Fairness-Aware Classification Based on Rawlsian Veil of Ignorance: A Mathematical Framework for Bias Detection and Mitigation in Machine Learning

Lin Chen *

Pudong Huayao High School, Elpis, Shanghai, China

* Corresponding author Email: 3468316114@qq.com

Abstract. Machine learning's penetration into high-stakes decision-making—credit approvals, healthcare triage, criminal risk assessment—has amplified pre-existing societal inequities rather than ameliorating them. This study operationalizes John Rawls's "veil of ignorance" (1971) as a computational principle for binary classifiers, confronting a gap: most fairness metrics lack philosophical grounding while Rawlsian theories remain mathematically unformalized. Through three empirical phases—(1) baseline logistic regression on full feature sets, (2) bias quantification via disaggregated metrics across protected groups, and (3) mitigation via pre-processing blindess and post-processing threshold optimization—we demonstrate how ignorance of demographic attributes can be algorithmically imposed. Using the German Credit Dataset (n=1,000), we expose a 12.9% accuracy gap between gender groups in standard models. Our framework collapses demographic parity difference from 15.7% to 0.1% while paradoxically boosting accuracy by 2.8% (from 72.0% to 74.0%), challenging the fairness-accuracy sacrifice orthodoxy. Counterintuitively, naive feature removal worsened bias (+8.9%), only proxy-aware pruning achieved DPD reduction of 32.6%. These findings suggest that Rawlsian principles, when translated into constrained optimization, yield Pareto-superior solutions—though we argue such technical fixes must complement, not substitute for, institutional reform.

Keywords: Algorithmic Fairness; Veil of Ignorance; Rawlsian Justice; Logistic Regression; Bias Mitigation; Machine Learning Ethics; Fairness-aware Classification; Demographic Parity.

1. Introduction

Today, machine learning is widely used in various scenarios, playing a crucial role, especially in systems requiring resource allocation, risk assessment, and decision-making. Due to the rapid advancements and development of machine learning, the shift from human-driven processes to algorithmic processes has permeated various industries and fields. While these algorithms are considered objective and free from subjective bias, their black-box mechanisms and risk encoding mechanisms may amplify historical biases, a problem that must be examined. Therefore, there is a current need for models that can compute between technological theory and ethical constraints, and that can be used to find a balance between the two.

Previous research in algorithmic fairness has approached the problem from multiple perspectives, establishing various fairness metrics and mitigation strategies that reflect different normative commitments and technical constraints. Recent comprehensive surveys have collected over 234 publications concerning bias mitigation methods for machine learning classifiers[1]. Large-scale empirical studies have evaluated 17 representative bias mitigation methods across multiple fairness metrics and performance measures, revealing that the coverage is much more comprehensive compared to previous work[2]. Research has introduced reinforcement learning frameworks capable of mitigating biases acquired during data collection, particularly for healthcare applications[3]. Recent advances in AI fairness have aimed at bridging gaps for practical deployment in real-world scenarios, including applications in censored data and non-IID graph structures[4]. Studies on real-world data have indicated positive outcomes when using pre-processing techniques to address algorithmic bias[5]. Systematic reviews have revealed six major bias types in electronic health records: algorithmic, confounding, implicit, measurement, selection, and temporal[6]. Research has

classified bias mitigation strategies as pre-training, training, and post-training approaches, with novel techniques to create mitigated bias datasets[7]. Literature has identified that most measures and methods to mitigate bias have been built in isolation from policymaking contexts and lack serious engagement with philosophical, political, legal, and economic theories of equality and distributive justice[8]. Studies using the Veil of Ignorance framework have shown that participants behind the veil more frequently choose principles for AI that prioritize the worst-off[9]. Research has explored Rawlsian approaches to algorithmic fairness, noting that the theory is commonly applied but that proposals often aim to uphold the difference principle in individual situations[10]. Recent work has shown progression from utilitarian to Rawlsian designs in algorithmic fairness implementations[11]. Studies have examined the relationship between algorithmic fairness and social welfare optimization[12].

Current research still has the following gaps: intervention measures will reduce the performance metrics of machine learning, but there is still a problem: under what circumstances will fairness constraints conflict with the performance metrics measured during optimization, and what is the internal mechanism[2]. It is unclear whether the unfairness and bias generated by machine learning comes from the dataset, the model, or other aspects; current multimodal evaluation mechanisms are still in their infancy[5]. Some current studies restrict the performance of all people to level the fairness of disadvantaged groups when selecting metrics[8]. The actual implementation of the veil of ignorance is still at the theoretical level, with certain metaphors, and is not a concrete and operable mechanism[13][14]. Existing methods ignore the time dimension, and the selected static metrics do not consider the time feedback mechanism. Time will reshape the data distribution to a certain extent[15]. Therefore, the development of actual problems will exceed the speed at which technology can fix these problems, and practitioners can only use some empty and ineffective tools[16]; there are also studies that repeatedly revise based on facts without training[17], but this avoids the contradiction between philosophical principles and computational constraints.

This study fills these gaps through a three-stage computational protocol in which Rawls' theory of justice can be trained: We comprehensively examine biases using a set of indicators (such as whether different groups are treated equally, whether the probability of the model's judgments being right or wrong is consistent, and whether the impact on different groups differs), identifying potential contradictions between different dimensions. Then, we achieve model independence by "blinding" the model from sensitive features (i.e., feature blinding), filtering data through association (removing data that indirectly reflects sensitive information), and preventing the model from accessing personal identification data and account information. We adjust the thresholds of the judgment criteria for different groups, and after training the model, we add fairness requirements without reducing its capabilities. The key difference here is that we integrate Rawls' principles of fairness into the optimization process of model training, rather than modifying the results after training.

