AI & SOCIETY (2026) 41:2783–2801
https://doi.org/10.1007/s00146-025-02452-1

RESEARCH

A fairness-focused approach to recidivism prediction: implications for accuracy, trust, and equity

Michael Mayowa Farayola[1] · Irina Tal[1] · Takfarinas Saber[2] · Regina Connolly[1] · Malika Bendechache[2]

Received: 10 February 2025 / Accepted: 20 June 2025 / Published online: 12 July 2025

Abstract

Fairness in AI has emerged as a critical priority, particularly in high-stakes domains like criminal justice, where algorithmic decisions profoundly affect individuals and demographic groups. This study addresses persistent challenges of bias in recidivism prediction by evaluating the integrated application of fairness-enhancing techniques across the AI pipeline. Specifically, we analyze the integrated use of Reweighing, Adversarial Learning, Disparate Impact Remover, Exponentiated Gradient Reduction, Reject Option Classification, and Equalized Odds optimization. These methods are assessed using key fairness metrics, Statistical Parity Difference (SPD), Disparate Impact (DI), Equal Opportunity Difference (EOD), and Predictive Equality Difference (PED), on two real-world recidivism datasets: COMPAS, a well-known publicly available dataset, and RisCanvi, a curated English dataset derived from a publicly available Spanish dataset. Our findings reveal trade-offs between fairness improvements and predictive accuracy, with fairness performance varying across integrated mitigation strategies. While individual techniques often fall short of effectively reducing bias, their integration across pre-processing, in-processing, and post-processing stages demonstrates a more comprehensive and robust capacity to address fairness concerns. In some instances, fairness mitigation strategies exacerbate disparities instead of reducing them, highlighting the complexity of achieving equitable AI outcomes. Some methods struggle with specific datasets, leading to worsened SPD, DI, EOD, or PED values, potentially reinforcing biases rather than eliminating them. Using multi-objective and bi-objective optimization frameworks, we identify optimal configurations that balance fairness and accuracy, contributing to a deeper understanding of fairness–accuracy trade-offs in AI systems. This study makes a novel contribution by systematically evaluating multi-phase fairness integration in recidivism prediction, bridging critical gaps in the literature. However, our results also underscore the risks of ineffective or counterproductive fairness interventions when not carefully tailored. Our findings highlight the necessity of tailoring bias mitigation strategies to specific datasets and contexts, providing actionable insights for developing equitable, diverse, and inclusive AI systems that are fair and reliable in the criminal justice system.

Keywords Fairness · Artificial intelligence · Criminal justice system · Recidivism · Trust · Multi-objective optimization · Equity

✉ Michael Mayowa Farayola
michael.farayola2@mail.dcu.ie

Irina Tal
irina.tal@dcu.ie

Takfarinas Saber
takfarinas.saber@universityofgalway.ie

Regina Connolly
regina.connolly@dcu.ie

Malika Bendechache
malika.bendechache@universityofgalway.ie

[1] Dublin City University, Dublin, Ireland

[2] University of Galway, Galway, Ireland

1 Introduction

Artificial intelligence (AI) is increasingly being integrated into the criminal justice system, particularly for predicting recidivism risk-the likelihood that an individual will re-offend upon release (Desmarais et al. 2016; Green 2020; Rodolfa et al. 2020; Farayola et al. 2023). This growing reliance on AI is driven by its potential to enhance judicial decision-making by improving efficiency, consistency, and objectivity, thereby mitigating human biases and promoting equitable treatment (Berk and Bleich 2014; Wang et al. 2022; Pelegrina et al. 2023). However, the application of AI in high-stakes legal decisions raises critical concerns

regarding fairness, ethics, and public trust. A central issue is the risk of perpetuating historical and systemic biases, which disproportionately impact marginalized demographic groups, ultimately undermining the principles of fairness and equality that are foundational to the criminal justice system.

Fairness is recognized as an essential requirement for trustworthy AI, as outlined by the European Commission (AI 2019; Thiebes et al. 2021). Fairness is particularly crucial in the criminal justice domain, where AI-driven decisions carry profound and lasting consequences for individuals and society. Ensuring that AI systems treat all demographic groups equitably is imperative to prevent outcomes that unfairly disadvantage specific populations. Consequently, fairness in AI models predicting recidivism risk is not merely a technical challenge but a broader societal and ethical necessity that aligns with the fundamental tenets of justice and equity.

AI-driven recidivism models analyze diverse datasets, including criminal records, socioeconomic backgrounds, demographic attributes, and psychological factors, to generate risk scores that inform judicial decisions on bail, sentencing, parole, and rehabilitation. Ideally, these models promote consistency and evidence-based decision-making. However, in practice, they often replicate or amplify the existing social inequities. For example, studies have shown that tools such as the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) systematically overestimate recidivism risk for Black individuals while underestimating it for White individuals (Angwin et al. 2016). Such disparities can result in disproportionately punitive outcomes for marginalized populations, reinforcing cycles of poverty, incarceration, and social exclusion. These systemic biases highlight the urgent need for frameworks that enhance the equity and trustworthiness of AI-based recidivism prediction models.

Bias in AI systems originates from multiple sources, including historical data, algorithmic design, and deployment practices (Ferrara 2023; Farayola et al. 2023). Historical data often reflect entrenched societal inequalities, and AI models trained on such data may inadvertently learn and perpetuate these biases. Moreover, algorithmic design choices and deployment practices can further exacerbate disparities. In response, the literature has proposed various bias mitigation strategies categorized into three key phases: pre-processing, in-processing, and post-processing. Pre-processing techniques focus on correcting biases in datasets before training, ensuring a more equitable foundation. In-processing methods modify training algorithms to counteract biases during development, while post-processing approaches adjust model outputs to promote fairer outcomes (Pagano et al. 2023; Berk et al. 2023; Wan et al. 2023).

Despite these efforts, existing approaches often fall short in addressing fairness comprehensively. A key limitation is the tendency to apply fairness-enhancing techniques in isolation, targeting only one phase-pre-processing, in-processing, or post-processing—rather than considering their interconnected nature (Jain et al. 2019; Corbett-Davies et al. 2017; Jain et al. 2020; Biswas and Mukherjee 2021; Kobayashi and Nakao 2020; Miron et al. 2021). This fragmented approach overlooks the dynamic nature of bias, which can manifest and propagate across multiple stages of model development. Consequently, isolated mitigation strategies often fail to ensure fairness across the entire AI lifecycle. Moreover, this segmented approach risks undermining stakeholder trust, as addressing bias at one stage does not guarantee fairness throughout the system. To achieve equitable outcomes, rebuild trust, and effectively mitigate bias, an integrated approach is necessary—one that accounts for fairness considerations across all stages of AI system development.

Another key challenge is the trade-off between fairness and predictive accuracy. Fairness-enhancing techniques can sometimes reduce predictive accuracy, a critical concern in the criminal justice domain where erroneous predictions may have severe consequences, such as wrongful detention or premature release (Zhang and Sun 2022; Agarwal et al. 2018). Addressing this fairness–accuracy trade-off is essential to developing AI models that are not only equitable but also reliable and practical for real-world decision-making.

The deployment of biased AI models in the criminal justice system carries profound ethical implications, with far-reaching societal consequences (Angwin et al. 2016). Biased recidivism predictions can reinforce the existing disparities, shaping how individuals are perceived and treated within and beyond the judicial system. For example, being classified as "high risk for recidivism" may lead to stigmatization, limiting opportunities for employment, housing, and education (Jacobs and Gottlieb 2020; Graffam et al. 2014). This, in turn, can increase an individual's likelihood of re-offending, creating a self-fulfilling cycle of disadvantage.

Criminal justice stakeholders, including judges, parole officers, policymakers, and civil rights advocates, hold divergent views on AI's role in judicial decision-making. While some regard AI as a tool for enhancing consistency and efficiency, others raise concerns about its fairness and predictive reliability (Watamura et al. 2025; Farber 2024). These concerns underscore the need for AI models that balance fairness and accuracy. Achieving this balance is a complex challenge that requires integrating high predictive performance with equitable treatment across diverse populations. This study investigates a multi-stage approach that applies fairness interventions across all phases of the AI

modeling process, aiming to determine whether fairness can be improved without significantly compromising accuracy.

To validate this approach, we leverage two widely used recidivism datasets: the RisCanvi dataset and the COMPAS dataset. Validation across multiple datasets is crucial in this field, where the ethical and societal implications of AI predictions are significant. Demonstrating that fairness-enhancing techniques yield consistent and reliable outcomes across diverse datasets strengthens confidence in their real-world applicability. Specifically, we evaluate the integrated use of fairness-enhancing methods such as Reweighing (Krasanakis et al. 2018), Adversarial Learning (Wadsworth et al. 2018), Disparate Impact Remover (Feldman et al. 2015), Exponentiated Gradient Reduction (Agarwal et al. 2018), Reject Option Classification (Kamiran et al. 2012), and Equalized Odds optimization (Pleiss et al. 2017).

This study hypothesizes that integrating multiple fairness-enhancing techniques, each addressing biases at different stages of AI development, will provide a comprehensive and robust framework for mitigating bias throughout the AI lifecycle. By testing this approach on multiple datasets, we contribute to the broader discourse on AI fairness, exploring whether fairness-enhancing measures can maintain predictive stability while improving equity. This approach is motivated by limitations identified in prior research and the pressing need for a holistic strategy to enhance algorithmic fairness and rebuild public trust in AI-based recidivism risk assessment (Farayola et al. 2023; Caton and Haas 2023; Chakraborty et al. 2020; Raza et al. 2023; Chen et al. 2023; Zhang and Sun 2022; Farayola et al. 2024).

