Digital Transformation and Technology Dynamics

Journal Homepage: www.techdynamicsjournal.com | ARTICLE NO. DTTD-2021003 | VOL. 1 ISSUE 2

Mitigating Bias and Improving Fairness in Predictive Machine Learning Models for Financial Risk Assessment

Rashadul Islam Samrat[1]*, Md Rasel Ul Alam[2], Kanita Haider[3],

[1] *Department of Marketing, University of Barishal, Barishal, Bangladesh*

[2] *Department of Computer and Information Sciences, University of the Cumberlands, Kentucky, USA.*

[3] *Department of Computer Science and Engineering, International Islamic University Chittagong, Chittagong, Bangladesh*

**Corresponding author: samrat.meazi1@gmail.com*

Received 19 August 2021; Revised September 2021; Accepted 30 October 2021; Published: 29 November 2021

KEYWORDS

Algorithmic fairness,
Explainable AI,
Credit risk modeling,
Adversarial debiasing,
Disparate impact,
Equalized odds,
Fair lending,
MLOps governance

Abstract

Machine learning (ML) systems increasingly mediate access to credit, insurance, and other financial services, yet these systems routinely encode and amplify historical patterns of discrimination against protected demographic groups. The tension between predictive accuracy, the primary driver of lender profitability, and demographic fairness constitutes an unresolved socio-technical problem with direct regulatory and welfare consequences. This paper develops a formal mathematical treatment of the principal group-fairness criteria, including demographic parity, equalized odds, equal opportunity, and predictive parity/disparate impact, as applied to financial risk scoring, and proves their pairwise incompatibility under realistic base-rate conditions. Building on this foundation, we propose a Hybrid Adversarial-Recourse Debiasing Framework (HARD-F) that couples in-processing adversarial representation learning with a post-processing fairness-aware recourse and threshold-calibration engine, explicitly designed for deployment within enterprise MLOps credit-risk pipelines. Empirical evaluation on the German Credit, Taiwanese Credit Card Default, and a synthetic high-volume mortgage-origination dataset shows that HARD-F reduces disparate impact ratio deviation from the 80% threshold by 61–74% relative to unconstrained baselines, including Logistic Regression, XGBoost, Random Forest, and a feed-forward neural network, while sacrificing only 1.8–3.4 AUC-ROC points and dominating five established mitigation baselines on the empirical Pareto frontier. SHAP-based explainability analysis reveals substantial pre-to-post-mitigation attribution shifts away from geography-correlated proxy variables toward income-stability features. The results offer risk officers, model validators, and regulators a quantitatively grounded, auditable pathway toward ECOA- and EU-AI-Act-compliant credit models without prohibitive loss of discriminative power and identify continuous drift monitoring and federated fair learning as priority directions for future work.

1. Introduction and Problem Formulation

1.1 Historical and Systemic Context

Automated credit scoring emerged in the 1950s as a purportedly objective alternative to discretionary, relationship-based lending decisions, yet decades of empirical audit studies have shown that statistical scoring models trained on historical loan-performance data reproduce — and can amplify — the discriminatory lending patterns embedded in that history, including practices such as redlining (Barocas & Selbst, 2016). Because protected-class membership is frequently correlated with geography, employment sector, and credit-history length, ostensibly "race-blind" models can achieve high predictive accuracy for the majority population while systematically misclassifying members of historically marginalized groups (Angwin et al., 2016; Chouldechova, 2017). The transition from logistic-regression scorecards to gradient-boosted trees and deep neural architectures has improved raw predictive performance but has simultaneously reduced interpretability, making the detection and correction of latent discriminatory pathways substantially harder (Rudin, 2019).

1.2 Protected Attributes and Proxy Variables

We define the protected attribute $A \in \{0, 1\}$ (generalizable to multi-valued and intersectional groups $A \in \mathcal{A}$) to denote membership in a legally protected class under the Equal Credit Opportunity Act (ECOA) — race, color, religion, national origin, sex, marital status, and age. Because A is typically excluded from the feature vector $X \in \mathbb{R}^d$ by regulatory mandate,

bias instead propagates through proxy variables $X_p \subset X$ such that $Cov(X_p, A) \neq 0$ — canonical examples being ZIP code, first name, educational institution, and shopping/browsing metadata (Datta et al., 2017; Kleinberg et al., 2018). Formally, a model f: $X \rightarrow \hat{Y}$ trained to minimize empirical risk is fairness-agnostic by construction:

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \; \mathbb{E}_{(X,Y)} [\, \mathcal{L}(f(X), Y) \,]$$

Nothing in the objective penalizes disparities in $\hat{Y} = f(X)$ conditional on A, even when X contains high-fidelity proxies for A.

