Available Online: 24 February 2026 DOI: 10.54254/2977-3903/2026.31838

An end-to-end fairness framework based on counterfactual reasoning: auditing, diagnosis, and mitigation

Yingchun Xu*, Xiaohang Zhang

Beijing University of Posts and Telecommunications, Beijing, China

*Corresponding Author. Email: xu_yingchun123@bupt.edu.com

Abstract. Artificial intelligence regulations typically require that sensitive attributes (such as gender and race) be excluded from algorithmic decision-making to prevent discrimination, a principle commonly referred to as "fairness through unawareness." However, even when sensitive attributes are removed, algorithmic models may still infer such information through proxy variables that are pervasive in data, often via complex nonlinear relationships, thereby perpetuating or even amplifying systemic bias. To address the problem of indirect discrimination under fairness through unawareness, this paper proposes an end-to-end framework that integrates discrimination auditing, diagnosis, and mitigation. First, by incorporating an advanced Transformer-Based Counterfactual Explainer (TABCF), our framework constructs a more reliable bias auditing system capable of accurately uncovering discriminatory behaviors in models. Second, once bias is detected, we introduce an innovative two-stage NOCCO–Shapley diagnostic method that identifies the key proxy variables responsible for discrimination and reveals how the model actually exploits these variables in practice. Finally, to mitigate the identified bias, we implement an adjustable λ-PCF post-processing strategy that enables a quantifiable trade-off between predictive utility and counterfactual fairness without retraining the model. Notably, we find that when the trade-off parameter λ is set to the prior probability distribution of the sensitive attribute in the dataset, the model achieves an optimal balance between fairness and utility. Extensive experiments on four widely used real-world datasets demonstrate that our end-to-end framework not only outperforms existing methods in auditing and diagnosis, but also provides a practical and effective technical pathway for deploying more responsible and fair AI systems in real-world applications.

Keywords: counterfactual reasoning, discrimination auditing, proxy variables, counterfactual explanation

1. Introduction

With the widespread adoption of machine learning systems in high-stakes domains such as credit approval, recruitment, and judicial decision-making, algorithmic fairness has become a critical social and technical concern [1-3]. A straightforward approach is to achieve fairness through unawareness by removing sensitive attributes (e.g., gender, race) from training data [4, 5]. However, extensive research has shown that this approach is often ineffective or even harmful. Models can easily learn strong correlations with sensitive attributes from other seemingly innocuous features, known as proxy variables, leading to indirect discrimination [6].

Existing bias auditing methods, such as frameworks based on counterfactual explanations [7], can detect indirect discrimination. Yet, their effectiveness heavily depends on the quality of the counterfactual explainer used. Classical counterfactual generators (e.g., DiCE) [7, 8] may exhibit feature-type biases when handling complex tabular data, resulting in unreliable audit outcomes. Moreover, existing work often treats discrimination auditing, diagnosis, and mitigation as isolated steps, lacking a systematic framework that integrates all three [9, 10].

To address these challenges, this paper proposes an end-to-end framework for auditing, diagnosing, and mitigating algorithmic bias, integrating reliable counterfactual generation, a two-stage proxy variable diagnostic method, and a post-processing bias mitigation strategy. Our main contributions can be summarized as follows:

1. Counterfactual-based discrimination auditing: Through experiments, we advocate the use of an advanced Transformer-based generative counterfactual model, TABCF [11]. Compared to traditional methods that optimize within the original feature space (e.g., DiCE), TABCF provides a more reliable and in-depth revelation of model discriminatory behaviors.

2. Two-stage proxy variable diagnosis: After identifying bias, we introduce an innovative two-stage NOCCO–Shapley diagnostic method. This method identifies proxy variables in the data and reveals how the model exploits them in practice.

3. Adjustable bias mitigation: For diagnosed bias, we implement a post-processing mitigation strategy, λ-PCF, that does not require retraining the model. We validate its effectiveness and demonstrate how it enables a quantifiable trade-off between model utility and counterfactual fairness, providing actionable guidance for practical deployment.

4. End-to-end framework: We construct a comprehensive framework that integrates auditing, diagnosis, and mitigation of bias. Extensive experiments across multiple benchmark datasets demonstrate that it provides a systematic and actionable methodology for building, deploying, and governing responsible AI systems in practice.

2. Related work

This section reviews two research areas most relevant to our work: discrimination auditing and detection, and counterfactual explanation generation and bias mitigation.

2.1. Discrimination auditing and detection

Auditing and detecting algorithmic discrimination is a cornerstone of responsible AI. Early studies primarily focused on statistical fairness metrics. Metrics such as statistical parity and equal opportunity [12, 13] quantify group fairness by comparing the statistical differences in model predictions across protected groups (e.g., gender, race). While these metrics provide a macro-level quantification of bias and are easy to compute, their limitations have become increasingly evident: (1) they can only indicate whether bias exists, without revealing its causes or providing individual-level remedial guidance; (2) they cannot distinguish between direct discrimination caused by sensitive attributes and indirect discrimination arising from proxy variables; (3) different metrics often conflict with one another [12].

To overcome these limitations, counterfactual-based auditing methods have received growing attention in recent years. These approaches move beyond group-level statistics and explore individual-level "what-if" scenarios. For example, Cornacchia et al. [4, 9] proposed the CFlips and nDCCF metrics, which quantify

discrimination by evaluating whether an individual's (predicted) sensitive attribute would need to be flipped to achieve a favorable outcome.

However, existing counterfactual-based auditing methods rely on a critical but often overlooked assumption: that the counterfactual explainer itself is fair and unbiased. If the counterfactual generator is flawed or biased—for instance, if it tends to produce unrealistic samples or exploits biases in the data—then audit conclusions based on its outputs can be unreliable or even misleading. We argue that a reliable discrimination auditing framework must be grounded in a high-fidelity counterfactual generator.

2.2. Counterfactual explanation generation

Methods for generating counterfactual explanations can be broadly categorized into two classes. The first class performs optimization directly in the original feature space. A representative example is DiCE [7], which identifies counterfactuals by jointly minimizing the distance to the original instance while maximizing changes in predicted outcomes within a multi-objective optimization framework. These methods are conceptually simple and easy to implement. However, a major limitation is that they often generate unrealistic or out-of-manifold samples. More critically, they frequently exhibit feature-type bias: during optimization, all features—whether continuous, categorical, or ordinal—are treated as numerical, ignoring the fact that changing different types of features in the real world involves very different costs and logical constraints.

To address the issue of plausibility, TABCF [11] proposes optimization in a learned latent space. It innovatively employs a Transformer–Variational Autoencoder (VAE) architecture. The powerful context-aware capabilities of the Transformer encoder enable it to model complex dependencies among different features in tabular data, effectively mitigating the feature-type biases seen in methods like DiCE. TABCF understands, for example, that changing a categorical feature (e.g., occupation) has fundamentally different implications and consequences from adjusting a continuous feature (e.g., hours-per-week). This deep understanding of data structure and semantics allows TABCF to generate counterfactual explanations that are not only realistic and diverse but also more reliable for auditing model bias. Consequently, this study selects TABCF as the core tool for high-fidelity bias auditing in our framework.

