Review

Bias and Fairness in Automated Loan Approvals: A Systematic Review of Machine Learning Approaches

Suraiyo Raziyeva [1] and Meraryslan Meraliyev [2]

[1]Department of Computer Science, SDU University, Almaty, Kazakhstan
[2]Department of Computer Science, SDU University, Almaty, Kazakhstan

DOI: 10.47344/jbzmnx25

Abstract

Artificial intelligence (AI) is increasingly transforming credit approval processes, enabling financial institutions to assess risk more efficiently and at greater scale. As these systems become more embedded in lending decisions, concerns around fairness, bias, and accountability have grown significantly. Many of these concerns stem from the use of historical data, proxy variables, and model optimization choices that can unintentionally reinforce existing social and economic inequalities. This work presents a systematic overview of the types and sources of bias in AI - driven loan approval systems and critically examines how machine learning techniques attempt to address them. It also highlights emerging solutions, including explainable AI, federated learning, human-in-the-loop frameworks, and intersectional fairness approaches. Despite ongoing advancements, unresolved challenges remain - particularly the need for dynamic fairness monitoring and for addressing intersectional biases affecting individuals from multiple marginalized groups. To bridge these gaps, the paper emphasizes the importance of interdisciplinary collaboration among AI developers, regulatory bodies, and social scientists. It advocates embedding fairness as a core design principle in the development and deployment of future AI systems. Overall, this study contributes to the growing effort to develop more transparent, inclusive, and socially responsible financial technologies.

Keywords: AI bias, fairness techniques, loan approval, financial inclusion, regulatory compliance, algorithmic fairness, proxy bias.

I. Introduction

Artificial Intelligence (AI) has significantly transformed decision-making in the banking sector, particularly through the automation of lending approvals. These systems are often praised for their efficiency and scalability; however, they also raise critical concerns regarding fairness and bias. Historical data used to train such systems may reflect past discriminatory practices, leading to models

Email: suraiyo.raziyeva@sdu.edu.kz ORCID: 0009-0008-1832-7155
Email: meraryslan.meraliyev@sdu.edu.kz ORCID: 0000-0003-2627-0837

Received: April 3, 2025. Reviewed: April 16, 2025. Accepted: April 22, 2025.

that perpetuate unfair outcomes. As a result, automated credit scoring and lending decisions may disproportionately disadvantage underrepresented groups [1].

Bias in AI systems may also stem from algorithmic design choices, such as optimization objectives or feature selection. For instance, if the training dataset is imbalanced, models may systematically favor majority groups. Variables like ZIP codes can inadvertently encode socioeconomic or demographic biases, reinforcing existing inequalities. In some cases, the use of such features may even violate regulatory frameworks, including the General Data Protection Regulation (GDPR) and the Equal Credit Opportunity Act (ECOA) [1], [2]. Beyond ECOA and GDPR, emerging international frameworks—such as the EU AI Act and Canada's Directive on Automated Decision-Making—signal a global shift toward standardizing fairness and transparency in automated financial systems. These evolving policies reflect the growing global consensus on the need for algorithmic accountability in financial decision-making.

This paper considered various contemporary fairness metrics used to detect and assess bias in AI-driven lending systems, as well as a range of strategies designed to mitigate such biases [3]. Lending discrimination can originate from multiple sources, including proxy variables that are correlated with race, gender, or economic status, as well as algorithmic priorities that favor accuracy at the expense of fairness. These biases, if left unaddressed, may deepen economic inequalities by restricting equitable access to credit.

In deploying AI for credit decisions, it is essential to consider not only model performance but also fairness, ethical accountability, and compliance with legal standards. Financial institutions must adhere to anti-discrimination laws and data protection regulations such as the GDPR and ECOA [1]. These frameworks aim to ensure equitable treatment within algorithmic decision-making. However, many current machine learning techniques struggle to balance fairness and accuracy effectively.

The goal of this work is to provide a comprehensive examination of machine learning approaches aimed at promoting fairness in AI-powered lending systems [4]. It categorizes bias mitigation methods into three primary classes: pre-processing, in-processing, and post-processing. It further analyzes the strengths and limitations of each approach in addressing algorithmic unfairness. The study also examines the broader societal implications of discriminatory lending and contributes to the ongoing discourse by including real-time fairness monitoring techniques and intersectional fairness considerations.

By offering an in-depth analysis of fairness-centered machine learning strategies, this paper contributes to the growing literature on ethical AI in financial services. It synthesizes key trends in recent work and clearly identifies persistent gaps in current research. Notably, we find a lack of robust methods for continuous ("live") fairness monitoring in deployed lending systems and a deficiency of techniques to address intersectional biases—cases where compounded disadvantages (e.g., being a minority and low-income) fall through the cracks of one-dimensional fairness metrics. We also observe a disconnect between high-level ethical principles and their practical implementation in AI lending algorithms. To help bridge these gaps, we propose future directions such as developing dynamic equity monitoring tools and more holistic fairness metrics that account for intersecting social categories. Ultimately, the insights from this systematic review are intended to inform the design of more transparent, inclusive, and accountable AI-driven credit decision platforms, guiding both researchers and practitioners toward solutions that balance performance with fairness and uphold regulatory and social standards.

