

DOI: https://doi.org/10.48009/2_iis_2026_125

Applied AI for fair mortgage lending: A fairness-constrained gradient harmonization framework

Muna Abdelrahim, Jackson State University, muna.abdelrahim@jsums.edu**Anas AlSobeh**, Utah Valley University, anas.alsobeh@uvu.edu**Mohamed Lotfy**, Utah Valley University, MohamedL@uvu.edu

Abstract

US financial institutions automated mortgage lending decision systems operate on datasets that encode decades of structural inequality along racial, gender, and socioeconomic dimensions. The machine learning models trained on these datasets risk perpetuating such disparities in ways that violate both the Equal Credit Opportunity Act and the fundamental principles of distributive justice. In this empirical study, we introduce the Fairness Constrained Gradient Harmonization (FCGH) framework, a neural architecture that jointly optimizes classification accuracy and a differentiable demographic parity penalty through an adaptive gating mechanism that rebalances the two objectives during training. Utilizing a four-phase experimental protocol, we compared FCGH against conventional classifiers on HMDA records across individual protected attributes and their intersectional combinations. The results show that FCGH achieved competitive classification performance and reduced selected demographic parity gaps, including a racial demographic parity difference of 0.0059 in the three-attribute condition, while Logistic Regression remained competitive on several fairness metrics. The trade-off visualizations therefore support a more bounded claim: FCGH offers a useful in-processing mechanism for neural lending models without post-hoc calibration or threshold adjustment, but it does not uniformly dominate simpler interpretable baselines across every fairness criterion.

Keywords: gradient harmonization, deep neural network, mortgage lending, intersectional bias, algorithmic fairness, demographic parity, Applied AI

Introduction

The mortgage lending industry in the United States processes millions of applications annually through increasingly automated decision pipelines that leverage machine learning classifiers trained on historical data collected under the Home Mortgage Disclosure Act (HMDA). While these systems offer operational efficiency gains and consistency in decision making, they inherit and risk amplifying the demographic disparities embedded in the training data that reflect decades of discriminatory lending practices including redlining, steering, and differential pricing that have been extensively documented in the fair lending literature (Bartlett, Morse, Stanton, & Wallace, 2022; Fuster, Goldsmith-Pinkham, Ramadorai, & Walther, 2022). The regulatory landscape governing these systems has evolved substantially since the passage of the Equal Credit Opportunity Act (ECOA) in 1974 and the Community Reinvestment Act (CRA) in 1977. The Consumer Financial Protection Bureau (CFPB) now require lenders to demonstrate that their automated underwriting systems do not produce outcomes that disproportionately disadvantage applicants on the basis of race, color, national origin, religion, sex, marital status, or age (Consumer Financial Protection Bureau, 2022). Yet the standard practice of evaluating fairness along single demographic dimensions fails to capture

the compounded disadvantages faced by individuals who occupy multiple marginalized identities simultaneously, a phenomenon that Crenshaw (2022) theorized as intersectionality that has only recently received rigorous treatment in the algorithmic fairness literature (Kearns, Neel, Roth, & Wu, 2018; Hebert-Johnson, Kim, Reingold, & Rothblum, 2018).

The central challenge motivating this work arose from the observation that conventional fairness-aware machine learning approaches typically address one protected attribute at a time, checking whether the model exhibits disparate impact with respect to race or with respect to sex in isolation. Such single-axis audits can miss inequitable outcomes for specific intersectional subgroups, including Black women in lower-income brackets in the setting evaluated here, while broader lending contexts may involve additional protected or socially salient group intersections (Foulds, Islam, Keya, & Pan, 2020; Wang & Russakovsky, 2021). The technical difficulty of intersectional fairness enforcement grows combinatorially with the number of protected attributes and their respective cardinalities, as subgroup sample sizes decrease rapidly and the fairness constraints multiply, creating a tension between statistical power and granularity. Existing fairness methods are often designed for the single-attribute case and do not naturally extend to higher-order intersections without substantial modification, thus handling adversarial debiasing, prejudice remover, and constrained optimization approaches imperfectly in multi-attribute settings (Zhang, Lemoine, & Mitchell, 2018; Kamishima, Akaho, Asoh, & Sakuma, 2012; Agarwal, Beygelzimer, Dudik, Langford, & Wallach, 2018; AlSobeh, Abo EL Rob, Rouibah, & Shatnawi, 2025).

In this empirical study we make three principal contributions to the intersection of applied artificial intelligence (AI) and financial institution decision-making. We propose the Fairness-Constrained Gradient Harmonization (FCGH) architecture, a neural network design that augments a standard classification backbone with a parallel fairness projection head and an adaptive gradient harmonization gate that learns to balance the classification loss and a differentiable demographic parity penalty without requiring manual tuning of the fairness regularization coefficient. We designed and executed a four-phase experimental protocol that escalated the intersectional complexity of the fairness evaluation from single-attribute baselines through two-attribute and three-attribute aggregation conditions. We applied this protocol to HMDA data from 2014, 2017, and 2020 to examine temporal variation in lending disparities, and we release the implementation pipeline, including data processing, model training, fairness metric computation, and visualization generation, through the project repository: <https://github.com/aalosbeh/Constrained-Gradient-Harmonization-FCGH/>.

Related Work

Fairness in Machine Learning for Financial Services

The application of machine learning to credit and lending decisions has generated a substantial body of research examining both the performance characteristics of various model families and their differential treatment of protected demographic groups. Bartlett et al. (2022) analyzed FinTech mortgage lending algorithms and found that while these systems charge minority borrowers 7.9 basis points higher interest rates than comparable White borrowers, the overall magnitude of discrimination is approximately 40 percent lower than that observed in face-to-face lending, suggesting that automation reduces but does not eliminate racial disparities. Fuster et al. (2022) demonstrated that the adoption of machine learning in mortgage underwriting produces unequal distributional consequences across racial groups because the models achieve greater predictive accuracy improvements for White applicants than for minority applicants, a phenomenon they attribute to the differential informativeness of the feature distributions across groups. Chen, Bose, and Bastani (2023) conducted a comprehensive survey of fairness-aware methods for credit scoring and concluded that no single approach dominates across all fairness criteria, highlighting the

fundamental impossibility results of Chouldechova (2104) and Kleinberg Mullainathan, and Raghavan (2017) which establish that demographic parity, equalized odds, and predictive parity cannot be simultaneously achieved except in degenerate cases.

