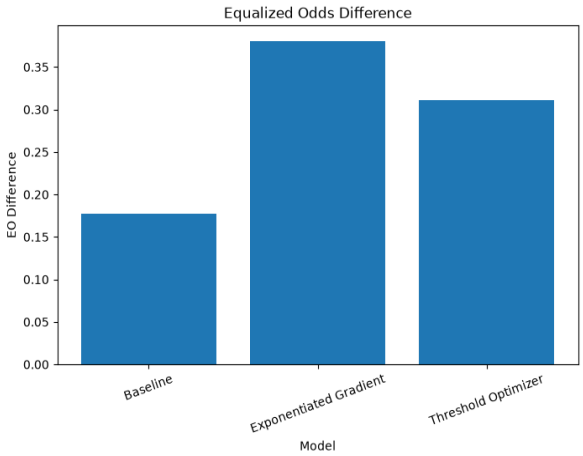

Fig. 6. Sex-based equalized odds difference across the baseline and two mitigation approaches.

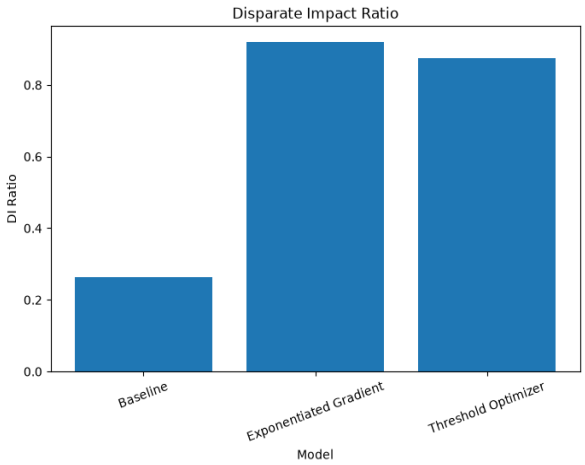

Fig. 7. Sex-based disparate impact ratio across the baseline and two mitigation approaches.

C. Age-Group Fairness After Mitigation

Because the mitigation procedures use Sex as the fairness-sensitive training feature, the AgeGroup results serve as a post-mitigation audit rather than the explicit optimization target.

TABLE IV FAIRNESS METRICS ACROSS AGEGROUP

Model	DPD	EOD	DIR
Baseline	0.2632	0.1060	0.6184
Exponentiated Gradient	0.2130	0.0749	0.6950
Threshold Optimizer	0.1592	0.0273	0.7827

For AgeGroup, both mitigation approaches reduce DPD and EOD and increase DIR. Threshold Optimizer produces the largest improvement across all

three age-related fairness measures. The corresponding values were 0.1592 DPD, 0.0273 EOD, and 0.7827 DIR.

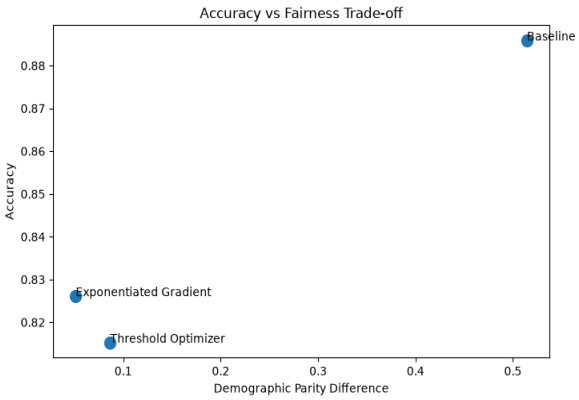

Fig. 8. Accuracy versus Sex-based demographic parity difference for the baseline and mitigation approaches.

D. Overall Performance-Fairness Comparison

The complete comparison is summarized in Table V. The results show that there is no single method that dominates the baseline across every reported metric. Both mitigation approaches reduce demographic-parity disparity across Sex and improve the Sex DIR substantially. However, both also increase Sex EOD and reduce aggregate accuracy and precision.

For AgeGroup, Threshold Optimizer gives the strongest fairness results among the evaluated methods according to DPD, EOD, and DIR. Exponentiated Gradient also improves the age-group measures, although to a lesser extent.