A. Theoretical Frameworks of Fairness

In recent discussions surrounding AI-driven financial decisions, particularly automated loan approvals, several ethical frameworks provide valuable insights into the concept of fairness. From a utilitarian perspective, fairness is assessed by evaluating whether algorithmic outcomes maximize overall welfare or utility. This approach considers decisions fair if they enhance aggregate societal benefit or utility, often prioritizing accuracy and profitability in credit scoring. However, this might inadvertently result in unequal distributions of benefits and harms [30], [31].

Alternatively, Rawlsian fairness, rooted in John Rawls's notion of justice as fairness, emphasizes distributive justice aimed explicitly at improving conditions for the least advantaged. Rawls's Difference Principle asserts that any inequality is justified only if it benefits those most disadvantaged [32], [33]. Applying this framework to lending implies that fairness-oriented interventions should protect vulnerable borrowers and mitigate disparities that could otherwise exacerbate their disadvantages [34].

Furthermore, the emerging computational justice framework proposes integrating egalitarian principles directly into AI systems. This framework stresses the importance of equitable access, representation, and outcomes, ensuring algorithmic lending practices provide fair opportunities across diverse socioeconomic backgrounds and proactively address biases [35], [36].

Collectively, these ethical perspectives—utilitarianism, Rawlsian fairness, and computational justice—serve as complementary lenses, enriching our understanding of fairness in AI-driven financial services. They guide the design and assessment of algorithms, promoting fairness that encompasses both efficiency and equity, particularly benefiting vulnerable populations.

Received: April 3, 2025. Reviewed: April 16, 2025. Accepted: April 22, 2025.

II. Methodology

To carry out a comprehensive and objective evaluation of fairness-centered, AI-driven loan approval systems, this study adopts the systematic literature review (SLR) methodology. This approach enables a structured examination of existing scholarly work, facilitating the identification of prevailing trends, emerging challenges, and methodological limitations associated with fairness-aware AI in financial decision-making.

A. Search Strategy

The literature search focused on peer-reviewed journal articles, conference proceedings, and preprint research published between 2020 and 2024. Key academic databases were utilized, including Scopus, Google Scholar, IEEE Xplore, the ACM Digital Library, and arXiv—chosen for their extensive coverage of publications related to artificial intelligence, machine learning, and financial technologies. A combination of keywords and Boolean operators was applied to ensure the inclusion of relevant and high-quality studies. The following search terms were used:

- `("fairness-aware machine learning" OR "bias mitigation in AI")`
- AND `("loan approval" OR "credit scoring" OR "financial decision-making")`
- AND `("pre-processing" OR "in-processing" OR "post-processing")`

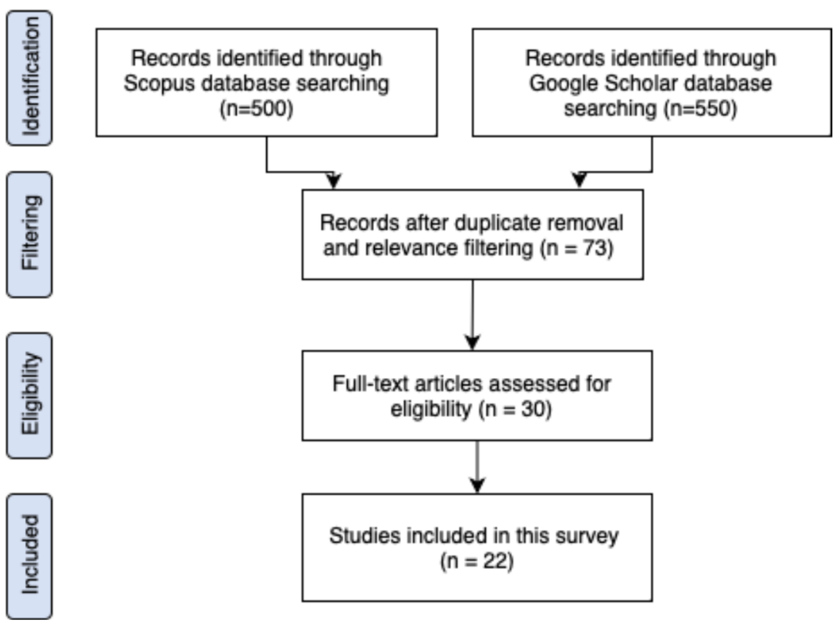

Fig. 1. Systematic Literature Review (SLR) Process: from initial identification to final selection.

Figure 1 illustrates the step-by-step review process applied in identifying relevant literature for this study. Initially, database searches yielded 500 records from Scopus and 550 from Google Scholar, totaling 1050 records.