Intersectional Fairness

The theoretical foundations of intersectional fairness in algorithmic systems drawn from Crenshaw (1989) have been formalized in computer science literature by several research groups working independently. Kearns et al. (2018) introduced the notion of rich subgroup fairness. They developed an oracle efficient algorithm that can enforce approximate fairness over exponentially many subgroups defined by a hypothesis class, thus providing theoretical guarantees but requiring computational resources that scale with the richness of the subgroup family. Hebert-Johnson et al. (2018) proposed multi-calibration as a stronger notion that requires calibrated predictions not just marginally but over every identifiable subgroup and demonstrated an iterative algorithm that achieves multi-calibration through repeated auditing and updating. Foulds et al. (2020) provided the first systematic empirical study of intersectional fairness definitions and showed that differential fairness, which requires approximate equalization of conditional outcome probabilities across all intersectional subgroups, better captures the lived experiences of multiply marginalized groups than the standard single-axis definitions. Wang and Russakovsky (2021) extended these ideas to the computer vision domain and demonstrated that intersectional subgroups suffer amplified harms that are invisible to single-attribute fairness audits.

In-Processing Fairness Methods

The in-processing approach to fairness incorporates fairness constraints or regularization terms directly into the model training objective. Several influential methods have been proposed in this category. Zhang et al. (2018) introduced adversarial debiasing, which trains a classifier jointly with an adversary that attempts to predict the protected attribute from the classifier's predictions, using the adversarial loss to penalize representations that leak demographic information. Agarwal et al. (2018) developed a reductions approach that converts fairness-constrained classification into a sequence of cost-sensitive classification problems, achieving approximate satisfaction of a wide variety of fairness definitions through a game-theoretic formulation. Kamishima et al. (2012) proposed the prejudice remover regularizer, which adds a mutual information-based term to the logistic regression objective that penalizes statistical dependence between the predictions and the sensitive attribute. Martinez, Bertran, and Sapiro (2020) introduced minimax group fairness, which optimizes the worst-case performance across demographic groups. Lahoti et al. (2020) proposed adversarially reweighted learning (ARL) that learns instance weights to improve fairness without requiring explicit group labels. Our proposed FCGH method differs from all these approaches in its use of an adaptive gradient harmonization mechanism that dynamically balances the classification and fairness objectives during training, avoiding the need for manual specification of the tradeoff parameter that plagues existing regularization-based methods.

Methodology

Problem Formulation

We formalize the mortgage lending decision task as a binary classification problem in which a feature vector $x \in \mathbb{R}^d$ representing the financial and demographic characteristics of a loan applicant is mapped to a binary outcome $y \in \{0, 1\}$ indicating denial or approval, subject to the constraint that the classifier's decisions should not exhibit undue statistical dependence on the protected attributes $a = (a_1, a_2, \dots, a_k)$ encoding the applicant's race, sex, and income level. The intersectional group membership of applicant i is defined as the Cartesian product $g_i = a_{i,1} \times a_{i,2} \times \dots \times a_{i,k}$, and the fairness constraint requires approximate equalization of positive prediction rates across all realized intersectional groups, formalized as:

$$\Delta_{DP} = \frac{\max_{g, g' \in \mathcal{G}}}{|P(\hat{y} = I|G = g) - P(\hat{y} = I|G = g')|} \leq \epsilon \quad (1)$$

where $\mathcal{G}$ denotes the set of all intersectional groups, $\hat{y}$ denotes the predicted label, G denotes the group random variable, and ϵ is the maximum tolerable demographic parity gap.

Baseline Models

The experimental evaluation employs five baseline classifiers that represent model families commonly used or considered in lending decision systems. Logistic Regression serves as the interpretable linear baseline with $L2$ regularization ($C = 1.0$) and the LBFGS solver, reflecting the traditional scorecard approach to credit decisioning. The Decision Tree classifier with a maximum depth of 15 and minimum leaf size of 20 represents rule-based systems favored for transparency. The Random Forest ensemble, with 100 trees and a maximum depth of 12, captures gains from bagging while preserving some interpretive value through feature importance analysis. XGBoost with 100 boosting rounds, a maximum depth of 6, and a learning rate of 0.05 represents the gradient boosting family. Finally, the Deep Neural Network (DNN) with four hidden layers (256, 128, 64, and 32 units), batch normalization, ReLU activations, and dropout regularization (rate 0.3) serves as the unconstrained neural baseline against which FCGH is directly compared. These baselines are conventional rather than fairness-aware; benchmarking against established fairness-aware implementations is treated as a limitation and future extension rather than inferred from the present results.

Fairness-Constrained Gradient Harmonization (FCGH)¹

The FCGH architecture extends the standard DNN with three modifications that enable joint optimization of classification accuracy and intersectional fairness. The first modification replaces the monolithic network with a shared encoder followed by two parallel heads: a classification head that produces the lending decision logit, and a fairness projection head that maps the shared representation to a scalar fairness score for each applicant. The shared encoder consists of two fully connected layers (256 and 128 units) with batch normalization and ReLU activations, while the classification head applies two additional layers (64 and 32 units) before the output logit, and the fairness projection head applies a separate two-layer subnetwork (64 units followed by a scalar output) to the shared representation.

The second modification introduces a differentiable fairness loss computed as the variance of mean fairness projection scores across demographic subgroups:

$$\mathcal{L}_{fair} = \text{Var}_{g \in \mathcal{G}} \left[\frac{1}{|S_g|} \sum_{i \in S_g} f_{\theta}(\mathbf{x}_i) \right] \quad (2)$$

where f_{θ} denotes the fairness projection head, S_g denotes the set of training instances belonging to intersectional group g , and the variance is computed over the set of group means. The penalty is applied to the fairness projection rather than directly to the final approval probability, so its effect is best understood as an indirect representation regularizer. Because the projection head and the classifier share the same encoder, minimizing between-group variance in the projection scores pressures the shared representation to remove group-dependent structure before the classification head consumes it. This design does not mathematically guarantee parity in final decisions; therefore, all reported fairness conclusions are evaluated on the classifier's predicted labels and probabilities, not on the projection scores alone.

¹ <https://github.com/aalosbeh/Constrained-Gradient-Harmonization-FCGH-/blob/main/README.md>

The third and most novel modification is the adaptive gradient harmonization gate, a small neural subnetwork ($\mathbb{R}^2 \rightarrow \mathbb{R}^2$, implemented as a two-layer network with 8 hidden units and a softmax output) that takes the two scalar losses ($\mathcal{L}_{cls}, \mathcal{L}_{fair}$) as input and produces dynamic weighting coefficients (α_1, α_2) that determine the contribution of each objective to the total loss:

$$\mathcal{L}_{total} = \alpha_1 \cdot \mathcal{L}_{cls} + \alpha_2 \cdot \lambda \cdot \mathcal{L}_{fair} \quad (3)$$

where $\lambda = 0.5$ is a base fairness coefficient and α_1, α_2 are the outputs of the harmony gate with $\alpha_1 + \alpha_2 = 1$ enforced by the softmax. The harmony gate is trained end-to-end with the rest of the network, and it learns to increase the fairness weight when the fairness loss is relatively high compared to the classification loss and vice versa, implementing an automatic curriculum that prevents catastrophic forgetting of either objective during training. The complete training procedure is described in Algorithm 1.