To the best of our knowledge, this paper is the first comprehensive study to systematically integrate and evaluate multiple fairness-enhancing techniques across all three phases of fairness intervention-pre-processing, in-processing, and post-processing-using two recidivism datasets, with a focus on fairness in recidivism prediction within the criminal justice system. We categorize these integrations into four distinct model types: PI, PP, IP, and PIP. The PI model combines pre-processing and in-processing techniques; the PP model incorporates pre-processing and post-processing methods; the IP model integrates in-processing and post-processing strategies; and the PIP model unifies methods across all three phases. This classification provides a structured framework for systematically assessing each model's impact on fairness and predictive accuracy.

Additionally, we employ a multi-objective optimization (MOO) methodology (Sharma and Kumar 2022) using the Pareto front concept to identify optimal models among the PI, PP, IP, and PIP configurations. This framework identifies non-dominated models, those that excel in at least one fairness metric, such as statistical parity, disparate impact, or equal opportunity, while maintaining strong predictive performance. By highlighting trade-offs between fairness and accuracy, the Pareto front provides a nuanced perspective on the effectiveness of each fairness-enhancing approach.

Our methodology offers criminal justice stakeholders actionable insights by presenting models that balance fairness and accuracy. By showcasing models along the Pareto front, stakeholders can select the most appropriate approach based on their priorities—whether emphasizing fairness, accuracy, or a balanced trade-off. Ultimately, this framework ensures that AI-driven recidivism risk assessments are optimized for equity and reliability, thereby enhancing the fairness and credibility of high-stakes judicial decision-making.

1.1 Research questions and hypothesis

This study contributes to the ongoing discourse on AI fairness by addressing the fairness-accuracy trade-off identified in Liu and Vicente (2022), Plecko and Bareinboim (2024), Mehrabi et al. (2021), Farayola et al. (2024). It aims to validate the applicability and generalizability of a multi-stage fairness-improving approach, integrating techniques across pre-processing, in-processing, and post-processing phases, on recidivism datasets, especially the RisCanvi dataset. This validation is essential to assess the robustness and applicability of fairness techniques in diverse, real-world contexts, particularly within high-stakes domains like criminal justice.

The central hypothesis of this study posits that combining fairness-improving techniques across different stages of the AI pipeline yields improved fairness without a significant loss of predictive accuracy, even when applied to a diverse dataset. To test this hypothesis, the study addresses the following research questions:

- RQ1: Can the integrated application of fairness-improving techniques across pre-processing, in-processing, and post-processing stages achieve comparable fairness and accuracy outcomes?
- RQ2: How does the effectiveness of various fairness metrics (e.g., Disparate Impact, Statistical Parity Difference, and Equal Opportunity Difference) vary when applying the integrated approach?
- RQ3: Which fairness-improving techniques or combinations are most effective in maintaining predictive accuracy?
- RQ4: Does the integration of fairness techniques across the AI pipeline enhance the predictive stability of the models?
- RQ5: What are the implications of validating fairness-enhanced models on new datasets for their real-world applicability in recidivism prediction?

2 Related works

The literature on fairness in machine learning underscores the urgent need to address biases in AI, especially in high-stakes domains like criminal justice system. Recent studies highlight the limitations of applying fairness-enhancing techniques at isolated stages of the AI development pipeline and advocate for a comprehensive, multi-stage approach to ensure fairness and maintain public trust in AI systems.

Farayola et al. (2023) provide a detailed examination of fairness-enhancing techniques applied within recidivism prediction models, noting the limitations of interventions focused on a single stage, whether pre-processing, in-processing, or post-processing. They argue that isolated interventions address only specific biases and do not comprehensively tackle fairness throughout the AI pipeline. Their findings support the need for multi-phase interventions, a key motivation for this study, to address fairness challenges at multiple stages of model development and build stakeholder trust in judicial AI applications.

Similarly, Caton and Haas (2024) provide a broad overview of fairness definitions, metrics, and mitigation strategies in their survey on machine learning fairness. They identify a persistent trade-off between fairness and accuracy, noting that isolated fairness interventions, whether at the pre-processing, in-processing, or post-processing stage, tend to yield suboptimal results by compromising either fairness or predictive accuracy. Caton et al. emphasize the importance of integrated fairness strategies to better balance these metrics, especially in high-stakes applications where both fairness and predictive performance are critical. Their survey informs this study's multi-stage approach, where combining techniques across all development stages aims to create a fairer and more accurate model for recidivism prediction.

Chen et al. (2023) present an empirical evaluation of various fairness-improving methods, demonstrating that single-phase interventions often struggle to balance multiple fairness metrics, such as disparate impact and statistical parity. By contrast, they show that integrated fairness approaches can enhance performance across multiple metrics without significantly compromising accuracy. Their findings underscore the practical advantages of adopting multi-phase interventions across the AI pipeline to achieve more consistent and equitable model performance; however, their work is limited to benchmarking 17 representative bias mitigation methods on eight widely adopted software decision tasks. The study systematically evaluates trade-offs between fairness and machine learning performance but does not explore fairness interventions specific to recidivism prediction which are the focus of our research.

Zhang and Sun (2022) propose a fairness framework that incorporates causal analysis to refine fairness interventions. They argue that causal relationships between features and outcomes, often overlooked in traditional fairness methods, are vital to mitigating bias effectively. By leveraging causal insights, their framework achieves a more nuanced mitigation of bias across multiple stages of development. This work reinforces the importance of a comprehensive fairness strategy, particularly in sensitive domains like criminal justice, where the integration of fairness techniques across all pipeline phases is essential for robust and equitable outcomes.

Kozodoi et al. (2022) examine fairness interventions in the financial sector, particularly credit scoring. Their analysis, grounded in the separation criterion, explores the trade-offs between fairness, accuracy, and profitability across various fairness-enhancing methods. Using multi-objective optimization, they identify non-dominated solutions that achieve balanced fairness and profitability, illustrating a practical application of the Pareto front in fairness optimization. While focused on profitability, their work contributes to a broader understanding of fairness trade-offs. In contrast, our study emphasizes a comprehensive fairness framework for recidivism prediction, a domain where diverse fairness metrics and stakeholder trust are crucial and profitability is not the primary consideration.

Raza et al. (2023) explore the intersection of fairness in AI and equity in healthcare, underscoring the need for multi-stage fairness interventions in sensitive domains. Through a case study on diabetic retinopathy prediction, they demonstrate that fairness techniques applied solely at one phase (e.g., pre-processing) are insufficient to address deep-seated inequities in healthcare data. By applying fairness interventions across pre-, in-, and post-processing stages, they achieve significant improvements in fairness and health equity. This finding supports the broader applicability of a multi-phase fairness strategy, validating its importance across critical domains beyond criminal justice.

Wang et al. (2024) extend the scope of fairness research by introducing a fairness-enhancing approach for large language models that integrates structural fairness mechanisms with active learning. A distinguishing aspect of their work lies in its exploration of the trade-offs between fairness and accuracy in large-scale models, offering valuable insights into how active learning can be employed to enhance structural fairness. While their study centers on language models, the integration of fairness-aware layers and iterative optimization resonates with the multi-phase approach of this study. Their method, which integrates fairness-aware loss functions with active learning for informed data selection, provides a scalable framework for ongoing fairness improvement, relevant to high-stakes AI applications where fairness is critical.

Chakraborty et al. (2020) take a two-stage approach, integrating pre-processing and in-processing fairness techniques to mitigate bias while maintaining model performance. By leveraging multi-objective optimization, their approach balances fairness and accuracy in fairness-sensitive datasets like COMPAS. However, their method does not include post-processing interventions, which are essential for refining predictions and addressing fairness issues post-deployment. Moreover, their approach lacks configurations that combine pre- and post-processing techniques (PP models) or in- and post-processing techniques (IP models), limiting its scope. In contrast, this study expands the methodology to integrate fairness-enhancing techniques across all three phases, pre-processing, in-processing, and post-processing, addressing bias in a more comprehensive manner and improving the fairness–accuracy trade-off in recidivism prediction.

Farayola et al. (2024) propose a fairness framework for recidivism prediction models by integrating fairness techniques across pre-processing, in-processing, and post-processing stages. Their study employs multi-objective optimization to balance fairness and accuracy. Key methods include Reweighing, Disparate Impact Remover, Adversarial Learning, Exponential Gradient Reduction, Reject Option-Based Classification, and Equalized Odds Optimization. The findings suggest that the combination of Disparate Impact Remover, Adversarial Learning, and Equalized Odds Optimization (`DIR+AL+EO`) yields notable fairness improvements with minimal accuracy loss. This multi-stage approach aligns with our work by demonstrating the effectiveness of integrated fairness interventions, particularly when guided by optimization frameworks to identify balanced, non-dominated models. However, a key limitation of their study is its reliance on a single dataset, the COMPAS dataset, for evaluating fairness-improving combinations. This constraint limits the generalizability of their approach, as datasets vary in attributes and inherent biases. In contrast, our study applies the integrated fairness-improving approach to the RisCanvi dataset, broadening the evaluation and enhancing both its applicability and potential generalizability across diverse recidivism prediction contexts.

These studies collectively highlight the importance of an integrated fairness approach across the AI development pipeline. Building on this foundation, our research aims to validate the robustness and generalizability of a multi-stage fairness framework by applying it to a new dataset, RisCanvi dataset. This approach provides a replicable model that seeks to enhance fairness while preserving accuracy, contributing to a broader body of knowledge focused on equitable outcomes in the criminal justice system when predicting recidivism risk, where fairness and predictive performance must be carefully balanced to maintain public trust.

3 Background

This section provides an overview of fairness in AI and the contemporary methodologies employed to improve fairness.