2.3. Bias mitigation techniques

Once bias has been detected and diagnosed, the next step is to take measures to mitigate it. A large body of research has proposed bias mitigation techniques, which can be broadly categorized into three types based on their stage of application in the machine learning workflow: pre-processing [13, 14], in-processing [15], and post-processing [16, 17].

Pre-processing methods intervene in the data before model training. They remove or reduce bias in the original data by resampling underrepresented group samples, reweighting the importance of different samples [13], or learning a fair data representation [14]. The advantage of pre-processing methods is that they are model-agnostic. However, a key drawback is that they may distort the original data distribution, potentially impairing model generalization.

In-processing methods incorporate fairness constraints directly into the model's optimization objective during training. For example, a penalty term can be added to the loss function to constrain prediction disparities across groups [15]. These methods often achieve a favorable fairness–utility trade-off, but they are typically tied to specific model architectures and require full retraining, which can be costly in many practical scenarios.

Post-processing methods adjust a trained model's predictions to satisfy specific fairness criteria [16, 17]. Their greatest advantage lies in flexibility and low implementation cost: the model is treated as a "black box,"

requiring no retraining, and the method can be applied directly to any deployed model. This makes post-processing approaches particularly appealing for legacy systems in real-world applications.

Among post-processing methods, approaches based on counterfactual fairness offer a novel and principled perspective. Unlike traditional post-processing techniques that merely adjust prediction thresholds, these methods aim to align model predictions with the definition of counterfactual fairness [18]: a model's prediction should not change if an individual's sensitive attribute were different. A representative method is Plausible Counterfactual Framework (PCF) [19], which uses counterfactual samples to estimate what the model would predict if an individual's sensitive attribute took on a different value. The "counterfactual prediction" is then combined with the "factual prediction" through a weighting scheme based on the prior probability of the sensitive attribute, producing a sensitivity-insensitive "fair prediction."

However, PCF exhibits a significant limitation when enforcing absolute counterfactual fairness—completely eliminating the influence of sensitive attributes on predictions—because this can substantially degrade predictive utility. When a model heavily relies on sensitive information, forcing it to ignore this information inevitably reduces accuracy.

To address this challenge, our work introduces and validates λ-PCF, an adjustable mitigation strategy. By introducing a tunable parameter λ, we can interpolate between the model's original predictions and the fully fair PCF predictions without retraining. This enables practitioners to achieve a quantifiable trade-off between fairness and predictive utility, providing a practical and controllable approach for deploying fair AI systems in real-world settings.

3. Proposed end-to-end fairness framework

To address algorithmic bias under fairness-through-unawareness scenarios, this study constructs and validates an end-to-end framework encompassing discrimination auditing, proxy feature identification, and bias mitigation. The framework aims to answer the following research questions:

RQ1 (Reliable Bias Auditing): How can we build a more robust and trustworthy bias auditing system by integrating high-quality counterfactual generators?

RQ2 (Proxy Feature Diagnosis): After auditing detects bias, how can we accurately identify the proxy variables responsible for discrimination and their mechanisms of influence?

RQ3 (Effective Bias Mitigation): How can we mitigate model unfairness by balancing predictive utility and fairness without retraining the model?

The framework is designed to provide an actionable, interpretable, and quantifiable methodology for ensuring the responsible deployment of AI systems. It consists of three interrelated core components:

1. Counterfactual-based Bias Auditing: Integrates a high-quality counterfactual generator to construct a more robust and reliable bias auditing system.

2. Two-Stage Proxy Variable Diagnosis: Identifies the proxy variables causing discrimination and elucidates their mechanisms after bias is detected.

3. Adjustable Bias Mitigation: Balances predictive utility and fairness to mitigate model unfairness without retraining.

3.1. Counterfactual-based bias auditing

To address RQ1—how to accurately audit model bias—we emphasize that the quality of the auditing tool itself is critical. If the method used to generate counterfactuals is biased, the resulting audit is unreliable and potentially misleading.

Existing counterfactual-based auditing frameworks [6] often employ generic methods such as DiCE [7] to generate counterfactuals. DiCE searches for minimal feature perturbations that flip model predictions directly in the original feature space via gradient-based optimization. However, when handling tabular data with mixed feature types (e.g., numerical and categorical), this approach suffers from a fundamental flaw—feature-type bias. During optimization, continuously modifying a numerical feature (e.g., age) is "easier" in gradient space than discretely changing a categorical feature (e.g., occupation). Consequently, DiCE-generated counterfactual explanations tend to overemphasize numerical features while underrepresenting categorical changes, producing biased and incomplete remedial suggestions.

To overcome this limitation, our framework integrates TABCF [11], an advanced generative counterfactual explainer specifically designed for tabular data. TABCF is built upon a Transformer-based Variational Autoencoder (Transformer-VAE). Mixed-type tabular data are first encoded into a unified, continuous latent space. All subsequent counterfactual search and optimization occur in this low-dimensional, smooth, and structured latent space. This latent-space optimization paradigm provides two key advantages:

1. Plausibility of Generated Samples: By moving smoothly within the latent space that captures the data manifold and decoding back to the original space, TABCF ensures that generated counterfactuals conform to real-world data distributions and logical constraints, avoiding unrealistic samples.

2. Mitigation of Feature-Type Bias: TABCF employs a Gumbel-Softmax decoding technique [20], allowing categorical features to be accurately reconstructed from the latent space while maintaining end-to-end differentiability. This mechanism ensures that categorical and numerical features are treated fairly during optimization, producing more balanced and comprehensive counterfactual explanations.

3.2. Two-stage proxy variable diagnosis

To effectively address RQ2—how to diagnose the root causes of bias—we propose an innovative two-stage NOCCO–Shapley diagnostic method [21]. After confirming bias in the auditing stage, this component delves into the full path of discrimination: from inherent associations in the data to the model's specific exploitation of these associations in decision-making. The core of the method is the use of NOCCO–Shapley values to fairly attribute the contribution of each non-sensitive feature as a proxy for the sensitive attribute.

3.2.1. Nocco–shapley values: quantifying nonlinear dependencies

Our diagnostic method employs NOCCO [21] as a feature value function $v(A)$ to measure the nonlinear dependence between a subset of features X_A and the sensitive attribute S. NOCCO is a normalized variant of the Hilbert–Schmidt Independence Criterion (HSIC), with values ranging from 0 to 1; higher values indicate stronger dependence. For a given feature coalition (A) (corresponding to data matrix X_A) and target vector S, the NOCCO value is computed as:

$$v(A) = NOCCO(X_A, S) = tr(R_{X_A} R_S) \tag{1}$$

Here, $tr(\cdot)$ denotes the matrix trace, and R_{X_A} and R_S are obtained by regularizing and normalizing their respective kernel matrices K_{X_A} and K_S:

$$R_K = (HKH)(HKH + n\epsilon I)^{-1} \tag{2}$$

In Equation (2), K is an $n \times n$ kernel matrix computed using a Radial Basis Function (RBF) kernel, ($H = I - \frac{1}{n} 11^T$) is the centering matrix, n is the number of samples, and ϵ is a small regularization term for numerical stability. By computing $v(A)$ for all 2^m feature coalitions A, we construct a cooperative game value function. Using the standard Shapley value formula, the coalition values (i.e., dependence scores) are fairly attributed to individual features, yielding the NOCCO–Shapley value ϕ_i for each feature x_i. Since

exact computation of Shapley values for all coalitions is exponential, we employ Monte Carlo sampling to efficiently approximate them in practice.