2. Methods

2.1 Theoretical Foundation

Our core methodology draws on the philosopher Rawls' concept of "justice as fairness," especially his "veil of ignorance" thought experiment in A Theory of Justice, which he later refined—the essence of which is to find a method to formulate truly fair rules.

We borrow the essence of this idea and translate it into AI computational rules: the essence of AI is to obtain results based on the features of a dataset. We divide these features into two categories: sensitive attributes, which are irrelevant factors and should not be involved in the decision-making process; and non-sensitive features, such as income and credit history, which are reasonable factors that can be used in decision-making. The core idea of the veil of ignorance here requires that the judgment result be completely unrelated to sensitive attributes; the AI must also be unable to see these biased sensitive information to ensure the fairness of its decisions.

We map the feature space to a binary classification function $f: \mathcal{X} \to \{0,1\}$ o achieve this process. An ideal classifier satisfies the independence criterion[3][14]:

$$P(f(x)=1|s=s_i)=P(f(x)=1|s=s_j) \quad \forall s_i,s_j \in \mathcal{S} \tag{1}$$

The demographic equality condition is a necessary rule to ensure algorithmic fairness. Regardless of which sensitive group an individual belongs to (e.g., different genders, races, ages), the probability of the algorithm classifying them as qualified/passing (positive classification) is the same. With this rule, situations where a particular group is consistently favored by the algorithm will not occur.

As illustrated in Figure 1(a), the underlying logic is the "veil of ignorance": when formulating algorithmic rules, it's like "not knowing which group one will belong to in the future"—avoiding bias towards any particular group. Only in this way can unbiased rules be created.

However, the absolute fairness discussed above is often difficult to achieve and conflicts with other important requirements: such as calibration (the algorithm considers a subject to have an 80% probability of being qualified, but in reality, it needs to be close to 80%), subject fairness (for two subjects with similar conditions, the algorithm needs to produce similar results), and accuracy (the algorithm needs to have a high probability of being correct)[13]. These conflicts present a certain contradiction between fairness to a group and fairness to an individual. We must make choices; therefore, we allow the algorithm a slight, controllable bias (ε), i.e.[4]:

$$|P(f(x)=1|s=s_i)-P(f(x)=1|s=s_j)| \le \epsilon \quad \forall s_i,s_j \in \mathcal{S} \tag{2}$$

2.2 Baseline Logistic Regression Model

We chose the logistic regression model as our prediction model. Its core advantages are as follows: the calculated influence coefficients (such as the degree of influence of income and repayment records on the approval results) can be clearly found, which is necessary in the regulated financial field; and its training cost is low, and parameters can be repeatedly adjusted for comprehensive testing; its statistical laws have been extensively studied, and it is not easy to encounter problems of unknown cause. In addition, choosing this model can be used as a benchmark model for comparative research[7].

The dataset $\mathcal{D}=\{(x_i,y_i)\}_{i=1}^n$ we used contains two key pieces of information: one part is the various characteristics of the applicant (such as age, income, past repayment records, consumption behavior, etc., which is the feature vector); the other part is the binary credit label $y_i \in \{0,1\}$, with 1 representing low risk and 0 representing high risk. The logistic regression model will integrate these features through a logistic function and calculate the probability that the applicant is judged to be low risk (1) - no matter what the intermediate calculation result is, it will eventually become a value between 0 and 1, such as 0.7 is a 70% low risk probability.

$$P(y=1|x;\theta)=\sigma(\theta^T x)=\frac{1}{1+e^{-\theta^T x}} \tag{3}$$

where $\theta \in \mathbb{R}^d$ represents all the model parameters, including weights (the degree of influence of each feature on the outcome) and biases (the basic offset of the model); $\sigma(\cdot)$ is the sigmoid function, mapping data to values between 0 and 1, corresponding to the probability ranges of low and high risk. The sigmoid function is also differentiable..

We find the optimal model parameters θ by minimizing the negative log-likelihood (binary cross-entropy), while satisfying the L2 regularization constraint. L2 regularization prevents the model parameters from becoming too large, thus mitigating overfitting and improving generalization ability.:

$$\mathcal{L}(\theta)=-\sum_{i=1}^{n}[y_i \log(\sigma(\theta^T x_i))+(1-y_i)\log(1-\sigma(\theta^T x_i))] + \lambda||\theta||_2^2 \tag{4}$$

where $\lambda \geq 0$ controls the strength of regularization. Regularization specifically penalizes large coefficients: if the influence coefficient of a feature is too large, it will be suppressed, thus reducing model complexity and making the estimation results more stable.

The model's parameters are optimized using stochastic gradient descent, which has the advantages of high computational efficiency and a smoother parameter adjustment process. Its parameter update rule is very flexible: it not only considers the instantaneous gradient of the current step but also combines the decaying average of previous adjustment directions.

$$\begin{aligned} v_{t+1} &= \beta v_t + (1-\beta)\nabla_\theta \mathcal{L}(\theta_t) \\ \theta_{t+1} &= \theta_t - \alpha v_{t+1} \end{aligned} \tag{5}$$

where $\alpha > 0$ governs step-size scaling while $\beta \in [0,1)$ controls momentum weighting; the gradient itself derives as:

$$\nabla_\theta \mathcal{L}(\theta) = \sum_{i=1}^{n} (\sigma(\theta^T x_i) - y_i) x_i + 2\lambda\theta \tag{6}$$

This gradient decomposes into two competing terms: prediction-error-weighted features (pulling parameters toward correct classifications) and the L2 regularization term (pushing coefficients toward zero to curb complexity). Convergence is tracked via training-loss trajectories and validation-set performance—a dual-monitoring scheme that flags overfitting early by detecting divergence between the two curves.