3.1 Fairness in AI

The issue of fairness and the methods to ensure its implementation in the design of AI systems, particularly for binary classification tasks, has garnered significant attention, since concerns about the equity of AI systems first arose. To address these challenges, researchers have proposed various metrics to measure and assess fairness in AI. These measures typically fall in two primary categories: individual fairness and group fairness.

Individual fairness emphasizes that individuals in similar situations should receive comparable treatment and outcomes from an algorithm, regardless of differences in protected attributes. This principle requires that individuals who are alike in relevant aspects are treated equally (Beutel et al. 2019). Notable examples of individual fairness include the Generalized Entropy Index and counterfactual fairness (Caton and Haas 2023.

Group fairness evaluates the outcomes of a classification algorithm across two or more protected groups, such as racial categories, to ensure parity based on statistical metrics. Unlike individual fairness, this paper focuses on group fairness and considers established metrics in the literature to evaluate the fairness of AI models used to predict recidivism risk.

3.2 Fairness metrics

This section outlines significant group fairness metrics, emphasizing their attributes and relevance to assessing fairness across demographic groups in AI models predicting recidivism risk (Chouldechova 2017; Mehrabi et al. 2021; Kleinberg et al. 2016; Pleiss et al. 2017). Many studies tend to focus on specific fairness metrics without considering the broader implications or effects of alternative metrics on different demographic groups. To address this limitation, our integrative approach incorporates multiple fairness metrics at various stages of AI system development. This comprehensive strategy enhances fairness assessments across multiple dimensions while addressing the trade-offs inherent in balancing different fairness metrics.

The literature has explored multiple fairness metrics during a single stage of AI development, often revealing conflicts between metrics (Chouldechova 2017; Mehrabi et al. 2021; Kleinberg et al. 2016; Pleiss et al. 2017). Our study builds upon this foundation by integrating several fairness-enhancing techniques and evaluating their impact

on key fairness metrics. Here, $G = g_i$ and $G = g_j$ denote the privileged and unprivileged groups, respectively.

Next, we will define one-by-one the fairness metrics of interest in the context of the study.

- **Statistical/demographic parity difference (SPD):** Statistical or demographic parity ensures that all groups defined by a protected attribute, such as race or gender, have an equal likelihood of receiving a positive classification from a model. Let $\hat{y} \in \{0, 1\}$ represent predictions and g_i and g_j represent groups based on a sensitive attribute G. Parity is achieved when the probability $P(\hat{y} = 1)$ is not influenced by G (Gao et al. 2022). This can be expressed as follows:

$$P(\hat{y} = 1 \mid G = g_i) - P(\hat{y} = 1 \mid G = g_j).$$

A value of zero indicates no disparity in positive outcomes across groups.
- **Disparate Impact (DI):** Disparate Impact measures the ratio of positive classification probabilities between privileged and unprivileged groups. It aligns with legal fairness principles, such as the 80% rule, which identifies discrimination when the ratio falls below 0.8. To ensure fairness, the ratio should ideally be close to 1 (Jo et al. 2023). The formula is as follows:

$$\frac{P(\hat{y} = 1 \mid G = g_i)}{P(\hat{y} = 1 \mid G = g_j)}.$$

- **Equal Opportunity Difference (EOD):** Equal Opportunity Difference evaluates fairness by comparing true positive rates (TPR) across groups. It ensures that qualified individuals in all groups have an equal chance of being correctly classified as positive (Beutel et al. 2019). The difference is calculated as follows:

$$P(\hat{y} = 1 \mid y = 1,\ G = g_i) - P(\hat{y} = 1 \mid y = 1,\ G = g_j).$$

A score of zero represents perfect fairness, while deviations indicate biases favoring or disadvantaging certain groups.
- **Predictive Equality Difference (PED):** Predictive Equality Difference measures the disparity in false positive rates (FPR) across groups, ensuring that no demographic group is disproportionately affected by incorrect positive predictions (Castelnovo et al. 2022). The metric is defined as follows:

$$P(\hat{y} = 1 \mid y = 0,\ G = g_i) - P(\hat{y} = 1 \mid y = 0,\ G = g_j).$$

A PED of zero indicates fairness, while deviations suggest unequal treatment among groups.

3.3 Fairness–metrics trade-off

Fairness metrics often conflict, requiring trade-offs when optimizing multiple objectives simultaneously Kleinberg et al. (2016); Chouldechova (2017); Chen et al. (2023). Key trade-offs include the following:

- **Statistical Parity vs. Equal Opportunity:** Statistical Parity ensures equal positive outcomes across groups but does not consider whether individuals are qualified. Equal opportunity ensures fairness for qualified individuals (true positives). Optimizing for Statistical Parity might inadvertently favor higher-risk individuals in some groups, violating Equal Opportunity.
- **Equal Opportunity vs. Predictive Equality:** Equal Opportunity seeks to equalize true-positive rates, while Predictive Equality focuses on equal false-positive rates. Improving one can worsen the other, as emphasizing true positives for qualified individuals may increase false positives.
- **Disparate Impact vs. Equal Opportunity:** Disparate Impact ensures proportional positive outcomes across groups, potentially conflicting with Equal Opportunity's focus on treating qualified individuals fairly. Balancing these may reduce fairness for those who clearly qualify or disqualify for certain outcomes.

These trade-offs highlight the complexity of achieving fairness in machine learning. In high-stakes contexts like criminal justice, stakeholders often prioritize metrics such as Predictive Equality to reduce unjust detentions, even at the expense of true-positive rate parity (Equal Opportunity).

3.4 Fairness-improving methods

Several fairness-enhancing techniques are designed to operate at specific phases of the AI development pipeline: pre-processing, in-processing, or post-processing (Farayola et al. 2023). In this study, we selected six fairness techniques: Reweighing, Disparate Impact Remover, Exponentiated Gradient Reduction, Adversarial Learning, Equalized Odds Optimization, and Reject Option-Based Classification. These techniques were chosen based on their empirical performance and comparative analyses, as identified in Zhang and Sun (2022).

Each method was selected for its capacity to improve at least one fairness metric widely used in the literature, including Statistical Parity Difference (SPD), Disparate Impact (DI), Equal Opportunity Difference (EOD), and Predictive Equality Difference (PED). Moreover, as detailed by Bellamy et al. (2019), Reweighing and Disparate Impact Remover are especially effective at reducing SPD and improving DI, making them well suited for correcting disparities in dataset composition. Conversely, Exponentiated Gradient Reduction and Adversarial Learning focus on fairness during the training process and have demonstrated strong performance in mitigating classification error imbalances. Post-processing methods such as Equalized Odds Optimization and Reject Option-Based Classification are designed to mitigate bias at the evaluation stage of AI development outputs, thereby further aligning EOD and PED across privileged and unprivileged groups.

Importantly, no single method can comprehensively address all fairness concerns across the pipeline. Hence, the integration of these techniques is not arbitrary but reflects a deliberate design to leverage their complementary strengths.

- **Reweighing (Re):** A pre-processing method that assigns weights to correct biases in sensitive attributes while preserving the original data. Reweighing ensures balanced representation across groups and targets metrics like statistical parity and disparate impact (Bellamy et al. 2019; Krasanakis et al. 2018).
- **Disparate Impact Remover (DIR):** DIR modifies sensitive attribute values to achieve equitable decision distributions. It is a pre-processing method designed to address disparate impact while preserving relative rankings (Feldman et al. 2015).
- **Exponential Gradient Reduction (EGR):** An in-processing technique that creates a randomized classifier by solving fairness constraints as cost-sensitive classification tasks. EGR targets metrics like demographic parity and Equal Opportunity Difference (Agarwal et al. 2018).
- **Adversarial Learning (AL):** This in-processing technique trains models to minimize biases by introducing adversarial examples. AL improves fairness by reducing the influence of sensitive attributes on predictions (Wadsworth et al. 2018).
- **Reject Option-Based Classification (ROC):** A post-processing technique that adjusts predictions near decision boundaries to reduce discrimination. ROC targets metrics, such as statistical parity and Equal Opportunity Difference (Kamiran et al. 2012).
- **Equalized Odds Optimization (EO):** EO uses linear programming to modify labels, aligning predictions with fairness criteria. It focuses on metrics like Equal Opportunity Difference and Predictive Equality Difference (Pleiss et al. 2017).

By applying these methods across all stages of the pipeline, this study aims to mitigate the trade-offs between fairness metrics and accuracy, offering a more holistic approach to fairness in machine learning.

4 Methodology

This section outlines the empirical study designed to assess how integrating different fairness-enhancing methods influences fairness outcomes in AI systems. The study focuses on the criminal justice use case, more specifically, recidivism prediction, and uses two datasets described next.

4.1 Datasets

4.1.1 COMPAS dataset

The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) dataset serves as the foundation for this study (Angwin et al. 2016). It comprises 6172 records from Broward County, Florida, capturing criminal history details across 11 features. The target variable, 2-year recidivism, indicates whether an individual reoffended within 2 years of release. For this analysis, a subset of 5278 records was used, focusing on individuals identified as African–American or Caucasian. Race was treated as the protected attribute, with Caucasians designated as the privileged group and African–Americans as the unprivileged group. Within the subset, 2795 records represented non-recidivists (label 0), and 2483 corresponded to recidivists (label 1), revealing a slight imbalance in class distribution.

Categorical variables were transformed into numeric formats to ensure compatibility with fairness models. Age was encoded as 0 (“Less than 25”), 1 (“25–45”), and 2 (“Greater than 45”). Recidivism risk text was mapped to 0 ("Low"), 1 (“Medium”), and 2 ("High"), while gender was encoded as 0 (“Female”) and 1 (“Male”). For consistency, race was standardized as 0 (African–American) and 1 (Caucasian), and charge severity was represented as 0 ("Misdemeanor") and 1 ("Felony"). Random sampling techniques were employed to address class imbalance, resulting in a more balanced dataset to facilitate fair evaluations across demographic groups.

