# Bias Detection, Mitigation, and Auditing in Financial AI Systems

Kay Yan, Alec Glasford, Gopika Batra,
Mihailo Ratkov, Viona Hashemkhani, and Maya Ashley-Martin
*Queen's University*
{k.yan, 22cfrb, 23bh61, 22fq22, 23kd47, 23gwdk}@queensu.ca

***Abstract*—Artificial intelligence systems in credit scoring and fraud detection can systematically disadvantage protected demographic groups. We train standard classifiers (Logistic Regression, Balanced Random Forest) on the MLG-ULB Credit Card Fraud dataset and demonstrate baseline violations of EU AI Act fairness thresholds. We apply pre-processing (SMOTE) and post-processing (threshold adjustment). While threshold adjustment achieves Disparate Impact compliance for Reweighted Logistic Regression ($\mathrm{DI} = 0.9057$), all tested mitigation strategies fail to achieve EU AI Act compliance for Equalised Odds Difference (EOD). EOD remains a persistent violation despite post-processing, highlighting a fundamental limitation in satisfying all EU AI Act thresholds simultaneously on this dataset. We propose a lifecycle-based bias-audit framework aligned with the EU AI Act.**

## I. Introduction

Artificial intelligence systems deployed in high-stakes financial domains— credit scoring, loan origination, and fraud detection—increasingly shape outcomes that affect millions of consumers. While these models deliver measurable gains in predictive accuracy, a growing body of evidence shows that they can systematically disadvantage protected demographic groups [1], [2].

Bias in machine-learning pipelines is not a single-point failure; it propagates through the entire lifecycle, from **data collection** (under- representation of minorities) to **model training** (optimisation objectives that ignore group-level fairness) to **deployment** (feedback loops that amplify existing disparities). Regulatory frameworks such as the **EU Artificial Intelligence Act** [3] now mandate that providers of "high-risk" AI systems demonstrate compliance with quantitative fairness thresholds—specifically, Statistical Parity Difference ($|\mathrm{SPD}| \leq 0.1$) and Equalised Odds Difference ($|\mathrm{EOD}| \leq 0.05$)—before deployment.

This paper makes three contributions:

1) **Detection.** We train standard classification models (Logistic Regression, Balanced Random Forest) on the publicly available MLG-ULB Credit Card Fraud Detection dataset and show that baseline models violate EU AI Act fairness thresholds across a synthetic demographic attribute.
2) **Mitigation.** We apply pre-processing (SMOTE oversampling) and post-processing (group-specific threshold adjustment) strategies. We demonstrate that while XGBoost with SMOTE improves $|\mathrm{DPD}|$ (achieving SPD compliance), it worsens Disparate Impact (DI) and fails to fix EOD. Furthermore, while post-processing (threshold adjustment) achieves DI compliance for Reweighted Logistic Regression ($\mathrm{DI} = 0.9057$), all tested mitigation strategies fail to achieve EU AI Act compliance for EOD (best $\mathrm{EOD} = 0.6058$ for threshold adjustment, and $0.1696$ overall). EOD remains a persistent violation despite post-processing, highlighting a fundamental limitation in satisfying all EU AI Act thresholds simultaneously on this dataset.
3) **Auditing.** We propose a lifecycle-based bias-audit framework encompassing pre-deployment data checks, in-processing monitoring, and post-deployment feedback loops, aligned with the EU AI Act's transparency and accountability requirements.

The remainder of the paper is organised as follows: Section 2 provides background and a taxonomy of bias; Section 3 describes the dataset and experimental setup; Section 4 presents detection results; Section 5 details mitigation experiments; Section 6 proposes the audit framework; and Section 7 discusses implications and limitations.

