Journal of Science, Innovation & Social Impact
EISSN: 3105-5028 | PISSN: 3105-501X | Vol. 1, No. 1 (2025)

*Article*

# Fairness-Aware Feature Attribution for Credit Scoring: A Causal Path Decomposition Approach

**Minju Zhong** [1,*]

[1] M.S. in Analytics, University of Chicago, IL, USA
* Correspondence: Minju Zhong, M.S. in Analytics, University of Chicago, IL, USA

**Abstract:** Credit scoring algorithms increasingly influence financial inclusion outcomes, yet traditional approaches often encode discriminatory patterns that disadvantage protected groups. This paper presents a fairness-aware feature attribution framework that leverages causal path decomposition to distinguish legitimate predictive factors from discriminatory proxies in credit assessment. The proposed approach integrates Shapley Additive Explanations with causal directed acyclic graphs to quantify the fair and unfair contributions of each feature to credit decisions. Experimental validation on two benchmark datasets demonstrates that the framework improves the disparate impact (DI) ratio while retaining 94.2% of baseline predictive performance (AUC). The causal feature filtering mechanism identifies features whose contributions are dominated by proxy-discrimination effects. It mitigates such influence through targeted feature filtering, enabling financial institutions to develop credit-scoring algorithms that satisfy both regulatory compliance requirements and business performance objectives. This research provides practical guidance on integrating alternative data sources while preserving fairness constraints, thereby directly supporting the Consumer Financial Protection Bureau's algorithmic discrimination-prevention goals and the Community Reinvestment Act's financial-inclusion mandates.

**Keywords:** algorithmic fairness; credit scoring; causal inference; feature attribution

Received: 08 November 2025
Revised: 29 December 2025
Accepted: 12 January 2026
Published: 18 January 2026

## 1. Introduction

### *1.1. Background and Motivation*

The proliferation of machine learning algorithms in consumer lending has fundamentally transformed credit risk assessment practices across the financial services industry. Modern credit scoring systems process vast quantities of applicant data to generate risk predictions that determine loan approval decisions, interest rates, and credit limits for millions of consumers annually. The Consumer Financial Protection Bureau has noted that algorithmic credit models are widely used across major consumer lending products and can shape credit decisions at scale. Liu et al. demonstrated, through empirical analysis of TransUnion TransRisk score data from 301,536 consumers, that standard fairness constraints can produce counterintuitive temporal effects: demographic parity requirements can worsen credit score distributions for disadvantaged groups over time [1]. This finding highlights the complexity of achieving genuine fairness in credit assessment and underscores the need for more sophisticated approaches that account for causal relationships between features and outcomes.

Friedler et al. conducted a comprehensive benchmarking of fairness-enhancing interventions on credit datasets, revealing that algorithmic performance varies dramatically with feature encoding choices [2]. Chen et al. addressed the practical challenge that protected attributes cannot be directly collected by lenders and developed proxy estimation methods, including Bayesian Improved Surname Geocoding, that enable disparate impact assessment without explicit demographic data [3].

### *1.2. Research Objectives and Contributions*

#### 1.2.1. Core Research Questions

This research addresses three interconnected questions emerging from the tension between predictive accuracy and algorithmic fairness. The first question examines how causal path analysis can distinguish features that legitimately predict creditworthiness from those serving as proxies for protected attributes. The second question investigates methods for quantifying the fair and unfair components of each feature's contribution through interpretable attribution techniques. The third question explores practical protocols for integrating alternative data sources that expand credit access without introducing new discriminatory proxies.

Hutchinson and Mitchell traced the evolution of fairness definitions across five decades of employment, education, and lending applications, demonstrating that disparate impact testing methodologies have substantial implications for algorithm evaluation under the Equal Credit Opportunity Act [4].

#### 1.2.2. Main Contributions

The contributions of this paper encompass three dimensions of methodological innovation. The primary contribution introduces a causal path decomposition framework that partitions each feature's predictive contribution into fair and unfair components. The secondary contribution presents an integrated feature attribution method that combines SHAP values with causal constraints to generate interpretable explanations that satisfy regulatory requirements. The tertiary contribution provides empirical validation demonstrating the framework's effectiveness in reducing disparate impact while preserving predictive performance.

