STAR MOUNTAIN
INTERNATIONAL PUBLISHING GROUP PET.LTD

Article

Algorithmic Fairness in AI-Driven Loan Origination through Bias Mitigation Across Demographic Groups

Aisha Patel[1, *], Lukas Schmidt[2], María Elena Fernández[3]

[1]School of Computer Science, University of Birmingham, Birmingham, B15 2TT, United Kingdom; a.patel@cs.bham.ac.uk

[2]Department of Finance, Vienna University of Economics and Business, Vienna, 1020, Austria; lukas.schmidt@wu.ac.at

[3]Escuela de Ingeniería, Pontificia Universidad Católica de Chile, Santiago, 7820436, Chile; mfernandez@ing.puc.cl

***Correspondence:** Aisha Patel. Email: a.patel@cs.bham.ac.uk

Abstract

The adoption of artificial intelligence in credit underwriting has transformed loan origination, yet these systems can inherit and amplify biases that disproportionately affect protected demographic groups. This article provides a critical analysis of algorithmic fairness in AI-driven loan origination, examining bias sources across the machine learning lifecycle and evaluating three categories of mitigation techniques: pre-processing, in-processing, and post-processing. Key fairness metrics including demographic parity, equalized odds, and individual fairness are assessed in the context of fair lending regulations. Empirical evidence from fintech lending audits and adversarial learning evaluations illustrates the practical challenges and trade-offs inherent in fairness-aware model design. Findings indicate that no single mitigation technique is universally optimal; the choice depends on the specific fairness definition, regulatory context, and performance requirements. Persistent challenges include the tension between group-level and individual fairness, the difficulty of detecting proxy discrimination, and the need for dynamic fairness monitoring in evolving systems. The analysis outlines a multi-layered framework for operationalizing fairness in AI-driven lending, encompassing technical, regulatory, and organizational dimensions.

Keywords: algorithmic fairness; bias mitigation; credit scoring; loan origination; machine learning; fair lending; demographic parity

ARTICLE INFORMATION

Received: 4 March 2026; Accepted: 10 May 2026; Published: 24 June 2026

CITATION

Patel A, Schmidt L, Fernández ME. Algorithmic Fairness in AI-Driven Loan Origination through Bias Mitigation Across Demographic Groups.
Advances in Digital Finance. 2026; 1(1): 3.

1. Introduction

The integration of artificial intelligence (AI) and machine learning (ML) into credit underwriting has fundamentally transformed the landscape of consumer lending. Financial institutions increasingly rely on algorithmic models to assess creditworthiness, automate loan origination, and scale decision-making across millions of applicants (Bahlool et al., 2026). By 2023, private credit had reached nearly $2 trillion in assets under management, with 73% of financial institutions reporting the use of ML models and tools in their operations (Adadi & Berrada, 2018). These systems promise enhanced efficiency, consistency, and predictive accuracy, particularly for thin-file borrowers who lack traditional credit histories (Raziyeva & Meraliyev, 2025).

Yet the same technologies that enable greater financial inclusion also carry significant risks of perpetuating or amplifying historical inequalities. ML models trained on biased data can produce discriminatory outcomes, even when protected characteristics such as race, gender, or age are explicitly excluded from the model inputs (Coots et al., 2026). This phenomenon—known as algorithmic bias—has drawn increasing scrutiny from regulators, researchers, and civil society organisations. The Consumer Financial Protection Bureau's 2022 algorithmic fairness guidance explicitly calls for causal attribution tools to identify discrimination in AI-driven credit decisions (Suresh & Guttag, 2021). Recent incidents of alleged gender-based disparities in automated credit decisions have intensified public and regulatory concern (Adadi & Berrada, 2018).

The tension between algorithmic efficiency and fairness is particularly acute in loan origination, where decisions carry profound consequences for individuals' economic opportunities and well-being. Biased lending algorithms can systematically disadvantage protected groups, reinforcing patterns of exclusion that fair lending laws were designed to eliminate (Bartlett et al., 2022). The Equal Credit Opportunity Act and the Fair Housing Act establish the legal foundation for non-discriminatory lending, yet the opacity of complex ML models poses novel challenges for enforcement and accountability (Coots et al., 2026).