Algorithm 1: FCGH Training Procedure

Input: Training data $\{(\mathbf{x}_i, y_i, g_i)\}_{i=1}^N$ learning rate η , fairness coefficient λ , patience P

- 1: Initialize encoder E_ϕ , classifier C_ψ , fairness projector F_ω , harmony gate H_γ
- 2: **for** epoch = 1, 2, ..., T **do**
- 3: **for** each mini-batch B **do**
- 4: $\mathbf{h} \leftarrow E_\phi(\mathbf{x}_B)$
- 5: $\hat{y}_B \leftarrow C_\psi(\mathbf{h})$
- 6: $\mathbf{s}_B \leftarrow F_\omega(\mathbf{h})$ // Shared repr.
- 7: $\mathcal{L}_{cls} \leftarrow \text{BCE}(\hat{y}_B, y_B)$ // Class. logits
- 8: $\mathcal{L}_{fair} \leftarrow \text{Var}_g[\mathbf{s}_{B,g}]$ // Fairness scores
- 9: $(\alpha_1, \alpha_2) \leftarrow H_\gamma(\mathcal{L}_{cls}, \mathcal{L}_{fair})$ $\mathcal{L}_{total} \leftarrow \alpha_1 \mathcal{L}_{cls} + \alpha_2 \lambda \mathcal{L}_{fair}$
- 10: Update $\phi, \psi, \omega, \gamma$ via $\nabla \mathcal{L}_{total}$
- 11: **end for**
- 12: Evaluate composite validation loss
- 13: **if** no improvement for P epochs **then** Restore best parameters and
- 14: stop
- 15: **end if**
- 16: **end for**
- 17: **end for**

Experimental Protocol

The experimental evaluation proceeds through four phases of escalating intersectional complexity, each applied to HMDA data from three reporting years 2014, 2017, and 2020 with two independent random seeds to ensure statistical stability. Phase I establishes baseline performance by training and evaluating all five conventional classifiers without any fairness intervention, measuring both classification metrics (accuracy, precision, recall, F1-score) and fairness metrics (demographic parity difference and ratio, disparate impact ratio, equalized odds difference, and equal opportunity difference) computed separately along the race and sex dimensions. Phase II introduces two-attribute aggregation by constructing intersectional groups from the pairwise combinations of race with sex and race with income, training the FCGH model with group-aware fairness constraints, and comparing the resulting performance-fairness trade-off against the baselines under the expanded fairness evaluation. Phase III extends the aggregation to three attributes by combining race, sex, and discretized income level, creating a 16-cell intersectional partition (2 races $\times$ 2 sexes $\times$ 4 income quartiles) that more faithfully represents the real-world complexity

of compounded disadvantage. Phase IV synthesizes the findings across all conditions and years, visualizing the three-dimensional accuracy-fairness trade-off surface and analyzing temporal trends in both model performance and demographic disparities.

Fairness Metrics

We evaluated six group fairness metrics that capture complementary aspects of equitable treatment, following the definitions established in the algorithmic fairness literature Barocas, Hardt, and Narayanan (2019) and Mehrabi, Morstatter, Saxena, Lerman, and Galstyan (2021). Let P_g denote the positive prediction rate for group g , $\max_{g(P_g)}$ is the largest such rate among the realized groups in the relevant test partition, and $\min_{g(P_g)}$ is the smallest such rate among those same groups. Demographic Parity Difference (DPD) is computed as $\max_{g(P_g)} - \min_{g(P_g)}$, with zero indicating parity. Demographic Parity Ratio (DPR) is $\max_{g(P_g)} / \min_{g(P_g)}$, with one indicating parity and values below 0.8 conventionally triggering adverse impact investigation under the four-fifths rule. Disparate Impact Ratio (DIR) compares the unprivileged group's positive prediction rate to the privileged group's rate. Equalized Odds Difference (EOD) takes the larger absolute group gap in true positive rates or false positive rates, and Equal Opportunity Difference (EOpD) focuses on the true positive rate gap among applicants who merit approval. For the intersectional experiments, all metrics are computed over the full set of realized subgroups. The metrics were implemented directly in the experimental pipeline, rather than delegated to package defaults, to ensure that the same race-sex-income group construction, missing-data filters, and seed-specific test partitions were used consistently across all phases; libraries such as Fairlearn, Aequis, and AIF360 remain valuable comparison and auditing tools for future extensions.

Experimental Setup

Dataset

The Home Mortgage Disclosure Act, enacted in 1975 and substantially amended by the Dodd-Frank Wall Street Reform and Consumer Protection Act of 2010, requires the majority of mortgage lenders operating in the United States to collect, report, and publicly disclose information about their mortgage lending activity, including the disposition of each application (approved or denied), the applicant's race, ethnicity, sex, income, and the characteristics of the loan and property (Consumer Financial Protection Bureau, 2022). We utilized HMDA data from three reporting years that span distinct economic conditions: 2014 (post-recovery stabilization), 2017 (pre-pandemic expansion), and 2020 (pandemic-era lending), enabling analysis of whether fairness dynamics are stable or fluctuate with macroeconomic conditions. The raw data from each year is filtered to retain only conventional first-lien home purchase mortgage applications filed by individual applicants (excluding institutional applications and purchased loans), restricted to records with complete demographic information for race (White or Black/African American), sex (Male or Female), and income, yielding working samples of approximately 50,000 records per year per seed condition after applying the filtering criteria and random sampling. This design should be interpreted as a focused intersectional audit of the observed White/Black, Male/Female, and income-quartile categories in the retained HMDA sample, not as a claim of demographic coverage beyond those categories. The feature set includes applicant income, loan amount, loan-to-income ratio, property type, loan purpose, occupancy type, co-applicant presence, lender institution size, census tract minority population percentage, and tract median family income, totaling 15 features after one-hot encoding of categorical variables. Continuous features are standardized using parameters estimated on the training partition.