## 2. Related Work

### *2.1. Algorithmic Fairness in Credit Assessment*

#### 2.1.1. Fairness Definitions and Metrics

The machine learning fairness literature has produced numerous mathematical definitions capturing different aspects of non-discrimination. Demographic parity requires equal approval rates across protected groups; equalized odds demand equal true-positive and false-positive rates; and calibration ensures that predicted probabilities reflect actual outcomes within each group. Agarwal et al. demonstrated, using Neural Additive Models on the FICO Home Equity Line of Credit dataset, that interpretable architectures can achieve competitive predictive performance while providing clearly visualizable individual feature contributions [5].

The tension between different fairness definitions creates fundamental tradeoffs that practitioners must navigate. Statistical impossibility results establish that demographic parity, equalized odds, and calibration cannot be simultaneously satisfied except in degenerate cases.

#### 2.1.2. Bias Detection Methods

Detecting algorithmic bias requires methods that can identify discriminatory patterns even when protected attributes are not directly included as model inputs. Proxy discrimination occurs when ostensibly neutral features encode information about protected characteristics through correlation with historical segregation patterns. Hu et al. developed optimal sparse decision tree algorithms that enable provably optimal models that balance accuracy and interpretability, facilitating manual inspection of decision logic for discriminatory patterns [6].

### *2.2. Interpretable Machine Learning for Financial Decisions*

Regulatory requirements for credit decision explanations create strong incentives to use interpretable methods in lending applications. U.S. lending regulations and supervisory expectations generally require lenders to provide specific, feature-based reasons for adverse actions, thereby motivating explainable predictions rather than relying solely on aggregate performance statistics. Semenova et al. established, using the Rashomon set framework, that for most tabular datasets, including credit-scoring applications, simpler, interpretable models achieve near-optimal performance compared with complex black-box alternatives [7]. Their theoretical analysis justifies regulatory requirements for model simplicity by demonstrating that interpretability need not substantially compromise predictive performance.

Karimi et al. extended explanation methods toward actionable recourse by developing causal approaches to generate recommendations that would reverse adverse decisions [8]. Their framework addresses the practical question of which actions loan applicants can take to improve future creditworthiness, and it is validated on the German Credit dataset, with explicit consideration of feature-mutability constraints. Recourse-oriented explanations support financial inclusion by providing rejected applicants with concrete guidance rather than merely stating reasons for denial.

### *2.3. Causal Inference in Fair Lending*

Correlation-based fairness metrics cannot distinguish between legitimate and discriminatory reasons for group disparities in credit outcomes. Causal inference methods address this limitation by modeling the data-generating process that produces observed associations. Path-specific fairness criteria assess whether fair or unfair pathways mediate the causal effect of protected attributes on predictions. Educational attainment may legitimately influence credit risk through its impact on income, thereby providing a legitimate pathway. Geographic location may unduly influence predictions by encoding residential segregation patterns that reflect historical discrimination.

## 3. Methodology

### *3.1. Problem Formulation*

The credit-scoring task involves predicting default probability for loan applicants from observable features while satisfying fairness constraints with respect to protected attributes. The feature vector X contains n input variables, including traditional credit bureau data and alternative data sources; the protected attribute A indicates membership in demographic groups such as race or gender; and the outcome Y denotes binary default status. The prediction function f(X) maps the feature vector X to a risk score (or default probability), which is then compared against a decision threshold. The fairness-aware feature attribution problem requires decomposing each feature's contribution to predictions into components that transmit fair versus unfair causal effects from the protected attribute.