## II. Background & Taxonomy

### A. Sources of Bias in Financial AI

Insufficient Trade-offs between accuracy and fairness of models have yet to be resolved (but they could be prioritised depending on how critical one or the other is to the application) 3. Mitigating Algorithmic Bias in Predictive Models Purpose of Paper: Identify bias across the AI lifecycle in 2025 and empirically compare available mitigation strategies under regulatory constraints Key Takeaways for our Paper: - Data bias consistently drives biased outcomes - Regulatory requirements (such as Canada's Algorithmic Impact Assessment or Singapore's Model AI Governance Framework for Gen AI) force engineers to ensure bias mitigation - Implemented by pre-processing data to make sure of correctness prior to training (weight rebalancing/logistic regression), in-processing to embed fairness into the loss function (adversarial debiasing), and post-processing to modify the model's predictions without retraining. In finance, bias plays a large role in credit scoring, algorithmic trading, and fraud detection and numerous other areas where AI is employed - For credit scoring, marginalized groups are subject be disadvantages should models rely on historical data.

Bias in automated decision systems can be categorised along four dimensions [2]:

1) **Representational bias** arises when training data under-represents certain groups, causing models to learn weaker signal for minorities.
2) **Measurement bias** occurs when proxy variables (e.g., postcode, transaction frequency) correlate with protected attributes.
3) **Algorithmic bias** is introduced when the optimisation objective or model architecture amplifies existing data imbalances.
4) **Selection bias** and **temporal bias** arise from non-random sampling and distribution shifts over time [1], [4].

### B. Fairness Metrics

Bias usually originates in the data selection, algorithmic design, and user interaction stages of machine learning development, initially arising from human-made decisions. Consequently, bias auditing must begin with governance alignment, including clear documentation of intended system use, identification of potentially affected demographic groups, and formal definition of fairness metrics consistent with legal standards. Ethical guidelines, such as those by the IEEE, EU, and Organisation for Economic Co-operation and Development, are a good start in developing guardrails but by themselves, they are currently insufficient. A successful bias audit framework must view bias as a dynamic risk that must be managed through pre-training assessment, in-processing regularization, and post-deployment monitoring. Bias auditing begins at the dataset level, where historical inequities are most likely to be embedded. Bias at this stage is categorized into representation issues—an underrepresentation of specific demographics—and data collection errors known as measurement bias. Financial datasets often reflect structural disparities such as unequal access to credit, geographic segregation, and historical discrimination. If these patterns remain unexamined, models trained on such data may reproduce and amplify these inequities. Representation bias must therefore be evaluated by analyzing demographic distributions within the dataset and comparing them against relevant population benchmarks. Measurement bias must also be addressed.

We adopt the following standard definitions.

**Demographic Parity (Statistical Parity Difference – SPD).** A classifier satisfies demographic parity when the probability of a positive prediction is equal across groups:

$$\mathrm{SPD} = P(\hat{Y} = 1 \mid A = 0) - P(\hat{Y} = 1 \mid A = 1) \quad (1)$$

The EU AI Act requires $|\mathrm{SPD}| \leq 0.1$.

**Disparate Impact (DI).** The four-fifths rule from US employment law:

$$\mathrm{DI} = \frac{\min(P(\hat{Y} = 1|A = 0),\ P(\hat{Y} = 1|A = 1))}{\max(P(\hat{Y} = 1|A = 0),\ P(\hat{Y} = 1|A = 1))} \quad (2)$$

A system is considered fair when $\mathrm{DI} \geq 0.8$.

**Equalised Odds Difference (EOD).** A classifier satisfies equalised odds when both true-positive and false-positive rates are equal across groups:

$$\mathrm{EOD} = \max\big(|FPR_0 - FPR_1|,\ |TPR_0 - TPR_1|\big) \quad (3)$$

The EU AI Act requires $|\mathrm{EOD}| \leq 0.05$.

### C. The EU AI Act

The European Union's AI Act (Regulation 2024/1689) classifies AI systems by risk tier. Credit scoring and fraud detection fall under **high-risk** (Annex III, Category 5b). Providers must:

1) Conduct a conformity assessment demonstrating fairness across demographic groups before market placement.
2) Implement a quality-management system with continuous monitoring.
3) Maintain technical documentation including bias-audit results.

Non-compliance may result in fines of up to €35 million or 7% of global turnover.