TABLE V PREDICTIVE PERFORMANCE AND FAIRNESS ACROSS BASELINE AND MITIGATION APPROACHES

Approach	Accuracy	Precision	Recall	F1	Sex DPD	Sex EOD	Sex DIR	Age DPD	Age EOD	Age DIR
Baseline	0.8859	0.8716	0.9314	0.9005	0.5144	0.1775	0.2637	0.2632	0.1060	0.6184
Exponentiated Gradient	0.8261	0.8070	0.9020	0.8519	0.0512	0.3800	0.9188	0.2130	0.0749	0.6950
Threshold Optimizer	0.8152	0.7742	0.9412	0.8496	0.0865	0.3113	0.8749	0.1592	0.0273	0.7827

IX. DISCUSSION

The baseline results show a model with relatively strong aggregate predictive performance on the single held-out test set. An accuracy of 88.59%, recall of 93.14%, and F1-score of 90.05% indicate that the baseline classifier performs well in aggregate under the experimental configuration used in this study.

The fairness audit changes the interpretation of that aggregate result. Although the baseline performs well overall, its Sex DPD and Sex DIR indicate a pronounced difference in positive prediction rates across the evaluated groups. The AgeGroup audit also reveals measurable disparity. This demonstrates why reporting only overall accuracy can provide an incomplete description of a healthcare prediction model.

The mitigation experiments show that fairness and predictive performance need to be interpreted jointly. Exponentiated Gradient provides the strongest reduction in Sex DPD and brings Sex DIR close to one, but its Sex EOD increases and its accuracy declines to 82.61%. Threshold Optimizer produces a similarly large reduction in Sex DPD and reaches a Sex DIR of 0.8749 while yielding the strongest improvement in the three AgeGroup fairness measures. Its accuracy, however, is lower at 81.52%.

The different behavior of DPD and EOD is particularly important. The mitigation objective is demographic parity, so improvements in positive prediction-rate parity do not guarantee equalized odds. In this experiment, both mitigation approaches improve demographic-parity-related quantities but worsen Sex EOD relative to the baseline. This is consistent with the broader literature's observation that multiple fairness criteria can conflict [3], [8], [9].

The age-group results provide an additional insight. Even though AgeGroup was not the fairness constraint passed to the mitigation procedures, both methods improve the measured age-related fairness statistics.

Threshold Optimizer produces the strongest age-group reduction among the evaluated methods. This result should not be interpreted as evidence that the method guarantees age fairness in general; it is an empirical result for this dataset and this experiment.

The observed fairness improvements are therefore better understood as metric-specific reductions in measured disparity rather than proof that bias has been eliminated. The experiment also does not establish why the original disparities occur. A larger and more robust investigation would be required to separate data imbalance, feature distributions, model specification, and other potential causes.

X. LIMITATIONS

Several limitations must be considered when interpreting the results.

First, the study uses 918 observations from a single dataset and one 80:20 train-test split. The reported values therefore represent one experimental partition rather than an uncertainty-adjusted estimate of expected performance on unseen populations.

Second, only one primary predictive model is evaluated: logistic regression. The mitigation methods are alternative fairness treatments of that predictive workflow rather than a broad model comparison across multiple algorithm families.

Third, no external validation dataset is used. The findings should therefore not be generalized to other hospitals, populations, geographic regions, or deployment environments.

Fourth, the fairness analysis is limited to Sex and a binary AgeGroup constructed at 50 years. Other demographic characteristics are not assessed.

Fifth, the model input representation retains Sex-derived information as a predictive feature and retains Age while AgeGroup is used for fairness auditing. This means the experiment evaluates fairness of a model that has access to these demographic variables; it does not evaluate a demographic-blind predictor.

Sixth, the mitigation procedures use Sex as the sensitive feature for the fairness constraint, while AgeGroup is evaluated after mitigation. The methods therefore do not constitute a simultaneous optimization against both demographic attributes.

Seventh, the study reports point estimates without confidence intervals or statistical significance tests. The experimental design does not include resampling-based uncertainty analysis, repeated trials, or cross-validation.

Finally, fairness metrics themselves capture different notions of equity and can conflict. No single value reported here should be interpreted as a complete representation of fairness, and no result supports a claim that the model is suitable for direct clinical decision-making.