4.1.2 RisCanvi dataset

The anonymized RisCanvi dataset used in this research is a carefully curated resource derived from the fifth installment in a longitudinal research series investigating recidivism among prison populations in Catalonia (CEJFE 2020).

Over 3 decades, this series has examined the post-release trajectories of incarcerated individuals, with the latest dataset focusing on those released in 2015 and followed for up to 5 years, ending in December 2019. In accordance with international recommendations, data collection concluded before March 2020 to prevent potential biases introduced by the COVID-19 pandemic.

This dataset was curated to ensure accuracy, completeness, and relevance for fairness-aware machine learning research. It includes N = 3,814 individuals, categorized into final releases (67.4%), conditional releases (30.7%), suspended sentences (1.9%), and expulsion cases (3.1%). Expulsion cases were excluded from recidivism monitoring, as these individuals do not reenter the penitentiary system. Recidivism is defined as penitentiary recidivism, involving new entries into the penitentiary system, either pre-trial or post-conviction. Additionally, the dataset supports focused studies on specific subpopulations, including individuals expelled from Spanish territory.

The RisCanvi dataset is geographically specific to Catalonia, an autonomous region of Spain, offering tailored insights into its penitentiary and judicial systems. This resource enables in-depth analysis of recidivism trends and the effectiveness of targeted interventions, providing critical support for advancing evidence-based policy and criminal justice practices within the region.

On average, inmates who undergo only one evaluation spend approximately 3 months in prison, whereas those with four evaluations remain incarcerated for an average of 2 years before being released on parole or regaining their freedom. Notably, evaluations cease once an inmate is released, ensuring that recidivism tracking accurately reflects post-release outcomes.

As part of the curation process, systematic data cleaning and pre-processing techniques were applied to address missing values. Missing data were imputed using corresponding values from either the previous or subsequent evaluations. For items with a yes/no/uncertain response format, missing values were replaced with "uncertain." Cases with irreparable missing data were excluded from the analysis, resulting in a final dataset of 2885 individuals.

The final class distribution consists of 2391 individuals who did not commit penitentiary recidivism and 464 individuals who did. To mitigate class imbalance, random sampling techniques were applied, creating a more balanced dataset for fair evaluations across demographic groups.

To facilitate accessibility for English-speaking researchers, the dataset and its documentation, originally in Spanish, were translated into English. Furthermore, the dataset has been structured to ensure ease of use for researchers across different disciplines, with clearly defined variable descriptions and standardized formats for seamless integration into various analytical workflows.

This curated dataset is publicly available for further research and reproducibility at the following repository: RisCanvi GitHub Repository.

4.2 Binary label dataset transformation

To assess fairness using AIF360's suite of fairness metrics (Bellamy et al. 2019), the datasets were converted into Binary Label Datasets via AIF360's 'BinaryLabelDataset' class. This conversion standardized the data, ensuring compatibility with fairness algorithms and enabling the computation of metrics, such as statistical parity difference, disparate impact, predictive equality difference, and equal opportunity difference. Once transformed, the Binary Label Datasets served as inputs for pre-processing, in-processing, and post-processing fairness interventions.

4.3 Experimental setup

The study explores the integration of fairness-enhancing methods at various stages of the AI pipeline: pre-processing, in-processing, and post-processing (see Fig. 1). Several configurations were implemented, including PI (Pre-Processing + In-Processing), PP (Pre-Processing + Post-Processing), IP (In-Processing + Post-Processing), and PIP (Pre-Processing + In-Processing + Post-Processing). These setups allowed for a comprehensive analysis of how the inclusion of fairness techniques at different pipeline stages impacted fairness outcomes while preserving predictive accuracy.

Pre-processing techniques, namely Reweighing (Re) and Disparate Impact Remover (DIR), modified the training data to address bias before model training. In-processing methods such as Adversarial Learning (AL) and Exponentiated Gradient Reduction (EGR) enforced fairness constraints during training. Post-processing techniques, including Equalized Odds Optimization (EO) and Reject Option-Based Classification (ROC), adjusted model predictions after training to ensure fairness in outcomes.

To enhance the reliability of the results, each experiment was repeated ten times using different random seeds.

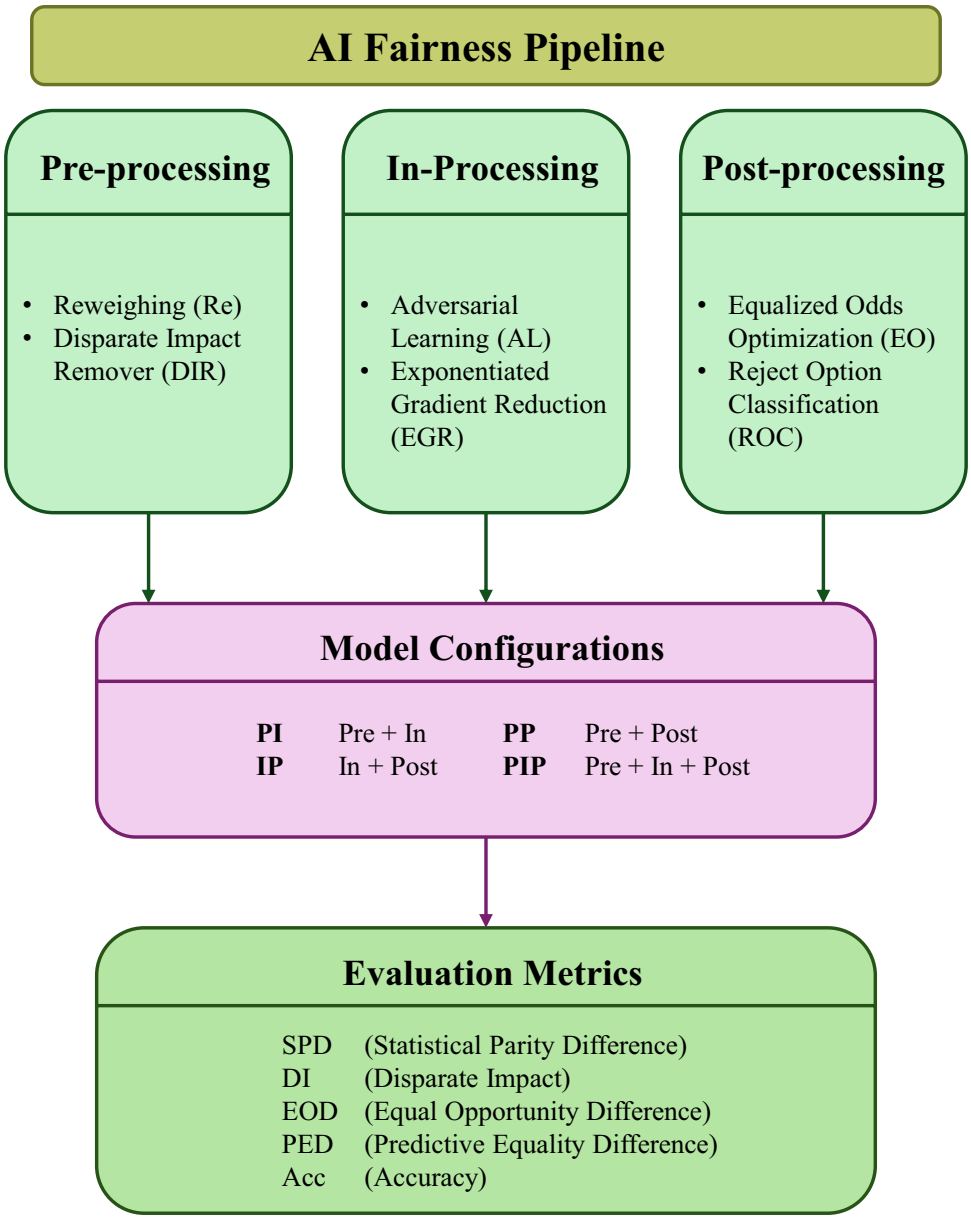

Fig. 1 Fairness-enhancing techniques integrated across the AI fairness pipeline

This repetition minimized the influence of randomness, providing consistent and reproducible insights into the effectiveness of fairness-enhancing techniques.

4.4 Handling of pre-processing techniques

Pre-processing methods aimed to reduce bias in the dataset before the training phase.

4.4.1 Reweighing (Re)

The reweighing method assigned weights to instances in the dataset, balancing the representation of privileged and unprivileged groups. By adjusting instance weights according to the protected attribute (race), it ensured equitable treatment of both groups during model training.

4.4.2 Disparate impact remover (DIR)

The Disparate Impact Remover adjusted feature values to reduce bias, promoting fair treatment of both demographic groups. A repair level of 1 was used, signifying complete bias correction based on race. The modified dataset was subsequently utilized for training models in the next stages.

4.5 Handling of in-processing techniques

In-processing methods integrated fairness constraints directly into the model training process.

4.5.1 Adversarial learning (AL)

Adversarial Learning incorporated a secondary adversarial network tasked with predicting the protected attribute (i.e., race), while the primary classifier focused on predicting recidivism. The adversarial network actively guided the primary classifier to avoid discriminatory patterns. The model was trained for 50 epochs using a multi-layer neural network with a batch size of 128. The weight parameter for adversary loss was set to 0.2, and the debias parameter was enabled to ensure fairness.

4.5.2 Exponentiated gradient reduction (EGR)

Exponentiated Gradient Reduction enforced fairness constraints by adjusting model parameters to optimize both fairness and accuracy. Logistic Regression served as the base classifier, and demographic parity was applied as the fairness constraint. The protected attribute (race) was retained in the dataset for evaluation, with training configured to a maximum of 20 iterations for consistency.