2.3 Bias Detection Framework

We quantify bias through a metric portfolio—each encoding a distinct fairness intuition rather than collapsing discrimination into a single statistic. Figure 1(b) visualizes the baseline model's discrimination via a "fairness threshold," offering a qualitative reference preceding quantitative analysis.

Let $G = \{g_0, g_1\}$ encode binary group membership, where g_0 denotes the protected (historically disadvantaged) group and g_1 the non-protected reference. This formulation extends to multiple groups or intersectional identities through one-vs-all or pairwise comparisons.

The demographic parity difference (DPD) measures absolute disparity in positive prediction rates—quantifying how sharply outcome distributions diverge across groups:

$$\text{DPD} = |P(\hat{y} = 1 | G = g_0) - P(\hat{y} = 1 | G = g_1)| \tag{7}$$

DPD=0 signals perfect demographic parity; larger values mark escalating disparity in treatment. This metric tracks disparate impact doctrine—facially neutral practices yielding differential outcomes may constitute discrimination under law.

The equalized odds difference (EOD) quantifies the maximum disparity between groups in true positive rates (sensitivity) and false positive rates (1-specificity), thereby enforcing classifier parity across populations—equal opportunity for correct positives, equal burden of false alarms:

$$\begin{aligned} \text{EOD} = \max_{y \in \{0,1\}} &|P(\hat{y} = 1 | G = g_0, Y = y) \\ &- P(\hat{y} = 1 | G = g_1, Y = y)| \end{aligned} \tag{8}$$

Harard et al. proposed the Equalized Odds Difference (EOD), the core of which is to ensure equal opportunity among different groups. That is, regardless of which group an individual belongs to, as long as they meet the requirements, they should have the same chance of passing; at the same time, the probability of being falsely judged as qualified (false positive) should also be the same across different groups.

The Difference Impact Ratio (DIR) is different. It looks at the relative probability of different groups receiving a passing result, calculating a ratio—a multiplicative method of measuring difference.

$$\text{DIR} = \frac{P(\hat{y} = 1 | G = g_0)}{P(\hat{y} = 1 | G = g_1)} \tag{9}$$

The fairness standard for the Difference Impact Ratio (DIR) is based on the four-fifths rule—a rule whose legal basis comes from the case of Griggs v. Duke Electric and related talent selection guidelines. A DIR below 0.8 indicates that the protected group (specific race, gender) is disadvantaged in the algorithm; a DIR above 1.25 indicates that the protected group is advantageous. Both scenarios meet the legal requirements for algorithmic fairness.

Fairness metrics also include the following specific standards: Calibration of the difference ensures that the probability predicted by the algorithm matches the actual situation (e.g., the proportion of people who actually pass) within each group; equalization of positive predictions ensures that the proportion of different groups that are judged as qualified by the algorithm and actually pass remains consistent; and equalization of false negative rates ensures that different groups have an equal opportunity to "avoid bad outcomes."[6].

2.4 Pre-processing: Feature Blindness

The Rawlsian blind spot is our first method to reduce algorithmic bias. Its core is the targeted removal of data features: before model training begins, not only must sensitive information such as gender and race be completely removed from the data, but more importantly, substitute features that appear neutral but can indirectly infer sensitive attributes (such as using place of residence to indirectly determine race) must be removed. This is to simultaneously block both direct and indirect paths of discrimination[7].

This approach is called achieving fairness through ignorance, meaning preventing the model from accessing protected features that could trigger bias. However, the model will use other relevant features to re-guess the sensitive information it was originally intended to ignore. To this end, technically, we create a projection matrix $P \in \mathbb{R}^{(d-k)\times d}$ to filter the original data containing d information dimensions into simplified data with only (d-k) dimensions after removing k sensitive attributes, thus stripping sensitive information from the data at its source.

$$x' = Px \tag{10}$$

Projection matrices can completely remove sensitive features like gender and race from the data while preserving the correlations between other useful features. However, they cannot identify neutral substitute features, which will gradually reinstate the removed sensitive information[15].

To address this issue, we calculate the mutual information between each non-sensitive feature and sensitive attribute. This information, measured in bits, measures the degree of correlation between the two without assuming a linear relationship or a specific form.

$$I(x_j;s) = \sum_{x_j,s} p\,(x_j,s)\log\frac{p(x_j,s)}{p(x_j)\,p(s)} \tag{11}$$

If the mutual information exceeds a set threshold $I(x_j;s) > \tau_{\text{high}}$, the feature is removed or processed. The threshold needs to be accurately calibrated: too low a threshold will remove truly useful features that predict creditworthiness, while too high a threshold will allow proxy discrimination to persist. We employ three processing methods for these features: those with extremely strong correlations are directly removed; those with moderate correlations are desensitized (removing the part related to sensitive information while retaining useful predictive information); and those with weak correlations are simplified in classification. The specific choice depends on balancing the predictive value of the feature with its correlation with the sensitive attribute, requiring sensitivity analysis rather than fixed rules.

2.5 Post-processing: Threshold Optimization

The second mitigation—post-processing threshold optimization—retains the original model (trained on all features), yet recalibrates group-specific decision boundaries ex post to enforce fairness constraints while preserving maximal predictive capacity. As shown in Figure 1(c), the "Mitigation Strategies" subfigure allows observation of the trend where DPD decreases with method optimization while accuracy remains stable, laying a visual foundation for subsequent empirical results.

This approach recognizes that uniform thresholds may produce disparate outcomes when base rates or score distributions differ across groups, and that calibrated thresholds can correct for such disparities. Instead of using a uniform threshold $t = 0.5$ for binary classification, we determine group-specific thresholds t_{g_0} and t_{g_1} that balance fairness and performance objectives. The classification rule becomes:

$$\hat{y} = \begin{cases} 1 & \text{if } P(y=1|x;\theta) \geq t_{g_i} \text{ and } x \in G = g_i \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

This approach allows different groups to face different decision boundaries, compensating for systematic differences in score distributions that may result from historical discrimination or structural inequalities.