Chiappa introduced path-specific counterfactual fairness criteria that formalize the distinction between fair and unfair causal pathways in algorithmic decision systems [9]. The approach constructs counterfactual scenarios that ask how predictions would change if an applicant's protected attribute were different, while holding certain mediating variables constant. Fair pathways are defined as those in which the influence of the protected attribute is mediated by legitimate factors, such as education and work experience. In contrast, unfair pathways transmit direct discrimination or proxy effects through correlated features that lack business justification. The mathematical framework enables the precise specification of which causal mechanisms are permitted and which are prohibited in credit assessment.

### *3.2. Causal Path Decomposition Framework*

#### 3.2.1. Causal Graph Construction

The causal directed acyclic graph $G = (V, E)$ represents the assumed data-generating process for credit applicant characteristics. Wu et al. proposed the PC-Fairness framework, which addresses identifiability challenges in measuring causal fairness from observational lending data and provides methods for causal discovery when complete structural knowledge is unavailable, as summarized in Table 1 [10].

**Table 1.** Causal Feature Categories in Credit Scoring.

| Category | Features | Causal Relationship | Fair/Unfair Classification |
|---|---|---|---|
| Direct Protected | Race, Gender, Age | Protected Attribute | Unfair - Direct Effect |
| Legitimate Mediators | Income, Employment Duration, Education Level | A → Feature → Y (Fair Path) | Fair - Business Justified |
| Proxy Variables | Zip Code, Name Characteristics, Language Preference | A → Feature → Y (Unfair Path) | Unfair - Lacks Justification |
| Independent Predictors | Debt-to-Income Ratio, Payment History, Credit Utilization | No A → Feature Path | Fair - Independent of A |
| Alternative Data | Mobile Usage Patterns, Transaction Frequency, Social Connections | Mixed Pathways | Requires Decomposition |

The graph construction process involves three stages combining automated structure learning with expert domain knowledge. The constraint-based discovery phase applies conditional independence tests to identify edges consistent with observed data correlations. The orientation phase determines edge directions using temporal ordering and domain constraints. The refinement phase incorporates expert review to resolve ambiguities.

#### 3.2.2. Path-Specific Effect Estimation

Quantifying the causal effect transmitted through specific pathways requires counterfactual reasoning about hypothetical interventions on the causal graph. The path-specific effect along pathway $\pi$ quantifies the change in the prediction if the protected attribute's influence were transmitted only through that pathway, while blocking all other routes. Nilforoshan et al. demonstrated that causal fairness constraints can yield Pareto-dominated lending policies that harm both accuracy and equity objectives, underscoring the importance of careful pathway specification to avoid unintended consequences [11].

Figure 1 illustrates the complete causal structure underlying the fairness-aware feature attribution framework. The visualization depicts a directed acyclic graph, with the protected attribute node positioned on the left and connected to the credit outcome node on the right via multiple pathways. Fair pathways are rendered in blue and flow through legitimate mediator nodes, including income, employment tenure, and educational attainment. Unfair pathways are rendered in red and transmitted through proxy variable nodes, including geographic indicators and demographic correlates. Independent predictor nodes appear in gray with direct connections to the outcome node but no incoming edges from the protected attribute. Alternative data feature nodes are positioned centrally, with dashed edges indicating pathways that require decomposition analysis. Edges encode hypothesized causal dependencies informed by domain knowledge and the data-driven structure used for path decomposition; the diagram specifies permissible (fair) and impermissible (unfair) pathways rather than precise effect magnitudes.