4.6 Handling of post-processing techniques

Post-processing methods adjusted predictions to ensure fairness in the final model outcomes.

4.6.1 Equalized odds optimization (EO)

Equalized Odds adjusted prediction labels to align true-positive and false-positive rates across privileged and unprivileged groups. This ensured balanced error rates and equitable outcomes for both groups.

4.6.2 Reject option-based classification (ROC)

Reject Option-Based Classification modified predictions for individuals near the decision boundary, prioritizing fairness for unprivileged individuals. A grid search optimized classification thresholds, using equal opportunity difference and average odds difference as guiding metrics to balance fairness and accuracy.

4.7 Model integration and metric computations

The fairness-enhancing techniques, pre-processing (Re, DIR), in-processing (AL, EGR), and post-processing (EO, ROC), were combined to develop fairness-enhanced models (PI, PP, IP, PIP). These models were evaluated using metrics such as Statistical Parity Difference (SPD), Disparate Impact (DI), Equal Opportunity Difference (EOD), and Predictive Equality Difference (PED), alongside accuracy metrics. This comprehensive evaluation captured the trade-off between fairness and predictive performance for each method.

4.8 Experimental evaluation

The datasets were split into a 70:30 train-test ratio, ensuring stratified sampling for proportional representation of the privileged and unprivileged groups. The same random state was used across all experiments to maintain consistency and comparability. Each model was trained and evaluated under identical conditions, with the Binary Label Dataset structure ensuring uniform fairness assessments. Predictions were analyzed using both fairness and accuracy metrics to assess the effectiveness of each configuration in promoting fairness while preserving predictive accuracy.

5 Multiple objective optimization framework

5.1 Balancing fairness and accuracy

Selecting the optimal model from the PI, IP, PP, and PIP configurations requires a careful equilibrium between fairness metrics and predictive accuracy. This process is challenging due to inherent trade-offs between these objectives and the absence of a universally accepted fairness metric. To address this complexity, a multi-objective optimization framework is adopted, enabling the identification of fairness-enhancing integrations that reconcile these competing priorities effectively.

5.2 Multi-objective approach and the pareto front

The multi-objective optimization framework resolves the fairness–accuracy trade-off by addressing diverse objectives without collapsing them into a singular metric. This study emphasizes identifying "best" approaches, which are characterized as non-dominated solutions within the set of PI, IP, PP, and PIP models. Non-dominated solutions excel on at least one fairness metric without being outperformed across all others. This approach leverages the Pareto front concept (Sharma and Kumar 2022), which systematically

evaluates trade-offs and identifies models that best balance competing objectives.

By employing this strategy, the study circumvents the limitations of single-objective optimization. Models situated on the Pareto front represent viable options that balance fairness and accuracy effectively, ensuring their applicability across diverse contexts.

5.3 Formulating the objectives

The multi-objective framework defines accuracy as a maximization objective to capture the predictive capability of models. Fairness metrics, on the other hand, require optimization toward specific target values. For Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD), and Predictive Equality Difference (PED), the optimal value is zero, signifying reduced disparities across demographic groups. For Disparate Impact (DI), the ideal value is one, indicating equitable treatment between privileged and unprivileged groups.

To unify these objectives within a single optimization framework, the study employs transformations. Absolute value transformations are applied to SPD, EOD, and PED, normalizing them around zero, while an inverse transformation adjusts DI values below one toward its target. This standardization ensures coherence and comparability across fairness objectives, enabling seamless integration with accuracy optimization.

5.4 Non-dominated solutions and pareto front analysis

The evaluation of non-dominated solutions involves assessing their performance across multiple objectives. Let $O = [O_1(x), \ldots, O_k(x)]$ represent the vector of k objective values (e.g., Accuracy, SPD, DI, EOD, PED) for each technique x in the set of techniques. A technique x_1 dominates another technique x_2, written as $x_1 \succ x_2$, if $\forall i \in \{1, \ldots, k\}, O_i(x_1) \geq O_i(x_2)$ and $\exists j \in \{1, \ldots, k\}$ such that $O_j(x_1) > O_j(x_2)$. A technique is non-dominated if no other technique outperforms it across all objectives. The set of all

Table 1 Many-objective optimization. Star (*) indicates a standardized value for minimization. Non-dominated approaches for both dataset and best metric result are in bold

Model	Riscanvi					COMPAS				
	SPD	DI	EOD	PED	Acc	SPD	DI	EOD	PED	Acc
NONE	0.059	1.224*	0.011	0.045	0.72	0.25	2.13	0.025	0.18	0.68
Re	0.046	1.155*	0.011	0.039	0.71	0.097	1.243	0.094	0.031	0.66
DIR	0.079	1.288*	0.056	0.067	0.71	0.284	2.092	0.281	0.211	0.67
EGR	0.082	1.274*	0.135	0.083	0.71	0.036	1.093	0.080	0.088	0.67
AL	0.064	1.237	0.059	0.055	0.71	0.053	1.174	0.115	0.085	0.66
EO	0.016	1.048*	0.033	0.004	0.70	0.040	1.125	0.014	0.015	0.64
ROC	0.079	1.288*	0.056	0.066	0.71	0.270	2.100	0.287	0.183	0.68
Re+EGR	0.058	1.189	0.068	0.067	0.71	0.033	1.086*	0.026	0.032	0.66
DIR+EGR	0.135	1.536*	0.169	0.142	0.73	0.055	1.142*	0.038	0.004	0.67
Re+AL	0.064	1.238*	0.059	0.055	0.71	0.061	1.168	0.052	0.004	0.67
DIR+AL	0.076	1.297	0.0820	0.0640	0.72	0.009	1.020*	0.028	0.067	0.67
Re+EO	0.010	1.035*	0.020	0.002	0.71	0.034	1.095	0.001	0.002	0.64
Re+ROC	0.086	1.340	0.092	0.075	0.71	0.081	1.230	0.091	0.002	0.67
DIR+EO	0.010	1.037*	0.007	0.0001	0.74	0.032	1.097	0.007	0.005	0.64
DIR+ROC	0.023	1.086	0.028	0.011	0.74	0.288	2.160	0.300	0.207	0.68
EGR+EO	0.009	1.024*	0.008	0.001	0.67	0.036	1.095	0.003	0.003	0.65
EGR+ROC	0.021	1.066	0.011	0.034	0.71	0.023	1.059	0.015	0.042	0.67
AL+EO	0.011	1.036*	0.015	0.0002	0.71	0.036	1.099	0.014	0.016	0.66
AL+ROC	0.024	1.092*	0.028	0.013	0.74	0.079	1.159	0.037	0.033	0.68
Re+EGR+EO	0.005	1.014*	0.033	0.010	0.69	0.034	1.088	0.001	0.0002	0.65
Re+EGR+ROC	0.021	1.066	0.011	0.034	0.71	0.023	1.059	0.015	0.042	0.67
Re+AL+ROC	0.024	1.092*	0.028	0.013	0.74	0.079	1.159	0.037	0.033	0.68
Re+AL+EO	0.011	1.036*	0.015	0.0002	0.71	0.036	1.099	0.014	0.016	0.66
DIR+EGR+ROC	0.080	1.314	0.076	0.093	0.74	0.055	1.142	0.038	0.004	0.67
DIR+EGR+EO	0.007	1.021*	0.011	0.0003	0.70	0.037	1.097	0.002	0.003	0.66
DIR+AL+ROC	0.031	1.081*	0.0002	0.026	0.66	0.030	1.056	0.006	0.032	0.68
DIR+AL+EO	0.010	1.034*	0.006	0.002	0.72	0.038	1.089	0.001	0.001	0.66

non-dominated techniques forms the Pareto front, wherein each solution represents an optimal balance that cannot be improved in one objective without compromising another.

Through Pareto front analysis, this framework identifies models that achieve an effective equilibrium between fairness and accuracy. These insights form a robust foundation for selecting models suitable for real-world deployment. The framework ensures that selected models not only address fairness comprehensively, but also maintain strong predictive performance, aligning ethical considerations with practical utility in fairness-enhanced AI systems.

Table 1 presents the results of a many-objective optimization evaluation, comparing the performance of various fairness-enhancing techniques, individually and in combination, across two datasets: *RisCanvi* and *COMPAS*. Each model is evaluated using four widely recognized group fairness metrics: Statistical Parity Difference (SPD), Disparate Impact (DI), Equal Opportunity Difference (EOD), and Predictive Equality Difference (PED), along with overall classification accuracy (Acc). For fairness metrics, lower values indicate better fairness performance, with Disparate Impact (DI) being standardized for minimization (as indicated by a star '*').

The table is organized to reflect parallel evaluations across the two datasets. Within each dataset column block, fairness metrics are presented in the same order for ease of comparison. Models are sorted by technique type, including baseline (NONE), single interventions (e.g., Re, DIR), and multi-intervention combinations (e.g., Re + EGR + EO). Bold values highlight non-dominated approaches, models that achieved optimal or near-optimal trade-offs across multiple metrics without being outperformed on all objectives. This allows for quick identification of models that maintain strong fairness across several dimensions while preserving predictive accuracy. The inclusion of multiple combinations emphasizes the importance of composite fairness strategies, as no single technique consistently performs best across all metrics or datasets.

5.5 Implementation of many-objective optimization

In our research, we simultaneously optimize five objectives—Accuracy (Acc), Statistical Parity Difference (SPD), Disparate Impact (DI), Equal Opportunity Difference (EOD), and Predictive Equality Difference (PED)—using many-objective optimization along the Pareto front to achieve a balance across these diverse objectives within a unified framework. Table 1 presents our findings, identifying four optimal integrations of fairness-enhancing techniques across PI (pre-processing and in-processing), IP (in-processing and post-processing), PP (pre-processing and post-processing), and PIP (pre-processing, in-processing, and post-processing) models for the COMPAS dataset: `Re+EGR+EO`, `DIR+EGR+EO`, `DIR+AL+ROC`, and `DIR+AL+EO`. For the RisCanvi Dataset: `DIR+EO` and `DIR+AL+EO`. Significantly, the "NONE" model, which does not include any fairness interventions, does not appear on the Pareto front, emphasizing the benefits of integrating fairness-enhancing techniques to improve both fairness and accuracy.