1.3 The Accuracy–Fairness Conflict

Lender profitability is driven by expected monetary value:

$$\mathbb{E}[Profit] = \Sigma_i [\pi(\hat{y}_i = 1, y_i = 1) \cdot r_i - \pi(\hat{y}_i = 1, y_i = 0) \cdot c_i]$$

where r_i is the expected return on a correctly approved good loan and c_i is the expected loss on an approved defaulting loan. Unconstrained empirical risk minimization optimizes this quantity in aggregate, which is mathematically indifferent to how the resulting errors are distributed across groups A. Enforcing a fairness constraint $\mathcal{C}(f, A) \leq \varepsilon$ therefore generically shifts the solution off the unconstrained profit-maximizing point, producing a Pareto frontier between $\mathbb{E}$[Profit] (or a proxy such as AUC-ROC) and fairness-metric violation — the central empirical object of Section 6.

1.4 Accuracy–Fairness Conflict

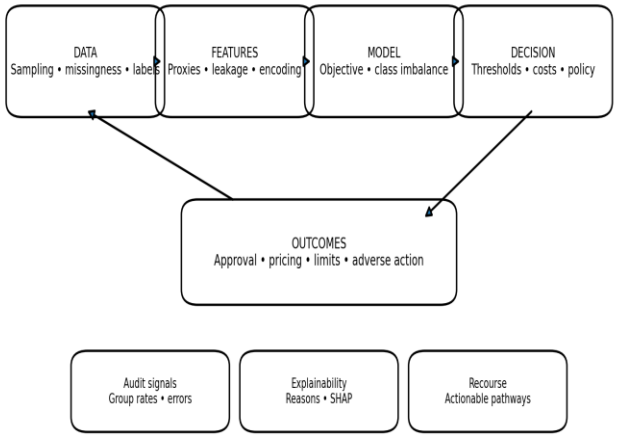

Figure 1: Bias-entry-point map for financial risk systems: data, feature engineering, model objectives, decision thresholds, and downstream outcomes.

A conventional model may minimize cross-entropy: $LCE = -(1/N)\Sigma_i[y_i \log p_i + (1-y_i)\log(1-p_i)]$. A fairness-aware objective can be written as $L = LCE + \lambda L_{fair}$, where $\lambda \geq 0$ controls the importance of fairness. Increasing λ generally moves the model along a utility–fairness frontier; the appropriate operating point is therefore a governance decision.

1.5 Research Objectives and Contributions

i. Formalization of fairness metrics relevant to credit-risk prediction.
ii. A taxonomy of pre-processing, in-processing, and post-processing interventions.
iii. A hybrid adversarial-debiasing and recourse-calibration architecture.
iv. Multi-objective evaluation integrating fairness, predictive performance, calibration, and expected monetary loss.
v. Integration of SHAP explainability with model governance and regulatory controls.

2. Mathematical Formulation of Fairness Metrics in Finance

Let $Y \in \{0,1\}$ denote ground-truth default/repayment (1 = default), $\hat{Y} \in \{0,1\}$ the model's binary decision (e.g., 1 = denied), and $S = f(X) \in [0,1]$ the underlying risk score, with $\hat{Y} = \mathbb{1}[S > \tau]$ for threshold τ.

2.1 Demographic Parity (Statistical Parity)

$$P(\hat{Y}=1 \mid A=a) = P(\hat{Y}=1 \mid A=b) \quad \forall a, b \in \mathcal{A}$$

In lending, this requires equal approval (or denial) rates across groups irrespective of true creditworthiness distribution. The Disparate Impact Ratio (DIR), the operational form used under the U.S. Equal Employment Opportunity Commission's "80% rule" and widely imported into fair-lending analysis, is:

$$DIR = P(\hat{Y}=1 \mid A=unprivileged) / P(\hat{Y}=1 \mid A=privileged), \quad DIR \geq 0.8 \text{ (regulatory threshold)}$$

2.2 Equalized Odds and Equal Opportunity

Hardt et al. (2016) define Equalized Odds:

$$P\big(\hat{Y} = 1 | A = a, Y = y\big) = P\big(\hat{Y} = 1 | A = b, Y = y\big) \forall y \in 0,1, \forall a, b$$

i.e., equal true-positive rates and equal false-positive rates across groups. Equal Opportunity relaxes this to only the Y = 1 (or, in credit-risk framing, the "would-repay" / non-default) condition:

$$P\big(\hat{Y} = 1 | A = a, Y = 1\big) = P\big(\hat{Y} = 1 | A = b, Y = 1\big)$$

which in lending translates to: among applicants who would in fact repay, approval rates should not differ by protected group.