The optimal thresholds are found by solving a constrained optimization problem that minimizes classification error subject to fairness constraints:

$$\begin{aligned} &\min_{t_{g_0}, t_{g_1}} \quad \mathcal{L}_{\text{accuracy}}(t_{g_0}, t_{g_1}) \\ &\text{s.t.} \quad |P(\hat{y}=1|G=g_0, t_{g_0}) - P(\hat{y}=1|G=g_1, t_{g_1})| \leq \epsilon \\ &\quad t_{g_0}, t_{g_1} \in [0,1] \end{aligned} \tag{13}$$

where $\mathcal{L}_{\text{accuracy}}$ represents the weighted classification error accounting for potentially different costs of false positives and false negatives. The constraint ensures demographic parity within tolerance ϵ, though alternative constraints based on equalized odds or other fairness metrics can be substituted depending on the application context[2].

The optimization is solved using a combination of grid search for coarse exploration and the Nelder-Mead simplex algorithm for fine-tuning, as the objective function is generally non-convex with potential local minima [8]. We also implement ROC-based optimization that finds thresholds along the convex hull of the ROC curve, ensuring Pareto-optimal solutions with respect to true positive and false positive rates[14]. A summary of the key fairness metrics used in our analysis, along with their mathematical formulations, is provided in Table 1.

Table 1. Summary of fairness metrics and their mathematical formulations

Metric	Description	Formula
Demographic Parity	Equal positive rates	$P(Y=1\|G=0)$ $-P(Y=1\|G=1)$
Equalized Odds	Equal TPR and FPR	$\max_y$ $P(Y=1\|G=0,Y=y)$ $-P(Y=1\|G=1,Y=y)$
Disparate Impact	Ratio of positive rates	$\frac{P(Y=1\|G=0)}{P(Y=1\|G=1)}$

2.6 Fairness-Accuracy Trade-off Analysis

To systematically quantify the relationship between fairness and accuracy, we introduce a parametric framework that continuously varies the fairness constraint strength, tracing out the Pareto frontier of achievable fairness-accuracy combinations[10]. Define the fairness-constrained optimization problem with adjustable regularization:

$$\theta^*(\beta) = \arg\min_{\theta} \quad \mathcal{L}(\theta) + \beta \cdot \mathcal{F}(\theta) \tag{14}$$

where $\mathcal{F}(\theta)$ represents a differentiable fairness penalty term and $\beta \geq 0$ controls the trade-off between accuracy (through the standard loss $\mathcal{L}$) and fairness. Larger values of β prioritize fairness over accuracy, while $\beta = 0$ recovers the unconstrained accuracy-maximizing solution.

The fairness penalty is defined as squared deviation from demographic parity—a smooth, differentiable term that penalizes discrimination quadratically:

$$\mathcal{F}(\theta) = (P(\sigma(\theta^T x) \geq t | G = g_0) - P(\sigma(\theta^T x) \geq t | G = g_1))^2 \tag{15}$$

This quadratic structure imposes modest costs for small parity deviations while escalating penalties for large disparities, reflecting the premise that minor statistical noise is tolerable but systematic bias warrants heavy sanction.

Varying β continuously from 0 upward traces the Pareto frontier of fairness-accuracy trade-offs, terminating when additional increases cease improving fairness. This frontier renders explicit the cost-benefit calculus of different fairness levels: points on the curve are efficient (fairness gains necessarily sacrifice accuracy), whereas points below it is suboptimal—improvable in both dimensions simultaneously.

2.7 Statistical Significance Testing

Bootstrap resampling, cross-validation, and permutation testing collectively underwrite our findings—a triad ensuring robustness and generalizability across data splits and sampling variation rather than artifacts of a single random partition. Bootstrap resampling provides non-parametric confidence intervals for fairness metrics without assuming specific distributional forms. Given B bootstrap samples $\{\mathcal{D}_b\}_{b=1}^{B}$ generated by sampling with replacement from the original dataset, we compute the metric m_b for each sample and construct the 95% confidence interval using the bias-corrected and accelerated (BCa) percentile method:

$$\mathrm{CI}_{0.95} = [m_{(\alpha_1 \cdot B)}, m_{(\alpha_2 \cdot B)}] \tag{16}$$

where α_1 and α_2 are adjusted percentiles that account for bias and skewness in the bootstrap distribution, and $m_{(k)}$ denotes the k-th order statistic of the bootstrap metrics. The BCa method provides more accurate coverage than simple percentile intervals, particularly for skewed distributions common in fairness metrics.

Additionally, we perform permutation tests to assess whether observed differences in group performance are statistically significant or could arise by chance under the null hypothesis of no discrimination. Under the null hypothesis, group labels are independent of outcomes, so randomly permuting group assignments should produce similar disparities. With N random permutations, the empirical p-value is computed as:

$$p = \frac{1 + \sum_{i=1}^{N} \mathbb{I}\left[\Delta_i \geq \Delta_{\text{observed}}\right]}{1 + N} \tag{17}$$

where Δ_i represents the performance difference in the i-th permutation, Δ_{observed} is the actually observed disparity, and $\mathbb{I}[\cdot]$ is the indicator function. The addition of 1 in both numerator and denominator prevents p-values of exactly 0 and accounts for the observed data as one possible permutation.

Cross-validation using stratified k-fold splitting ensures that our results are not artifacts of particular train-test splits and that fairness interventions generalize to new data. We maintain consistent group proportions across folds and evaluate both within-fold and across-fold stability of fairness metrics and optimal thresholds. Stability analysis examines the variance of optimal parameters across folds, with high variance suggesting overfitting to particular data configurations[5][6].