5.6 Bi-objective optimization for targeted fairness metrics

Bi-objective optimization focuses on balancing two objectives, typically aiming to achieve a trade-off between a prioritized fairness metric and model accuracy. This approach is particularly useful in high-stakes domains, where decision-makers may emphasize specific fairness considerations alongside the need for predictive reliability.

When focusing on optimizing Statistical Parity Difference or Disparate Impact or equal opportunity difference alongside accuracy for RisCanvi dataset, models like `AL+ROC`, `DIR+EO`, `DIR+ROC`, `DIR+AL+EO`, and `Re+AL+ROC` stand out as optimal choices. For Predictive Equality and accuracy, four models emerge as optimal: `AL+ROC`, `DIR+EO`, `DIR+ROC`, and `Re+AL+ROC`.

When focusing on optimizing Statistical Parity Difference or Disparate Impact alongside accuracy for COMPAS dataset, models like `DIR+AL`, `EGR+ROC`, and `Re+EGR+ROC` stand out as optimal choices. Likewise, when prioritizing Equal Opportunity Difference in combination with accuracy, the non-dominated solutions include `DIR+EGR+EO`, `DIR+AL+EO`, and `DIR+AL+ROC`. For Predictive Equality and accuracy, five models emerge as optimal: "Re+AL," `DIR+EGR`, `DIR+EGR+ROC`, `DIR+AL+EO`, and `DIR+EGR+EO`.

6 Discussion

This section examines the integration of fairness-enhancing techniques to achieve a balance between fairness and predictive accuracy in recidivism prediction models. While single-phase fairness methods often have limitations, the integration of techniques across pre-processing, in-processing, and post-processing stages demonstrates promising improvements. A preliminary assessment of individual fairness techniques reveals their varying strengths and trade-offs, as summarized in Table 1.

6.1 Integrations enhancing fairness and accuracy

RQ1: Can the integrated application of fairness-improving techniques across pre-processing, in-processing, and post-processing stages achieve comparable fairness and accuracy outcomes?

The results demonstrate that integrated approaches consistently outperform individual fairness-enhancing techniques in achieving a balanced trade-off between fairness and predictive accuracy. Several combinations of methods achieved substantial improvements across multiple fairness metrics without significantly compromising classification accuracy.

For instance, on the RisCanvi dataset, the integration of `DIR+EGR+EO` resulted in an SPD of 0.007, a DI of 1.021, an EOD of 0.011, and a PED of 0.0003, with an accuracy of 0.70. This represents a substantial improvement over the baseline model (NONE), which had an SPD of 0.059, DI of 1.224, an EOD of 0.011, and a PED of 0.045 while achieving higher accuracy (0.72) but offering minimal fairness guarantees. Another competitive integration, Re+EGR+EO, yielded even lower SPD (0.005) and DI (1.014), with only a minor drop in accuracy to 0.69.

On the COMPAS dataset, similar patterns emerged. The model DIR+AL+EO achieved exceptional fairness outcomes with an SPD of 0.038, DI of 1.089, a near-perfect EOD of 0.001, and a PED of 0.001, while maintaining an accuracy of 0.66. Compared to the baseline, which recorded an SPD of 0.25, DI of 2.13, EOD of 0.025, PED of 0.18, and an Acc of 0.68, the integrated approaches reduced disparities by significant margins across all group fairness dimensions.

Top-performing models that combined fairness techniques across all three pipeline stages, such as `Re+EGR+EO` and `DIR+AL+EO`, consistently appeared as non-dominated solutions, achieving balanced improvements across all fairness metrics. In contrast, single techniques like EGR, while effective on specific metrics (e.g., DI = 1.093 on COMPAS), showed weaker performance on others (e.g., EOD = 0.080, PED = 0.088), highlighting the advantage of multi-stage integration.

These results confirm that no single technique is sufficient to address the complex and multidimensional nature of algorithmic bias. Instead, integrated models allow complementary strengths to emerge leading to broader fairness improvements. These results validate the hypothesis that integrated fairness techniques are effective in addressing bias comprehensively and can yield comparable outcomes across datasets. The findings underscore the value of a multi-phase approach in improving both fairness and accuracy, demonstrating its applicability to real-world scenarios.

6.2 Impact of dataset characteristics on metrics

RQ2: How does the effectiveness of various fairness metrics (e.g., Disparate Impact, Statistical Parity Difference, and Equal Opportunity Difference) vary when applying the integrated approach?

The dataset-specific impact of fairness techniques is evident in the performance variability of key metrics across RisCanvi and COMPAS datasets. Metrics, such as DI and SPD, showed greater variability, reflecting the influence of dataset characteristics on fairness outcomes. For instance, `Re+EO` achieved a DI of 1.035 on RisCanvi but a slightly higher DI of 1.095 on COMPAS, indicating that DI may be more sensitive to the underlying data distribution. Conversely, metrics like PED and EOD exhibited more consistent behavior across datasets, suggesting that they may be more reliable in evaluating fairness across different contexts. These observations highlight the importance of selecting fairness metrics that align with the specific characteristics and fairness objectives of the dataset. The findings also emphasize the need for practitioners to account for dataset variability when evaluating fairness-enhancing techniques, ensuring that models are tailored to the nuances of the data in real-world applications.

6.3 Effectiveness of techniques in maintaining accuracy

RQ3: Which fairness-improving techniques or combinations are most effective in maintaining predictive accuracy?

Integrated fairness techniques demonstrate better performance in maintaining accuracy compared to individual methods, particularly on new datasets. For example, on the RisCanvi dataset, DIR+AL+EO achieved an accuracy of 0.72 while significantly improving fairness metrics such as SPD (0.010) and DI (1.034). On the COMPAS dataset, combinations like EGR+EO and Re+EGR preserved accuracy levels of 0.65–0.67 while effectively reducing disparities across fairness metrics. In contrast, individual techniques often lead to greater trade-offs between fairness and accuracy. For instance, while EO alone achieved a DI of 1.048 on RisCanvi, its accuracy dropped to 0.70. The results suggest that combining techniques across the AI pipeline not only mitigates fairness disparities but also reduces the likelihood of substantial accuracy losses, making integrated

approaches more effective for maintaining performance on new datasets.

6.4 Predictive stability across datasets

RQ4: Does the integration of fairness techniques across the AI pipeline enhance the predictive stability of the models?

The integration of fairness techniques contributes to greater predictive stability across datasets, ensuring consistent performance even in diverse contexts. For instance, combinations like `DIR+AL+EO` and `Re+EGR+EO` exhibited similar trends in fairness and accuracy on both RisCanvi and COMPAS datasets, with only marginal differences in key metrics. This stability contrasts with the variability observed in individual techniques, where performance often fluctuated significantly between datasets. For example, DIR alone achieved an SPD of 0.079 on RisCanvi but a higher SPD of 0.284 on COMPAS, reflecting inconsistencies in its effectiveness. The robustness of integrated approaches stems from their ability to address fairness concerns comprehensively across the AI pipeline, mitigating the compounding effects of bias at different stages. These findings underscore the importance of adopting holistic strategies for fairness mitigation, particularly in high-stakes applications where predictive stability is critical for trust and reliability.

6.5 Real-world applicability and implications

RQ5: What are the implications of validating fairness-enhanced models on new datasets for their real-world applicability in recidivism prediction

The validation of fairness-enhanced models on diverse datasets such as RisCanvi and COMPAS reinforces their practical applicability in real-world contexts. The results demonstrate that integrated fairness techniques are adaptable to varying data distributions, making them more reliable for deployment in sensitive domains like criminal justice. For instance, combinations like `DIR+AL+EO` achieved strong fairness outcomes across both datasets, highlighting their generalizability. However, the dataset-specific variability observed in metrics like DI and SPD emphasizes the need for careful calibration of fairness techniques to align with the characteristics of the data and the fairness objectives of the application. These findings suggest that real-world deployment should prioritize integrated approaches that balance fairness and accuracy while being tailored to specific datasets. This ensures that fairness-enhanced models can address systemic biases effectively while maintaining stakeholder trust and confidence in their predictions.

7 Why do integrated fairness-enhancing techniques produce different results across datasets?

In the course of our analysis, a crucial question emerged repeatedly: why do different integrated fairness-enhancing techniques produce varying results? The variability in results arises from a complex interplay between dataset-specific characteristics, the design objectives of the fairness techniques, and the sensitivity of the fairness metrics employed. By examining the outcomes from both the COMPAS and RisCanvi datasets, we identify five key factors that can contribute to these differences.

7.1 Dataset-specific biases

Each dataset encodes distinct systemic and historical biases. COMPAS, for example, exhibits stronger correlations between demographic attributes (e.g., race) and the target variable (i.e., recidivism), leading to pronounced disparities in metrics like Disparate Impact (DI) and Statistical Parity Difference (SPD). Techniques like Disparate Impact Remover (DIR) are more effective on COMPAS due to its imbalanced feature distributions but show reduced impact on RisCanvi, where biases are subtler and less correlated with features. This highlights the importance of tailoring fairness techniques to the dataset's specific biases.