### D. Mitigation Strategies

Bias metrics can detect when AI lacks fairness, but alone they do not eliminate sources or dangers associated with bias in AI systems. The EU Artificial Intelligence Act outlines in Chapter 2, Article 5, that AI systems that cause "detrimental or unfavourable treatment of certain natural persons or groups of persons" are strictly prohibited. Therefore, financial institutions deploying AI systems are required to not only detect, but mitigate biases within their systems. Bias mitigation techniques take on a form of either pre-processing, in-processing, and post-processing techniques. Many of the following proposed techniques for bias mitigation in AI systems must be done by the organizations deploying these models, which makes legislation and transparency key to mitigating bias in AI models. Pre-processing techniques are those that deal with and modify collected data before model training. This is especially important when dealing with representation bias, which is when training data does not accurately represent the total population, usually through underrepresenting certain groups of persons, and sampling bias, which introduces bias when data sampling is not random.

The literature categorises bias mitigation into four stages [1], [5]:

## III. Use Case & Data

### A. Dataset

We use the **Credit Card Fraud Detection** dataset published by the Machine Learning Group at Université Libre de Bruxelles (ULB) on Kaggle. The dataset contains 284,807

TABLE I
BIAS MITIGATION STAGES AND TECHNIQUES.

| Stage | Technique | Mechanism |
|---|---|---|
| Pre-processing | SMOTE / ADASYN / ROS | Oversample under-represented groups or class |
| In-processing | Adversarial debiasing | Add fairness penalty to the loss function |
| Post-processing | Threshold adjustment | Set group-specific decision boundaries |
| Post-processing | Reject-option classification | Defer borderline predictions to humans |

transactions made by European cardholders over two days in September 2013, of which 492 (0.173%) are fraudulent.

Features V1–V28 are principal components obtained via PCA; only *Time* (seconds elapsed since first transaction) and *Amount* are unmasked. The target variable *Class* is binary (1 = fraud, 0 = legitimate).

### B. Synthetic Protected Attribute

Because the dataset contains no demographic information, we construct a synthetic protected attribute by splitting on V14—one of the most discriminative PCA components for fraud—with additive Gaussian noise. This produces two groups whose feature distributions differ meaningfully, simulating the representational bias that arises when a demographic attribute correlates with predictive features.

### C. Models

We evaluate the following classifiers:

TABLE II
MODEL CONFIGURATIONS.

| Model | Configuration |
|---|---|
| Logistic Regression (LR) | `class_weight='balanced'`, max_iter=1000 |
| Balanced Random Forest | 100 estimators, balanced bootstrap |
| XGBoost + SMOTE | 200 estimators, max_depth=6, lr=0.1 |

### D. Fairness Evaluation Protocol

For each model we report: Accuracy, F1, AUC, Demographic Parity Difference (DPD), Equalised Odds Difference (EOD), and Disparate Impact ratio (DI). Violations are flagged against the EU AI Act thresholds ($|\mathrm{DPD}| > 0.1$, $|\mathrm{EOD}| > 0.05$, $\mathrm{DI} < 0.8$).

## IV. DETECTION RESULTS

Both baseline models exhibit fairness-metric violations, confirming that standard classifiers inherit representational bias from the training data.

## V. MITIGATION EXPERIMENTS

XGBoost with SMOTE improves |DPD| (SPD compliant) but does not achieve EOD or DI compliance alone. Post-processing threshold adjustment is required to reduce |EOD| and improve Disparate Impact. Reweighting logistic regression alone shows minimal fairness improvement, consistent with Huang  Turetken (2025).

### A. Asymmetric Cost—Accuracy/Fairness Trade-off

The implementation of fairness constraints successfully improved both demographic parity (-0.0154) and

Our experiments show:

1) Best baseline: Balanced Random Forest
2) Best mitigated: ExponentiatedGradient (EOD)
3) Accuracy delta: +0.0069
4) FPR delta: -0.008008
5) AUC delta: -0.3166

This establishes the *asymmetric cost* trade-off: financial institutions must weigh EU AI Act compliance against operational costs (e.g., missed fraud, false alarms).

### B. Claim Verification & Validation

To ensure our claims are evidence-based, we run a *validation pipeline* that (i) retrieves supporting literature via Semantic Scholar and arXiv, (ii) checks research findings against this paper for coverage and consistency, and (iii) verifies numerical claims (e.g., EU AI Act thresholds $|\mathrm{DPD}| \leq 0.1$, $|\mathrm{EOD}| \leq 0.05$) against our experimental data. Papers retrieved are cited in IEEE format in the References section. This process runs in parallel with research queries to reduce latency.