XI. CONCLUSION

This study examined fairness and bias mitigation in a logistic regression heart disease risk prediction model trained on a 918-record dataset. Following categorical encoding, standardization, and an 80:20 stratified split, the baseline model achieved 88.59% accuracy, 87.16% precision, 93.14% recall, and 90.05% F1-score.

The fairness audit showed substantial disparities, particularly across Sex, where the baseline model produced a DPD of 0.5144 and a DIR of 0.2637. AgeGroup also showed measurable disparity with DPD of 0.2632 and DIR of 0.6184.

Two mitigation methods were then evaluated. Exponentiated Gradient reduced Sex DPD to 0.0512 and increased Sex DIR to 0.9188, while Threshold Optimizer reduced Sex DPD to 0.0865 and increased Sex DIR to 0.8749. Threshold Optimizer produced the strongest age-group fairness results, reducing AgeGroup DPD to 0.1592 and EOD to 0.0273 while increasing DIR to 0.7827.

These improvements came with a reduction in predictive accuracy to 82.61% for Exponentiated Gradient and 81.52% for Threshold Optimizer. In addition, Sex EOD increased for both mitigation approaches, demonstrating that improving one fairness criterion does not guarantee improvement across all fairness dimensions.

The principal contribution of this study is therefore not a claim that demographic bias has been eliminated, but a quantitative demonstration of how fairness auditing can reveal disparities that are hidden by aggregate performance measures and how mitigation can alter both fairness and predictive performance. Further work involving additional validation, broader demographic analysis, uncertainty estimation, and alternative feature specifications would be required before making stronger claims about generalizability or practical deployment.

REFERENCES

[1] M. Padilla Rodriguez and M. Nafea, "Centralized and federated heart disease classification models using UCI dataset and their Shapley-value based interpretability," *arXiv preprint arXiv:2408.06183*, 2024.

[2] A. K. Dubey, "An improved auto categorical PSO with ML for heart disease prediction," Engineering, Technology & Applied Science Research, vol. 12, no. 2, pp. 8567–8573, 2022.

[3] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," ACM Computing Surveys, vol. 54, no. 6, pp. 1–35, 2021.

[4] S. Barocas and A. D. Selbst, "Big data's disparate impact," California Law Review, vol. 104, pp. 671–732, 2016.

[5] S. Shevkari, S. V. Choudhari, and A. N. Tonpe, "Analysis of implementing green accounting (GA) in India—Need of the modern era," International Journal of Scientific Development and Research, vol. 10, no. 11, 2025.

[6] R. J. Chen et al., "Algorithmic fairness in artificial intelligence for medicine and healthcare," Nature Biomedical Engineering, vol. 7, pp. 719–742, 2023.

[7] N. Norori, Q. Hu, F. M. Aellen, F. D. Faraci, and A. Tzovara, "Addressing bias in big data and AI for health care: A call for open science," Patterns, vol. 2, no. 10, Art. no. 100347, 2021.

[8] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in Advances in Neural Information Processing Systems (NeurIPS), 2016, pp. 3315–3323.

[9] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," Big Data, vol. 5, no. 2, pp. 153–163, 2017.

[10] M. Kalyani, M. Manaswini, S. Shevkari, J. Sinkar, L. K. Tyagi, and Y. K. Sharma, "LeafNet: An intelligent system for plant disease classification with deep learning models," in Proc. 2026 13th Int. Conf. Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2026, pp. 1–6, doi: 10.23919/INDIACom70271.2026.11525518.

[11] F. Kamiran and T. Calders, "Classifying without discriminating," in Proc. 2nd Int. Conf. Computer, Control and Communication, 2009, pp. 1–6.

[12] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker, "Fairlearn: A toolkit for assessing and improving fairness in AI," Microsoft, Tech. Rep. MSR-TR-2020-32, 2020.

[13] H. Weerts et al., "Fairlearn: Assessing and improving fairness of AI systems," Journal of Machine Learning Research, vol. 24, no. 257, pp. 1–8, 2023.

[14] fedesoriano, "Heart Failure Prediction Dataset," Kaggle, 2021. [Online]. Available: Kaggle dataset