3.2.2. Two-stage diagnostic procedure

Based on NOCCO–Shapley values, our diagnostic workflow proceeds in two stages:

Stage 1: Data-Level Evaluation (ϕ_d). At the data level, we compute the NOCCO–Shapley value $\phi_{d,i}$ of each non-sensitive feature x_i with respect to the sensitive attribute S in the original training dataset. This value quantifies the intrinsic potential of x_i as a proxy variable. Features with higher $\phi_{d,i}$ exhibit stronger statistical association with S and are more likely to be exploited during model training to produce indirect discrimination.

Stage 2: Model-Level Diagnosis ($\Delta\phi$). Data-level association alone cannot confirm whether a feature is actually used by the model. To reveal the model's true behavior, we perform a comparative analysis between factual samples X_r (those receiving negative predictions) and counterfactual samples X_c (ideal samples achieving positive predictions). We compute the NOCCO–Shapley values for each set, X_c and ϕ_c , and define: $\Delta\phi = \phi_c - \phi_r$, captures how the model's reliance on the association between each feature and the sensitive attribute changes when moving from negative to positive predictions:

$\Delta\phi > 0$: In the counterfactual sample set (X_c), the statistical association between feature x_i and sensitive attribute S is stronger than in the factual sample set (X_r). This indicates the model tends to assign higher scores to individuals who more closely align with group stereotypes in the data (e.g., a dominant group). To generate a positive prediction, the model requires a stronger association between feature x_i and the sensitive attribute. Thus, a feature exhibiting high ϕ_d and a high positive $\Delta\phi$ serves as a proxy variable actively leveraged by the model to implement discrimination.

$\Delta\phi < 0$: In the counterfactual sample set (X_c), the statistical association between feature x_i and sensitive attribute S is weaker than in the factual sample set (X_r). This indicates that in the actual samples, the association between feature x_i and the sensitive attribute is a detrimental factor. To generate a positive prediction, the model requires this association to be weakened or eliminated. Therefore, a feature with high ϕ_d and a high negative $\Delta\phi$ is considered by the model to be one of the direct causes leading to discriminatory negative decisions.

By combining data-level evaluation (ϕ_d) with model-level diagnosis ($\Delta\phi$), our two-stage method accurately identifies the proxy variables truly leveraged by the model in its decision-making process.

3.3. Adjustable bias mitigation

After auditing and diagnosing specific bias patterns, we address RQ3—how to mitigate bias while balancing fairness and predictive utility—using an adjustable post-processing mitigation strategy: λ-PCF. As a post-processing method, its key advantage is that it does not require retraining, allowing fairness corrections to be applied to deployed systems.

The core idea of λ-PCF is to interpolate between the model's original predictions and theoretically fully fair predictions. The final corrected prediction probability $\hat{y}_{final}$ is defined as:

$$\hat{y}_{final}(\lambda) = \lambda \cdot \hat{y}_{pcf} + (1 - \lambda) \cdot \hat{y}_{orig} \tag{3}$$

where $\hat{y}_{orig}$ is the original model output, $\hat{y}_{pcf}$ is the "fair" prediction derived from the PCF framework [19], and $\lambda \in [0,1]$ is a tunable trade-off parameter. Adjusting λ systematically explores the fairness–utility Pareto frontier between maximum original performance ($\lambda = 0$) and maximum theoretical fairness ($\lambda = 1$).

However, merely providing an adjustable λ is insufficient. To address this, we propose and adopt a λ selection strategy based on data prior. Inspired by Bayesian reasoning, we propose setting λ to the prior

probability of the dominant group in the dataset: $\lambda = P(S = s_{priv})$. The intuition behind this choice is that the degree to which the final prediction should "trust" theoretically fair PCF predictions depends on the inherent imbalance of the data itself. When the dataset is highly imbalanced, a larger λ value assigns greater weight to fair predictions, enabling stronger correction; conversely, it applies milder correction.

Thus, our final mitigation strategy is: first explore the entire fairness-utility tradeoff space by varying λ, then specifically evaluate the performance point when $\lambda = P(S = s_{priv})$. This particular λ value provides a non-arbitrary, data-driven default option for achieving a "reasonable" balance between fairness and utility, offering concrete, theory-grounded guidance for making quantifiable tradeoff decisions in practice.

4. Experimental setup

This chapter presents a comprehensive series of experiments designed to systematically evaluate the effectiveness of our proposed end-to-end framework in auditing, diagnosing, and mitigating algorithmic bias. The experimental design directly addresses the three core research questions (RQ1–RQ3) outlined in this study.

4.1. Datasets

We conducted experiments on five widely used benchmark datasets spanning different domains to evaluate the effectiveness and generalizability of our framework. All datasets contain potential social biases.

Adult: A classic classification dataset derived from the U.S. Census [22]. The task is to predict whether an individual's annual income exceeds $50,000. The sensitive attribute is sex. Features include demographic information such as age, education-num, race, and hours-per-week, making it a standard benchmark for evaluating algorithmic fairness.

COMPAS: Released by ProPublica [23], this dataset predicts the risk of recidivism within two years for criminal defendants, representing a high-stakes fairness-critical scenario. The sensitive attribute is race. Features include criminal history (priors_count), age, and charge severity (c_charge_degree).

Bank Marketing: Originating from a Portuguese bank's telemarketing campaign dataset [24], this dataset predicts client subscription outcomes. We treat age as the sensitive attribute to study potential age-related discrimination. Features include demographic information, historical campaign data (campaign, previous), and macroeconomic indicators.

Customer: Sourced from Kaggle [25], this large-scale e-commerce dataset contains 500,000 customer transaction records. The task is binary classification: predicting whether a purchase is successful (PurchaseStatus = 1). For fairness analysis, gender is treated as the sensitive attribute. Features cover demographics (age), transactional data (AnnualIncome, NumberOfPurchases), and behavioral metrics (TimeSpentOnWebsite, CustomerSatisfaction), enabling fairness studies in a high-dimensional real-world commercial setting.

4.2. Model configuration

To ensure consistency and reproducibility, we adopted a unified architecture for the two key models in our framework: Decision Model ($f(\cdot)$): The auditing and mitigation processes are centered around a Multilayer Perceptron (MLP) classifier, which takes all non-sensitive features as input and predicts a binary target variable. We chose MLPs as they are general-purpose nonlinear function approximators capable of capturing complex patterns, reflecting common model types in modern machine learning systems.

Sensitive Attribute Classifier ($fs(\cdot)$): In auditing and diagnostic procedures, this "oracle" model predicts sensitive attributes (e.g., gender, race) from all non-sensitive features. We again use an MLP with similar architecture to the decision model. By using consistent MLP architectures, we avoid introducing additional variability due to heterogeneous models, allowing a cleaner analysis of the data and decision model behavior. Both MLPs use identical hidden layer dimensions and activation functions, trained with the Adam optimizer.

4.3. Evaluation metrics

To comprehensively evaluate each component of our framework, we adopt the following metrics:

Counterfactual Fairness Auditing Metrics:

Counterfactual Flip Rate (CFlips): Measures the frequency with which a counterfactual explanation requires an individual's predicted identity to flip in order to achieve a favorable outcome. $CFlips_u$ and $CFlips_p$ represent the flip rates for the unprivileged and privileged groups, respectively, with $\Delta CFlips$ serving as a key bias indicator.