B. Inclusion and Exclusion Criteria

Studies were selected as suitable if they comprehensively addressed bias mitigation methods, specifically in AI-driven loan approval systems. Preference was given to studies explicitly employing fairness-aware techniques categorized into pre-processing,

Received: April 3, 2025. Reviewed: April 16, 2025. Accepted: April 22, 2025.

in-processing, or post-processing methods. Priority was also given to studies providing empirical evaluations of fairness-accuracy trade-offs and discussions aligned with regulatory and ethical frameworks, such as the General Data Protection Regulation (GDPR) and the Equal Credit Opportunity Act (ECOA).

Studies were excluded from the review if they addressed fairness exclusively in non-financial contexts, such as healthcare or employment. In addition, works lacking quantitative evaluations or those that did not propose explicit bias mitigation strategies were omitted. Non-peer-reviewed sources, including editorial commentaries and opinion articles, were also excluded to ensure the inclusion of rigorously vetted research.

Initially, 73 records were retained after duplicate removal and initial filtering based on relevance. These underwent a full-text eligibility assessment, after which 30 articles remained. Following rigorous quality assessment and further evaluation for relevance and context, a final set of 22 high-quality and contextually relevant studies was selected for detailed comparative analysis.

C. Data Extraction Process

Following the selection of relevant studies, key information was carefully extracted to support a meaningful comparison of fairness-aware machine learning techniques. This included identifying the specific types of bias each study addressed - such as historical, sample, algorithmic, proxy, or societal bias. The review also categorized the mitigation methods based on the stage at which they were applied: before training (pre-processing), during training (in-processing), or after model deployment (post-processing). Particular attention was given to empirical findings, especially how each approach balanced fairness with predictive accuracy, using metrics like statistical parity difference, equalized odds, and disparate impact ratio. Finally, the analysis considered how well each technique aligned with existing financial regulations and ethical standards.

D. Quality Assessment

Each study was assessed based on clarity of methodology, empirical rigor, practical applicability, and regulatory alignment to ensure the robustness and relevance of findings. After this detailed quality assessment and eligibility evaluation, a final set of 22 studies were included in the comparative survey.

The collected data were analyzed to evaluate the comparative strengths, limitations, and practical applicability of each technique. This analysis provided insights into how different bias mitigation strategies perform in real-world contexts and highlighted inherent trade-offs.

III. Taxonomy of Bias and Fairness Techniques

A. Categorization of Bias in AI-Based Loan Approvals

Recent developments in fairness research have introduced new ways to reduce bias. For instance, federated learning allows models to be trained across different datasets without centralizing data [5]. This approach keeps personal information private and helps reduce geographical and demographic biases. We also have real-time fairness monitoring systems. These systems continuously check and adjust model performance after deployment to address any changes in bias [6]. Finally, new intersectional fairness metrics consider the combined effects of biases on people who belong to multiple disadvantaged groups [7]. This promotes more inclusive lending practices.

Table I provides a clear comparison of different bias mitigation techniques used in fairness-aware machine learning, grouped by when they are applied in the modeling process: before training (pre-processing), during training (in-processing), or after the model has been trained (post-processing) [8]. The table outlines how each method typically performs in terms of improving fairness, the extent to which it affects model accuracy, and how complex it is to implement. As shown, pre-processing techniques like reweighing and data balancing are relatively easy to apply and don't significantly impact accuracy, but they may offer only modest improvements in fairness [4]. In contrast, in-processing methods, such as adversarial debiasing, tend to be more effective in reducing bias but are also more complex and may involve trade-offs in performance [8], [9]. Post-processing approaches are often useful for adjusting outcomes in already trained models, though they can sometimes lead to inconsistencies or reduced interpretability [11]. This comparison helps highlight the practical choices researchers and developers must make when selecting fairness techniques, depending on their goals, constraints, and the context in which the model will be used.

B. Fairness-Aware Machine Learning Techniques

To mitigate the impact of bias in AI-driven lending, various fairness-aware machine learning techniques have been developed. These techniques can be classified based on when they are applied during the model development process:

Received: April 3, 2025. Reviewed: April 16, 2025. Accepted: April 22, 2025.

TABLE I
SUMMARY OF BIAS TYPES AND CORRESPONDING MITIGATION TECHNIQUES IN AI-BASED LOAN APPROVAL SYSTEMS.

Bias Type	**Application in Loan Approvals**	**Key Features and Challenges**	**Performance Analysis/Remarks**
Historical Data Bias	AI models learn from past loan decisions.	Reflects systemic inequalities in training data, leading to biased outcomes.	Replicates historical discrimination: requires data rebalancing or de-biasing techniques.
Sample Bias	Models trained on non-representative datasets.	Leads to poor generalization for underrepresented groups.	Reduces model accuracy for diverse populations: mitigated by re-sampling or balanced datasets.
Algorithmic Bias	Introduced during feature selection or optimization processes.	Favors majority groups due to imbalanced cost functions or feature correlations with protected attributes.	Reduces fairness, mitigated by fairness-aware algorithms and optimization strategies.
Proxy Bias	Use of correlated attributes as substitutes for sensitive ones.	Leads to indirect discrimination when proxies represent protected characteristics.	Compromises fairness: requires removal or neutralization of proxy variables.
Amplification Bias	AI systems intensify existing disparities in training data.	Magnifies inequities by reinforcing patterns of inequality present in datasets.	Increases systemic bias: mitigated through fairness monitoring and algorithmic adjustments.
Societal Bias	Societal prejudices embedded in datasets influence AI decisions.	Reflects societal inequities like gender or racial discrimination, influencing model outcomes.	Preserve inequality addressed through societal-level reforms and ethical AI frameworks.