This article provides a critical analysis of algorithmic fairness in AI-driven loan origination, with a specific focus on bias mitigation techniques across demographic groups. It examines the sources of bias across the ML lifecycle, evaluates the effectiveness of pre-processing, in-processing, and post-processing mitigation strategies, and assesses key fairness metrics in the context of fair lending regulations. The analysis draws on recent empirical studies, including audits of fintech lending platforms and evaluations of adversarial learning approaches, to illustrate the practical challenges and trade-offs inherent in fairness-aware model design. The article concludes by outlining a multi-layered framework for operationalizing fairness in AI-driven lending.

2. Materials and Methods

2.1 Analytical Framework

This study takes a qualitative, critical synthesis approach. The analysis is structured around three interconnected questions: where does bias enter the machine learning lifecycle, what techniques are available to mitigate it, and what fairness metrics make sense in a lending context? Each question points to a different set of literatures and a different set of practical constraints.

2.2 Literature Search and Selection

The literature review was conducted across Scopus, Web of Science, IEEE Xplore, and Google Scholar. Search terms included algorithmic fairness, bias mitigation, credit scoring, loan origination, machine learning, fair lending, demographic parity, equalized odds, and adversarial learning. The search covered work published between 2018 and 2026, with an emphasis on empirical evaluations of mitigation techniques applied to lending data. Systematic reviews and meta-analyses were included where they helped map the broader landscape.

2.3 Bias Source Categorisation

Bias is categorised according to where it emerges in the machine learning pipeline: data-level biases (sampling, measurement, historical, and label bias); algorithmic biases (optimisation, proxy discrimination, and specification bias); and deployment-level biases (feedback loops, concept drift, and evaluation bias). This framework helps show that fairness is not something to be solved at a single point but rather something that demands attention throughout.

2.4 Mitigation Technique Classification

Mitigation techniques are grouped into three families following the standard taxonomy (Suresh & Guttag, 2021; Mehrabi et al., 2021): pre-processing, which acts on the training data; in-processing, which intervenes during model training; and post-processing, which adjusts model outputs. Each approach is assessed for its effectiveness in reducing disparity, its impact on predictive accuracy, its computational demands, and its fit with regulatory requirements.

2.5 Case Study Selection

Two empirical case studies are explored in detail: a large-scale audit of algorithmic bias in a U.S. fintech lending platform (Coots et al., 2026), and an evaluation of adversarial learning for bias mitigation in credit scoring (Adadi & Berrada, 2018). These cases were chosen because they represent different approaches to fairness assessment and different stages of the mitigation pipeline.

3. Results

3.1 Where Bias Creeps In

Bias can enter AI-driven lending systems at almost any stage. Historical bias is perhaps the most straightforward: the training data reflect past discriminatory practices—redlining, unequal access, biased loan officer decisions—and the model learns from them as if they were neutral facts (Bartlett et al., 2022). Sampling bias occurs when certain demographic groups are underrepresented in the training data, leading to models that perform poorly for those populations (Mehrabi et al., 2021). Measurement bias arises from proxy variables that correlate with protected characteristics—zip code, employment history, or even the type of phone someone uses—which can serve as a backdoor for discrimination (Coots et al., 2026). Label bias comes from the historical loan approval decisions themselves, which were made by human loan officers who brought their own biases to the table (Raziyeva & Meraliyev, 2025).

Algorithmic biases are more subtle. Even with a clean dataset, the choices made during model development can amplify disparities. Optimisation bias emerges when models are trained purely to maximise predictive accuracy, without regard for fairness (Suresh & Guttag, 2021). Proxy discrimination happens when the model picks up on variables that are highly correlated with protected attributes, effectively bypassing regulations that prohibit the direct use of such characteristics (Coots et al., 2026). Model specification bias results from decisions about functional form, feature selection, or hyperparameter tuning that inadvertently privilege certain groups (Mehrabi et al., 2021).

Deployment-level biases are often overlooked, yet they can be just as consequential. Feedback loops form when model decisions influence future data—if a model disproportionately denies loans to a particular group, those individuals will lack the credit history needed to qualify in the future, perpetuating the cycle (Raziyeva & Meraliyev, 2025). Concept drift arises when the relationship between features and outcomes shifts over time, potentially undoing hard-won fairness guarantees (Bahlool et al., 2026). Evaluation bias reflects the metrics used to assess model performance; if those metrics don't capture fairness concerns, the assessment is incomplete.