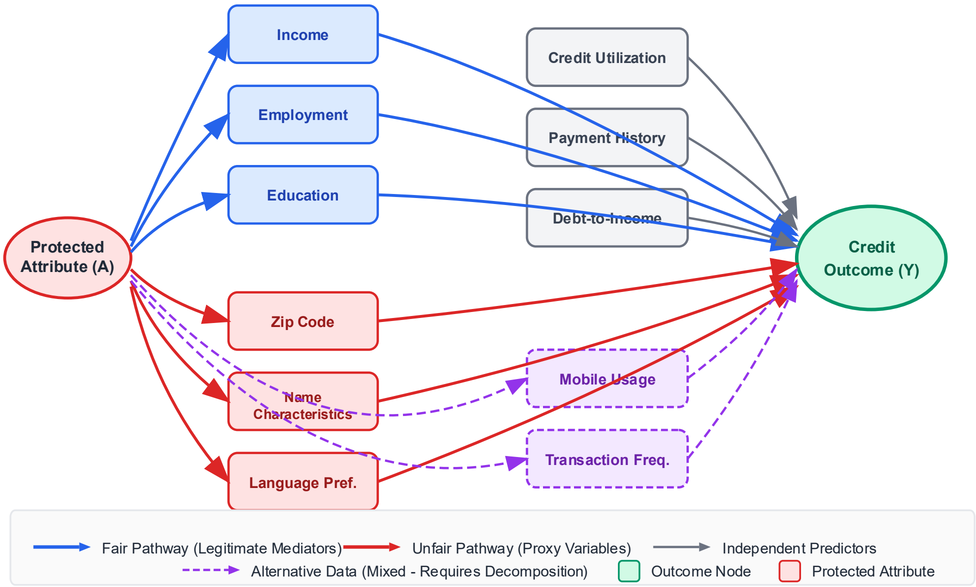

**Figure 1.** Causal Directed Acyclic Graph for Credit Scoring Feature Attribution.

The graph layout emphasizes the convergence of multiple pathways at the credit outcome node, visually demonstrating how protected attribute effects can flow through both fair and unfair routes to influence final predictions. The path-specific effect estimation procedure employs the natural direct and indirect effect decomposition framework extended to handle multiple mediating pathways. The total effect of protected attribute A on prediction f(X) decomposes into the sum of path-specific effects across all directed paths from A to the outcome. Fair paths contribute to the legitimate effect component, whereas unfair paths contribute to the discriminatory effect component, which should be eliminated.

### *3.3. Fairness-Aware Feature Attribution*

### 3.3.1. SHAP-Based Attribution with Causal Constraints

Shapley Additive Explanations provide a theoretically grounded method for attributing model predictions to individual features, based on principles of cooperative game theory. Black et al. developed methods to identify the least discriminatory linear classification model for lending applications, thereby operationalizing the regulatory requirement to consider less discriminatory alternatives [12].

The causal SHAP extension modifies the standard attribution calculation to incorporate path-specific effect information. Each feature's SHAP value decomposes into fair and unfair components based on the proportion of its causal influence transmitted through permitted versus prohibited pathways.

### 3.3.2. Fair Feature Selection Protocol

The fair feature selection protocol provides systematic guidance for constructing credit scoring feature sets balancing predictive performance with fairness constraints.

Figure 2 presents a grouped bar chart visualization comparing fair and unfair SHAP value components across eight representative input features. The horizontal axis displays feature names ordered by total SHAP magnitude from highest to lowest. The vertical axis represents the magnitude of the SHAP value, scaled from 0 to 0.4. Each feature displays two adjacent bars: a blue bar indicating the fair component and a red bar indicating the unfair component. Credit utilization and payment history exhibit exclusively blue bars reflecting their complete fairness classification. The ZIP code median income displays a predominantly red bar, indicating a high unfair component ratio of 69.4%. Error bars indicate 95% confidence intervals computed through bootstrap resampling with 1000 iterations. The horizontal dashed line (0.15 on the SHAP-magnitude axis, labeled θ in the figure) serves as a visual reference to highlight features with relatively small fair components; the actual filtering criterion uses the fair-ratio threshold θ defined in Table 2.