1) Pre-Processing Techniques: As a pre-processing method, transformation techniques are applied to manipulate the dataset before machine learning models consume the data to get unbiased data and equal representation. These include approaches to resampling, creating synthetic data, and finally, removing proxy variables so that the dataset fairly represents all demographic groups. Such methods mainly target historical and sample bias but must be carefully tuned to ensure the utility of the data. Increasing data for underrepresented groups should re-balance the data set and improve fairness and minimal prediction accuracy reduction [5].

2) In-Processing Techniques: In-processing methods include fairness constraints in the process of training a model. These algorithms optimize fairness metrics, such as demographic parity or Equalized Odds, and traditional accuracy metrics. The models balance fairness and performance by embedding these constraints into the learning objective. In-processing techniques go well with modern AI frameworks, which consider fairness-aware algorithms one of the strong approaches toward handling algorithmic bias [9].

3) Post-Processing Techniques: Post-processing methods take an already fitted model and perform whatever adjustments necessary to the model outputs so that the resulting predictions satisfy some notion of fairness, be it demographic parity or Equalized Odds. While post-processing methods allow for easy retrofitting of fairness into many existing models, they come at a small cost in the overall predictive accuracy of the model. In practice, these methods are used when it is infeasible to modify the process by which models are trained or the data on which they are trained [11].

C. Empirical Findings: Fairness - Accuracy Trade-offs

Empirical studies of fairness-aware machine learning techniques reveal a complex interplay between predictive accuracy and fairness outcomes. In the context of automated loan approvals, achieving both high performance and equity across demographic groups is rarely straightforward. This subsection synthesizes empirical findings from the reviewed literature to highlight observed trade-offs, using metrics such as statistical parity difference (SPD), equalized odds, and disparate impact ratio (DIR).

Received: April 3, 2025. Reviewed: April 16, 2025. Accepted: April 22, 2025.

Pre-processing techniques, such as reweighing and massaging the training data, often lead to modest improvements in fairness metrics like SPD and DIR, particularly in datasets with high initial bias. However, these improvements can come at the cost of reduced predictive accuracy, especially when the modified data distribution diverges from the underlying population [3], [11].

In-processing methods, including adversarial debiasing and fairness-constrained optimization, tend to achieve a better balance between fairness and accuracy. Several studies demonstrate that models trained with fairness constraints are capable of maintaining comparable AUC scores while significantly reducing disparate impact [15]. However, these methods often require extensive tuning and longer training times.

Post-processing techniques such as reject option classification or calibrated equalized odds adjustments are shown to be effective when model retraining is not feasible. While these approaches can achieve compliance-level improvements in fairness indicators like DIR (moving from below 0.8 to above the threshold), they sometimes reduce classification confidence, especially near decision boundaries [26].

D. Comparative Analysis of Fairness Metrics and Mitigation Techniques

Bias mitigation strategies in machine learning for loan approval systems are commonly grouped into three categories: pre-processing, in-processing, and post-processing. Each group offers distinct advantages and limitations depending on the context in which it is applied.

Pre-processing techniques attempt to transform the training data to remove bias before model training. Methods such as reweighting and sampling adjustments are often model-agnostic and relatively easy to apply. However, they may risk distorting the original data distribution, which could reduce model performance in certain applications.

In-processing techniques modify the learning algorithm itself by incorporating fairness constraints or altering the loss function. These methods, such as adversarial debiasing or fairness-constrained optimization, tend to offer a strong balance between accuracy and fairness. However, they often require access to and modification of the model's internals, making them less applicable to proprietary or black-box systems.

Post-processing techniques alter the model output without changing the model or data. Examples include reject option classification and calibrated equal odds. These methods are typically easy to implement but may offer limited fairness correction, especially if bias is deeply embedded in the model structure.

evaluation is further complicated by the choice of metric. The most widely used measures include:

- Statistical Parity Difference (SPD) – measures the difference in positive outcomes between privileged and unprivileged groups. It is simple but may ignore performance discrepancies.
- Disparate Impact (DI) – computes the ratio of favorable outcomes across groups. It is widely used in regulatory contexts but sensitive to class imbalance.
- Equal Opportunity Difference – measures the true positive rate gap between groups, emphasizing fairness in correctly approved applications.
- Average Odds Difference – considers both true positive and false positive rates across groups, offering a more nuanced view of fairness.