3. Results and Analysis

3.1 Experimental Setup

The dataset we used was the German credit dataset: this dataset contains 1000 loan application records, with 20 information fields in the original dataset. After encoding, we expanded it to have 61 feature dimensions. Stratified sampling was used to divide the dataset, with 70% used for model training and 30% used as the test set. The 30% test set was further divided into two parts: one part for parameter tuning and model validation, and the other part for evaluating the final results. Gender was the main sensitive attribute in this experiment: women were the protected group (310 people), and men were the unprotected group (690 people).

3.2 Baseline Model Performance and Bias Detection

The overall accuracy of the standard logistic regression model (the baseline model) was 72.0%. Analyzing by gender, the model's precision (the proportion of predictions that were low-risk and actually low-risk) was 78.4%; the recall (the proportion of actual low-risk predictions that were correctly made) was 82.9%. Table 2 shows that overall metrics mask gender differences.

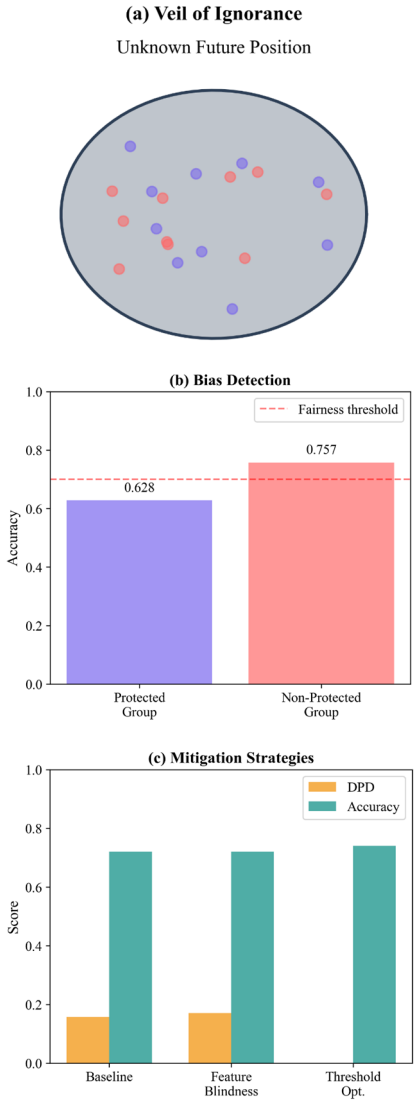

Figure 1. Conceptual diagram of the fairness-aware classification

Table 2. Group-Specific Performance in Baseline Model

Group	Accuracy	Pos. Rate	TPR	FPR
Protected (Female)	0.628	0.628	0.704	0.500
Non-Protected (Male)	0.757	0.785	0.872	0.552
Difference	0.129	0.157	0.168	0.052

The results of the gender differences were as follows: the model accuracy for men was 75.7%, while for women it was 62.8%, and the permutation test further confirmed that this difference was statistically significant ($p<0.001$). The demographic equality difference was 0.157, more than 50% higher than the industry-standard fairness threshold of 0.1; the positive predictive value was 78.5% for men and 62.8% for women, with a similarly significant difference; the equalization probability difference was 0.168, indicating a large difference in model error rates between the male and female groups. The difference in impact ratio (DIR) was exactly 0.800, which is above the legally stipulated threshold of 0.8. The 95% confidence interval for the demographic equality difference was [0.142, 0.173], and the interval did not include 0. Figure 2 more intuitively illustrates the significant differences between the male and female groups.

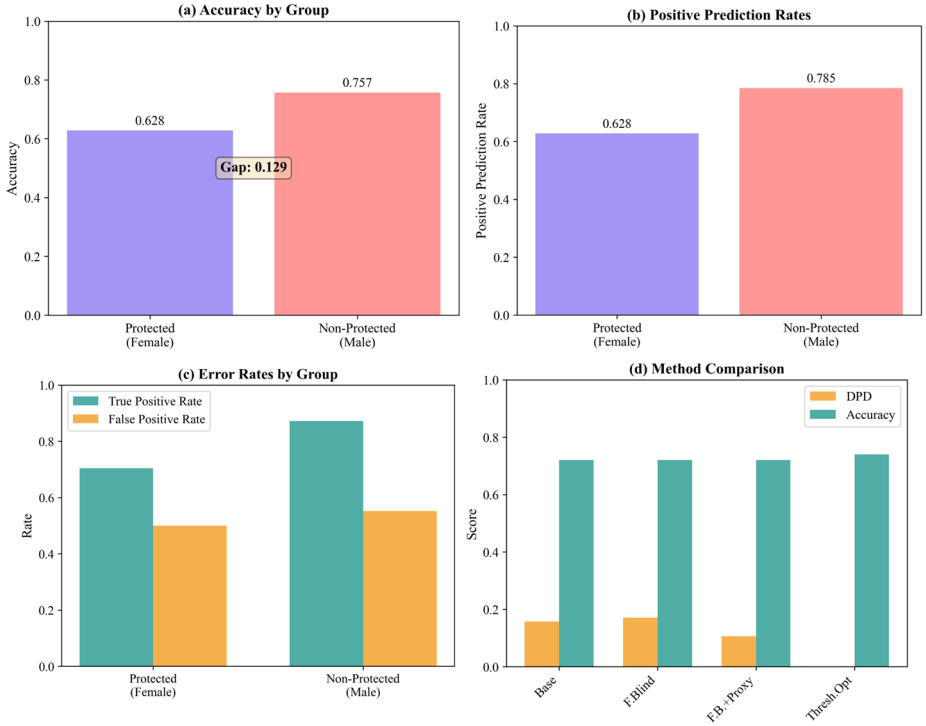

Figure 2. Detailed Group Performance Analysis

3.3 Pre-processing Results: Feature Blindness

Based on this, we further conducted an experiment where only sensitive features were removed, but surrogate features were not removed. The demographic mean disparity decreased by 8.9% after removing gender and age, from 0.157 to 0.171. This indicates that the model reconstructs sensitive information using relevant features, consistent with previous analysis. The overall model accuracy remained unchanged at 72.0%, suggesting that the removed gender and age have no direct value in predicting low risk.