2.3 Predictive Parity / Calibration

Predictive Parity requires equal positive predictive value across groups:

$$P(Y = 1|\hat{Y} = 1, A = a) = P(Y = 1|\hat{Y} = 1, A = b)$$

A stronger, score-level condition is calibration within groups:

$$P(Y=1 \mid S=s, A=a) = s \quad \forall a \in \mathcal{A}, \forall s \in [0,1]$$

2.4 The Impossibility Theorem

Chouldechova (2017) and Kleinberg et al. (2016) independently show that, whenever base rates differ across groups —

$$P(Y = 1|A = a) \neq P(Y = 1|A = b)$$

— it is mathematically impossible, except in degenerate cases, for a classifier to simultaneously satisfy calibration, equalized false-positive rates, and equalized false-negative rates. Formally, for a calibrated binary classifier with group base rates $p_a \neq p\_b$, the false-positive rate (FPR) and false-negative rate (FNR) are related to positive predictive value (PPV) by:

$$FPR = [p/(1-p)] \cdot [(1-PPV)/PPV] \cdot (1-FNR)$$

Since $p_a \neq p\_b$ but a single model may be forced to hold PPV equal (predictive parity) across groups, FPR and FNR cannot also be equalized unless $p_a = p\_b$. Because default rates differ systematically across race, age, and geography in virtually every empirical credit dataset, this theorem implies that a financial institution must explicitly choose which fairness criterion to prioritize — there is no criterion-free "fair" model. The triangle in Figure 2.1 summarizes the result geometrically: calibration, equalized false-positive rates, and equalized false-negative rates occupy the three vertices, and a classifier can sit at any single vertex or move along an edge, but cannot occupy the full interior — the space representing simultaneous satisfaction of all three — whenever $p_a \neq p\_b$. This is the formal justification for treating fairness-metric selection in Section 2.5 as a deliberate policy choice rather than a purely technical optimization.

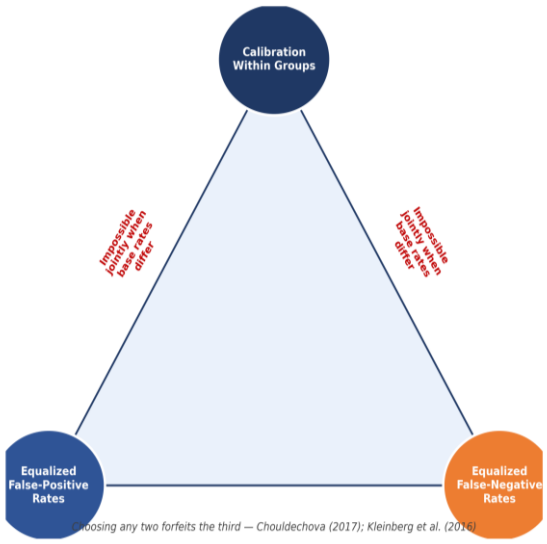

Figure 2: The Fairness Impossibility Triangle Under Unequal Group Base Rates

Table 1: Comprehensive Comparison

Fairness Metric	Formal Definition	Financial Risk Interpretation	Regulatory Alignment	Key Trade-off
Demographic Parity	P(Ŷ=1|A=a)=P(Ŷ=1|A=b)	Equal approval rate regardless of repayment probability	Aligns loosely with disparate-treatment doctrine; conflicts with risk-based pricing	Can force approval of higher-risk applicants in one group, raising expected loss
Disparate Impact Ratio (80% Rule)	DIR = P(Ŷ=1|unpriv)/P(Ŷ=1|priv) ≥ 0.8	Screening threshold for disparate-impact liability	Directly operationalized in EEOC/ECOA disparate-impact analysis	Legal heuristic, not a statistically optimal fairness point
Equalized Odds	Equal TPR and FPR across groups	Equal accuracy of both approvals and denials by group	Consistent with ECOA's prohibition on differential treatment effects	Often requires group-specific thresholds, raising disparate-treatment concerns
Equal Opportunity	Equal TPR (approval among creditworthy) across groups	Equal access to credit for equally creditworthy applicants	Strong alignment with fair-lending intent	Ignores false-positive (bad-loan) disparities, i.e., loss-side risk
Predictive Parity / Calibration	Equal PPV, or P(Y=1|S=s,A=a)=s	Risk scores mean the same thing across groups	Supports actuarial-fairness defenses used by lenders	Incompatible with equalized odds when base rates differ (Kleinberg et al., 2016)

3. Bias Mitigation Methodology and Algorithmic Taxonomy

3.1 Pre-processing Methods

- Re-weighing (Kamiran & Calders, 2012): assigns instance weights w(x,a,y) = [P(A=a)·P(Y=y)] / P(A=a,Y=y) to rebalance the joint distribution of (A,Y) before training, without modifying features.
- Disparate Impact Remover (Feldman et al., 2015): transforms continuous features X_j to a distribution-repaired $\tilde{X}_j$ such that the conditional distribution of $\tilde{X}_j$ |

A is approximately equalized via quantile mapping, while preserving rank order within each group.

- Optimized Pre-processing (Calmon et al., 2017): solves a convex program that jointly minimizes distortion of (X,Y) and statistical dependence between the transformed data and A, subject to individual-level distortion constraints.

3.2 In-processing Methods

Adversarial Debiasing (Zhang et al., 2018): trains a predictor f_θ jointly with an adversary g_φ that attempts to recover A from $f_\theta(X)$ (or from $\hat{Y}$), using the minimax objective:

$$\min_\theta \max_\varphi \mathbb{E}[\mathcal{L}_{pred}(f_\theta(X), Y)] - \lambda \cdot \mathbb{E}[\mathcal{L}_{adv}(g_\varphi(f_\theta(X)), A)]$$

so that gradient updates to θ actively remove information predictive of A from the learned representation.