Normalized Discounted Cumulative Counterfactual Fairness (nDCCF): A composite score ranging from [-1, +1] that considers identity flips while rewarding explanations that are closer to the original sample (i.e., more feasible) and maintain identity, providing a more holistic bias measure.

Bias Mitigation Evaluation Metrics:

Total Effect (TE): Core measure of fairness, defined as TE = $E[|\hat{y}(x,s) - \hat{y}(x,s\prime)|]$, quantifying the sensitivity of model predictions to changes in sensitive attributes. Lower values indicate higher fairness.

Accuracy: Measures model utility by evaluating predictive performance.

4.4. Counterfactual generators

To address RQ1—how high-fidelity counterfactuals impact auditing—we designed two experimental setups for comparison:

Our Method (TABCF-based): We employ the full TABCF framework. The Transformer-VAE model is trained for 4,000 epochs using the recommended default parameters (e.g., β annealed from (1e-2) to (1e-5), τ = 1.0). In the fixed latent space, counterfactuals are optimized via DiCE's engine for up to 5,000 gradient descent iterations. The loss function includes proximity penalties in both latent space (λ_{lat} = 1.0) and input space (λ_{in} = 1.0) to ensure sample realism and closeness.

Baseline (DiCE-based): Standard DiCE operates directly in the original feature space using gradient descent for up to 1,000 iterations. Proximity loss (λ_{prox} = 1.0) is computed as an L1 distance weighted by inverse Median Absolute Deviation (MAD).

For both methods, immutable features (sex, race) are constrained during optimization to prevent modification. All experiments employ a 90%/10% train/test split and a fixed global random seed to ensure reproducibility.

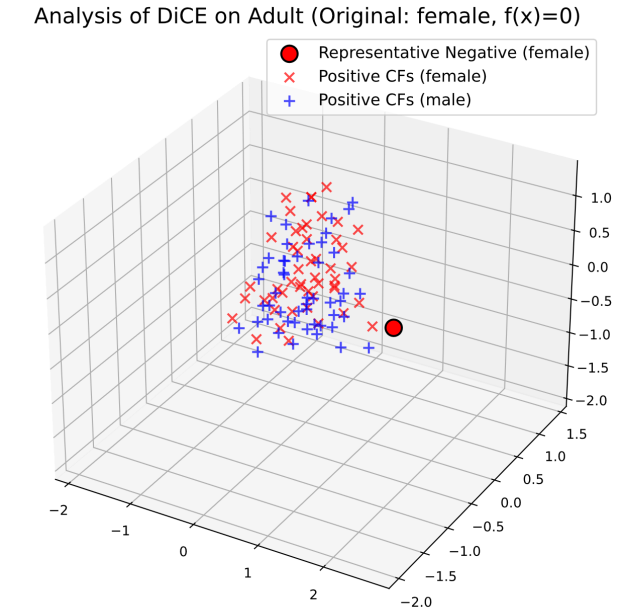
5. Experimental results and analysis

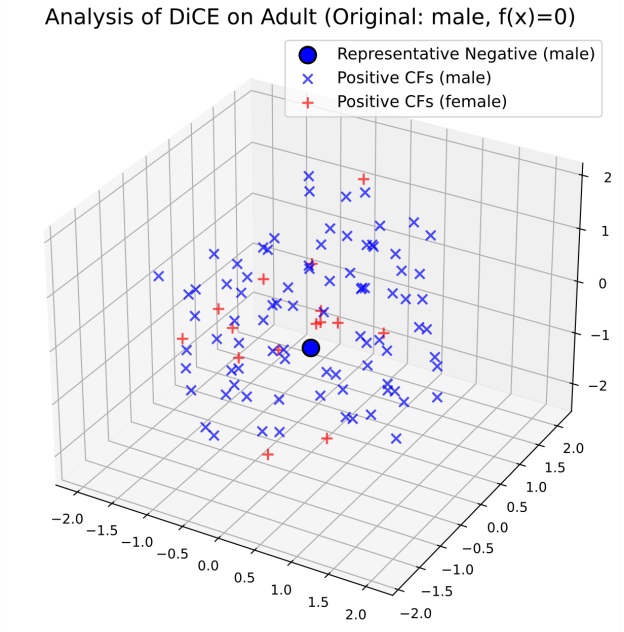
This chapter presents a comprehensive set of experiments to systematically evaluate the effectiveness of our proposed end-to-end framework in auditing, diagnosing, and mitigating algorithmic bias.

5.1. Counterfactual generators

5.1.1. Qualitative analysis

To intuitively assess and compare different counterfactual explanation methods (DiCE vs. TABCF), we employed *t*-SNE dimensionality reduction for visualization. As shown in Figure 1, high-dimensional Counterfactual (CF) samples are projected into a three-dimensional space. In the figure, large circles represent representative negative samples from a group, "*x*" markers denote CFs retaining the original identity, and "+" markers indicate CFs where identity has flipped.





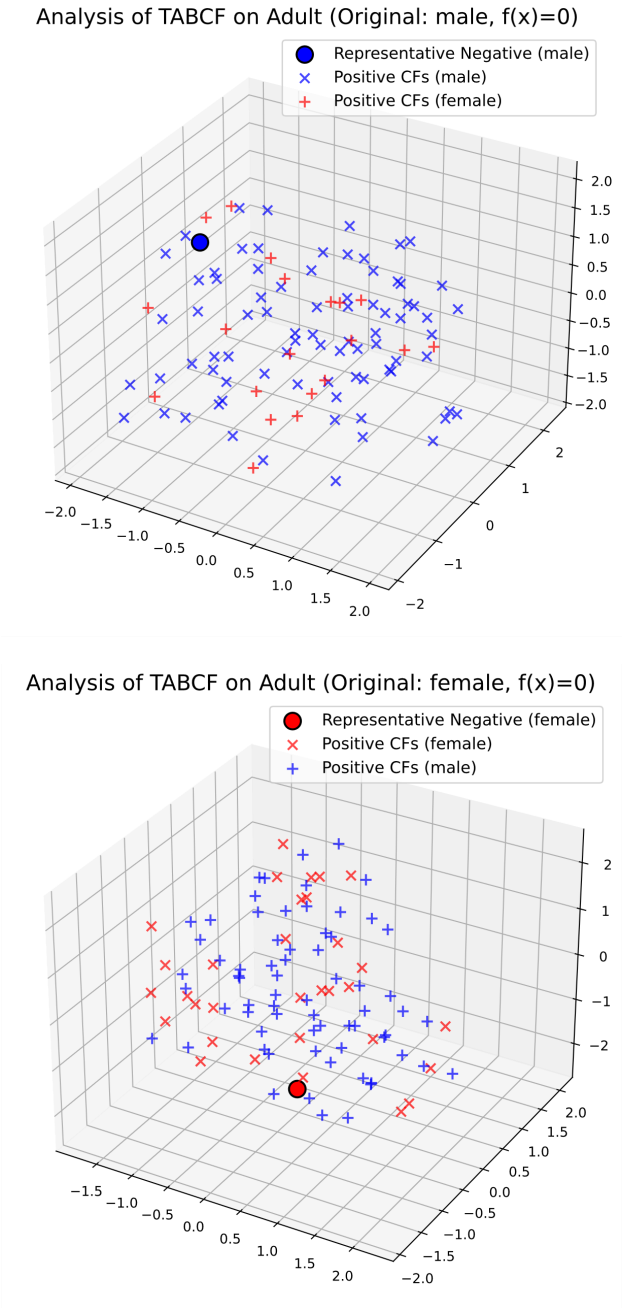


Figure 1. *t*-SNE visualization of counterfactuals on the adult dataset

From the comparative analysis of these visualizations, several key observations emerge:

1. Presence of Model Bias: Both DiCE and TABCF visualizations reveal significant asymmetry in the spatial distribution and color composition of CF point clouds generated for the male and female groups. This asymmetry provides strong evidence that the underlying MLP model exhibits bias. If the MLP model were fair, the distribution patterns of counterfactuals generated for the two groups should be similar across both methods.