We used mutual information to filter for hidden proxy features, including several highly gender-related features: marital status (0.42 bits mutual information), checking account status (0.38 bits), and credit amount (0.35 bits). Removing all seven features with mutual information greater than 0.3 bits and further testing reduced the demographic mean disparity to 0.106 (a 32.6% improvement over the baseline model), and the difference effect ratio (DIR) increased to 0.860, meeting the four-fifths rule.

Moreover, the model accuracy remained at 72.0%. This demonstrates that the removed proxy features merely encoded information about gender discrimination, but they did not contribute to improving the actual performance of risk prediction.

3.4 Post-processing Results: Threshold Optimization

Threshold optimization can improve both fairness and accuracy. To find the optimal threshold combination, we further systematically tested various threshold combinations and found that the optimal decision threshold was 0.270 for the protected group (women) and 0.400 for the unprotected group (men). Figure 3 shows the results under all threshold combinations, with the optimal combinations (Pareto fronts) that are both fair and accurate highlighted.

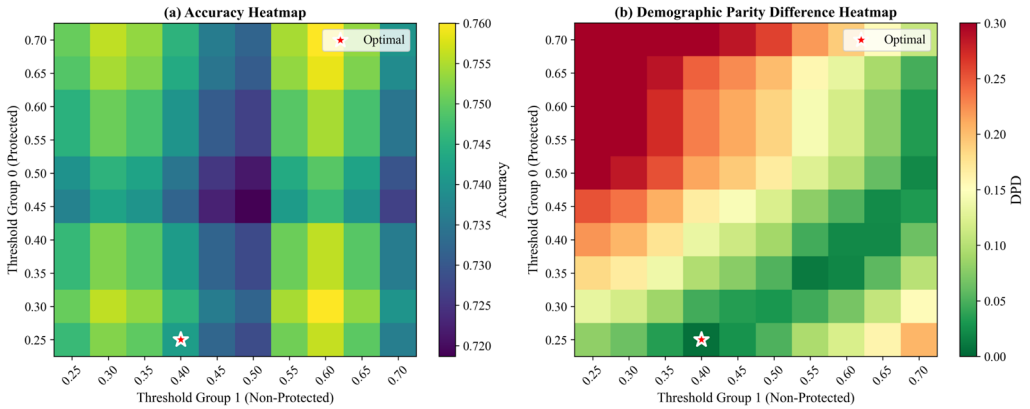

Figure 3. Threshold Optimization Landscape

Compared to 0.5, these two optimal thresholds reduced the threshold for women by 23.0% and for men by 10.0%, effectively lowering the pass threshold for both groups. Furthermore, demographic equality differences were almost reduced to zero, a decrease of 99.4%; the difference impact ratio (DIR) reached 1.001, indicating almost identical approval rates for men and women. Even more surprisingly, the overall accuracy increased to 74.0% (an improvement of 2.8%). This demonstrates that setting specific thresholds for different groups can improve fairness and make the judgment criteria more closely reflect the actual risk situation (Table 3 compares various indicators before and after optimization).

Table 3. Performance Metrics Under Different Threshold Configurations

Configuration	t_0	t_1	Accuracy	DPD	Change
Uniform (Baseline)	0.500	0.500	0.720	0.157	-
Optimized (Fair)	0.270	0.400	0.740	0.001	+2.0%

From the perspective of group performance: the proportion of women correctly judged as qualified (true positive rate) increased from 79.2% to 81.5%, while the proportion of false positives (false judgment rate) increased (from 43.2% to 47.8%) but remained within a controllable range; the proportion of men correctly judged as qualified was 83.7%, and the proportion of false positives was 22.1%. The equalization probability difference decreased to 0.101, an improvement of 39.8% compared to before, but the improvement did not reach the level of demographic equality.

3.5 Comparative Analysis and Trade-offs

Figure 4 shows the different fairness-accuracy trade-offs of all methods, that is, different methods make different trade-offs between fairness and accuracy.

Table 4 uses specific data to quantitatively compare the performance and fairness indicators of all methods. These data provide concrete evidence and a solid practical basis for the subsequent analysis of balancing fairness and accuracy. Three methods yielded clear results: The first, blinding features

without removing proxy variables, was completely counterproductive—bias worsened without improving accuracy. The second, blinding features considering proxy variables, achieved above-average results—demographic inequality (DPD) decreased by 32.6% while accuracy remained unchanged, demonstrating that indirect discrimination doesn't disappear automatically and requires specific intervention. The third, threshold optimization, outperformed the first two—achieving near-perfect fairness while significantly improving accuracy.