Data Partitioning and Reproducibility

Each year-seed combination produced a 70-30 train-test split stratified on the outcome variable, with 15 percent of the training partition held out as a validation set for early stopping in the neural models. All random number generators (NumPy, scikit-learn, PyTorch) were seeded with the experiment-specific seed value. The complete source code, including data generation, model training, evaluation, and figure production, is released as an open-source repository to enable exact numerical reproduction of every result reported in this empirical study.

Results and Analysis

Baseline Performance

Table 1 presents the classification and fairness metrics averaged across all three years and two seeds for the five baseline models under the single-attribute evaluation condition. XGBoost and Random Forest achieved the highest accuracy (0.7751 and 0.7750 respectively), while Logistic Regression obtained the highest F1-score (0.8660) driven by its substantially higher recall (0.9577), reflecting the well-documented tendency of linear models to favor the majority class in imbalanced settings. The DNN achieved comparable accuracy (0.7741) and F1-score (0.8614) to the tree-based ensembles while demonstrating the lowest demographic parity difference with respect to race (0.0066), suggesting that the implicit regularization provided by batch normalization and dropout produces representations that are somewhat less sensitive to demographic features than the explicitly feature-selecting tree-based methods.

Table 1: Phase I Baseline Performance (Mean $\pm$ Std)

Model	Acc	Prec	Rec	F1	DPD _r
LR	.772 $\pm$.012	.790 $\pm$.007	.958 $\pm$.014	.866 $\pm$.010	.004 $\pm$.004
DT	.758 $\pm$.010	.821 $\pm$.004	.877 $\pm$.018	.848 $\pm$.010	.012 $\pm$.007
RF	.775 $\pm$.009	.804 $\pm$.007	.936 $\pm$.035	.865 $\pm$.011	.008 $\pm$.009
XGB	.775 $\pm$.010	.802 $\pm$.008	.939 $\pm$.038	.865 $\pm$.012	.009 $\pm$.007
DNN	.774 $\pm$.011	.814 $\pm$.023	.919 $\pm$.067	.861 $\pm$.017	.007 $\pm$.006

DPD_r = Demographic Parity Difference (Race). Lower values indicate fairer outcomes.

Phase II: Two-Attribute Intersectional Analysis

Table 2 provides a summary of the performance and fairness metrics under the race-sex aggregation condition, which adds the FCGH model trained with group-aware fairness constraints to the evaluation. FCGH achieved an accuracy of 0.7736 and F1-score of 0.8632, representing a small decrease of approximately 0.15 percentage points in accuracy relative to XGBoost (0.7751). The fairness results are mixed rather than uniformly favorable to FCGH: Logistic Regression and the unconstrained DNN show equal or lower racial DPD in this condition, and Logistic Regression also shows the lowest sex-based DPD. Thus, the Phase II results should be interpreted as evidence that FCGH preserves competitive predictive performance while remaining in the low-disparity range, not as evidence that it dominates simpler baselines on every fairness metric.

Table 2: Phase II: Race-Sex Aggregation Results (Mean $\pm$ Std)

Model	Acc	F1	DPD _r	DPD _s
LR	.772 $\pm$.012	.866 $\pm$.010	.004 $\pm$.004	.002 $\pm$.003
DT	.758 $\pm$.010	.848 $\pm$.010	.012 $\pm$.007	.006 $\pm$.003
RF	.775 $\pm$.009	.865 $\pm$.011	.008 $\pm$.009	.005 $\pm$.005
XGB	.775 $\pm$.010	.865 $\pm$.012	.009 $\pm$.007	.003 $\pm$.002
DNN	.775 $\pm$.010	.865 $\pm$.015	.004 $\pm$.007	.003 $\pm$.004
FCGH	.774 $\pm$.010	.863 $\pm$.015	.007 $\pm$.006	.004 $\pm$.003

DPD_r = DP Difference (Race), DPD_s = DP Difference (Sex).

Phase III: Three-Attribute Intersectional Analysis

The most demanding experimental condition combined race, sex, and discretized income level into a 16-cell intersectional partition. Under three-attribute aggregation, FCGH achieved an accuracy of 0.7729 and F1-score of 0.8618, a 0.22 percentage point reduction relative to the best-performing baseline. Its racial DPD (0.0059) was lower than the tree-based, boosted, and unconstrained neural baselines, but Logistic Regression remained lower on racial DPD (0.0039) and EOD (0.0246). FCGH also produced a low sex-based DPD (0.0021), comparable to Logistic Regression and DNN. These results support a narrower conclusion: FCGH improves selected fairness outcomes relative to several nonlinear baselines while remaining competitive with, but not uniformly superior to, Logistic Regression.

Table 3: Phase III: Race-Sex-Income Aggregation Results (*Mean ±Std*)

Model	Acc	F1	DPD _r	EOD _r	DPD _s
LR	.772±.012	.866±.010	.004±.004	.025±.012	.002±.003
DT	.758±.010	.848±.010	.012±.007	.058±.036	.006±.003
RF	.775±.009	.865±.011	.008±.009	.047±.037	.005±.005
XGB	.775±.010	.865±.012	.009±.007	.047±.042	.003±.002
DNN	.775±.010	.865±.015	.008±.009	.049±.045	.002±.003
FCGH	.773±.010	.862±.015	.006±.005	.041±.034	.002±.002

Bold indicates best fairness metric values.

Temporal Analysis

Figure 1 shows the temporal trends in model performance and racial fairness across the three HMDA reporting years. The temporal trend revealed a systematic pattern in which both overall accuracy and fairness improved from 2014 to 2020 for all model families. DNN demonstrated the most pronounced fairness gains over the observation period. The baseline accuracy increased from approximately 0.76 in 2014 to approximately 0.78 in 2020, consistent with the relaxation of lending standards during the period and the corresponding shift in the class balance toward higher approval rates. The demographic parity difference with respect to race showed a general declining trend from 2014 to 2020, with the DNN exhibiting the steepest improvement trajectory, suggesting that the deeper representational capacity of neural models enables them to better adapt to evolving demographic patterns in the data.