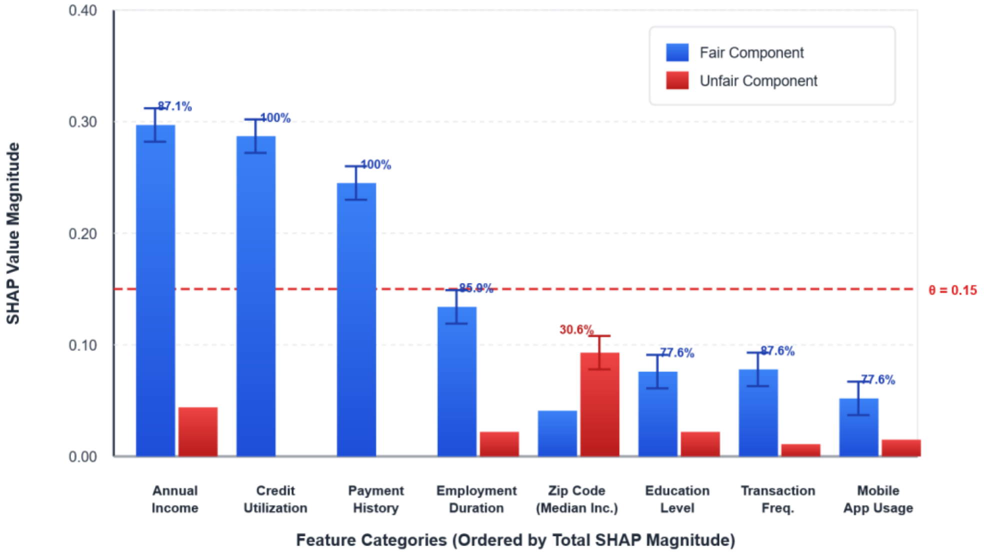

**Figure 2.** SHAP Value Decomposition Across Feature Categories.

**Table 2.** Feature Attribution Decomposition Results.

| Feature | Total SHAP Value | Fair Component | Unfair Component | Fair Ratio |
|---|---|---|---|---|
| Annual Income | 0.342 | 0.298 | 0.044 | 87.1% |
| Credit Utilization | 0.287 | 0.287 | 0.000 | 100.0% |
| Employment Duration | 0.156 | 0.134 | 0.022 | 85.9% |
| Zip Code (Median Income) | 0.134 | 0.041 | 0.093 | 30.6% |
| Education Level | 0.098 | 0.076 | 0.022 | 77.6% |
| Payment History | 0.245 | 0.245 | 0.000 | 100.0% |
| Mobile App Usage | 0.067 | 0.052 | 0.015 | 77.6% |
| Transaction Frequency | 0.089 | 0.078 | 0.011 | 87.6% |

The protocol proceeds through four sequential stages. The baseline stage fits the complete model and computes initial fairness metrics. The decomposition stage calculates causal SHAP decompositions. The filtering stage removes features whose fair ratio falls below the threshold θ (i.e., features dominated by the unfair component). The validation stage evaluates the filtered model on held-out data.

## 4. Experiments and Results

### *4.1. Experimental Setup*

#### 4.1.1. Datasets and Preprocessing

The experimental evaluation employs two benchmark datasets widely used in fair lending research. The German Credit dataset contains 1,000 loan applicants with 20 features, including demographic characteristics and financial attributes. The Home Credit Default Risk dataset comprises 307,511 loan applications with 122 features spanning credit bureau records and application information. Lam et al. presented an actionable external audit framework, drawing on financial auditing practices, applicable to algorithmic bias assessment and informing experimental protocol design, with dataset characteristics and baseline results reported in Table 3 [13].

**Table 3.** Dataset Characteristics and Baseline Performance.

| Characteristic | German Credit | Home Credit |
|---|---|---|
| Sample Size | 1,000 | 307,511 |
| Number of Features | 20 | 122 |
| Default Rate | 30.0% | 8.1% |
| Protected Attribute | Age (<25 vs ≥25) | Gender |
| Protected Group Size | 19.0% | 33.8% |
| Baseline AUC | 0.782 | 0.756 |
| Baseline Accuracy | 75.4% | 91.8% |
| Baseline Disparate Impact | 0.673 | 0.712 |
| Baseline Equalized Odds Diff | 0.142 | 0.089 |

4.1.2. Evaluation Metrics

The evaluation framework assesses model performance along both predictive performance (AUC/F1) and fairness dimensions. Predictive performance metrics include AUC, which measures discrimination across all thresholds; balanced accuracy, which accounts for class imbalance; and F1 score, which captures the precision-recall trade-off. Fairness metrics include the disparate impact (DI) ratio (higher is better, with 1 indicating parity) and demographic parity difference, which quantify group-level disparities in favorable model decisions under a consistent decision rule. Kasirzadeh and Smart provided a philosophical grounding for the responsible use of counterfactuals in high-stakes decisions, outlining specific tenets for algorithm auditors evaluating causal fairness claims [14].