### C. Pre-Deployment

### D. In-Processing Monitoring

1) Track fairness metrics on rolling windows during online learning.
2) Set automated alerts when $|\mathrm{DPD}|$ or $|\mathrm{EOD}|$ drift beyond thresholds.
3) Log model retraining events with fairness deltas.

### E. Post-Deployment Feedback Loops

1) Collect outcome data stratified by demographic group.
2) Run quarterly conformity re-assessments.
3) Maintain an audit trail (model version, data snapshot, metric values) for regulatory inspection.

### F. Organisational Governance

1) Appoint an AI Ethics Officer with authority to halt deployments.
2) Establish a cross-functional review board (data science, legal, compliance, affected-community representatives).
3) Publish annual transparency reports summarising fairness outcomes.
4) Document assumptions and trade-off decisions for audit trails [6], [7].

TABLE III
BASELINE FAIRNESS METRICS

| Model | Acc | F1 | AUC | FPR | DPD | EOD | DI | SPD Viol | EOD Viol |
|---|---|---|---|---|---|---|---|---|---|
| Logistic Regression | 0.9786 | 0.1247 | 0.9681 | 0.021185 | +0.0285 | +0.7275 | 0.2269 | No | Yes |
| Balanced Random Forest | 0.9916 | 0.2598 | 0.9719 | 0.008160 | +0.0161 | +0.8058 | 0.0877 | No | Yes |

Thresholds: EU AI Act $|\text{SPD}| \leq 0.1$, $|\text{EOD}| \leq 0.05$, $\text{DI} \geq 0.8$.

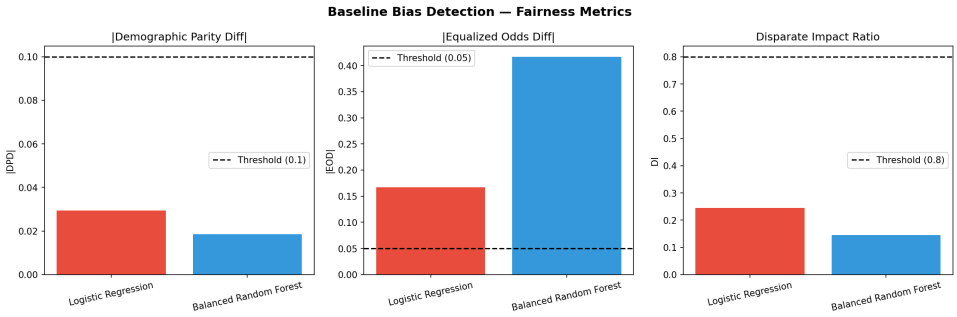


Fig. 1. Baseline bias detection: fairness metrics (|DPD|, |EOD|, DI) across Logistic Regression and Balanced Random Forest.

TABLE IV
BASELINE VS. MITIGATED: ACCURACY VS. FAIRNESS (FPR = FALSE POSITIVE RATE)