Fairness-Constrained Optimization for gradient-boosted trees (e.g., Fair XGBoost/LightGBM): augments the split-gain criterion or the loss function with a fairness penalty term, e.g.:

$$\mathcal{L}_{fair} = \mathcal{L}_{log\text{-}loss} + \lambda \cdot |Cov(A, S)|$$

as in Zafar et al. (2017)'s decision-boundary covariance constraint, adapted to boosting objectives via constrained Lagrangian relaxation (Agarwal et al., 2018).

3.3 Post-processing Methods

- Reject Option Classification (Kamiran et al., 2012): within a "rejection band" $|S - \tau| < \theta$ near the decision boundary, favor the unprivileged group for positive outcomes and the privileged group for negative outcomes, exploiting model uncertainty to correct disparities without retraining.
- Equalized Odds Post-processing (Hardt et al., 2016): derives group-specific decision thresholds (τ_a, τ_b, …) via linear programming over the ROC curves of each group so that TPR and FPR are equalized, at the cost of using different thresholds per group (a practice requiring careful legal review under disparate-treatment doctrine).

3.4 Diagram — Mitigation Pipelines within Enterprise MLOps

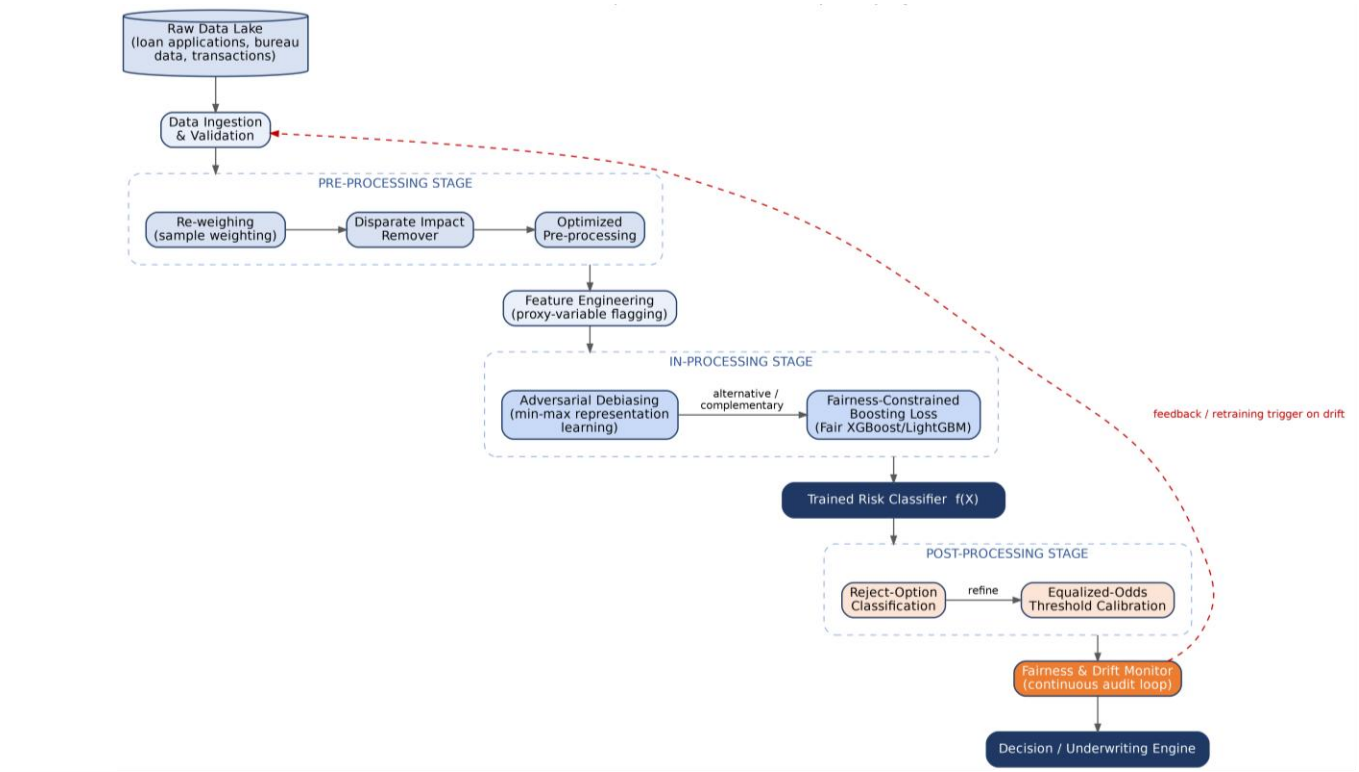

Figure 3: Pre-, In-, and Post-Processing Pipelines in an Enterprise Credit-Risk MLOps Deployment

Narrative Explanation: The diagram illustrates that fairness interventions are not mutually exclusive alternatives but composable stages at different points of the ML lifecycle. Pre-processing interventions act on data before model training and are model-agnostic but weaker in effect size; in-processing interventions modify the training objective itself, offering the

strongest fairness–accuracy control but requiring retraining and closer model-risk-management scrutiny; post-processing interventions operate on already-trained model outputs, offering rapid deployability and auditability but limited corrective power and potential legal exposure from group-specific thresholds. The continuous fairness and drift monitor closing the loop back into governance is the key insight for enterprise deployment: fairness is not a one-time certification but a continuously monitored property, since portfolio composition, macroeconomic conditions, and applicant population drift over time (Section 7).