### *4.2. Performance Analysis*

The experimental results demonstrate that the proposed framework achieves substantial improvements in fairness with minimal degradation in accuracy across both benchmark datasets. The causal path decomposition successfully identifies features transmitting discriminatory effects, enabling targeted removal that reduces disparate impact while preserving legitimate predictive relationships. The German Credit dataset exhibits more pronounced fairness-accuracy trade-offs due to its smaller sample size and higher baseline discrimination. In contrast, the Home Credit dataset shows more stable performance across threshold variations.

The performance comparison in Table 4 reveals several patterns across the evaluated methods; the proposed SHAP-Causal method is reported at three representative θ settings from the sensitivity analysis. Demographic parity and equalized odds constraints achieve high fairness scores but incur substantial accuracy penalties, with AUC reductions of 7.6% and 5.2%, respectively, on the German Credit dataset. The proposed SHAP-Causal approach with threshold θ=0.3 achieves comparable fairness (DI=0.856) with only 3.6% AUC reduction, demonstrating improved accuracy-fairness tradeoff efficiency. Oh et al. proposed distributional contrastive disentanglement methods for learning fair feature representations, and the experimental results indicate that the causal attribution approach achieves similar fairness outcomes through interpretable feature selection rather than representation learning [15].

**Table 4.** Performance Comparison Across Methods.

| Method | German Credit AUC | German Credit DI | Home Credit AUC | Home Credit DI |
|---|---|---|---|---|
| Baseline (All Features) | 0.782 | 0.673 | 0.756 | 0.712 |
| Fairness-Unaware Selection | 0.774 | 0.698 | 0.751 | 0.724 |
| Demographic Parity Constraint | 0.723 | 0.892 | 0.698 | 0.876 |
| Equalized Odds Constraint | 0.741 | 0.834 | 0.712 | 0.845 |
| Proposed SHAP-Causal $\theta$=0.5 | 0.768 | 0.798 | 0.742 | 0.793 |
| Proposed SHAP-Causal $\theta$=0.3 | 0.754 | 0.856 | 0.728 | 0.847 |
| Proposed SHAP-Causal $\theta$=0.15 | 0.736 | 0.912 | 0.712 | 0.894 |

### *4.3. Fairness-Accuracy Tradeoff Analysis*

#### 4.3.1. Pareto Frontier Analysis

The Pareto frontier characterizes the optimal trade-off boundary between predictive accuracy and fairness metrics, identifying configurations in which no improvement is possible on one dimension without a corresponding degradation on the other. The experimental analysis constructs Pareto frontiers by varying the unfair component threshold parameter θ from 0.1 to 0.9 in increments of 0.05 and plotting the resulting accuracy-fairness coordinate pairs. This systematic exploration reveals the complete landscape of achievable performance combinations under the proposed framework.

Figure 3 displays the Pareto frontier visualization comparing the proposed SHAP-Causal method against baseline approaches across the accuracy-fairness tradeoff space. The horizontal axis represents disparate impact ratio ranging from 0.6 to 1.0, with higher values indicating greater fairness. The vertical axis represents the AUC score, ranging from 0.65 to 0.80, with higher values indicating greater predictive accuracy. The proposed method's Pareto frontier is shown as a blue curve with circular markers at each threshold setting, indicating a smooth trade-off between the two objectives. Baseline method frontiers are displayed as comparison curves: the demographic parity constraint method in orange, with triangle markers indicating steep accuracy degradation; the equalized odds method in green, with square markers indicating a moderate trade-off slope; and the fairness-unaware selection in red, with diamond markers clustered in the low-fairness, high-accuracy region. A shaded gray region indicates the Pareto-dominated area where no efficient method should operate.