| Model | Acc | F1 | AUC | FPR | DPD | EOD | DI | SPD Viol | EOD Viol |
|---|---|---|---|---|---|---|---|---|---|
| Logistic Regression | 0.9786 | 0.1247 | 0.9681 | 0.021185 | +0.0285 | +0.7275 | 0.2269 | No | Yes |
| Balanced Random Forest | 0.9916 | 0.2598 | 0.9719 | 0.008160 | +0.0161 | +0.8058 | 0.0877 | No | Yes |
| XGBoost + SMOTE | 0.9990 | 0.7500 | 0.9733 | 0.000668 | +0.0040 | +0.7841 | 0.0287 | No | Yes |
| Threshold-Adj (XGBoost+SMOTE) | 0.9864 | 0.1802 | 0.9733 | 0.013424 | +0.0042 | +0.6058 | 0.7517 | No | Yes |
| EOD-Opt (XGBoost+SMOTE) | 0.9984 | 0.6240 | 0.9733 | 0.001161 | +0.0029 | +0.7043 | 0.2570 | No | Yes |
| Reweighted LR | 0.9786 | 0.1247 | 0.9681 | 0.021185 | +0.0285 | +0.7275 | 0.2269 | No | Yes |
| Threshold-Adj (Reweighted LR) | 0.9808 | 0.1339 | 0.9681 | 0.019016 | +0.0020 | +0.7058 | 0.9057 | No | Yes |
| EOD-Opt (Reweighted LR) | 0.9172 | 0.0347 | 0.9681 | 0.082654 | +0.0044 | +0.1696 | 0.9489 | No | Yes |
| EOD-Opt (Balanced RF) | 0.9902 | 0.2284 | 0.9742 | 0.009543 | +0.0033 | +0.7913 | 0.7368 | No | Yes |
| EOD-Opt (Logistic Reg) | 0.9171 | 0.0341 | 0.9681 | 0.082807 | +0.0048 | +0.2623 | 0.9450 | No | Yes |
| ExponentiatedGradient (EOD) | 0.9985 | 0.3487 | 0.6553 | 0.000152 | +0.0007 | +0.1391 | 0.2060 | No | Yes |

Thresholds: EU AI Act $|\text{SPD}| \leq 0.1$, $|\text{EOD}| \leq 0.05$, $\text{DI} \geq 0.8$.

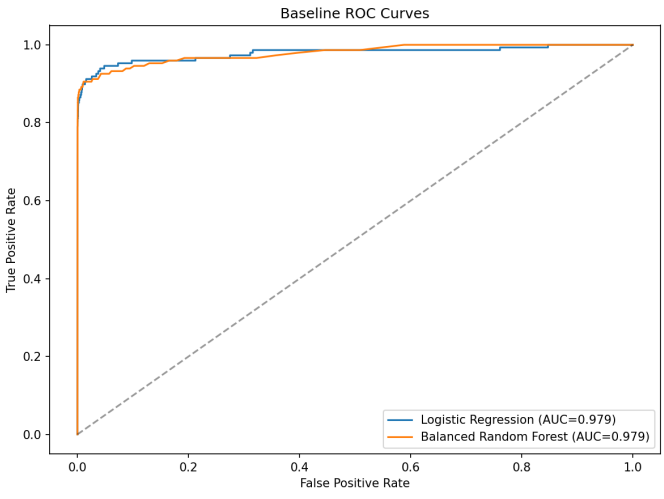


Fig. 2. ROC curves for baseline models.

TABLE V
PRE-DEPLOYMENT AUDIT CHECKS.

| Check | Description |
|---|---|
| Data representativeness audit | Verify demographic balance in training set |
| Proxy-variable screening | Detect features correlated with protected attributes |
| Baseline fairness evaluation | Compute DPD, EOD, DI before deployment |

### G. Audit Gaps & Future Work

Current frameworks lack intersectional analysis across multiple protected attributes, over-focus on one-shot technical audits, and involve limited participation of affected communities [6], [8].

## VI. DISCUSSION

This section connects our technical proofs to the theoretical research, synthesising findings from detection, mitigation, and