In practice, there is often a trade-off between fairness and accuracy. For example, post-processing techniques may improve DI but slightly reduce ROC-AUC. In contrast, in-processing methods like adversarial debiasing may maintain accuracy but increase computational cost.

The comparative analysis conducted in this review suggests that fairness-aware learning remains a multidimensional problem where technical performance must be balanced with ethical and legal considerations.

E. Case Studies on Bias Mitigation in AI-Based Loan Approval Systems

Bias mitigation in AI-driven loan approval systems has been the focus of active study through numerous case studies that detail the best strategies for the implementation of fairness and transparency. Figure II provides a summary of representative case studies that have examined various fairness-aware strategies in credit decision-making. The table highlights not only the technical contributions but also the domain-specific applications and the diversity of fairness concerns addressed in the literature. These case studies present a range of real applications and tests of methods related to bias mitigation within a wide array of domains relating to financial services. Below are summaries of key recent findings: Purificato et al. (2023) and Lorenzo (2019) both incorporate explainable

Received: April 3, 2025. Reviewed: April 16, 2025. Accepted: April 22, 2025.

AI techniques—such as the Trust, Reliance Scale,[1] and SHAP values—to promote transparency and user trust, enabling clearer justification for credit decisions [12], [13]. In contrast, Dattachaudhuri et al. (2022) focus on rule extraction and rule pruning toward the same goal and outline the manner in which interpretability can complement and exist separately from typical model structures [11]. Regulatory focus enhances the work of Nadeem et al. (2023), who connect fairness interventions directly into organizational practice and emphasize the need for embedding algorithmic remedies within larger-scale institutional change [14]. A second group of studies addresses fairness from the systems-level or technical perspective. Wang et al. (2023) introduce a human-in-the-loop framework that enables real-time correction of bias through interactive feedback [15], while Zhou et al. (2022) address bias, privacy, and regional imbalance simultaneously through federated learning [16].

Parra et al. (2022) address the problem from the data perspective, looking at how proxy variables such as ZIP codes may introduce bias inadvertently and describing how to nullify their effect [17]. Finally, there are a few papers that focus on fairness-performance trade-offs. Karimova (2024) investigates optimization techniques that balance fairness and precision in micro-lending in small businesses [18], while Cozerenco and Szafarz (2015) adopt a co-financing strategy in microfinance for reducing gender bias, illustrating a policy intervention specific to the field rather than a technical one [19]. Overall, the papers illustrate that while fairness is the common target, the paths towards its realization are very distinct depending on the field, purpose, and limitations of the credit system.

IV. Emerging Trends and Potential Improvements

Recent developments in AI-driven financial decisions have marked a milestone toward a solution for bias and for the promotion of fairness. Specifically, fairness-aware frameworks with integrated XAI tools have become a game-changing solution for improving transparency and rebuilding stakeholder trust. Ashraf and Faheem 2021 [21] proposed XAI-based fairness detection frameworks. These frameworks ensure equity as well as accountability within credit ratings by eliminating algorithmic biases. On the other hand, Garcia et al. (2023) [22] also put great emphasis on systematic bias detection and fairness metrics as being very important in order to reduce both societal and algorithmic discrimination in loan approval systems. Federated learning has also been highlighted as a means of training AI on various data without the sharing of private information. Zhou and Tang (2022) [25] have demonstrated their potential in trying to reduce geographical and population biases while maintaining strong information security in financial institutions. Interpretable machine learning for finance has also seen developments in tools such as the SHAP technique, which is designed to give insight into how decision-making is derived by algorithms. SHAP has been shown to increase transparency and build trust in AI-driven credit decisions, according to Lorenzo 2019 [24], which was one of the concerns that stakeholders had. Other innovative techniques have recently emerged as human-in-the-loop frameworks that solve the challenge of bias dynamically at deployment. Wang et al. (2021) [26] have proposed iterative correction mechanisms that use user feedback to improve fairness in AI. Recently, critical activities have taken place to handle gender bias in financial services. Likewise, Parra et al. (2022) [24] highlighted the role that decorrelating variables has in lessening proxy bias via the removal of disparities within input features. In conclusion, many meaningful advances have been made in the development of fairness-aware machine learning techniques. However, ensuring ethical and unbiased outcomes in AI-based lending remains a complex and evolving challenge. The emerging trends discussed - such as real-time fairness monitoring, intersectional fairness, and the integration of explainability—reflect a growing awareness of both technical and societal dimensions of algorithmic bias. However, these innovations also highlight the need for ongoing interdisciplinary collaboration, stronger regulatory guidance, and scalable, context-sensitive solutions. As financial institutions continue to adopt AI technologies, it is essential that fairness is treated not as an optional add-on but as a core design principle embedded throughout the development and deployment process.

A. Research Directions

Despite these advances, important research gaps and opportunities remain. First is an increasing need for the actual development of real-time fairness monitoring systems that can automatically detect and adapt to changing biases. Such systems can help financial institutions uphold the concept of fairness as time changes with data distribution. Moreover, further work is to be done on improving the so-called fairness-aware algorithms, such as Random Forest models, which balance between fairness and predictive performance, as was shown in the work of Karimova (2024) [4], in order to reach the optimal solution for such a diverse and complex financial context [27].