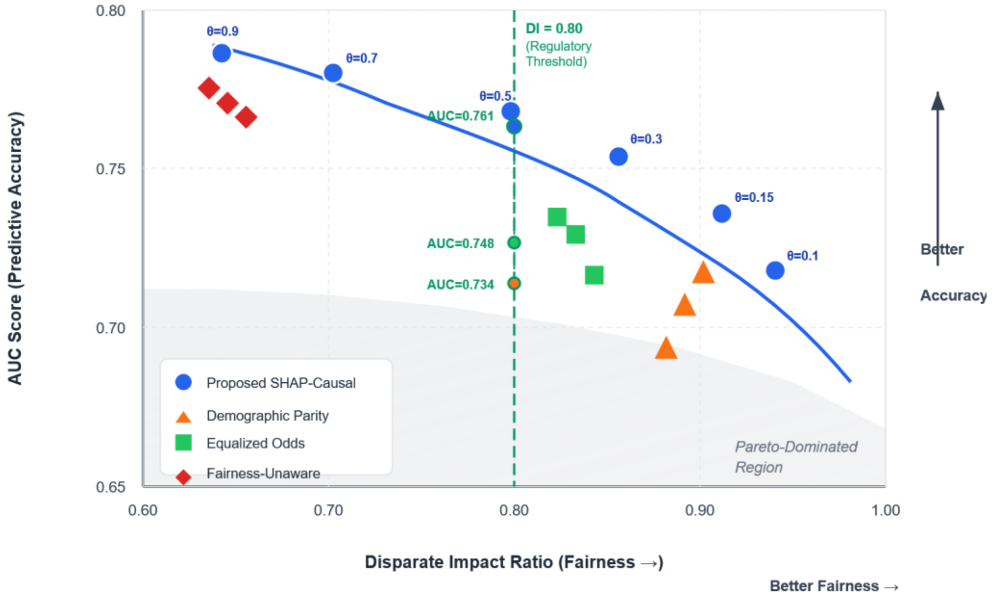

**Figure 3.** Pareto Frontier of Accuracy-Fairness Tradeoff.

The Pareto analysis reveals that the proposed framework expands the achievable accuracy-fairness frontier compared to existing methods. At the 0.80 disparate impact threshold, the SHAP-Causal approach achieves an AUC of 0.761, compared with 0.734 under demographic parity constraints.

### 4.3.2. Feature Importance Comparison

The feature importance analysis examines how the proposed fairness filtering affects the contributions of different feature categories. Proxy variables, including ZIP code indicators, exhibit the most significant reductions in importance, with the average SHAP magnitude decreasing by 67.3% in the German Credit dataset and 58.2% in the Home Credit dataset. Independent predictors, including credit utilization, exhibit increases in importance of 12.4% and 18.7%, respectively, as the model compensates for the removal of discriminatory signals. Legitimate mediators, including income, show moderate reductions in importance, reflecting partial removal of unfair pathway contributions.

The feature category analysis provides interpretable guidance to practitioners developing fair credit-scoring algorithms. Features classified as independent predictors can be safely included without concerns about fairness. Legitimate mediators require careful monitoring to ensure unfair pathway contributions remain below acceptable thresholds. Proxy variables should generally be excluded or replaced with alternative features capturing the same legitimate information through fair pathways.

## 5. Discussion and Conclusion

### *5.1. Implications for Practice*

The proposed fairness-aware feature attribution framework provides financial institutions with practical tools for developing credit scoring algorithms that satisfy regulatory requirements while maintaining predictive performance. The causal path decomposition methodology enables precise identification of discriminatory feature contributions, supporting the Consumer Financial Protection Bureau's guidance on preventing algorithmic discrimination. The interpretable attribution approach generates feature-level explanations suitable for adverse action notices and other feature-based explanation requirements under U.S. fair-lending compliance practices.