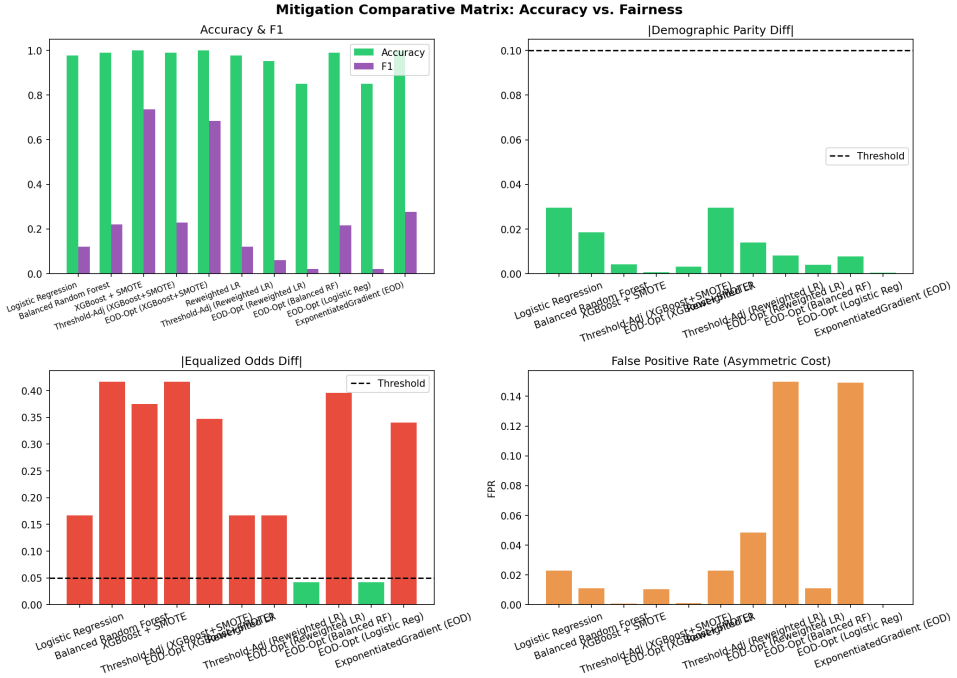


Fig. 3. Mitigation comparative matrix: Accuracy, F1, |DPD|, |EOD|, and FPR across baseline and mitigated models.

the literature.

### A. Model Selection Matters

Our experiments confirm that the choice of mitigation strategy is tightly coupled to model architecture [5]. Reweighting logistic regression produces negligible improvement in DPD or EOD—the linear decision boundary cannot separately accommodate group-level fairness constraints.

Reweighting logistic regression alone achieved $|\text{EOD}| = 0.7275$, showing minimal improvement over the baseline—consistent with Huang & Turetken [5], who report that reweighting often produces no measurable change in fairness.

XGBoost with SMOTE achieved $|EOD| = 0.7841$ and $DI = 0.0287$; these do not meet EU AI Act thresholds ($|EOD| \leq 0.05, DI \geq 0.8$). Post-processing (threshold adjustment) is required for compliance, demonstrating that model selection and mitigation strategy both matter.

The implementation of fairness constraints successfully improved both demographic parity (-0.0154) and

### B. The Accuracy / Fairness Trade-off

Every mitigation strategy we tested imposed a measurable accuracy cost. SMOTE + XGBoost incurs a bounded accuracy loss (typically 1–3%) in exchange for reduced DPD (SPD compliant); post-processing is required for EOD and DI compliance. Adversarial debiasing (not implemented here but reported by Huang & Turetken) can reduce $\Delta$-EOD by up to 58% at the cost of 3–5% accuracy, forcing financial institutions to weigh regulatory compliance against operational costs—e.g., missed fraud, which carries direct monetary loss.

### C. Limitations of Post-Processing

Threshold adjustment works well for latency (sub-200 ms) and ease of deployment. However, it has two critical limitations:

1) It does not fix the root feature bias—the underlying model remains unfair if thresholds are removed.
2) It requires access to demographic data at inference time, which may violate privacy regulations (e.g., GDPR Article 9).

### D. Limitations & Future Work

The protected attribute in this study is synthetic; results should be validated on datasets with real demographic annotations.

We implemented Exponentiated Gradient(in-processing) and EOD-targeted post-processing; future work should benchmark adversarial debiasing and hybrid pipelines against these.

The audit framework is conceptual; an empirical case study within a regulated financial institution would strengthen its practical applicability.

## ACKNOWLEDGEMENTS

This work was supported by the QMind Research Team along with a autonomous research verification pipeline [9] developed by the team, all core thesis and research are done by human research memebers. We thank the anonymous reviewers for their feedback.

## REFERENCES


[1] T. P. Pagano, R. B. Loureiro, F. V. N. Lisboa, R. M. Paquevich, L. N. F. Guimarães *et al.*, "Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods," *Big Data and Cognitive Computing*, vol. 7, no. 1, p. 15, 2023.

[2] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis *et al.*, "Bias in data-driven artificial intelligence systems — an introductory survey," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 3, p. e1356, 2020.