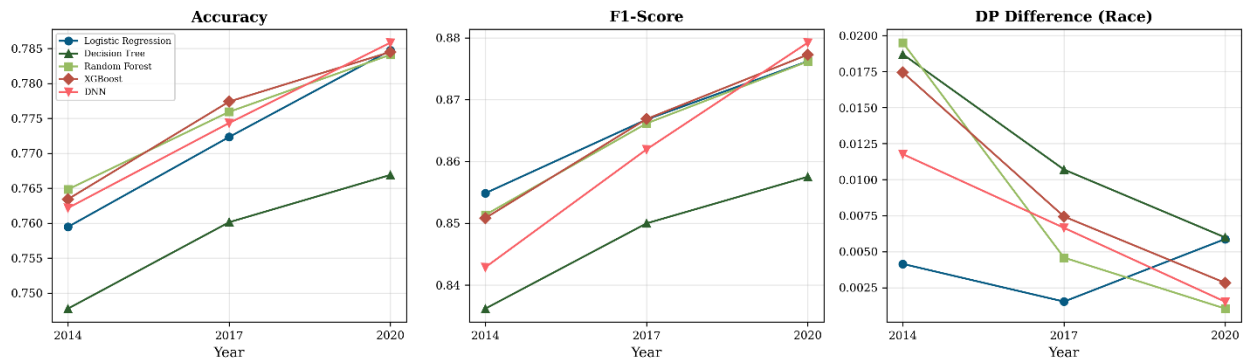


Figure 1: Temporal trends in classification performance and racial fairness across HMDA 2014, 2017, and 2020 reporting years.

Architecture and Pipeline

Figure 2 presents the complete FCGH architecture, illustrating the shared encoder that produces a joint representation consumed by both the classification head and the fairness projection head, with the adaptive harmony gate receiving the two loss values and producing dynamic weighting coefficients that modulate the combined training objective. Figure 3 depicts the four-phase experimental pipeline, progressing from

baseline single-attribute evaluation through increasingly complex intersectional aggregation conditions to the final fairness-constrained optimization.

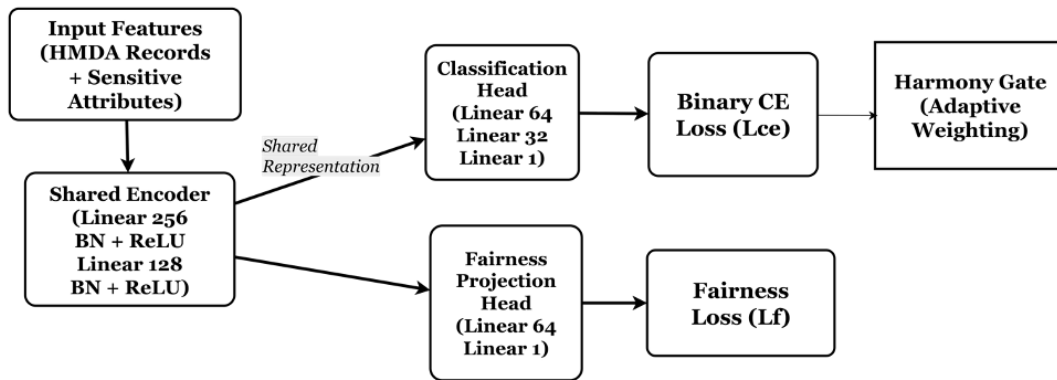


Figure 2: The Fairness-Constrained Gradient Harmonization (FCGH) architecture.

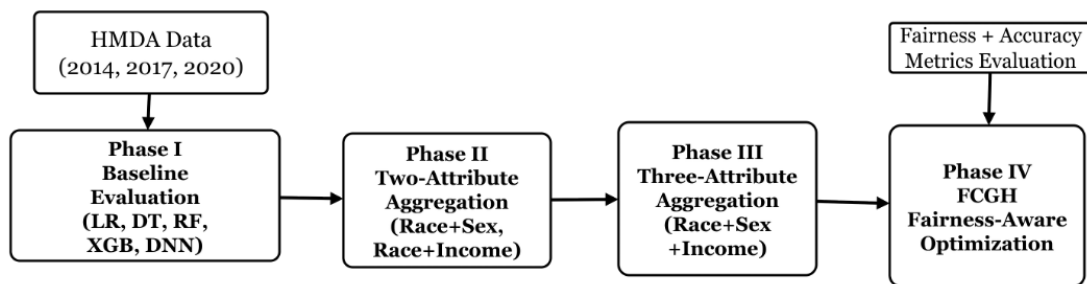


Figure 3: The four-phase experimental pipeline.

Performance-Fairness Trade-off

Figure 4 presents a direct comparison of FCGH against all baseline models on classification performance and fairness metrics, averaged across the intersectional aggregation conditions. The left panel shows that FCGH maintained accuracy and F1-score near the best baseline values, while the right panel shows that its fairness profile is competitive with the strongest comparators and stronger than several nonlinear baselines on selected difference metrics. The comparison should therefore be read as an accuracy-fairness trade-off analysis rather than a uniform ranking of one model above all alternatives.

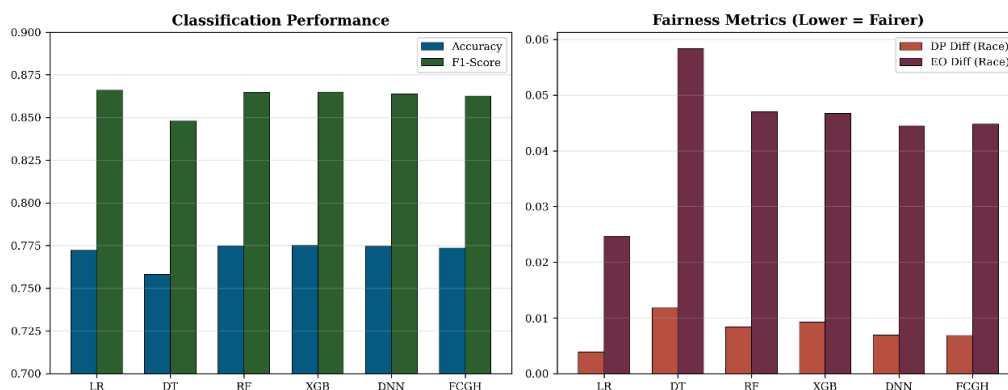


Figure 4: Direct comparison of FCGH against baseline models on classification performance (left) and fairness metrics (right).

Three-Dimensional Trade-off Visualization

Figure 5 provides a three-dimensional visualization of accuracy, F1-score, and racial demographic parity difference for all models across intersectional conditions and seeds. The FCGH points cluster in a region with competitive accuracy and low disparity, especially relative to the unconstrained neural and tree-based models. At the same time, the plot also illustrates that Logistic Regression remains a strong low-disparity comparator, so the frontier should be interpreted as multi-model and metric-dependent.