[1]This scale, introduced by Purificato et al. (2023), is a previously established framework used to assess user confidence in AI decisions. It is not newly developed in this manuscript.

Received: April 3, 2025. Reviewed: April 16, 2025. Accepted: April 22, 2025.

TABLE II
CASE STUDIES ON BIAS MITIGATION IN AI-BASED LOAN APPROVAL SYSTEMS.

Case Study	**Authors (Year)**	**Key Contributions**	**Application**
Fair Lending Tool for Credit Decisions	Erasmo Purificato et al. (2023)	Developed fairness-aware frameworks incorporating Explainable AI (XAI) and Trust & Reliance Scale for evaluating AI systems.	Fair credit assessment and monitoring
Transparent Decision Support for Credit Risk	Abhinaba Dattachaudhuri et al. (2022)	Proposed a neural network-based decision support system enhancing transparency through rule extraction and pruning techniques.	Transparent credit scoring
Algorithmic	Jason Jia-Xi	Critiqued existing legal frameworks like ECOA and advocated harm-based frameworks to address AI-driven discrimination.	Credit underwriting fairness
Gender Bias in Loan Allocations	Ayesha Nadeem et al. (2022)	Examined gender biases in loan allocations and proposed integrating fairness-aware algorithms with organizational changes for equitable outcomes.	Gender bias mitigation in loan allocations
Interactive Approach to Bias Mitigation in ML	Hao Wang et al. (2021)	Introduced an iterative human-in-the-loop approach for bias mitigation, combining feedback loops and interactive visualization.	Iterative bias detection in ML applications
Federated Learning for Loan Approvals	Zhou et al. (2022)	Proposed federated learning approaches to preserve privacy while improving fairness and reducing geographical biases in decentralized training datasets.	Privacy-preserving credit risk evaluation
Addressing Proxy Bias in Lending	Parra et al. (2022)	Highlighted the role of proxy variables like ZIP codes in embedding discrimination and proposed mitigation strategies such as variable decorrelation.	Mitigating proxy bias in lending
Bias in Microfinance Systems	Cozarenco & Szafarz (2015)	Analyzed biases in microfinance systems, particularly gender-related biases, and emphasized co-financing mechanisms to address discriminatory patterns.	Gender fairness in microfinance
SHAP for Fairness in Credit	Lorenzo (2019)	Demonstrated the use of SHAP for interpreting credit scoring models, emphasizing its application in detecting and mitigating algorithmic bias.	Explainable AI for credit scoring
Fairness Metrics in Loan Decisions	Karimova (2024)	Introduced fairness-aware optimization techniques using Random Forest models to evaluate and reduce biases in small business loan decisions.	Optimizing fairness in small business loans

In the future, fairness metrics would need further refinement for capturing intersectional biases that disproportionately affect people who belong to more than one disadvantaged group. Interdisciplinary collaboration between practitioners in AI, regulators, and social scientists is badly needed to develop ethical frameworks that align technological innovation with societal values. Increasing federated learning beyond the currently explored use cases might advance inclusive global financial systems and safeguard data privacy.

Received: April 3, 2025. Reviewed: April 16, 2025. Accepted: April 22, 2025.

B. Discussion

AI-powered loan decision engines have turned a new leaf and become truly transformative forces that change the paradigm of operation scale and efficiency in the realm of finance. However, recurrent bias raises problems in ethical and regulatory paradigms, raising some serious questions about algorithm fairness. Recent work suggests some ways to diminish these; at a similar scale, there continue to be many challenges to reaching equal, transparent AI systems. One of the most relevant developments is the development of frameworks and tools for fairness. These frameworks allow explainable AI and transparency that relates to a clear look at how AI comes up with its decisions to stakeholders; for example, Lorenzo has shown that it can be done through better interpretability of credit scoring using SHAP [29]. On the other hand, Purificato et al. (2023) have proposed the Trust, Reliance Scale, an approach to embed fairness in AI systems so that the latter could be held accountable for the decisions made [29]. The key issue is gender bias in loan distribution. Nadeem et al. (2022) have proposed a socio-technical fairness framework that can serve to overcome systemic and algorithmic biases, especially affecting provisions for financial services related to women [15]. Another point of interest is proxy bias. Examples of strategies, such as variable de-correlation, that may be used for eliminating the indirect discrimination given by features like ZIP codes or levels of education are provided by Parra et al. (2022) [16]. Other techniques, such as federated learning, took the cause further in terms of fairness. Zhou and Tang 2022 showed how Federated learning decreases geographical and demographic biases by taking away the centralization of the training of the model [25]. This allows inclusiveness while providing protection against sensitive information. Iterative bias mitigation through human-in-the-loop frameworks discussed by Wang et al. 2021 allows dynamic changes in deployment for evolving biases [26]. Advantages, however, have many problems. The problem of bias in AI models is dynamic and thus can always be subject to the detection, capturing, and correction of constantly evolving disparities by certain monitoring mechanisms. Karimova, 2024 [4]. The paper proposes a fairness-aware optimization of Random Forest algorithms that strikes a balance between predictive accuracy and fairness in granting loans to small businesses [17]. At a wider scale, such implementation still remains resource-intensive. Another important and complex issue to mention is the intersectionality of biases: persons having more than one disadvantaged group they identify with. Most of the existing fairness metrics are not correctly positioned to handle these interactions in compound ways. This calls for future research in developing holistic metrics that address nuances around intersectional fairness. Integrating ethical frameworks: this aligns technological advancements with placing value on society [28]. In return, this would require collaboration among AI researchers, regulators, and social scientists. Much has been done, yet significant steps to handle the biases of AI in loan approval systems are long in coming. Dynamic fairness-monitoring systems, intersectional fairness metrics, and robust frameworks pave the way for equity in financial decision-making. In this regard, further innovation will have to negotiate a delicate balance between ethical and regulatory imperatives if the gap between technical capability and societal expectation is to be bridged [29].