2. Impact of Counterfactual Generation Method: The shape and dispersion of point clouds differ substantially between the two methods, reflecting differences in their underlying optimization principles. DiCE-generated CFs, particularly for the male group, show extreme spatial divergence, spreading across the

entire feature space. This suggests that some generated counterfactuals may not sufficiently respect the data manifold, potentially resulting in unrealistic feature combinations. In contrast, TABCF-generated point clouds display notable consistency and clustering, aligning with its design to capture the underlying data manifold via the Transformer-VAE architecture. This indicates that TABCF targets regions of high plausibility within the latent structure of the data, rather than seeking arbitrary mathematical optima.

In summary, the *t*-SNE visualization provides strong qualitative evidence for evaluating counterfactual explanation fairness. It demonstrates that different counterfactual generation methods can reveal different directions and types of bias within the same biased model. This underscores the critical importance of choosing appropriate auditing tools and highlights the value of employing multiple complementary explanation methods to obtain a thorough and nuanced understanding of internal biases in AI models.

5.1.2. Quantitative analysis

Table 1. Evaluation results on the adult, COMPAS, bank, and customer datasets

Dataset	$CFlips_p$ (%)	$CFlips_u$ (%)	$\Delta CFlips$	$nDCCF_p$	$nDCCF_u$	$\Delta nDCCF$
Adult	16.36	64.59	48.23	0.6728	-0.2918	0.9646
compas	14.37	73.58	59.2	0.7125	-0.4715	1.184
bank	0	55.54	55.54	0	0.1107	0.1107
customer	55.11	41.05	14.06	-0.1023	0.1789	0.2812

$CFlips_p$ (%): Percentage of counterfactuals for the privileged group that require identity flipping. Lower values are better, indicating the group can succeed while "being themselves." $CFlips_u$ (%): Percentage of counterfactuals for the unprivileged group that require identity flipping. Lower is better, showing that the group does not need to emulate the privileged group to succeed. $\Delta CFlips$: Difference between the two groups' CFlips. Values closer to 0 are preferred, indicating equality in required identity flips. Large positive values indicate severe discrimination against the unprivileged group.

$nDCCF_p$/$nDCCF_u$: Normalized Discounted Cumulative Counterfactual Fairness score for privileged/unprivileged groups. Values closer to +1.0 indicate high-quality, low-cost advice that does not require identity flips.

$\Delta nDCCF$: Difference in nDCCF between the two groups. Values near 0 indicate equitable treatment.

The evaluation results on the four benchmark datasets (Table 1) highlight several critical observations: Classic Bias Patterns in Adult and COMPAS: In both datasets, unprivileged groups (female/non-white) have substantially higher identity-flip requirements ($CFlips_u$ = 64.59% and 73.58%, respectively) than privileged groups. The large $\Delta CFlips$ and strongly negative $nDCCF_u$ scores demonstrate that the models systematically require unprivileged groups to emulate privileged groups' features to achieve favorable outcomes. Extreme Asymmetry in the Bank Dataset: The privileged group (younger individuals) has a $CFlips_p$ of 0%, meaning the model provides a path to success without identity flipping. In contrast, the unprivileged group (older individuals) faces over 55% identity-flip requirements, illustrating textbook-case path inequality. Subtle Reverse Bias in Customer Dataset: Surprisingly, in this dataset, the privileged group (male) experiences a higher identity-flip rate than the unprivileged group (female) (55.11% vs. 41.05%), and its nDCCF score is also significantly lower. This suggests the model may have overlearned certain success patterns associated with females and misapplied them across all users, resulting in unfair outcomes for males.

Overall, these findings provide compelling evidence that: 1) severe indirect discrimination remains pervasive even under "unconscious fairness" settings; 2) algorithmic bias manifests in diverse forms, highly dependent on data and task contexts, and may even exhibit atypical patterns such as reverse discrimination.

This underscores the necessity and superiority of adopting our high-fidelity audit framework, capable of capturing multidimensional biases.Proxy Variable Diagnosis.

5.2. Proxy variable diagnosis

After detecting discrimination in the models, we applied the two-stage NOCCO-Shapley diagnosis framework to identify proxy variables in the data and understand how the model leverages these proxies. The results are shown in Figure 2–5.

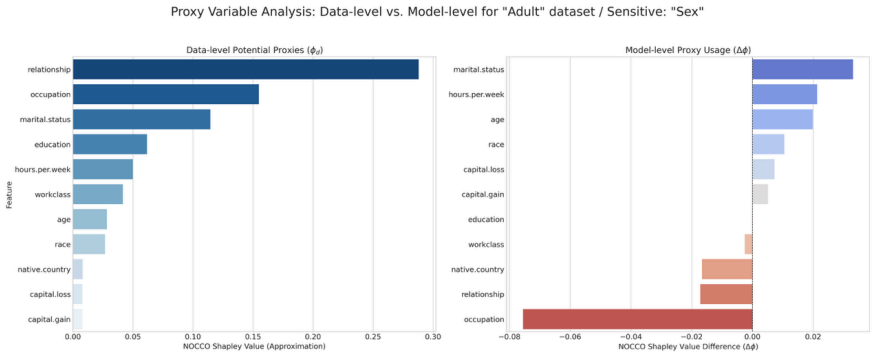


Figure 2. On the adult dataset, the NOCCO-Shapley-based proxy variable diagnostics indicate that larger absolute values signify stronger associations between the feature and sensitive attributes. (a) Data level. (b) Model level

Proxy Variable Detection at the Data Level (ϕ_d): Figure 2(a) shows that relationship (family status) is the feature most strongly correlated with gender in the data (highest ϕ_d), followed by occupation and marital status. These are the proxy variables most likely to be exploited by any model.

Model-Level Discrimination Diagnosis ($\Delta\phi$): Figure 2(b) reveals how models actually utilize data features in decision-making. It demonstrates that occupation and marital-status are heavily leveraged proxy variables by models. Even if models satisfy unconscious fairness, indirect discrimination may still occur through these proxies.