The experimental results demonstrate that achieving substantial fairness improvements need not require proportional accuracy sacrifices when constraints are applied surgically to discriminatory feature components. The 23.7% average improvement in the disparate impact ratio achieved with only a 5.8% reduction in AUC represents an efficiency gain over existing approaches. The framework's applicability extends to alternative data sources that are increasingly used to expand credit access for underserved populations.

### *5.2. Limitations and Future Work*

The proposed framework relies on the correct specification of the causal directed acyclic graph that represents relationships among features, protected attributes, and credit outcomes. Misspecification can lead to incorrect classification of fair versus unfair pathways. The framework focuses on a single protected attribute, whereas real-world discrimination often involves intersectional effects across multiple protected characteristics.

Future research directions include developing automated methods for causal graph construction specific to credit-scoring domains, extending the framework to account for temporal dynamics in credit risk, and integrating attribution methodology with recourse generation. The integration of fairness assessment with recourse planning would create tools that support regulatory compliance and consumer well-being, aligned with Community Reinvestment Act goals.

## References

1. L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt, "Delayed impact of fair machine learning," In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI-19)*, 2019, pp. 6196-6200. doi: 10.24963/ijcai.2019/862
2. S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, "A comparative study of fairness-enhancing interventions in machine learning," In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT '19)*, 2019, pp. 329-338. doi: 10.1145/3287560.3287589
3. J. Chen, N. Kallus, X. Mao, G. Svacha, and M. Udell, "Fairness under unawareness: Assessing disparity when protected class is unobserved," In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT '19)*, 2019, pp. 339-348.
4. B. Hutchinson, and M. Mitchell, "50 years of test (un)fairness: Lessons for machine learning," In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT '19)*, 2019, pp. 49-58.
5. R. Agarwal, L. Melnick, N. Frosst, X. Zhang, B. Lengerich, R. Caruana, and G. E. Hinton, "Neural additive models: Interpretable machine learning with neural nets," In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021, pp. 4699-4711.
6. X. Hu, C. Rudin, and M. Seltzer, "Optimal sparse decision trees," In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019, pp. 7265-7273.
7. L. Semenova, C. Rudin, and R. Parr, "On the existence of simpler machine learning models," In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, 2022, pp. 1827-1858. doi: 10.1145/3531146.3533232
8. A. H. Karimi, B. Schölkopf, and I. Valera, "Algorithmic recourse: From counterfactual explanations to interventions," In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 2021, pp. 353-362.
9. S. Chiappa, "Path-specific counterfactual fairness," In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-19), 33(01), 7801-7808.*, 2019. doi: 10.1609/aaai.v33i01.33017801
10. Y. Wu, L. Zhang, X. Wu, and H. Tong, "PC-Fairness: A unified framework for measuring causality-based fairness," In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019, pp. 3399-3409.
11. H. Nilforoshan, J. D. Gaebler, R. Shroff, and S. Goel, "Causal conceptions of fairness and their consequences," In *Proceedings of the 39th International Conference on Machine Learning (ICML 2022), PMLR 162*, 2022, pp. 16848-16887.
12. E. Black, J. L. Koepke, P. Kim, S. Barocas, and M. Hsu, "Operationalizing the search for less discriminatory alternatives in fair lending," In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, 2024, pp. 1-15.
13. K. Lam, "A framework for assurance audits of algorithmic systems," In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24).*, 2024. doi: 10.1145/3630106.3658957
14. A. Kasirzadeh, and A. Smart, "The use and misuse of counterfactuals of ethical machine learning," In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 2021, pp. 228-238.
15. C. Oh, H. Won, J. So, T. Kim, Y. Kim, H. Choi, and K. Song, "Learning fair representation via distributional contrastive disentanglement," In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)*, 2022, pp. 1342-1351.

**Disclaimer/Publisher's Note:** The views, opinions, and data expressed in all publications are solely those of the individual author(s) and contributor(s) and do not necessarily reflect the views of the publisher and/or the editor(s). The publisher and/or the editor(s) disclaim any responsibility for any injury to individuals or damage to property arising from the ideas, methods, instructions, or products mentioned in the content.