V. Conclusion

This study investigated the sources of bias in AI-driven loan approval systems and critically evaluated fairness-aware machine learning techniques designed to mitigate such biases. It examined a spectrum of approaches: pre-processing, in-processing, and post-processing methods-and discussed their comparative effectiveness, implementation challenges, and the trade-offs they present between model fairness and predictive accuracy. Special attention was given to how these techniques align with regulatory frameworks and ethical standards, particularly in high-stakes financial decision-making.

In addition, the study identified a number of emerging tools and frameworks-such as explainable AI (XAI), federated learning, and human-in-the-loop architectures-that hold promise for making AI systems more transparent, accountable, and adaptable to real-world complexities. These advances support auditing and improving credit scoring algorithms.

Despite this progress, the study revealed several ongoing limitations. Chief among them is the lack of dynamic fairness monitoring systems that can detect and adjust for bias as models interact with evolving data environments. Moreover, current mitigation strategies often fall short in addressing intersectional biases-situations where individuals experience multiple, overlapping forms of disadvantage based on race, gender, socioeconomic status, and other factors.

To effectively confront these issues, the paper underscores the need for sustained interdisciplinary collaboration among AI developers, policymakers, legal scholars, and social scientists. Such collaboration is essential for translating abstract fairness principles into actionable design practices and policy guidelines. The paper strongly advocates for embedding fairness not as a secondary consideration, but as a core tenet throughout the life cycle of AI systems-from data collection and model training to deployment and evaluation.

Meeting this challenge will require more than algorithmic innovation; it will demand a deep ethical commitment and policy coherence to ensure that technological advancements do not exacerbate existing inequalities. Ultimately, fostering fairness in AI-

Received: April 3, 2025. Reviewed: April 16, 2025. Accepted: April 22, 2025.

based credit decision-making is not only a matter of compliance or performance-it is central to building systems that are trustworthy, inclusive, and aligned with broader goals of financial equity and social justice in an increasingly automated world.

REFERENCES

[1] J. Scott, *et al.*, "Revealing and mitigating racial bias and discrimination in financial services," *Journal of Social Equity in Finance*, vol. 11, no. 3, pp. 178–201, 2023.

[2] A. Marshall, *et al.*, "Variable reduction, sample selection bias, and bank retail credit scoring," *Journal of Financial Modeling and Analytics*, vol. 6, no. 1, pp. 56–78, 2010.

[3] S. Priya and R. Kumari, "Loan approval prediction using machine learning," *Journal of Predictive Analytics in Finance*, vol. 7, no. 3, pp. 156–178, 2024.

[4] N. Karimova, "Application of AI in credit risk scoring for small business loans," *Journal of Financial AI Applications*, vol. 10, no. 1, pp. 89–112, 2024.

[5] A. O. Abhulimen, *et al.*, "Ensuring fairness in AI-driven financial services," *International Journal of Financial Ethics*, vol. 12, no. 1, pp. 67–85, 2024.

[6] H. Wang, *et al.*, "An interactive approach to bias mitigation in machine learning," *Journal of Machine Learning Ethics*, vol. 7, no. 2, pp. 178–201, 2021.

[7] A. Khaleghi, "Towards achieving gender equality in automated loan approval processes," *Journal of Financial Technology Ethics*, vol. 9, no. 4, pp. 112–134, 2020.

[8] S. Priya and R. Kumari, "Loan approval prediction using machine learning," *Journal of Applied Data Science*, vol. 8, no. 1, pp. 89–110, 2024.

[9] J. J. X. Wu, "Algorithmic fairness in consumer credit underwriting," *Journal of Financial AI Compliance*, vol. 6, no. 2, pp. 78–96, 2024.

[10] S. Krishnaraj, *et al.*, "Comparing machine learning techniques for loan approval prediction," *Journal of Data-Driven Finance*, vol. 10, no. 1, pp. 245–267, 2024.