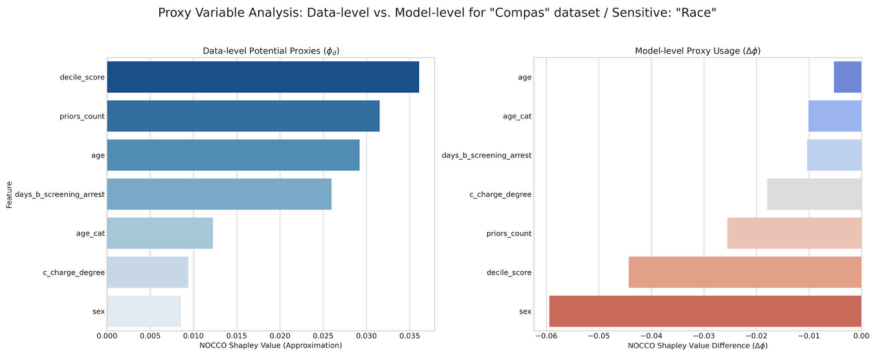


Figure 3. On the compas dataset, the diagnostic results based on NOCCO-Shapley proxy variables indicate that a larger absolute value signifies a stronger association between the feature and the sensitive characteristic. (a) Data level. (b) Model level

Data-level potential (ϕ_d): Figure 3(a) shows that decile_score (risk score) and priors_count (prior convictions) are the two strongest proxies associated with race, reflecting significant statistical disparities across racial groups and highlighting high-risk sources of discrimination.

Model-level discrimination ($\Delta\phi$): Figure 3(b) shows that sex is highly utilized by the model despite being weakly correlated with race in the data, while decile_score is also penalized. This suggests intersectional discrimination, where individuals possessing a combination of sensitive traits (e.g., certain race and gender) may be disproportionately penalized, whereas age-related features may be rewarded. Such complex, multi-variable discrimination patterns are not detectable through simple auditing, demonstrating the unique value of our diagnostic framework.

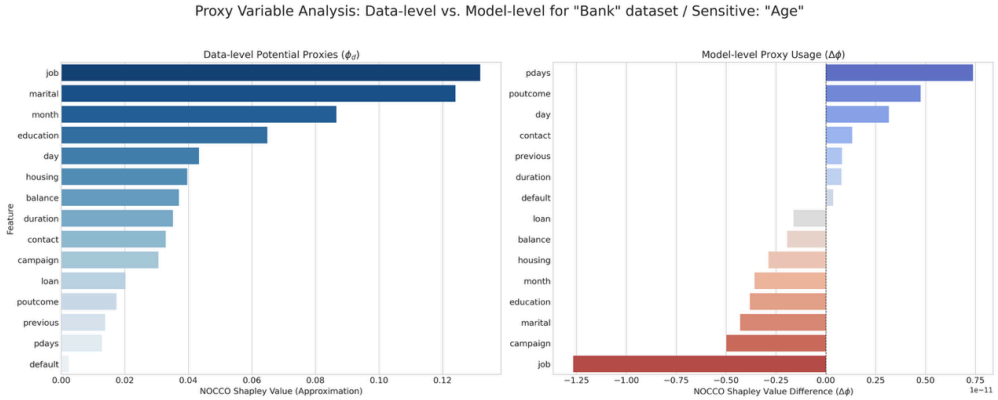


Figure 4. On the bank dataset, the diagnostic results based on NOCCO-Shapley proxy variables indicate that a larger absolute value signifies a stronger association between the feature and the sensitive feature. (a) Data level. (b) Model level

Data-level potential (ϕ_d): Figure 4(a) shows that job and marital are the strongest proxies for age, aligning with common socio-economic knowledge and indicating the main potential sources of age bias.

Model-level discrimination ($\Delta\phi$): Figure 4(b) reveals complex model behavior. The model heavily penalizes job (the strongest data-level proxy) along with marital and education. This may indicate an attempt to "mask" direct demographic labels. However, the model simultaneously exploits and rewards features related to historical marketing activity, such as pdays (days since last contact) and poutcome (outcome of the previous campaign), highlighting a shift from demographic to behavioral proxies.

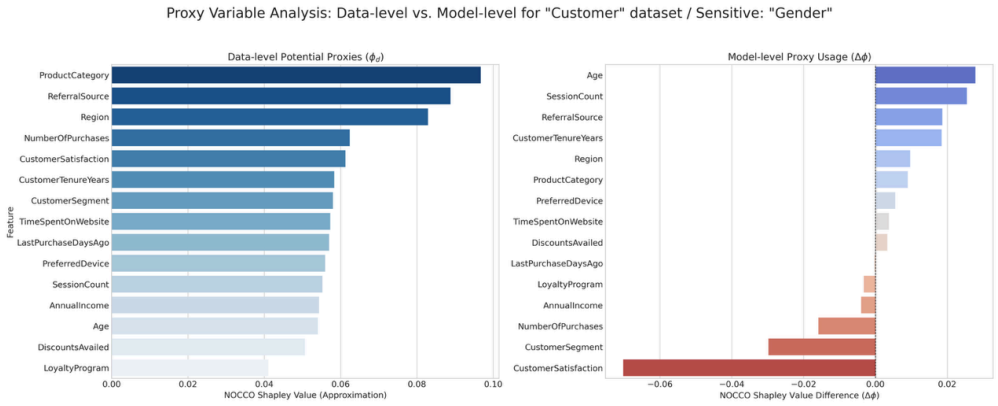


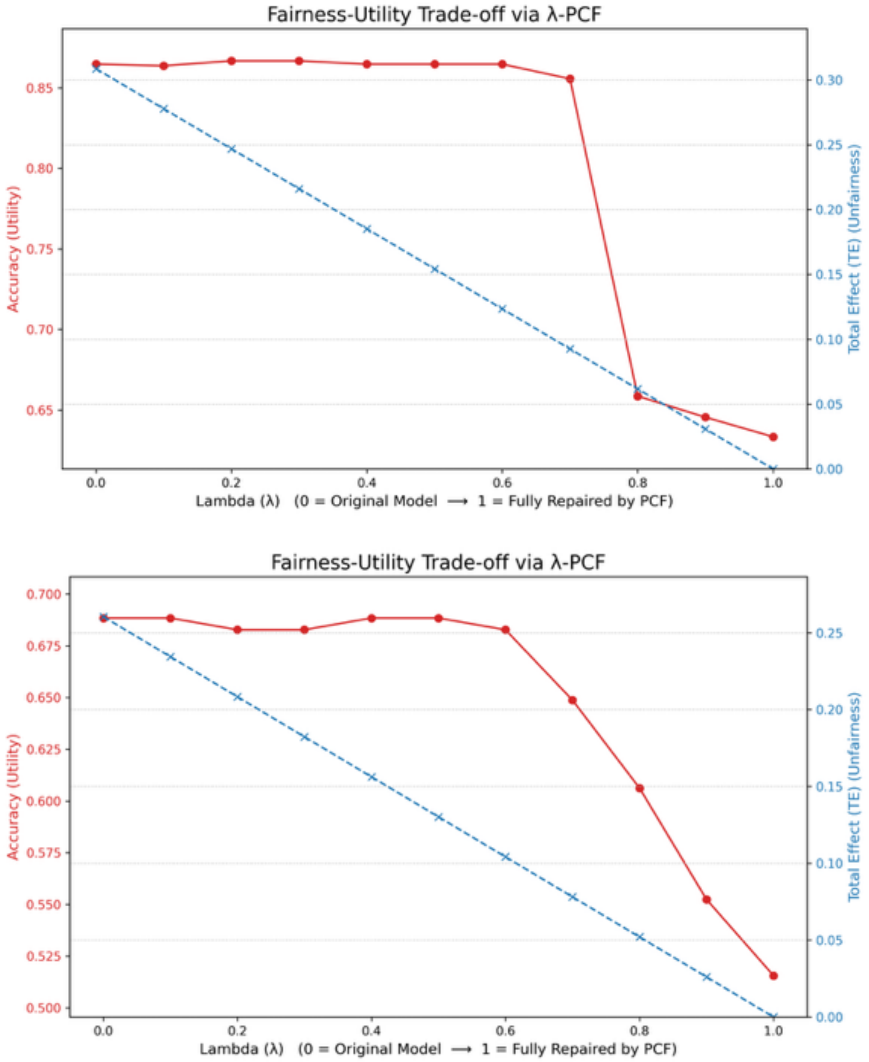
Figure 5. On the customer dataset, based on NOCCO-Shapley proxy variable diagnostics, a higher absolute value indicates a stronger association between the feature and sensitive attributes. (a) Data level. (b) Model level