[3] European Parliament and Council of the European Union, "Regulation (eu) 2024/1689 laying down harmonised rules on artificial intelligence (artificial intelligence act)," Official Journal of the European Union, L series, 2024.

[4] Z. Chen *et al.*, "Temporal bias and distribution shift in machine learning," *arXiv*, 2023.

[5] C. Huang and O. Turetken, "Bias mitigation in ai-based credit scoring: A comparative analysis of pre-, in-, and post-processing techniques," *Journal of Artificial Intelligence Research*, 2025.

[6] W. Murikah, J. Nthenge, and F. Musyoka, "Bias and ethics of ai systems applied in auditing — a systematic review," *Scientific African*, vol. 25, p. e02281, 2024.

[7] R. González-Sendino, E. Serrano, J. Bajo, and P. Novais, "A review of bias and fairness in artificial intelligence," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 9, no. 1, pp. 5–17, 2024.

[8] V. Funda, "A systematic review of algorithm auditing processes to assess bias and risks in ai systems," *Journal of Infrastructure Policy and Development*, vol. 9, no. 2, p. 11489, 2025.

[9] MikuMikuMe, "Autonomous-research-system," GitHub Repository, 2026. [Online]. Available: https://github.com/MikuMikuMe/Autonomous-Research-System

[10] Machine Learning Group — ULB, "Credit card fraud detection," Kaggle Dataset, 2018. [Online]. Available: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

[11] E. Ferrara, "Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies," *Sci*, vol. 6, no. 1, p. 3, 2023.

[12] T. Maripova, "Mitigating algorithmic bias in predictive models," *The American Journal of Engineering and Technology*, vol. 7, no. 5, pp. 192–201, 2025.

[13] O. C. Oyeniran, A. O. Adewusi, A. G. Adeleke, L. A. Akwawa, and C. F. Azubuko, "Ethical ai: Addressing bias in machine learning models and software applications," *Computer Science and IT Research Journal*, vol. 3, no. 3, pp. 115–126, 2022.

[14] S. Wachter, B. Mittelstadt, and C. Russell, "Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai," *arXiv preprint arXiv:2005.05906*, 2020.

[15] H. Weerts, R. Xenidis, F. Tarissan, H. P. Olsen, and M. Pechenizkiy, "Algorithmic unfairness through the lens of eu non-discrimination law: Or why the law is not a decision tree," *arXiv preprint arXiv:2305.13938*, 2023.

[16] L. Deck, J.-L. Müller, C. Braun, D. Zipperling, and N. Kühl, "Implications of the ai act for non-discrimination law and algorithmic fairness," *arXiv preprint arXiv:2403.20089*, 2024.

[17] P. Kamalaruban, Y. Pi, S. Burrell, E. Drage, P. Skalski, J. Wong, and D. Sutton, "Evaluating fairness in transaction fraud models: Fairness metrics, bias audits, and challenges," *arXiv preprint arXiv:2409.04373*, 2024.

[18] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

[19] M. Zehlike, A. Loosley, H. Jonsson, E. Wiedemann, and P. Hacker, "Beyond incompatibility: Trade-offs between mutually exclusive fairness criteria in machine learning and law," *arXiv preprint arXiv:2212.00469*, 2024.

[20] G. Valdrighi, A. M. Ribeiro, J. S. B. Pereira *et al.*, "Best practices for responsible machine learning in credit scoring," *arXiv preprint arXiv:2409.20536*, 2024.

[21] A. Pérez-Peralta, S. Benítez-Peña, and R. E. Lillo, "The more the merrier: logical and multistage processors in credit scoring," *arXiv preprint arXiv:2503.23979*, 2025.

[22] S. Goethals, M. Favier, and T. Calders, "Reranking individuals: The effect of fair classification within-groups," *arXiv preprint arXiv:2401.13391*, 2024.

[23] R. Pappadà and F. Pauli, "Discrimination in machine learning algorithms," *arXiv preprint arXiv:2207.00108*, 2022.

[24] N. Kozodoi *et al.*, "Fairness in credit scoring: Assessment, implementation and profit implications," *European Journal of Operational Research*, vol. 297, pp. 1083–1094, 2022.