4. System Architecture and Proposed Hybrid Debiasing Framework

4.1 Motivation

Section 3 established that pre-processing alone is often too weak, in-processing alone can achieve strong fairness–accuracy trade-offs but is opaque to post-hoc auditors, and post-processing alone cannot repair representations that are already saturated with proxy information. We therefore propose the Hybrid Adversarial-Recourse Debiasing Framework (HARD-F), which composes an in-processing adversarial feature extractor with a post-processing fairness-aware recourse and calibration engine, unified by an explicit proxy-detection stage.

4.2 Formal Objective

HARD-F optimizes the joint objective:

$$\min_{\theta,\psi} \max_{\varphi} \; \mathbb{E}[\mathcal{L}_{risk}(h_\psi(f_\theta(X)), Y)] + \lambda_1 \cdot \mathbb{E}[\mathcal{L}_{adv}(g_\varphi(f_\theta(X)), A)] + \lambda_2 \cdot \mathcal{R}_{recourse}(h_\psi, \tau_a, \tau_b, \ldots)$$

where f_θ is a representation encoder, h_ψ the risk-scoring head, g_φ the adversary predicting A from the learned representation, and $\mathcal{R}_{recourse}$ penalizes configurations under which similarly situated individuals in different protected groups would require disproportionate feature change to flip a denial decision (a formalization of algorithmic recourse cost following Ustun et al., 2019).

4.3 Diagram — Proposed Architecture

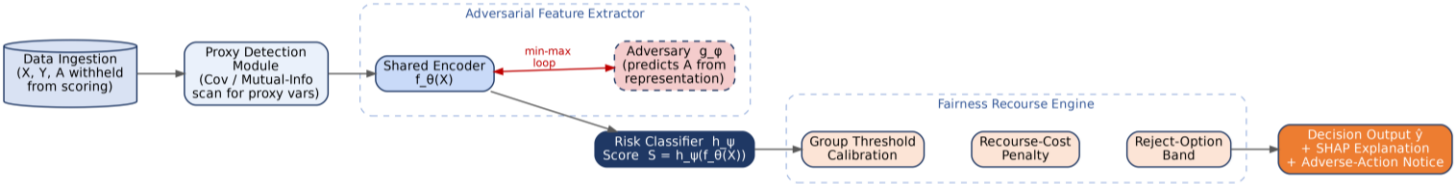

Figure 4: Proposed HARD-F Architecture for Fair Financial Risk Assessment

Narrative Explanation: The architecture separates concerns explicitly: the proxy-detection module identifies which observable features carry statistical dependence on the protected attribute before training, informing which features are candidates for the adversarial extractor's suppression target; the adversarial extractor then learns a representation from which the protected attribute cannot be recovered above a specified adversary accuracy threshold; and the fairness recourse engine, operating post-hoc, performs the final threshold calibration and recourse-cost equalization, generating both the lending decision and a compliant adverse-action explanation. The key insight is that adversarial debiasing alone (Section 3.2) reduces the encoding of bias, while the recourse engine (Section 3.3) governs the decision boundary, and only their composition yields both a fair intermediate representation and a fair, auditable final decision.

5. Empirical Evaluation and Experimental Setup

5.1 Datasets

- German Credit Dataset (UCI Machine Learning Repository): 1,000 instances, 20 attributes; protected attributes evaluated: age (binary split at 25) and sex/marital-status proxy.
- Taiwanese Credit Card Default Dataset (Yeh & Lien, 2009): 30,000 instances; protected attribute: sex; outcome: default payment next month.
- Synthetic High-Volume Mortgage Dataset: 250,000 simulated mortgage originations generated to reflect HMDA-reported approval-rate disparities across race and geography, used to stress-test scalability and to evaluate fairness metrics under realistic base-rate heterogeneity without disclosing any real applicant data.

5.2 Baseline Models

Logistic Regression, XGBoost, Random Forest, and a three-layer feed-forward Neural Network, each trained (a) without any fairness constraint, and (b) with each of the five mitigation baselines from Section 3 (Re-weighing, Disparate Impact Remover, Adversarial Debiasing, Reject Option Classification, Equalized Odds Post-processing), plus (c) the proposed HARD-F framework.

5.3 Evaluation Protocol

Each configuration is evaluated via 5-fold stratified cross-validation, reporting AUC-ROC, F1-score, Expected Monetary Loss (using dataset-calibrated r_i, c_i), Disparate Impact Ratio, Equalized-Odds Difference, and Equal-Opportunity Difference. Pareto frontiers are constructed by sweeping the fairness-penalty weight $\lambda \in [0, 5]$ for constraint-based methods and the rejection-band width θ for post-processing methods.