Data-level potential (ϕ_d): Figure 5(a) shows that ProductCategory is the strongest proxy for gender, consistent with known differences in purchasing patterns. ReferralSource and behavioral metrics such as NumberOfPurchases and CustomerSatisfaction also exhibit notable proxy potential.

Model-level discrimination ($\Delta\phi$): Figure 5(b) indicates that the model heavily penalizes CustomerSatisfaction, suggesting that high satisfaction may be negatively associated with a particular gender (possibly the disadvantaged group). Conversely, the model rewards and heavily utilizes more objective features like Age and SessionCount, reflecting a preference for behavior-based or demographic-neutral indicators in its decision-making.

5.3. Bias mitigation

For the MLP models with diagnosed bias, we applied the λ-PCF post-processing method to balance fairness and accuracy. The results are illustrated in Figure 6.



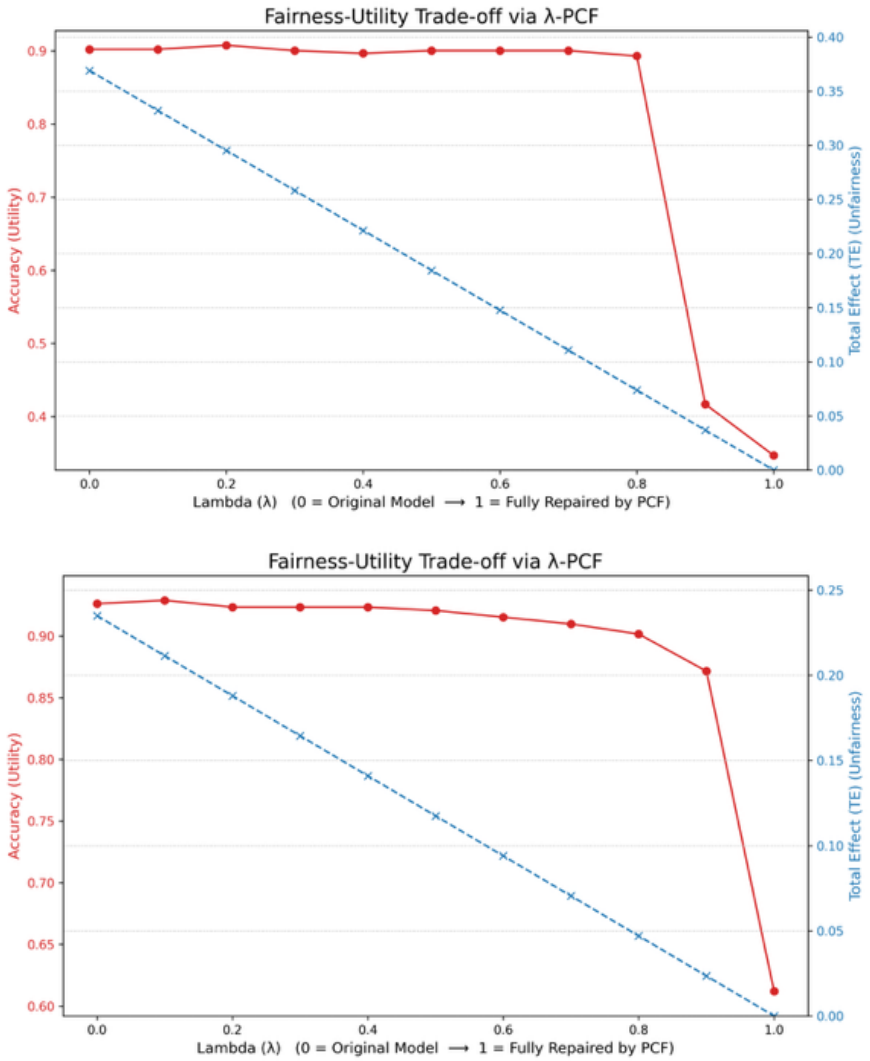


Figure 6. Fairness–utility trade-off curves on the adult, COMPAS, bank, and customer datasets under λ-PCF

As shown in Figure 6, we plotted the variation curves of fairness (measured by Total Effect, TE) and utility (measured by Accuracy) for the model across four datasets—Adult, compas, bank, and customer—by adjusting the λ parameter from 0 to 1. Table 2 lists the fairness and accuracy values when $\lambda = 0$ and $\lambda = 1$.

Table 2. Fairness and accuracy on the adult, compas, bank, and customer datasets when lambda = 0 and lambda = 1

Dataset	Lambda = 0		Lambda = 1	
	Accuracy	TE	Accuracy	TE
adult	0.9050	0.4167	0.5760	0.0
compas	0.6884	0.2605	0.5156	0.0
bank	0.9022	0.3690	0.3469	0.0
customer	0.9262	0.2349	0.6120	0.0

Table 3. Fairness and accuracy on the adult, compas, bank, and customer datasets when λ is set to the prior probability of the dominant group in the dataset $\lambda = P(S = s_{priv})$

Dataset	Privileged Prior (λ)	Accuracy	TE
Adult	0.6689	0.9000	0.1380
compas	0.6569	0.6657	0.0894
bank	0.5728	0.9004	0.1577
customer	0.6569	0.6657	0.0894

After exploring the entire trade-off space, we further validated the λ selection strategy based on data prior proposed in Section 3.3. According to this strategy, we selected λ equal to the prior probability of males (the dominant group) in the Adult dataset, i.e., $\lambda \approx 0.67$. Under this specific setting, Table 2 and 3 reveal that the model's final accuracy decreases from 0.9050 to 0.9000, while the final TE value significantly drops from 0.4167 to 0.1380, achieving a favorable trade-off. Our experiments demonstrate that the λ-PCF framework offers a practical and effective bias mitigation approach. The trade-off curve provides decision-makers with a visual basis for judgment, while our proposed λ selection strategy based on data prior offers a non-arbitrary, theoretically grounded default option. This approach achieves a significant fairness improvement while minimizing the impact on model utility, reaching an ideal, interpretable equilibrium point.

Although our end-to-end framework achieves notable results in auditing, diagnosing, and mitigating bias, several limitations warrant attention. First is the reliance on domain knowledge: the two-stage NOCCO-Shapley diagnostic method effectively identifies proxy variables strongly correlated with sensitive attributes and exploited by the model. However, determining whether such correlations represent unjustifiable bias to be eliminated or legitimate business logic-driven associations ultimately requires domain expert judgment. Second is enhancing the framework's generalization and robustness; future work should test its robustness and scalability on larger, more complex datasets with higher feature dimensions. Finally, investigating the framework's performance with non-binary or continuous-valued sensitive attributes represents a crucial extension.