[11] A. Dattachaudhuri, *et al.*, "Transparent decision support system for credit risk evaluation," *Neural Networks and Decision Making*, vol. 11, no. 4, pp. 67–89, 2022.

[12] E. Purificato, *et al.*, "Responsible AI techniques in loan approval processes," *AI in Finance Review*, vol. 14, no. 3, pp. 99–121, 2023.

[13] L. Belenguer, "AI bias: Exploring machine-centric solutions," *AI Ethics and Development Quarterly*, vol. 13, no. 2, pp. 112–130, 2022.

[14] A. Castelnovo, "Towards responsible AI in banking," *AI Governance and Ethics Journal*, vol. 9, no. 1, pp. 56–74, 2023.

[15] A. Nadeem, *et al.*, "Gender bias in AI-based decision-making systems," *Journal of Gender and Technology*, vol. 7, no. 3, pp. 198–215, 2022.

[16] C. M. Parra, *et al.*, "Likelihood of questioning AI-based recommendations due to perceived racial/gender bias," *Journal of Social and Ethical AI*, vol. 7, no. 3, pp. 201–224, 2022.

[17] T. Ndayisenga, "Bank loan approval prediction using ML techniques," *Journal of Predictive Analytics in Finance*, vol. 9, no. 3, pp. 145–167, 2021.

[18] B. Low, *et al.*, "Risk-informed and AI-based bias detection using Gen-Z survey data," *Journal of Social Justice and AI*, vol. 10, no. 4, pp. 178–198, 2023.

[19] F. Lorenzo, "Techniques for trustworthy AI in loan approval," *International Journal of Financial Technology*, vol. 8, no. 1, pp. 45–62, 2019.

[20] A. Marshall and R. McManus, "Variable reduction, sample selection bias, and bank retail credit scoring," *Journal of Banking Analytics*, vol. 15, no. 2, pp. 89–115, 2010.

[21] A. Cozarenco and A. Szafarz, "Gender biases in bank lending: Lessons from microcredit in France," *Journal of Microfinance Studies*, vol. 18, no. 1, pp. 67–84, 2015.

[22] A. Kartal, "Mitigating digital discrimination in credit decisions," *AI Fairness Review*, vol. 14, no. 2, pp. 89–112, 2022.

[23] R. Brody, *et al.*, "The potential for biases in resolving loan problems," *Journal of Financial Risk Management*, vol. 13, no. 4, pp. 201–224, 2021.

[24] C. M. Parra, *et al.*, "Likelihood of questioning AI-based recommendations due to perceived racial/gender bias," *Journal of Ethical AI Practices*, vol. 11, no. 1, pp. 145–167, 2022.

Received: April 3, 2025. Reviewed: April 16, 2025. Accepted: April 22, 2025.

[25] N. Zhou, *et al.*, "Bias, fairness, and accountability in AI/ML," *Journal of Fairness in Machine Learning*, vol. 8, no. 1, pp. 123–145, 2021.
[26] H. Wang, *et al.*, "An interactive approach to bias mitigation in machine learning," *Journal of Machine Learning Fairness*, vol. 8, no. 3, pp. 156–178, 2021.
[27] T. Ndayisenga, "Bank loan approval prediction using ML techniques," *Journal of Predictive Analytics in Finance*, vol. 9, no. 3, pp. 145–167, 2021.
[28] T. Beck, P. Behr, and A. Madestam, "Sex and credit: Is there a gender bias in lending?," *Journal of Financial Inclusion Studies*, vol. 9, no. 2, pp. 112–134, 2017.
[29] R. Brody, *et al.*, "The potential for biases in resolving loan problems," *Journal of Credit Risk Studies*, vol. 11, no. 4, pp. 201–224, 2021.
[30] D. E. Rigobon, "From Utilitarian to Rawlsian Designs for Algorithmic Fairness," *arXiv preprint arXiv:2302.03567*, 2023.
[31] D. Card and N. A. Smith, "On Consequentialism and Fairness," *Frontiers in Artificial Intelligence*, vol. 3, p. 34, 2020.
[32] J. Rawls, *A Theory of Justice*, revised ed., Harvard University Press, 1999.
[33] S. Buijsman, "Navigating Fairness Measures and Trade-offs," *AI and Ethics*, vol. 4, pp. 1323–1334, 2024.
[34] J. Kuppler et al., "Formalizing Trade-offs Beyond Algorithmic Fairness: Lessons from Ethical Philosophy and Welfare Economics," *AI and Ethics*, vol. 1, pp. 529–544, 2021.
[35] N. Thieme, "Will Computer Algorithms Support Equity or Reinforce Inequality?," *Undark Magazine*, 2018.
[36] J.-C. Bélisle-Pipon, "AI, Universal Basic Income, and Power: Symbolic Violence in the Tech Elite's Narrative," *Frontiers in Artificial Intelligence*, vol. 3, article 1488457, 2025.

Received: April 3, 2025. Reviewed: April 16, 2025. Accepted: April 22, 2025.