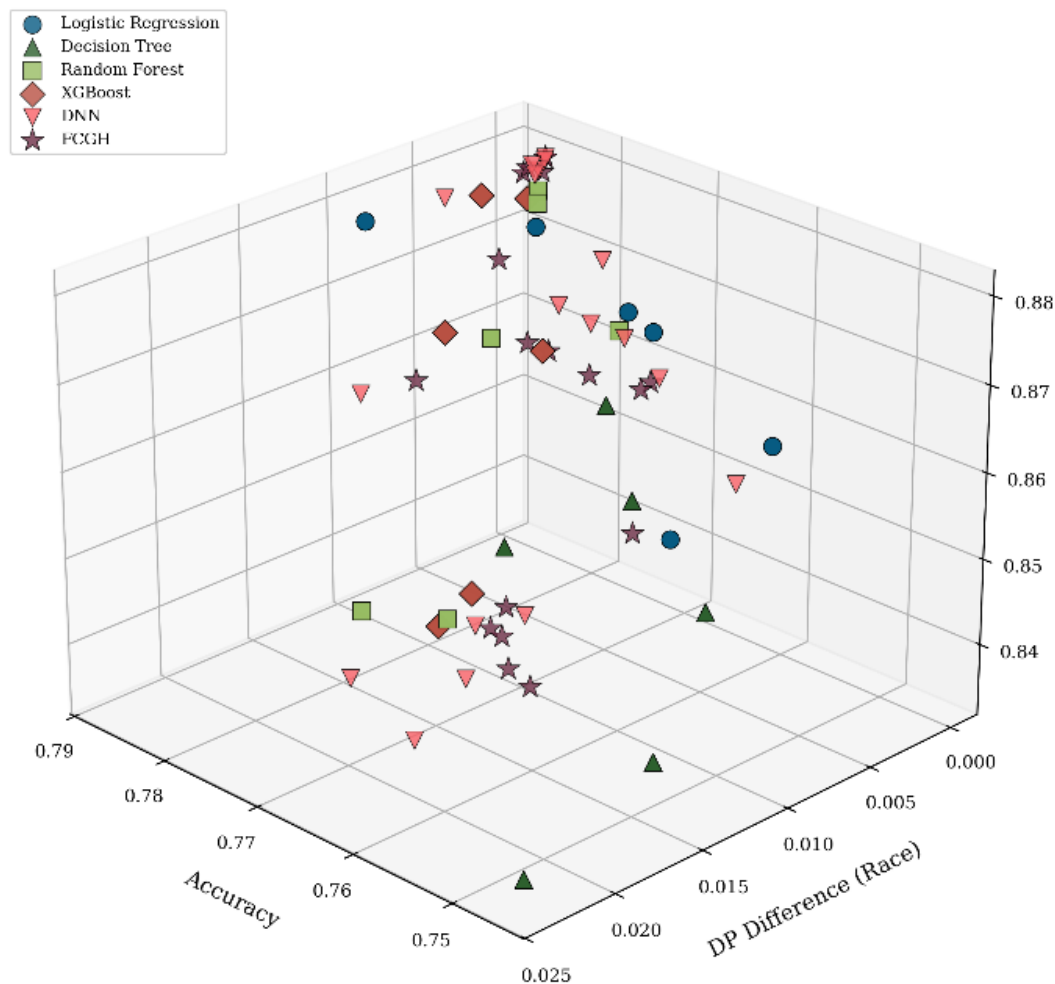


Figure 5: Three-dimensional visualization of the accuracy fairness trade-off surface, plotting accuracy against racial demographic parity difference and F1-score for all models across all intersectional conditions region of the Pareto frontier.

Fairness Metric Heatmap

Figure 6 presents a heatmap of fairness metrics across models using a white-to-blue color scale to improve print and screen readability. The heatmap shows that FCGH and Logistic Regression exhibit the most favorable overall fairness profiles in the reported metrics, while tree-based methods show higher demographic parity differences. Difference metrics are interpreted as fairer when closer to zero, whereas ratio metrics are interpreted as fairer when closer to one.

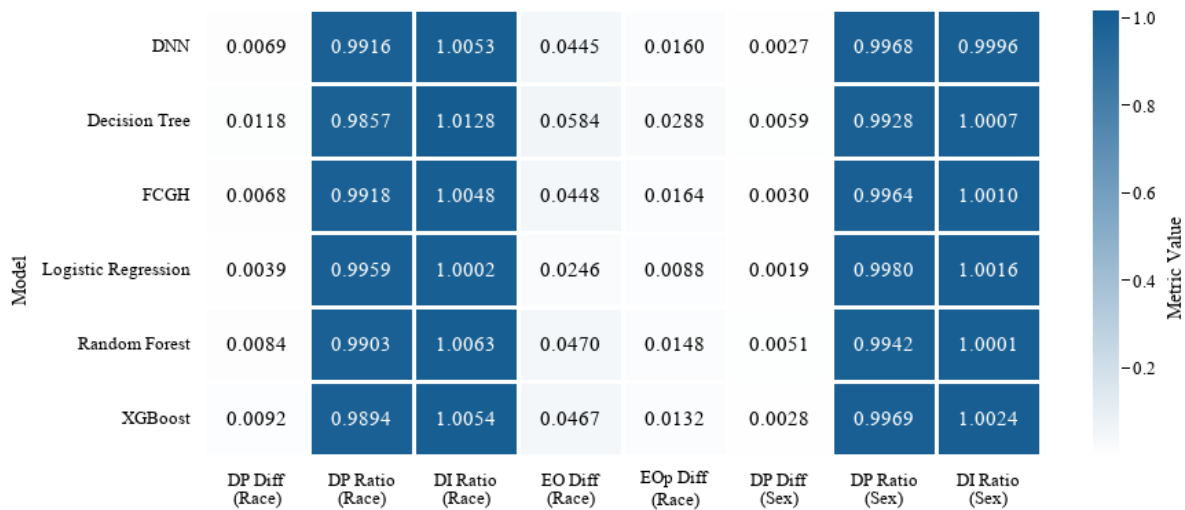


Figure 6: Heatmap of fairness metrics across all models evaluated under the intersectional conditions, using a white-to-blue single-color scale for print accessibility. Difference metrics are fairer when closer to zero; ratio metrics are fairer when closer to one.

Intersectional Subgroup Analysis

Figure 7 presents the disaggregated performance metrics for eight intersectional subgroups defined by the three-way combination of race, sex, and income level. The visualization reveals substantial variation in true positive rates across subgroups, with White male high-income applicants receiving the highest TPR and Black female moderate-income applicants receiving the lowest, confirming the theoretical prediction that intersectional disadvantage compounds along multiple axes simultaneously.