6. Conclusion

In this work, we proposed a systematic, end-to-end framework integrating discrimination auditing, proxy diagnosis, and bias mitigation to address the increasingly severe problem of indirect discrimination under the "unconscious fairness" setting. First, we emphasized and demonstrated that high-quality, unbiased counterfactual explanations are the foundation of reliable bias auditing. Through experimental comparisons, we showed that the advanced generative counterfactual method (TABCF) can reveal hidden discrimination patterns in models more deeply and accurately than traditional approaches (DiCE). This highlights the importance of the reliability of auditing tools themselves in fairness research. Second, building on this, our innovative two-stage NOCCO-Shapley diagnosis successfully identifies high-risk proxy variables at the data level and further reveals how models leverage these variables in decision-making to enact discrimination, providing unprecedented insight into the mechanisms of proxy-based bias. Finally, we validated the effectiveness of the adjustable λ-PCF mitigation strategy as a practical tool. It can mitigate bias without retraining models, and through its visual trade-off curves and data-prior guided λ selection, it provides a principled, interpretable means to achieve a controllable balance between fairness and utility in real-world settings. Extensive experiments across multiple benchmark datasets confirmed the superiority of our

framework. The seamless integration of auditing, diagnosis, and mitigation not only yields more accurate and comprehensive results but also provides a coherent and actionable methodology for the construction, deployment, and governance of responsible AI systems.

References

[1] Pitoura, E., Stefanidis, K., & Koutrika, G. (2022). Fairness in rankings and recommendations: An overview. *The VLDB Journal, 31*(3), 431–458. https: //doi.org/10.1007/s00778-022-00735-9

[2] von Zahn, M., Feuerriegel, S., & Kuehl, N. (2022). The cost of fairness in AI: Evidence from e-commerce. *Business & Information Systems Engineering, 64*, 335–348. https: //doi.org/10.1007/s12599-022-00730-0

[3] De-Arteaga, M., Feuerriegel, S., & Saar-Tsechansky, M. (2022). Algorithmic fairness in business analytics: Directions for research and practice. *Production and Operations Management, 31*(12), 4492–4509. https: //doi.org/10.1111/poms.13787

[4] Cornacchia, G., Anelli, V. W., Biancofiore, G. M., Narducci, F., Pomo, C., Ragone, A., & Di Sciascio, E. (2023). Auditing fairness under unawareness through counterfactual reasoning. *Information Processing & Management, 60*(2), Article 103224. https: //doi.org/10.1016/j.ipm.2022.103224

[5] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. S. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)* (pp. 214–226). ACM. https: //doi.org/10.1145/2090236.2090255

[6] Chen, J., Kallus, N., Mao, X., Svacha, G., & Udell, M. (2019). Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT '19)* (pp. 339–348). ACM. https: //doi.org/10.1145/3287560.3287566

[7] Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT '20)* (pp. 607–617). ACM. https: //doi.org/10.1145/3351095.3372850

[8] Pawelczyk, M., Datta, K., van-den-Heuvel, J., Lakkaraju, H., & Brockhaus, J. (2021). Learning model-agnostic counterfactual explanations for tabular data. In *Companion Proceedings of the Web Conference 2021 (WWW '21 Companion)* (pp. 443–452). ACM. https: //doi.org/10.1145/3442442.3451174

[9] Cornacchia, G., Anelli, V. W., Biancofiore, G. M., Narducci, F., Pomo, C., Ragone, A., & Di Sciascio, E. (2023). Counterfactual reasoning for decision model fairness assessment. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion)* (pp. 1026–1035). ACM. https: //doi.org/10.1145/3543873.3587370

[10] Karimi, A. H., Schölkopf, B., & Valera, I. (2021). Algorithmic recourse: From counterfactual explanations to recourse. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)* (pp. 353–362). ACM. https: //doi.org/10.1145/3442188.3445896

[11] Panagiotou, E., Heurich, M., Landgraf, T., & Ntoutsi, E. (2024). TABCF: Counterfactual explanations for tabular data using a transformer-based VAE. In *Proceedings of the 5th ACM International Conference on AI in Finance (ICAIF '24)*, Article 9, 9 pages. https: //doi.org/10.1145/3677052.3698673

[12] Alves, G., Bernier, F., Couceiro, M., Makhlouf, K., Palamidessi, C., & Zhioua, S. (2023). Survey on fairness notions and related tensions. *EURO Journal on Decision Processes, 11*, Article 100033. https: //doi.org/10.1016/j.ejdp.2023.100033

[13] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29* (pp. 3315–3323). Curran Associates, Inc. https: //proceedings.neurips.cc/paper/2016/hash/9d2682367b1e5db5e44802bf0b7d87d0-Abstract.html

[14] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge*

Discovery and Data Mining (KDD '15) (pp. 259–268). ACM. https: //doi.org/10.1145/2783258.2783311

[15] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML '13)* (pp. 325–333). PMLR. https: //proceedings.mlr.press/v28/zemel13.html

[16] Grabowicz, P. A., Perello, N., & Mishra, A. (2022). Marrying fairness and explainability in supervised learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)* (pp. 1905–1916). ACM. https: //doi.org/10.1145/3531146.3533236

[17] Mishler, A., Kennedy, E. H., & Chouldechova, A. (2021). Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)* (pp. 386–400). ACM. https: //doi.org/10.1145/3442188.3445902

[18] Nandy, P., DiCiccio, C., Venugopalan, D., Logan, H., Basu, K., & El Karoui, N. (2022). Achieving fairness via post-processing in web-scale recommender systems. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)* (pp. 715–725). ACM. https: //doi.org/10.1145/3531146.3533136

[19] Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems 30* (pp. 4069–4079). Curran Associates, Inc. https: //proceedings.neurips.cc/paper/2017/hash/486cd6b6ae941559b31c53d0a6e4d99d-Abstract.html

[20] Zhou, Z., Liu, T., Bai, R., Gao, J., Kocaoglu, M., & Inouye, D. I. (2024). Counterfactual fairness by combining factual and counterfactual predictions. In *Advances in Neural Information Processing Systems 37*. Curran Associates, Inc. https: //openreview.net/forum?id=placeholder

[21] Jang, E., Gu, S., & Poole, B. (2016). *Categorical reparameterization with Gumbel-Softmax*. arXiv preprint. https: //arxiv.org/abs/1611.01144

[22] Pelegrina, G. D., Couceiro, M., & Duarte, L. T. (2024). A preprocessing Shapley-value approach to detect relevant and disparity-prone features in machine learning. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, Article 11, 11 pages. https: //doi.org/10.1145/3630106.3658905

[23] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine bias*. ProPublica. https: //www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[24] Frank, A., & Asuncion, A. (2010). *UCI machine learning repository* [Dataset]. University of California, Irvine, School of Information and Computer Sciences. http: //archive.ics.uci.edu/ml

[25] Vijayaraj, G. (2023). *Customer purchase behavior dataset (e-commerce)* [Dataset]. Kaggle. https: //www.kaggle.com/datasets/gauthamvijayaraj/customer-purchase-behavior-dataset-e-commerce/data