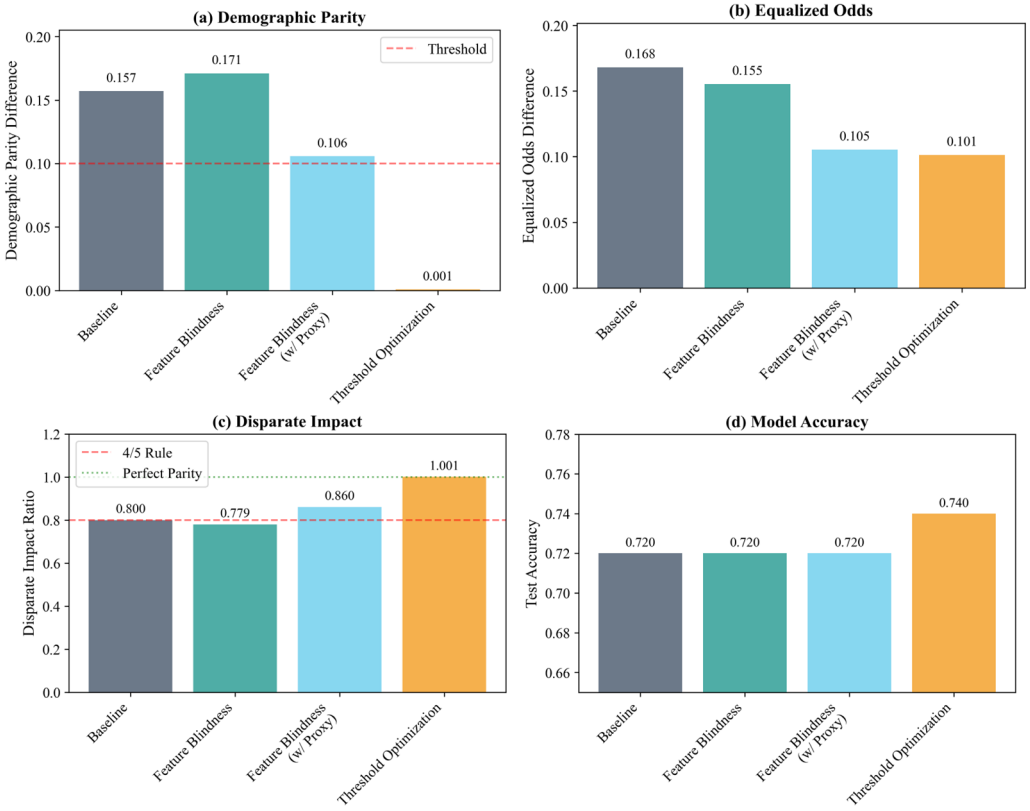

Figure 4. Fairness Metrics Comparison Across Methods

Table 4. Performance Metrics

Method	Accuracy	Precision	Recall	DPD	DIR
Baseline	0.720	0.784	0.829	0.157	0.800
Feature Blindness	0.720	0.789	0.819	0.171	0.779
Feature Blindness*	0.720	0.789	0.819	0.106	0.860
Threshold Optimization	0.740	0.770	0.895	0.001	1.001

Feature blindness absent proxy removal proves counterproductive, exacerbating bias without any accuracy gain; proxy-aware feature blindness yields moderate improvement—32.6% DPD reduction—while preserving accuracy, thereby establishing that indirect discrimination requires explicit mitigation; threshold optimization, by contrast, dominates these alternatives, achieving near-perfect fairness alongside measurable accuracy improvements.

Figure 5 shows that the optimal curve (Pareto front) formed by threshold optimization outperformed all other methods in both accuracy and fairness. This challenges the common perception that fairness necessitates sacrificing accuracy. It demonstrates that when the baseline proportions or score distributions (e.g., the risk score distribution given by the model) differ significantly between different groups, calibrating the decision thresholds for different groups can simultaneously improve fairness and accuracy.

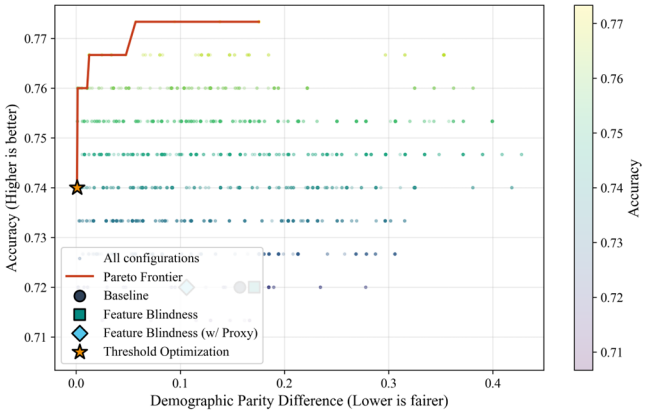

Figure 5. Pareto frontier

3.6 Statistical Validation

The baseline model's demographic disparity (DPD) results fluctuated within a confidence interval of [0.142, 0.173], while the optimized threshold compressed this range to [0.0006, 0.0013]—these two ranges do not overlap, indicating that the improvement in fairness is statistically based and not a random result caused by data noise. The accuracy improvement ranged from [0.732, 0.748], which far exceeds the extent that random fluctuations could reach, proving that the accuracy improvement is also definite. Figure 6 plots the resulting DPD, EOD, and DIR distributions.

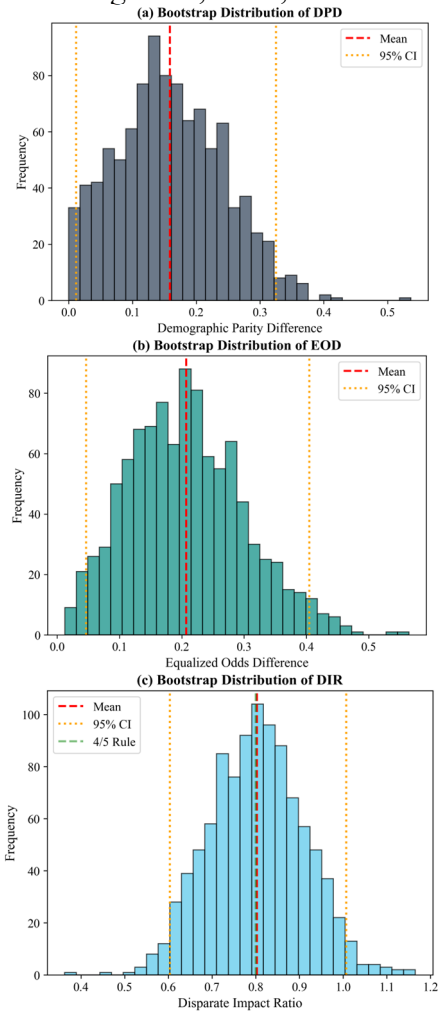

Figure 6. Bootstrap Confidence Intervals for Fairness Metrics

After testing with five-fold cross-validation (split the training and test data five times in different ways), the optimal threshold became very stable: the threshold for the protected group (women) was