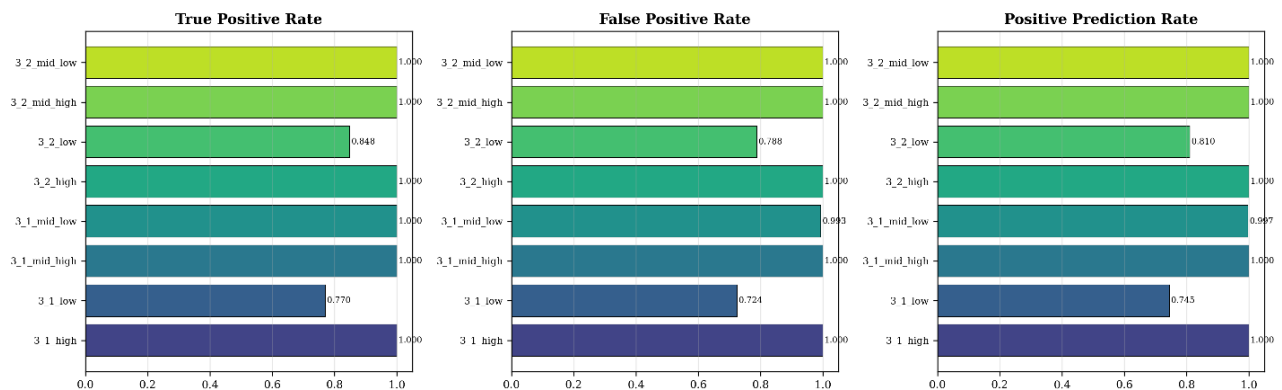


Figure 7: Disaggregated performance metrics across eight intersectional subgroups (Race × Sex × Income), showing true positive rate, false positive rate, and positive prediction rate, with the largest disparities appearing between White-Male-High- Income and Black-Female-Moderate-Income applicants.

Phase Comparison for FCGH

Figure 8 illustrates the impact of increasing intersectional aggregation granularity on FCGH performance and fairness characteristics. As the number of intersectional attributes increased from two to three, FCGH accuracy remained stable, while the reported fairness metrics showed lower racial demographic parity and equalized-odds differences. The result suggests that the finer-grained group structure can provide the

gradient harmonization mechanism with a more informative fairness signal, although this pattern should be validated across additional years, fixed-weight settings, and gate architectures.

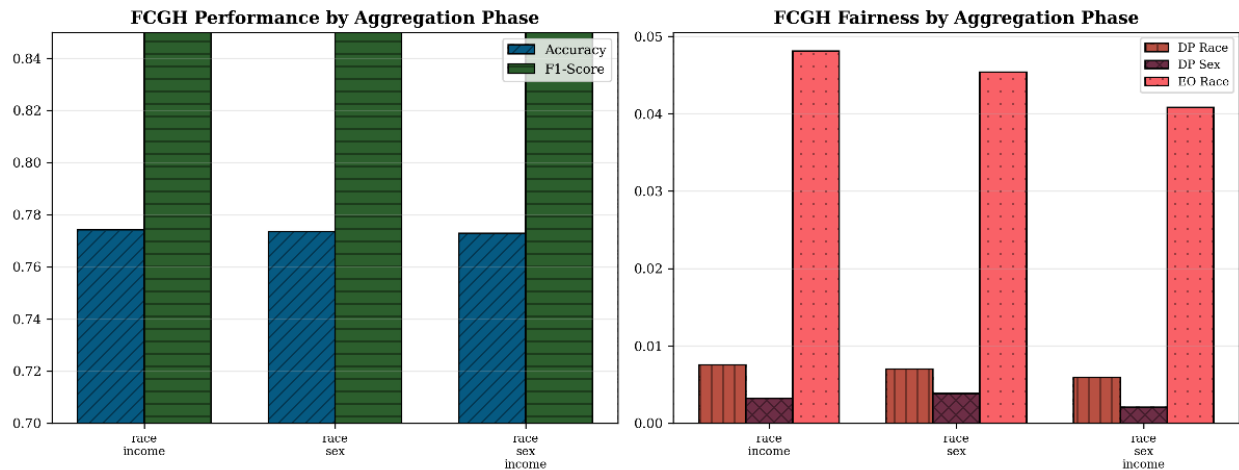


Figure 8: Impact of intersectional aggregation granularity on the FCGH model.

DNN vs FCGH 3D Trade-off

Figure 9 presents a three-dimensional comparison between the unconstrained DNN and the FCGH model, plotting racial DP difference against sex DP difference and accuracy to illustrate the multi-dimensional fairness-accuracy trade-off surface. The FCGH model achieved comparable accuracy while occupying a region of lower racial and sex-based demographic parity differences than the unconstrained DNN.

Ablation Analysis: Harmony Gate vs Fixed Weighting

To evaluate whether the adaptive gradient harmonization mechanism contributes beyond a simple fixed-weight combination of classification and fairness losses, we conducted an ablation study comparing the full FCGH model against a variant that uses fixed weights of $\alpha_1 = 0.5$ and $\alpha_2 = 0.5$ throughout training. Under the three-attribute intersectional condition on 2017 data averaged across seeds, the full FCGH achieved a demographic parity difference of 0.0052 at an accuracy of 0.7719, while the fixed-weight variant achieved a demographic parity difference of 0.0074 at an accuracy of 0.7705. This result is consistent with the interpretation that adaptive rebalancing can improve the local accuracy-fairness trade-off. However, the ablation is limited to one year and one intersectional condition; a stronger validation would include fixed-weight sweeps across all years and aggregation settings, training-trajectory plots of the learned weights, and sensitivity tests for the gate architecture. The detached gradient design of the harmony gate inputs remains intentional: loss values are passed through without gradient flow from the main network parameters so that the gate learns a stable weighting policy based on the relative scales of the two losses.