7.2 Mechanisms of fairness techniques

Different fairness techniques target different dimensions of bias, resulting in complementary, but dataset-dependent effects. For instance: pre-processing techniques (e.g., Reweighing, DIR) modify input data to balance group representation, leading to strong improvements in SPD and DI. However, their impact on metrics like Equal Opportunity Difference (EOD) may be limited by dataset-specific error distributions. In-processing techniques (e.g., Adversarial Learning, Exponential Gradient Reduction) enforce fairness constraints during training, excelling in metrics like Predictive Equality (PE) but being sensitive to the strength of correlations between features and sensitive attributes. Post-processing techniques (e.g., Equalized Odds Optimization) adjust decision thresholds to refine fairness in predictions, showing consistent improvements in PE and EOD but being less effective on datasets with residual feature imbalances.

7.3 Metric sensitivity to dataset properties

Fairness metrics capture different aspects of bias, and their sensitivity varies across datasets. For example: DI

and SPD, which focus on group-level disparities, are more sensitive to feature and demographic distribution, leading to variability between datasets. Likewise, EOD and PED, which evaluate error rate disparities, are influenced by both datasets bias and model generalization, making them more stable across datasets. Techniques like EGR+EO consistently improved EOD and PED on both datasets, highlighting their robustness.

7.4 Interactions in integrated techniques

Combining fairness techniques creates cumulative or competing effects depending on dataset characteristics. For instance: Re+EGR+EO showed strong synergy across both datasets, reducing DI, SPD, and EOD while maintaining accuracy. This is because reweighing balances the dataset by adding weights, EGR optimizes fairness constraints during training, and EO refines predictions. DIR+AL+EO, while effective on COMPAS, had diminished synergy on RisCanvi potentially due to weaker feature-label correlations, which limited the adversarial component's impact.

7.5 Trade-offs between fairness and accuracy

Integrated techniques must balance fairness improvements with accuracy preservation. Some combinations, such as Re+EGR+EO, achieved minimal accuracy loss while improving fairness across all metrics. Others, like DIR+AL+EO, demonstrated slightly reduced accuracy on RisCanvi, thus reflecting the cost of fairness and accuracy trade-offs when mitigating biases in the models.

8 Potential implications of the research findings

The findings of this study have far-reaching implications for society, the general public, and policymakers, particularly in deploying AI systems in high-stakes domains such as the criminal justice system for predicting the risk of recidivism. This research contributes to advancing fairness and trust in AI-driven decision-making by addressing systemic biases and improving fairness in recidivism prediction models.

For society, integrating fairness-enhancing techniques into AI systems is a crucial step toward promoting equity and mitigating systemic biases. Recidivism prediction models that fail to address fairness risk perpetuating cycles of inequality, disproportionately impacting marginalized communities. The findings of this study highlight that fairness-enhanced models can help disrupt these cycles by producing more equitable outcomes across demographic groups. In doing so, they contribute to fostering a society where AI-driven decisions align with democratic principles and ethical values, ensuring that individuals are treated fairly, regardless of their background.

Similarly, public trust in AI systems depends on their ability to deliver fair and unbiased outcomes, particularly in high-stakes applications such as predicting recidivism risk in the criminal justice system. By demonstrating the effectiveness of integrated fairness techniques across multiple datasets, this research provides evidence that AI systems can be designed to uphold fairness and equity while maintaining predictive reliability. This assurance has the potential to empower the public to trust and engage with AI technologies, fostering broader societal acceptance of their deployment in critical domains.

For policymakers, the findings underscore the importance of designing regulations and standards that prioritize fairness in AI systems. The variability in results across datasets highlights the need for policies that mandate fairness evaluations across diverse data contexts to ensure that AI models generalize effectively and do not perpetuate existing inequities. Additionally, this research provides a foundation for developing standardized frameworks for fairness metrics and mitigation techniques, enabling consistent and transparent evaluation of AI systems in practice. Policymakers can leverage these insights to establish guidelines for the responsible deployment of AI, ensuring that systems align with societal values and ethical principles.

Beyond the criminal justice system, the implications of this research extend to other high-stakes domains, such as healthcare, education, and employment, where algorithmic decisions significantly impact individuals and communities. The proposed fairness-enhancing framework offers a scalable and adaptable approach for addressing biases across various applications, providing a foundation for equitable and trustworthy AI systems in diverse contexts. Moreover, the emphasis on multi-dataset validation and fairness–accuracy trade-offs highlights the need for rigorous evaluation protocols to assess the generalizability and robustness of fairness-enhancing techniques.

Finally, this research demonstrates that fairness in AI systems is a technical challenge and an ethical imperative. Fairness-enhancing techniques play a pivotal role in advancing societal progress by addressing systemic inequities and fostering trust in AI-driven decision-making. The findings presented here provide actionable insights for researchers, practitioners, and policymakers, contributing to developing AI systems that align technological advancements with societal values of fairness, justice, and trust.

9 Potential challenges to the validity of the research

While our study demonstrates the effectiveness of integrated fairness-enhancing techniques in improving algorithmic fairness and balancing predictive accuracy, a few potential challenges to the validity of our research must be acknowledged to contextualize our findings and guide future research.

9.1 Limited dataset scope

This study focuses on two datasets, COMPAS and RisCanvi, both of which, while relevant and used in recidivism prediction research, may not capture the full range of demographic, legal, or geographic variability present in real-world criminal justice systems. The generalizability of our findings to other jurisdictions remains an open question. Future research should incorporate additional datasets with varying cultural, legal, and socio-political contexts to test the robustness and adaptability of integrated fairness approaches.

9.2 Absence of qualitative and stakeholder insights

This study adopts a quantitative approach centered on measurable fairness metrics and predictive accuracy. However, algorithmic fairness is inherently multidimensional, encompassing legal, ethical, and societal perspectives that may not be captured by metrics alone. The exclusion of insights from legal professionals, ethicists, and directly affected communities limits our ability to assess the practical implications and social acceptability of integrated fairness-enhancing interventions. Future work should incorporate participatory design methods and expert consultation to align fairness outcomes with stakeholder expectations and legal standards.

9.3 Static fairness evaluation

Our evaluation occurs at a single point in time based on historical data. We do not account for potential feedback loops that arise when AI systems influence future data (e.g., recidivism predictions affecting sentencing decisions, which then affect future data distributions). This static view overlooks the dynamic and evolving nature of fairness in deployed systems. Future work should explore longitudinal and causal analyses to understand how fairness interventions behave over time and under deployment.

9.4 Fairness techniques and metrics

This study focuses on six fairness-enhancing techniques and evaluates their impact using group fairness metrics, Statistical Parity Difference (SPD), Disparate Impact (DI), Equal Opportunity Difference (EOD), and Predictive Equality Difference (PED), which are commonly used to quantify disparities between demographic groups. While effective at capturing population-level bias, these metrics may obscure unfair treatment at the individual or subgroup level. Furthermore, fairness metrics can conflict with one another, making it difficult to achieve uniformly fair outcomes across all dimensions.

A further potential challenge to the validity of our research lies in the selection of a specific subset of fairness-enhancing techniques. Although the chosen methods were selected for their empirical performance and coverage across different stages of the AI pipeline, they do not represent the full range of available approaches. As the field of algorithmic fairness continues to evolve, future research should explore alternative fairness techniques to better account for context-sensitive and nuanced manifestations of bias.

10 Conclusion

The integration of fairness-enhancing techniques in AI systems holds significant promise for mitigating biases in recidivism prediction models within the criminal justice system. This study evaluates the performance of integrated techniques across pre-processing, in-processing, and post-processing stages using two recidivism datasets, COMPAS and RisCanvi, and employs a multi-objective optimization framework using the Pareto Front to select models that effectively balance fairness and predictive accuracy. The findings demonstrate that integrated approaches often outperform individual methods, yielding substantial improvements in fairness metrics, including Statistical Parity Difference (SPD), Disparate Impact (DI), Equal Opportunity Difference (EOD), and Predictive Equality Difference (PED), while maintaining minimal accuracy trade-offs. Integrated approaches, such as the combination of Reweighing, Exponential Gradient Reduction, and Equalized Odds Optimization (`Re+EGR+EO`) and as well as Disparate Impact Reduction and Adversarial Learning (`DIR+AL+EO`) exhibited robust performance across datasets, underscoring their potential for real-world deployment.

However, while the results indicate that integration can enhance fairness, performance varied significantly across datasets. The effectiveness of fairness-enhancing strategies is highly dependent on dataset characteristics, feature distributions, and demographic representation. For example, COMPAS benefited more from methods addressing

feature imbalances, whereas RisCanvi responded better to interventions targeting error distributions. These findings underscore the need for dataset-aware fairness strategies to improve generalizability and reliability across contexts.

Limitations and future work: Several limitations of this study also highlight promising directions for future research. First, the evaluation was conducted using only two datasets, COMPAS and RisCanvi, which may limit the generalizability of the results. Future work should include a broader range of datasets with diverse legal, cultural, and demographic characteristics to assess the robustness of integrated fairness-enhancing interventions. Second, this study adopted a purely quantitative approach, without incorporating the perspectives of key stakeholders, such as legal professionals, ethicists, and policymakers. Including these voices through participatory design and interdisciplinary consultation will help ensure that technical definitions of fairness align with real-world values and legal norms. Third, while this work evaluated a carefully selected subset of fairness-enhancing techniques, it did not explore the full spectrum of available methods. Future research should investigate alternative fairness strategies.

Finally, future studies should systematically investigate when and why integrated fairness approaches fail to produce expected improvements and explore the conditions under which integration may introduce new or unintended biases. Developing dynamic, data-aware fairness mechanisms and understanding their evolution across the AI pipeline will help advance more robust and adaptable solutions.

This work contributes to the field of AI fairness by providing a holistic framework, demonstrating the value of multi-dataset validation, and addressing fairness–accuracy trade-offs in a structured manner. Future research should extend this framework to other domains, incorporate real-time fairness adjustments, and assess the long-term societal impacts of fairness-aware AI. Achieving fairness in AI requires balancing technical innovation with ethical responsibility. By aligning AI systems with principles of inclusion, transparency, and justice, this work takes a step toward building more trustworthy and equitable technologies.