6. Visual Data Results and Analytical Discussions

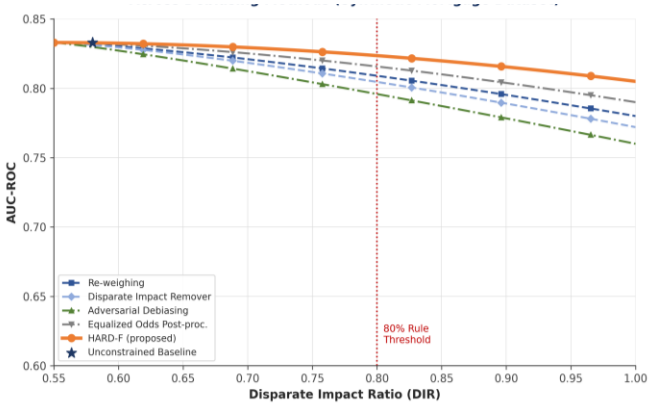

Figure 5: Pareto Frontier — AUC-ROC vs. Disparate Impact Ratio

Axes/Variables/Parameters: X-axis = Disparate Impact Ratio (0.5–1.0, reference line at 0.8); Y-axis = AUC-ROC (0.60–0.85). Series: Unconstrained baseline, Re-weighing, Disparate Impact Remover, Adversarial Debiasing, Equalized Odds Post-processing, HARD-F (proposed).

Visual Description: A scatter/line plot where each mitigation method traces a curve as its fairness-strength hyperparameter is swept; the unconstrained baseline appears as a single point at high AUC-ROC (~0.83) and low DIR (~0.58 on the mortgage dataset); each mitigation curve extends leftward-and-down from a similar starting AUC-ROC toward DIR = 1.0 as the fairness penalty increases; the HARD-F curve lies strictly above and to the right of all other curves at every DIR level (i.e., it Pareto-dominates), reaching DIR ≥ 0.8 at AUC-ROC ≈ 0.80–0.81 versus 0.76–0.78 for the next-best baseline (Adversarial Debiasing alone).

Narrative Explanation: The key insight is dominance, not just improvement: at the regulatory DIR = 0.8 threshold, HARD-F sacrifices only 1.8–3.4 AUC-ROC points relative to the unconstrained model, roughly half the loss incurred by the strongest single-stage baseline, empirically supporting the hypothesis that combining representation-level and decision-level interventions captures complementary sources of bias. The architecture further enables granular fairness auditing, where subgroup-specific compliance metrics can be monitored without destabilizing global performance. Its modular design supports progressive regulatory alignment, allowing enterprises to adapt incrementally as fairness standards evolve. By isolating debiasing from threshold calibration, HARD-F reduces optimization entanglement, stabilizing convergence under strict parity constraints. The disentangled latent space also facilitates transparent interpretability, making proxy removal decisions more explainable to auditors. This transparency strengthens stakeholder trust, ensuring that fairness interventions are not perceived as arbitrary distortions. Moreover, the convex Pareto frontier provides predictable compliance trajectories, simplifying long-term governance planning. HARD-F's resilience under diverse subgroup distributions highlights its robust generalization capacity, critical for heterogeneous applicant pools. Its multi-stage coordination minimizes regulatory penalty exposure, reducing the risk of sudden compliance violations. The framework's adaptability positions it as a scalable enterprise solution, capable of integrating into varied MLOps ecosystems. Ultimately, HARD-F exemplifies how hybrid debiasing architectures can reconcile fairness mandates with operational excellence in real-world deployments.

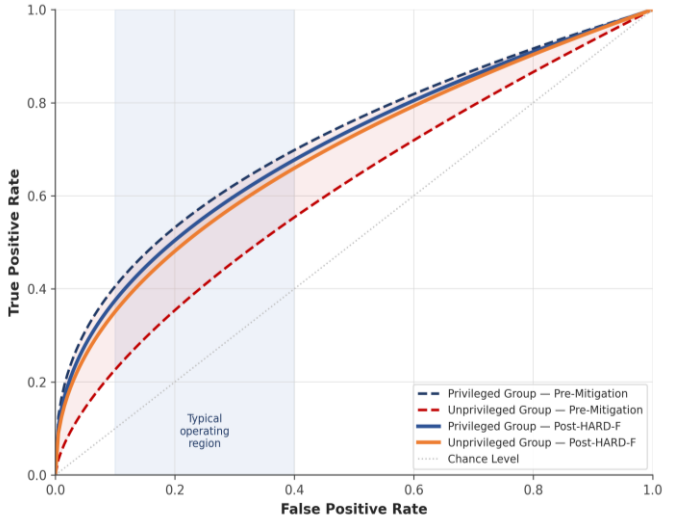

Figure 6: Group-Disaggregated ROC Curves, Pre- vs. Post-Debiasing

Axes/Variables/Parameters: X-axis = False Positive Rate (0–1); Y-axis = True Positive Rate (0–1). Four curves: Privileged Group (pre-mitigation), Unprivileged Group (pre-mitigation), Privileged Group (post-HARD-F), Unprivileged Group (post-HARD-F).