3.2 Fairness Metrics

Different fairness metrics operationalise different visions of what a fair lending system should look like. Demographic parity requires approval rates to be approximately equal across protected groups, regardless of risk (Mehrabi et al., 2021). It is intuitive and maps relatively directly onto disparate impact doctrine, but it can force models to ignore legitimate risk differences—trading off accuracy for equity (Bartlett et al., 2022). Equalized odds requires that false positive and false negative rates be equal across groups, ensuring that model errors are not systematically biased (Hardt et al., 2016). It is more permissive than demographic parity, allowing approval rates to differ when justified by actual risk differences, but it requires access to ground truth labels that may themselves be biased (Suresh & Guttag, 2021). Equal opportunity, a variant of equalized odds, requires only that true positive rates be equal—ensuring that qualified applicants from all groups have an equal chance of approval (Hardt et al., 2016). Individual fairness shifts the focus from groups to individuals, requiring that similar people receive similar outcomes (Dwork et al., 2012). It avoids the group-level trade-offs but requires a meaningful similarity metric, which is difficult to define and operationalise in practice (Mehrabi et al., 2021). Predictive parity requires that positive predictive values be equal across groups—the proportion of approved applicants who actually repay their loans should be consistent across demographic groups (Chouldechova, 2017). It makes sense from a risk management perspective but can conflict with other fairness criteria.

The choice of metric is normative, not technical. In the U.S. regulatory context, disparate impact analysis has been the dominant legal framework, which gives demographic parity and related metrics particular weight (Coots et al., 2026). But regulators are increasingly aware that no single metric captures the full picture; a multi-metric approach is needed.

3.3 Bias Mitigation Techniques

Pre-processing techniques aim to cleanse the training data of bias before the model is trained. Common methods include re-weighting, which assigns different weights to instances from different groups; re-sampling, which over-samples underrepresented groups or under-samples overrepresented

ones; and data transformation, which maps the data to a representation that preserves task-relevant information while removing group-specific information (Mehrabi et al., 2021; Suresh & Guttag, 2021). These methods are computationally efficient and model-agnostic, which makes them attractive in practice. But they may not be sufficient when bias arises from algorithmic choices rather than data alone, and they can reduce the overall representativeness of the training data.

In-processing techniques build fairness constraints directly into the training process. Adversarial learning is perhaps the most promising approach in this category: a primary model is trained to predict outcomes, while an adversary tries to predict protected attributes from the model's predictions, forcing the primary model to learn representations that are not informative of group membership (Adadi & Berrada, 2018). Other in-processing techniques include constrained optimisation, where fairness metrics are incorporated as constraints in the loss function, and fairness-aware regularisation, which penalises models that exhibit disparate impact (Mehrabi et al., 2021). In-processing methods can achieve better fairness-accuracy trade-offs than pre-processing, but they are often computationally intensive and model-specific.

Adversarial learning has been tested in credit scoring contexts with promising results. A dual-network architecture—one network for prediction, one for adversarial debiasing—can learn task-relevant representations while mitigating bias. Evaluations on a Chinese P2P lending dataset and the German Credit dataset show that this approach performs well across multiple fairness criteria and achieves higher predictive accuracy than widely used machine learning baselines (Adadi & Berrada, 2018). The adversarial mechanism improves fairness in credit scoring decisions while maintaining a better fairness-efficiency trade-off.

Post-processing techniques adjust the model's outputs after training. Common methods include threshold adjustment, which sets different decision thresholds for different groups; calibration, which adjusts predicted probabilities to be well-calibrated across groups; and reject-option classification, which changes decisions for instances near the decision boundary (Mehrabi et al., 2021). Post-processing is model-agnostic and easy to implement, making it suitable for retrofitting fairness into existing systems. But it cannot address bias that is deeply embedded in the model's representations, and it may require access to protected attributes at deployment time, raising privacy and regulatory concerns.

3.4 Evidence from Fintech Audits

Recent audits of fintech lending platforms have provided concrete evidence of algorithmic bias and its consequences. One study, examining approximately 80,000 personal loans from a major U.S. fintech platform, found that loans made to men and Black borrowers yielded lower profits than loans to other groups—meaning these borrowers benefited from relatively favourable pricing (Coots et al., 2026). The disparities were traced to miscalibration in the platform's underwriting model, which overestimated risk for women and underestimated risk for Black borrowers. The study also found that one could correct this

miscalibration by including race and gender in the underwriting model. This is a deeply uncomfortable finding: colour-blind models can inadvertently perpetuate disparities, but fairness-aware models that explicitly consider protected characteristics raise their own legal and ethical questions.