Section title of first appendix

An appendix contains supplementary information that is not an essential part of the text itself but which may be helpful in providing a more comprehensive understanding of the research problem or it is information that is too cumbersome to be included in the body of the paper.

Acknowledgements This work was supported in part by the Taighde Éireann—Research Ireland under Grant Nos. 13/RC/2094_P2 (Lero) and 13/RC/2106_P2 (ADAPT) and is co-funded under the European Regional Development Fund (ERDF).

Funding Open Access funding provided by the IReL Consortium.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., Wallach, H.: A reductions approach to fair classification. In: International Conference on Machine Learning, pp. 60–69 (2018). PMLR

AI, H.: High-level expert group on artificial intelligence. European Commission. Available at: https://ec.europa.eu/digital-single (2019)

Angwin, J., Larson, J., Mattu, S., Kirchner, L.: How We Analyzed the COMPAS Recidivism Algorithm (2016). https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Berk R, Bleich J (2014) Forecasts of violence to inform sentencing decisions. J Quant Criminol 30(1):79–96. https://doi.org/10.1007/s10940-013-9195-0

Beutel A, Chen J, Doshi T, Qian H, Woodruff A, Luu C, Kreitmann P, Bischof J, Chi EH (2019) Putting fairness principles into practice: Challenges, metrics, and improvements. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 453–459

Bellamy RK, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, Lohia P, Martino J, Mehta S, Mojsilović A et al (2019) Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM J Res Dev 63(4/5):1–4

Berk RA, Kuchibhotla AK, Tchetgen Tchetgen E (2023) Fair risk algorithms. Annu Rev Stat Appl 10:165–187

Biswas A, Mukherjee S (2021) Ensuring fairness under prior probability shifts. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 414–424. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3461702.3462596

Castelnovo A, Crupi R, Greco G, Regoli D, Penco IG, Cosentini AC (2022) A clarification of the nuances in the fairness metrics landscape. Sci Rep 12(1):4209. https://doi.org/10.1038/s41598-022-07939-1

Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A (2017) Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp 797–806

Caton S, Haas C (2023) Fairness in machine learning: a survey. ACM Comput Surv 300–360 (2023) https://doi.org/10.1145/3616865

Caton S, Haas C (2024) Fairness in machine learning: A survey. ACM Comput Surv 56(7):1–38

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/10.48550/arXiv.1610.07524

Chakraborty J, Majumder S, Yu Z, Menzies T (2020) Fairway: a way to build fair ml software. In: Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, pp 654–665

Chen Z, Zhang JM, Sarro F, Harman M (2023) A comprehensive empirical study of bias mitigation methods for machine learning classifiers. ACM Trans Softw Eng Methodol 32(4):1–30

Formació Especialitzada (CEJFE), C.: Taxa de reincidència penitenciária 2020. Accessed: 2024-04-17 (2020). https://cejfe.gencat.cat/ca/recerca/opendata/presons/taxa-reincidencia-2020/index.html

Desmarais SL, Johnson KL, Singh JP (2016) Performance of recidivism risk assessment instruments in us correctional settings. Psychol Serv 13(3):206. https://doi.org/10.1037/ser0000075

Farber S (2024) Ai in terrorism sentencing: evaluating predictive accuracy and ethical implications. Crim Justice Stud 37(4):352–370. https://doi.org/10.1080/1478601X.2024.2415558

Ferrara E (2023) Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. Science 6(1):3. https://doi.org/10.3390/sci6010003

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

Farayola, M.M., Malika, B., Saber, T., Connolly, R., Tal, I.: Enhancing algorithmic fairness: Integrative approaches and multi-objective optimization application in recidivism models. In: Proceedings of the 19th International Conference on Availability, Reliability and Security, pp. 1–10 (2024). https://doi.org/10.1145/3664476.3669978

Farayola, M.M., Tal, I., Malika, B., Saber, T., Connolly, R.: Fairness of ai in predicting the risk of recidivism: Review and phase mapping of ai fairness techniques. In: Proceedings of the 18th International Conference on Availability, Reliability and Security, pp. 1–10 (2023). https://doi.org/10.1145/3600160.3605033

Green, B.: The false promise of risk assessments: epistemic reform and the limits of fairness. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 594–606 (2020). https://doi.org/10.1145/3351095.3372869

Graffam J, Shinkfield AJ, Lavelle B (2014) Recidivism among participants of an employment assistance program for prisoners and offenders. Int J Offender Ther Comp Criminol 58(3):348–363. https://doi.org/10.1177/0306624X12470526

Gao, X., Zhai, J., Ma, S., Shen, C., Chen, Y., Wang, Q.: Fairneuron: improving deep neural network fairness with adversary games on selective neurons. In: Proceedings of the 44th International Conference on Software Engineering, pp. 921–933 (2022). https://doi.org/10.1145/3510003.3510087

Jo, N., Aghaei, S., Benson, J., Gomez, A., Vayanos, P.: Learning optimal fair decision trees: Trade-offs between interpretability, fairness, and accuracy. In: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, pp. 181–192 (2023)

Jacobs LA, Gottlieb A (2020) The effect of housing circumstances on recidivism: Evidence from a sample of people on probation in san francisco. Crim Justice Behav 47(9):1097–1115. https://doi.org/10.1177/0093854820942285

Jain B, Huber M, Elmasri R, Fegaras L (2020) Using bias parity score to find feature-rich models with least relative bias. Technologies 8(4):68. https://doi.org/10.3390/technologies8040068

Jain, B., Huber, M., Fegaras, L., Elmasri, R.A.: Singular race models: addressing bias and accuracy in predicting prisoner recidivism. In: Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments, pp. 599–607 (2019). https://doi.org/10.1145/3316782.3322787

Kozodoi N, Jacob J, Lessmann S (2022) Fairness in credit scoring: Assessment, implementation and profit implications. Eur J Oper Res 297(3):1083–1094

Kamiran, F., Karim, A., Zhang, X.: Decision theory for discrimination-aware classification. In: 2012 IEEE 12th International Conference on Data Mining, pp. 924–929 (2012). https://doi.org/10.1109/ICDM.2012.45 . IEEE

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

Kobayashi, K., Nakao, Y.: One-vs.-one mitigation of intersectional bias: A general method to extend fairness-aware binary classification. arXiv preprint arXiv:2010.13494 (2020)

Krasanakis, E., Spyromitros-Xioufis, E., Papadopoulos, S., Kompatsiaris, Y.: Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In: Proceedings of the 2018 World Wide Web Conference, pp. 853–862 (2018). https://doi.org/10.1145/3178876.3186133

Liu S, Vicente LN (2022) Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach. CMS 19(3):513–537. https://doi.org/10.1007/s10287-022-00425-z

Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR) 54(6):1–35

Miron M, Tolan S, Gómez E, Castillo C (2021) Evaluating causes of algorithmic bias in juvenile criminal recidivism. Artificial Intelligence and Law 29(2):111–147. https://doi.org/10.1007/s10506-020-09268-y

Plecko, D., Bareinboim, E.: Fairness-accuracy trade-offs: A causal perspective. arXiv preprint arXiv:2405.15443 (2024)

Pelegrina, G.D., Couceiro, M., Duarte, L.T.: A statistical approach to detect sensitive features in a group fairness setting. arXiv preprint arXiv:2305.06994 (2023)

Pagano TP, Loureiro RB, Lisboa FV, Peixoto RM, Guimarães GA, Cruz GO, Araujo MM, Santos LL, Cruz MA, Oliveira EL et al (2023) 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 7(1):15. https://doi.org/10.3390/bdcc7010015

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., Weinberger, K.Q.: On fairness and calibration. Advances in neural information processing systems **30** (2017)

Raza, S., Pour, P.O., Bashir, S.R.: Fairness in machine learning meets with equity in healthcare. In: Proceedings of the AAAI Symposium Series, pp. 149–153 (2023)

Rodolfa, K.T., Salomon, E., Haynes, L., Mendieta, I.H., Larson, J., Ghani, R.: Case study: predictive fairness to reduce misdemeanor recidivism through social service interventions. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 142–153 (2020). https://doi.org/10.1145/3351095.3372863

Sharma S, Kumar V (2022) A comprehensive review on multi-objective optimization techniques: Past, present and future. Archives of Computational Methods in Engineering 29(7):5605–5633

Thiebes S, Lins S, Sunyaev A (2021) Trustworthy artificial intelligence. Electron Mark 31(2):447–464

Wang, C., Han, B., Patel, B., Rudin, C.: In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. Journal of Quantitative Criminology, 1–63 (2022) https://doi.org/10.1007/s10940-022-09545-w

Watamura, E., Liu, Y., Ioku, T.: Judges versus artificial intelligence in juror decision-making in criminal trials: Evidence from two preregistered experiments. PloS one 20(1), 0318486 (2025) https://doi.org/10.17605/OSF.IO/N6TM2

Wang, C., Li, J., Zhang, R.: A method to enhance structural fairness in large language models with active learning (2024)

Wadsworth, C., Vera, F., Piech, C.: Achieving fairness through adversarial learning: an application to recidivism prediction. arXiv preprint arXiv:1807.00199 (2018) https://doi.org/10.48550/arXiv.1807.00199

Wan M, Zha D, Liu N, Zou N (2023) In-processing modeling techniques for machine learning fairness: A survey. ACM Trans Knowl Discov Data 17(3):1–27

Zhang, M., Sun, J.: Adaptive fairness improvement based on causality analysis. In: Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, pp. 6–17 (2022). https://doi.org/10.1145/3540250.3549103