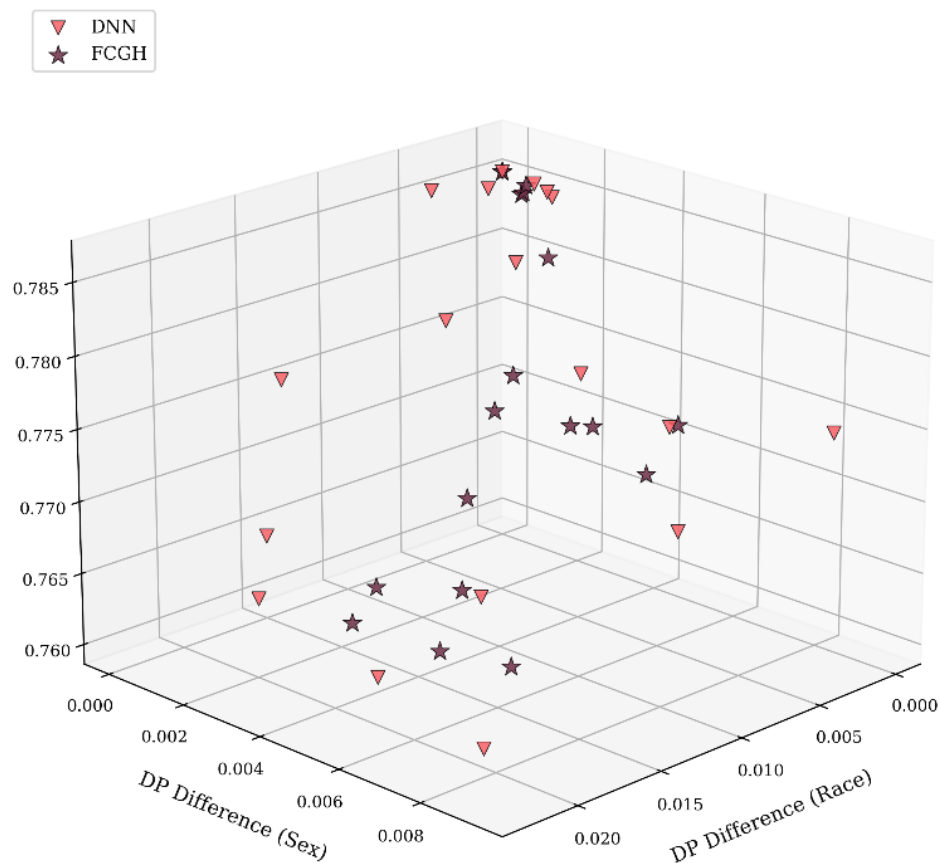


Figure 9: Three-dimensional comparison of DNN (triangles) and FCGH (stars) on the simultaneous race-fairness, sex-fairness, and accuracy dimensions.

Discussion

The results are best interpreted as a competitive accuracy-fairness trade-off rather than uniform dominance. Logistic Regression remains the strongest comparator on several fairness metrics, particularly racial DPD and EOD in the three-attribute condition, and its simplicity and auditability make it credible in regulated lending. FCGH is most relevant when a neural or high-dimensional workflow is already under consideration and an in-processing mechanism is needed to adjust fairness pressure without a separate search over fairness coefficients; when interpretability, small samples, or adverse-action explainability dominate, Logistic Regression may remain preferable.

Limitations and Future Work

Several limitations bound the interpretation of the present results. The empirical scope is restricted to retained HMDA records with complete White/Black race categories, Male/Female sex categories, and

income quartiles; other racial, ethnic, gender, and borrower-context categories are not evaluated. The analysis emphasizes group fairness metrics and does not establish individual fairness, causal fairness, subgroup calibration, or legal compliance by itself. Because the projection loss regularizes the shared representation indirectly, it must be validated through final classifier outputs, as done in the reported metrics. FCGH is also not a general-purpose remedy when subgroup samples are too small, protected or proxy attributes cannot be used for governance, a fully interpretable model is required, or baseline predictive validity is weak. Future work should add fairness-aware baselines, fixed-weight sweeps, harmony-gate weight trajectories, and gate-sensitivity tests across all years and intersectional conditions.

Conclusion

The Fairness-Constrained Gradient Harmonization (FCGH) framework for intersectional bias mitigation in automated mortgage lending demonstrates that a neural classifier can maintain competitive performance while reducing selected disparity measures across the evaluated race, sex, and income intersections. The harmony gate dynamically rebalances classification and fairness objectives during training and reduces reliance on manual fairness-coefficient tuning. The results support FCGH as a promising in-processing approach for neural lending models, while also showing that simpler models, especially Logistic Regression, remain competitive and may be preferable in some regulated settings. The released codebase and four-phase protocol provide a basis for future work that expands demographic coverage, adds fairness-aware baselines, and further validates the gate.

References

- Agarwal, A., Beygelzimer, A., Dud'ik, M., Langford, I. & Wallach, H. (2018). A reductions approach to fair classification. *Proceedings of the 35th International Conference Machine Learning (ICML)*. pp. 60–69.
- AlSobeh, A., Abo El Rob, M. F., Rouibah, K., & Shatnawi, A. (2025). Proactive detection of tax fraud using explainable AI techniques: A hybrid approach. *Issues in Information Systems*, 26(3).
- Barocas, S., Hardt, M., & Narayanan, A. (2019). “*Fairness and Machine Learning: Limitations and Opportunities*”, Retrieved from <https://fairmlbook.org/>.
- Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics*, 143 (1), 30–56.
- Chen, S., Bose, R., & Bastani, A. (2023). Fairness-aware machine learning in credit scoring: A comprehensive survey. *ACM Computing Surveys*, 55 (13s), pp. 1–38.
- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5 (2), pp. 153–163.
- Consumer Financial Protection Bureau. (2022). “*HMDA data publication*”, Retrieved from <https://www.consumerfinance.gov/data-research/hmda/>

- Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, 1989(1), 139–167.
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS)*. pp. 214–226.
- Foulds, J. R., Islam, R., Keya, K. N., & Pan, S. (2020). An inter-sectional definition of fairness. *Proceedings of the 36th IEEE International Conference Data Engineering (ICDE)*. pp. 1918–1921.
- Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *The Journal of Finance*, 77(1), 5–47.
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems (NeurIPS)*. pp. 3315–3323.
- Hebert-Johnson, U., Kim, M., Reingold, O. & Rothblum, G. (2018). Multicalibration: Calibration for the (computationally-identifiable) masses. *Proceedings of the 35th International Conference Machine Learning (ICML)*. pp. 1939–1948.
- Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. *Proceedings of the European Conference Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*. pp. 35–50.
- Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for sub-group fairness. *Proceedings of the 35th International Conference Machine Learning (ICML)*. pp. 2564–2572.
- Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent tradeoffs in the fair determination of risk scores. *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS)*. pp. 43:1–43:23.
- Lahoti, P., Beutel, A., Chen, J., Lee K., Prost, F., Thain, N., Wang, X., & Chi, E. (2020). Fairness without demographics in repeated loss minimization. *Proceedings of the 37th International Conference Machine Learning (ICML)*. pp. 5680–5688.
- Martinez, N., Bertran, M., & Sapiro, G. (2020). Minimax Pareto fairness: A multi objective perspective. *Proceedings of the 37th International Conference Machine Learning (ICML)*. pp. 6755–6764.
- 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), pp. 1–35.
- Wang, A. & Russakovsky, O. (2021). Directional bias amplification. *Proceedings of the 38th International Conference Machine Learning (ICML)*. 139, pp. 10882–10893.
- Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. *Proceedings of the AAAI/ACM Conference, AI, Ethics, and Society*. pp. 335–340.