Evidence from a Chinese auto equity lender tells a somewhat different story. The study found that cognitive biases decreased significantly when loan officers used algorithmic lending decisions, substantially reducing disparities in loan-to-value ratios between local and nonlocal borrowers without worsening default differentials (Du et al., 2023). The researchers found that discretionary adjustments made by loan officers remained modest, while advisory credit scores alone had no discernible bias-reducing effect. This suggests that automation—specifically, nudging via algorithmic defaults—is more effective than mere information provision in combating discrimination and promoting financial inclusion.

4. Discussion

4.1 Interpretation of Findings

Several insights emerge from this analysis. Bias can enter the lending pipeline at multiple points, which means fairness cannot be solved at a single stage. Pre-processing, in-processing, and post-processing all have a role to play, but none is a silver bullet. The choice of technique depends on the fairness definition one adopts, the regulatory context one operates in, and the performance requirements one has to meet.

No single fairness metric is universally appropriate. Demographic parity, equalized odds, equal opportunity, individual fairness, and predictive parity all capture something important, but they also conflict with one another. The tension between group-level and individual fairness, and between fairness and accuracy, reflects fundamental value judgments that cannot be resolved through technical means alone.

The fintech audit evidence shows that algorithmic bias is not a theoretical worry—it has real consequences for real people. The finding that colour-blind models can perpetuate disparities, while fairness-aware models that explicitly consider protected characteristics may be legally and ethically controversial, highlights just how difficult it is to operationalise fairness in practice.

Adversarial learning stands out as a promising in-processing technique for credit scoring. It achieves a better fairness-efficiency trade-off than alternative approaches, and its dual-network architecture can learn task-relevant representations while mitigating bias, performing well across multiple fairness criteria.

4.2 Comparison with Existing Literature

These findings align with prior systematic reviews that identified fairness, explainability, and regulatory accountability as critical barriers to AI adoption in high-stakes credit decisions (Bahlool et al., 2026). A recent systematic review of 43 peer-reviewed studies found that performance, fairness, and explainability are predominantly addressed in isolation, with little attention to how they interact in regulated deployment settings (Bahlool et al., 2026). This article extends that work by providing a comprehensive evaluation of bias mitigation techniques specifically in loan origination, drawing on recent empirical evidence from fintech audits and adversarial learning evaluations.

The findings also support a recent systematic review of AI-based credit assessment, which traced developments across five domains: machine learning and hybrid models, alternative data, real-time learning, Explainable AI, and fairness-oriented governance (Raziyeva & Meraliyev, 2025). The review identified four gaps: the absence of streaming-ready frameworks, insufficient validation of behavioural datasets, lack of standardised explainability metrics, and weak integration of fairness and governance requirements. This article addresses the fairness and governance gap by offering a framework for operationalising fairness in AI-driven lending.

4.3 Persistent Challenges

Technical challenges remain. Proxy discrimination is hard to detect and harder to mitigate. The tension between group-level and individual fairness is not easily resolved. And fairness guarantees can degrade over time as systems evolve and concept drift sets in (Mehrabi et al., 2021; Suresh & Guttag, 2021).

Regulatory challenges are equally formidable. Fair lending regulations vary across jurisdictions, auditing complex machine learning models is difficult, and regulators have been slow to provide clear guidance on acceptable fairness metrics and thresholds (Bahlool et al., 2026). The Consumer Financial Protection Bureau's 2022 algorithmic fairness guidance was a step forward, but implementation remains uneven.

Organisational challenges are often overlooked. Fairness requires interdisciplinary collaboration among AI developers, regulatory bodies, and social scientists—collaboration that does not come naturally. Embedding fairness as a core design principle, rather than treating it as an afterthought, requires organisational commitment that many institutions have not yet made. And fairness can conflict with other organisational priorities, such as profitability and efficiency (Raziyeva & Meraliyev, 2025).

4.4 Policy Implications

For policymakers, this analysis points to several priorities. Regulators need to provide clearer guidance on fairness metrics and thresholds. While disparate impact analysis provides a legal framework, it does not specify which fairness metrics should be used or what constitutes an acceptable level of disparity. Auditing and accountability mechanisms need to be strengthened to ensure that AI-driven lending systems are subject to meaningful oversight. Interdisciplinary collaboration among AI developers, regulatory bodies, and social scientists should be encouraged. And dynamic fairness monitoring should be required to ensure that fairness guarantees remain valid as systems evolve.