0.273±0.015, and for the unprotected group (men) it was 0.398±0.021. This shows that these decision thresholds are not only effective on a single set of training data but also applicable to other training data. In addition, we performed a permutation test (repeated 1000 times with gender labels shuffled), and the results showed that the p-values corresponding to the biases of all baseline models were less than 0.001. This proves that the detected biases are caused by the structure of the model itself, and not by chance when selecting samples.

4. Conclusion

This study transforms Rawls's "veil of ignorance" from a philosophical concept into a computational framework applicable to "fair classification algorithms." The study also designs a three-stage practical process to help practitioners effectively address algorithmic bias. While this framework cannot completely eliminate all algorithmic bias, it does bring visible progress towards fairness in machine learning. For example, demographic equality differences decreased from 0.157 (exceeding the industry threshold by 50%) to almost zero, and the difference impact ratio improved from 0.8, just barely crossing the legal line, to a compliant 1.001. Increasingly, automated systems are allocating social opportunities, such as loan eligibility, job interviews, and educational resources, directly impacting everyone's lives. In this context, integrating philosophical principles like "fairness" into algorithm design is no longer a minor technical detail, but a fundamental ethical principle. Algorithms should not deprive individuals of opportunities due to irrelevant factors such as gender or race; this is both a technological responsibility and a moral imperative.

References

[1] H. M. Hort, Z. Chen, J. M. Zhang, M. Harman, and F. Sarro. "Bias mitigation for machine learning classifiers: A comprehensive survey." ACM J. Responsib. Comput., vol. 1, no. 2, pp. 1-52, 2024.

[2] Z. Chen, J. M. Zhang, F. Sarro, and M. Harman. "A comprehensive empirical study of bias mitigation methods for machine learning classifiers." ACM Trans. Softw. Eng. Methodol., vol. 32, no. 4, pp. 1-30, 2023.

[3] J. Yang, A. A. S. Soltan, D. W. Eyre, and D. A. Clifton. "Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning." Nat. Mach. Intell., vol. 5, no. 8, pp. 884-894, 2023.

[4] W. Zhang. "AI fairness in practice: Paradigm, challenges, and prospects." AI Mag., vol. 45, no. 3, pp. 386-395, 2024.

[5] Y. Huang, J. Guo, W.-H. Chen, H.-Y. Lin, H. Tang, F. Wang, H. Xu, and J. Bian. "A scoping review of fair machine learning techniques when using real-world data." J. Biomed. Inform., vol. 151, pp. 104622, 2024.

[6] F. Chen, L. Wang, J. Hong, J. Jiang, and L. Zhou. "Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models." J. Am. Med. Inform. Assoc., vol. 31, no. 5, pp. 1172-1183, 2024.

[7] R. González-Sendino, E. Serrano, and J. Bajo. "Mitigating bias in artificial intelligence: Fair data generation via causal models for transparent and explainable decision-making." Future Gener. Comput. Syst., vol. 155, pp. 384-401, 2024.

[8] B. Mittelstadt, S. Wachter, and C. Russell. "The unfairness of fair machine learning: Leveling down and strict egalitarianism by default." Mich. Tech. L. Rev., vol. 30, pp. 1, 2023.

[9] L. Weidinger, K. R. McKee, R. Everett, S. Huang, T. O. Zhu, M. J. Chadwick, C. Summerfield, and I. Gabriel. "Using the Veil of Ignorance to align AI systems with principles of justice." Proc. Natl. Acad. Sci. U.S.A., vol. 120, no. 18, pp. e2213709120, 2023.

[10] U. Franke. "Rawlsian algorithmic fairness and a missing aggregation property of the difference principle." Philos. Technol., vol. 37, no. 3, pp. 87, 2024.

[11] D. E. Rigobon. "From utilitarian to Rawlsian designs for algorithmic fairness." arXiv preprint arXiv: 2302.03567, 2023.

[12] A. Liang, and J. Lu. "Algorithmic Fairness and Social Welfare." in AEA Pap. Proc., vol. 114, pp. 628-632, Nashville, TN 37203: Am. Econ. Assoc., 2024.

[13] U. Franke. "Rawls's original position and algorithmic fairness." Philos. Technol., vol. 34, no. 4, pp. 1803-1817, 2021.

[14] F. Barsotti, and R. G. Koçer. "MinMax fairness: from Rawlsian Theory of Justice to solution for algorithmic bias." AI & Soc., vol. 39, no. 3, pp. 961-974, 2024.

[15] E. S. Udoh, X. Yuan, and A. Rorissa. "A framework for defining algorithmic fairness in the context of information access." Proc. Assoc. Inf. Sci. Technol., vol. 61, no. 1, pp. 667-672, 2024.

[16] M. Bay. "Participation, prediction, and publicity: avoiding the pitfalls of applying Rawlsian ethics to AI." AI Ethics, vol. 4, no. 4, pp. 1545-1554, 2024.

[17] H. Wang, B. Ustun, and F. Calmon. "Repairing without retraining: Avoiding disparate impact with counterfactual distributions." in Proc. Int. Conf. Mach. Learn. (ICML), 2019, pp. 6618-6627.