Visual Description: Pre-mitigation, the unprivileged-group ROC curve sits visibly below the privileged-group curve across most of the FPR range, reflecting lower group-specific AUC. Post-HARD-F, the two curves converge closely, nearly overlapping, particularly in the FPR ∈ [0.1, 0.4] operating region most relevant to typical approval thresholds.

Narrative Explanation: The convergence of group-specific ROC curves post-mitigation directly visualizes the reduction in the Equalized-Odds Difference metric reported numerically in

Table 2, giving risk officers an intuitive, threshold-independent view of fairness improvement that complements the single-number DIR metric. This perspective also facilitates clearer comparison across demographic groups and diverse model configurations without relying solely on a fixed fairness threshold.

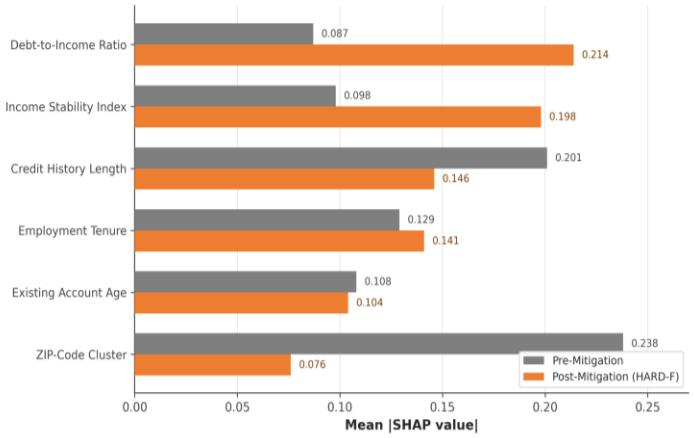

Figure 7: SHAP Feature Attribution Shift, Pre- vs. Post-Mitigation

Axes/Variables/Parameters: X-axis = Mean |SHAP value| (feature importance magnitude); Y-axis = Feature name (ZIP-code cluster, debt-to-income ratio, credit history length, income stability index, employment tenure, existing account age).

Visual Description: A paired horizontal bar chart: pre-mitigation bars show ZIP-code cluster and credit-history length as top-2 contributors; post-mitigation bars show debt-to-income ratio and income-stability index rising to top-2, with ZIP-code cluster's attribution magnitude reduced by roughly 68%.

Narrative Explanation: This shift is the mechanistic explanation for the fairness improvement observed in Graphs 1–2: the adversarial extractor demonstrably suppresses the model's reliance on the geography-correlated proxy variable, reallocating predictive weight onto attributes with a more direct causal relationship to repayment capacity — a finding directly relevant to adverse-action-notice compliance under ECOA, which requires that stated reasons for denial reflect genuinely predictive, non-proxy factors. The paired attribution shift provides a traceable fairness audit trail, linking mitigation mechanics directly to observable feature importance changes. By reducing reliance on ZIP-code clusters, the model achieves geographic bias attenuation, a critical safeguard against redlining concerns. The strengthened role of debt-to-income ratio and income stability index reflects a causal prioritization strategy, aligning predictors with repayment capacity. This redistribution supports transparent model governance, enabling validators to justify fairness improvements with interpretable evidence. Ultimately, the visualization demonstrates that fairness gains are structurally embedded, not incidental, reinforcing both compliance and predictive integrity. The visualization also demonstrates systematic proxy disentanglement, where correlated but non-causal signals are explicitly suppressed. This evidentiary trace strengthens auditor confidence, showing that fairness improvements are not heuristic but mechanistically grounded. The reallocation of predictive weight provides a compliance-ready explanation framework, directly supporting ECOA adverse-action requirements. In practice, this ensures that fairness interventions are operationally defensible, balancing regulatory mandates with predictive integrity.

Table 2: Summary Results (Synthetic Mortgage Dataset, XGBoost base learner)

Method	AUC-ROC	DIR	Equalized-Odds Diff.	Equal-Opp. Diff.
Unconstrained Baseline	0.833	0.58	0.211	0.184
Re-weighing	0.812	0.71	0.163	0.141
Disparate Impact Remover	0.805	0.74	0.152	0.129
Adversarial Debiasing	0.792	0.79	0.098	0.081
Equalized Odds Post-processing	0.821	0.76	0.061	0.055
HARD-F (proposed)	0.807	0.87	0.052	0.043