For financial institutions, the lesson is that fairness cannot be an afterthought. It needs to be integrated into the machine learning lifecycle from the outset. This means investing in fairness-aware tools and techniques, training data scientists and model developers in fairness principles, and engaging with regulators and affected communities.

4.5 Limitations and Future Research

This analysis focuses on U.S. and European regulatory contexts, and findings may not generalise to jurisdictions with different fair lending frameworks. The field is evolving rapidly, and some findings may quickly become dated. And the analysis does not quantify the economic costs of bias mitigation or the benefits of fairness-aware lending.

Future research should focus on quantitative analyses of the fairness-accuracy trade-off in real-world lending contexts, comparing mitigation techniques across diverse datasets and regulatory regimes. Comparative studies of fairness metrics and their implications for disparate impact analysis would be valuable. And the governance of AI-driven lending systems—including the role of internal audit, external oversight, and community engagement—deserves more attention..

5. Conclusion

Algorithmic fairness in AI-driven loan origination is a complex challenge spanning technical, regulatory, and organisational dimensions. Bias can enter lending systems at multiple points across the machine learning lifecycle, and no single mitigation technique or fairness metric is universally optimal. Pre-processing, in-processing, and post-processing each have their place, but the choice depends on the specific fairness definition, regulatory context, and performance requirements.

The evidence from fintech audits shows that algorithmic bias is not a hypothetical concern—it has concrete consequences for protected groups. Colour-blind models can inadvertently perpetuate disparities, while fairness-aware models that explicitly consider protected characteristics raise their own legal and

ethical questions. Adversarial learning has emerged as a promising in-processing technique, achieving better fairness-efficiency trade-offs than alternative approaches.

Operationalising fairness in AI-driven lending requires a multi-layered framework that includes technical measures (bias detection and mitigation, fairness metrics, dynamic monitoring), regulatory measures (clear guidance, auditing and accountability, enforcement), and organisational measures (investment in fairness-aware tools, interdisciplinary collaboration, community engagement).

The stakes are high. AI-driven lending systems can expand access to credit and promote financial inclusion, but they also carry the risk of perpetuating historical inequalities. Ensuring algorithmic fairness in loan origination is not just a technical problem—it is a fundamental question of justice and equity in an increasingly automated world. Addressing it requires sustained commitment from researchers, practitioners, regulators, and communities alike.

Author Contributions

Conceptualization, A.P.; methodology, A.P. and L.S.; investigation, M.E.F.; writing—original draft preparation, A.P.; writing—review and editing, A.P., L.S., and M.E.F. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Funding

This research received no external funding.

Ethics Approval

Not applicable

Acknowledgements

The authors thank the Consumer Financial Protection Bureau and the Financial Action Task Force for publicly available guidance on algorithmic fairness and fair lending. AI language tools were used to

assist with language polishing. The authors also gratefully acknowledge the open access policies of their respective institutions.

Conflicts of Interest

The authors declare no conflicts of interest to report regarding the present study.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access, 6,* 52138-52160. https://doi.org/10.1109/ACCESS.2018.2870052

Bahlool, R., Hewahi, N., & Elmedany, W. (2026). Performance, fairness, and explainability in AI-based credit scoring: A systematic literature review. *Journal of Risk and Financial Management, 19*(2), 104. https://doi.org/10.3390/jrfm19020104

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics, 143*(1), 30 - 56. https://doi.org/10.1016/j.jfineco.2021.05.047

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

Coots, M., Bartlett, R., Nyarko, J., & Goel, S. (2026). Algorithmic bias in lending: Evidence from a fintech audit. *arXiv.* 2512.20753.

Du, X., Li, Y., & Zhang, J. (2023). Does FinTech reduce human biases? Evidence from advisory vs. automated FinTechs in lending. *Journal of Financial Economics.*

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp. 214-226). ACM. https://doi.org/10.1145/2090236.2090255

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

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), 1-35. https://doi.org/10.1145/3457607

Raziyeva, S., & Meraliyev, M. (2025). Bias and fairness in automated loan approvals: A systematic review of machine learning approaches. *Journal of Emerging Technologies and Computing.* https://doi.org/10.47344/jbzmnx25

Suresh, H., & Guttag, J. V. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. https://doi.org/10.1145/3465416.3483305