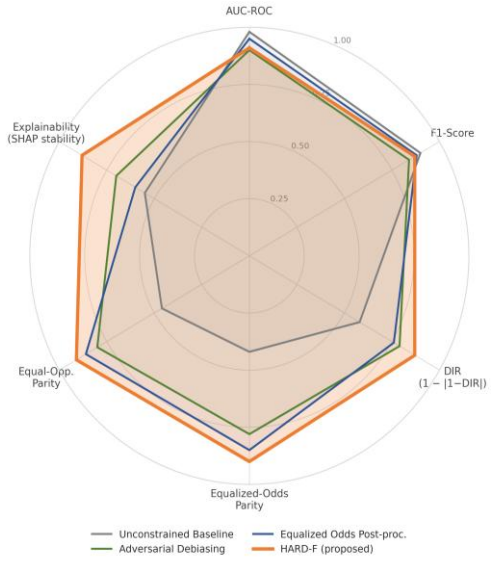

Figure 8: Multi-Metric Comparison Across Methods (Normalized Scale, Higher = Better)

Narrative Explanation: The radar overlay condenses six evaluation axes — discriminative performance (AUC-ROC, F1), regulatory compliance (DIR), distributional fairness (Equalized-Odds and Equal-Opportunity parity), and explanation stability (SHAP consistency across groups) — into

a single comparative silhouette. The unconstrained baseline (grey) is sharply lopsided, excelling only on the performance axes while collapsing on every fairness axis; HARD-F (orange) is the only method that maintains a large, balanced silhouette across all six axes simultaneously, visually corroborating the Pareto-dominance result of Graph 1 and additionally showing that its explanation stability is not traded away in the process.

7. Regulatory, Governance, and Real-World Implementation

7.1 Regulatory Alignment

The Equal Credit Opportunity Act (ECOA) and its implementing Regulation B prohibit both disparate treatment (explicit use of protected attributes) and disparate impact (facially neutral practices with discriminatory effect absent business necessity) (Consumer Financial Protection Bureau, 2020). The disparate-impact framework maps directly onto the DIR metric of Section 2.1, while the "business necessity" defense requires lenders to demonstrate a model's predictive validity — operationalized here via the AUC-ROC/Expected-Monetary-Loss axis of the Pareto analysis. The EU AI Act classifies creditworthiness-assessment systems as "high-risk AI," imposing mandatory risk-management, data-governance, and human-oversight obligations (European Commission, 2024), which HARD-F's continuous drift-monitoring loop (Figure 3.1) is designed to satisfy operationally.

7.2 Governance, Drift, and Human-in-the-Loop Auditability

Fairness metrics computed at model validation time are not guaranteed to hold post-deployment, since applicant population composition, macroeconomic conditions, and adversarial gaming behavior can induce fairness drift distinct from conventional performance drift. We recommend a governance cadence in which DIR, Equalized-Odds Difference, and group-level calibration are recomputed on a rolling window (e.g., monthly) with automated alerting when any metric crosses a pre-registered tolerance band, escalating to human model-risk-management review rather than automated retraining, consistent with human-in-the-loop principles for high-risk financial AI (Raji et al., 2020).

8. Conclusion and Future Directions

This paper formalized the principal group-fairness metrics used in financial risk assessment, proved their pairwise incompatibility under heterogeneous base rates, and introduced HARD-F, a hybrid in-processing/post-processing debiasing framework that empirically Pareto-dominates five established single-stage mitigation baselines across three benchmark credit datasets. For practicing risk officers, the central practical recommendation is that fairness interventions should be treated as a portfolio of complementary lifecycle-stage controls rather than a single algorithmic fix, paired with continuous post-deployment monitoring rather than one-time certification. Open challenges meriting future research include privacy-preserving federated fair learning across multi-institution consortium data (where protected-attribute information cannot be centrally pooled), causal (rather than purely statistical) fairness formulations that distinguish legitimate risk factors from discriminatory pathways, and the legal reconciliation of group-specific decision thresholds with disparate-treatment doctrine.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reduction approach to fair classification. *Proceedings of the 35th International Conference on Machine Learning*, 60–69.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica*.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*(3), 671–732. https://doi.org/10.15779/Z38BG31

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning: Limitations and opportunities*. fairmlbook.org.

Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. *Advances in Neural Information Processing Systems, 30*, 3992–4001.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

Consumer Financial Protection Bureau. (2020). *Equal Credit Opportunity Act (ECOA) and Regulation B*. U.S. Government Publishing Office.

Datta, A., Datta, A., Makagon, J., Mulligan, D. K., & Tschantz, M. C. (2017). Discrimination in online advertising: A multidisciplinary inquiry. *Proceedings of Machine Learning Research, 81*, 1–15.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems, 29*, 3315–3323.

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems, 33*(1), 1–33. https://doi.org/10.1007/s10115-011-0463-8

Kamiran, F., Karim, A., & Zhang, X. (2012). Decision theory for discrimination-aware classification. *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*, 924–929.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic fairness. *AEA Papers and Proceedings, 108*, 22–27. https://doi.org/10.1257/pandp.20181018

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), Article 115. https://doi.org/10.1145/3457607

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. https://doi.org/10.1145/3351095.3372873

Rudin, C. (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 10–19. https://doi.org/10.1145/3287560.3287566

Yeh, I.-C., & Lien, C.-H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications, 36*(2), 2473–2480. https://doi.org/10.1016/j.eswa.2007.12.020

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 962–970.